САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

СТРУКТУРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

Межвузовский сборник

Издается с 1978 года

Выпуск 11

Под редакцией А. С. Герда и И. С. Николаева

Издательство Санкт-Петербургского государственного университета

УДК 80+618.31
ББК 81.1
С83

Печатается по постановлению
Редакционно-издательского совета
филологического факультета
Санкт-Петербургского государственного университета

С83 **Структурная и прикладная лингвистика.** Вып. 11: межвуз. сб. / под ред. А. С. Герда и И. С. Николаева. — СПб.: Изд-во С.-Петерб. ун-та, 2015. — 304 с.

Сборник (вып. 10 вышел в 2014 г.) содержит статьи по широкому кругу проблем теоретической и прикладной лингвистики, по применению математических методов в языкознании.

Для специалистов по теории языка, прикладной и теоретической лингвистике.

ББК 81.1

ОТ РЕДАКЦИОННОЙ КОЛЛЕГИИ

В 2016 году мы отмечаем юбилеи выдающихся профессоров кафедры математической лингвистики Александра Сергеевича Герда и Григория Яковлевича Мартыненко.

А. С. Герд — известный специалист по прикладной лингвистике и лексикографии, редактор большого числа словарей. Более сорока лет он заведует кафедрой математической лингвистики и почти все это время является ответственным редактором межвузовского сборника «Структурная и прикладная лингвистика». Этот сборник, основанный Александром Сергеевичем в 1978 году, сегодня входит в список изданий РИНЦ и ВАК. В нем публикуют свои работы не только преподаватели, студенты и аспиранты отделения «Прикладная и математическая лингвистика» СПбГУ, но и наши коллеги из других университетов и научных организаций России.

Г. Я. Мартыненко — известный специалист по квантитативной лингвистике и стилеметрии, автор ряда монографий по числовой гармонии текста, семиотике и статистике. Он почти сорок лет работает на нашей кафедре и воспитал целое поколение ученых в области квантитативной лингвистики. Кроме того, Григорий Яковлевич — автор художественных книг, он ведет активную концертную деятельность как исполнитель классических арий и народных песен.

Члены редакционной коллегии, авторы настоящего сборника, преподаватели и сотрудники кафедры математической лингвистики, сердечно поздравляют Александра Сергеевича Герда и Григория Яковлевича Мартыненко, желают им крепкого здоровья и успешной плодотворной деятельности во всех сферах их интересов. Мы посвящаем 11-й выпуск сборника «Структурная и прикладная лингвистика» нашим юбилярам.

А. С. Герд

НЕРЕШЁННОЕ В МОДЕЛИРОВАНИИ ЛОГИКО-ПОНЯТИЙНЫХ СИСТЕМ[1]

Аннотация. Построение моделей знания — актуальная проблема ряда наук. В статье анализируется отражение проблем историзма, теории, индивидуальных концепций в структуре логико-понятийных систем.

Ключевые слова. Теория знания, онтология, моделирование.

Alexander S. Gerd

UNSOLVED PROBLEMS OF LOGICAL AND CONCEPTUAL SYSTEMS MODELING

Abstract. Knowledge modeling is considered to be one of the most challenging problems of various disciplines. The paper is an attempt to analyze the essential problems of historicism, theory and bespoke solutions in the framework of logical and conceptual systems.

Keywords. Knowledge theory, ontology, modeling.

Постановка вопроса о нерешенном в моделировании логико-понятийных систем (далее — ЛПС) вновь особенно ярко высвечивает старую проблему антиномии между чисто теоретическим языкознанием и прикладной лингвистикой в целом. В последний раз она наиболее явственно обозначилась на конференции «Диалог», среди докладов которой практически не было ни одного доклада по актуальным проблемам прикладной лингвистики [Компьютерная линг-

[1] Работа выполнена в рамках научно-исследовательского проекта РГНФ № 15-04-12055 «Структурная типология словарных статей в словарях русского языка разных типов и способы их формального представления в компьютерных системах».

вистика …]. Разумеется, в основе конкретной прикладной задачи всегда лежит та или иная теоретическая концепция.

Проблема моделирования человеческого знания с каждым днем становится все более актуальной. Если в 1988 г., когда мы публиковали свою статью [Герд, 1988], она представлялась еще достаточно одинокой и периферийной среди моря других вопросов, то сегодня эти проблемы вышли на первый план. Опубликована прекрасная обобщающая книга В. Ш. Рубашкина [Рубашкин]. В Москве в стенах Института русского языка им. В. В. Виноградова РАН под руководством С. В. Шелова регулярно проводится симпозиум «Терминология и знание» [Терминология и знание …].

Оговоримся сразу, что общие проблемы теории знания, когнитивного терминоведения, когнитологии и прикладные вопросы моделирования конкретных ЛПС — это не одно и то же [Табанакова].

Вряд ли кто-либо усомнится в том, что теория знания как такового принадлежит многим дисциплинам — от философии до биологии, — у каждой из которых к ней свое отношение.

В то же время моделирование ЛПС отдельных отраслей науки и техники — это каждый раз прикладная задача, так же как, например, и проектирование, и создание нового словаря или разработка новой базы данных.

Не повторяя сказанного и написанного нами ранее, отметим только те вопросы моделирования конкретных ЛПС как фрагментов специального научного и технического знания, которые представляются спорными и сегодня.

Одна из самых сложных проблем моделирования знания и конкретной ЛПС — это отражение историзма. Как правило, ЛПС строится как синхронически статичная система того или иного периода, чаще всего периода, современного для ее создателей и исследователей.

Возможен чисто формальный подход, когда модель знаний и ЛПС строятся, например, по векам; модель физики или географии XVIII века, отдельно — для XIX века и т. д. Однако такой чисто формальный подход чаще всего войдет в противоречие с этапами развития и истории той или иной науки.

Таким образом, более целесообразно сначала изучить и знать историю развития данного конкретного научного знания в данной стране. При этом следует учитывать, что концептуально наука

неделима и всегда исторически связана с развитием ее в других странах.

При таком подходе сначала выделяются основные этапы эволюции данной науки, а затем в соответствии с ними подбираются источники и строятся отдельные ЛПС для каждого периода.

Так, в ряде наших работ были выделены основные этапы, периоды развития ихтиологии в России и затем построены отдельные ЛПС для каждого периода [Герд, 1988; Герд, 2005].

Казалось бы, создать модель современного знания для совсем новой науки гораздо проще. Однако здесь наиболее трудный вопрос — где хронологически начинается современное знание и, следовательно, его терминология.

Таким образом, именно для научного и технического знания путь моделирования его ЛПС по отдельным хронологическим периодам в тесной связи с этапами истории развития данной науки — путь наиболее естественный и достоверный, хотя и здесь постоянно будет вставать вопрос о нижней границе того или иного периода. При всей значимости великих научных открытий большинство наук развиваются все же эволюционно, а не путем революционных скачков.

В то же время семантика и терминология любой современной фундаментальной науки и отрасли знания нередко заключают в себе множество компонентов, унаследованных от прошлого. Это и устаревшие термины, широко известные ранее, но употребляемые и сегодня, и термины, без которых нельзя воссоздать концепцию того или иного ученого.

И здесь мы переходим к другому, не менее сложному вопросу — об отражении идей, концепций русских ученых в модели знания. Более других об этом писал В. М. Лейчик [Лейчик].

Как и насколько возможно отразить в ЛПС единого научного знания разные мнения и гипотезы?

Обычно считается, что единая модель ЛПС основывается на знании наиболее авторитетном, общепринятом, стандартном. Но, как известно, в каждой науке есть разные подходы к отдельным проблемам.

Каким путем здесь идти: то ли строить модели знания отдельных ученых, то ли по отдельным проблемам, и как их вписать в общую модель того или иного периода?

Один из возможных путей: в единую модель вводятся условные маркеры, отсылающие пользователя к уже имеющимся, построенным ранее моделям отдельных ученых или проблем.

Именно с отражением разных концепций и теорий связан и гораздо более общий вопрос практики современного научного и технического перевода. Нередко в различных странах с одним и тем же особенно интернациональным по форме термином связывается разное содержание.

Таким образом, в той или иной национальной модели ЛПС какой-то узел в одной стране может быть заполнен, а в другой — вообще оставаться пустым.

Сказанное представляет особую опасность для корпусной лингвистики. Корпус специальных текстов, содержащий лексику таких текстов на том или ином языке, в котором нет ссылок на источники и точное значение термина в данном языке в данной стране, представляет большую опасность как «друг переводчика».

Только зная, что понимается под этим термином в том или ином языке в данной стране, можно более уверенно вести перевод. Как известно, под одним и тем же по форме термином в испанском языке в самой Испании и в Латинской Америке часто понимается разное.

Дело не в самом корпусе, а в его источниках и методах его построения.

На этапе выборки терминов из источников все время встает вопрос: а относится ли этот термин к данной науке? При этом мы часто забываем, что в наше время кристально чистых рафинированных наук не бывает.

Во всех отраслях знания идет постоянный процесс взаимовлияния, интеграции и смешения с данными близких и смежных наук. В каждой науке возникают новые направления, отрастают новые ветви.

Яркими примерами могут служить лингвистика и география.

Лингвистика — это и история языка, и этнография, и психология, и логика, и математика, и информатика, и социология, и биология.

В географии окончательно утвердились такие новые направления, как историческая география, культурная география, этногеография.

Исследования по языкознанию, географии переполнены терминами смежных наук. Вряд ли стоит говорить о таких целостных науках, как биохимия, геохимия.

Насколько и как современные методы моделирования научного знания и построения ЛПС учитывают эти процессы? Если и учитывают, то в весьма слабой степени.

Все сказанное свидетельствует о том, что в моделировании процессов научного знания и техники лучше идти вперед медленно поспешая, малыми объемами по отдельным тематическим группам понятий и отдельным лексическим группам терминов. В результате, может быть, мы не получим сразу глобальной общей системы научного знания, но зато будем иметь достаточно точные, глубокие и детальные практически полезные микросистемы отдельных ЛПС.

Литература

Герд А. С. Логико-понятийное моделирование терминосистем и машинный фонд русского языка // Отраслевая терминология и ее структурно-типологическое описание: межвуз. сб. науч. тр. Воронеж, 1988.

Герд А. С. Прикладная лингвистика. СПб., 2005.

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21). М., 2015.

Лейчик В. М. Терминоведение. М., 2006.

Рубашкин В. Ш. Онтологическая семантика. Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М., 2012.

Табанакова В. Д. Моделирование научно-исследовательского текста. Тюмень, 2004.

Терминология и знание. Материалы IV Международного симпозиума (Москва, 6–8 июня 2014 г.) / отв. ред. С. Д. Шелов. М., 2014.

Г. Я. Мартыненко

СТИЛЕМЕТРИЯ: ВОЗНИКНОВЕНИЕ И СТАНОВЛЕНИЕ В КОНТЕКСТЕ МЕЖДИСЦИПЛИНАРНОГО ВЗАИМОДЕЙСТВИЯ

Часть 2. Первая половина XX века: расширение междисциплинарных контактов стилеметрии

К 100-летию выхода в свет статьи Николая Александровича Морозова «Лингвистические спектры: Средство для отличия плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд»

Аннотация. Данная работа является продолжением статьи, опубликованной в предыдущем выпуске сборника «Структурная и прикладная лингвистика». В первой половине XX века расширилась география стилеметрии, основанной В. Диттенбергером, обогатилась и ее методическая база и междисциплинарные контакты. Стилеметрия пришла в Россию благодаря усилиям Н. А. Морозова. В работах А. А. Чупрова, А. А. Маркова, Андрея Белого, в успехах русской формальной школы, в математических достижениях английской биометрической школы существенно расширился круг филологических проблем, решаемых с помощью математических методов. Появились новые метрические дисциплины, в частности возникла искусствометрия — сначала при поиске кульминации в композиции вербального и музыкального текста с помощью золотого сечения, а затем при поиске критериев измерения эстетичности текста.

Ключевые слова. Стилеметрия, теория совокупности, теория устойчивости, цепи Маркова, Н. А. Морозов, русская формальная школа, динамика текста, квантитативная лингвистика, композиция текста, эстетические измерения, искусствометрия.

Gregory Ya. Martynenko

STYLOMETRY: EMERGENCY AND EVOLUTION IN CONTEXT OF INTERDISCIPLINARY INTERACTION

Part II. The First Half of the 20th Century: The Expansion of Interdisciplinary Contacts

Abstract. This article continues a historical review on the emergence and development of stylometrics which was published in the previous volume of "Structural and Applied Linguistics". Being founded by Wilhelm Dittenberger, stylometrics expanded its geography in the first half of 20th century and enriched its methodological framework and interdisciplinary contacts. In Russia stylometrics appeared due to endeavours by N. A. Morozov. Later, the works by A. A. Chuprov, A. A. Markov, Andrei Bely, as well as the successes of Russian formal school and mathematical achievements of English biometrical school significantly expanded the range of philological problems solved by means of mathematical methods. Later, there emerged new metric disciplines. In particular, art-metrics appeared first in the process of searching for culmination in verbal and musical text composition with help of the golden section. And then it was further developed in the search for criteria for measuring text aesthetics.

Keywords. Stylometrics, population theory, stability theory, Markov chains, Nikolai Morozov, Russian formalist school, the dynamics of the text, quantitative linguistics, text composition, aesthetic measurements, art-metrics.

Начало XX века характеризуется беспрецедентно радикальными сдвигами в области научного и художественного творчества. Речь идет об отходе от классических схем, переоценке ценностей, декадансе, обновлении художественного и научного языка, становлении новых и даже экстравагантных научных парадигм, возникновении различных форм модернизма и авангардизма.

Уходило в прошлое и традиционное представление о гармонии природы и человеческого существования, идея единения с природой и сопричастности к вечным ценностям. Разброд и шатание, бунт, ниспровержение всего и вся, неприятие всего затхлого и омертвевшего.

Первая четверть XX века — это сложный и бурный период в истории европейской культуры с колоссальным приливом творческой энергии и поиском новых путей во всех областях искусства. Всеми цветами радуги переливался нескончаемый поток сталкивающихся противоречивых идей. В этих условиях формировалась какая-то новая гармония, уникальный и парадоксальный сплав течений, школ, манер, не вмещавшихся в традиционные рамки реализма, импрессионизма, романтизма и прочих течений.

Экспансия идей и методов естественных наук и математики в гуманитарные науки и искусство в это время была тотальной.

Вдогонку за антропометрией, биометрией, психометрикой, стилеметрией, эконометрией, которые зародились в XIX веке, устремились социометрия, наукометрия, библиометрия, искусствометрия, информетрия, документометрия, клиометрика, текстометрия (а к концу века — медиаметрия и даже киберметрия).

Но на фоне этой первой, основной тенденции проявила активность и вторая — экспансия гуманитарных наук и искусства в естественные науки, то есть наметилась тенденция гуманизации естественных наук.

Возникают обширные сферы междисциплинарной деятельности. Это теория систем и системный анализ, математическое моделирование, синергетика, социодинамика и теория ценозов.

Начало XX века — время формирования фундаментальных учений, впитавших в себя достижения предыдущего столетия и развивших эти достижения с привлечением новых идей и познавательных принципов. Остановимся на основных научных направлениях, которые впитывала стилеметрия для решения собственных задач.

1. Теория статистической совокупности

В первой части данной статьи [Мартыненко, 2014] внимание было акцентировано на возникновение, благодаря усилиям Рюмелина и Мебиуса, предпосылок для создания общей теории сообществ. В России эти идеи были поддержаны выдающимся лесоводом Георгием Федоровичем Морозовым (1867–1920), который создал единое учение о лесе как биогеоценотическом, географическом, социальном и историческом явлении, образующем единый природный комплекс.

Вслед за Морозовым и другими исследователями выдающийся русский статистик А. А. Чупров (1874–1926) построил теорию статистической совокупности, которая впоследствии, уже в конце XX века, была «подключена» к общей теории систем и общей теории ценозов (ценологии) [Чупров, 1910]. Чупров подчеркивал, что его теория исключительно междисциплинарна, она пронизывает и естественные, и общественные науки. «Давно ли обществоведение с увлечением играло понятием организма? А ныне ботаника и зоо-

логия оперируют понятиями общества и сообщества, и вместе с тем в область науки об органической природе проникают те же формы переработки эмпирического материала, каким пользуется статистика в обществоведении» [Чупров, 1914]. Суть концепции Чупрова заключается в следующем. Отталкиваясь от противопоставления общих и собирательных понятий, он вслед за Рюмелином приходит к заключению, что статистика проявляет интерес к собирательным совокупностям, локализованным в определенных рамках времени и пространства. В качестве такой совокупности может выступать, в частности, текст, представляющий собой целостное образование собирательного типа, состоящее из единиц, не обязательно однородных.

Еще один камень в здание теории сообществ заложил русский писатель и стиховед Юрий Николаевич Тынянов [Тынянов], выдвинувший идею системности художественной литературы, в которой он различал синхронические и диахронические литературно-художественные системы. Под синхронической системой Тынянов понимал совокупность всех художественных произведений данной эпохи, представленных разными жанрами и авторами. Диахроническая система по Тынянову — это последовательность синхронических систем, в которой формируется эволюция художественной литературы. Тынянов сетовал на то, что усилия филологов направлены преимущественно в сторону изучения выдающихся авторов и выдающихся произведений, ее ядра. При этом периферия литературы, ее литературный быт, в котором рождаются новые тенденции, остается вне внимания исследователей. Таким образом, в трудах Тынянова теория литературы и словесность в целом обогатились не только системными представлениями, но и идеей ядра и периферии. Эта идея вполне согласуется с теорией элиты В. Парето.

Идеи Тынянова активно разделял Роман Якобсон. Конкретизируя концепцию Тынянова, он подчеркивал, что синхроническое описание литературной продукции данной эпохи предполагает и обращение той части литературной традиции, которая в данную эпоху сохраняет жизненность [Якобсон].

2. Теория устойчивости статистических рядов

В начале XX века были продолжены теоретико-статистические исследования в области устойчивости статистических распределений, начатые Лексисом [Мартыненко, 2014]. А. А. Чупров, обратившись к лингвистическому материалу, пересмотрел некоторые выводы Лексиса. Вопреки Лексису русскому ученому удалось показать, что в явлениях действительной жизни ряды с «патологической» устойчивостью, не предусмотренной теорией вероятностей, встречаются достаточно часто. Примечательно, что этот факт был им установлен на текстовом материале: в серии выборок из произведений Гёте, в распределении сильных и слабых слогов в гекзаметрах Овидия и Вергилия, а также букв в строках научных трудов самого Лексиса [Чупров, 1910, с. 281]. Эти выводы Чупрова породили в дальнейшем активную переписку с академиком А. А. Марковым, в которой они обсуждали теорию дисперсии Лексиса. В конечном итоге это привело к рождению эпохальной теории. Она возникла в результате проведения следующего эксперимента. Марков наблюдал за порядком следования гласных и согласных букв в тексте романа «Евгений Онегин». Успехом испытания он считал гласность буквы (неудачей — согласность). Осуществив 10 000 испытаний, он разбил их на 100 серий по 100 последовательных букв в каждой и подсчитал коэффициент дисперсии Лексиса. Оказалось, что число гласных среди 100 последовательных букв русского текста гораздо более стабильно, чем было бы в том случае, если бы А. С. Пушкин просто выбирал буквы путем независимых случайных испытаний, соблюдая вероятности их появления, характерные для русского текста [Марков, 1913]. Полученная цепь испытаний была названа цепью Маркова. Это было важнейшим достижением математики, возникшим при решении филологических проблем.

3. Андрей Белый и математическое стиховедение

Серьезное воздействие на гуманитарную мысль начала XX века и последующих лет оказал русский писатель, поэт и ученый Андрей Белый (Борис Николаевич Бугаев; 1880–1934). Особенно велик его вклад в стиховедение. Андрей Белый — пионер в организации крупномасштабных статистических исследований русской поэзии,

явившихся источником идей и стимулом для многих поколений стиховедов. Именно он поставил проблемы, которые не теряют актуальности и в наши дни. Можно, наверное, сказать, что он является основоположником математического стиховедения (стихометрики).

А. Белый впервые определил задачи научного исследования стиха и убедительно продемонстрировал это на практике многоаспектным описанием ритмики четырехстопного ямба у 34 поэтов. Он впервые поставил вопрос о сходстве поэтов по ритмическим компонентам. «Он сдвинул изучение русского стихосложения с мертвой точки, — писал впоследствии В. М. Жирмунский о Белом, — сосредоточив внимание исследователей не на однообразных и абстрактных метрических схемах, а на живом многообразии реального ритма русского стиха, отклоняющегося в различных направлениях от той или иной из указанных схем» [Жирмунский].

Но Белый был не только кабинетным ученым. Он был великим ученым практиком. Он организовал кружок исследователей стиха, итогом работы которого явился «Учебник ритма», к сожалению, оставшийся неопубликованным.

На фоне масштабных достижений Белого в науке о стихе значительно скромнее выглядят его успехи в области изучения художественной прозы. Здесь можно прежде всего отметить его выдающееся исследование — книгу «Мастерство Гоголя» [Белый, 1934] — его лебединую песню, в которой он, помимо ритма, «расшевелил» многие доселе неведомые пласты стиля на материале произведений Гоголя, Сологуба, самого Белого, Блока, Маяковского. Его заветной мечтой было построение совокупности словарей, которые бы представляли разнообразие и единство русской литературы.

Не следует забывать, что Борис Николаевич сам был выдающимся писателем, одним из самых ярких представителей орнаментальной прозы, новатором и экспериментатором в искусстве слова.

Научные достижения Белого высоко ценились не только в филологической среде. Так, А. А. Чупров, говоря о применении статистических методов в гуманитарных науках, отмечает, что в книге «Символизм» [Белый, 1910] «в интересно задуманных и искусно выполненных статистических исследованиях… восходит к общей характеристике ритма поэта на основании взятых на выдержку „порций“ его стихов и поднимается затем до картины ритма разных эпох русского стихосложения (ритмический перелом у Жуковского)»

[Чупров, 1910]. Мы позволили себе привести эту цитату не только потому, что статистические опыты Белого были высоко оценены выдающимся статистиком-современником, но и потому, что филологам эта оценка, по-видимому, неизвестна.

4. Русская формальная школа

Математическое стиховедение Андрея Белого и статистические штудии Чупрова и Маркова были тем источником, из которого черпали свое вдохновение представители великой русской формальной школы, сложившейся в промежутке между 1910 и 1920 гг. В 1916 г. русские формалисты объединились в Общество изучения поэтического языка (ОПОЯЗ). В этот кружок вошли историки и теоретики литературы (В. Б. Шкловский, Б. М. Эйхенбаум, Ю. Н. Тынянов), лингвисты (Р. О. Якобсон, Е. Д. Поливанов, Л. П. Якубинский), стиховеды (С. И. Бернштейн, О. М. Брик).

Эта группа без всяких оговорок превратила литературоведение в науку, и притом науку мирового значения, создав условия для возникновения пражской школы структурной лингвистики, тартуско-московской школы структурной лингвистики и всего европейского структурализма в целом.

Идейным вдохновителем формальной школы был Виктор Борисович Шкловский. История формальной школы начинается с его ранних статей «Воскрешение слова» [Шкловский, 1914] и «Искусство как прием» [Шкловский, 1925], в которых резко критиковался подход к искусству как к «системе образов» и выдвигался тезис об искусстве как совокупности приемов. Это и было названо формальным методом.

Формализм поначалу был очень шумным течением, так как развивался параллельно с русским футуризмом и являлся разновидностью авангарда, но в науке. Предельно точно охарактеризовав сущность формального метода, Борис Викторович Томашевский отмечает, что: «При изучении явлений вовсе не нужно априорного определения сущностей. Важно различать их проявления и осознавать их связи. Такому изучению литературы посвящают свои труды формалисты. Именно как науку, изучающую явления литературы, а не ее „сущность“, мыслят они поэтику» [Томашевский]. И еще одно меткое замечание о междисциплинарных связях такого метода: «Да,

формалисты мечтают о создании специфической науки о литературе, науки, связанной с примыкающими к литературе отраслями человеческих знаний» [Там же].

Круг тем и интересов формалистов был огромен. Они уделяли большое внимание динамике текста. Построив теорию сюжета, занимались изучением ритма и синтаксиса стиха с помощью статистических методов, изучали звуковой, синтаксический и лексический повторы, создавали справочники стихотворных размеров, подчеркивали важность системного изучения художественной литературы и ее эволюции.

5. Проблема спорного авторства (стилеметрии) в России

В конце XIX и начале XX века пионерские работы немецкого филолога В. Диттенбергера [Dittenberger] были продолжены многими последователями [Zeller; Čada; Ritter].

В России первым стилеметром был Николай Александрович Морозов (1854–1946) — революционер-народник, разносторонний ученый, распространявший методы естественных наук на гуманитарные дисциплины, включая словесность. В статье «Лингвистические спектры: Средство для отличия плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд» Николай Александрович [Морозов] первым в России поставил задачу решения спорного авторства на основе статистического метода. Задачу атрибуции текста он сформулировал на основе обобщения опыта предшественников и указал на ряд методических трудностей, ожидающих исследователя при решении этой сложной задачи.

Отметим некоторые положительные черты работы Морозова.

1. Вслед за Диттенбергером Николай Александрович полагал, что служебные части речи или, как он говорил, «распорядительные частицы», не зависят от тематики текста, а определяются исключительно «складом» (стилем) речи говорящего.

2. При идентификации (или различении) текстов Морозов опирался не только на служебные, но и на знаменательные части речи. Использовал он также и лексические приметы: термины, диалектную и политическую лексику, редкие и необычные слова, а также синонимические дуплеты (*так как — потому что, потом — затем, полагать — считать* и т. п.).

3. Морозов атрибутировал не только древние тексты, но и современные.

4. При сравнении текстов ученый использовал, не только средние величины, но и их колебания в разных произведениях. Такие структуры Морозов называл лингвистическими спектрами. Их можно рассматривать как «прообразы» статистических распределений.

Приведем несколько примеров таких спектров, заимствованных из работы Морозова (рис. 1).

Выбор текстов Морозовым достаточно произволен. При замене одних текстов на другие (частично или полностью) картина

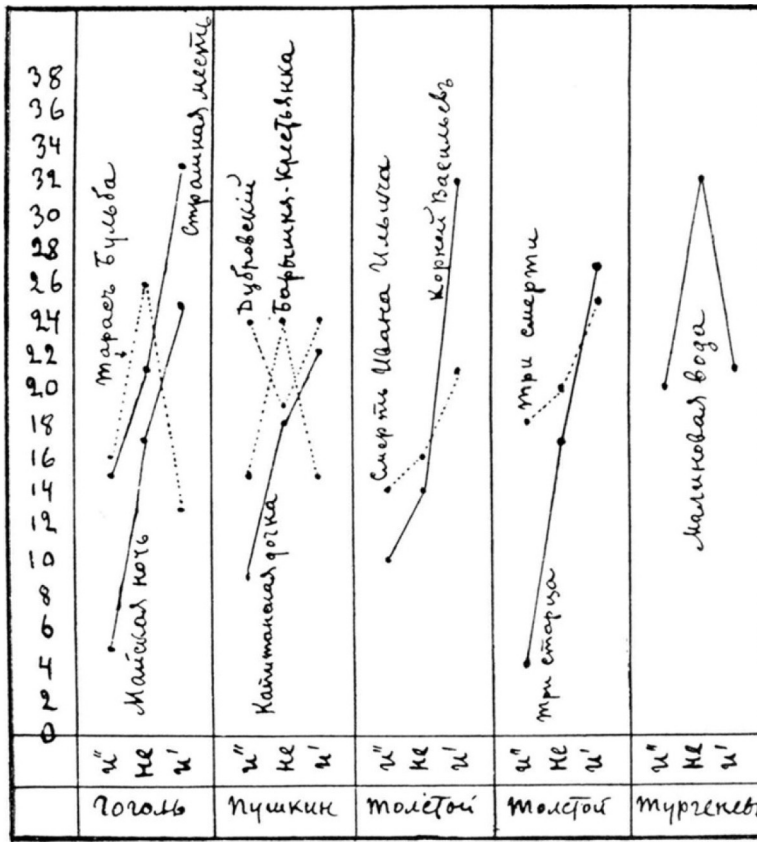

Рис. 1. Спектр частицы «не» и двух версий союза «и» (первая версия — союз, объединяющий имена, вторая — глаголы и предложения) [Морозов, с. 109]

получится совсем иной. И это лишает атрибуцию всякого смысла. Именно за это упрекал автора академик Марков, говоря о недостоверности эксперимента Морозова [Марков, 1916]. Многие исследователи, особенно текстологи, и по сей день к работе Морозова относятся сдержанно-скептически, но уже по лингвистическим основаниям. Так, А. Л. Гришунин пишет: «Наш собственный опыт работы по способу Морозова показал, что замена одних отрывков (тысяч слов) на другие дает совершенно различные результаты, причем диапазон колебаний чрезвычайно велик. Кроме того, многие служебные слова на 1000 слов встречаются в мизерном количестве (1–2–3 случая), а из таких малых цифр выводить какую-либо закономерность нельзя. Очевидно, одной тысячи слов недостаточно для таких наблюдений. Нужны значительно большие объемы» [Гришунин]. Однако Гришунин с похвалой отзывается об эффективности использования при атрибуции синтаксических дублетов (*потому что — так как, нежели — чем* и т. п.).

6. Искусствометрия (измерение динамики текста)

Основоположником отечественной искусствометрии является музыковед Эмилий Карлович Розенов. Основное внимание в своих исследованиях он обратил на динамику текста.

В статье, написанной в 1904 г., были продолжены опыты по формированию искусствометрии на основе исследований Цейзинга, о которых мы говорили выше. Основой статьи послужил доклад Розенова «О применении закона „золотого деления“ к музыке», сделанный им на заседании Московского научно-музыкального кружка 15 октября 1903 г. и опубликованный в «Русской музыкальной газете» (1904, № 25–28), а также в «Известиях С.-Петербургского Общества музыкальных собраний» (СПб., 1904, июнь — июль — август, с. 1–19).

Отличительной чертой статьи Розенова является то, что здесь золотое сечение в явном виде используется для анализа не только музыкального, но и словесного текста. До Розенова эта эстетическая характеристика использовалась только при анализе музыки. Причем Розенов исследует оба типа звучащего текста с единых позиций и в единых терминах. К такому методу анализа Розенова, по его же словам, побудили обратиться «бедность, шаткость, и разрозненность музыкальной эстетики» и желание разгадать «таинственные

творческие законы природы, руководящие музыкальным формовоплощением художественно-эмоциональных идей через посредство человеческого гения» [Розенов, с. 119]. Это очень важная мысль, ибо для Розенова закон природы, воплощенный в выдающихся произведениях искусства, превращается в эстетический закон.

В центре внимания Розенова — динамическая развертка текста и разыскание в нем точки-кульминации, которая может «1) служить моментом раздела между главными частями произведения и установить этим пропорциональные размеры частей по отношению к целому; она может 2) подчеркнуть кульминационный пункт возрастающего по напряжению ожидания и может 3) отметить главную, основную мысль произведения, поместив ее на столь заметное для непосредственного чувственного восприятия место» [Там же. С. 125].

Далее Розенов на материале текстов «Бородино», «Умирающий гладиатор», «Демон», «Три пальмы» М. Ю. Лермонтова, «Кубок» Ф. Шиллера, «То было раннею весной» А. К. Толстого демонстрирует эффективность своей методики. Тот же вывод делается для ряда произведений Баха, Бетховена, Моцарта, Вагнера и Глинки, народных песен.

В заключение Розенов отмечает, что закон золотого сечения проявляется далеко не во всех случаях. Наиболее четко он «проявляется у гениальных авторов, а у последних — преимущественно в эпоху их полной зрелости и главным образом в лучших, наиболее одухотворенных творениях их» [Там же. С. 156].

Анализу словесного текста посвящена и статья Павла Александровича Флоренского (1882–1937), в которой золотое сечение рассматривается как точка поворота от кульминации к развязке на материале трагедий Софокла [Флоренский, с. 488].

Несколько позднее Леонид Леонидович Сабанеев (1881–1968) предпринял детальное исследование проявления золотого сечения. Им было исследовано более 2000 произведений русских и зарубежных композиторов.

Сабанеев исходил из посылки, что музыкальное произведение во времени делится на части некоторыми вехами, которые облегчают восприятие сложного целого [Сабанеев]. Такими вехами, по Сабанееву, являются: изменение структуры мелодии, интонационные кульминационные пункты, изменение тональности и др. При этом

в большинстве случаев такие изменения, переломы, переключения делят текст по закону золотого сечения.

Интересно, что в динамике музыкальных произведений Сабанеев обнаруживает не только классическое, а целую серию золотых сечений. Каждое сечение отражает свое музыкальное событие в развитии музыкальной темы. В изученных им 1770 сочинениях 42 композиторов он зафиксировал 3275 золотых сечений. Причем в подавляющем числе произведений золотое сечение проявляется.

Наиболее всесторонне Сабанеевым были изучены этюды Шопена. Все они, кроме трех, содержат золотое сечение (всего было выявлено 154 таких сечения). Сабанеев обратил внимание также и на то, что в ряде случаев зеркальная симметрия сочетается с золотой. В таких случаях произведение распадается на несколько равных частей, в каждой из которых можно выделить золотое сечение.

Характерно, что Сабанеев, как и Розенов, указывает на то, что золотое сечение чаще всего обнаруживается в высокохудожественных произведениях, написанных выдающимися композиторами. Причем весьма примечателен тот факт, что в *произведениях композиторов XX века золотая пропорция встречается значительно реже, чем у их предшественников*. Это было следствием отхода от классических традиций, массового распространения модернизма и авангардизма.

Исследования Розенова и Сабанеева позднее были продолжены Львом (Лео) Абрамовичем Мазелем (1907–2000). В своей книге [Мазель] он отмечает наличие в произведении некоторого «кульминационного взлета», высшей точки, причем такое построение характерно не только для произведения в целом, но и для его частей. Мазель подчеркивает, что кульминация редко располагается в центре произведения, она обычно асимметрично смещена. Например, в восьмитактных мелодиях Бетховена, Шопена, Скрябина высшая точка располагается на сильной доле шестого такта или на последней мелкой доле пятого такта, то есть в точке золотого сечения. По мнению Мазеля, доля таких восьмичленных мелодий, в которых подъем занимает пять тактов, а последующий спуск — три, необычайно велика. Если автор пишет гармонично, то наверняка это проявляется в установленной числовой закономерности. Рисунок мелодии имеет такой «профиль»: от длительного периода нарастания через кульминацию к более короткому спаду.

Значительный вклад в теорию и практику формирования структуры повествования на материале кинематографа внес великий кинорежиссер Сергей Михайлович Эйзенштейн (1898–1948). Он понимал необходимость введения точных методов анализа, чем объясняется, в частности, его увлечение русским авангардом, достижениями русских формалистов, а также золотым сечением. Специфической особенностью работ Эйзенштейна было то, что и золотое сечение, и логарифмическая спираль (линия типа латинской *S*) изучались им как структурные схемы, соотносящиеся с общими законами природы. Логарифмическую спираль в версии Хогарта Эйзенштейн обсуждает в контексте соединения двух противоположных начал — *инь* и *ян*, характерных для китайской модели мира. Такая спираль рассматривается Эйзенштейном как модель развития.

Можно отметить также интерес Эйзенштейна к динамической организации временных искусств: музыки, словесных произведений, кинофильмов. Например, анализируя произведения Пушкина, он отмечает характерные точки поэтического текста, в которых проявляется закон золотого сечения [Иванов, с.193].

Характерные переломы композиции в точках золотого сечения Эйзенштейн использовал и при «конструировании» фильма «Броненосец „Потемкин“». Он разбил ленту на пять частей. В первых трех действие разворачивается на броненосце. В двух последних — в Одессе, где поднимается восстание. Этот переход в город происходит точно в точке золотого сечения. В каждой части также есть свой перелом, соответствующий золотому сечению.

Таким образом, Эйзенштейн сознательно использовал в своих шедеврах идею золотого сечения, рассматривая связанные с ним структуры как элемент творческого метода.

Заключая этот раздел, мы хотели бы обратить внимание на еще одно малозаметное достижение. Оно принадлежит Роману Осиповичу Якобсону (1896–1982), выдающемуся деятелю русской формальной школы. Речь идет о симметрийных идеях, которые занимали Якобсона в течение всей жизни (вспомним хотя бы теорию параллелизма, сочетание связей по сходству и по смежности, принцип бинарности в теории оппозиций и др.). Причем Якобсон объединял эти идеи с физическими принципами, математическими и биологическими идеями.

Особенно ярко это проявилось при анализе творчества выдающегося немецкого поэта Фридриха Гёльдерлина (1770–1843) [Якобсон, с. 364–386].

Якобсон говорит о регулярном противопоставлении трех заключительных строк в стихотворении «Die Aussicht» пяти предшествующим. При этом меньший отрезок (Minor) относится к большему (Major), как больший к целому, то есть здесь наблюдается золотая пропорция $8:5 \approx 5:3$, которая связывает неравные отрезки восьмистишия. Это две синтаксически подобные группы с пятью глаголами каждая. Глаголы зеркально симметрично расположены в полустишиях пятистрочно и трехстрочно.

Якобсон при этом подчеркивает, что Гёльдерлин использует прием пропорционирования *сознательно*, как пример сложного и преднамеренного формообразования. При этом обращает на себя внимание то, что это пропорционирование отражает и определенные динамические закономерности организации стиха.

Таким образом, Якобсон продолжает традицию изучения гармонических структур, идущую от искусствометрии Цейзинга и психометрики Фехнера.

7. Квантитативная лингвистика

Основателем квантитативной лингвистики принято считать американского лингвиста и психолога Джорджа Кинсли Ципфа (1902–1950). В основе лингвистических взглядов Ципфа лежат психологические установки весьма общего характера. «Весь опыт — это реакция, структурно организованная в своем источнике. Любая реакция есть выражение, как только она осознается, а любое выражение — это язык, как только нам удается расшифровать его. То, что мы обычно называем языком, это весьма частная область поведения, чей код достаточно хорошо известен» [Zipf, p. 309]. Это утверждение Ципфа ставит язык в один ряд с другими системами коммуникации и перекликается с концепцией Соссюра о языке как одной из разновидностей знаковых систем. Ципф говорит о том, что для всех форм поведения человека, включая речевое, характерно стремление к равновесию, которое регулируется законом экономии усилий — стремлением поддерживать равновесие между формой и поведением [Там же. P. 303]. Под формой Ципф имеет в виду

внутреннее строение объекта, а под поведением — его встречаемость (повторяемость). Этот закон Ципф распространял не только на лингвистические феномены, но и на различные социальные и даже обыденные явления.

Например, Ципф, проведя статистический анализ браков, заключенных в 1931 г. в 20 кварталах Филадельфии, показал, что 70% брачных союзов были зарегистрированы между людьми, проживавшими друг от друга на расстоянии, не превышающем 30% размера этой территории.

Любопытен и следующий пример. Ципф предложил научное обоснование такому обыденному явлению, как беспорядок на рабочем столе. Он предположил, что здесь имеет место такая закономерность: человек стремится к тому, чтобы все необходимые вещи были рядом, «под рукой», «на подхвате», загромождая ими рабочее пространство.

Этот пример Ципфа можно истолковать и так. Между иерархической значимостью объекта и его локализацией в пространстве существует сильная зависимость. Если речь идет, например, о линейной организации синтаксиса текста, то слова, тесно связанные с «хозяином» (синтаксической доминантой), стремятся расположиться рядом с ним. В том же ключе может рассматриваться и закон Отто Бехагеля, согласно которому более распространенный (более сложный) член предложения располагается относительно глагола после более простого. Это одна из интерпретаций принципа наименьшего усилия, предложенного Ципфом в 1949 г.

Этот принцип перекликается с принципом равновесия Парето, согласно которому любые ресурсы (люди, товары, время, знания, финансы и др.) самоорганизуются так, чтобы свести к минимуму затраченную работу, и, таким образом, 20–30% ресурса создают 70–80% совокупного результата. Например, 20% самых частых слов обычно составляют 80% словоупотреблений текста.

Ципф предложил также эмпирический закон, связывающий ранг слова в частотном словаре с частотой слова [Там же. P. 484–490]. Этот закон, имевший в окончательной версии вид неравносторонней гиперболы, принципиально ничем не отличался от закона, который был предложен пятью десятилетиями раньше итальянским экономистом Вильфредо Парето (1848–1923). Возможно, Ципф не знал о достижении итальянца. Но в этом Ципф не был одинок.

Аналогичный закон в 1913 г. вывел немецкий физик Феликс Ауэрбах (1856–1933), упорядочивая города по численности населения.

Так же как и закон Парето, закон Ципфа, сформулированный для рангового распределения, имеет вид формулы неравносторонней гиперболы вида

$$k_r = \frac{k_{\max}}{r\gamma},$$

где r — ранг слова, k_r — частота слова ранга r, $k_{\max}$ — частота самого частого слова, а γ — коэффициент, характеризующий неравномерность (рассеяние и концентрацию) распределения частот. Прежде эту формулу уже использовал американец Альфред Лотка в наукометрии при построении распределения ученых по их научной продуктивности (числу написанных статей) и цитируемости [Lotka]. В формуле Лотки γ — величина постоянная, равная 2. Но намного раньше Ципфа (в 1918 г.) формулу гиперболы использовал английский ботаник и биогеограф Джон Кристофер Уиллис (1868–1958), который исследовал распределения биологических родов по числу видов, а также распределение видов по численности популяций [Willis].

8. Первые шаги лингвистической статистики

Первая половина XX века — время начала рабочего освоения статистического исследования речевых произведений. В обиход статистического анализа постепенно вводились такие понятия, как вероятность, статистические распределения, средняя величина, дисперсия, статистическая ошибка и др. Эти и многие другие статистические понятия широко использовались в книгах Юла [Jule], Росса [Ross], Рида [Reed] и др. В этих работах впервые были поставлены вопросы о тексте как статистической совокупности, об устойчивости и вариативности стилевых характеристик текста, о значимости различий между средними величинами, а также впервые было показано, как строятся статистические распределения.

9. Эстетические измерения

Эстетические измерения — интересная и запутанная область взаимодействия психологии и искусствоведения [Семиотика и искусствометрия]. Деятелей искусства, искусствоведов, специалистов

по эстетике всегда волновали идеи экспериментальной проверки эстетических теорий и объективного измерения эстетических предпочтений, относящихся к произведениям искусства или тем или иным артефактам различной формы. Пионером в этой области является основоположник психофизики Густав Теодор Фехнер (1801–1887) [Мартыненко, 2009], проверявший гипотезу о золотом сечении для артефактов прямоугольной формы с разным соотношением сторон.

В такого рода психологической эстетике исследователи разделились на две группы.

Одна из групп исследователей пришла к выводу, что визуальными переменными объектов, предпочитаемых художниками, являются простота и симметричность. Исследователи второй группы придерживались противоположного убеждения, считая такими переменными сложность и асимметричность. Между этими полярными точками зрения и колеблются варианты эстетических измерений; более четких позиций нет и по сей день.

В качестве примера приведем концепцию гарвардского математика Гаррета Биркгофа (1911–1996). Он предложил следующую формулу эстетической меры:

$$M = \frac{O}{C},$$

где M — эстетическая мера, O — упорядоченность, а C — сложность. Эстетическая мера есть ощущение ценности, чувство удовольствия. Эта мера, по мнению Биркгофа, соответствует идее единства в многообразии и понятию хорошей формы в гештальтпсихологии. Предпочтительными зрительными характеристиками объектов являются простота и симметричность [Birkhoff].

Впрочем, спустя некоторое время была предложена противоположная формула, согласно которой эстетическая мера есть не отношение, а произведение упорядоченности и сложности [Eysenk].

В общем, в выборе меры согласия нет, но для нас эти усилия интересны тем, что они продолжают серию попыток построить обобщающие индексы. Такие усилия характерны в целом и для стилеметрии.

10. Заключение

В первой половине XX века продолжалось развитие стилеметрии, представленной в конце XIX века как решение проблем атрибуции и датировки текста лингвостатистическими методами, в том числе и в России. Причем наметился переход от атрибуции древних текстов, преимущественно связанных с именем Платона, к современным текстам.

1. Опираясь на идеи Рюмелина и других приверженцев теории сообществ, А. А. Чупров построил теорию статистической совокупности на основании противопоставления общих и собирательных понятий с учетом эффекта локализации группы объектов в определенных рамках времени и пространства. К представлениям Чупрова тесно примыкают соображения Ю. Н. Тынянова касающиеся литературно-художественных систем (синхронических и диахронических), их ядра и периферии.

2. На основе филологических данных получили развитие важные математико-статистические идеи. В частности, А. А. Чупров, рассматривая теорию статистической устойчивости Лексиса, опроверг утверждение последнего, что в реальной действительности статистические ряды с поднормальной дисперсией отсутствуют. Такие ряды Чупров нашел, проводя лингвистические эксперименты. Вслед за Чупровым академик А. А. Марков, исследуя последовательность чередования гласных и согласных в романе «Евгений Онегин», открыл свои знаменитые цепи. Оба достижения связали теорию дисперсии Лексиса с цепями Маркова и теорией статистического вывода английской биометрической школы, в частности с критерием хи-квадрат Карла Пирсона [Мартыненко, 2014].

3. Параллельно лингвисты неспешно осваивали статистическую технику. Особенно они преуспели в фонетике и даже основали дисциплину фонометрия [Там же]. О словарных изысканиях мы уже говорили в связи с именем Ципфа. Возникла квантитативная лингвистика как раздел науки о статистических законах в языке и речи. Первым стал закон Ципфа, перекликающийся с аналогичными законами в биологии (Уиллис), экономике (Парето), науковедении (Лотка).

4. Бурно развивались, благодаря усилиям Андрея Белого и представителей русской формальной школы (Шкловский, Тынянов, То-

машевский, Жирмунский), математическое стиховедение и лингвистическая поэтика.

5. В области динамических искусств были продолжены исследования композиции текста на основе золотого сечения: в словесности (Розенов, Флоренский, Якобсон), музыке (Розенов, Сабанеев), кинематографе (Эйзенштейн).

6. Психологические исследования, инициированные Г. Фехнером в области эстетического восприятия, были продолжены в области экспериментальной эстетики путем построения мер эстетичности. Это исследовательское направление продолжило развитие индексного метода, предложенного А. Кетле [Там же], в рамках эстетических измерений, тесно связанных со стилеметрией [Birkhoff].

Литература

Белый А. Мастерство Гоголя. Л., 1934.

Белый А. Символизм. СПб., 1910.

Гришунин А. Л. Опыт обследования употребительности языковых дублетов в целях атрибуции // Вопросы текстологии. М., 1960. С. 146–195.

Жирмунский В. М. Введение в метрику. Теория стиха. Л., 1925.

Иванов В. В. Очерки по истории семиотики в СССР. М., 1976.

Мазель Л. А. Строение музыкальных произведений. М., 1960.

Марков А. А. Об одном применении статистического метода (О статье Морозова «Лингвистические спектры») // Известия Отд. рус. языка и словесности Имп. Академии наук. СПб., 1916. Т. X. № 4. С. 239–242.

Марков А. А. Опыт статистического исследования текста романа «Евгений Онегин» // Изв. Имп. Академии наук. СПб., 1913. Серия 6. Т. 7.

Мартыненко Г. Я. Введение в теорию числовой гармонии текста. СПб., 2009.

Мартыненко Г. Я. Стилеметрия: возникновение и становление в контексте междисциплинарного взаимодействия. Часть 1. Первые шаги: XIX век // Структурная и прикладная лингвистика. СПб., 2014. Вып. 10. С. 3–23.

Морозов Н. А. Лингвистические спектры: Средство для отличия плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд // Известия Отделения рус. языка и словесности Имп. Академии наук. СПб., 1915. Т. XX. Кн. 4. С. 93–134.

Розенов Э. К. Закон золотого сечения в поэзии и музыке // Статьи о музыке. Избранное. М., 1982.

Сабанеев Л. Л. Этюды Шопена в освещении золотого сечения. Опыт позитивного обоснования // Искусство, ГАХН. М., 1927. Т. II–III. Вып. 2–3.

Семиотика и искусствометрия / под ред. Ю. М. Лотмана и В. М. Петрова. М., 1972.

Томашевский Б. Теория литературы. Поэтика. М.-Л., 1925.

Тынянов Ю. Н. Архаисты и новаторы. М., 1929.

Флоренский П. В. Разбор некоторых суждений о законе Цейзинга // У водораздела мысли. Сочинения в 4 т. М., 2000. Т. 3 (1).

Чупров А. А. Закон больших чисел в современной науке // Статистический вестник. М., 1914. Кн. 1–2. С. 1–21. № 2. С. 3–21.

Чупров А. А. Теория статистики. М., 1910.

Шкловский В. Б. Воскрешение слова. СПб., 1914.

Шкловский В. Б. О теории прозы. М., 1925.

Якобсон Р. О. Избранные работы. М., 1985.

Birkhoff G. D. Aesthetic Measure. Cambridge, Mass., Harvard Univ. Press., 1932.

Čada F. Datovani Platonova Faidra // Listy filologicke. 1901. 28. P. 173–193.

Dittenberger W. Sprachliche Kriterien fur die Chronologie der Platonishe Dialogie // Hermes, Zeitschrift fur klassisch Philologie. B. XVI. Berlin, 1881. P. 321–345.

Eysenk H. J. An Experimental Study of the Good Gestalt // Psychological Rewiew. 1942. N 49. P. 344–364.

Jule G. U. The Statistical study of Literary Vocabulary. Camdridge, 1944.

Lotka A. J. The Frequency Distribution of Scientific Productivity // J. Washington Acad. Sci., 1926. Vol. 16. N 12. P. 317–323.

Reed D. W. A Statistical Approach to Quantitative Linguistic Analysis // Word, 1949. N 5. P. 235–247.

Ritter C. Die Sprachstatistic in Anwendung auf Platoo und Ghethe // Neue Jahrbucher fur das klassische Altertum, 1903. S. 241–261, 313–325.

Ross A. S. Philological probability problems // Journal of the Royal Statistical Society. Ser. B12, 1950. P. 19–50.

Willis J. C. Age and area. Cambdridge, 1922.

Zeller E. Uber die Unterscheidung einer Gestalt den Ideenlehre in den platonischen Scyriften. Berlin, 1887.

Zipf G. K. Human behavior and the principle of least effort. Addison-Wesley Press,1949.

С. Д. Шелов, А. Э. Цумарев

ТЕОРИЯ И ПРАКТИКА ОПРЕДЕЛЕНИЙ СПЕЦИАЛЬНОЙ ЛЕКСИКИ В ИСТОРИИ АКАДЕМИЧЕСКИХ ТОЛКОВЫХ СЛОВАРЕЙ РУССКОГО ЯЗЫКА[1]

Аннотация. В статье анализируются принципы и практика определений (толкований) лексики терминологического и профессионального характера в основных толковых словарях русского языка, начиная со «Словаря Академии Российской» (1789–1794) и кончая семнадцатитомным «Словарем современного русского литературного языка (1948–1965; БАС), а также четырехтомным «Словарем русского языка» под редакцией А. П. Евгеньевой (1981–1984; МАС). Особое внимание уделено инструкциям по толкованию лексики этого типа в МАС и БАС и вопросу о том, могут ли эти толкования отличаться от дефиниций той же лексики в терминологических словарях и чем именно. Приводятся разнообразные иллюстрации, охватывающие лексику различных научных дисциплин, областей знания и сфер профессиональной деятельности.

Ключевые слова. Терминологическая и профессиональная лексика, толковый словарь, толкование специальной лексики, история русской лексикографии.

Serguey D. Shelov, Aleksei E. Tsumarev

THEORY AND PRACTICE OF SPECIAL WORD DEFINITIONS IN THE HISTORY OF ACADEMIC EXPLANATORY DICTIONARIES OF THE RUSSIAN LANGUAGE

Abstract. The article deals with regulations and practice of defining terms and professional words in the principal explanatory dictionaries of the Russian language beginning with the "Slovar Akademii Rossiyskoi" (Dictionary of the Russian Academy) (1789–1794) and ending in "Slovar Russkogo jazyka" (Dictionary of the Russian Language / Ed. A. P. Evgenieva. Vol. 1–4. Moscow, 1981–1984) and "Slovar sovremennogo russkogo literaturnogo jazyka" (Dictionary of the Modern Standard Russian Language. Vol. 1–17.

[1] Статья подготовлена в рамках совместного проекта российских и белорусских ученых, поддержанного РГНФ (проект 15-24-01001 а/м).

Moscow, Leningrad, 1948–1965). The article focuses on particular instructions published before the dictionaries come out of press on how to define terms and professional words. Special attention is drawn to the question whether definitions of the same terms and professional words in explanatory dictionaries could differ from these in the terminological dictionaries, in what and why.

Keywords. Terminological and professional lexis, explanatory dictionary, explanation of special words, history of the Russian lexicography.

Слова терминологического или профессионального характера из сферы науки, техники, производства, традиционных ремесел и промыслов, искусства и спорта, получившие достаточное распространение в общем языке, традиционно включаются в общие толковые словари русского языка, в которых эти слова, как и все другие лексические единицы словаря, получают словарное определение[2]. Изучение истории правил и способов определения лексики этого типа в русской лексикографии конца XVIII в. — 1950-х гг. показывает, что они складывались постепенно[3].

В предисловии к первому академическому толковому словарю — «Словарю Академии Российской» (1789–1794) — об использовании в толковании «сословов» (т. е. синонимов) и «определений» говорится следующее: «При объяснении знаменования слов главным пособием употребила Академия сословы или слова тожде значащие, после коих для большего утверждения как точного, так и в иной смысл преходящего, или распространенного знаменования, присовокуплены определения…» [САР, с. XIII]. Обращение к самим словарным статьям показывает, что для описания семантики специальных слов авторами словаря использовались различные средства и способы. Так, практиковались отсылка к синониму (см. ниже *апроши*), толкование с указанием синонима (синонимов; см. *архитрав*) или без указания такового (см. *атом*); толкование могло сопровождаться

[2] В данной работе мы будем использовать термины «толкование», «определение» и «дефиниция» как синонимичные. В работе [Шелов, 2003] предлагается считать термин «толкование» более общим, чем термины «определение» и «дефиниция».

[3] Толкованию специальной лексики терминологического и профессионального характера в филологических толковых словарях посвящено значительное количество исследований, см., например, сборник [Проблематика определений …], а также работы [Сороколетов, 1957; Сороколетов, 1962; Щерба; Гречко; Паламарчук; Толикина; Правдин; Малина; Соболева; Барсукова; Четырина; Крысин; Шелов, 2014а; Шелов, 2014б] и др. В настоящей статье нас интересует главным образом исторический аспект этого вопроса.

комментарием с примерами обозначаемого словом явления (см. *амфивия*), например:

АПРО́ШИ, -шей, с. м. множ. Реч. военн.: Франц. зр. Прикопы[4].

АРХИТРА́В, -ва, с. м. Греколат. Речение архитектурное. Брус положенной сверх столбов, соединяющий оные и составляющий одну из трех частей антаблемента. Зр. Переводина, перекладина.

А́ТОМ, -ма, с. м. Греч. Собственно значащее, несекомый, неразделимый. В физике под сим именем разумеются тела, которые в рассуждении их малости разделены быть не могут.

АМФИ́ВИЯ, -вии, мн. ср. Греч. означающее животных двустихийных или водоземных, т. е. могущих жить в воде и на суши: *Крокодил, тюлени* и пр. суть животныя *Амфивии*. Новейшие же естественной истории писатели в обширнейшем знаменовании пред древними употребляют сие слово, и разумеют чрез оное животных снабженных единовместилищным сердцем и красною холодною кровью. Оне дышут легким, вместо же костей снабжены хрящем. В сем смысле *Скат*, *Пресмыкающиеся*, *Миноги* и пр. суть животные амфивии. Зр. Двустихийный.

Отдельного замечания была удостоена биологическая лексика (лексика «трех царств природы»): «Объяснения слов, изъявляющих произведения трех царств природы в России, не токмо составлены из отличительных их знаков и свойств, но прилагаемо было тщание замечать главнейшие их употребления; для большего же разумения и введения оных в употребление присовокуплены и латинские названия из Линнеевой системы» [САР, с. XIV]. Следует полагать, что речь здесь идет о включении в толкование сведений, раскрывающих (часто с избыточным креном в сторону энциклопедизма) актуальность в жизни человека обозначаемой словом реалии, см., например, следующую словарную статью:

БАРВЕНА́, -ны. с. ж. Mullus barbatus. Рыба отменного вкуса и красивого вида, величиною от шести до семи вершков, покрытая белоблестящею чешуею, из-под коей алая кожа везде отливает и производит весьма приятный алый цвет, который у сонной живее бывает, нежели у живой. Спина и голова у ней выпуклисты; а на нижней челюсти два длинные красные уса. Ловится у нас в Севастопольской гавани.

Как мы увидим далее, установка на отражение в толковании полезных для человека свойств обозначаемого каким-либо словом

[4] Зр. — зри (т. е. смотри).

растения, животного и т. п., впервые заданная в «Словаре Академии Российской», будет в дальнейшем развита в отечественной толковой лексикографии XX в.[5]

В толковом «Словаре церковнославянского и русского языка» (1847) был сформулирован ряд лексикографических правил, касающихся составления этого словаря. Относительно толкований авторы ограничились следующей формулировкой: «Определения значений слов составлять со всею логическою точностью и краткостью» [СЦРЯ, с. XII]. Описание специальной лексики в этом словаре действительно в значительной степени отличается от соответствующих мест словаря 1789–1794 гг.: словарная статья становится более структурированной, а толкование — лапидарным, ср.:

АПРО́ШИ, -ей, с. ж. мн. Воен. Земляные работы, производимые осадным войском, чтобы приблизиться к крепости; прикопы.

АРХИТРА́В, -а, с. м. Архит. 1) Брус, положенный сверх столбов, дверей и окон, и соединяющий их. 2) Третья и нижняя часть антаблемента, лежащая непосредственно на капителях колонн; переводина, перекладина.

А́ТОМ, -а, с. м. В физике: тело, которое, по причине его малости, не может быть разделено.

АМФИ́БИЯ, и **АМФИ́ВИЯ**, -и, с. ж. Земноводное животное.

БАРВЕНА́, -ы́, с. ж. *Mullus barbatus*, рыба. *Краснобородка*.

В 1891–1895 гг. вышли в свет выпуски А–Д неоконченного «Словаря русского языка» под редакцией Я. К. Грота. Характеризуя этот лексикографический труд, академик В. В. Виноградов отмечал: «Грот придавал особенное значение четкости и точности семантических определений слова. И в этом отношении составленная им часть академического словаря (А–Д включительно) является до сих пор непревзойденным образцом (выше по качеству лишь выпуск на И — до слова *идеализироваться*, обработанный проф. Л. В. Щербой)» [Виноградов, с. 231].

Во введении к словарю (в параграфе IV «Определение значений

[5] Известно второе издание этого лексикографического труда — «Словарь Академии Российской, по азбучному порядку расположенный» (1806–1822), который историками словарного дела характеризуется следующим образом: «Несмотря на устранение некоторых недочетов первого издания, Словарь в новой редакции „не имел того живого теоретического и практического значения, как первооригинал“» [История русской лексикографии, с. 126]. Это издание в настоящей работе не рассматривается.

слова») способы толкования, применяемые в нем, остаются нераскрытыми, вместе с тем исследователь гротовского словаря С. А. Эзериня обращает внимание на то, что в работе «К соображению будущих составителей русского словаря» Я. К. Грот писал: «Что касается в особенности до технических терминов, то словарь, конечно, не обязан во всей подробности объяснять или описывать выражаемые им предметы, что составляет дело науки. При именах растений достаточно, кажется, как и сделано в нашем академическом словаре, объяснять их латинским названием, прибавляя по-русски только слово *растение*»; см. [Эзериня, с. 79]. Исследователь отмечает также, что «подобно словарям-предшественникам, Словарь Грота активно прибегает к подтолкованию специальной лексики через ее синонимические эквиваленты из общелитературного языка, диалектов, из сферы разговорной речи» [Там же].

Общий вывод, к которому приходит С. А. Эзериня, говоря о разработке толкований единиц специальной лексики в «Словаре русского языка» под редакцией Я. К. Грота, таков: «Несмотря на многочисленные достижения в толковании специальной лексики, вероятно, из-за весьма сжатых сроков издания, „Словарь русского языка“ так и не смог до конца преодолеть несбалансированность в отношении объема и типа дефиниций терминов — от исчерпывающего энциклопедизма в описании одних терминологических групп, в первую очередь терминов филологии и гуманитарных дисциплин (e. g. *давнопрошедший*, *дательный*, *давность* (юрид.), *дежурный*), до минимализма, неточности, неясности в описании некоторых других терминов (e. g. *атом*, *гуява* (бот.), *деклинация* (астрон.), *долерит* (минерал.) и др.)» [Там же]. Ряд удачных толкований специальной лексики в словаре Я. К. Грота проанализирован в работе [Цумарев].

После смерти в 1893 г. Я. К. Грота работа над словарем (со значительной его переработкой) была продолжена (до 1920 г. она велась под руководством А. А. Шахматова, затем до 1937 г. — под редакцией В. И. Чернышева и Л. В. Щербы). Т. А. Коровaненко указывает на то, что «словарь Шахматова задуман как тезаурус, в котором каждая категория лексики получала свой тип толкования»: для основной массы обиходно-бытовой лексики использовался собственно филологический родо-видовой тип толкования; для терминологической и этнографической лексики — энциклопедический и описательно-терминологический; для устарелой, областной, иноязычной, экс-

прессивной лексики — синонимический; для лексики регулярных моделей словообразования — отсылочный. Т. А. Корованенко отмечает: «Описательно-энциклопедический тип толкований широко использовался в Словаре Даля и был взят на вооружение Шахматовым как один из наиболее частотных типов определения значения слова. Большинство таких определений представляло собой цитату — заимствование из терминологических источников. Ср. *Желудь*… Бот. *glans*, односемянный плод с деревянистым или кожистым околоплодником, снабженный по основании чешуйками или листоватыми возросшими прицветниками…» [История русской лексикографии … , с. 260][6].

В конце 1930-х гг. работа над многотомным словарем русского языка была признана неудовлетворительной, в проекте нового словаря в духе того времени подчеркивалась необходимость «принять решительные меры по перестройке всей работы над Словарем со стороны Президиума АН СССР, чтобы это важное государственное дело организовать заново и начать выполнять его по-настоящему, как этого требует партия и правительство» [Проект Словаря…, с. 3]. Глава III проекта была посвящена вопросам разработки семантической и стилистической характеристики слов. Об определении терминов говорится в параграфах 74–80 проекта. Прежде всего выдвигается положение о том, что «определения научных и технических терминов не могут быть однотипными; характер определения зависит от содержания данного понятия, от его семантической ценности и от соотношения между специальным, научно-техническим его значением и значением общелитературным» [Там же. С. 29].

Далее уточняется, что значение таких слов, как *заяц*, *яблоко*, *железо*, *аспирин*, *барометр*, *гроза* и т. п. «определяется на основе общелитературного его понимания, согласованного, однако, с научным определением». При этом не следует ограничиваться отнесением обозначаемых такими словами понятий к семейству, роду, группе по соответствующей научной классификации, но необходимо давать описание наиболее существенных признаков этих понятий. «Так, при определении слова *заяц* надо указать наиболее существенное из таких его признаков, как „зверек из породы грызунов, с длин-

[6] Следует упомянуть источник, относящийся к шахматовскому словарю: [Словарь русского языка: Инструкция для редакторов]. Поскольку эта инструкция не была воплощена в жизнь, далее она не рассматривается.

ными ушами, с тонким слухом, очень пугливый". Не годится общее определение типа „млекопитающее из отряда грызунов" или общее указание на место в научной классификации с латинскими названиями типа „род. из сем. зайцев или заячьих, Lepus timidus, грызун". При определении слова *вобла* недостаточно указания на то, что это „рыба из семейства карповых"; следует прибавить, что она — небольшая, одного семейства с плотвой, водится в Каспийском море, что это дешевый пищевой продукт, употребляется преимущественно в вяленом виде» [Там же. С. 29–30].

Общеупотребительным терминам противопоставлены термины ограниченного употребления. О них сказано: «Научно-технические термины, не получившие сравнительно широкого распространения в литературном языке и не связанные в литературном употреблении с яркими конкретными признаками, получают сжатое определение на основе научно-технических словарей и справочников: в таком случае после определения помещается ссылка на использованный источник» [Там же. С. 30].

Значительным событием в лингвистике 1930-х гг. стал выпуск «Толкового словаря русского языка» под редакцией Д. Н. Ушакова. В формальном отношении этот словарь не является академическим: гриф Академии наук на нем отсутствует. Однако по существу он не только развил традиции академических словарей XIX в. как словарей литературного языка, но и оказал большое влияние на последующие академические толковые словари русского языка (так, хорошо известно, что на базе четырехтомного «Толкового словаря русского языка» Д. Н. Ушакова был создан однотомный «Словарь русского языка» С. И. Ожегова). По этой причине авторы настоящей статьи сочли правильным включить этот словарь в список рассматриваемых лексикографических изданий.

В этом словаре корпусу словарных статей предшествует подробный раздел «Как пользоваться словарем». Принцип лингвистического толкования слов (в его противопоставлении энциклопедическому принципу) выражен здесь максимально определенно и решительно: «Толковый словарь — не энциклопедический словарь, задачи того и другого не совпадают: первый есть словарь языка и толкует слова́, второй — объясняет предметы, понятия. В иных случаях и в некоторых отношениях бывает трудно провести тончайшую грань между объяснением предмета и объяснением называющего его слова, но эта

грань есть и должна быть. От словаря языка требуется дать все то, что достаточно для понимания слова, а не для знакомства с самим предметом; поэтому от него нельзя требовать не только исчерпывающих, но и полных сведений о предмете. В особенности это касается научных и технических слов. **Определения их в толковом словаре не должны противоречить науке и действительности, но в то же время могут и не совпадать с научными определениями, так как могут не передавать всех научных признаков понятия**» (выделено нами. — *С. Ш., А. Ц.*) [СУ, с. XXIV].

Интересно привести, в частности, словарные статьи из [СУ], посвященные упомянутым выше словам *заяц* и *вобла*:

ВО́БЛА… Рыба одного семейства с плотвой, водящаяся в Каспийском море.

ЗА́ЯЦ… **1.** Млекопитающее отряда грызунов. *Охота на зайцев.* || Жарко́е из этого животного <…>.

Как видим, толкование в этих двух случаях строится не вполне одинаково: в первом случае семейство, к которому относится вобла (карповые), прямо не называется, но при этом имеется указание на ареал ее обитания; во втором случае прямо указывается биологический отряд. При этом словарные статьи не содержат «описания наиболее существенных признаков».

Таким образом, в лексикографической практике 1930-х гг., очевидно, могла применяться различная стратегия толкования терминологической лексики: без включения энциклопедических сведений или с широким включением последних.

В целом следует сказать, что включению и описанию лексики терминологического и профессионального характера в филологических словарях в первые десятилетия советской лексикографии придавалось огромное значение. Об этом, в частности, ярко свидетельствуют следующие слова члена-корреспондента АН СССР В. И. Чернышева, который, говоря о подготовительной работе к изданию «Словаря современного русского литературного языка» в 17 томах, писал: «Научная терминология, обработанная специалистами, является важным и ценным отличием нашего Словаря от всех доселе выходивших академических словарей русского языка» [Чернышев, с. 54]. Он же особо отмечал участие в этой работе академиков, членов-корреспондентов, профессоров, включая академиков В. Л. Комарова, С. И. Вави-

лова, В. И. Вернадского, В. Р. Виль-ямса, Б. Д. Грекова, И. М. Губкина, И. И. Мещанинова, А. Е. Ферсмана и др.

Для дальнейшего исследования толкований специальной лексики в общефилологических толковых словарях полезно обратиться к опыту в этой области, который отражен в Инструкциях для составления «Словаря современного русского литературного языка» (в трех томах) [Инструкция 1] и «Словаря современного русского литературного языка» (в пятнадцати томах) [Инструкция 2]. Иными словами, теперь мы рассмотрим инструктивный материал двух важнейших в истории отечественной лексикографии изданий — Малого академического словаря [МАС] и Большого академического словаря [БАС].

Общие положения, касающиеся толкования лексики в МАС, приводятся в главе IV Инструкции 1 «Смысловая и стилистическая характеристика слов», точнее в той ее части, которая называется «Смысловая характеристика слов» [Инструкция 1, с. 30–39]. Здесь речь идет об определении различных значений слов русского языка, в частности отмечается: «Самым важным при построении словарной статьи является правильное выделение и определение отдельных значений слова, а также установление верного соотношения между значением и оттенками значения» [Там же. С. 31]. При обсуждении толкования любого типа лексики говорится: «Определение должно представлять собой ясное, сжатое и при этом достаточно полное истолкование значения (или оттенка значения). Оно должно быть выражено при помощи общелитературной лексики. В определении следует избегать просторечных, специальных и редких слов, а также слов, употребленных в переносном значении. Совершенно недопустимы в определении слова, которые не войдут в Словарь и не получат в нем истолкования» [Там же. С. 32]. Эти положения Инструкции 1 в целом совпадают с правилами определения понятий и терминов, принятыми в логике, или чрезвычайно близки им. Исключение составляет лишь пункт, говорящий о том, что в определении «следует избегать… специальных слов» [Кондаков, с. 467; Свинцов, с. 154–176; Шелов, 2003, с. 204–205].

Далее говорится о применении различных типов определений, которые включают а) сжатое истолкование слова, б) краткую характеристику и в) краткое определение энциклопедического характера [Инструкция 1, с. 33]. Эти типы определений поясняются следующим образом.

Сжатое истолкование слова применяется в отношении «глаголов, прилагательных, наречий, отвлеченных существительных и т. п., например:

КАТИ́ТЬ… **1.** Двигать в одном направлении округлый предмет, вращая и не отрывая его от поверхности» [Там же].

Краткая характеристика используется для пояснения значений существительных, «обозначающих конкретные предметы или явления, например:

ВО́ЖЖИ, -е́й, *мн.* Часть упряжи, веревка (или длинный ремень), прикрепляемая с обеих сторон к узде и служащая для управления запряженной лошадью.

КА́ПЛЯ, -и, *ж.* Маленькая частица жидкости, имеющая округлую форму» [Там же].

Наконец, **краткие определения энциклопедического характера** применяются «лишь по отношению к словам, обозначающим особо важные политические и философские понятия (*ленинизм*, *материализм*, *государство* и т. п.)»; иногда они приводятся в цитатном виде из работ классиков марксизма-ленинизма (вполне возможно, что использование этого типа толкований было продиктовано соображениями идеологического порядка, характерными для времени создания Инструкции 1), например в МАС:

БУРЖУАЗИ́Я… Господствующий класс капиталистического общества, являющийся собственником орудий и средств производства и живущий капиталистическим доходом, получая прибавочную стоимость эксплуатацией наемного труда.

КЛАСС… **2.** «Классами называются большие группы людей, различающиеся по их месту в исторически определенной системе общественного производства, по их отношению (большей частью закрепленному и оформленному в законах) к средствам производства, по их роли в общественной организации труда, а следовательно, по способам получения и размерам той доли общественного богатства, которой они располагают». Ленин, Великий почин.

Такая трактовка задачи толкования различных слов в филологическом словаре оставляет много неясного.

Во-первых, неясно различение между **сжатым истолкованием слова**, **краткой характеристикой и кратким определением энци-**

клопедического характера. Что, собственно, их различает? Ответа на этот вопрос ни цитированный выше фрагмент Инструкции, ни вся Инструкция в целом, ни приведенные примеры не содержат, вследствие чего обесценивается и ориентация этих типов определений на различные группы лексики. Остается без ответа и вопрос о том, почему, например, в отношении отвлеченных существительных следует применять сжатое истолкование слова, а для толкования существительных, обозначающих конкретные предметы или явления (*вожжи*, *капля*), следует использовать краткую характеристику.

Во-вторых, неясно, как соотносить перечисленные типы определений, включающие сжатое истолкование слова, краткую характеристику и краткие определения энциклопедического характера, с «**определениями, имеющими описательный характер**» (эта формулировка также используется в Инструкции 1). В самом деле, являются ли «определения, имеющие описательный характер» разновидностью одного из перечисленных выше типов толкования или это самостоятельный тип определения?

Наконец, в-третьих, остается без ответа и вопрос о том, как соотносятся выделенные выше типы словарных определений с такими твердо установленными и хорошо изученными в логике типами определений, как родо-видовые (иначе называемыми классификационными), контекстуальные, операциональные и т. п.

Как отмечается в работах [Шелов, 2014а; Шелов, 2014б], все приведенные выше примеры вполне могут быть истолкованы как обычные **родо-видовые (классификационные) определения**, в составе которых для определяемых слов (и понятий) имеются формулировки ближайших родовых понятий и видовых признаков.

Особого внимания заслуживают содержащиеся в приложении к Инструкции 1 рекомендации по толкованию названий растений и животных [Инструкция 1, с. 53–57].

Некоторые из этих рекомендаций повторяют уже сказанное, распространяя его на частный лексический материал. Так, здесь сообщается, что «в определении названий растений и животных не должно содержаться специальных ботанических и зоологических терминов, не имеющих широкого распространения в общем языке» [Там же. С. 53]. Но в других случаях эти рекомендации относятся именно к данному тематическому классу лексики. В частности, отвечая на

вопрос, в каких случаях следует, а в каких не следует при определении названий животных и растений фиксировать их отнесенность к соответствующему семейству (согласно научной таксономии видов, родов и семейств этих биологических объектов), авторы пишут: «Указание на семейство дается в Словаре только в том случае: а) если семейство в целом имеет важное хозяйственное значение. Например: злаки, бобовые, грызуны (общее название для животных с ценным мехом и для животных, приносящих вред в хозяйстве); б) для наименований редких животных и растений, мало известных в пределах нашей страны, например:

КАБАРГА́… Горное безрогое животное из сем. оленей…» [Там же].

Аналогично, при решении вопроса о толковании названий редких растений авторы пишут: «Для растений редких или не встречающихся в пределах нашей страны указывается преимущественная область их распространения, что является наиболее характерным признаком для этих растений. Например:

БАОБА́Б… Гигантское тропическое растение с чрезвычайно толстым стволом» [Там же].

Также мы находим рекомендации по составлению словарных определений отдельных подтипов этой лексики (названий деревьев, цветов, животных и т. п.), например: «Для деревьев обязательно указывается, лиственное оно или хвойное; если дерево является вечнозеленым, указание на это присоединяется к определению „лиственное“ или „хвойное“. Для плодовых деревьев в определении указывается только „плодовое“:

ЕЛЬ… Вечнозеленое хвойное дерево…
ЛАВР… Южное вечнозеленое лиственное дерево…» [Там же. С. 54].

Авторы Инструкции 1 идут по линии детализации классов специальной лексики, для каждого из которых разрабатывается своя конкретная схема толкования. Такие схемы представлены для названий плодовых растений, овощей, животных (с уточнением «простейшие», «кишечнополостные», «моллюски», «рыбы», «земноводные») и т. п. Иногда приводятся и образцы подобных толкований, которыми, по мысли авторов Инструкции 1, иллюстрируются соответствующие схемы, например:

«В определении рыб указываются следующие характерные признаки:

а) промысловая (горбуша, килька, треска и др.);

б) употребляемая в пищу в консервированном (анчоусы, сардины, шпроты), засоленном (сельдь), вяленом виде (вобла, тарань и др.);

в) ценная своим мясом, икрой (все осетровые и лососевые рыбы)…» [Там же. С. 56–57].

Пользуясь современными понятиями, можно было бы сказать, что составители Инструкции 1 задают некоторые **фреймы для толкования наименований флоры и фауны**. Именно в них, можно полагать, содержатся конкретные рекомендации по толкованию общего понятия, доступного для среднего носителя русского языка и названного термином или профессиональным словом соответствующей области знания или производства.

Заметим также, что содержание приложения к разделу «Смысловая характеристика слов», по сути дела, касается проблемы соотношения специальной и общей семантики слова, научного и «наивного», обиходного понятия, передаваемого словом, которое по форме является частью общей лексики, а по содержанию представляет собой и единицу общего словарного запаса русского языка, и термин или профессиональное слово той или иной области знания (в данном случае является названием представителя флоры или фауны).

При этом приложение, разумеется, дает рекомендации именно по представлению «наивного понятия», «наивной семантики» слова общего языка, а не научного понятия, стоящего за тем же по форме словом, называющим представителя растительного или животного мира. По этой же причине одно из самых частых словосочетаний в приложении — это «характерные признаки»: ср. «в определении следует указывать **характерные признаки** для отдельных групп растений», «в определении указываются **характерные признаки** животных», «в определении птиц указываются следующие **характерные признаки**» и т. п. При этом, упоминая «характерные признаки», авторы совсем не обязательно соотносят их с научными характеристиками представителей флоры и фауны, с помощью которых эти представители описываются в научных номенклатурах и таксономиях, начиная с номенклатуры К. Линнея. Более того, иногда авторы прямо предупреждают о нежелательности подобного подхода,

см. уже приводившуюся цитату: «В определении названий растений и животных не должно содержаться специальных ботанических и зоологических терминов, не имеющих широкого распространения в общем языке» [Там же. С. 53]. В качестве элементов языка толкования, согласно Инструкции 1, допускаются лишь те слова и выражения, которые сами функционируют в общем языке: «В состав определения включаются те термины ботанических и зоологических классификаций, которые **одновременно являются словами общего языка** (грибы, кустарники, мхи, рыбы, птицы и др.)» (выделено нами. — *С. Ш., А. Ц.*) [Там же][7].

Приложение содержит наиболее разработанное описание приемов толкования названий различных представителей флоры и фауны, причем эти приемы сохраняют свое значение и по сей день, а соответствующий раздел, как говорилось, можно охарактеризовать как задание фреймов для отдельных классов слов этого типа.

Общие положения, касающиеся толкования лексики в филологическом словаре, в Инструкции 2 во многом повторяют соответствующие общие положения Инструкции 1. В то же время, согласно Инструкции 2, используемые при разработке БАС толкования и их типы значительно отличаются от тех, о которых сообщается в Инструкции 1. Так, в Инструкции 2 различаются: а) описательные толкования значений слова, б) синонимические определения и в) соотносительные (или отсылочные) толкования. В качестве описательных приводятся следующие толкования знаменательных слов:

КАЛАНЧА́… Сторожевая дозорная вышка пожарной части;

ЖА́ЖДА… Сильное желание пить;

ЗЕЛЁНЫЙ… Один из цветов солнечного спектра, находящийся между жёлтым и голубым. || Имеющий цвет травы, листвы;

КРА́СИТЬ… Покрывать или пропитывать краской, красящим веществом; окрашивать. [Инструкция 2, с. 28–29]

---

[7] За этим высказыванием следует как будто бы противоречащее ему примечание: «Исключение составляют термины, обозначающие важные в зоологических и ботанических классификациях группы растений или животных, для которых нет соответствий в общем языке (простейшие, бактерии, кишечнополостные)» [Там же]. Однако единицы *простейшие, бактерии, кишечнополостные* сами являются заголовочными словами МАС. Возможно, включение субстантивированных слов *простейшие* и *кишечнополостные* в МАС спорно, но формального противоречия между двумя высказываниями нет, так как «с точки зрения словаря» единицы *простейшие, бактерии, кишечнополостные* оказываются лексическими единицами общего языка.

К числу описательных толкований авторы относят также «определения, вскрывающие различные грамматические взаимоотношения слов языка». Они применяются к словам служебным, например:

И, *союз*. **1.** *Объединительный*. Образует из двух слов интонационное целое для выражения единого понятия, одной идеи. **2.** *Соединительный*. Употребляется для соединения двух равноправных синтаксических единиц;

ИЗ и **ИЗО**, *предлог. Употребляется с род. падежом*. Сочетание с предлогом из выражает: Пространственные отношения. **1.** Обозначает место, откуда направляется чье-либо движение…;

ЖЕ и **Ж. 1.** *Частица*. Усиливает, подчеркивает, выделяет слово, после которого ставится… [Там же. С. 29]

Включение подобных толкований в одну группу описательных толкований наряду с толкованиями, «указывающими на основные признаки называемых словом предметов или явлений действительности», вряд ли целесообразно. Приведенные толкования союза *и* и предлога *из* (*изо*) устроены принципиально иначе, полностью отличны от других описательных толкований, так как они не описывают обозначаемое понятие и не указывают его признаки, — их задача указать на правильное использование соответствующих знаков. В логике такие толкования соответствуют, скорее всего, типу гетерогенных контекстуальных определений, в которых не столько раскрывается то, что обозначает знак, сколько то, как его правильно употреблять (подробнее см.: [Шелов, 2003, с. 51–59, 73–74]). Однако мы не будем останавливаться на этом вопросе, поскольку нас интересуют в первую очередь толкования полнозначных (знаменательных) слов.

Заслуживают специального обсуждения упомянутые Инструкцией 2 **синонимические определения**, которые, по мнению ее авторов, раскрывают «значение данного слова путем сопоставления его с близкими или идентичными по смыслу». Они применяются в словаре для описания значений, стилистически или экспрессивно окрашенных, а также значений устарелых или областных, имеющих в литературном языке более употребительные или нейтральные синонимы. В качестве иллюстраций синонимических определений приводится следующий материал:

КОРЮ́ЧИТЬСЯ… *Простореч*. Сгибаться, корчиться.

ЖА́МКАТЬ… *Простореч*. Жать, сдавливать, стискивать.

КО́ЧЕТ… *Обл.* Петух.
КОШ… **2.** *Обл.* Шалаш… **4.** *Обл.* Артель. Ватага.
ИЗМЕ́ННЫЙ… **1.** *Устар.* Непостоянный, изменчивый. **2.** Изменнический, предательский.
ИЗЪЯ́ТИЕ… **1.** Удаление, устранение…
ЖА́ЛОСТЛИВЫЙ… **1.** Отзывчивый, сострадательный… [Инструкция 2, с. 29]

Среди этих примеров собственно синонимичных определений, когда одно слово строго и полностью синонимично другому, вообще говоря, всего два, а именно:

КО́ЧЕТ… *Обл.* Петух.
КОШ… **2.** *Обл.* Шалаш… **4.** Обл. Артель. Ватага.

Полное семантическое равенство означаемых слов общего языка со словами-историзмами, словами-диалектизмами, словами-регионализмами и т. п. позволяет эффективно прибегать к данному типу толкований.

Иное дело — толкование с помощью перечислительных конструкций, отдельные члены которых представлены «квазисинонимичными» единицами, которые в работе [Шелов, 2003, с. 45–50] трактуются как **перечислительные определения**. Они действительно нередки и в терминологической, и в общей лексикографии, где с их использованием связан важнейший вопрос об интерпретации собственно перечислительной конструкции, которую образуют соответствующие синонимы или квазисинонимы. В самом деле, если *жать*, *сдавливать*, *стискивать* — абсолютные (полные) синонимы, то зачем их перечисление в рамках одной перечислительной конструкции? С другой стороны, если хотя бы два члена перечислительной конструкции не являются полными синонимами, то возникает другой принципиальный вопрос: как трактовать перечисление в составе определяющего выражения? Так, при использовании перечислительных конструкций «сгибаться, корчиться», «жать, сдавливать, стискивать», «отзывчивый, сострадательный» соответственно для слов *корючиться*, *жамкать*, *жалостливый* возникают вопросы: «корючиться» — это только сгибаться или только корчиться? Или так можно назвать и то, и другое? Или «корючиться» — это нечто третье, что лишь как-то связано с действиями, обозначенными словами *сгибаться* и *корчиться*? Аналогично «жамкать» — это то же, что

жать? Или то же, что сдавливать? Или то же, что стискивать? Или это — и то, и другое, и третье? Или это — ни то, ни другое, ни третье, а лишь нечто, что как-то связано с действиями, обозначенными глаголами *жать*, *сдавливать* и *стискивать*?

Толкование общеязыковых лексических единиц с помощью однородных перечислительных конструкций, членами которых являются синонимические или квазисинонимичные выражения, рассматривал Х. Касарес [Касарес], который предложил, как представляется, достаточно интересное решение данного вопроса и которому удалось, с нашей точки зрения, выявить основное в этом типе дефиниций: указание на **общее в сходном** (но не тождественном), что выражается различными синонимичными или «квазисинонимичными» членами перечислительной конструкции. Неизменное семантическое сходство элементов сочинительной конструкции обусловливает трактовку изучаемого материала, согласно которой в основе подобных дефиниций лежит структурная формула: определяемое — это то, что является общим для всех, большинства или некоторых из объектов, названных перечислительной конструкцией [Шелов, 2003, с. 78–82].

К **соотносительным (или отсылочным) определениям** авторы Инструкции 2 относят определения, объясняющие либо «значения производного слова путем отсылки к значению или значениям слова основного (для данной словообразовательной группы)», либо «совпадающие значения словообразовательных вариантов путем их взаимного соотнесения по формуле: „То же, что…“» [Инструкция 2, с. 30]. Этот тип определений поясняется следующими примерами:

ЗАГОНЯ́ТЬСЯ… *Страдательное к* **1.** Загонять;
ИДЕАЛИСТИ́ЧЕСКИЙ… **1.** Относящийся к идеализму (в 1-м значении);
ИЗБЁНКА… *Разг. Уничижительное к* изба (в 1-м значении);
ИЗМЯ́ТОСТЬ… Состояние измятого;
КАЛЕ́НИЕ… **1.** *Действие по значению глагола* калить;
КВА́КНУТЬ… *Однократное к* квакать.
КО́ВКОСТЬ… Качество, свойство ковкого;
ЛЫ́ЖНИЦА… **1.** *Женск. к* лыжник;
РУЧЕЁК… Уменьшительное к ручей. [Там же. С. 30–39]

Заметим, что при всем разнообразии соответствующих примеров ключевым словом для приведенной цитаты, касающейся так

называемых соотносительных (отсылочных) определений в целом, является «**отсылка**» — слово, которое точно разъясняет суть текста, поясняющего значение определяемого слова. Значение приведенных слов не становится ясным непосредственно из того текста, который приводится как будто бы в качестве определения. Оно становится ясным из толкования, которое приписано другим словам, связанным с определяемым словом известными словообразовательными отношениями; смысловое содержание этих, вообще говоря, других слов, должно быть найдено в другом месте, к которому и **отсылает** соответствующее определение. Никакого собственно определения, толкования, дефинирования в отсылочной статье не происходит, а приведенный способ якобы объяснения значения слов нецелесообразно считать определением (толкованием). При всей разнотипности и многоликости соответствующих пояснений все они, с нашей точки зрения, являются лишь **пояснительными отсылками**, важность которых для любого филологического словаря невозможно отрицать.

Инструкция 2 содержит актуальный и в наши дни методический материал о толковании названий растений и животных. Этот материал во многом аналогичен указаниям Инструкции 1, ср.: «Для грибов в определении должны быть указаны цвет шляпки, размер или форма ножки, а также несъедобность:

ВОЛНУ́ШКА, -и, *ж.* Гриб с мохнатой шляпкой розоватого цвета.
ДОЖДЕВИ́К, -а, *м.* Несъедобный шарообразный гриб…». [Там же. С. 54].

В связи с тем, что в отношении этого типа лексики две инструкции во многом подобны, подробнее останавливаться на указаниях по его толкованию не имеет смысла.

Таким образом, предпринятый краткий обзор положений о толковании специальной лексики научного и профессионального характера в общих толковых словарях конца XVIII в. — 1950-х гг. свидетельствует о том, что в русской лексикографии постепенно росло осознание специфичности профессионально-терминологической лексики как объекта общей лексикографии и, как следствие, понимание необходимости различать научные дефиниции и толкования в общефилологическом словаре. Теоретические положения инструктивных материалов к МАС и БАС (1950-е), включая типологию толкований,

содержат много спорного и вряд ли выдерживают критику: анализ конкретного материала этих словарей показывает, что подавляющее большинство толкований, которые, согласно обеим Инструкциям, относятся к различным типам, на самом деле являются родо-видовыми (классифицирующими) определениями с явной формулировкой ближайшего родового понятия и его дифференциальных признаков. Изучение соответствующих инструкций и в большей степени анализ реального дефиниционного материала МАС и БАС позволяют говорить, что фактически в этих словарях используются в первую очередь родо-видовые, реже перечислительные и (значительно реже) синонимические определения; предположительно можно говорить и о спорадическом использовании для различных грамматических слов языка определений, которые в логике относятся к гетерогенным контекстуальным определениям (подробнее см. [Шелов, 2003, с. 51–59, 71–78]). Самостоятельным инструментом представления семантики словообразовательно производных слов, отличным от собственно дефиниционной системы, является развитый аппарат отсылок в словаре. В то же время безусловным лексикографическим достижением этого периода являются рекомендации по правилам толкования отдельных тематических групп и классов лексики (в особенности названий различных представителей флоры и фауны), которые с точки зрения дальнейшего развития лингвистики можно было бы считать своеобразным фреймовым описанием соответствующей лексики. Актуальной (как в теоретическом, так и в практическом плане) является разработка вопроса о соотношении родо-видового определения термина (или профессионального слова), которое принято в специальной литературе, и родо-видового определения, отражающего детерминологизацию соответствующего термина (или профессионального слова) и пригодного для общего толкового словаря.

Литература

Барсукова Е. А. Научный термин в общем и терминологическом толковых словарях русского и английского языков: На материале медицинской и компьютерной терминологий: Автореф. дис. … канд. филол. наук. М., 2004.

[БАС] Словарь современного русского литературного языка: в 17 т. М.; Л., 1948–1965.

Виноградов В. В. Толковые словари русского языка // Избранные труды. Лексикология и лексикография. М., 1977. С. 206–242.

Гречко В. А. Терминологическая лексика в академических словарях // Современная русская лексикография. 1976. Л., 1977. С. 91–100.

[Инструкция 1] Инструкция для составления «Словаря современного русского литературного языка (в трех томах)» / Общ. ред. С. Г. Бархударова и А. П. Евгеньевой. М., 1953.

[Инструкция 2] Инструкция для составления «Словаря современного русского литературного языка» (в пятнадцати томах). М.; Л., 1958.

История русской лексикографии / отв. ред. Ф. П. Сороколетов. СПб., 1998.

Касарес Х. Введение в лексикографию / пер. с испанского Н. Д. Арутюновой. М., 1958.

Кондаков Н. И. Логический словарь-справочник: 2-е изд., испр. и доп. М., 1975.

Крысин Л. П. Терминологическая лексика в современных лингвистических словарях // Терминология и знание: Материалы II Международного симпозиума / отв. ред. С. Д. Шелов. М., 2010.

Малина З. М. Терминологические фразеологизмы и их презентация в словарях русского языка: автореф. дис. … канд. филол. наук. М., 1999.

[МАС] Словарь русского языка: в 4 т. / под ред. А. П. Евгеньевой. 2-е изд. М., 1981–1984.

Паламарчук Е. А. Терминологическая лексика в общеязыковом (филологическом) словаре // Проблематика определений терминов в словарях разных типов. Л., 1976. С. 250–257.

Правдин М. Н. Словарное толкование, наглядность и здравый смысл // Лингвистическая семантика и логика: Сб. науч. трудов. М., 1983. С. 28–48.

Проблематика определений терминов в словарях разных типов. Л., 1976.

Проект Словаря современного русского литературного языка. М.-Л., 1938.

[САР] Словарь Академии Российской. СПб., 1789–1794.

Свинцов В. И. Логика: учебник для факультетов журналистики и редакционно-издательских факультетов полиграфических институтов. М., 1987.

Словарь русского языка: Инструкция для редакторов / сост. С. П. Обнорский. М.; Л., 1936.

Соболева В. Ю. Функционально маркированная специальная лексика в словаре и тексте: автореф. дис. … канд. филол. наук. Самара, 2000.

Сороколетов Ф. П. О месте производственной терминологии в толковом словаре русского языка // Лексикографический сборник. 1957. № 1. С. 121–135.

Сороколетов Ф. П. Смысловая характеристика терминов в толковых словарях // Лексикографический сборник. 1962. № 5. С. 125–131.

[СУ] Толковый словарь русского языка: в 4 т. / под ред. Д. Н. Ушакова. М., 1935–1940.

[СЦРЯ] Словарь церковнославянского и русского языка. СПб., 1847.

Толикина Е. Н. Термин в толковом словаре (к проблеме определения) // Проблематика определений терминов в словарях разных типов. Л., 1976. С. 45–57.

Цумарев А. Э. «Словарь русского языка» под редакцией Я. К. Грота и вопросы описания специальной лексики в толковых словарях // Труды Института русского языка им. В. В. Виноградова. М., 2014. Вып. 1. С. 388–433.

Чернышев В. И. Принципы построения академического словаря современного русского литературного языка // Русский язык в школе. 1939. № 2. С. 50–55.

Четырина А. М. Лексикографическое представление фрагмента языковой картины мира: религиозная лексика в «Словаре церковнославянского и русского языка»: автореф. дис. … канд. филол. наук. СПб., 2008.

Шелов С. Д. Вопросы толкования специальной лексики в толковых словарях (из истории теории и практики отечественной лексикографии) // Труды Института русского языка им. В. В. Виноградова. М., 2014. Вып. 1. С. 434–453 (а).

Шелов С. Д. О различии в толковании специальной лексики в терминологических и филологических словарях // Терминология и знание: Материалы IV Международного симпозиума / отв. ред. С. Д. Шелов. М., 2014. С. 185–200 (б).

Шелов С. Д. Термин. Терминологичность. Терминологические определения. СПб., 2003.

Щерба Л. В. Языковая система и речевая деятельность. Л., 1971. С. 265–304.

Эзериня С. А. Терминология в «Словаре русского языка» под редакцией Я. К. Грота // От буквы к словарю: сборник научных статей к 200-летию со дня рождения академика Я. К. Грота / отв. ред. О. А. Старовойтова. СПб., 2013. С. 75–84.

В. И. Шадрин

ЛОКАЛИЗАЦИЯ ИНФОРМАЦИОННОГО ТЕКСТА И ПРОБЛЕМА ДЕГУМАНИЗАЦИИ ДЕЯТЕЛЬНОСТИ ПЕРЕВОДЧИКА

Аннотация. Статья посвящена рассмотрению феномена «локализация» применительно к переводческой деятельности в процессе электронной обработки текста. Автор полагает, что современные способы локализации текстов с помощью компьютерных технологий имеют тенденцию дегуманизации языка как средства человеческого общения. Для преодоления этой тенденции предлагается развивать гуманитарный компонент перевода путем использования экономических возможностей компаний-заказчиков и совершенствования программ подготовки переводчиков.

Ключевые слова. Локализация, глобализация, дегуманизация перевода, фрагментарное чтение, гуманитарный компонент, получатель информации.

Victor I. Shadrin

LOCALIZATION OF NON-LITERARY TEXT AND DEHUMANIZATION OF TRANSLATOR'S ACTIVITIES

Abstract. The paper deals with the problem of localization viewed upon as an aspect of translator's activity in electronic text processing. The author argues that modern technologies of electronic text localization show a tendency of dehumanization of language as a means of communication between people. To overcome this tendency a certain humanitarian component should be developed on the basis of economic assistance of client companies and adaptation of translator education programs to suit the aim.

Keywords. Localization, globalization, dehumanization of translation, fragmentary reading, humanitarian component, receiver of information.

В современном переводоведении термин «локализация» является широко распространенным наименованием, употребляемым

при описании процессов перевода и адаптации документов в соответствии с требованиями новых «мест действия» (стран, регионов и соответствующих им языков), а также в области компьютерного программирования и построения страниц сети Интернет. Термин, заимствованный из сферы производства, оказался особенно полезным для обозначения одного из параметров электронной обработки текстов коллективами профессионалов [Brown, Jule, p. 58–67; Baker, p. 254; Петрова, с. 259–266].

Терминологическая ценность этого слова заключается в его эффективности с точки зрения точности номинации одного из аспектов процесса перевода. Как таковая, дискуссия по поводу понятия «локализация» обогащает теорию перевода, которая до настоящего времени часто ассоциировалась с изучением индивидуального творчества переводчика, а также с такими понятиями, как «верность оригиналу» (fidelity/faithfulness), стремление к совершенству в переводе и т. п. Исследование этих проблем почти не имеет отношения к анализу параметра «эффективность перевода». Единственная проблема заключается в том, что исследования феномена локализации, рассматриваемые скорее в практической, чем в теоретической плоскости, имеют тенденцию *дегуманизировать* язык как средство человеческого общения, что не всегда приносит положительные результаты. В постоянно нарастающем темпе мы лихорадочно соединяем слова друг с другом, чтобы выполнить работу точно в срок, строго следовать полученным инструкциям, при этом мы объясняем мир «сиюминутным» способом без опоры на предшествующий опыт и без взгляда в будущее. Локализация не предполагает активного сотрудничества с людьми на интерактивной основе, а лингвистические результаты производимых исследований часто вызывают скуку. Целью данной статьи является представление перечисленных недостатков феномена локализации в виде одной глобальной проблемы; кроме того, будет сделана попытка предложить некоторые пути ее решения.

Для начала приведем два примера, не имеющих ничего общего с областью высоких электронных технологий. Студентам-переводчикам был дан следующий текст для перевода на английский язык:

Los alumnos que hayan estudiado en el extranjero y deseen iniciar estudios en las universidades españolas deberán convalidar u homologar sus estudios.

В итоге был получен вполне адекватный перевод:

Students who have studied outside Spain and wish to enter a program at a Spanish university must **convalidar** or **homologar** their foreign studies.

В процессе анализа этого перевода возникает закономерный вопрос: что делать со словами, выделенными полужирным шрифтом? Через несколько секунд поиска в Интернете студенты обнаружили параллельные тексты, в которых термин «аккредитация», как более обобщенный, совмещал в себе значения обоих испанских слов. Еще через несколько секунд учащиеся получили описания правил употребления специальных терминов *homologación* и *convalidación* в официальных документах.

Так, термином *homologación* обозначаются академические степени и дипломы, оформленные в виде документов, а термин *convalidación* отражает набор реальных дисциплин, изученных студентом за время обучения в вузе (в российской терминологии — «приложение к диплому»). Каким образом могли переводчики с помощью одного английского термина покрыть семантическое пространство, обозначенное двумя испанскими словами? Студенты начали поиск путей решения этой проблемы, но у них ничего не получалось до тех пор, пока не было предложено рассмотреть ее с позиций студентов-иностранцев. Был поставлен вопрос: чего они желают? Если им не нужна «аккредитация» в юридическом смысле этого слова (то есть официальное утверждение учебного плана), то в этом случае наличие двух испанских терминов является избыточным, и вместе они могут обозначать понятие обычной аккредитации. В том случае, когда аккредитация им необходима, они будут использовать оба испанских термина, сопровождая их описаниями соответствующих реалий. Таким образом, информация, содержащаяся в исходном тексте, может быть либо сокращена, либо увеличена без осуществления ненужной процедуры поиска двух английских эквивалентов для двух испанских терминов. В данном случае студенты-переводчики не думали о будущих читателях переведенного текста, а занимались интенсивным поиском словарных эквивалентов в лексиконах двух языков.

Анализ приведенного примера не добавляет ничего нового к существующим функциональным теориям перевода: то, о чем здесь говорится, обсуждается в научном мире на протяжении многих

десятилетий. Теория локализации, как и все прочие, призывает нас думать о будущем читателе результата процесса перевода. Тем не менее в рамках данной теории этого, как правило, не происходит: причина такого положения вещей та же, что и у студентов-переводчиков в примере, приведенном выше. Свободный доступ к изобилию информации (само это изобилие представляет собой актуальную проблему современности) означает, что переводчики работают над текстом *ради самого текста* с целью создания той или иной терминологической системы на переводном языке по заказу чиновников, которым она необходима, а отнюдь не для широкого круга читателей, нуждающихся в помощи при изучении научных и иных достижений иностранных авторов. В результате методика преподавания перевода сводится к элементарным поискам терминологических соответствий в двух языках, что дезориентирует учащихся, создавая у них неадекватное представление о роли переводчика в современном обществе. В этом заключается одна из разновидностей дегуманизации: *текст как объект изучения является более важным для переводчика, чем читатель как личность*.

В качестве второго примера дегуманизации можно привести буклет — инструкцию из мира рекламы, которая начинается следующим словами:

Welcome to Dragon NaturallySpeaking, the world's most acclaimed large-vocabulary continuous-speech dictation system.

With Dragon NaturallySpeaking you can dictate to your computer instead of using the computer to enter and revise text.

Буквально через пару страниц читаем тот же самый текст с незначительными изменениями.

Welcome to Dragon NaturallySpeaking, the world's most acclaimed large-vocabulary continuous-speech dictation system.

With Dragon NaturallySpeaking you can dictate to your computer instead of typing.

Что же происходит в этом случае? С какой целью дважды воспроизводятся почти идентичные тексты? Может быть, с целью уточнения смысла и устранения стилистических погрешностей? Ответы на поставленные вопросы дать крайне затруднительно, поскольку в обоих текстах содержится та же самая информация, а уточнения смысла и улучшения стиля не наблюдается. Этот факт не может

не озадачить читателя, поскольку он не понимает цели, ради которой приводится почти прямой повтор этого отрезка текста: как правило, при чтении человек воспринимает текст *линейно*, начиная с введения и заканчивая заключением. Оказывается, все дело в том, что читателю и не предлагается «линейное» ознакомление с текстом. Подобные «произведения» представляют собой всего лишь пособия (файлы) для облегчения работы в сети Интернет и предполагают *фрагментарное чтение* в зависимости от желания пользователя. По моему мнению, подобные тексты не следует рассматривать в качестве элементов культуры издательской деятельности, поскольку их читатель фактически является всего лишь оператором электронной поисковой системы, который погружается в фрагменты текста для мгновенного получения информации путем нажатия на ту или иную кнопку.

Этот пример имеет прямое отношение к феномену локализации, который в данном случае понимается как процесс адаптации некоего «обобщенного» текста ко всему разнообразию новых целей и сфер применения, возникающих в современном мире. Действительно, простой здравый смысл подсказывает, что гораздо целесообразнее адаптировать существующие материалы, чем создавать новые тексты для каждого конкретного случая. В крупных межъязыковых проектах исходная информация является подчеркнуто обобщенной («интернационализированной»), сознательно освобожденной от локальной зависимости с целью облегчения задачи ее перевода на большое количество языков одновременно.

В этом случае тексты становятся информационными объектами, то есть отрезками письменной речи, которые можно модифицировать и комбинировать электронным способом, причем они создаются в таком формате, который исключает всякую линейность восприятия. Таким образом, писательская деятельность превращается в процесс постоянной модификации уже написанного, авторы напрасно растрачивают силы, посвящая свою жизнь модернизации («синхронизации» по терминологии Е. С. Петровой [Петрова, с. 262]) и улучшению существующих материалов, а не созданию новых произведений (текстов). Перевод представляет собой следующий шаг в развитии этого процесса, так как он призван совершенствовать уже переведенные тексты. Это очень прогрессивный и экономически выгодный подход. Что может вызывать у нас сомнения в его эффективности?

В обоих приведенных примерах наблюдается серьезная обработка текста, которая локализуется в пространстве между отправителем и получателем информации во время акта коммуникации. Совершенно очевидно, что этот процесс имеет место в коммуникативных актах любого типа. В то же время в нашем случае такое посредничество предполагает, что получатель информации не имеет представления о ее отправителе, так как отправитель информации превратился в авторский коллектив «переписчиков текстов», которых мало интересуют сведения о получателе их продукции (в данном случае получатель информации — «типичный» представитель той или иной точки локализации: определенная страна, регион и т. п.).

С позиций отправителя отсутствие данных о получателе информации является важным фактором в деле организации проектов локализации. Специализированные агентства рассылают внешние заказы внештатным переводчикам, которые получают работу в виде готовых текстов и глоссариев к ним: и то и другое является для переводчиков обязательным. Использование банков переводческих данных значительно способствует укреплению позиций этого отсталого подхода, принуждая переводчиков повторять существующие материалы и исключая любую возможность серьезных размышлений о будущих читателях. В результате переводы становятся исключительно консервативными как с точки зрения стиля изложения, так и в отношении используемой терминологии. Иначе говоря, в переводах такого типа «нужные» слова всегда говорятся в «нужном» месте, а тот, кто их произносит, не должен знать, кому эти слова адресованы.

Это «сокрытие» адресата становится особенно интенсивным при смене формата перевода. В проектах программного обеспечения локализации обычно выделяются сочетания слов, звучащие естественно на том или ином языке и подлежащие переводу без особых затруднений со стороны специалистов, которые тем самым освобождаются от работы над тонкостями и сложностями процесса перевода.

Документы могут выдаваться переводчикам в виде текстовых файлов в формате программы Word с той же самой целью: максимально облегчить их работу над кодами исходных и переводных текстов. Таким путем переводчики отстраняются не только от наиболее выгодных сделок по проектам локализации, но и от любых потенциальных получателей переводимой информации. Их работа не

может быть предназначена для кого-либо, потому что они не знают, каким будет их потенциальный читатель: в настоящее время переводчики всего лишь повторяют то, что было уже многократно повторено. В результате мы получаем довольно однообразный и скверный перевод.

Таким образом, существующая практика локализации перечеркивает перспективы ее теории. Один из корифеев локализации Б. Эсселинк [Esselink] описывает будущий мир перевода, в котором специалисты работают исключительно над незначительными задачами обновления существующих текстов, при этом заказы от компаний, имеющих дело с локализацией, поступают эпизодически, в зависимости от необходимости более равномерного распределения объемов работы. В этом случае мы представляем себе переводчика будущего как домохозяйку, сидящую в четырех стенах, ежедневно переводящую отрывки из проектов. Такая фрагментарность заказов с неизбежностью означает, что переводчики никогда не увидят ни одного проекта в целом виде и не узнают, для чего конкретно была предназначена их работа.

Действительно, основной объем работы по локализации выполняется позднее, на этапах постредактирования, оценки качества перевода, многократной модернизации, которые неразрывно связаны с решением организационных вопросов в ходе реализации того или иного проекта. Обычно эта работа выполняется профессиональными сотрудниками компании, среди которых мы *можем надеяться* отыскать отправителя информации, заинтересованного в личности ее получателя, однако в реальном мире основными факторами, влияющими на процессы локализации, являются срок исполнения работы (the deadline), юридические формальности и требования рынка. Чрезвычайно редко в качестве такого фактора выступает конкретный человек.

Со стороны получателя информации локализация создает таких читателей, которых она заслуживает. Мы «читаем» с помощью поискового устройства с целью найти определенную страницу в сети Интернет. Никто не просматривает инструкцию систематически, страницу за страницей; мы обычно делаем три попытки решить проблему немедленно, затем, в случае неудачи, мы обращаемся за консультацией к справочному файлу Интернета, вновь читая с помощью указательного пальца. Для нас линейность повествования —

это потеря времени, *наша единственная цель — мгновенное получение информации*. Таким образом, мы превращаем сами себя в палец, способный нажимать на нужную кнопку.

Предположим, что технологии со стороны получателя информации дают возможность пользователям стать активными участниками процесса ее получения. При этом отсутствие повествования может предоставлять пользователю известную степень свободы действий: все могут создавать свои собственные варианты перевода (за исключением банальных «локализаторов»). В последнее время появилось много неофициальных программ, действующих параллельно в сети Интернет, что дает возможность проводить многочисленные неформальные дискуссии, посвященные практике программного обеспечения; сеть изобилует хакерами и пиратами, плетущими свои тайные интриги. Все это происходит исключительно по той причине, что официальная локализация как процесс не включает в себя человеческий фактор. Дегуманизация заключается не в самой технологии, которая может служить человеку универсально; она состоит в тех *способах применения* технологии, которые мы выбираем.

Подводя итоги, можно высказать предположение о том, что чем больше процедур обработки текста разделяют отправителя и получателя информации, тем выше степень дегуманизации коммуникативного процесса и тем больше людей будет стремиться к менее опосредованным способам человеческого общения. Что же следует предпринять? Часть ответа на этот вопрос дает сама практика процесса локализации. Компания Microsoft в высшей степени заинтересована реализовывать свою продукцию на местных языках и в рамках местных культур, глобальная сеть Интернет активно адаптируется к языкам и культурам своих пользователей, что позволяет надеяться на то, что человеческий фактор не будет полностью исключен из планов создателей компьютерных программ.

Мы живем в мире глобальных свершений, которые стремятся быть одновременно и локализованными, и гуманизированными (это в полной мере относится к достижениям переводческой индустрии). Несмотря на то что процесс локализации является довольно противоречивым, он, несомненно, открывает хорошие перспективы для развития тех или иных языков и культур. В этом заключается положительная роль феномена локализации в переводческой деятельности, которую надо оценить по достоинству.

С другой стороны, для развития *гуманитарного компонента перевода* следует использовать экономические возможности транзаказчиков и компаний, занимающихся проблемами локализации переводческой продукции. Гуманизация лучше всего проецируется на тот мир, в котором различные локализаторы и получатели информации делают попытку представить тексты как результат деятельности живых людей, а не просто в качестве обезличенных информационных объектов.

В то же время мы должны *сознательно принуждать* себя думать о безликих локализаторах-глобалистах как о живых людях, а не просто носителях информационных технологий, поскольку они своей кипучей деятельностью этого вполне заслужили. Изменение взглядов на проблемы локализации и дегуманизации в переводе (если таковому суждено произойти) должно, естественно, начинаться с программ подготовки переводчиков. Преподаватели могут оказать обществу реальную помощь в деле воспитания будущих переводчиков в духе приверженности человеческим ценностям наряду с формированием у них широкого спектра чисто технических умений и навыков.

Литература

Петрова Е. С. Локализация в межъязыковом и внутриязыковом переводе // Университетское переводоведение, вып. 2. Вторые Федоровские чтения. СПб., 2001.

Baker M. In Other Words. London; New York, 1992.

Brown G., Jule G. Discourse Analysis. Cambridge, 1983.

Esselink B. A Practical Guide to Localization. Amsterdam; Philadelphia, 2000.

А. В. Ачкасов

ЭКОЛОГИЧЕСКАЯ ВАЛИДНОСТЬ И РЕПРЕЗЕНТАТИВНОСТЬ ДАННЫХ КОМПЬЮТЕРНОГО МОНИТОРИНГА В ИЗУЧЕНИИ ПРОЦЕССА ПЕРЕВОДА

Аннотация. Специализированные методы компьютерного мониторинга позволяют испытуемым выполнять перевод в привычной для них среде и фиксировать все виды их активности, включая использование дополнительных ресурсов. Данные мониторинга характеризуются высокой экологической валидностью, унифицированностью (независимостью от индивидуальных психологических особенностей испытуемых), равномерностью распределения (фиксируются все действия переводчика) и репрезентативностью в отношении обращения переводчика к разным ресурсам. Методы компьютерного мониторинга фиксируют внешние проявления процесса перевода, поэтому возможность их использования для моделирования мыслительных процессов переводчика остается спорной.

Ключевые слова. Компьютерный мониторинг, процесс перевода, экологическая валидность, репрезентативность данных, унифицированность данных, равномерность распределения данных.

Andrei V. Achkasov

ECOLOGICAL VALIDITY AND REPRESENTATIVENESS OF COMPUTER MONITORING DATA IN TRANSLATION PROCESS RESEARCH

Abstract. Methods of computer monitoring allow users perform translation in regular environments and collect data on all types of translators' activity including the use of support resources. Monitoring data are characterized by high ecological validity, consistency (independence from individual psychological characteristics of subjects), uniformity of distribution and representativeness with respect to the use of different resources. The collected data represent external manifestations of the translation process. Their use for modelling mental processes remains controversial.

Keywords. Computer monitoring, translation process, ecological validity, data representativeness, consistency, uniformity of distribution.

Методы интроспекции используются в исследовании процесса перевода начиная с 1980-х годов. Наиболее известным и популярным является метод вербализации, или «мышления вслух» (think-aloud method; think-aloud protocol; далее — TA). Наряду с TA или в качестве дополнения к нему используются методы ретроспекции (интервью или опросники) и совместного обсуждения перевода несколькими переводчиками во время его выполнения (collaborative translation protocol). В 1990-е годы в изучении процесса перевода стали активно использоваться количественные методы, такие как запись активности пользователя на клавиатуре, или кейлоггинг (keylogging, keystroke logging; далее KL), запись экрана (screen recording), окулография, или айтрекинг (eye-tracking). Перечисленные методы нередко используются в разных комбинациях. Использование более двух методов сбора эмпирических данных получило наименование триангуляции процесса перевода [Hansen, 2003]. Существуют отдельные исследования, выполненные с применением методов томографии и энцефалографии, в частности в рамках проекта EYE-to-IT, финансируемого Европейской комиссией. В проекте были использованы методы айтрекинга, KL и энцефалографии.

Основными методами изучения процесса перевода являются TA и KL, остальные выполняют вспомогательные функции. Хотя есть исследования, выполненные целиком на материале данных ретроспекции и айтрекинга, они не позволяют моделировать перевод как процесс. Метод записи экрана позволяет получить максимально полный набор данных о действиях переводчика, однако сопряжен с проблемой сегментации данных (определение набора фиксируемых действий, фиксация и интерпретация «незавершенных» действий), а его применение ограничено трудоемкостью этого процесса. Указанные методы позволяют заполнить пробелы в наборах данных TA и KL, интерпретировать периоды «молчания» в TA или «бездействия» в KL, используются для категоризации и сегментации протоколов TA и т. д.

Основные методологические проблемы, связанные с применением TA и KL, касаются аспектов их репрезентативности, экологической валидности и интерпретации. В период, когда интроспекция была основным методом изучения процесса перевода, дискуссия развернулась вокруг методов интерпретации данных TA [Шадрин]. Основные усилия были направлены на выработку объективных

стратегий сегментации и категоризации протоколов ТА, появилось несколько конкурирующих интерпретационных моделей. С началом широкого применения методов компьютерного мониторинга, которые позволяют осуществлять сбор данных в естественных для переводчика условиях и минимизируют воздействие условий проведения эксперимента, на первый план вышли вопросы репрезентативности и экологической валидности. Основной вектор дискуссии в этом контексте направлен на обсуждение обратной зависимости между репрезентативностью и экологической валидностью данных. Методы компьютерного мониторинга позволяют получать более объективные данные о действиях переводчика, однако они фиксируют лишь его внешнюю активность, поэтому не могут быть эффективно интерпретированы для понимания «внутренних феноменов», мыслительных процессов, протекающих во время перевода. И напротив, данные, полученные при помощи методов интроспекции, дают существенно больше содержательной информации, однако имеют крайне низкую экологическую валидность, зависят от условий эксперимента и индивидуальных психологических особенностей испытуемых.

Обоснование экологической валидности ТА связано с тезисом Эрикссона и Саймона о том, что методы интроспекции не влияют на «ход мысли» (sequence of thoughts) испытуемых [Ericsson, Simon]. В исследованиях процесса перевода на основе ТА этот довод сводится к тому, что, так как перевод на родной язык представляет собой речевой вид деятельности по перекодировке информации, то «озвучивание мыслей» для переводчика является естественным процессом [Göpferich, Jääskeläinen].

Основные факторы, снижающие экологическую валидность методов интроспекции, и прежде всего ТА, неоднократно становились предметом обсуждения. Метод ТА может применяться только в условиях лабораторного эксперимента, что само по себе влияет на эмоциональное состояние, концентрацию внимания и речевое поведение переводчика. Устная вербализация действий в процессе письменного перевода не может быть «естественной», так как переводчик вынужден выполнять сразу два типа речевой деятельности. «Мышление вслух» является, в сущности, созданием устного текста-комментария к процессу создания письменного текста, что как минимум заставляет переводчика перераспределять

внимание, влияет на его речевое поведение и замедляет процесс перевода [Jakobsen, 2003; Hansen, 2005]. Озвучивание действий, которое в конечном итоге является формой их осмысления, ведет к снижению интуитивности перевода, может менять ход решения переводческих задач.

С целью повышения экологической валидности ТА минимизируется контакт между испытуемыми и наблюдателями, так как «в случае их прямого взаимодействия испытуемые пытаются адаптировать устную речь к социальной норме» [Hansen, 2013, p. 90], что может исказить данные. Вместе с тем для увеличения продуктивности ТА наблюдатель может вмешиваться в ход эксперимента, стимулируя испытуемого напоминаниями («продолжайте говорить») и вопросами. Согласно Эрикссону и Саймону, такая стимуляция оказывает минимальное воздействие на испытуемого [Ericsson, Simon].

Другой способ повышения экологической валидности ТА состоит в проведении предварительных тренингов, в ходе которых испытуемые «привыкают» к процессу вербализации своих действий и получают инструкции о том, что и как желательно озвучивать в ходе эксперимента. Результаты такого подхода остаются спорными, так как характер вербализации зависит от психологических особенностей переводчика.

По мере совершенствования мониторинговых программ и методов автоматической обработки данных высказывалось мнение, что KL и другие средства компьютерного мониторинга являются полноценной заменой ТА. Серьезной критике при этом были подвергнуты методы интроспекции, вплоть до полного отрицания их эвристического потенциала. Методологические проблемы ТА сжато обобщил Г. Тури: «Валидность данных интроспекции для изучения когнитивных процессов часто подвергалась сомнению, но большая часть возражений всегда удивительным образом опровергалась. Например, общим местом стало утверждение, что именно процесс перевода может наиболее эффективно исследоваться с помощью протоколов вербализации … и что „мышление вслух“ в ходе выполнения перевода является почти естественным процессом. <…> Даже если это так, в данном контексте мои аргументы направлены не на критику *психолингвистической валидности* данных, а на *релевантность* этого метода с точки зрения переводоведения» [Toury, p. 63].

Критика вызвала реакцию сторонников ТА, которые обновили старый аргумент о том, что ТА является одним из немногих методов, позволяющих «проникнуть в ход мысли испытуемого» (reveal a person's sequence of thoughts). Было даже высказано мнение, мотивированное развитием KL, что ТА является не чисто качественным методом, как это принято считать, а находится где-то между качественными и количественными методами, хотя и тяготеет к качественным методам [Sun]. KL в контексте таких исследований оценивается как вспомогательный метод, который не может быть использован для моделирования содержательных аспектов процесса перевода.

Процесс логгинга (logging) в самом общем виде представляет собой автоматическую фиксацию всех событий на компьютере. Этот принцип лежит в основе мониторинговых программ, фиксирующих активность пользователя, включая активность на клавиатуре и использование вспомогательных ресурсов. Программы этого класса позволяют перехватывать информацию из буфера обмена, делать снимки экрана через определенные промежутки времени, фиксировать обращение и ресурсам Интернета и т. д.

Первые версии программ KL, разработанные специально для мониторинга действий испытуемого в процессе перевода, имели ограниченный функционал. Это, в частности, касается наиболее известной и широко используемой программы Translog. При работе с этой программой перевод осуществляется в специальной среде, поэтому испытуемые не могут использовать привычные текстовые редакторы или программы класса «переводческая память». Translog фиксирует всю активность на клавиатуре, включая изменения, вносимые переводчиком в текст (удаление, добавление текста, вырезание и вставка символов и сегментов текста), движение курсора, а также время нажатия клавиш на клавиатуре, паузы между нажатиями. Кроме того, программа позволяет осуществлять видеозапись экрана. Несомненное достоинство программы — возможность автоматической обработки данных. Процесс перевода может быть представлен в виде линейной записи (графически или в виде набора символов), с указанием пауз и типов действий переводчика. В программе есть опция, позволяющая обобщать статистические данные по разным типам действий переводчика, а также опция записи экрана. В процессе мониторинга не фиксируется обращение пользователя к дополнительным ресурсам.

Сегодня существуют программы, позволяющее испытуемым работать в привычной среде и фиксирующие использование других приложений: веб-браузеров, словарей, средств работы с корпусами текстов и т. д. Приложение Inputlog, созданное для изучения процесса письма (создания письменных текстов), в дополнение к перечисленным функциям, позволяет делать запись экрана, распознавать речь, автоматически собирать статистическую информацию и синхронизировать разные данные [Van Waes, Leijten, Van Weijen]. Фиксировать все перечисленные типы данных позволяют и мониторинговые программы широкого назначения (за исключением функции распознавания речи), однако они не имеют функции автоматической обработки информации.

Высокая экологическая валидность экспериментов с применением KL неоднократно отмечалась как одно из главных достоинств этого метода. Компьютерный мониторинг процесса перевода протекает в естественных условиях, а использование программного обеспечения никак не влияет на процесс перевода. К общим факторам, влияющим на экологическую валидность KL, относят использование незнакомой пользователю операционной системы, непривычные для пользователя настройки компьютера и даже физические особенности клавиатуры [Saldanha, O'Brien, p. 135]. Эти факторы могут замедлять процесс перевода и влиять на использование тех или иных ресурсов, однако их влияние может быть снижено в процессе отбора кандидатов для участия в эксперименте, а с учетом возросшего уровня компьютерной грамотности и высокой адаптивности современных пользователей ими вообще можно пренебречь.

Использование в процессе эксперимента среды перевода Translog, которая представляет собой упрощенный текстовый редактор, лишенный многих привычных пользователю функций, оказывает более существенное влияние на экологическую валидность эксперимента. Эта среда не только создает ряд неудобств для пользователя, которые могут влиять на характер и последовательность его действий, но и является постоянным напоминанием о лабораторных условиях выполнения перевода. По мнению ряда исследователей, значение влияния особенностей работы в непривычной для переводчика среде минимально или несущественно. А. Якобсен отмечает, что в ходе эксперимента переводчики «забывали, что уча-

ствовали в эксперименте» и отмечали, что «выполнение перевода в среде Translog мало чем отличалось от выполнения обычного перевода» [Jakobsen, 1999, p. 15]. Другие исследователи полагают, что выполнение перевода в среде Translog и аналогичных программ существенно снижает валидность эксперимента. В. Нойнциг считает, что выполнение перевода в условиях, когда переводчику известно, что он участвует в эксперименте и при этом не ведется запись его обращения к другим ресурсам (что также может влиять на его поведение), делает условия перевода «нереалистичными» [Neunzig, p. 96]. Среда Translog не позволяет, в частности, менять исходный текст, в то время как, работая со стандартным текстовыми редакторами, переводчики нередко переводят «поверх» исходного текста, форматируют его, используют выделения и т. д. [Antônio, Silva, p. 76]. По мнению Б. Димитровой, «о высокой степени экологической валидности эксперимента можно говорить при использовании программного обеспечения, которое позволяет испытуемым работать в привычной среде. Большинство логгинг-программ создаются преимущественно для исследовательских целей, по крайней мере, более ранние версии имели функции простого текстового процессора, без возможностей форматирования или поиска. Не позволяли они исследовать и то, как испытуемые используют функции проверки грамматики и орфографии» [Dimitrova, p. 75–76]. Применение программ, предоставляющих переводчикам для работы привычную среду (текстовый редактор, программы класса «переводческая память»), таких как Inputlog или BB FlashBack, позволяет утверждать, что проблема экологической валидности, связанная с техническими особенностями программного обеспечения, полностью решена.

Не менее серьезной является проблема осведомленности переводчика об участии в эксперименте. «…Ощущение, что за тобой наблюдают, сказывается на результатах исследования. Оно создает нервозность, стрессовую ситуацию, влияющую на мыслительные процессы испытуемых, и степень этого влияния не может быть измерена точно» [Hansen, 2005, p. 97]. Этот вопрос почти не исследован, что обусловлено этическим аспектом осведомленности переводчика об участии в эксперименте. Сложившаяся на сегодня практика требует привлекать к участию в эксперименте только волонтеров и подробно информировать их о том, какие виды их активности будут

фиксироваться. Данные о влиянии осведомленности переводчика об участии в эксперименте на его активность были получены в ходе эксперимента по изучению распределения типов действий в процессе терминологического поиска, проведенного в Лаборатории письменного перевода СПбГУ [Ачкасов, 2011b]. Хотя этот аспект анализа не входил в основные задачи эксперимента, было установлено, что существуют определенные различия между действиями тех испытуемых, которые выполняли перевод в режиме стандартного переводческого теста (группа 1), и тех, которые были проинформированы о целях и условиях эксперимента (группа 2). За первые 30 минут перевода испытуемые группы 2 перевели существенно меньший объем текста (в среднем на 20%), в их активности было больше периодов бездействия, они чаще возвращались к исходному сегменту и более тщательно верифицировали терминологические эквиваленты. К середине эксперимента эти различия уже не наблюдались, скорость перевода и распределений действий переводчиков обеих групп выровнялись. Различия сохранились лишь по двум параметрам: переводчики группы 2 в большей степени использовали рекомендованные ресурсы, в частности специализированный корпус, и не пользовались социальными сетями. Хотя использование социальных сетей для общения, не связанного с выполнением перевода, не имеет отношения к содержанию эксперимента, такая форма «отвлечения» от основного типа деятельности является типичной для письменных переводчиков и поэтому также влияет на экологическую валидность его результатов. Указанные наблюдения не являются бесспорными, так как при регулярном выполнении перевода в условиях эксперимента указанные отличия могут нивелироваться.

KL не рассматривается сегодня как полноценная альтернатива ТА главным образом по причине недостаточной репрезентативности данных мониторинга для анализа мыслительных процессов переводчика. Критика этого аспекта KL сводится к двум основным аргументам: 1) в ходе мониторинга не всегда фиксируются действия переводчика, выполняемые за рамками специализированного программного обеспечения; 2) фиксируется не процесс перевода, а только процесс ввода текста. Б. Димитрова отмечает: «Хорошо известно, что в отношении распределения времени в процессе перевода собственно запись переведенного текста не является преобладающим видом дея-

тельности. Большую часть времени в процессе перевода переводчик занят другими видами деятельности, такими как чтение, просмотр разнообразных ресурсов и т. д. Поэтому при исследовании процесса перевода логгинг-данные надо использовать с осторожностью. Более того, важно понимать, что при логгинге фиксируется не процесс перевода per se, а процесс записи перевода» [Dimitrova, p. 75].

Применение программ с функционалом Inputlog или даже мониторинговых программ широкого профиля снимает первую проблему, связанную с репрезентативностью данных, полученных методом KL. Такие действия, как чтение, просмотр дополнительных ресурсов, использование специализированных корпусов, поиск лингвистической и фоновой информации в Интернете и т. д., отражаются в мониторинговых данных. Обращение переводчика, например, к интернет-ресурсу может быть отражено в виде текста, который переводчик ввел в веб-браузере, ссылки на ресурс, скриншота страницы ресурса и информации в буфере обмена, в случае если переводчик ее скопировал с ресурса.

Существует мнение, что часть данных, фиксируемых в процессе KL, не имеют прямого отношения к процессу перевода. Это, в частности, относится к операциям, которые переводчики совершают с исходным текстом: форматирование, выделение, добавление комментариев и т. п. [Antônio, Silva, p. 76]. Этот вопрос является как минимум спорным, так как указанные операции могут отражать «ход мысли» переводчика. Определенное значение могут иметь характер и способы перевода выделенных фрагментов текста, лингвистический и информационный поиск, сопряженный с переводом этих фрагментов. В ходе перевода «поверх исходного текста» переводчик может использовать разные стратегии, например переводить текст не последовательно, а обращаться к его разным частям, начинать с поиска терминологических эквивалентов и замещать исходные термины переводными и т. д. Традиционно полагают, что перевод «поверх текста» характерен для непрофессиональных переводчиков, однако на практике так работают и многие профессионалы — в силу многолетней привычки или того, что использование программ класса «переводческая память» неэффективно для тех текстов, с которыми они работают.

Так как современные мониторинговые программы позволяют испытуемому работать в любой среде, появились исследования

процесса перевода и с использованием переводческой памяти [Christensen], а также исследования в области постредактирования машинного перевода. Хотя редактирование совпадений из переводческой памяти или машинного текста не является переводом в традиционном смысле, следует учитывать, что именно эти виды работы сегодня преобладают на рынке перевода. Применение мониторинговых программ позволяет, кроме того, исследовать процесс перевода в условиях ограниченного времени, что в принципе невозможно с использованием ТА.

Второй аспект репрезентативности данных мониторинга связан с тем, что они фиксируют лишь внешние проявления деятельности переводчика, позволяют «точно измерять действия, выполняемые в процессе перевода на микроуровне» [Saldanha, O'rien, p. 135], измерять распределение времени на выполнение разных видов деятельности, в то время как «внутренняя работа» испытуемого не отражается в данных KL. Иными словами, эти данные показывают «что» и «когда» делал переводчик, позволяют восстановить последовательность его действий, но не дают ответа на вопрос, «почему» он поступал именно так. Действительно, сопоставление данных ТА и KL, полученных в ходе одного эксперимента, позволяет утверждать, что периоды интенсивной вербализации дают более содержательные данные о мотивации действий переводчика. Это можно проиллюстрировать конкретным примером на материале одного предложения (подробнее о проведенном эксперименте см. [Ачкасов, 2011a]).

Оригинал: **Map your brandscape to discover white space.**

Перевод: **Сделать карту бренда, чтобы выявить на ней белые пятна.**

	ТА	KL
1	Brandscape — неологизм, в словарях нет смысла проверять. <5>. Можно перевести буквально «ландшафт бренда», учитывая, что там еще и глагол *map*	Idle
2	<4> наносить на карту; чертить карту; производить съемку местности; <2> бизнес, составлять схему, изображать, отражать	Multitran (map)

	TA	KL
3	<3> Составить карту ландшафта <2> сделать карту <15> закартографировать ландшафт бренда <3> странная метафора, но в брендинге по ходу все так	*Составить карту ландшафта бренда*
4	<5> В широком смысле “to sketch or plan”, например “to map out a new career” <2>	Dictionary.com (map)
5	Сделать карту бренда? <5>	~~Составить~~ {Сделать} карту ~~ландшафта~~ бренда
6	Чтобы выявить белые пятна	*, чтобы выявить белые пятна*
7	<6> Ага, «карта бренда», есть совпадения в Интернете, а “brand mapping” — это «картирование бренда». <2> В общем, лучше «сделать карту бренда», по смыслу понятно и более естественно звучит	Google (карта бренда) → http://www.purebrand.ru/brandmapping/
8		, чтобы выявить {на ней} белые пятна

В приведенном примере вербализация представлена относительно протяженными отрезками текста. Данные KL позволяют просмотреть ресурсы, к которым обращался переводчик, тогда как TA дает представление о выборочности внимания переводчика. Так, в сегменте № 2 испытуемый зачитывает варианты перевода глагола *map* (словарная статья), а в данных мониторинга зафиксированы только скриншот и время, которое переводчик оставался на сайте словаря. В целом из данных TA понятен ход рассуждений переводчика, мотивирован ряд принятых решений (сегменты № 1, 3, 7). Данные KL позволяют получить полное представление о ресурсах, к которым обращался переводчик, и о последовательности ввода текста и редактирования. Последний этап редакторской правки в TA не отражен.

Такое соотношение данных TA и KL не является типичным, что обусловлено интенсивностью вербализации, которая зависит как минимум от трех факторов: интенсивности ментальной активности, квалификации переводчика и его индивидуальных особенностей.

Неоднократно отмечалось [Ericsson, Simon], что периодам «молчания» в TA соответствует наибольшая мыслительная активность

испытуемых. Проблему «молчания» неоднократно пытались решить, запрещая переводчикам использовать информационные и лингвистические ресурсы при переводе, «чтобы интенсифицировать процессы решения проблем» [Bernardini, 2001, p. 178], или стимулируя вербализацию устными напоминаниями, что не только влияет на экологическую валидность эксперимента, но и снижает качество перевода, так как переводчик вынужден принимать решения без необходимой ему лингвистической или фоновой информации. На интенсивность и распределение вербализации в ТА влияет и квалификация переводчика. Активное использование словарей и других ресурсов характерно для неопытных переводчиков [Jensen, p. 113; Lee-Jahnke]. В целом чем опытнее переводчик и чем лучше он владеет тематикой перевода, тем ниже степень вербализации и выше степень автоматизации процесса перевода. Третий фактор, влияющий на интенсивность вербализации, — индивидуальные особенности переводчика. Данные ТА гораздо более персонализированы, чем данные KL. Существуют «индивидуальные различия в способности к вербализации, что может существенно повлиять на получаемые данные» [Bernardini, 1999, p. 182]. Хотя два разных переводчика, сталкиваясь с одной и той же проблемой (например, отсутствие в словаре или глоссарии терминологического эквивалента), выполняют сходные действия по решению этой проблемы и нередко в итоге приходят к сходным или идентичным результатам, они могут совершенно по-разному озвучивать свои действия.

Все перечисленные факторы не влияют на данные мониторинга, поэтому можно говорить о существенно большей степени их равномерности и унифицированности по сравнению с данными ТА. Равномерность проявляется в наличии данных KL независимо от интенсивности вербализации, периодов «молчания» и квалификации переводчика. Так, при переводе выражения «foundational brand» были зафиксированы следующие данные ТА и KL (см. с. 71).

На основе данных ТА можно утверждать, что: 1) эквивалент «основополагающий» показался переводчику «странным»; 2) он хотел отложить решение этого вопроса; 3) рассматривал возможность опустить компонент *foundational*; 4) в итоге использовал эквивалент «базовый семинар».

По данным мониторинга можно восстановить ход действий переводчика, который: 1) ввел фразу «на этом семинаре»; 2) нашел рус-

1	На этом семинаре <29>	*на этом семинаре*
		Multitran (foundational)
2	основополагающем семинаре? Странно, это потом. Можно вообще опустить <12>	Multitran (основополагающий)
3	базовый семинар <15>	На этом {базовом} семинаре
		Google («базовый семинар» = 41 100 hits)

скоязычный эквивалент слова *foundational* в словаре; 3) выполнил обратный перевод найденного эквивалента (*основополагающий*); 4) отредактировал первоначальную версию перевода («на этом базовом семинаре»); 5) после редактирования верифицировал эквивалент «базовый семинар». В ТА все это не отражено и лишь зафиксированы периоды «молчания» (29, 12 и 15 секунд).

В целом можно утверждать, что данные мониторинга фиксируют действия переводчика как минимум более равномерно, кроме того, они более репрезентативны в отношении фиксации обращений переводчика к разным ресурсам. Даже в тех случаях, когда переводчик озвучивает подобные действия, данные KL оказываются более полными, так как фиксируют ресурс, к которому обращался переводчик, и текст запроса. Унифицированность данных мониторинга характеризуется их независимостью от индивидуальных психологических особенностей испытуемых и степени их готовности к эксперименту. Благодаря равномерности и унифицированности данных мониторинга они более объективно показывают сходство и различия переводческих стратегий и в этом отношении легче поддаются обобщению.

Единственным бесспорным преимуществом данных интроспекции по сравнению с данными мониторинга, таким образом, является их содержательность в отношении мотивации деятельности испытуемых, которая и позволяет интерпретировать собственно процесс перевода, а не его «запись». Действительно, данные KL a priori не могут отражать аргументацию переводчика и его комментарии. Этот же аргумент, однако, в несколько модифицированном виде неоднократно высказывался и в отношении ТА. Вербализован может быть не процесс перевода как таковой, а его определенные результирующие стадии, состояния обработки информации (information

states): «Так как вербализация отражает состояния обработки информации, то глубинные процессы мышления моделируются исследователем» [Dimitrova, p. 69]. Из этого следует, что данные ТА дают более эксплицитные описания состояний обработки информации и, следовательно, больше возможностей для содержательного обобщения данных.

Однако именно на этапе категоризации и обобщения данных возникают новые претензии к ТА. Этот вопрос неоднократно поднимался в исследовательской литературе, и его подробное рассмотрение не входит в задачи настоящего исследования. В этом отношении достаточно сослаться на мнение Бернардини: «Основная проблема исследований на основе ТА состоит в отсутствии единой исследовательской парадигмы, что ведет к произвольной трактовке методологических вопросов (проведение исследования, анализ данных, обобщение данных) и появлению многочисленных исследований, в которых предлагается категоризация данных в теоретическом вакууме» [Bernardini, 1999]. На практике это проявляется в том, что категоризация данных ТА дает тривиальные результаты. Все индивидуальные особенности ТА формализуются, например как разные типы индикаторов (problem indicators), и обобщаются в виде «переводческих стратегий», при этом данные интроспекции, характеризующие индивидуальные особенности переводчика или его отношение к конкретным проблемам, нивелируются.

В данной статье обобщены результаты дискуссий, касающихся экологической валидности и репрезентативности данных компьютерного мониторинга в изучении письменного перевода. Чаще всего обсуждение этих вопросов сводится к анализу отдельных аспектов валидности и репрезентативности данных, условий их обеспечения в рамках конкретных экспериментов или методов их обобщения и интерпретации. В настоящее время назрела необходимость комплексного обсуждения этих вопросов.

Литература

Ачкасов А. В. Аспекты репрезентативности TAPs и LOGGING DATA при изучении процесса перевода // Проблемы современного переводоведения: сборник статей в честь 60-летия профессора В. И. Шадрина / сост. А. В. Ачкасов; отв. ред. И. В. Недялков. СПб., 2011а. С. 11–27.

Ачкасов А. В. Распределение типов действий в процессе терминологического поиска (на материале лог-данных) // Университетское переводоведение. Вып. 11. Материалы XI Международной научной конференции по переводоведению «Федоровские чтения» (20–23 октября 2010 г.). СПб., 2011b. С. 48–66.

Шадрин В. И. «Мышление вслух» как метод исследования процесса перевода // Материалы II Международной научной конференции по переводоведению «Федоровские чтения» (23–25 октября 2000 г.). СПб., 2001. С. 387–394.

Antônio I., Silva L. On a more robust approach to triangulating retrospective protocols and key logging in translation process research // Psycholinguistic and Cognitive Inquiries into Translation and Interpreting / eds A. Ferreira, J. W. Schwieter. Amsterdam; Philadelphia, 2015. P. 175–201.

Bernandini S. Think-aloud protocols in translation research: Achievements, limits, future prospects // Target. 2001. Vol. 13, N 2. P. 241–263.

Bernardini S. Using think-aloud protocols to investigate the translation process: Methodological aspects // Research Centre for English and Applied Linguistics. Working Papers. 1999.Vol. 6. P. 182.

Christensen T. P. Studies on the Mental Processes in Translation Memory-assisted Translation — the State of the Art // trans-kom. 2011. Vol. 4, N 2. P. 137–160.

Dimitrova B. Expertise and Explication in the Translation Process. Amsterdam; Philadelphia, 2005.

Ericsson K. A., Simon H. A. Protocol analysis: Verbal reports as data. 2nd ed. Cambridge, 1993.

Göpferich S., Jääskeläinen R. Process Research into the Development of Translation Competence: Where Are We, and Where Do We Need to Go? // Across Languages and Cultures. 2009. Vol. 10, N 2. P. 169–191.

Hansen G. Controlling process: Theoretical and methodological reflections on research in to translation processes // Triangulating Translation: Perspectives in process oriented research / ed. by F. Alves. Amsterdam; Philadelphia, 2003. P. 25–42. (Benjamins Translation Library; vol. 45.)

Hansen G. Experience and Emotion in Empirical Translation Research with Think-Aloud and Retrospection // Meta: Translators' Journal. Montréal, 2005. Vol. 50, N 2. P. 511–521.

Hansen G. The Translation Process as Object of Research // The Routledge Handbook of Translation Studies / eds C. Millán-Varela, F. Bartrina. London, New York, 2013. P. 88–101. (Routledge Handbooks in Applied Linguistics)

Jakobsen A. L. Effects of think aloud on translation speed, revision, and segmentation // Triangulating Translation: Perspectives in Process Oriented Research / ed. by F. Alves. Amsterdam; Philadelphia, 2003. P. 69–95.

Jakobsen A. L. Logging target text production with Translog // Probing the Process In Translation: Methods and Results / ed. by G. Hansen. Copenhagen, 1999. P. 9–20.

Jensen A. Time Pressure in Translation // Probing the Process in Translation: Methods and Results / ed. by G. Hansen. Copenhagen, 1999. P. 113.

Lee-Jahnke H. L'introspection à haute voix: recherché appliquée // J. Delisle, H. Lee-Jahnke. Enseignement de la traduction et traduction dans l'enseignement. Ottawa, 1998. P. 164–165.

Neunzig W. Empirical Studies in the Didactics of Translation — The Computer as an Instrument for Standardising Input and Ensuring Environmental Validity // Investigating translation / eds A. Beeby, D. Ensinger, M. Presas. Amsterdam; Philadelphia, 2000. P. 91–98.

Saldanha G., O'Brien S. Research Methodologies in Translation Studies. Manchester, 2013.

Sun S. Think-Aloud-Based Translation Process Research: Some Methodological Considerations // Meta: Translators' Journal. Montréal, 2011. Vol. 56, N 4. P. 928–951.

Toury G. Experimentation in Translation Studies: achievements, prospects and some pitfalls // Empirical Research in Translation and Intercultural Studies: Selected papers of the TRANSIF Seminar, Savonlinna 1988 / ed. by S. Tirkkonen-Condit. Tübingen, 1991. P. 45–66.

Van Waes L., Leijten M., Van Weijen D. Keystroke logging in writing research. Observing writing processes with Inputlog // German as a Foreign Language (GFL). 2009. N 2–3. P. 41–64.

И. В. Азарова, Е. Л. Алексеева

ИСПОЛЬЗОВАНИЕ АППАРАТА КРИТИЧЕСКОГО ИЗДАНИЯ ЧЕТВЕРОЕВАНГЕЛИЯ ДЛЯ СОЗДАНИЯ КОРПУСА СЛАВЯНСКИХ ПЕРЕВОДОВ ЕВАНГЕЛИЯ[1]

Аннотация. В статье рассматриваются этапы изучения славянских переводов Евангелия и результаты исследования, приведенные в издании Четвероевангелия в группе А. А. Алексеева (Славянском проекте). На базе разработанной аннотации критического аппарата издания создается структурированный корпус славянских вариантов Евангелия в программе Паратекст.

Ключевые слова. Славянский перевод, Евангелие, критический аппарат, схема аннотации, структурированный корпус.

Irina V. Azarova, Elena L. Alekseeva

APPARATUS OF THE CRITICAL EDITION OF THE SLAVIC TETRAEVANGELION AS THE BASIS FOR THE CORPUS OF THE SLAVIC VERSIONS OF THE GOSPELS

Abstract. Having identified the milestones of the studies into the history of the Slavic version of the Gospels and having summed up the essential features of the critical edition of the Slavic Gospels prepared by Prof. Anatoly Alexeev and his team, the article goes on to consider the possibilities of transforming the electronic version of the edition into a structured corpus of the Slavic Gospels.

Keywords. Slavic version, Gospels, critical apparatus, annotation pattern, structured corpus.

[1] Работа выполняется при финансовой поддержке СПбГУ, НИР 2/14 «Четвероевангелие в славянской традиции» (шифр ИАС 31.38.285.2014).

1. Изучение славянских переводов Евангелия

История масштабного изучения славянского перевода Евангелия насчитывает более 100 лет, и точка здесь пока не поставлена. Назовем основные вехи.

1896 — Воскресенский Григорий Александрович (1849–1918, профессор Московской духовной академии, чл.-корр. РАН) удостоен степени доктора богословия за представленные им труды [Воскресенский, 1894; Воскресенский, 1896]. На основании исследования более 100 рукописей славянского Евангелия им выделены четыре редакции славянского перевода: первая редакция — древняя, вторая — полные апракосы, третья — Чудовский Новый Завет, четвертая — поздняя.

1915 — по инициативе профессора Петроградской духовной академии Ивана Евсеевича **Евсеева** (1868–1921) при академии создана Комиссия по научному изданию славянской Библии, в которую вошел весь цвет русской историко-филологической науки того времени. Комиссия намеревалась, начиная с Ветхого Завета, за 60 лет издать полный текст славянской Библии в критическом освещении. Первая мировая война и революция сделали невозможным осуществление этого плана: хотя в 1918 г. Комиссия перешла в ведение Академии наук, к 1921 г. ее деятельность практически прекратилась [Дудинов].

1929 — на Первом съезде филологов-славистов в Праге ставится задача критического издания славянского Евангелия [Meillet, Vaillant].

1935–1936 — работы чешского слависта Йозефа **Вайса** (Josef Vajs, 1865–1959, профессор Пражского университета, действительный член Чешской АН) преследовали цель реконструировать старославянский текст Евангелий и установить греческий оригинал кирилло-мефодиевского перевода [Evangelium, 1935a; Evangelium, 1935b; Evangelium, 1936a; Evangelium, 1936b]. Реконструкция текста основана на данных восьми древнейших рукописей с частичным использованием данных еще шести рукописей. Хотя издание имело большое значение для своего времени, оно было подвергнуто серьезной критике [Garzaniti].

1970 — Лидией Петровной **Жуковской** (1920–1994, доктор филологических наук, главный научный сотрудник РГБ) защищена

докторская диссертация, посвященная типологии славянских рукописей Евангелия. Л. П. Жуковская выполнила от начала до конца полную классификацию примерно 500 рукописей Евангелия XI — начала XV в. [Жуковская, 1968; Жуковская, 1976].

1985 — Анатолий Алексеевич **Алексеев** (профессор СПбГУ, заведующий кафедрой библеистики) впервые применил к славянскому материалу метод американского исследователя греческой новозаветной традиции Э. Колвелла, предложившего вместо сравнения рукописей с эталоном сравнивать их попарно между собой и оценивать их близость по общему тексту, а не по различиям [Colwell; Алексеев, Лихачева]. С помощью калькулятора был проведен кластерный анализ 19 рукописей Апокалипсиса по 246 узлам разночтений. Позднее под руководством А. А. Алексеева было исследовано еще несколько славянских памятников этим же методом, но уже с использованием компьютера [Алексеев, Кузнецова; Пичхадзе; Афанасьева; сюда же можно отнести: Гребенников; Миронова, 2005; Миронова, 2012; Рогозина].

1985 — А. А. Алексеев опубликовал Проект текстологического исследования кирилло-мефодиевского перевода Евангелия [Алексеев]. Основные положения таковы.

- Данные древнейших списков славянского Евангелия далеки от единодушия, поэтому успехов в реконструкции не достигнуто.
- Так как традиция является контролируемой, разночтения оказываются в самых различных сочетаниях. Л. П. Жуковская, исследовав 150 рукописей, признала невозможным выявить закономерность в хаотическом смешении разночтений, поэтому нельзя построить генеалогическую стемму.
- Отсюда следует, что нужно подготовить критическое издание. Но каковы должны быть принципы отбора списков и разночтений? Следует:
  — взять все списки XI — начала XV в. и по методу Колвелла провести их группировку;
  — выявить средний, типичный список и положить его в основу издания;
  — центральные списки выделенных групп использовать в критическом аппарате.
- Выделение группировок списков существенно упрощает нахождение чтений, которыми группы противопоставлены

друг другу, что делает возможным качественный анализ чтений и установление генетических отношений между группами, поэтому реальной становится задача реконструкции архетипа.

«При охвате 500 списков необходимо будет провести свыше 100 000 расчетов. Долгие годы уйдут на коллации рукописей в древлехранилищах СССР и славянских стран, на подготовку материалов для машины, на обработку и интерпретацию полученных после пересчетов данных. Однако уклониться от этой задачи нельзя…» [Алексеев, с. 94].

1993 — начало работы над Славянским проектом при финансовой поддержке Немецкого библейского общества.

Были проведены коллации двух пассажей — из Евангелия от Иоанна и Евангелия от Матфея по 1154 рукописям и четырем изданиям. Из этих 1154 рукописей 532 находятся в Санкт-Петербурге, 375 — в Москве, 46 — в других городах России, 77 — на Украине, 62 — на Афоне, 35 — в Сербии и 27 — в других странах.

В результате коллаций в выбранном пассаже из Евангелия от Иоанна было выявлено 330 узлов разночтений — мест, где текст в рукописях не совпадал, в пассаже из Евангелия от Матфея — 545 узлов.

Кластерный анализ для текста Евангелия от Иоанна позволил выделить восемь групп рукописей, для Евангелия от Матфея — пять.

1998 — опубликовано [Евангелие от Иоанна в славянской традиции].

2005 — опубликовано [Евангелие от Матфея в славянской традиции] при финансовой поддержке Синодальной библиотеки РПЦ и РГНФ.

2014 — работа над изданием Евангелия от Марка и Евангелия от Луки финансируется СПбГУ.

2. Основные принципы представления материала в изданиях Славянского проекта

В качестве базового текста было выбрано Мариинское Евангелие[2]. Привлечение большого круга источников, в том числе новых, позволило создать новую теорию текста, согласно которой первона-

[2] Глаголическая рукопись XI в., хранящаяся в Российской государственной библиотеке под шифром Григ. 6 / Муз. 1689 (ф. 87).

чальный перевод охватывал Четвероевангелие и не ограничивался апракосом, как считалось ранее. Мариинское Евангелие — единственная древняя рукопись Тетра, в которой сохраняется первичный текст; хотя в нем есть поновления, но он не прошел систематической правки по новому греческому оригиналу и находится в тесной связи с утраченным архетипом. Положенное в основу издания, Мариинское Евангелие сделало возможным реконструкцию истории славянского евангельского текста.

В критическом аппарате издания Евангелия от Иоанна последовательно цитируются 28 рукописей, в издании Евангелия от Матфея — 27.

Текст Мариинского Евангелия представлен в кириллической транслитерации без нормализации орфографии. Орфография источников, цитируемых в критическом аппарате, нормализована за исключением тех случаев, когда невозможна однозначная интерпретация:

- ошибка в рукописи: рече: *ре же* (Архангельское Евангелие, Ин 12:41);
- неоднозначное сокращение: чл҃в(к)а: *чловѣ(к)* (Мирославово Евангелие, Мф 9:9);
- региональные особенности орфографии: знаѭтъ мѧ моѭ҃: з. *и ме моє* (сербская рукопись Литургического тетра, Ин 10:14).

Во всех случаях, когда в аппарате воспроизводится орфография оригинала, используется особый шрифт.

Особенности цитирования чтений апракосов

В отличие от Четвероевангелия, или Тетра, где полностью и последовательно воспроизводится текст всех четырех Евангелий — от Матфея, Марка, Луки, Иоанна, в апракосах (лекционариях) отрывки из Евангелий следуют в том порядке, в каком они читаются в церкви в течение церковного года.

Церковный календарь и с ним вместе церковная служба состоят из двух циклов, неподвижного и подвижного, каждый из которых охватывает год целиком.

Неподвижный цикл по происхождению представляет собою старый римский и затем византийский гражданский календарь, он начинается 1 сентября. По неподвижному циклу отмечаются

праздники, привязанные к конкретным календарным датам (Рождество, Преображение, памяти святых…).

Подвижный (пасхальный) календарный цикл делится на пять периодов:

1) от Пасхи до Пятидесятницы (Духова дня), продолжительностью 50 дней;

2) от Пятидесятницы до Нового лета (т. е. нового года), приходящегося на начало сентября, продолжительностью 16–17 недель в зависимости от даты Пасхи;

3) от Нового лета до Великого поста, продолжительностью 16–17 недель в зависимости от начала поста (его начало на семь недель предшествует Пасхе);

4) Великий пост продолжительностью 6 недель;

5) Страстная седмица.

В апракосах сначала идут чтения подвижного цикла, затем неподвижного — месяцеслов.

Полный апракос содержит чтения на каждый день года за исключением Великого поста, когда литургии служатся только по субботам и воскресеньям, а краткий — только на субботы и воскресенья за исключением первого и пятого периодов, когда в нем есть чтения на каждый день, поэтому в кратком апракосе текст Евангелий представлен не полностью.

В апракосах некоторые отрывки из Евангелия повторяются 2–3 раза, при этом текст, как правило, неидентичен, а иногда имеет и существенные различия. В издании предусмотрены пометы, показывающие, какой части апракоса принадлежит данный отрывок: *s* — перикопы от Пасхи до Великого поста, *f* — чтение Великого поста, *p* — Страстные Евангелия, *mt* — перикопы утрени, *m* — перикопы месяцеслова.

В апракосе нередко начало перикопы повторяет последний стих предыдущей, в таких случаях для соответствующих стихов используются индексы *i* — начало перикопы и *e* — конец перикопы.

Например, в аппарате Евангелия от Матфея для стиха 24:43 приведены следующие разночтения для Юрьевского Евангелия (полный апракос), в котором этот стих встречается три раза [Евангелие от Матфея в славянской традиции, с. 133]:

храма своего: **храмины своєѣ** Ju^{si}, **клѣти** Ju^{se}, **домоу своєго** Ju^{f}.

Помета *si* в данном случае означает начало перикопы, читаемой в 11-ю пятницу по Пятидесятнице, *se* — конец перикопы, читаемой в 16-ю субботу по Пятидесятнице, *f* — перикопу, читаемую в Великий вторник на литургии (в критическом аппарате приводятся подробные указания деления текста апракосов на перикопы).

Последовательное отражение в издании при помощи особых помет и шрифтов целого ряда сведений делает возможной его конвертацию в структурированный корпус цитируемых в издании рукописей, что позволит подвергнуть их дальнейшей компьютерной обработке, уточнить состав выделенных групп, проверить их стабильность на протяжении всего текста и начать, таким образом, новый этап исследования рукописей.

3. Структурированный корпус славянских вариантов Евангелия

Преобразование текста издания в структурированный корпус предполагается выполнить с использованием программы Паратекст (http://paratext.org/), которая предназначена для работы с библейскими текстами и содержит специальные маркеры для аннотации фрагментов текста, обеспечивающие навигацию по библейским текстам (книгам, главам, стихам, фрагментам стихов) и характеристику текстовой структуры (основной заголовок, вступление, поэтические строки, обычный абзац и т. п.). Более того, можно вводить новые теги, которые позволят проводить поиск и редактирование, как в нашем случае, фрагментов критического издания.

На рис. 1 приведен пример аннотации критического издания к тексту Мф 2:14 в программе Паратекст. В верхней части окна показан текст Типографского Евангелия[3] (ТЕ) с маркерами стандартной разметки (номера стихов Мф 2:14–21). В данном варианте визуализации теги специально выделены цветом и размером, чтобы не мешать восприятию текста.

В нижней части окна представлены поля критического аппарата, привязанные к тексту ТЕ, при этом размеры окон могут по желанию

[3] В тексте Мариинского Евангелия отсутствует начало Евангелия от Матфея (1:1–5:23), поэтому в издании этот большой пассаж приводится по Типографскому Евангелию (РГАДА, ф. 381, 1, XII в.), очень близкому по тексту Мариинскому.

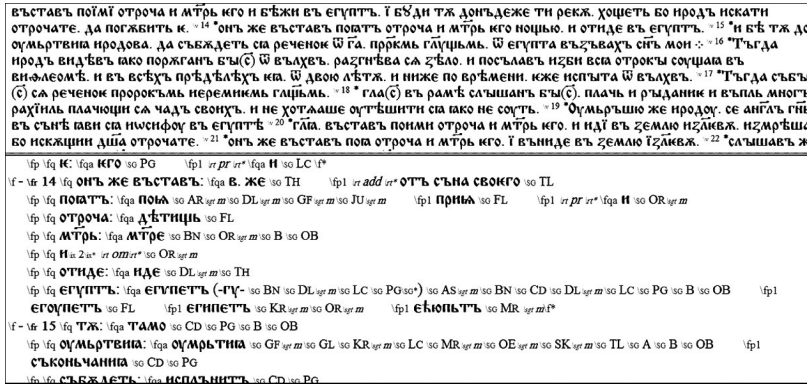

Рис. 1. Пример аннотации критического издания к тексту Мф 2:14 в программе Паратекст

изменяться. Часть тегов отображают смысловую структуру аппарата. Тег ***fq*** воспроизводит в критическом аппарате слова (фразы) из ТЕ, которые являются узлами разночтения (меняются в других группах текстов). Тег ***fqa*** показывает, как представлен этот фрагмент в нормализованной орфографии в других рукописях, вслед за этим полем приводятся сокращенные обозначения представителей групп рукописей, которые обозначаются тегом ***sg***. Таким образом, для узла разночтения *отроча* в Мф 2:14 в ркп. Fl (РНБ, F. п. 1.14, которая представляет периферию Древнего и Преславского текстов наряду с Баницким и Вукановым Евангелиями XIII века) в нормализованной орфографии стоит *дѣтищь*. Для апракосов приводится комплексное обозначение: кроме сиглы рукописи указывается тип чтения при помощи тега ***sgt***, например ***sg *Ar**sgt *m* — перикопа месяцеслова Архангельского Евангелия (в справочной части приводятся границы перикопы 2:13–25, которая читается 26 декабря, и список апракосов, в которых встречается данный текст: Ar As Dl Gf Ju Kr Mr Or SK). В приведенном фрагменте аппарата Мф 2:14 для слова *поıатъ* из ТЕ указаны два варианта замен: *поѣ* (в апракосах ***sg *Ar**sgt *m **sg *Dl**sgt *m **sg *Gf**sgt *m **sg *Ju**sgt *m*) и *приѣ* (в ркп. ***sg *Fl*), кроме того, в апракосе ***sg *Or**sgt *m* к этому слову добавлено слово *и.* Последнее разночтение вводится специальным тегом ***rt****, который оформляет* добавление после указанного фрагмента (*add*), пропуск фрагмента (*om*), добавление впереди указанного фрагмента (*pr*), перестановку частей цитируемого разночтения (*tr*). Тег ***fp*** отмечает очередной узел

разночтения в линейной структуре исходного стиха TE, а тег ***fp1*** разделяет варианты чтений, причем замены приводятся в алфавитном порядке текстовых замен в нормализованной орфографии, после которых приводятся другие типы узлов (*add*, *om*, *pr*, *tr*).

Структурированное представление критического аппарата облегчает проверку и редактирование чтений при помощи просмотра указателя слов и контекстов (рис. 2).

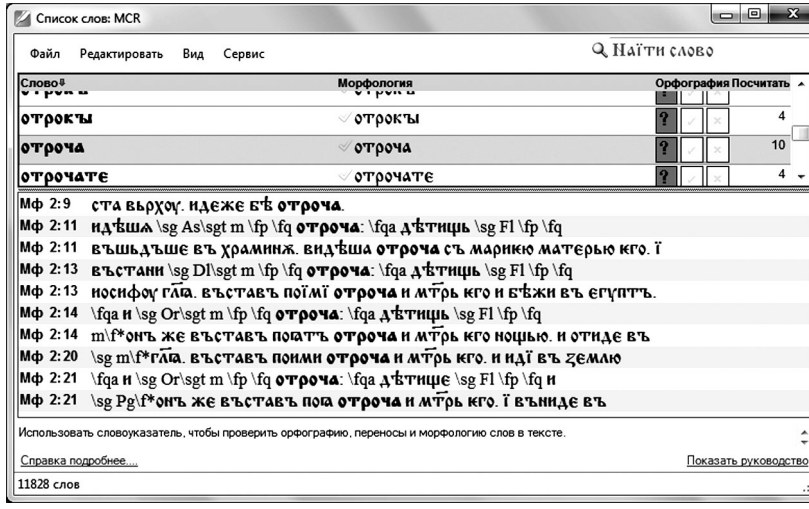

Рис. 2. Окно указателя слов и контекстов в программе Паратекст

В программе Паратекст есть возможность автоматически преобразовать текст в формат настольно-издательской программы InDesign, что потребуется для публикации Четвероевангелия, причем каждый из тегов преобразуется в стиль абзаца или символов InDesign, которые легко корректировать для печати.

Каждый из списков, включенных в критический аппарат издания, будет представлен в корпусе в виде орфографически нормализованного текста, в котором будут обеспечены следующие варианты представления (навигации) по стихам соответствующих глав Евангелия:

- изолированный просмотр текста каждой из 28 рукописей;
- подстрочное представление базового Мариинского Евангелия (для первых пяти глав Евангелия от Матфея — Типографского Евангелия) и каждой из 28 рукописей;

- представление узлов разночтений в виде структуры с гиперссылками, причем будет возможно выбрать типы отображаемых разночтений из заданного списка: пропуск фрагмента текста, вставка фрагмента текста, замена незнаменательного слова, замена знаменательного слова, перестановка фрагментов текста.

Литература

Алексеев А. А., Лихачева О. П. К текстологической истории древнеславянского Апокалипсиса // Материалы и сообщения по фондам Отдела рукописной и редкой книги Библиотеки Академии наук СССР. 1985. Л., 1987. С. 8–22.

Алексеев А. А. Проект текстологического исследования Кирилло-Мефодиевского перевода Евангелия // Советское славяноведение. 1985. № 1. С. 82–95.

Алексеев А. А., Кузнецова Е. Л. ЭВМ и проблемы текстологии древнеславянских текстов // Лингвистические задачи и обработка данных на ЭВМ. М., 1987. С. 111–120.

Афанасьева Е. В. К истории текста и языка древнейшего славянского перевода книги Иова: автореф. дис. … канд. филол. наук. Л., 1988.

Воскресенский Г. А. Евангелие от Марка по основным спискам четырех редакций рукописного славянского евангельского текста с разночтениями из 108 рукописей Евангелия XI–XVI вв. Сергиев Посад, 1894.

Воскресенский Г. А. Характеристические черты четырех редакций славянского перевода Евангелия от Марка по 112 рукописям Евангелия XI–XVI вв. М., 1896.

Гребенников А. О. Исследование устойчивости лексико-статистических характеристик текста: автореф. дис. … канд. филол. наук. СПб., 1998.

Дудинов П. А. Комиссия по научному изданию Славянской Библии при Петроградской духовной академии. Организация, деятельность, результаты // Христианское чтение. СПб., 1996. № 13. С. 36–72.

Евангелие от Иоанна в славянской традиции / под ред. А. А. Алексеева. СПб., 1998.

Евангелие от Матфея в славянской традиции / под ред. А. А. Алексеева. СПб., 2005.

Жуковская Л. П. Текстология и язык древнейших славянских памятников. М., 1976.

Жуковская Л. П. Типология рукописей древнерусского полного апракоса XI–XIV вв. в связи с лингвистическим изучением их // Памятники древнерусской письменности: Язык и текстология. М., 1968. С. 199–332.

Миронова Д. М. Классификация славянских рукописей Евангелия от Матфея // Евангелие от Матфея в славянской традиции / под ред. А. А. Алексеева. СПб., 2005. С. 163–168.

Миронова Д. М. Оценка текстологической значимости узлов разночтений (на материале 525 списков славянского Евангелия от Матфея XI–XVI вв.) // Структурная и прикладная лингвистика. Вып. 9 / под ред. А. С. Герда. СПб., 2012. С. 261–275.

Пичхадзе А. А. Применение статистики при текстологическом исследовании памятника с контролируемой традицией (Древнеславянский Паримейник) // Лингвистические задачи и обработка данных на ЭВМ. М., 1987. С. 121–140.

Рогозина Е. А. Разметка структуры содержания в корпусе агиографических текстов СКАТ // Информационные технологии и письменное наследие [материалы международной научной конференции] / отв. ред. В. А. Баранов. Уфа; Ижевск, 2010. С. 172–175.

Colwell E. C. Studies in Methodology in Textual Criticism of the New Testament. Leiden: Brill, 1969. P. 56–62.

Evangelium sv. Marka. Text rekonstruovaný / ed. by J. Vajs. Praha, 1935a.

Evangelium s. Matthaei palaeoslovenice / ed. by J. Vajs. Pragae, 1935b.

Evangelium s. Ioannis palaeoslovenice / ed. by J. Vajs. Pragae, 1936a.

Evangelium s. Lucae palaeoslovenice / ed. by J. Vajs. Pragae, 1936b.

Garzaniti M. Die slavische Version der Evangelien. Forschungsgeschichte und zeitgenössische Forschung. Köln, Weimar, Wien, 2001. S. 153–163.

Meillet A., Vaillant A. Communication sur ľopportunité de publier des éditions critiques des textes vieux-slaves // Sborník prací I. sjezdu slovanských filologů v Praze 1929. Svazek II. Praha, 1932. S. 594–598.

С. С. Волков

ИСТОРИЯ ТЕРМИНОЛОГИЙ ГУМАНИТАРНЫХ НАУК КАК ЛИНГВИСТИЧЕСКАЯ И УЧЕБНАЯ ДИСЦИПЛИНА

Аннотация. Статья посвящена вопросам формирования терминологии гуманитарных наук (риторика и грамматика) в русском языке. Рассматриваются лингвистический и дидактический аспекты этой проблемы. Особое внимание уделяется процессам сложения языка для специальных целей филологии в XVII–XVIII веках.

Ключевые слова. Термин, филология, язык для специальных целей, русский язык, словесная культура XVII–XVIII веков, тропы.

Sergey S. Volkov

THE HISTORY OF THE TERMINOLOGY OF THE HUMANITIES AS A LINGUISTIC AND ACADEMIC DISCIPLINE

Abstract. In the article the questions of formation of the terminology of the Humanities (philology, rhetoric, grammar) in the Russian language are concerned, the linguistic and didactic aspects of the problem being shown off. Special attention is paid to the process of the philological language for special purposes formation in XVII–XVIII centuries.

Keywords. Terms, rhetoric, grammar, didactics, language for special purposes, Russian language and verbal culture, antonomasia.

I

Одним из императивов современного университетского образования, настойчиво устанавливаемых государственными образовательными стандартами, является изучение студентами языков для специальных целей (ЯСЦ), столь необходимых для качественного профессионального общения, овладения новыми научными знаниями и практическим опытом. Язык для специальных целей (англ.

language for special purposes, LSP; нем. Fachsprache) — это «исторически сложившаяся, относительно устойчивая для данного периода автономная экзистенциальная форма национального языка, обладающая своей системой взаимодействующих социолингвистических норм, представляющая собой совокупность некоторых фонетических, грамматических и, преимущественно, специфических лексических средств общенародного языка, обслуживающая речевое общение определенного социума, характеризующегося единством профессионально-корпоративной деятельности своих индивидов и соответствующей системой специальных понятий» [Коровушкин, с. 12]. Вопросам ЯСЦ посвящена обширная литература (труды А. С. Герда, В. М. Лейчика, С. Д. Шелова, О. В. Фельде, В. П. Коровушкина, А. В. Суперанской, В. Д. Бондалетова, Б. Л. Богородского, также Т. Рёльке, Дж. Фишмана, Т. Хатчинсона, А. Уотерса и многих других), обучение ЯСЦ уверенно нашло свое место в современной образовательной константе «Communication. Culture. Knowledge». Несмотря на это, пока не выработано единого понимания ЯСЦ, идет интенсивная дискуссия о том, каково, так сказать, «оптимальное», непротиворечивое содержание этого понятия и в чем отличие ЯСЦ от разнообразных «подъязыков», «регистров» и пр. (вопрос, заметим, чисто терминологический), но вряд ли вызовет сомнение утверждение о том, что ядро ЯСЦ составляют системы терминов (терминосистемы) — сознательно конструируемые совокупности терминов, удовлетворяющих языковым, логическим и собственно терминологическим требованиям [Лейчик, 2001, с. 55]. Соответственно сказанному, среди задач, которые ставятся при изучении ЯСЦ, весьма актуальна задача личностно-ориентированного освоения учащимися «всеобщего» репертуара лексических знаков каждого ЯСЦ — *системы терминов*. Термином, вслед за А. С. Гердом, будем считать номинативную единицу какого-либо конкретного естественного или искусственного языка, обладающую в результате стихийно сложившейся или особой сознательной коллективной договоренности специальным терминологическим значением [Герд, с. 2–3]. Различия между термином и не-термином исследовал Б. Н. Головин: эти различия, по его мнению, заключаются в том, что термин соотнесен не с отдельным предметом, а с понятием; нуждается в дефинировании; его значение соотносится со значениями других терминов в пределах соответствующей терминологической

системы и с определенной профессиональной деятельностью [Головин, с. 7].

Синхронному существованию *терминосистемы* как относительно «упорядоченной» и нормированной совокупности (системы) терминов» в диахронии предшествует так называемая *терминология* — языковое образование парадигматического типа, представляющее собой *стихийно* сложившуюся совокупность лексических единиц разной степени терминологичности: прототерминов, терминов, предтерминов, терминоидов, номенов, терминонимов и терминоэлементов (см. [Лейчик, 2001, с. 55]). Терминология, таким образом, представляет собой живую, изменяющуюся, формирующуюся, складывающуюся лексическую систему. История терминологий, изучение процессов приобретения «потенциальными» терминами или прототерминами качеств терминологичности, моменты преодоления этими лексическими единицами, как пишет об этом В. М. Лейчик, «порога терминологизации» [Лейчик, 2009, с. 84] актуальны и как объект научного исследования, и как один из аспектов обучения студентов ЯСЦ. Особый интерес в учебном процессе с точки зрения изучения динамики терминологий представляет наблюдение:

- *эволюции* знаков ЯСЦ (их фонетические, графические, семантические трансформации, также изменение инвентаря морфем и терминоэлементов);
- *развития (прогресса)* знаков ЯСЦ (рост их численного состава, усложнение их внутренних системных связей);
- *совершенствования* ЯСЦ (сознательное упорядочение и организация терминологии в процессе ее систематизации и нормализации, редукция дублетов, избыточных и «неугодных» терминологических единиц, создание новых терминов) [Фельде, с. 52–53].

Представляется, что совокупность фактов и материалов, связанных с формированием терминологий, в том числе и терминологии гуманитарных наук (филологии и лингвистики), представляет не только исследовательский интерес, но также значима дидактически, так как позволяет учащимся получить новые знания в профессиональной сфере, узнать новое о глубоких качественных и количественных изменениях в научном знании. Ярче и яснее выступают многие сложные вопросы истории русского литературного язы-

ка, русской национальной культуры, становления национальной нормы. Добавим, что история культуры, как справедливо отмечал Б. Л. Богородский [Богородский], всегда остается неполной без привлечения лингвистических данных.

II

Проиллюстрируем сказанное кратким очерком истории филологического термина, с которым студенты гуманитарных специальностей часто встречаются на практических занятиях по культуре речи и литературному редактированию, в лекционных курсах по лексикологии, стилистике, риторике и пр. Наш очерк, акцентируем, посвящен истории *лексической единицы* (термина), а не истории научного понятия, которая, собственно, является объектом другой дисциплины — истории науки.

Начнем с простых и известных всем заимствованных из греческого или латинского языка наименований тропов речений, таких как, например, *антономазия* (также *антономасия*).

Антономазия (лат. antonomasia или pronominatio, греч. ἀντονομασία, букв. — переименование) — замена имени собственного нарицательным или имени нарицательного именем собственным, один из самых известных тропов. Авторы риторических трактатов единодушны во мнении: не следует достойным риторам злоупотреблять антономазией, но Порфирий Крайский (1707–1768), преподававший риторику в Славяно-греко-латинской академии, об этом тропе отзывается с бóльшей симпатией: «Хорошо украсишь речь с помощью антономазии, если заменишь имя нарицательное именем собственным» [Смирнова, с. 49].

Наиболее интересным является, безусловно, ранний этап терминологизации лингвистической единицы, когда слово, так сказать, борется за место под солнцем в системе языка, встраивается во множество связей и отношений в терминологии филологии. Хронологически этот период совпадает с первым (1620–1695) и вторым (1695–1750) периодами в истории становления русской словесной культуры по В. И. Аннушкину [Аннушкин, 2011, с. 16–17].

Когда же все-таки появился термин *антономазия* в русском языке? Наши разыскания показывают, что сравнительно поздно. В «Словаре русского языка XI–XVII веков» это слово отсутствует,

и только обратившись к электронной коллекции риторических текстов ИЛИ РАН, обнаруживаем одну из наиболее ранних фиксаций слова *антономазия* в письменных источниках — в тексте первой русской риторики начала XVII века — «Риторике» 1620 года. Этот источник особо интересен для изучения терминов на этапе формирования терминосистемы, так как «количество обнаруженных списков (34) и их датировка позволяют предполагать, что „Риторика“ активно переписывалась и изучалась в общественно-речевой и, прежде всего, школьной практике всего XVII столетия, начиная, по меньшей мере, с конца второго десятилетия» [Аннушкин, 1985, с. 141]. Таким образом, система наименований риторических понятий в «Риторике» (1620) активно входила в общественную и коммуникативную практику, приобретая тем самым свойство *конвенциональности* (от лат. conventio — соглашение, договоренность), столь важное для формирования системы терминов. Вот какие интересные материалы предоставляет нам «Риторика» (1620):

> Что есть антономия? — Антономасия то есть егда начертание или ознаменование, или истиннословие, или описание ставим вместо каковаго имения, яко же есть вместо Омира — творца стихов, вместо Аристотеля — философа, вместо Спаса — Христа, вместо гневу — злость отричнутую. Латынским языком антономасию нарицают пермутацио, сииречь пременение [Аннушкин, 1999, с. 69–70].

В этом контексте, возможно, мы видим сам акт рождения, творения термина, главная цель которого — номинировать новое понятие. В качестве результата номинации нового для русской словесной культуры понятия наблюдаем группу «избыточных» терминоподобных единиц, как заимствованных, так и исконных: *антономия*, *антономасия*, *пермутацио*, *пременение* (или явление «разноименства» — так называл это Ю. С. Сорокин). В этом случае, в отличие от упорядоченных, «готовых» *терминосистем*, как указывает О. В. Фельде, отражающих определенные теории области знаний или производственной деятельности, *терминологии* могут характеризоваться избыточностью наименований. Особенно ярко, по мнению этого исследователя, вариативность и, добавим, сосуществование и конкуренция терминологических единиц представлены в ЯСЦ XVIII в. [Фельде, с. 51]. Так и русскоязычные риторики XVII–XVIII вв. характеризуются избыточностью терминологических

обозначений — стилистической «пестротой», смешением старых и новых слов, церковнославянизмов и лексических калек [Богданов, с. 77], что обусловлено как способностями русского языка этого периода к номинации новых научных понятий, так и, как мы покажем ниже, особенностями решения переводчиками своих задач.

В. И. Аннушкин в своем исследовании, посвященном первой русской риторике, доказал, что ее текст представляет собой анонимный перевод западноевропейского риторического трактата — латинской «Риторики» Филиппа Меланхтона[1] в сокращенной переработке Луки Лоссия (Лотце) из Люнебурга (см. [Аннушкин, 1999; Lossius]. Для нас важно, что этот неизвестный переводчик знал (или, возможно, делал вид, что знал, — укажем, например, на такое необычное переводческое решение, как *трантланцо*, вместо, как мы предполагаем, лат. translatio — 'перенос, метафора') греческий, латинский и польский языки (см. [Вомперский, с. 14]) и владел терминологическим аппаратом риторики. Всеми силами он старался решить новую культурную задачу: адаптировать новый ЯСЦ риторики и «культурные экзотизмы», в нем содержащиеся, к средствам русского языка того периода. Для достижения лучшего понимания текста и, возможно, преследуя потенциальные учебные задачи, он или прибегал к заимствованиям (возможно, даже уже известным книжным людям Московского государства той поры), или создавал русские «пуристические» переводческие эквиваленты (преимущественно с прозрачной внутренней формой).

Каков лингвистический статус приведенных выше единиц? Какова их дальнейшая судьба? Остаются ли они далее релевантными для терминологии гуманитарных наук? Полноценные ли это термины, терминологические ли дублеты, аналоги или только окказиональные переводческие эквиваленты (см. [Рецкер, с. 10–18])? Этот принципиальный вопрос может быть решен только путем анализа фактов, полученных в результате применения метода *исторической ретроспекции*, а именно: исследования традиции, широкого привлечения языковых материалов периода (т. е. XVII–XVIII вв.), анализа данных исторических словарей и картотек, фактов других языков (европейских и, что очень важно, родственных славянских) и, самое

[1] Филипп Меланхтон (1497–1560) — немецкий гуманист, теолог и филолог, сподвижник М. Лютера.

главное, внимательных наблюдений за формой и семантикой. Должна получить рассмотрение дихотомия «общеязыковое — специальное (терминологическое)». Эти важные аспекты научного изучения терминологии, к сожалению, абсолютно игнорируются в некоторых «публикациях», посвященных истории риторических терминов (см., например, [Лемешев].

1. Лексическая единица *антономия* не отмечается ни одним из доступных нам исторических словарей XVII–XVIII вв., контексты его употребления отсутствуют в «Картотеке „Словаря русского языка XVIII века“» и Национальном корпусе русского языка, что позволяет считать ее переводческим окказионализмом, своеобразным гапаксом XVII в. Вполне возможно, что это просто ошибка писца или переписчика, своего рода lapsus calami: в термине *антономасия* выпущены две буквы (-ас-), и вместо «Что есть *антономия*?» читать следует «Что есть антономасия?»[2], тем более что В. И. Аннушкин полагает, что протограф «Риторики» (1620) представляет собой записи писца или писцов под диктовку неизвестного переводчика [Аннушкин, 2011, с. 30]. У Меланхтона—Лоссия: «Quid est antonomasia?» [Lossius, p. 181].

2. *Пермутацио* — лат. ‘перемена, изменение, обмен’ — представляет собой прямое единичное, то есть характерное только для перевода данного источника, переводческое заимствование из текста «Elementorum Rhetorices…» Меланхтона—Лоссия, ср.: «Quid est Antonomasia? — Est cum definitionem, aut Etymologiam, aut descrptionem pro aliquo nomine ponimus, ut pro Homero Poëtā, pro Aristotele Philosophū, pro Christo Seruatorem, pro ira, bilem efferuescentem. (Latine nomines permutatio)» [Lossius, p. 181]. Лексическая единица *пермутацио* (не регистрируется «Словарем русского языка XI–XVII вв.» и не входит в словник «Словаря русского языка XVIII века», который, следует заметить, довольно последовательно отражает риторические новации XVIII в.

3. Совершенно иная ситуация с аналогом *пременение* (премѣнение, прѣмѣнение, прѣмѣненье и др.). Это «старое» слово хорошо известно русскому языку XVII в., равно как и языку более ранних периодов (одно из первых употреблений — в «Изборнике великого князя Святослава Ярославича 1073 г.), оно фиксируется

[2] Благодарим М. Г. Маматову, любезно подсказавшую нам эту мысль.

историческими словарями в значениях ‘перемена, изменение’ [Срезневский, с. 1671], ‘смена’, ‘замена’, ‘превращение’ и пр. По данным «Словаря русского языка XI–XVII вв.», слово *пременение* регистрируется в значении ‘гипербола’ уже в переводе «Епископа Брунона Вирцбургского толкования на Псалтырь», выполненном Дмитрием Герасимовым (1535)[3], одним из «культурных героев» XVI столетия. Перевод сделан с оригинала на латинском языке:

Птицы … премѣнениемъ чл҃комъ уподобляются (aves … per iperbolen [hyperbolen] hominibus comparantur) [Словарь русского языка … , с. 273].

Предположим, что перед нами один из первых шагов на пути вхождения слова *пременение* в круг системы обозначений языка словесных наук. Наше предположение подтверждается фактами нарастающей активности реализации системы употреблений этой лексической единицы в филологическом (риторическом) дискурсе. Слово робко вступает в круг риторических наименований: в качестве семантического неологизма, то есть получив новое, «риторическое» содержание, оно появляется в текстах русских риторик XVII–XVIII веков, тем самым начинается новая ветвь его развития. Так, например, в «Риторике» (1620) это не только дублет слова *антономазия*, но также и наименование риторической фигуры — *антиметаболы*:

Пременение есть егда сопротивными предлагаетъся, яко же: недругом милостивым, другом же немилостивым уставися; труд есть самое прохладство; потребно, чтоб еси жил, не жити, чтоб еси ел; грекове нарицают антиметаволи и метафесис [Аннушкин, 1999, с. 86].

В «Риторике» Михаила Ивановича Усачева (1699) в объяснении, что представляет собой аллегория и чем аллегория отличается от метафоры, обнаруживаем следующее:

Аллигория, есть *пременение*, егда иное глаголется, иное разумеется, или есть слово имже иное реченми, а иное разумом познавается, обаче неразное, на приклад: В поучении моем разгореся огнь, то есть любве Божия (цит. по [Матвеев, с. 41]).

Аллигория от метафоры разделяется, яко метафора есть *пременение* или пренесение во едином речении, аллигория же во многих, и да явнее

[3] В книге Л. В. Балашовой «Русская метафора: прошлое, настоящее, будущее» (М., 2014) этот же текст служит источником для демонстрации метафорических переносов средневековья.

реку: яко аллигория из метафор составляется из двух, или трех, или многих [Там же].

Это значит, что и метафора, и аллегория есть разновидности *пременения*, то есть изменения значения слова.

Вызывает интерес попытка М. И. Усачева использовать вслед за «Риторикой» (1620) слово *пременение* для наименования риторических фигур. *Пременение* указывается Усачевым в качестве русского дублета к греческому термину πολύπτωτον (русск. полиптотон или полиптот, буквально — многопадежие, ‘многократное употребление слова в разных грамматических формах в пределах одного контекста’)[4], притом слово *пременение* находится в контекстуальной связи со словом *предложение*:

> Полиптотон, предложение, или *пременение* (выделено нами. — *С. В.)*, есть, егда во отменном падеже едино речение многажды полагается, например: сего ты человеком именуеши аще бы был он человек, человеческий а не скотский живот препровождал бы. Или богатый всуе богатым наричется, аще желает богатство, зане богатство требуемых доволство требуемых именуется [Усачев, с. 128 об. — 129).

Возможно, потенциал номинации, возможность выбора точного наименования в складывающемся в то время языке для специальных целей гуманитарных наук были еще весьма скромными. Число вовлеченных в эту сферу, «приспособленных» к задачам филологии языковых единиц было невелико, опции выбора наиболее подходящего (то есть точного и удобного) средства для номинации были ограниченны. Возможно, такое употребление объясняется влиянием какой-то еще, неизвестной нам традиции или какого-то неизвестного, но авторитетного образца. Возможно, выбор Усачева основывался на актуализации уже ставшего общепонятным и общепринятым семантического элемента ‘*изменение*’ в значении слова, ведь *полиптотон*, как указывает Никола Коссен,

> назвали так потому, что, когда она [фигура речи] основывается на одной и той же части речи — чаще всего в начале — эта часть речи изменяется по падежам, родам или числам. По падежам — таким образом: «Сенат — совет с высшей властью; сенату поручается забота о государстве; на сенат обра-

[4] Благодарим А. К. Филиппова и А. С. Смирнову за предоставленную возможность использовать их рабочие материалы к словарной статье «Наклонение» в готовящемся к печати словаре-справочнике «Риторика М. В. Ломоносова».

щаются взоры всего города в сомнительных и опасных обстоятельствах»… А по числам — так: «Приятными всегда были частным лицам постановления, принятые ради их выгоды. Приятное сделал сенаторам старший Африканец, освободив скамьи этого сословия от простого народа» [Caussinus, с. 406.2–407.1].

Поэтому изменение формы слова (по падежам, родам и т. д.), являющееся характерным признаком этой фигуры, превратилось в яркий и понятный мотивирующий признак, позволивший осуществить номинацию с помощью именно этого слова, ср. у Усачева:

Разумеи же полиптотон быти пременением речения не токмо во отменных [падежах], но и во отменных глагола временем[5], наприклад: преступил еси закон гражданский, преступаеши церковный, не уступишся преступити естественнаго, и не токмо во отменных временех, но и во отменных родех и числех пременение бывает, наприклад: высокая луна, вышшее солнце, еще суть вышшия круги, но высочайшее емпирийское [Усачев, с. 128 об. — 129].

В заключение заметим, что, по данным «Словаря русского языка XI–XVII вв.», «Материалов» И. И. Срезневского, «Словаря русского языка XVIII века», «Лексикона треязычного» Ф. П. Поликарпова, слово *полиптотон* представляет собой новацию XVII века, поэтому использование вместе с ним мотивированного русского аналога вполне обосновано.

4. Исследования источников риторического трактата М. В. Ломоносова показали, что он был знаком не только с первой «Риторикой» (1620), но и с «Риторикой» Михаила Усачева (1699), к тексту которой, как доказал В. И. Аннушкин, восходят основные определения риторики и способов приобретения красноречия, которые находим в риторическом трактате М. В. Ломоносова [Аннушкин, 1999, с. 32–43]. Поэтому неудивительно, что слово *пременение* появляется в филологических трудах М. В. Ломоносова[6]. В «Кратком руководстве

[5] Так в рукописи.

[6] Нельзя не обратить внимание на то, что в «Кратком руководстве к риторике» (1744), оставшемся в рукописи, М. В. Ломоносов определяет антономазию следующим образом: «**Переменение** (выделено нами. — *С. В.)* имен собственных и нарицательных служит ко изобретению слов риторических: 1) Когда употреблено будет имя собственное вместо нарицательного… 2) Нарицательное вместо собственного… Показанная **перемена** (выделено нами. — *С. В.)* имен называется антономазия» [Ломоносов, с. 53].

к красноречию» (1747) *пременение* становится наименованием одной из фигур предложений — антиметаболы:

Лучшие фигуры предложений суть следующие двадцать шесть: определение, изречение, вопрошение, ответствование, обращение, указание, заимословие, умедление, сообщение, поправление, расположение, уступление, вольность, прохождение, умолчание, сомнение, заятие, напряжение, **пременение** (выделено нами. — *С. В.)*, присовокупление, желание, моление, восхищение, изображение, возвышение, восклицание [Ломоносов, с. 261–262].

Ср. также:

Пременение есть когда преложением речений противные идеи производятся. *Пример из 2 Димосфенова олинфического слова: Весьма погрешаете вы, афинеане, что чрез тое надеетесь произвести счастие из несчастия, чрез что из счастия несчастие сделалось.* Сюда принадлежит и следующее: *Ежели ты что хорошее сделаешь с трудом, труд минется, а хорошее останется, а ежели сделаешь что худое с услаждением, услаждение минется, а худое останется.* Также и сие: *Не для того мы живем на свете, чтобы насыщаться, но для того насыщаемся, чтобы жить* [Ломоносов, с. 280].

Авторитет М. В. Ломоносова как выдающегося поэта, филолога, просветителя, высокий, как теперь говорят, импакт-фактор «Краткого руководства к красноречию» и использование этой книги в качестве учебного пособия, большое количество переизданий (например, тираж издания Московского университета 1759 г. составил 1200 экземпляров, издания 1765 г. — 1200 экземпляров, издания 1776 г. — 1000 экземпляров и т. д.) [Сводный каталог… , с. 168–169] стали причиной завершения процесса терминологизации лексической единицы *пременение*. Выбор Ломоносовым именно такого наименования для одной из фигур вполне объясним:

1) это традиционно-книжное слово уже, так сказать, давно «действовало» в языке риторики. Заметим здесь, что в течение длительного времени старательно акцентировалась «демократическая» составляющая в языке и стиле Ломоносова, хотя на самом деле, как показывает наш материал, богатство и изысканность его индивидуального слога основываются на книжной и церковно-книжной лексике;

2) это, еще раз подчеркнем, исконное, незаимствованное слово обладало прозрачной, понятной внутренней формой, что соответствовало нормативным императивам, установленным немецкой просветительской школой Лейбница—Вольфа: термин должен быть

создан на основе национального языка и должен быть понятен его носителю: «Хороший термин должен содержать правильное исчерпывающее понятие о вещи, должен быть понятным и при этом не оскорблять ни языковую практику, ни вкус и благозвучие», — писал выдающийся немецкий нормализатор, грамматист и лексикограф И. К. Аделунг (цит. по [Филиппов, с. 56]).

Со второй половины XVIII в., во многом благодаря филологическим трудам Ломоносова, риторический термин *пременение* прочно закрепляется в языке риторики в значении ‘фигура предложения, антиметабола’. Употребление в таком значении находим, например, в «Кратком руководстве к оратории российской» преосвященного Амвросия (Серебренникова):

> Пременение… есть, когда преложением речений различной, или противной смысл раждается. На пр. Не для того живемъ чтобъ учиться; но для того учимся, чтобы приятнее жить. Еразм Ротер[дамский] Или: война дела домашния, домашния дела́ войну отягощали [Амвросий, с. 229].

История слова *антономазия* и его «временных» аналогов весьма интересна не только в исследовательском плане, но и дидактически. Период XVII–XVIII веков — ранний период в истории формирования терминологии гуманитарных наук — обладает большим образовательным потенциалом. Учащиеся могут получить информацию не только о деятельности таких просветителей, как Дмитрий Герасимов, преосвященный Амвросий Серебренников, М. В. Ломоносов и многие другие, но и о поступательном развитии русской словесной культуры. Острые культурно-языковые конфликты и потрясения этого периода, состояние «взбудораженности», по меткому выражению Ю. С. Сорокина, всей системы русского языка позволят привлечь их внимание к процессам формирования корпуса гуманитарной терминологии, научить ценить как узуальные, так и маргинальные явления. Приобретенный опыт, возможно, будет способствовать преодолению *антиисторизма* и *механицизма* (представление терминосистем как идеально действующих механизмов) при оценке прошлых языковых состояний.

Литература

Амвросий [*Серебренников*]. Краткое руководство к оратории российской, сочиненное в Лаврской семинарии в пользу юношества, красноречию обучающегося. М., 1778.

Аннушкин В. И. История русской риторики. Хрестоматия. М., 2011.

Аннушкин В. И. Первая русская «Риторика» XVII века: Текст. Перевод. Исследование. М., 1999.

Аннушкин В. И. Первая русская «Риторика» начала XVII века: автореф. дис. … канд. филол. наук. М., 1985.

Богданов К. А. О крокодилах в России: Очерки из истории заимствований и экзотизмов. М., 2006.

Богородский Б. Л. Об одном термине из «Слова о полку Игореве» (насадъ — носадъ) // Ученые записки Ленингр. гос. пед. ин-та им. А. И. Герцена. Кафедра рус. яз., 1955. Т. 104. С. 227–258.

Вомперский В. П. Риторики в России XVII–XVIII вв. М., 1988.

Герд А. С. Проблемы формирования научной терминологии: автореф. дис. … д-ра филол. наук. Л., 1968.

Головин Б. Н. Термин и слово // Термин и слово. Горький, 1980. С. 3–12.

Коровушкин В. П. Контрастивная социодиалектология как автономная лингвистическая дисциплина // Язык в современных общественных структурах: Материалы междунар. науч. конф. 21–22 апреля 2005 г. в Нижнем Новгороде. Нижний Новгород, 2005. С. 12.

Лейчик В. М. Проблема системности в отечественном терминоведении // Научно-техническая терминология. М., 2001. Вып. 2. С. 54–55.

Лейчик В. М. Терминоведение: предмет, методы, структура. Изд. 4-е. М., 2009.

Лемешев К. Н. Наименования фигур предложения в Риторике М. В. Ломоносова // Acta linguistica petropolitana. Труды Института лингвистических исследований. СПб., 2014. Т. X. Часть 1. С. 629–663.

Ломоносов М. В. Полное собрание сочинений: в 11 т. М.; Л., 1950–1983. Т. 7: Труды по филологии 1739–1758 гг. М.; Л., 1952.

Матвеев Е. М. Аллегория // Риторика М. В. Ломоносова: проект словаря / науч. ред. П. Е. Бухаркин, С. С. Волков, Е. М. Матвеев. СПб., 2013.

Рецкер Я. И. Теория перевода и переводческая практика. М., 2009.

Сводный каталог Русской книги гражданской печати XVIII века (1725–1800). Т. 2. М., 1964.

Словарь русского языка XI–XVII вв. М., 1992. Вып. 18.

Смирнова А. С. Антономазия // Риторика М. В. Ломоносова: проект словаря / науч. ред. П. Е. Бухаркин, С. С. Волков, Е. М. Матвеев. СПб., 2013.

Срезневский И. И. Материалы для словаря древнерусского языка по письменным памятникам. Т. 2. Л–П. СПб., 1902.

Усачев М. И. Риторика // ГИМ. Собрание Щукина. № 803.

Фельде О. В. Языки для специальных целей в историко-лингвистическом аспекте // Вестник Бурятского государственного университета. 2013. № 10.

Филиппов К. А. Наблюдения над метаязыком концептуальных грамматических текстов эпохи немецкого Просвещения // Материалы метаязыкового семинара ИЛИ РАН. 2014 г. / отв. ред. С. С. Волков, Е. М. Матвеев. СПб., 2015.

Caussinus N. De Eloquentia Sacra et Humana Libri XVI. Lugduni, 1637 / пер. А. А. Ветушко-Калевича.

Lossius Lucas. Erotemata Dialecticae et Rhetoricae Philippi Melanthonis… breuiter selecta et contracta per Lucam Lossium Luneburgensem ediscendi gratia. Francoforti, 1556.

А. О. Гребенников

ИНДИВИДУАЛЬНО-АВТОРСКИЙ ХАРАКТЕР РАЗЛИЧНЫХ ЗОН РАСПРЕДЕЛЕНИЯ В ЧАСТОТНЫХ СЛОВАРЯХ ЯЗЫКА ПИСАТЕЛЯ

Аннотация. В статье рассматривается сопоставление различных зон частотных словарей рассказов русских писателей (А. П. Чехова, Л. Н. Андреева, А. И. Куприна, И. А. Бунина), создаваемых на кафедре математической лингвистики филологического факультета СПбГУ. Подобный анализ позволяет выявить лексемы — ключевые слова, являющиеся одновременно и индивидуальными для каждого автора, и выражающими основные темы его творчества. Сопоставление с частотным словарем языка в целом подтверждает индивидуально-авторский характер выделенных единиц. Проводится попытка проанализировать реальное «наполнение» подобных ключевых слов низкочастотными единицами словаря.

Ключевые слова. Частотный словарь, язык писателя, лексикография, стилеметрия.

Alexander O. Grebennikov

AUTHOR'S INDIVIDUALITY IN DIFFERENT ZONES OF FREQUENCY DISTRIBUTION IN AUTHOR'S LEXICON DICTIONARY

Abstract. The article deals with frequency dictionaries of A. Chehov, L. Andreyev, A. Kuprin and I. Bunin published by the author. The high-frequency words from the dictionaries in question are compared thus allowing to identify lexical items both individual for every author and corresponding to the main themes of their works (keywords). Correlation of lexemes from low-frequency part of the distribution to those keywords is also analyzed.

Keywords. Frequency dictionary, lexicography, lexicon, stylometry.

Составление частотных словарей, в том числе частотных словарей языка писателя, является одной из актуальных лингвистических

задач. При их составлении собственно лингвистические интересы могут пересекаться с интересами литературоведения и «авторской» стилистики [Гребенников, 2006; Гребенников, 1998; Корпус древнерусских агиографических текстов …]. Одновременно значительный интерес представляет анализ получаемых частотных распределений, как по формальным (частота, доля, ранг и т. п.), так и по содержательным (выявление индивидуально-авторских характеристик) параметрам.

В качестве иллюстрации возможностей, предоставляемых частотными словарями языка писателя исследователям, занимающимся вопросами авторской стилистики, нами был произведен анализ слов знаменательных частей речи из верхних зон частотных словарей рассказов выдающихся русских писателей, создаваемых по единым принципам на кафедре математической лингвистики СПбГУ в рамках изучения проблемы «Эволюция лексического состава русской художественной прозы XX века» [Частотный словарь … , 1999; Частотный словарь … , 2003; Частотный словарь … , 2006; Частотный словарь … , 2012]. Данные словарные материалы обеспечивают хорошую базу для сравнения, так как анализируемые словари имеют сопоставимый объем; тексты, лежащие в их основе, принадлежат к одному жанру; принципы отбора материала и составления словарей также едины.

Можно предположить, что авторская склонность к употреблению той или иной лексической единицы неизбежно должна проявить себя в более высокой частоте этой единицы в тексте по сравнению с ее частотой в национальном языке и языке других писателей [Карпова, Кириллов].

Табл. 1 позволяет сопоставить данные, полученные для верхних зон частного распределения анализируемых словарей[1]. Все собственно индивидуально-авторские слова (то есть не повторяющиеся ни у одной пары анализируемых авторов; слов, не повторяющихся у различных пар, рассматриваемых по отдельности, гораздо больше) выделены в таблице полужирным курсивом.

Для подтверждения индивидуально-авторского характера выделенных единиц нами было проведено сравнение с данными частотного словаря языка в целом — частотного словаря русского языка С. А. Шарова [Ляшевская, Шаров]. Очевидно, что в данном случае, вследствие огромной разницы исходных выборок, собственно

[1] Здесь и далее приводится только часть полученных данных.

Таблица 1. **Сопоставление данных для верхних зон частного распределения словарей**

Словарь Л. Н. Андреева			Словарь А. И. Куприна			Словарь А. П. Чехова			Словарь И. А. Бунина		
Ранг	Лексема	Частота	Ранг	Лексема	Частота	Ранг	Лексема	Частота	Ранг	Лексема	Частота
6	быть	2779	8	быть	2965	9	быть	2411	16	сказать	380
27	рука	785	32	один	1010	23	говорить	916	21	только	314
31	один	720	35	рука	920	26	сказать	825	22	глаз	312
32	глаз	680	37	только	875	32	один	698	28	один	290
33	сказать	671	38	сказать	874	38	только	590	31	говорить	278
37	только	597	41	глаз	824	39	такой	564	36	рука	273
38	говорить	563	42	говорить	786	43	мочь	506	38	стать	247
41	мочь	539	47	такой	690	44	знать	489	43	день	208
42	такой	521	48	лицо	660	47	глаз	462	45	лицо	201
43	лицо	520	49	голова	647	48	человек	439	46	опять	198
45	другой	470	52	мочь	635	50	рука	430	49	потом	190
48	знать	438	53	большой	629	51	лицо	429	50	идти	182
49	голова	434	55	время	602	53	теперь	408	51	голова	181
51	отец	430	58	знать	594	54	два	404	52	знать	181
53	люди	423	60	другой	537	56	стать	384	53	такой	176
54	человек	408	66	человек	483	59	глядеть	369	54	стоять	173
56	стать	397	67	нога	479	60	думать	365	57	ночь	165
58	жизнь	383	68	два	476	62	большой	343	58	мочь	165
61	смотреть	348	69	вдруг	461	63	потом	341	59	сидеть	165
62	видеть	341	72	друг	435	64	жизнь	331	60	черный	164

66	теперь	317	75	теперь	424	66	идти	329	61	дом	162
67	слово	315	77	видеть	418	68	день	325	62	теперь	159
70	думать	311	79	слово	417	69	голова	320	63	очень	149
71	нога	309	80	идти	409	70	сидеть	315	64	белый	148
72	хотеть	298	81	день	404	72	видеть	310	67	вдруг	140
73	голос	291	82	голос	402	73	другой	310	69	большой	139
74	идти	287	83	очень	397	75	жить	307	70	быть	139
77	день	279	85	***пот***	391	77	дом	300	73	пойти	137
78	казаться	275	86	каждый	373	80	казаться	287	74	самый	136
80	очень	268	88	казаться	347	82	раз	284	75	думать	133
82	раз	264	89	люди	344	83	пойти	279	76	жизнь	132
83	самый	262	90	жизнь	340	85	спать	271	77	окно	132
85	потом	256	94	стать	327	86	дело	268	78	вечер	129
90	земля	239	97	белый	313	87	время	267	80	совсем	126
94	женщина	231	98	самый	312	88	очень	267	81	нога	122
95	дом	226	99	первый	309	89	***жена***	265	82	видеть	121
96	черный	226	100	***хороший***	308	93	люди	246	86	***ветер***	120
97	два	225	101	черный	307	95	бог	245	87	жить	120
101	ночь	213	103	ночь	304	96	***взять***	243	89	свет	116
103	стоять	208	106	всегда	292	99	***доктор***	235	90	глядеть	114
105	любить	206	108	сидеть	287	100	час	234	92	другой	112
106	маленький	205	109	раз	285	101	любить	226	94	два	108
107	***смерть***	204	110	дело	283	102	хотеть	226	96	друг	106
108	совсем	204	111	хотеть	281	103	***понимать***	225	97	дверь	105
109	ответить	203	112	много	276	104	***нужный***	218	98	***небо***	104

Окончание табл. 1

Словарь Л. Н. Андреева			Словарь А. И. Куприна			Словарь А. П. Чехова			Словарь И. А. Бунина		
110	страшный	202	113	думать	275	105	голос	217	100	время	102
111	время	199	114	маленький	275	106	***муж***	214	101	казаться	102
113	каждый	198	116	совсем	271	107	***есть***	211	102	бог	101
114	спросить	197	118	женщина	268	109	спросить	210	103	стол	101
116	вдруг	194	121	опять	264	110	выйти	209	106	ответить	100
117	***тихо***	192	122	час	261	111	***ходить***	205	107	***сад***	98
118	тело	191	123	спросить	260	112	***делать***	204	108	***волос***	97
121	пойти	189	124	место	259	113	***должный***	204	109	длинный	97
122	новый	186	125	***сторона***	253	114	ночь	203	110	***год***	96
123	плакать	184	127	глядеть	249	115	каждый	202	111	***старый***	96
125	***молчать***	182	128	минута	234	117	слово	199	112	***море***	95
126	быстро	178	130	длинный	241	118	опять	198	113	***поле***	95
127	***мысль***	177	134	дом	235	120	вдруг	197	114	темный	95
128	много	176	137	земля	228	122	молодой	195	117	***лес***	94
129	***стена***	176	138	***сейчас***	228	129	комната	188	117	молодой	94
130	темный	175	139	бог	227	130	минута	188	119	***солнце***	93
132	***палец***	171	140	***любовь***	225	131	***дать***	186	120	***красный***	93
134	место	169	141	душа	224	132	маленький	183	121	***дорога***	91
135	белый	168	142	стол	223	134	окно	181	122	смотреть	91
137	***новый***	186	143	страшный	223	135	стол	181	123	комната	90
138	***плакать***	184	144	быстро	222	136	***иметь***	178	124	человек	90
139	молчать	182	145	тело	222	137	душа	173	125	всегда	88
140	друг	162	146	господин	220	138	стоять	170	126	выйти	88

абсолютная частота является малоинформативной, соответственно основным анализируемым показателем становятся ранги (табл. 2–5).

Таблица 2. **Индивидуально-авторские слова Л. Н. Андреева**

Лексема	Словарь Л. А. Андреева		Словарь С. А. Шарова	
	Ранг	Частота	Ранг	Частота
молчать	139	182	813	145
мысль	127	177	299	332
новый	122	186	73	1218
палец	132	171	495	219
плакать	123	184	1165	104
понять	140	162	152	588
смерть	107	204	363	284
смеяться	149	154	1016	118
стена	128	175	400	261

Таблица 3. **Индивидуально-авторские слова А. И. Куприна**

Лексема	Словарь А. И. Куприна		Словарь С. А. Шарова	
	Ранг	Частота	Ранг	Частота
любовь	140	225	307	324
пот	85	391	3609	30
прекрасный	146	220	790	148
сейчас	137	228	102	897
сердце	147	220	432	245
сторона	125	253	121	768
хороший	100	308	199	471

Легко заметить, что единицы, отмеченные нами в качестве индивидуально-авторских, имеют в словаре языка автора, при иногда довольно схожей частоте, в целом гораздо более низкий (иногда более чем в четыре раза) ранг, что подтверждает их отнесенность к группе лексем, характеризирующих авторский стиль.

Обращает на себя внимание также соотнесение выделенных лексем с ключевыми темами творчества писателей, единодушно отмечаемыми критиками и литературоведами существенными особенностями авторской картины мира, такими как, например,

Таблица 4. **Индивидуально-авторские слова А.П.Чехова**

Лексема	Словарь А. П. Чехова		Словарь С. А. Шарова	
	Ранг	Частота	Ранг	Частота
доктор	99	235	829	143
жена	89	265	255	377
иметь	136	178	99	907
лето	140	169	889	135
муж	106	214	397	263
нужный	104	218	279	352
понимать	103	225	160	560
слышать	144	164	411	256
ходить	111	205	342	297
черт	148	156	1099	110

Таблица 5. **Индивидуально-авторские слова И. А. Бунина**

Лексема	Словарь И. А. Бунина		Словарь С. А. Шарова	
	Ранг	Частота	Ранг	Частота
ветер	86	120	847	140
вода	132	84	191	485
волос	108	97	842	142
высокий	134	81	193	484
год	110	96	28	3728
долго	129	85	417	253
дорога	121	91	300	330
красный	119	93	442	241
лежать	135	81	312	318
лес	116	94	512	212

трагическое мироощущение у Л. Н. Андреева, картины природы у И. А. Бунина, тема любви у А. И. Куприна, внимание к деталям у А. П. Чехова. Скажем, гиперболизм в изображении зла, а также темы одиночества и социального неблагополучия, отмечаемые критикой как особенность художественного видения Андреева, несомненно, проявляются в таких лексемах, как *плакать*, *страшный*, *темный*, *черный*.

Стоит заметить, что зачастую в лингвистических исследованиях низкочастотным фактам не уделяется достаточного внимания, при

этом именно они выявляют не общеязыковое, а авторское и специфическое [Герд]. Представляется логичным проследить, какое «наполнение» получают выделенные лексемы — ключевые слова в низкочастотной зоне распределения, куда, кроме слов с частотой собственно 1, традиционно относят единицы с частой менее 10.

Для примера проведем анализ низкочастотной лексики в словаре Л. Н. Андреева. Среди слов с частотами от 1 до 10 мы отметим слова, которые с наибольшей ясностью можно было бы считать входящими в «предметно-смысловой комплекс» некоторых выделенных выше индивидуально-авторских лексем. Результаты сопоставления представлены в табл. 6.

Таблица 6. **Сопоставление индивидуально-авторских лексем из высокочастотной зоны распределения в словаре Л. Н. Андреева с единицами из низкочастотной зоны**

Высокочастотные лексемы	Лексемы с частотой 1	Лексемы с частотой от 2 до 10
смерть — 204	вдовий вдовый выжить гибельный гробовой душегубец живодерня загубить задушенный зарезанный кончина кровопролитие мертвая мертвенно мертвенность мертвеющий мертво неживой омертвелый омертветь отживший отжить повеситься погибающий (сущ. и прил.)	погибший — 10 похороны — 10 смертный — 10 умерший — 7 вдова — 6 помереть — 6 похоронить — 6 убитый — 6 кладбище — 5 помирать — 5 предсмертный — 5 расстрелять — 5 удушить — 5 безжизненный — 5 мертвенный — 4 бойня — 4 покойный — 4 самоубийство — 4 могильный — 3 насмерть — 3 резня — 3 смертельно — 3 убиваться — 3 умертвить — 3 умерший (сущ.) — 3

Продолжение табл. 6

Высокочастотные лексемы	Лексемы с частотой 1	Лексемы с частотой от 2 до 10
смерть — 204	погибший (сущ.) подыхать покойница полусмерть помертвевший помертвелый посмертный траур трупный убиваемый убивающий убиенный умирание	покойный (сущ.) — 3 умирающий — 3 вымерший — 2 гибель — 2 гибнуть — 2 мертвецкая — 2 омертвевший — 2 отпевание — 2 панихида — 2 погибать — 2 погребальный — 2 подохнуть — 2 поминать — 2 похоронный — 2 саван — 2 сдохнуть — 2 скелет — 2 скончаться — 2 смертоносный — 2 траурный — 2 умиравший — 2 умирающая — 2
плакать — 184	всхлипывающий исплакать истерика наплакаться плачущий (сущ.) плакса прослезиться скорбеть	плачущий — 9 скорбь — 9 несчастье — 7 слезинка — 5 рыдающий — 4 всхлипнуть — 3 всхлипывать — 3 всхлипывание — 2 заплаканный — 2 зарыдать — 2 плакаться — 2
страшный — 202	грозиться грозящий жуткость запуганный запугать злобно-испуганный перепугаться пугливость спугнуть	угроза — 9 испугать — 8 страшиться — 7 кошмар — 6 угрожать — 6 ужасающий — 6 угрожающе -5 ужасаться — 5 панический — 4

Окончание табл. 6

Высокочастотные лексемы	Лексемы с частотой 1	Лексемы с частотой от 2 до 10
страшный — 202	страшный-престрашный ужасающе ужаснуть устрашать	ужаснуться — 4 угрожающий — 3 страшное — 3 кошмарный — 2 ужаснувшийся — 2 жуть — 2 страшить — 2
темный — 175	беспросветный затемненный потемки померкнуть потемнение темень сумрачно	тускло — 10 погасить — 7 сумрачный — 6 сумрак — 5 полутьма — 4 потемнеть — 4 потухший — 4 потускнеть — 4 непроницаемый — 3 затемнить — 2 полутемный — 2 стемнеть — 2

Примечание. В первом и третьем столбцах таблицы для лексемы указывается частота в частотном словаре языка автора.

Мы видим, что для каждого из высокочастотных семантических инвариантов можно легко найти более десятка только прямо соответствующих им по основному лексическому значению слов с низкой частотой, причем данное соответствие носит отчетливо выраженный индивидуально-авторский характер.

Полученные данные подтверждают большую роль, которую могут играть частотные словари при исследовании вопросов авторской стилистики и стилистики вообще, позволяя идентифицировать авторскую лексику на основании статистических данных. Дальнейшие исследования с расширением круга анализируемых авторов и увеличением количества обрабатываемых текстов позволят значительно углубить результат проделанного анализа.

Литература

Герд А. С. К вопросу о роли низкочастотных фактов в лингвистическом исследовании // Структурная и прикладная лингвистика: межвуз. сб. СПб., 1993. Вып. 4.

Гребенников А. О. Исследование устойчивости лексико-статистических характеристик текста: автореф. дис. … канд. филол. наук. СПб, 1998.

Гребенников А. О. Частотный словарь и образ мира писателя // Словоупотребление и стиль писателя: межвуз. сб. СПб., 2006. Вып. 3.

Карпова О. М., Кириллов М. А. Из опыта составления частотного словаря рассказов Ф. С. Фицджеральда. URL: http://fitzgerald.narod.ru/critics-rus/karpova-kirillov.html (дата обращения: 24.12.2015).

Корпус древнерусских агиографических текстов СКАТ: современное состояние и перспективы развития / Е. Л. Андреева [и др.] // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам: Материалы международной научной конференции (Ижевск, 13–17 июля 2006 г.) / отв. ред. В. А. Баранов. Ижевск, 2006.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М., 2009. URL: http://dict.ruslang.ru/freq.php (дата обращения: 24.12.2015).

Частотный словарь рассказов А. И. Куприна / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 2006.

Частотный словарь рассказов А. П. Чехова / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 1999.

Частотный словарь рассказов И. А. Бунина / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 2012.

Частотный словарь рассказов Л. Н. Андреева / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 2003.

А. В. Добров

КОМПЬЮТЕРНЫЙ СЕМАНТИКО-СИНТАКСИЧЕСКИЙ АНАЛИЗ ЯЗЫКОВЫХ ОБОЗНАЧЕНИЙ ДЕЙСТВИЙ ИЛИ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ГОСУДАРСТВЕННОЙ ВЛАСТИ[1]

Аннотация. В статье представлен опыт разработки семантических средств для распознавания обозначений деятельности или действий органов государственной власти с помощью лингвистического процессора AIIRE, включающего в себя также одноименную онтологию. Исследование осуществляется на материале новостных сообщений, полученных совместно с сотрудниками Центра технологий электронного правительства НИУ ИТМО. Данная работа является частью серии исследований «повестки дня», формируемой средствами массовой информации по тематике, связанной с развитием электронного правительства.

Ключевые слова. Автоматическое понимание текстов, онтологическая семантика, концептуальные отношения, лексическая неоднозначность.

Alexey V. Dobrov

SEMANTIC AND SYNTACTIC COMPUTER ANALYSIS OF LINGUISTIC DENOTATIONS OF ACTS OR ACTIVITIES OF PUBLIC AUTHORITIES

Abstract. The article describes the ongoing research aimed at creating semantic tools for recognition of denotations of public authorities acts or their activities with help of AIIRE natural language processor and ontology. The study is carried out on the material of news items, collected in collaboration with ITMO University Center for Electronic Government. This work is a part of a series of studies focused on the "Agenda" through the media resources on topics related to e-government development.

Keywords. NLU, ontological semantics, conceptual relations, lexical disambiguation.

[1] Работа выполняется в рамках проекта «Разработка инструмента опинионмайнинга и его апробирование на задачах обследования общественного мнения о деятельности органов власти» (НИР № 415825, Университет ИТМО).

Введение

В рамках проекта «Разработка инструмента опинион-майнинга и его апробирование на задачах обследования общественного мнения о деятельности органов власти» возникла задача выделения в текстах обозначений действий и различных видов деятельности органов государственной власти (далее ОГВ). Коллективом ООО «АИРЕ» была разработана система мониторинга, которая, совместно с сотрудниками Центра технологий электронного правительства НИУ ИТМО, была настроена для сбора данных о деятельности ОГВ.

Традиционные подходы к решению задачи выделения сущностей, основанные на шаблонах, оказались не вполне адекватными целям данного исследования, поскольку выделение шаблонов обозначений действий или различных видов деятельности ОГВ не даёт возможности в дальнейшем выполнять автоматическое выделение их оценок, и было решено использовать семантические технологии: лингвопроцессор AIIRE и одноимённую онтологию. В ходе компьютерного анализа языковых единиц, обозначающих действия или деятельность ОГВ, возникает необходимость выделения семантических отношений между понятиями указанной предметной области. Такие отношения позволяют строить семантические графы и разрешать неоднозначность различных видов (морфологическую, синтаксическую, лексическую).

В данной работе описываются методы компьютерного анализа языковых выражений, обозначающих действия или деятельность ОГВ, с использованием AIIRE.

Лишь в некоторых из работ, посвящённых исследованиям политической лексики, производится структурно-лингвистический анализ терминологии государственного управления. В этих работах основной акцент видится несколько смещённым в сторону обсуждения того, является ли данная терминология самостоятельной терминосистемой [Нгуен], как эта терминосистема соотносится с терминологией государственного управления, какова структура этой терминосистемы: как представлены её ядро и периферия, каковы особенности политической метафоры [Чудинов] и т. д.

Вопрос о терминологичности единиц политической лексики обусловлен широким распространением не свойственной терминам

синонимии, полисемии и омонимии. Согласно А. С. Герду, наименования учреждений следует относить не к терминам, а к идентификаторам [Герд, с. 4], но «неоднократные призывы филологов различать термины и номены во многом остаются на уровне абстрактных филологических требований, которые невозможно реализовать на практике» [Там же. С. 3].

Родо-видовые отношения между понятиями, стоящими за лексическими значениями единиц политической лексики, исследуются в работе Т. Н. Юдиной и А. В. Богомоловой [Юдина, Богомолова]. В этой работе рассматриваются теоретические вопросы организации компьютерной онтологии предметной области «Государственное управление».

Существующие в современной научной литературе сведения о семантических свойствах политической лексики, в особенности наименований ОГВ, не отражают фактического употребления различных языковых средств, выражающих действия или виды деятельности ОГВ. Как будет показано ниже, даже сами названия ОГВ часто строятся из таких средств. Среди них частотны отвлечённые процессуальные имена существительные (далее ПС), синтаксические и семантические свойства которых сами по себе на сегодняшний день нуждаются в исследовании.

ПС характеризуются явными особенностями уже на уровне так называемых присловных связей и требуют особого подхода при семантическом анализе синтаксических структур. Такой подход представлен в монографии В. П. Казакова [Казаков]. Вслед за Ю. Д. Апресяном В. П. Казаков отмечает, что ПС, образованные от глаголов (так называемые девербативы), наследуют семантические валентности этих глаголов [Апресян, с. 165]. Вместе с тем даже ПС остаются именами существительными, и потому, в отличие от глаголов, становятся «„опредмеченным“ наименованием — всё же действия, состояния, качества» [Золотова, с. 126]. Концепты, стоящие за значениями глагола и соответствующего ему девербатива, различны, вопреки традиционному в компьютерной лингвистике отождествлению этих концептов (так, концепты ‘бег’ и ‘бежать’ не дифференцируются в большинстве компьютерных онтологий).

В диссертационном исследовании Л. А. Вакарюк [Вакарюк] показано, что далеко не все ПС являются девербативами (ср. *акт*, *процесс*, *реверанс*, *лекция, навигация* и др.). При этом сохраняется

способность ПС непосредственно присоединять наречия времени и места (ср. *лекция вчера*, *навигация повсюду*) и присоединяться к лексемам процессуальной семантики (ср. *прервать спектакль*).

В. В. Богданов квалифицирует функцию девербатива в предложении как «непредметный аргумент», позицию которого могут занимать «отпредикатные существительные, инфинитивы, герундии и т. д.» [Богданов, с. 172]. ПС обозначают свёрнутые пропозиции, а «наличие пропозитивных существительных в некотором предложении свидетельствует о его полипропозициональном строении» [Там же. С. 173]. Следовательно, при наличии в словосочетании ПС это словосочетание может быть развёрнуто в предложение: ср. *продажа Минфином акций — Минфин продаёт акции*. При этом семантические отношения в словосочетании и соответствующем предложении идентичны или имеют взаимно-однозначное соответствие.

Метод межуровневого взаимодействия

Использование статистических эвристик для разрешения неоднозначности при автоматической обработке текста гораздо популярнее применения методов, основанных на правилах. Корпусные подходы (так называемые методы машинного обучения) имеют существенный успех в разрешении морфологической неоднозначности (например, Ян Хаджич и соавторы оценивают качество этих методов в 95% [Serial Combination …]). Корпусные методы более просты и объективны, чем методы, основанные на правилах, так как не требуют создания этих правил. Главный недостаток статистических эвристик состоит в том, что они не дают гарантии отсутствия ложноотрицательных результатов: некорректно исключённая версия морфологического анализа может привести к потере целостности всего синтаксического дерева.

Работа AIIRE основана на методе межуровневого взаимодействия, впервые предложенном Григорием Самуиловичем Цейтиным в 1985 г. [Цейтин]. Несмотря на то что метод был предложен уже тридцать лет назад, он до сих пор не был применён на практике ввиду сложности его алгоритмической реализации и отсутствия соответствующего программного обеспечения. Цель упомянутого метода — избавиться от искусственного разделения уровней лингвистического анализа и анализировать текст одновременно на всех

языковых уровнях. Такой подход позволяет устранять неоднозначность на более низких уровнях, используя правила более высоких уровней ещё до того, как неоднозначность более низких уровней успевает привести к так называемому комбинаторному взрыву.

Морфологическая неоднозначность может быть снята по результатам непосредственного синтаксического связывания (или не связывания) первых двух разобранных словоформ, в соответствии с ограничениями грамматики. Синтаксическая неоднозначность может быть снята при невозможности семантического связывания значений элементов синтаксического дерева [Dobrov].

Главный инструмент AIIRE для разрешения неоднозначности — онтология, содержащая в себе концепты (модели понятий), стоящие за значениями лексических единиц, и обеспечивающая возможность вычисления их семантических валентностей на основании тех отношений, которые установлены между концептами.

Концепты онтологии AIIRE и отношения между ними

Концепт в онтологии AIIRE — это набор атрибутов, где каждый атрибут представляет собой пару «отношение — объект». Для экономии вычислительных ресурсов в онтологии хранятся только прямые отношения, а обратные — вычисляются. Виды используемых отношений описаны в работе [Dobrov]. В онтологии AIIRE существуют строгие правила наследования и замещения отношений. Если один концепт наследует другой, то каждый атрибут наследуемого воспроизводится наследующим концептом, однако наследуемый атрибут может быть замещён другим атрибутом, если и отношение, и объект замещаемого атрибута наследуются или совпадают с отношением и объектом замещающего. Например, концепт ‘отрезок’ имеет атрибут <‘иметь размер’, ‘длина’>, в то время как его подкласс ‘период времени’ обладает замещением <‘иметь размер’, ‘длительность’>, что означает, что размер (длина) временного периода — это его длительность (концептом ‘длительность’ также наследуется концепт ‘длина’).

В систему анализа текстов AIIRE входит лингвопроцессор (далее ЛП) — программное средство, осуществляющее семантический анализ текста, то есть представление его семантики в виде семантического графа (далее СГ) на основании данных из онтологии и модулей

грамматики (морфологии и синтаксиса). ЛП производит семантические графы в качестве конечного представления текстовой семантики. Термин «семантический граф» используется в данной статье в широком смысле и не является синонимом термина «концептуальный граф», предложенного Дж. Совой [Sowa]. Рёбра СГ могут быть вершинами и иметь собственные рёбра, при этом вершины и рёбра СГ — концепты онтологии или наследующие их текстовые концепты.

Процедура нормализации СГ, то есть его приведения к единообразному виду, в котором обратные отношения заменяются прямыми, позволяет унифицировать семантические представления различных обозначений одних и тех же ситуаций (ср. *МФ продаёт акции*, *продажа акций МФ*; *продающий акции МФ*, *проданные МФ акции* и т. д.). В процессе нормализации концепты, не являющиеся корнями синонимических рядов (далее СР) — выделенными концептами, используемыми для идентификации СР, — заменяются на корни СР, что позволяет отождествлять, например, *продажа акций МФ*, *продажа акций Минфином*, *сбыт акций МФ* и т. д.

Названия ОГВ в онтологии

Компоненты названий ОГВ

Названия ОГВ, с точки зрения их разбора ЛП, можно разделить на три различных вида: полные названия (например, *Министерство финансов, Федеральная Антимонопльная Служба, Федеральное агентство связи*), слоговые аббревиатуры (*Минобр*, *Миобороны, Минэкономразвития*) и звуковые аббревиатуры (*ФАС*, *ФНС*, *МВД*).

Для корректного разбора ЛП естественно-языковых выражений все концепты, которые могут стоять за лексическими значениями единиц, составляющих эти выражения, должны содержаться в онтологии. Кроме того, в онтологии должны быть установлены отношения между этими концептами, позволяющие, с одной стороны, строить гипотезы о семантических связях между значениями лексических единиц в выражении и, с другой стороны, разрешать семантическую неоднозначность. Эти требования относятся и к названиям ОГВ. Вместе с тем название ОГВ, представленное сложным словосочетанием, обозначает единый концепт и часто является идиоматическим выражением. Онтология должна хранить идиомы и отношения между концептами, обозначаемыми словами, входящими в эти идиомы.

Например, название *Министерство здравоохранения* может быть разобрано ЛП только в том случае, если в онтологии обработаны концепты ‘министерство’ и ‘здравоохранение’ и определены концептуальные отношения, которые при разборе позволят ЛП связать эти концепты между собой. Для решения данной задачи в онтологии AIIRE для концепта конкретного ОГВ создается отношение ‘(о субъекте) осуществлять деятельность в предметной области’, направленное на концепт, представляющий собой вид деятельности ОГВ. Это отношение является частным случаем абстрактного отношения ‘(об объекте или процессе) принадлежать объекту или процессу’, обозначаемого сочетанием именной группы (далее ИГ) с несогласованным определением в родительном падеже (далее ИГРП). При этом концепт ‘министерство здравоохранения’ является подклассом концепта ‘министерство’ и наследует все его отношения, в том числе отношение ‘осуществлять деятельность в области…’, направленное на концепт ‘здравоохранение’, что позволяет ЛП разбирать словосочетание *министерство здравоохранения* как ‘министерство, осуществляющее деятельность в области здравоохранения’.

Аббревиатуры и проблема рода

В разборе ЛП аббревиатур имеются трудности, связанные с необходимостью определять род, который может не совпадать с родом, употребляемым в тексте. Склонение звуковых инициальных аббревиатур зависит от их опорного слова, то есть, например, если опорное слово — мужского рода, то и вся аббревиатура приобретает мужской род. В то же время если опорное слово — женского рода, но вся аббревиатура имеет морфологический облик слова, принадлежащего к мужскому роду, то указанное правило нарушается. К таким случаям относятся аббревиатуры, имеющие опорное слово женского рода и оканчивающиеся на согласный, например *ФАС* (*Федеральная Антимонопольная Служба*), *ЦИК* (*Центральная избирательная комиссия*). Опорное слово в таких случаях имеет женский род, но, так как аббревиатура оканчивается на согласный, она может употребляться в тексте как слово мужского рода (*ФАС произвёл проверку*, *ЦИК опубликовал результаты*).

Фактическое употребление аббревиатур в современном медиадискурсе вариативно, поэтому в подобных случаях в морфологическом словаре ЛП необходимо хранить два варианта рода аббревиатур.

Иногда AIIRE хранит в морфологическом словаре аббревиатуры в трёх экземплярах: мужского, женского и среднего рода. Данная ситуация типична для случаев омонимии: ср. *ФАС* (*Федеральная Антимонопольная Служба*) и *ФАС* (*Федеральное агентство связи*).

Названия ОГВ

Анализ синтаксических конструкций более 360 названий федеральных ОГВ показал, что в системе этих наименований используются всего два типа синтаксических структур: ИГРП (например, *Министерство финансов*, *Министерство иностранных дел*, *Минобр России*) и модифицированная предложным определением ИГ (далее ИГПО) (*Агентство по труду и занятости*).

Для разбора ИГРП ЛП осуществляет запрос, то есть процедуру поиска атрибута по абстрактному отношению, обозначаемому родительным падежом (далее РП), или любого подкласса этого отношения у концепта, обозначаемого ИГ в РП. При наличии такового проверяется, является ли концепт, обозначаемый ИГ в препозиции, подклассом или надклассом концепта, на который направлено отношение. Если связь найдена, то концепты, обозначаемые ИГ, семантически связаны найденным отношением, а концепт, обозначаемый ИГ в препозиции, замещён объектом отношения. Например, при разборе *министерство здравоохранения* ЛП обнаруживает отношение ‘(о предметной области) быть областью, в которой осуществляет деятельность субъект’, направленное на концепт ‘министерство здравоохранения’.

Аналогично ИГРП, ЛП обрабатывает ИГПО путём запроса отношения ‘обладание аргументом предложного отношения’.

В ЛП AIIRE используется внутренняя вспомогательная база данных идиом и соответствующих им СГ. Для её построения производится анализ каждой идиомы, и результат анализа сохраняется в базе идиом. В ходе разбора текста, если СГ содержит часть, соответствующую СГ идиомы, производится замена этой части на соответствующий концепт из базы идиом. Некоторые идиомы могут содержать в своем составе другие идиомы. Так, название *Министерство по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий* содержит такие выражения, как *гражданская оборона*, *чрезвычайная ситуация* и *стихийное бедствие*, каждое из которых рассматривается как идиома.

Обработка сложных идиоматических выражений требует наличия в базе идиом составных частей, являющихся идиомами. Поэтому построение базы идиом выполняется последовательно, от наименьшего количества словоформ в выражении к наибольшему.

Концепты действий, состояний и различных видов деятельности

Согласно Лингвистическому энциклопедическому словарю [Маслов], глаголы могут быть динамическими (предельными и непредельными) или статическими. Предельные динамические глаголы характеризуются завершённостью процесса, то есть обозначают действия субъекта, которые предполагают завершение. Непредельные динамические глаголы не предусматривают предела в протекании процесса и обозначают деятельность субъекта. Статические глаголы обозначают состояния (состоянием может быть, в частности, отношение субъекта к чему-либо). С концептуальной точки зрения, в онтологии AIIRE деятельность — частный случай состояния (субъект находится в состоянии осуществления какого-либо вида деятельности); далее действия, виды деятельности или состояния будут обозначаться ДС. Для предельных динамических глаголов в онтологии выделяются концепты, соответствующие процессу (несовершенный вид) и завершению процесса (совершенный вид). Внесение в онтологию концепта ПС покрывает все возможные варианты обозначения действия (ср. *суд приговорил*, *суд приговаривает*, *приговор суда*). Все три концепта должны сохранять валентности, в том числе связи с предлогами, что означает сохранение параллелизма трёх классификаций. Для этого ДС подразделяются на направленные (переходные глаголы) и ненаправленные, а также адресованные (управляющие дативом) и неадресованные. Для непредельных динамических глаголов указывается концепт, соответствующий завершению, который, в свою очередь, является действием (например, деятельность — *работать*, завершение — *доработать*, которому соответствует *дорабатывать*). Для статических глаголов в онтологии достаточно указания процесса, соответствующего состоянию.

При формировании параллельных классификаций в онтологии AIIRE созданы концепты, соответствующие различным семантическим классам значений глаголов и предлогов. В ходе вычисления СГ

для выражения с предложными группами с помощью правил наследования атрибутов производится необходимое связывание и снятие неоднозначности. Так, классы глаголов перемещения, передачи информации, наблюдения, соответствуют различным классам предлогов.

Синтаксические средства обозначения действий и деятельности ОГВ

Действия и деятельность ОГВ могут быть обозначены простыми двусоставными предложениями со сказуемым в действительном (*Минфин продаёт активы*) или страдательном залоге (*активы продаются Минфином*). На уровне синтаксической семантики приведённые примеры получают одинаковую интерпретацию в виде нормализованных СГ (рис. 1).

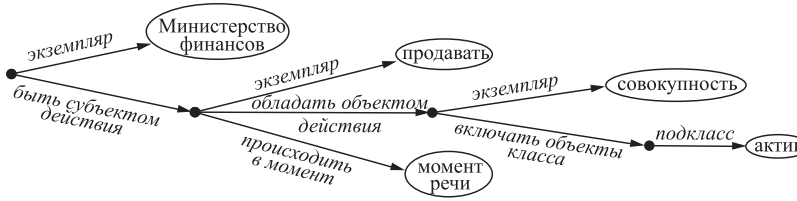

Рис. 1. Семантический граф (СГ)

Простые двусоставные предложения могут содержать в своей структуре распространители, обозначающие действия или деятельность ОГВ, например деепричастные обороты (ср. *продавая активы, Минфин постановил…*). В нормализованном СГ значения деепричастных оборотов отображаются так же, как и значения иных глагольных групп, при этом устанавливается связь субъекта основного ДС с ДС, обозначаемым деепричастным оборотом, а также отношение одновременности или предшествования между ними. Указанные распространители могут быть выражены группами ПС, и их значения встраиваются в СГ в соответствии с синтаксической позицией группы.

Субъект, равно как и объект действия или деятельности может быть выражен при помощи ИГ в РП (ср. *приговор суда*, *роспуск Госдумы*). Выбор субъектной или объектной интерпретации основыва-

ется исключительно на семантических валентностях, что не всегда позволяет разрешить неоднозначность: для выражения *приговор суда* возможна трактовка, в которой суд не приговаривает кого-либо, а сам является объектом приговора. Такая неоднозначность разрешается при помощи контекста или путём регистрации отдельных идиом. Контекстуальное разрешение указанной неоднозначности возможно в случаях, когда ПС образованы от переходного глагола и могут присоединять к себе ИГ в творительном падеже, обозначающую субъект ДС, и ИГ в РП в роли объекта (ср. *продажа активов Минфином*).

Действия и деятельность ОГВ могут быть выражены не только самими ИГ, но и их распространителями, в частности причастными группами в препозиции (*продающий активы Минфин*) и обособленными причастными оборотами в постпозиции (*Минфин, продающий активы*). Сходную с причастными группами функцию выполняют определительные придаточные предложения (*Минфин, который продаёт активы*). Действия или деятельность ОГВ могут также обозначаться группами страдательных причастий с обозначением субъекта, формально выраженным ИГ в творительном падеже (*проданные Минфином активы*; *активы, проданные Минфином*). Нормализованный СГ для ИГ, распространённой группой страдательного причастия, должен совпадать с СГ ИГ, распространённой группой действительного причастия, с учётом перестановки ИГ (то есть *проданные Минфином активы* после нормализации СГ получает такую же интерпретацию, как и *продавший активы Минфин*).

Выводы

В ходе выполнения работы были выделены различные синтаксические конструкции системы наименований ОГВ, а также синтаксические конструкции, обозначающие действия и деятельность ОГВ. Были разработаны принципы построения отношений между концептами, соответствующими названиям ОГВ и их видам деятельности. Онтология AIIRE была дополнена в соответствии с разработанными принципами. В ходе исследования были разработаны параллельные классификации ДС в соответствии с различными классами глаголов и ПС, представленных в имеющейся текстовой коллекции, обеспечивающие вычисление семантических валентностей

исследуемых единиц. Были рассмотрены и решены некоторые проблемы анализа идиоматических единиц в названиях ОГВ. Таким образом, была обеспечена корректная обработка ЛП AIIRE русскоязычных языковых единиц, обозначающих различные виды деятельности и действия органов государственной власти Российской Федерации.

Литература

Апресян Ю. Д. Лексическая семантика (синонимические средства языка). М., 1974.

Богданов В. В. Моделирование семантики предложения // Прикладное языкознание: Учебник. СПб, 1996.

Вакарюк Л. А. Структурно-семантический анализ имен существительных со значением процесса, не мотивированных глаголами (на материале русского языка): автореф. дис. … канд. филол. наук; 10.02.01. Черновцы, 1985.

Герд А. С. Основы научно-технической лексикографии. Л., 1986.

Золотова Г. А. Коммуникативные аспекты русского синтаксиса: моногр. М., 1982.

Казаков В. П. Синтаксис имен действия. СПб., 1994.

Маслов Ю. С. Глагол // Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. 2-е изд., доп. М., 2002.

Нгуен Т. Т. В. Терминология государственного управления в современном русском языке: автореф. дис. … канд. филол. наук. М., 2001.

Цейтин Г. С. Программирование на ассоциативных сетях // ЭВМ в проектировании и производстве. Л., 1985. Вып. 2.

Чудинов А. П. Российская политическая метафора в начале XXI века // Политическая лингвистика. Екатеринбург, 2008. Вып. 1(24). С. 86–93.

Юдина Т. Н., Богомолова А. В. УИС РОССИЯ: онтология предметной области «государственное управление» // Интернет и современное общество: сб. науч. статей. Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». СПб., 2011.

Dobrov A. V. Semantic and Ontological Relations in AIIRE Natural Language Processor // Computational Models for Business and Engineering Domains. Rzeszow, Sofia, 2014.

Serial Combination of Rules and Statistics: A Case Study in Czech Taggin / J. Hajic [et al.] // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001). Tolouse, 2001.

Sowa J. F. Conceptual Graphs: Draft Proposed American National Standard // International Conference on Conceptual Structures ICCS-99. Lecture Notes in Artificial Intelligence 1640. Berlin; New-York, 1999. P. 1–65.

В. П. Захаров

КОРПУСНО-ОРИЕНТИРОВАННЫЙ ПОДХОД К ПОСТРОЕНИЮ ТЕЗАУРУСОВ И ОНТОЛОГИЙ[1]

Аннотация. Статья представляет результаты исследования по автоматизированному построению лексического ядра тезауруса по узкой предметной области. Тезаурус строится на основе дистрибутивно-статистического анализа большого корпуса текстов (около 15 млрд словоупотреблений) с помощью инструментов системы Sketch Engine. Для парадигматических и синтагматических отношений лексических единиц тезауруса вычисляется количественный показатель силы связи. Изучено влияние объема корпуса на качество тезауруса.

Ключевые слова. Корпусная лингвистика, корпусы текстов, тезаурус, онтология, лексико-семантическое поле, термины, дистрибутивно-статистические методы, коллокации, автоматический анализ.

Victor P. Zakharov

CORPUS-BASED APPROACH TO THESAURUS AND ONTOLOGY CONSTRUCTION

Abstract. The paper presents the results of automatic thesaurus construction in a narrow subject area. Thesaurus is designed on the base of distributional and statistical analyses of a big text corpus of about 15 billion words by Sketch Engine system tools. Paradigmatic and syntagmatic relations in a thesaurus are evaluated quantifically. The corpus volume influence on the quality of a thesaurus is studied.

Keywords. Corpus linguistics, text corpora, thesaurus, ontology, terms, distributional and statistical methods, collocations, automatic analysis.

[1] Работа выполнена в рамках научно-исследовательского проекта РГНФ № 15-04-12029 «Программная разработка электронного ресурса с онлайн-версией русскоязычной вопросно-ответной системы».

Теоретические основания исследования

Данная статья посвящена методам автоматического выявления парадигматических связей, а именно методам наполнения лексико-семантических полей как основы построения тезаурусов и онтологий.

Самые разнообразные лингвопроцессоры явно или неявно опираются в своей работе на словари разного типа, чисто машинные словари или традиционные словари, приспособленные для компьютерного использования. В качестве примера можно привести словари систем морфологического анализа, базирующиеся на Грамматическом словаре А. А. Зализняка. Синтаксические парсеры также требуют различных словарей, в первую очередь словарей валентностей. Как «уровневые» анализаторы, так и прикладные системы используют словари сочетаемости, терминологические словари, фактографические справочники и т. п.

В большинстве случаев в этих словарях отражаются два аспекта функционирования языкового знака — синтагматика и парадигматика. Синтагматические связи между лексическими единицами нашли свое отражение, пусть и неполно, в традиционных словарях разного типа. Корпусная лингвистика часто объединяет разные типы сочетаемости под понятием «многословные единицы» (multiword expressions, MWE). К ним относятся аналитические формы (глаголы, компаративы, суперлативы), фразеологизированные единства, общепринятые словесные формулы (прагматемы) и т. д., вплоть до названий организаций, произведений, имен лиц и т. п. На уровне токенизации эти единицы разбиваются на отдельные слова, тем не менее в разных задачах мы должны рассматривать их как единое целое. Эти единицы могут быть как дистантными, так и неразрывными. Для их автоматического выделения из текста используются различные автоматизированные методы (автоматическое выделение коллокаций, извлечение именованных сущностей и т. д.).

Сложнее обстоит дело со словарями, которые отражают парадигматику. Здесь под парадигматикой мы понимаем не набор парадигм, а системные отношения между лексическими единицами. Слова в языке существуют не изолированно, они находятся в разнообразных связях и отношениях с другими значениями того же слова и со значениями других слов. Семантический уровень языка представ-

ляет собой упорядоченную систему, элементы которой находятся в отношениях взаимосвязи и взаимообусловленности. Имея в виду связи между значениями слов, говорят о лексико-семантической системе языка или подъязыка. Элементами ее являются отобранные по определенным правилам лексические единицы естественного языка, а структура изоморфна структуре логических связей между понятиями специальной области знаний и деятельности.

И если синтагматические отношения представлены в тексте, можно сказать, в явном виде и могут извлекаться из него, то парадигматика в тексте скрыта, и для ее выявления или требуются разнообразные знания человека, или должны разрабатываться гораздо более изощренные процедуры, чем в случае синтагматики.

Среди словарей, которые можно назвать парадигматическими, особое место занимают тезаурусы и онтологии, которые представляют собой словари понятий, фиксирующие человеческие знания. Это модели, формализующие план содержания, однако реально на уровне текста мы все равно вынуждены работать со словами. В конце XX — начале XXI в. было осознано, что многие задачи — как чисто лингвистические, например снятие синтаксической неоднозначности или разрешение референции, так и прикладные — не могут быть решены без обращения к семантике и, следовательно, без таких словарей. «Ключевым моментом системы семантического анализа является эффективная словарная поддержка. В этом смысле любая система семантического анализа является онтологически ориентированной. Поэтому основная проблема в создании реально работающих анализаторов — это проблема реально работающего понятийного словаря» [Рубашкин, с. 252].

Задачу моделирования понятийной системы можно разбить на две части: 1) выявление системы понятий, 2) выявление отношений между ними. Данная статья посвящена решению первой задачи. Эта задача может решаться «вручную», путем экспликации и формализации профессионального знания, накопленного в процессе человеческой деятельности, на основе знаний специалистов и с использованием имеющихся словарей, учебников и других пособий. Этот путь долгий и трудоемкий. Однако поскольку наши знания о мире так или иначе находят отражение в текстах, то можно поставить задачу извлечения системы понятий из текстов. Минимальный набор требований при этом следующий: множество этих автоматически

извлеченных понятий должно быть достаточно полным и сами понятия должны быть связаны между собой. Характер связей на этом первом этапе автоматически не устанавливается. В нашем случае можно говорить о принципе когнитивной однородности [Там же. С. 248], когда на каждом этапе решается одна задача. В данной работе это выявление множества основных взаимосвязанных понятий вокруг выбранного ядерного элемента (ключевого слова).

В лингвистике совокупность языковых единиц, объединенных общностью содержания и отражающих понятийную близость, известна под названием «поле» (семантическое поле, лексико-семантическое поле, функционально-семантическое поле и т. п.) [Уфимцева; Щур]. «Поле — совокупность содержательных единиц, покрывающая определенную область человеческого опыта и образующая более или менее автономную микросистему» [Ахманова]. При построении первых лингвистических тезаурусов (П. Роже, Ф. Дорнзайф, Р. Халлиг) уже отмечалась связь этого понятия с по́левой структурой лексики.

В трактовке В. Г. Адмони поле характеризуется наличием инвентаря элементов, связанных системными, то есть парадигматическими отношениями. По аналогии с практикой «тезаурусостроения» можно сказать, что в целом это ассоциативные связи, или отношение онтологической зависимости. В. Г. Адмони усматривает в поле центральную часть — ядро, элементы которого обладают полным набором признаков, определяющих данную группировку, и периферию, элементы которой обладают не всеми характерными для поля признаками, но могут иметь и признаки, присущие соседним полям [Адмони]. Поле предполагает непрерывность связей объектов множества, причем на некоторых участках поля создаются области, в которых связи особенно интенсивны, а признаки особенно сильно выражены. Тогда говорят о лексико-семантических группах — элементарных микрополях, объединяющих слова, обычно относящиеся к одной части речи и наиболее сильно связанные отношением семантической близости. В общем же случае для поля характерна нечеткость границ между частями речи. В теории баз знаний лексико-семантическая система в целом (множество связанных между собой полей) трактуется как онтология.

Методология исследования

Одним из старых и известных методов лингвистического исследования является дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах. Уже на заре компьютерной лингвистики предпринимались попытки на основе частотной информации о встречаемости лексических единиц в контекстах определенной величины получать по некоторой заданной формуле количественную характеристику их связанности, что впоследствии нашло выражение в методах выявления коллокаций и многословных единиц на основе мер ассоциации. Одновременно выдвигались идеи распространения этого метода и на парадигматический аспект языка, идеи о том, что парадигматические связи могут выводиться из связей синтагматических [Шайкевич, 1963; Арапов; Богданов; Пиотровский; Караулов; О некоторых лексико-семантических проблемах…; Гринев, Лейчик].

Принцип перехода от изучения текстуальных связей (синтагматических) к системным (парадигматическим) лежит в основе различных дистрибутивно-статистических методик [Шайкевич, 1976; Шайкевич, 1982; Pekar]. Считается, что два элемента связаны парадигматически, если оба они текстуально систематически связаны с каким-то третьим элементом. Значит, представляется разумным предположить, что сила парадигматической связи должна возрастать с увеличением числа и силы общих синтагматических связей [Шайкевич, 1976, с. 370].

Однако возможности вычислительной техники того времени не позволяли проверить эти идеи на практике. Далее, чтобы можно было говорить о закономерности статистических распределений, нужны очень большие массивы данных [Smrž, Rychlý]. Таковые появились только с развитием Интернета и созданием больших корпусов текстов. Одновременно стали появляться и соответствующие программные средства [The Sketch Engine; Sharoff; Сидорова; TTC: Terminology…]. Также важно учитывать наличие синтаксической связи между контекстно близкими элементами текста [Syntactic-Based Methods…; Pazienza, Pennacchiotti, Zanzotto].

Была поставлена задача: проверить работоспособность дистрибутивно-статистического метода и инструментария корпусной

лингвистики на базе больших корпусов, то есть автоматически сформировать лексическую основу тезауруса.

Инструмент и материал исследования

Для работы с корпусами требуется особый программно-лингвистический инструментарий. В данном эксперименте использована система Sketch Engine [The Sketch Engine], которая представляет собой корпусный менеджер, работающий с морфологически размеченным корпусом.

Sketch Engine, в числе прочих функций, формирует частотные списки лексических единиц, входящих в корпус, и эти частоты, безусловно, характеризуют лексический состав данного языка или подъязыка. Наряду с этим при отборе характеристической лексики относительные частоты единиц, включаемых в тезаурус, в текстах исследуемого корпуса должны существенно превосходить частоту этих слов в некотором фоновом неспециализированном корпусе. Такая методика была использована А. Я. Шайкевичем в «Словаре языка Достоевского» для выделения так называемых лексических маркеров [Шайкевич, Андрющенко, Ребецкая]. Подобная возможность имеется и в Sketch Engine (comparable frequency list). Однако наша задача другая, а именно выявление и вычисление парадигматического ядра тезауруса, или, иными словами, лексико-семантического поля.

Особенность системы Sketch Engine — наличие в ней специальных инструментов, реализующих методику дистрибутивно-статистического анализа: «Тезаурус» (Thesaurus) — построение тезауруса; «Кластеризация» (Clustering) — группировка единиц тезауруса в кластеры (лексико-семантические группы); «Дифференциация» (Sketch diff) — выявление сходства и различия в сочетаемости для пар слов; «Коллокации» (Collocations) — автоматическое выявление коллокаций; «Лексические шаблоны» (Word sketch) — выявление коллигаций (коллокаций, ограниченных синтаксической моделью). Все они разными способами выявляют парадигматические (то есть семантические) связи между терминами с количественным указанием силы этой связи.

«Тезаурус» в системе Sketch Engine (его можно охарактеризовать как дистрибутивный тезаурус) позволяет увидеть, какие слова имеют схожую дистрибуцию с заданным словом. Для вычисления

двигатель *(noun)*
ruTenTen [2011] freq = 2,066,742 (113.05 per million)

Lemma	Score	Freq	Cluster
мотор	0.560	590,265	агрегат [0.341, 415,228] движок [0.305, 224,399] дизель [0.265, 139,042]
автомобиль	0.295	5,415,679	машина [0.267, 5,899,922] авто [0.198, 763,847] самолет [0.193, 1,720,844] мотоцикл [0.185, 348,738]
насос	0.288	567,423	привод [0.285, 472,927] генератор [0.285, 391,057] вентилятор [0.216, 293,447] фильтр [0.211, 786,733] датчик [0.211, 647,667] компрессор [0.209, 220,470] кондиционер [0.193, 584,881]
прибор	0.285	1,583,195	механизм [0.272, 2,086,090] устройство [0.271, 3,545,701] оборудование [0.27, 4,428,495] аппарат [0.267, 1,958,038] блок [0.263, 2,003,819] система [0.251, 18,251,572] модель [0.244, 4,919,587] установка [0.242, 3,458,650] конструкция [0.234, 2,308,028] техника [0.229, 3,193,906] диск [0.225, 1,647,056] элемент [0.223, 3,537,867] камера [0.222, 1,550,599] инструмент [0.216, 2,264,279] компьютер [0.214, 2,497,874] корпус [0.213, 1,920,685] модуль [0.208, 969,885] деталь [0.206, 1,811,720] узел [0.205, 948,220] панель [0.193, 1,169,684] часть [0.187, 12,381,025] тип [0.185, 5,106,728] изделие [0.184, 2,161,851] компонент [0.182, 1,291,366] продукт [0.18, 4,639,169] схема [0.18, 2,604,256]
колесо	0.247	1,047,174	коробка [0.231, 840,107] шина [0.215, 595,074] труба [0.196, 1,423,759]
электродвигатель	0.235	132,171	турбина [0.189, 124,670]
котел	0.211	480,105	радиатор [0.191, 288,106]
тормоз	0.204	317,220	подвеска [0.202, 339,078]

Рис. 1. Гнездо тезауруса с выделенными кластерами для ключевого слова *двигатель*

парадигматического подобия слов рассматриваются наборы сочетаемости для пар слов с учетом синтаксического отношения (лексические шаблоны). Единицы семантического поля обладают общими синтагматическими и парадигматическими свойствами, что отражает их семантическую близость. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice [Rychlý] и с учетом лексико-синтаксических шаблонов [Kilgarriff, Rychlý]. Этот механизм в сочетании с инструментом «Кластеризация» позволяет строить тезаурус с кластерами лексических единиц, соответствующими лексико-семантическим группам (рис. 1).

Эксперименты и оценка результатов

Была поставлена задача: на основе корпусов текстов построить тезаурус для термина *двигатель* и оценить его силу (наполненность). В качестве исходного текстового материала был выбран корпус ruTenTen 2011 (18,28 млрд токенов, 14,55 млрд словоупотреблений). Количество употреблений заданного слова в корпусе составляет чуть более 2 млн (ipm = 113,05).

Инструмент «Тезаурус» по заданному корпусу порождает дистрибутивный тезаурус (табл. 1). Количество лексических единиц в тезаурусе при этом было установлено равным 60.

Данные представлены в трех столбцах: Lemma — слово; Score — значение статистической меры, показывающее степень семантической близости данного слова к ключевому; Freq — частота данного слова в корпусе. Слова в таблице упорядочены по значению статистической меры (Score).

На следующем шаге реализованный в системе инструмент «Кластеризация» автоматически формирует на множестве выданных слов лексико-семантические группы (кластеры, микрополя; см. рис. 1).

Анализ полученных кластеров, равно как и изучение имеющихся терминологических материалов по двигателям, показывает, что результат автоматического анализа и кластеризации хорошо соотносится с реальным миром машин и механизмов. Особенность системы: от каждого слова можно перейти ко всем контекстам его употребления в корпусе и уточнить релевантность его включения в тезаурус.

Встает вопрос о необходимом и достаточном объеме корпуса. Был построен тезаурус на подмножестве данного корпуса объемом в 998 млн словоупотреблений (в 15 раз меньше). Сравнение двух списков показало пересечение 52 слов из 60. Интересно сравнить слова, которые не совпали (табл. 2).

Даже неспециалист с уверенностью скажет, что слова из правого столбца более релевантны понятию *двигатель*. На основании этого сравнения можно заключить, что увеличение объема корпуса повышает качество тезауруса и уменьшает количество шума.

Существует, однако, извечное противоречие между точностью и полнотой. Чтобы достичь высокой полноты, мы построили новый тезаурус, задав количество лексических единиц в нем равным 600. Просмотр полученного списка показал его крайне низкую точность.

Таблица 1. **Результат работы инструмента «Тезаурус» для ключевого слова *двигатель***

Lemma	Score	Freq
мотор	0,560	590 265
агрегат	0,341	415 228
движок	0,305	224 399
автомобиль	0,295	5 415 679
насос	0,288	567 423
прибор	0,285	1 583 195
привод	0,285	472 927
генератор	0,285	391 057
механизм	0,272	2 086 090
устройство	0,271	3 545 701
оборудование	0,270	4 428 495
машина	0,267	5 899 922
аппарат	0,267	1 958 038
дизель	0,265	139 042
блок	0,263	2 003 819
система	0,251	18 251 572
колесо	0,247	1 047 174
модель	0,244	4 919 587
установка	0,242	3 458 650
электродвигатель	0,235	132 171
конструкция	0,234	2 308 028
коробка	0,231	840 107
техника	0,229	3 193 906
диск	0,225	1 647 056
элемент	0,223	3 537 867
камера	0,222	1 550 599
инструмент	0,216	2 264 279
вентилятор	0,216	293 447
шина	0,215	595 074
компьютер	0,214	2 497 874

Lemma	Score	Freq
корпус	0,213	1 920 685
датчик	0,211	647 667
котел	0,211	480 105
фильтр	0,211	786 733
компрессор	0,209	220 470
модуль	0,208	969 885
деталь	0,206	1 811 720
узел	0,205	948 220
тормоз	0,204	317 220
подвеска	0,202	339 078
авто	0,198	763 847
батарея	0,198	570 385
труба	0,196	1 423 759
кондиционер	0,193	584 881
панель	0,193	1 169 684
самолет	0,193	1 720 844
радиатор	0,191	288 106
турбина	0,189	124 670
аккумулятор	0,189	400 288
цилиндр	0,188	304 511
часть	0,187	12 381 025
лампа	0,186	684 590
мотоцикл	0,185	348 738
кузов	0,185	467 488
тип	0,185	5 106 728
изделие	0,184	2 161 851
компонент	0,182	1 291 366
автомат	0,180	769 619
схема	0,180	2 604 256
продукт	0,180	4 639 169

Большинство терминов характеризовались низким значением статистической меры (score) и не имели к двигателям никакого отношения либо только косвенное (через *автомобиль*). К 60 терминам из первого эксперимента новый тезаурус добавил чуть больше двух десятков: *мощность*, *клапан*, *редуктор*, *трансмиссия*, *топливо*, *бак*,

Таблица 2. **Отличия в лексическом наполнении двух тезаурусов, построенных на подмножествах корпуса ruTenTen 2011 разного объема**

ruTenTen 2011 sample (998 млн словоупотреблений)	ruTenTen 2011 (14,55 млрд словоупотреблений)
технология	*автомат*
станция	*аккумулятор*
сеть	*радиатор*
процессор	*тормоз*
объект	*турбина*
комплекс	*подвеска*
источник	*мотоцикл*
версия	*кузов*

подшипник, *регулятор*, *передача*, *запчасть*, *выключатель*, *сборка*, *стартер*, *ось*, *форсунка*, *тяга*, *ротор*, *поршень*, *карбюратор*, *электромотор*, *автоматика*, *моторчик*, *оборот*.

Однако, как известно, в каждой предметной области большая часть терминов, как правило, представлена словосочетаниями. Корпусные инструменты предоставляют возможность автоматического выявления коллокаций. В Sketch Engine есть два инструмента выявления устойчивых сочетаний: «Коллокации» и «Лексические шаблоны». Первый вычисляет силу связи между словами по всему корпусу, второй — в пределах заданной синтаксической формулы (шаблона). Остановимся на втором инструменте. Вся совокупность шаблонов называется грамматикой. В настоящий момент грамматика в Sketch Engine для русского языка[2] описывает следующие отношения: сочинительное, субъектное, объектное, атрибутивное, компаративное, обстоятельственное и отдельно сочетания с предлогами [Хохлова].

Инструмент «Лексические шаблоны» формирует список наиболее устойчивых сочетаний, вычисленных в соответствии со статистической мерой logDice отдельно по каждой синтаксической формуле, а также подсчитывает общее количество сочетаний в корпусе, соответствующих каждому отдельному шаблону. Данные второго вида, на наш взгляд, представляют отдельный интерес и заслу-

[2] Разработчик М. В. Хохлова (СПбГУ).

Таблица 3. **Количество связей по типам в корпусе ruTenTen 2011 для слова *двигатель***

Идентификатор связи (отношения)	Кол-во связей	Характеристика отношения (для заданного ключевого слова, КС)	Пример
gen_modifies	722 193	КС в родительном падеже	*мощность двигателя*
a_modifier	676 060	Определение при КС	*бензиновый двигатель*
prec_prep	306 413	Предложные сочетания с КС	*помимо двигателя*, с *двигателем*, *двигатель с*, *двигатель на* и др.
gen_modifier	212 003	КС управляет словом в родительном падеже	*двигатель внутреннего сгорания*
subject_of	161 924	Сказуемое при КС-подлежащем	*двигатель заводится*
и/или	126 893	Сочинительное отношение	*двигатели и трансмиссии*
object4_of	93 170	КС в винительном падеже	*глушить двигатель*
pp_с	47 763	КС в роли субъекта в конструкции с предлогом «с»	*двигатель с наддувом*
pp_obj_в	35 340	КС в роли объекта в конструкции с предлогом «в» (предлог управляет КС)	*стук в двигателе*
pp_в	32 436	КС в роли субъекта в конструкции с предлогом «в»	*двигатель в режиме*
inst_modifies	29 304	КС в творительном падеже	*оснащаться двигателем*
pp_obj_на	25 404	КС в роли объекта в конструкции с предлогом «на»	*форсунки на двигателе*
pp_на	24 685	КС в роли субъекта в конструкции с предлогом «на»	*двигатель на форсаже*
pp_obj_для	19 621	КС в роли объекта в конструкции с предлогом «для»	*масло для двигателя*

Продолжение табл. 3

Идентификатор связи (отношения)	Кол-во связей	Характеристика отношения (для заданного ключевого слова, КС)	Пример
pp_для	18007	КС в роли субъекта в конструкции с предлогом «для»	*двигатель для мотоблока*
pp_obj_от	13848	КС в роли объекта в конструкции с предлогом «от»	*привод от двигателя*
passive	12524	КС управляет пассивом	*двигатель оснащен*
pp_при	10564	КС в роли субъекта в конструкции с предлогом «при»	*двигатель при разгоне*
pp_от	8715	КС в роли субъекта в конструкции с предлогом «от»	*двигатель от перегрева*
object2_of	7765	КС в родительном падеже	*сопло двигателя*
pp_obj_к	6546	КС в роли объекта в конструкции с предлогом «к»	*запчасть к двигателю*
pp_по	5540	КС в роли субъекта в конструкции с предлогом «по»	*проверка двигателя по нагреву**
object3_of	4161	КС в дательном падеже	*навредить двигателю*
pp_obj_из	4129	КС в роли объекта в конструкции с предлогом «из»	*выжать из двигателя*
pp_к	3900	КС в роли субъекта в конструкции с предлогом «к»	*двигатель к мотоблоку*
pp_без	3683	КС в роли субъекта в конструкции с предлогом «без»	*двигатель без глушителя*
pp_obj_по	3341	КС в роли субъекта в конструкции с предлогом «по»	*механик по двигателям*

pp_до	3178	КС в роли субъекта в конструкции с предлогом «до»	*ресурс двигателя до капремонта*
pp_из	2400	КС в роли субъекта в конструкции с предлогом «из»	*двигатель из Японии*
pp_obj_о	2290	КС в роли объекта в конструкции с предлогом «о»	*подробности о двигателе*
pp_obj_над	2207	КС в роли объекта в конструкции с предлогом предлог «над»	*кабина над двигателем*
pp_obj_при	2024	КС в роли объекта в конструкции с предлогом «при»	*зажигание при неработающем двигателе*
pp_после	1903	КС в роли субъекта в конструкции с предлогом «после»	*двигатель после капремонта*
pp_у	1864	КС в роли субъекта в конструкции с предлогом «у»	*двигатель у мотоцикла*
pp_под	1838	КС в роли субъекта в конструкции с предлогом «под»	*двигатель под нагрузкой*
pp_obj_под	1701	КС в роли объекта в конструкции с предлогом «под»	*подушка под двигателем*
pp_за	1477	КС в роли субъекта в конструкции с предлогом «за»	*работа двигателя за счет*
pp_obj_между	1309	КС в роли объекта в конструкции с предлогом «между»	*муфта между двигателем и насосом*
pp_через	1278	КС в роли субъекта в конструкции с предлогом «через»	*в двигатель через форсунки*
pp_obj_за	1104	КС в роли объекта в конструкции с предлогом «за»	*уход за двигателем*

Окончание табл. 3

Идентификатор связи (отношения)	Кол-во связей	Характеристика отношения (для заданного ключевого слова, КС)	Пример
pp_obj_без	1 088	КС в роли объекта в конструкции с предлогом «без»	*планер без двигателя*
pp_перед	689	КС в роли субъекта в конструкции с предлогом «перед»	*двигатель перед пуском*
быть_adj	547	КС + *быть* + прилагательное	*двигатель был исправен*
pp_obj_внутри	534	КС в роли объекта в конструкции с предлогом «внутри»	*трение внутри двигателя*
pp_obj_до	523	КС в роли объекта в конструкции с предлогом предлог «до»	*от гайки до двигателя*
pp_путем	418	КС в роли субъекта в конструкции с предлогом «путем»	*пуск двигателя путем буксировки*
pp_obj_через	398	КС в роли объекта в конструкции с предлогом «через»	*протекать через двигатель*
pp_obj_перед	345	КС в роли объекта в конструкции с предлогом «перед»	*кабина перед двигателем*
pp_obj_благодаря	340	КС в роли объекта в конструкции с предлогом «благодаря»	*достигаться благодаря двигателю*

* В этом и в ряде других примеров КС не управляет предлогом, а лишь соседствует с ним. Для более точного анализа требуется хотя бы описание рамок валентностей, чего сегодня в шаблонах нет.

живают быть приведенными здесь (табл. 3; строки упорядочены по количеству связей).

Но вернемся к нашему *двигателю*. Просмотр полученных словосочетаний (word sketches) позволяет отобрать терминологические словосочетания для пополнения тезауруса. В процессе просмотра наполнения отдельных шаблонов выявилась интересная закономерность: некоторые синтаксические формулы (отношения) «добавляют» большое число терминологических словосочетаний, другие — наоборот.

К первым относятся сочетания «прилагательное + *двигатель*», отношение «и/или» (включая однородные члены через запятую), «существительное + *двигателя* (род. пад.)» (*мощность*, *охлаждение*, *оборот*, *вал*, *цилиндр*, *картер*, *прогрев*, *ремонт*, *вращение*, *блокировка*, *перегрев*, *тяга*). Вот далеко не полный список терминологических словосочетаний, выбранных с помощью инструмента «Лексические шаблоны» по модели a_modifier: ***двигатель*** *бензиновый*, *дизельный*, *шаговый*, *асинхронный*, *ракетный*, *реактивный*, *газотурбинный*, *N-цилиндровый*, *N-литровый*, *карбюраторный*, *поршневой*, *электрический*, *тяговый*, *двухтактный*, *паровой*, *турбореактивный*, *синхронный*, *маршевый*, *жидкостный*, *роторный*, *газовый*, *судовой*, *трехфазный*, *V-образный*, *приводной*, *форсированный*, *тепловой*, *четырехтактный*, *водородный*, *твердотопливный*, *инжекторный*, *подвесной*, *турбовинтовой*, *оппозитный*, *забойный*, *ионный*, *вертолетный*, *клапанный*, *прямоточный*, *линейный*, *ядерный*, *мотоциклетный*, *тормозной*, *гидравлический*, *ветряной*, *танковый*, *плазменный*, *лодочный*, *атомный*, *пороховой*, *гоночный*, *коллекторный*, *винтовой*, *гравитационный*, *пневматический*, *керосиновый*, *термоядерный*, *импульсный*, *водяной*, *фотонный*, *конденсаторный*, *радиальный*, *космический*, *роторно-поршневой*, *короткозамкнутый*, *шпиндельный*, *прыжковый*, *фазный*, *малогабаритный*, *ходовой*, *индукционный*, *нефтяной*, *роторно-лопастный*, *турбовальный*, *разгонный*, *морской*, *впрысковый*, *водометный*, *рулевой*, *турбовентиляторный*, *бесколлекторный*, *вентильный*, *корабельный*, *звездообразный*, *топливный*, *пружинный*, *инерционный*, *бесконтактный*, *бортовой*, *горизонтально-оппозитный*, *битурбированный*, *силовой*, *турбодизельный*, *реверсивный*, *биполярный*, *газопоршневой*, *кубовый*, *тепловозный*, *нижнеклапанный*, *низковольтный*, *сверхзвуковой*.

Но есть отношения и не столь продуктивные. Почти совсем нет терминов среди сочетаний «*двигатель* + существительное в родительном падеже». Они, как правило, содержат существительные, означающие разные механизмы, оснащенные двигателем (исключение — *двигатель внутреннего сгорания*). Совершенно очевидно, что далеко не все такие понятия должны быть включены в семантическое поле *двигатель*. В предложных сочетаниях pp_<предлог> и pp_obj_<предлог> безусловный приоритет надо отдать первым — тем, где *двигатель* выступает в роли субъекта словосочетания, или, если можно так сказать, в роли вышестоящего члена иерархического отношения (чаще всего это «целое—часть»). Например: ***двигатель*** *без турбонаддува*, *нейтрализатора*, *ключа*, *глушителя*, *прогрева*, *редуктора* или ***двигатель*** *с турбонаддувом*, *впрыском*, *ротором*, *наддувом*, *нагнетателем*, *воспламенением*, *турбокомпрессором*, *турбонагнетателем*, *охлаждением*, *зажиганием*, *смесеобразованием*, *распредвалом*, *карбюратором*, *впрыскиванием*, *трансмиссией*, *газораспределением*, *соплом* и т. д.

На следующем шаге мы использовали инструмент «Коллокации», применив его к словосочетаниям разного типа, выявленным с помощью языка регулярных выражений (в системе он называется CQL — corpus query language). В результате выяснилось, что большинство найденных терминов — кандидатов в коллокаты для новых, более длинных коллокаций — уже содержатся в нашем перечне. Это позволяет заключить, что в известном смысле наш тезаурус достиг степени насыщения (оговоримся, в пределах данного корпуса).

Заключение

Итак, было показано, что современная корпусная лингвистика располагает арсеналом средств, позволяющих автоматически на основе корпуса текстов формировать классы лексических единиц, находящихся в отношении онтологической зависимости. В частности, на примере термина *двигатель* была выявлена совокупность лексических единиц, образующих гнездо тезауруса, или лексико-семантическое поле для данного термина. Мы видим, что использование корпуса текстов и инструментов системы Sketch Engine позволяет выявлять в автоматизированном режиме синтагматические и пара-

дигматические связи и обеспечивает адекватное наполнение лексико-семантической системы.

Была продемонстрирована связь между объемом корпуса и качеством (точностью) формируемого тезауруса, а именно: чем больше объем корпуса, тем выше качество (точность) при заданном объеме дистрибутивного тезауруса.

Поскольку понятия в понятийной схеме выражаются не только отдельными словами, был задействован механизм выявления устойчивых словосочетаний (коллокаций и коллигаций), многие из которых являются терминами и были включены в тезаурус. Применение разных подходов к формированию тезауруса позволило достичь высокой степени его лексического насыщения.

Очевидно, что таким образом полученный тезаурус представляет собой входные данные для эксперта, задача которого заключается далее в описании отношений, связывающих эти слова с заглавным словом и между собой.

Исследование должно быть продолжено. Следует сравнить тезаурусы по одной и той же предметной области, формируемые для одних и тех же ключевых слов, но на корпусах разного типа: с одной стороны, на большом общеязыковом корпусе, с другой — на специализированном корпусе. Представляется, что именно последние дадут наилучший результат. Но первые имеют преимущество большого объема. Для подобных сравнений требуется разработка метрик, позволяющих получить численное значение точности и полноты тезауруса.

Программно-лингвистическое обеспечение корпусов также требует совершенствования. Необходимо повышать качество морфологической разметки, так как ошибки лемматизации ведут к тому, что словоформы одной и той же лексемы рассматриваются автономно, что искажает статистику.

Поскольку дистрибутивно-статистический анализ в системе Sketch Engine базируется на грамматике лексико-синтаксических шаблонов, то также важно заниматься совершенствованием этой грамматики.

Литература

Адмони В. Г. Синтаксис современного немецкого языка. Л., 1973.

Арапов М. В. Некоторые принципы построения словаря типа «тезаурус» // НТИ. Сер. 2. 1964. № 4. С. 40–46.

Ахманова О. С. Словарь лингвистических терминов. М., 1966.

Богданов В. В. Теоретические и практические аспекты тезаурусов // Инженерная лингвистика. Ч. 2. Л., 1971. С. 204–224.

Гринев С. В., Лейчик В. М. Некоторые аспекты тезаурусного представления знаний // НТИ. Сер. 2. 1993. № 10. С. 1–8.

Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981.

О некоторых лексико-семантических проблемах в «бестезаурусных» ИПС / В. Г. Войскунский [и др.] // Структурная и прикладная лингвистика: Межвуз. сб. Л., 1983. Вып. 2. С. 170–177.

Пиотровский Р. Г. Текст, машина, человек. Л., 1975.

Рубашкин В. Ш. Онтологическая семантика: Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М., 2012.

Сидорова Е. А. Подход к построению предметных словарей по корпусу текстов // Труды международной конференции «Корпусная лингвистика — 2008» (6–10 октября 2008 г., Санкт-Петербург). СПб., 2008. С. 365–372.

Уфимцева А. А. Теории «семантического поля» и возможности их применения при изучении словарного состава языка // Вопросы теории языка в современной зарубежной лингвистике. М., 1961. С. 30–63.

Хохлова М. В. Разработка грамматического модуля русского языка для специализированной системы обработки корпусных данных // Вестн. С.-Петерб. ун-та. Серия 9. Филология. Востоковедение. Журналистика. СПб., 2010. Вып. 2. С. 162–169.

Шайкевич А. Я. Дистрибутивно-статистический анализ в семантике // Принципы и методы семантических исследований. М., 1976. С. 353–378.

Шайкевич А. Я. Дистрибутивно-статистический анализ текстов: автореф. дис. … д-ра филол. наук. Л., 1982.

Шайкевич А. Я. Распределение слов в тексте и выделение семантических полей // Иностранные языки в высшей школе. М., 1963. Вып. 2.

Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А. Статистический словарь языка Достоевского. М., 2003.

Щур Г. С. Теория поля в лингвистике. М.-Л., 1974.

Kilgarriff A., Rychlý P. An Efficient Algorithm for Building a Distributional Thesaurus (and other Sketch Engine developments) // Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions — ACL 2007. P. 41–44.

Pazienza M., Pennacchiotti M., Zanzotto F. Terminology Extraction: an Analysis of Linguistic and Statistical Approaches // Knowledge Mining Series: Studies in Fuzziness and Soft Computing. Berlin, 2005. P. 255–279.

Pekar V. Linguistic Preprocessing for Distributional Classification of Words // Proceedings of the COLING-04 Workshop on Enhancing and Using Electronic Dictionaries. Geneva, 2004. P. 15–21.

Rychlý P. A Lexicographer-friendly Association Score // Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN. Brno, 2008. P. 6–9.

Sharoff S. Open-source corpora: Using the Net to Fish for Linguistic Data // International Journal of Corpus Linguistics. Amsterdam, 2006. Vol. 11, N 4. P. 435–462.

Smrž P., Rychlý P. Finding Semantically Related Words in Large Corpora // Text, Speech and Dialogue: Fourth International Conference (TSD-2001). LNCS (LNAI). Heidelberg, 2001. Vol. 2166. P. 108–115.

Syntactic-Based Methods for Measuring Word Similarity / P. Gamallo [et al.] // Text, Speech and Dialogue: Fourth International Conference TSD-2001. LNAI. Springer-Verlag, 2001. Vol. 2166 P. 116–125.

The Sketch Engine / A. Kilgarriff [et al.] // Proceedings of the XI EURALEX International Congress. Lorient, 2004. P. 105–116.

TTC: Terminology Extraction, Translation Tools and Comparable Corpora / H. Blancafort [et al.] // Proceedings of the XIV EURALEX International Congress. 2010. P. 263–268.

Т. И. Зубкова

ЯЗЫКОВОЕ СОЗНАНИЕ: НЕКОТОРЫЕ ОБЩИЕ ТЕНДЕНЦИИ

Аннотация. В статье обсуждается феномен языкового сознания под углом зрения психолингвистики. Рассматриваются различные факторы, влияющие на формирование этого процесса, такие как роль языковых клише или ментальных образцов. Обращается внимание на взаимоотношения языка и мышления.

Ключевые слова. Психолингвистика, язык, языковой, мышление, языковое сознание, универсальные тенденции, ментальный, клише, образец.

Tatiana I. Zoobkova

LINGUISTIC THINKING: SOME GENERAL TRENDS

Abstract. The paper discusses the phenomenon of linguistic thinking from a psycholinguistic point of view. Different factors involved in the problem are dealt with. These are the influence of linguistic clichés and mental patterns on the process, and some others. This problem is closely connected with the problem of interrelations between language and thought.

Keywords. Psycholinguistics, language, linguistic, thought, linguistic thinking, general trends, mental, cliché, pattern.

В отечественной лингвистике и психолингвистике термин «языковое сознание» встречается еще в 1960-е годы. В «Словаре лингвистических терминов» О. С. Ахмановой «сознание языковое» определяется как «особенности культуры и общественной жизни данного человеческого коллектива, определившие его психическое своеобразие и отразившиеся в специфических чертах данного языка» [Ахманова, с. 439]. Начиная с 1990-х годов в московской психолингвистической школе этот феномен интенсивно обсуждается и экспериментально

исследуется. Язык и культура — это «формы существования общественного сознания, которое бытует как „образ себя“ (образ своего этноса) и “образ другого“» [Уфимцева, с. 228]. Рассматривается национально-культурная специфика языкового сознания, а различие национальных сознаний коммуникантов признается одной из основных причин непонимания при межкультурном общении [Там же. С. 205]. Взаимосвязь мысли и слова, мышления и речи оказывается в центре психолингвистической теории [Там же, с. 242].

Язык структурирует картину мира как народа, так и отдельного человека, выявляет и личностное своеобразие человека, и специфическое содержание, наполняющее каждую культуру. Не вызывает сомнений, что при личностном или межкультурном общении наряду с возможностью непонимания нередко возникает ситуация взаимопонимания и единения в восприятии и интерпретации окружающего мира. Можно предположить, что в языковом сознании проявляются также и какие-то универсальные тенденции мышления, определяющие устойчивую и общую для всех картину мира.

Языковое сознание отражается и фиксируется в языковых эталонах и стереотипных словосочетаниях, анекдотах, пословицах и, конечно, в фольклоре, сказках. Сказки — это неотъемлемая часть культурного наследия любого народа. Они демонстрируют сходства и различия в восприятии и интерпретации окружающего мира разными этносами. Но особенно интересно сходство сказочного фольклора разных стран. В. Я. Пропп считал важнейшей и до сих пор не разрешенной проблемой именно сходство сказок по всему земному шару, в том числе и у народов, общение которых не может быть доказано исторически [Пропп, с. 21]. Представляется, что это подтверждает наличие общечеловеческих ментальных образцов и сюжетов, образующих единый образ мира.

В сказках многих народов есть сюжет о счастливом избавлении от злодеев, причем везет герою или героине, а другим персонажам, по-видимому, раньше не везло. Это истории про девушку-злодейку или старуху-ведьму, заманивающих мужчин, рассказы о женихе-разбойнике, который оказывается еще и людоедом, о злом животном, обернувшемся человеком с единственным желанием полакомиться своей невестой, в африканских сказках. Вот японский вариант сюжета: странствующий буддийский монах находит приют у приветливой старушки, она оставляет его ненадолго, он случайно

обнаруживает останки съеденных людей и в ужасе бежит, а ведьма — за ним, бежит и кричит: догадался, негодник! Сопоставим, казалось бы, несопоставимое и обратим внимание на поразительное совпадение линий развития сюжетов, что, очевидно, свидетельствует о существовании каких-то общечеловеческих мыслительных схем. Ленинград конца 1940-х годов. Дети много времени проводят во дворах, играют в игры, в которые теперь никто больше не играет, и рассказывают истории, услышанные от взрослых, а часто — сочиненные самими детьми по имеющимся образцам. Самые страшные из этих рассказов связаны с еще живыми воспоминаниями о блокаде — это рассказы о случаях людоедства. Эти истории невозможно забыть, их помнишь всю жизнь, вот одна из них: девочку заманивает к себе домой приветливая женщина и оставляет одну на некоторое время. Девочке становится страшно, она выскакивает в коридор, видит открытую на лестницу дверь и бежит, а женщина — за ней и кричит: догадалась!

Создается впечатление, что мы постоянно выстраиваем картину окружающей действительности, используя устойчивые языковые эталоны, включенные в языковое сознание. Например, еще один почти сказочный персонаж — добрый человек, волшебник, своего рода вариант исполняющей желания золотой рыбки. Исторически близкое время, конец XX века, продукты покупаются по талонам, которые привычно называют карточками. Язык сразу реагирует, появляется глагол «отоваривать». Женщины подолгу стоят в очередях, чтобы «отоварить» талоны, и рассказывают друг другу истории, как правило, со счастливым концом: одна женщина потеряла все карточки, она так плакала-плакала, но ее карточки нашел добрый человек и объявил по радио, ей не только карточки вернули, но и отоварили все без очереди. Если вслушаться в наши рассказы о событиях своей жизни, можно выявить привычные сюжетные линии, которые как бы сами выстраивают повествование и могут даже исказить содержание, чтобы подогнать его под имеющийся образец.

С помощью слова язык фиксирует в общественном и индивидуальном сознании события и обстоятельства жизни как языковой общности, так и отдельного человека. Языковое сознание изменчиво, наряду с общими, присущими ему тенденциями в языковом сознании отражаются постоянно происходящие изменения общественных настроений, корректируется толкование существующих

в обществе отношений, и, соответственно, возникают новые клише и языковые стереотипы. Целая эпоха стояла за словосочетанием «воинствующий атеизм», и атеизм был действительно воинствующим. Грозный смысл этого определения утрачивается, смягчается с изменением общественных отношений и общественного сознания, «атеизм» приобретает новый оттенок значения, связанный с общественным просвещением. Пропагандисты разъясняют, сколь ошибочно употребление предлога «для» в привычной фразе «Религия — это опиум для народа»: классики марксизма имели в виду, что это опиум народа, народ сам с помощью религии одурманивает себя. Меняются времена, и языковое сознание выстраивает иную картину мира с помощью новых языковых клише: вместо «воинствующих атеистов» появляются «православные активисты», они готовы вести борьбу с теми, кто — еще одно новое клише — «оскорбляет чувства верующих». Утверждение «Я православный коммунист» почти не вызывает удивления, не кажется бессмысленным.

В языковом сознании современных людей, особенно молодых, немалую роль играют запоминающиеся словосочетания на злободневную тему, например интернет-мемы. Воздействие слова дополняется и усиливается с помощью разного рода иллюстраций, создается определенный образ, одна или несколько картинок, юмористически интерпретирующих какую-то жизненную или политическую ситуацию.

Понятие языкового сознания непросто четко отграничить от понятия «языковая картина мира» или «образ мира». Но «языковое», «слово», «язык» — это, несомненно, своего рода центр, фокус, позволяющий выявить некоторый фрагмент действительности, часть вербального и невербального опыта, а вторая составляющая термина, «сознание», позволяет говорить об определенных общих тенденциях мышления, о закономерности существования того, что все-таки удачно называется «языковым сознанием».

Литература

Ахманова О. С. Словарь лингвистических терминов. М., 1966.
Пропп В. Я. Морфология сказки. М., 1969.
Уфимцева Н. В. Языковое сознание: динамика и вариативность. М., 2011.

О. А. Митрофанова

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КОРПУСА «НАРОДНЫХ РУССКИХ СКАЗОК А. Н. АФАНАСЬЕВА»[1]

Аннотация. В докладе рассматриваются результаты построения тематической модели для корпуса художественных текстов, а именно корпуса текстов русских народных сказок из собрания А. Н. Афанасьева. Тематическое моделирование осуществляется на основе алгоритма LDA (Latent Dirichlet Allocation). Автоматический анализ корпуса текстов проводится с помощью компьютерного инструмента gensim. Результаты экспериментов позволяют выявить особенности семантической организации исследуемого корпуса, охарактеризовать структуру автоматически сформированных тем и их качественный состав.

Ключевые слова. Тематическое моделирование, LDA, русскоязычные корпусы текстов.

Olga A. Mitrofanova

TOPIC MODELING OF „A. N. AFANASJEV'S RUSSIAN FAIRYTALES" CORPUS

Abstract. The paper presents results of topic modeling performed for a belletristic text corpus, namely, the corpus of Russian fairytales from A. N. Afanasjev's collection. Topic modeling is performed with the help of LDA (Latent Dirichlet Allocation) algorithm. Automatic processing of the corpus was carried out by means of gensim toolkit. Experimental results allow to figure out specific features in semantic structure of the corpus, to characterize the structure and content of automatically extracted topics.

Keywords. Topic modeling, LDA, Russian text corpora.

[1] Исследование выполнено при частичной финансовой поддержке гранта СПбГУ 30.38.305.2014 «Квантитативные лингвистические параметры определения стилевых характеристик и предметной области текстов».

1. Художественный текст как объект квантитативного анализа

Тексты художественной литературы являют собой богатый материал как для филологического, так и для структурно-лингвистического анализа. Классическими трудами в этой сфере стали исследования по формализации структуры сказочного сюжета [Пропп]. Анализ квантитативных параметров текста на уровне словаря, морфологии и синтаксиса успешно используется в стилистической диагностике [Мартыненко; Тулдава], в проверке атрибуционных гипотез при установлении авторства текста [Марусенко]. Статистические методы стали неотъемлемой частью работ в сфере авторской лексикографии [Поцепня; Шайкевич, Андрющенко, Ребецкая]. Методы компрессии, основывающиеся на синтаксических правилах и машинном обучении находят применение в издательском деле при автоматическом составлении рефератов и аннотаций для художественных текстов [Kazantseva, Szpakowicz].

В реестре формальных инструментов, доступных современным лингвистам, появились тематические модели — специфический класс статистических моделей, которые позволяют использовать количественные данные для исследования семантической организации текстов разных стилей и жанров. Целью нашего исследования является оценка продуктивности тематического моделирования в автоматической обработке текстов особого типа — русских народных сказок. Материал русскоязычных текстовых коллекций, в том числе представляющих художественный стиль, мало изучен с позиций тематического моделирования. Наше исследование направлено на частичное восполнение существующего пробела.

2. Тематическое моделирование как исследовательская задача

Построение тематической модели корпуса текстов — это вероятностный процесс, который предполагает установление связей между множеством документов, совокупностью слов и набором тем, отражающих содержание документов. Задача тематической модели заключается в выявлении скрытых факторов, определяющих внутреннюю организацию исследуемых текстов. Тематическое моделирование представляет собой разновидность нечеткой кластеризации, поскольку в результирующей модели слова или документы с той или иной вероятностью могут относиться к разным темам.

Для представления документов в тематической модели текстовой коллекции используется смесь вероятностных распределений на множестве тем. Обобщенная тематическая модель имеет следующий вид: $p(w|d) = \Sigma p(t|d)p(w|t)$, где $p(w|d)$ — известная частота появления слова w в документе d, $p(w|t)$ — неизвестная вероятность появления слова w в теме t, $p(t|d)$ — неизвестная вероятность появления темы t в документе d. Предполагается, что документ состоит из слов, которые выбраны случайным образом из смеси распределений, и в таком случае цель моделирования заключается в восстановлении компонент смеси. Это значит, что в тематической модели корпуса документов D должны быть определены множество тем T, распределения $p(w|t)$ для всех тем и $p(t|d)$ для всех документов.

Существует целый ряд тематических моделей различных типов: алгебраические (например, стандартная векторная модель (Vector Space Model, VSM), латентно-семантический анализ (Latent Semantic Analysis, LSA)), вероятностные (вероятностный латентно-семантический анализ (Probabilistic Latent Semantic Analysis, pLSA), латентное размещение Дирихле (Latent Dirichlet Allocation, LDA)). Наиболее перспективным в обработке текстовых коллекций признан алгоритм LDA. Доступна богатая библиография по тематическому моделированию англоязычных корпусов текстов, см. например: [Blei, Ng, Jordan; Knowledge Discovery …].

В построении тематических моделей находят применение одновременно и математический аппарат инженерии знаний, и основные теории структурной лингвистики. Темы, порождаемые в ходе тематического моделирования, — это по существу кластеры слов, для которых характерна семантическая близость в корпусе текстов. Данное наблюдение согласуется с дистрибутивной гипотезой о сходстве значений у совместно встречающихся слов, с практическими результатами дистрибутивно-статистического анализа, а также с исследованиями смысловой структуры текста и описания тематической сетки текста на основе повторяющихся в нем значений.

В практическом отношении результаты тематического моделирования представляют наибольший интерес применительно к автоматической обработке публицистических текстов (корпусов новостных сообщений, текстов социальных сетей и т. д.), а также научных текстов (текстов по биоинформатике и т. д.). Результат моделирования в этих случаях применим для оптимизации информационного поиска по

текстовым массивам, для мониторинга общественного мнения и т. п. [Карпович; Interval Semi-Supervised LDA … ; Vorontsov, Potapenko].

В зарубежной литературе описан опыт построения тематической модели LDA для анализа корпуса текстов английской поэзии, включающего около 4500 стихотворных произведений [Rhody]. Автор указывает на то, что темы могут быть интерпретированы как с точки зрения сюжета поэтических текстов, так и с позиций лексических группировок, характерных для языка поэзии (например, ряды эпитетов). Ранее с целью стилистической диагностики и характеристики сюжета мы проводили эксперименты по анализу русской прозы, где применялся более простой алгоритм нечеткой кластеризации [Mitrofanova]. Была также проведена работа с англоязычными художественными текстами, в ходе которой использовался инструмент TMT[2] [Митрофанова].

3. Лингвистические данные, их подготовка и предобработка

В экспериментах по тематическому моделированию задействован один из специализированных корпусных ресурсов, которыми располагает кафедра математической лингвистики СПбГУ, а именно корпус «Народных русских сказок А. Н. Афанасьева». Тексты корпуса соответствуют изданию [Народные сказки …]. Объем текстовой коллекции (за вычетом предисловия и комментариев) составляет порядка 500 тысяч словоупотреблений (с/у).

Тематическое моделирование требует, чтобы предварительно была реализована процедура предобработки корпуса, предполагающая следующие действия:

1) удаляются нетекстовые элементы;

2) создается стоп-словарь, включающий обороты, леммы служебных частей речи, местоимений, числительных, аббревиатур, списки латинских символов, знаков пунктуации и цифровых обозначений (при этом использовались словарные ресурсы НКРЯ[3]);

3) проводятся лемматизация и автоматическое разрешение морфологической неоднозначности с применением морфоанализатора mystem 3.0[4];

[2] URL: https://code.google.com/p/topic-modeling-tool/.

[3] URL: http://www.ruscorpora.ru/obgrams.html, http://dict.ruslang.ru/freq.php.

[4] URL: https://tech.yandex.ru/mystem/.

4) корпус разбивается на документы сообразно их первоначальной логической структуре (в нашем случае документ приравнивается к тексту отдельной сказки).

4. Компьютерная реализация LDA

На сегодняшний день существует ряд самостоятельных программ и компьютерных инструментов, ориентированных на статистическую обработку данных, машинное обучение, классификацию и кластеризацию, в составе которых предусмотрен блок тематического моделирования[5]. В нашем исследовании задача тематического моделирования решается с помощью gensim[6] — набора библиотек Python, позволяющих строить алгебраические и вероятностные модели текстовых корпусов. Gensim располагает разнообразными средствами количественной обработки текстовых данных, среди них построение матриц совместной встречаемости, вычисление коэффициента tf-idf, алгоритм word2vec, оценка близости контекстных векторов, построение модели LSA и т. п. В составе gensim имеются средства построения вероятностных тематических моделей корпусов текстов на основе LDA. Неоспоримым преимуществом gensim является простота его настройки для решения задач по автоматическому анализу русскоязычных текстов. Итак, построение тематической модели LDA осуществляется с помощью скрипта, составленного на основе компонентов gensim. Проводятся следующие процедуры анализа обрабатываемого корпуса текстов:

1) формируется словарь, который представляет собой список лемм с указанием порядковых номеров и частот в текстах;

2) удаляются стоп-слова, считываемые из пользовательского стоп-словаря, а также низкочастотные слова;

3) словарь преобразуется в матрицу;

4) реализуется модель LDA с разными параметрами, при этом пользователь самостоятельно задает число итераций, число тем, объем тем;

5) появляются результаты обработки: темы приводятся в виде списков лемм, также автоматически рассчитывается ряд количественных параметров тем.

---

[5] См., например: URL: http://www.cs.princeton.edu/~blei/topicmodeling.html.

[6] URL: http://radimrehurek.com/gensim/.

Поправимым недостатком модели и ее реализации является то, что метки темы не генерируются автоматически, а выбираются пользователем вручную.

5. Результаты экспериментов и их интерпретация

Был проведен анализ текстов из корпуса «Народных русских сказок А. Н. Афанасьева» с точки зрения их тематики. Эксперименты проводились со следующими параметрами: число итераций — 10; число тем, ожидаемых в выдаче, — 20; видимый пользователю объем тем в выдаче — 10 лемм.

Прежде чем перейти к обсуждению результатов, сделаем предварительные замечания об особенностях работы тематической модели и о том, как с ее помощью производится структурирование лингвистического материала. Используемая нами тематическая модель имеет вероятностный характер, в ней распределение лексики и документов по темам соответствует принципам нечеткой кластеризации (тема есть нечеткое множество лемм, которое описывает нечеткое множество текстов). Это может проявляться в том, что лексическое наполнение темы не всегда полностью соответствует метке, назначаемой вручную и обобщающей значения основной части лемм из темы: например, в теме, связанной с обозначениями сказочных артефактов, могут встретиться не только имена артефактов, но и другая лексика (см. ниже: артефакт *блюдечко* и не-артефакт *котенок*). В темах объединена лексика с общими дистрибутивными свойствами, это слова, которые встречаются в сходном контекстном окружении и имеют близкие или смежные (не обязательно совпадающие) значения. Состав тем исследовался на уровне ядра, а именно верхней части списка, границы которой задаются вручную (в нашем случае наиболее информативными являются первые 10 лемм). Связи между леммами в темах могут быть как парадигматическими (*старуха — баба*, *старик — дед* и т. д.), так и синтагматическими, например: *петух*, *лиса… — лес* (*Понесла меня лиса, понесла петуха за темные леса*) и т. д. В теме может присутствовать лексика из конкретного текста и в то же время — леммы из разных текстов с близким содержанием.

В ходе обработки полученных данных были выявлены два основных типа тем:

1) темы, связанные с общей структурой сказочного сюжета, например:

а) названия героев сказок (*поп*, *работник*, *казак*, *попадья*, *мужик*, *мертвец*, *лошадь*, *парень*, *батюшка* и т. д.),

б) названия животных, участвующих в сказочных ситуациях (*петух*, *лиса*, *коза*, *кот*, *лес*, *окошко* и т. д.),

в) обозначения сказочных артефактов (*короб*, *блюдечко*, *колодезь*, *сажа*, *заря*, *улица*, *котенок* и т. д.),

г) имена действий сказочных персонажей (*говорить*, *давать*, *сказать*, *взять*, *пойти*, *лететь*, *садиться* и т. д.);

2) темы, отражающие конкретные сказочные сюжеты, например сюжет сказок:

а) «Сестрица Аленушка, братец Иванушка» (*Аленушка*, *кипучий*, *пастись*, *подружка*, *сеть*, *пруд*, *оборачивать*, *зоря, наводить*, *пускать и* т. д.),

б) «Перышко Финиста ясна сокола» (*девица*, *красный*, *Василиса*, *дочь*, *избушка*, *Сокол*, *Ясный*, *сестра*, *море*, *золотой* и т. д.),

в) «Лисичка-сестричка и волк» (*волк*, *говорить*, *старик*, *лиса*, *медведь*, *мужик*, *пойти*, *старуха*, *приходить*, *давать* и т. д.),

а также сюжеты групп сказок о каком-либо персонаже, например:

г) об Иване-царевиче (*царевич*, *Иван*, *конь*, *богатырь*, *прекрасный*, *Баба-Яга*, *брат*, *сила*, *Елена*, *поехать* и т. д.),

д) об Алеше Поповиче (*князь*, *сила*, *Алеша*, *конь*, *Попович*, *поле*, *могучий*, *русский*, *добрый*, *посол* и т. д.).

Тем самым внутри тем, характеризующих корпус «Народных русских сказок А. Н. Афанасьева», присутствует наиболее значимая лексика, которая отражает сюжетные линии текстов. Это наблюдение подкрепляется тем фактом, что содержательная структура сказки подчинена строгому канону и может быть описана с точки зрения логического сценария [Мартемьянов], машиночитаемого алгоритма [Гаазе-Рапопорт, Поспелов, Семенова] или сюжетной грамматики [Олкер]. Однако выделенные нами темы представляют собой более крупные фрагменты содержания по сравнению с конструктивными единицами указанных выше представлений.

Важной особенностью построенной нами тематической модели является смешанность состава тем с точки зрения частеречной принадлежности лемм. Зачастую в темах соседствует именная и глагольная лексика, описывающая состояние действующих лиц, отдельные

сцены и ситуации (см. темы второго типа). Мы высказываем предположение о том, что это стилистически детерминированная черта, проявляющаяся при автоматической обработке художественных текстов.

6. Основные итоги и перспективы продолжения исследования

Проведенные нами эксперименты по тематическому моделированию корпуса «Народных русских сказок А. Н. Афанасьева» подтверждают применимость данного метода в исследовании сказок как разновидности художественных текстов.

Направления дальнейшего развития исследования связаны с улучшением качества предобработки текстов, оценкой оптимальных параметров построения тематических моделей, расширением текстовой коллекции, применением результатов тематического моделирования в стилистической диагностике.

Литература

Гаазе-Рапопорт М. Г., Поспелов Д. А., Семенова Е. Т. Порождение структур волшебных сказок. М., 1980.

Карпович С. Н. Русскоязычный корпус текстов СКТМ-ру для построения тематических моделей // Труды международной конференции «Корпусная лингвистика — 2015». СПб., 2015.

Мартемьянов Ю. С. Заметки о строении ситуации и форме ее описания // Логика ситуаций. Строение текста. Терминологичность слов. М., 2004.

Мартыненко Г. Я. Основы стилеметрии. Л., 1988.

Марусенко М. А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов. Л., 1990.

Митрофанова О. А. Тематическое моделирование художественного текста на основе алгоритма LDA // Материалы XLIII международной филологической конференции. Секция прикладной и математической лингвистики. СПб., 2014.

Народные русские сказки А. Н. Афанасьева: в 3 т. М., 1957–1958.

Олкер Х. Р. Волшебные сказки, трагедии и способы изложения мировой истории // Язык и моделирование социального взаимодействия. М., 1987.

Поцепня Д. М. Образ мира в слове писателя. СПб., 1997.

Пропп В. Я. Морфология «волшебной» сказки. М., 1998.

Тулдава Ю. А. Проблемы и методы квантитативно-системного исследования лексики. Таллинн, 1987.

Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А. Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. М., 2013. Т. 1.

Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research 3 (4–5), January 2003.

Interval Semi-Supervised LDA: Classifying Needles in a Haystack / S. Bodrunova [et al.] // MICAI–2013.

Kazantseva A., Szpakowicz S. Summarizing Short Stories // Computational Linguistics. 2010. Vol. 36, No. 1.

Knowledge Discovery through Directed Probabilistic Topic Models: a Survey / A. Daud [et al.] // Proceedings of Frontiers of Computer Science in China. 2010.

Mitrofanova O. Automatic Word Clustering in Studying Semantic Structure of Texts // Advances in Computational Linguistics: Research in Computing Science. Mexico, 2009. Vol. 41.

Rhody L. M. Topic Modeling and Figurative Language // Journal of Digital Humanities. 2012. Vol. 2, N 1.

Vorontsov K. V., Potapenko A. Additive Regularization of Topic Models // Machine Learning. 2014.

Д. М. Миронова

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА В ТЕКСТОЛОГИИ

Аннотация. В статье рассматриваются методы компьютерной текстологии, применяемые при классификации рукописей с контролируемой традицией. Проводятся анализ и сравнение методов, использующих кластерный анализ, применительно к славянской новозаветной традиции.

Ключевые слова. Кластерный анализ, классификация рукописей, контаминация, контролируемая традиция, Славянский Новый Завет, стемматология, узел разночтений.

Dina M. Mironova

CLUSTER ANALYSIS IN TEXTUAL CRITICISM

Abstract. The advance of computer technologies has made it possible to minimize manual work and calculations where pure computation is concerned, thereby providing the scholar with data for further research. Different methods of computer-assisted classification have been applied to the classification of ancient manuscripts depending on the type of the tradition. The paper deals with the variations of cluster analyses applied to New Testament Slavonic Gospels and analyses their efficiency and perspectives.

Keywords. Cluster analysis, manuscript classification, contamination, controlled tradition, Church Slavonic Gospels, stemmatology, variation unit.

Подход к выбору метода для текстологического исследования рукописной традиции определяется типом традиции. Для памятников, количество списков которых не превышает нескольких десятков экземпляров, широко используются методы кладистики и стемматологии, целью которых является представление истории текста в виде дерева отношений между рукописями. Однако для библейской

традиции, насчитывающей несколько тысяч свидетелей и характеризующейся контаминацией, невозможно построить дерево отношений между отдельными рукописями, и необходимо искать метод, позволяющий выявить группы наиболее близких между собою рукописей, и анализировать отношения уже между этими группами.

В 1985 году профессор СПбГУ А. А. Алексеев предложил план исследования Славянского Евангелия, включавший использование кластерного анализа для возможности обработки неограниченного количества рукописей контролируемой традиции [Алексеев, 1985]. Проект был успешно реализован, и автоматизация классификации рукописей при помощи кластерного анализа значительно оптимизировала обработку источников [Евангелие…, 1998; Евангелие…, 2005].

Впервые метод кластерного анализа был применен американскими текстологами Э. Колвеллом и Э. Тьюном для работы с греческой новозаветной традицией в 1960-е годы [Colwell], а к концу XX века возможности вычислительной техники позволили широко и эффективно применять данный метод в текстологии. Вариации кластерного анализа активно используются в современных текстологических исследованиях как для поиска групп рукописей с близким текстом, так и при построении стеммы, или дерева отношений между рукописями. При этом различные исследователи пользуются различными коэффициентами сходства или расстояния для классификации рукописей в зависимости от особенностей исследуемой традиции [Азарова, Алексеева; Алексеев, 1986; Миронова, 2005; Миронова, 2012; Пичхадзе; Galloway; Meyer; Mironova; Mulken; Nikolaenko; O'Hara, Robinson; Salemans; Wattel, Mulken; Wattel].

В 1970-е годы, немного позже выхода в свет работы Э. Колвелла и Э. Тьюна, голландский текстолог А. Деес разработал метод для построения дерева отношений между рукописями [Dees, 1975; Dees, 1976]. В отличие от предшественников — приверженцев представления рукописной традиции в виде ориентированного дерева — он не ставил задачи обязательного построения ориентированного дерева и обнаружения архетипа. Он предложил начать с построения неориентированного дерева, в котором все конечные узлы — реальные рукописи, а промежуточные узлы — гипотетические рукописи. В 1990-е годы голландский ученый-математик Э. Ваттель в сотрудничестве с текстологами развил метод А. Дееса, формально описал его, сделал ряд доработок и усовершенствований [Wattel, Mulken; Wattel], так-

же положив в основу кластерный анализ. Метод Ваттеля позволяет строить стеммы для любой рукописной традиции, как библейской, так и авторской, с любым количеством рукописей. Формально программа строит дерево не более чем для 250 рукописей, но практически идентичные рукописи можно представить как одну, что дает возможность построить дерево для гораздо большего числа источников.

Рассмотрим возможности и результаты применения кластерного анализа для классификации 525 рукописей Славянского Евангелия от Матфея при помощи методов Алексеева и Ваттеля.

Коэффициент сходства Алексеева выражается в процентах и вычисляется следующим образом: $K = F / V \cdot 100\%$, где K — коэффициент сходства, F — количество общих чтений для данной пары, V — количество узлов разночтений, по которым сравнивается эта пара [Алексеев, Кузнецова]. Простота и прозрачность вычислений уменьшают субъективное влияние исследователя на результат. Ваттель вводит коэффициент расстояния для пары рукописей: $K = N_p / (S_p + N_p)$, где K — коэффициент расстояния, N_p — количество узлов разночтений, где чтения рукописей пары различны, а S_p — количество узлов разночтений, где чтения рукописей пары совпадают. Алгоритм кластеризации включает вычисление коэффициента сходства или расстояния между всеми парами рукописей, построение исходной матрицы, где для каждой пары рукописей указан коэффициент сходства или расстояния, и формирование кластеров рукописей, начиная с пары с самым высоким коэффициентом сходства или самым низким коэффициентом расстояния. Процесс кластеризации заканчивается, когда все рукописи объединятся в один кластер.

Каждый метод предлагает свой способ визуализации итоговых кластеров. В методе Алексеева результатом является перестроенная матрица, в которой рукописи располагаются в том порядке, в каком они объединялись в кластеры. Двигаясь вниз по главной диагонали матрицы (рис. 1), мы обнаруживаем границы элементарных кластеров: внутри кластера процент сходства вдоль главной диагонали убывает, увеличение процента сходства указывает на начало нового кластера. Элементарные кластеры следуют в матрице в том порядке, в каком они объединялись в более крупные кластеры в процессе кластеризации.

В методе Ваттеля по результатам исходной матрицы строится неориентированное дерево, или стемма, где все рукописи являются

			1	2	3	4	5	6	7	8	9	10
V												
1:	B	:	0	98	88	79	77	78	72	63	64	68
2:	OB	:	98	0	88	80	78	79	73	64	65	69
3:	A	:	88	88	0	79	80	77	70	62	61	65
4:	Gl	:	79	80	79	0	91	89	81	72	68	81
5:	Me	:	77	78	80	91	0	88	81	70	65	76
6:	Tp	:	78	79	77	89	88	0	88	71	65	78
7:	Vls	:	72	73	70	81	81	88	0	69	66	74
8:	Trs	:	63	64	62	72	70	71	69	0	89	87
9:	Mss	:	64	65	61	68	65	65	66	89	0	83
10:	Jus	:	68	69	65	81	76	78	74	87	83	0

Рис. 1. Результат кластерного анализа (фрагмент итоговой матрицы). B, OB, A, Gl, Me, Tp, Vls, Trs, Mss, Jus — обозначения (сиглы) рукописей. Полученные кластеры: B, OB, A; Gl, Me, Tp, Vls; Trs, Mss, Jus

конечными узлами. Построение дерева начинается с пары с самым низким коэффициентом расстояния. Такая пара представляется на дереве как два конечных узла, соединенных с общим промежуточным узлом. Пара со следующим наименьшим значением коэффициента расстояния либо примыкает к данному промежуточному узлу, либо образует два конечных узла, восходящих к новому промежуточному узлу. После того как все рукописи окажутся представлены на дереве в виде конечных узлов, алгоритм кластеризации завершается соединением всех ветвей дерева в одном гипотетическом узле. Такое дерево не является ориентированным, и его можно интерпретировать как набор кластеров. Если рассматривать дерево как способ представления X групп (одна ветвь — одна группа), то дерево с X ветвями соответствует X кластерам Алексеева. Таким образом, метод Ваттеля может успешно использоваться для контаминированной новозаветной традиции.

Таблице кластеров, полученной по методу Алексеева, может соответствовать стемма Ваттеля, приведенная на рис. 2.

Результаты применения методов Алексеева и Ваттеля к рукописной традиции Славянского Евангелия показали, что оба метода дают

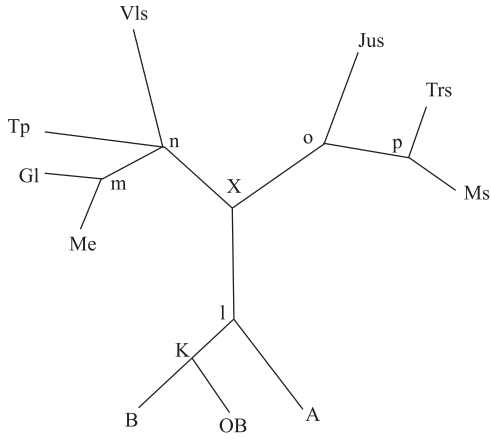

Рис. 2. Стемма Ваттеля, построенная с применением кластерного анализа. B, OB, A, Gl, Me, Tp, Vls, Trs, Mss, Jus — обозначения (сиглы) рукописей, k, l, m, n, o, p, X — гипотетические промежуточные узлы

одинаковое разбиение на кластеры, наблюдаются лишь небольшие вариации на периферии кластеров. Однако стоит отметить, что построение дерева связано также с высокими требованиями к памяти и скорости работы компьютера. Кроме этого, в матрице Алексеева показан процент сходства каждой пары рукописей, что делает ее более информативной, в силу чего для работы со славянской библейской рукописной традицией предпочтительнее оказывается метод Алексеева, позволяющий быстро и эффективно выявить редакции и определить их приметы [Алексеев, 1999, с. 54–62].

Литература

Азарова И. В., Алексеева Е. Л. От критического издания к структурированному корпусу славянских вариантов Евангелия // Труды международной конференции «Корпусная лингвистика — 2015» / отв. ред. В. П. Захаров, О. А. Митрофанова, М. В. Хохлова. СПб., 2015. С. 89–92.

Алексеев А. А. Опыт текстологического анализа славянского Евангелия // Старобългаристика. 1986. № 3. С. 8–19.

Алексеев А. А. Проект текстологического исследования кирилло-мефодиевского перевода Евангелия // Советское славяноведение. М., 1985. № 1. С. 82–94.

Алексеев А. А. Текстология славянской Библии. СПб., 1999.

Алексеев А. А., Кузнецова Е. Л. ЭВМ и проблемы текстологии древнеславянских текстов // Лингвистические задачи и обработка данных на ЭВМ. М., 1987. С. 111–121.

Евангелие от Иоанна в славянской традиции / под ред. А. А. Алексеева. СПб., 1998.

Евангелие от Матфея в славянской традиции / под ред. А. А. Алексеева. СПб., 2005.

Миронова Д. М. Классификация славянских рукописей Евангелия от Матфея // Евангелие от Матфея в славянской традиции / под ред. А. А. Алексеева. СПб., 2005. С. 163–168.

Миронова Д. М. Оценка текстологической значимости узлов разночтений // Структурная и прикладная лингвистика: Межвуз. сб. СПб., 2012. Вып. 9. С. 261–275

Пичхадзе А. А. Применение статистики при текстологическом исследовании памятника с контролируемой традицией (Древнеславянский Паримейник) // Лингвистические задачи и обработка данных на ЭВМ. М., 1987. С. 121–140.

Colwell E. C. Studies in Textual Criticism of the New Testament. Leiden, 1969.

Dees A. Considerations theorique sur la tradition manuscrite du Lai de l'ombre // Neophilologus. 1976. Vol. 60. P. 481–504.

Dees A. Sur une constellation de quatre manuscrits // Melanges de linguistique et de litterature offertes a Lein Geschiere. Amsterdam, 1975. P. 1–9.

Galloway P. Clustering Variants in the Lai de l'Ombre Manuscripts: Techniques and Principles // ALLC Journal. 1982. Vol. 3, N 1. P. 1–7.

Meyer E. Schone Historie und Ewangelien. Amsterdam, 2013.

Mironova D. Old Church Slavonic New Testament. The Problems of Automatic Classification // Studies in Stemmatology II. / eds P. van Reenen, M. van Mulken. Amsterdam, 2004. P. 241–268.

Mulken M. van. The Manuscript Tradition of the Perceval of Chretien de Troyes. A Stemmatological and Dialectological Approach. VU Amsterdam, 1993.

Nikolaenko D. Old Church Slavonic Versions of the Gospels. Computer-Aided Classification and the Choice of Variants // The Bible from Alpha to Byte. Proceedings of the 5th International Conference AIBI. Stellenbosch University. South Africa, 2000. P. 475–494.

O'Hara R., Robinson P. Computer-Assisted Methods of Stemmatic Analysis // The Canterbury Tales Project. Occasional Papers. Oxford, 1993. P. 53–74.

Salemans B. Building Stemmas with the Computer in a Cladistic, Neo-Lachamnian Way. Nijmegen, 2000.

Wattel E., Mulken M. van. Weighted Formal Support of a Pedigree // Studies in Stemmatology I. / eds P. van Reenen, M. van Mulken. Amsterdam, 1996. P. 135–167.

Wattel E. Constructing Initial Binary Trees in Stemmatology // Studies in Stemmatology II / eds P. van Reenen, M. van Mulken. Amsterdam, 2004. P. 145–166.

И. С. Николаев

ГЕОГРАФИЧЕСКАЯ ТЕРМИНОЛОГИЯ В БАЗЕ ДАННЫХ ПО ТОПОНИМИИ ИНГЕРМАНЛАНДИИ

Аннотация. В статье анализируются географические термины в базе данных по топонимии Ингерманландии (Ленинградская область). Рассмотрены основные типы терминов. Обсуждаются принципы использования местных географических терминов в топонимическом словаре Ингерманландии, который создается на кафедре математической лингвистики СПбГУ.

Ключевые слова. Топонимика, географические термины, топонимическая база данных, топонимический словарь, Ленинградская область, Ингерманландия.

Ilya S. Nikolaev

GEOGRAPHIC TERMINOLOGY IN THE TOPONYMIC DATABASE OF INGERMANLAND

Abstract. Geographical terms in the database of toponymy of Ingermanland (Leningrad region) are analyzed in the article. Main types of geographical terms are considered. We discuss general principles of inclusion of geographical terms into the toponymic vocabulary of Ingermanland, which is being created at the Department of Mathematical Linguistics of St. Petersburg State University.

Keywords. Toponymy, geographical terms, database of toponymy, toponymic vocabulary, Leningrad region, Ingermanland.

База данных по топонимии Ингерманландии, созданная коллективом исследователей на кафедре математической лингвистики СПбГУ, представляет собой оцифрованный вариант «бумажной» картотеки топонимов, собранной в ходе ежегодных экспедиций студентами и преподавателями кафедры в течение последних 35 лет [Николаев, Азарова, Герд; Азарова, Герд, Николаев]. Топонимии-

ческая база данных находится под управлением СУБД (системы управления базами данных) и позволяет проводить научные исследования, связанные с изучением историко-культурного ландшафта западной части Ленинградской области [Автоматизированная база данных … ; Типовые запросы в базе данных …]. На основе этой базы данных был создан научно-образовательный веб-ресурс, который предназначен для студентов отделения прикладной лингвистики СПбГУ и для широкого круга исследователей и краеведов [Научно-образовательный веб-ресурс … ; Nikolaev, Stolyarov].

В настоящее время идет работа по составлению топонимического словаря Ингерманландии по материалам базы данных и картотеки. В ходе этой работы были проанализированы географические термины, указывающие на тип объектов, поименованных топонимами и микротопонимами. В базе данных содержится 170 географических терминов. При создании словаря топонимов необходимо принять решение о том, какие термины могут быть использованы, а от каких следует отказаться. Обсуждению этого вопроса и посвящена эта статья.

Географическая терминология Ленинградской области изучена недостаточно, как и топонимия и микротопонимия в целом. Работ, посвященных региональной топонимии, очень немного. Небольшая научно-популярная книга «Следы времен минувших» [Попов] является одной из первых попыток описать топонимию Ленинградской области и единственной работой, охватывающей весь регион. Работы С. Кепсу в основном посвящены истории населенных пунктов Ингерманландии и названиям внутри Санкт-Петербурга [Kepsu; Кепсу]. Комплексное историко-географическое исследование населенных пунктов Карельского перешейка, инициированное Балашовым, представляет собой продолжающуюся серию работ, в которых уделено внимание и топонимии [Балашов, 2009; Балашов, 2012]. Имеется несколько топонимических словарей, но они покрывают только небольшие территории [Муллонен, Азарова, Герд; Рябов].

Работы по региональной географической терминологии Ленинградской области нам неизвестны. Можно обратиться к более общим работам, например к «Словарю народных географических терминов» [Мурзаев], в котором мы встретим большую часть русских терминов для объектов Ингерманландии. Диссертация В. М. Мокиенко посвящена лингвистическому, в том числе этимологическому, анализу географических терминов [Мокиенко].

Далее мы приводим список географических терминов Ингерманландии по материалам топонимической базы данных. Все термины приведены в той форме и орфографии, как они записаны в базе, и распределены на основные группы по обобщенным типам объектов, на которые они указывают (см. таблицу).

Список географических терминов
в топонимической базе данных

1. Населенные пункты и их части:

город	*окраина деревни*	*усадьба*
деревня	*поселение*	*хутор*
мыза	*поселок*	*часть деревни*

2. Объекты в населенных пунктах:

аллея	*настил*	*сад*
колодец	*памятник*	*стена (каменная)*
колонка	*парк*	*улица*
лестница	*пруд*	

3. Постройки (в населенных пунктах и за их пределами):

ангары	*конюшня*	*сушилка*
баня	*коровник*	*телятник*
бани	*мельница*	*ферма*
больница	*погреб*	*хозяйственное сооружение*
водонапорная вышка	*развалины здания*	*холодильник*
двор	*рига*	*хранилище*
дом	*сарай*	*церковь*
жилье	*свиноферма*	*часовня*
здание	*сельский совет*	*школа*
комплекс	*скотный двор*	*школа-интернат*

4. Участки земли, связанные с сельскохозяйственной деятельностью:

выгон	*пастбище*	*поляна-выгон*
граница	*пожня*	*сенозаготовка*
земля	*пожог*	*сенокос*
луг	*покос*	*участок земли*
межа	*поле*	*часть покоса*
надел	*полосы*	

5. Леса, лесные участки, деревья:

березняк
бор
боровина
вырубка
грибное место
делянка
дуб
ельник
заросль
кусты
лес
лесной массив
малинник
поляна
посадка
роща
торфяник
урочище
черничник
ягодник

6. Водные объекты:

болотистый ручей
болото
болота
озеро
река
речка
ручей

7. Части водных объектов и источники воды:

банка
брод
заболоченное место
заводь
залив
источник
ключ
мост
мостик
низовье
омут
остатки плотины
плес
плотина
порог
родник
русло (старое)
рыбное место

8. Объекты рельефа:

бугры
валун
возвышенность
глинник
гора
горки
гористая местность
долина
каменная гряда
каменный грот
каменоломня
камень
канава
карьер
курган
курганы
лог
ложище
лощина
низина
низины
овраг
подземный ход
скала
склон
подошва холма
холм
яма

9. Разновидности дорог и их части:

дорога
подъем на дороге
развилка
саночный тракт
спуск
тропинка
участок шоссе
часть дороги
шоссе

10. Организации, учреждения, предприятия и т. п.:

заповедник
воинская часть
военно-охотничья база
каменодробильня
клуб охотников
колхоз
комплекс
охотоводство
охотхозяйство
садоводство
совхоз
хозяйство

11. Захоронения и кладбища:

захоронения
кладбище
могила
могильник
погост

12. Общие термины:

место
местность

Типы и количество географических терминов в топонимической базе данных

Тип объектов	Количество терминов
1. Населенные пункты и их части	9
2. Объекты в населенных пунктах	11
3. Постройки (в населенных пунктах и за их пределами)	30
4. Участки земли, связанные с сельскохозяйственной деятельностью	17
5. Леса, лесные участки, деревья	20
6. Водные объекты	7
7. Части водных объектов и источники воды	18
8. Объекты рельефа	28
9. Разновидности дорог и их части	9
10. Организации, учреждения, предприятия и т. п.	12
11. Захоронения и кладбища	5
12. Общие термины	2

Термины, использованные в топонимической картотеке и базе данных, можно сгруппировать следующим образом по степени употребительности в географии и топонимике.

I. Общераспространенные стандартные и местные географические термины: *мыза*, *колодец*, *сарай*, *луг*, *бор*, *озеро*, *брод*, *дорога*, *колхоз*, *кладбище*. Эти термины зафиксированы в «Энциклопеди-

ческом словаре географических терминов» [Энциклопедический словарь …] и «Словаре народных географических терминов» [Мурзаев]. Они, как правило, входят в число стандартных терминов, применяемых на картах и топографических планах [Условные знаки …].

II. Местные географические термины, не вошедшие в вышеупомянутые словари и используемые в микротопонимии: *настил*, *рига*, *делянка*, *пожня*, *пожог*, *надел*, *валун*, *малинник*, *черничник*, *ягодник*.

III. Нестандартные термины, которые имеют отношение к микротопонимии: *ложище*, *глинник*, *каменодробильня*, *клуб охотников*, *охотоводство*. Эти термины единичны, поэтому, вполне возможно, созданы информантами или даже собирателями топонимии.

IV. Нестандартные термины неопределенного содержания: *место*, *местность*.

При выборе географических терминов для топонимического словаря следует прежде всего использовать стандартные и общераспространенные местные термины из группы I. При этом следует руководствоваться словарями географических терминов [Энциклопедический словарь … ; Мурзаев] и стандартными руководствами для картографов [Условные знаки …]. Однако следует принять во внимание и по возможности использовать местные термины для микротопонимии из группы II. Нестандартные термины из группы III надо использовать с осторожностью и только если не удастся подобрать эквиваленты из групп I и II. Так, вероятно, можно принять термины *глинник и каменодробильня*. Термины из группы IV необходимо заменить на более конкретные.

Анализ географической терминологии в топонимической базе данных Ингерманландии показал, что термины разнородны и часто нестандартны. Необходимо произвести унификацию географических терминов для топонимических экспедиций. Требуется продолжить сбор и изучение географических терминов и выявить термины, специфичные именно для Ингерманландии. Для этого также понадобится изучить терминологические системы прибалтийско-финских языков, распространенных на территории Ингерманландии.

Литература

Автоматизированная база данных по топонимии как основа модели формирования историко-культурного ландшафта Ингерманландии

/ А. С. Герд [и др.] // Труды международной конференции «Финно-угорская топонимия в ареальном аспекте». Петрозаводск, 2007. С. 143–154.

Азарова И. В., Герд А. С., Николаев И. С. Корпус данных в проекте «Комплексная модель формирования культурного ландшафта и историко-культурной зоны Ингерманландии на Северо-Западе России по данным топонимики» // Труды международной конференции «Корпусная лингвистика — 2006». СПб., 2007.

Балашов Е. А. Карельский перешеек. Земля неизведанная. Юго-Западный сектор. Часть 1. 8-е изд. СПб., 2012.

Балашов Е. А. Метаморфозы топонимии Карельского перешейка. 5-е изд. СПб., 2009.

Кепсу С. Петербург до Петербурга. СПб., 2000.

Мокиенко В. М. Лингвистический анализ местной географической терминологии: автореф. дис. … канд. филол. наук. Л., 1969.

Муллонен И. И., Азарова И. В., Герд А. С. Словарь гидронимов Юго-Восточного Приладожья. Бассейн реки Свирь. СПб., 1997.

Мурзаев Э. М. Словарь народных географических терминов. М., 1984.

Научно-образовательный веб-ресурс «Топонимия Ингерманландии (Ленинградская область)»: перспективы исследования / А. С. Герд [и др.] // Структурная и прикладная лингвистика: Межвуз. сб. СПб., 2012. Вып. 9. С. 148–158.

Николаев И. С., Азарова И. В., Герд А. С. Свод топонимов Ленинградской области в Санкт-Петербургском государственном университете // Новый Топонимический журнал. 2005. № 2 : [Материалы II городской топонимической конференции 14 февраля 2005 г.].

Попов А. И. Следы времен минувших. М., 1981.

Рябов Д. С. Топонимия Верхнего Пооредежья: словарь-справочник. СПб., 2010.

Типовые запросы в базе данных по топонимии Ингерманландии и их практическая реализация / И. С. Николаев [и др.] // Материалы XXX–VI Международной филологической конференции (12–17 марта 2007 г.). СПб., 2007. Вып. 10 : Прикладная и математическая лингвистика. С. 74–84.

Условные знаки для топографических планов. М., 1989.

Энциклопедический словарь географических терминов. М., 1968.

Kepsu S. Inkereen nimistön ja asetuksen vaiheita // Inkerin Teillä. Kalevala-seuran vuosikirja 69–70. SKS. Helsinki, 1990.

Nikolaev I., Stolyarov D. Linguistic Information System of Multicultural Russian-Fennic Region of Ingermanland // Proceedings of International Multidisciplinary Scientific Conferences on Social Sciences and Arts (SGEM 2014). 2014. P. 131–137.

Е. А. Рогозина

УТОЧНЕНИЕ И XML-РАЗМЕТКА СЮЖЕТНОЙ СХЕМЫ ЖИТИЙ В КОРПУСЕ АГИОГРАФИЧЕСКИХ ТЕКСТОВ СКАТ

Аннотация. СКАТ — электронный корпус агиографических церковнославянских текстов XV–XVII вв., созданный на кафедре математической лингвистики СПбГУ. Тексты житий представлены в формате XML. Предусмотрена разметка, отражающая структуру содержания — сюжетную схему. По мере расширения корпуса общая сюжетная схема уточняется и дополняется, что позволяет вводить более подробную XML-разметку, создавать оглавления и вести над текстами дальнейшую исследовательскую работу.

Ключевые слова. Агиографические тексты, корпус, структура содержания, сюжет, оглавление, XML.

Elena A. Rogozina

ELABORATION OF HAGIOGRAPHIC TEXT CONTENT STRUCTURE AND ITS XML-MARKUP IN „SKAT“ CORPUS

Abstract. SCAT is a digital corpus of Church Slavonic hagiographic texts of 15th–17th centuries, created and developed on the initiative of the Department of Applied Linguistics of Saint Petersburg State University. Hagiographic texts are transformed into XML. XML files represent content structure which is common for all texts. As the corpus is populated with new texts, the plot scheme is being elaborated and more precise XML-markup is being introduced. This allows creating table of contents for each legend to help with further investigation of the texts.

Keywords. Corpus, hagiographic texts, content structure, plot, table of contents, XML-encoding.

СКАТ — электронный корпус агиографических церковнославянских текстов XV–XVII вв., созданный на кафедре математической лингвистики СПбГУ. На данный момент корпус содержит более 50 житий общим объемом свыше 500 тысяч словоупотреблений.

Более подробную информацию о корпусе СКАТ можно найти на вебстранице, посвященной проекту[1]. На этой же странице публикуются обработанные тексты житий в форматах PDF и XML. Разметка осуществляется на основе международных норм оформления электронных изданий текста, в частности Text Encoding Initiative (TEI).

На сайте проекта представлен электронный словоуказатель, который позволяет осуществлять поиск словоформ по всему корпусу текстов [Азарова, Алексеева]. Для текстов корпуса также вводится морфологическая разметка с указанием не только частей речи, но и типа склонения, рода, падежа и числа для существительных и прилагательных; времени, спряжения, лица, числа для глагольных форм и т. д. Ведется работа и над синтаксической разметкой [Алексеева, Азарова].

Кроме того, для текстов корпуса проводится анализ содержательной структуры, позволяющий выявить общую для всех житий сюжетную схему, а затем с помощью XML-разметки обозначить смысловые части в текстах. Это дает возможность составлять оглавления житий, осуществлять поиск и сопоставление сходных элементов в разных текстах [Рогозина].

На первых этапах работы была выделена схема построения сюжета житий, состоящая из элементов трех уровней: основных сюжетных блоков, их компонентов и подвижных модулей, которые не имеют фиксированного положения и могут встречаться в любой части текста.

Основная схема на уровне блоков выглядела следующим образом:

ВСТУПЛЕНИЕ
РОДИТЕЛИ
РОЖДЕНИЕ И МЛАДЕНЧЕСТВО
УЧЕНИЕ
ВОЗДЕРЖАНИЕ
РЕШЕНИЕ УЙТИ В МОНАСТЫРЬ
МОНАСТЫРЬ
ОСНОВАНИЕ МОНАСТЫРЯ
РАСШИРЕНИЕ МОНАСТЫРЯ
КОНЧИНА
ПОХВАЛЬНОЕ СЛОВО
ЗАКЛЮЧЕНИЕ

[1] URL: http://project.phil.pu.ru/skat/.

В тексте конкретного жития могут быть представлены все или только некоторые из этих блоков.

Каждый блок в свою очередь делится на более мелкие компоненты. Например, блок РОДИТЕЛИ делится на компоненты: «Предки» — рассказ о том, от кого происходит род святого, «Родители» — рассказ о его родителях, «Молитва о сыне» и «Божественное явление».

Также выделяются пять подвижных модулей, для которых не фиксировано ни положение в тексте, ни количество повторений. На первом этапе были выявлены модули *НАСТАВНИК*, *УЧЕНИК*, *МОНАШЕСКИЙ ПОДВИГ*, *ЧУДО* и *ИНТЕРМЕДИЯ*.

Однако по мере расширения корпуса в него добавлялись новые тексты, и их анализ привел к уточнению сюжетной схемы. Хотя при написании житийных текстов использовался общий шаблон [Панченко], его реализация могла различаться от текста к тексту. Возможность доработки схемы была предусмотрена заранее, поэтому был выбран такой вариант XML-разметки, который позволял бы учитывать новые варианты и дополнять имеющуюся схему новыми элементами.

Следует отметить, что степень вариативности сюжетной схемы различается для разных частей текста. Текст жития можно условно разделить на две части: до прихода святого в монастырь и после (начиная с блока МОНАСТЫРЬ). Содержание второй части гораздо легче формализовать, чем первую часть. Это объяснимо, поскольку происхождение святых и их мирская жизнь могли значительно различаться, а процедуры приема в монастырь и основания новой обители достаточно схожи. Поэтому с расширением корпуса во вторую часть схемы были внесены лишь незначительные изменения.

Например, блок МОНАСТЫРЬ остается неизменным от текста к тексту и может содержать компоненты «Приход в монастырь», «Постриг/благословение», «Добродетельное служение», «Слава и людское внимание» и «Желание уединения и уход из монастыря».

На данный момент во вторую часть сюжетной схемы были добавлены всего два новых элемента. В блоке ОСНОВАНИЕ МОНАСТЫРЯ появился компонент «Разрешение на строительство». В нем описывается ситуация, когда святой запрашивает разрешение на использование земли у ее владельца или просит у государя разрешения на создание монастыря. В итоге структура блока выглядит следующим образом:

ОСНОВАНИЕ МОНАСТЫРЯ
Поиск места
Молитва
Знамение
Разрешение на строительство
Основание монастыря
Основание и/или освящение церкви
Начало строительства
Появление первых учеников
Добродетельное служение

Кроме того, был выделен еще один подвижный модуль — *ИКОНА*, в котором рассказывается о создании иконы для монастыря. Этот модуль может появляться внутри блока ОСНОВАНИЕ МОНАСТЫРЯ или РАСШИРЕНИЕ МОНАСТЫРЯ, если преподобный сам заказывает образ (как, например, в «Житии Антония Сийского»), или же модуль может использоваться после блока КОНЧИНА, если кто-то из последующих настоятелей монастыря заказывает икону с изображением преподобного уже после его смерти (например, в «Житии Феодосия Тотемского»).

С первой частью сюжетной схемы сложилась несколько иная ситуация. События, связанные с мирской жизнью святого, описаны в более свободной форме, и сюжетные линии значительно различаются для разных текстов. При расширении корпуса пришлось пересмотреть сформулированную ранее сюжетную схему, в частности ввести новый блок МИРСКАЯ ЖИЗНЬ.

В этот блок входят следующие компоненты: «Брак (состоявшийся/несостоявшийся)», «Посещение монастыря», «Добродетельная жизнь», «Смерть родителей», «Смерть жены» (в случае брака) и «Желание уйти в монастырь». Причем особенность блока заключается в том, что для него невозможно установить стандартную последовательность компонентов. Порядок компонентов различается от одного жития к другому. Приведем несколько примеров:

Житие Григория Пельшемского	**Житие Антония Сийского**	**Житие Феодосия Тотемского**
МИРСКАЯ ЖИЗНЬ	МИРСКАЯ ЖИЗНЬ	МИРСКАЯ ЖИЗНЬ
Добродетельная жизнь	Добродетельная жизнь	Брак
Несостоявшийся брак	Смерть родителей	Добродетельная жизнь
Смерть родителей	Брак	Желание уйти в монастырь
	Смерть жены	Смерть родителей

Несмотря на отсутствие фиксированного порядка этих компонентов, отнести их к подвижным модулям представляется нецелесообразным, поскольку они используются только внутри блока МИРСКАЯ ЖИЗНЬ, не повторяются по несколько раз и не встречаются в других блоках или на стыках элементов, как это происходит с модулями.

Выбранный с самого начала вариант XML-разметки позволяет указать в тексте даже такие компоненты, не имеющие четкой последовательности. В данном проекте для обозначения элементов содержания используются элементы разметки <milestones>, которые не содержат текста и не требуют построения строгой иерархии или последовательности. Достаточно отметить начало и конец компонента и указать в качестве атрибута его название, чтобы затем можно было автоматически составить оглавление даже для такого блока с нефиксированным порядком компонентов.

Помимо ввода нового блока, в первой части текста также пришлось уточнить состав блока УХОД В МОНАСТЫРЬ (ранее называвшегося РЕШЕНИЕ УЙТИ В МОНАСТЫРЬ). Теперь в него входят следующие компоненты:

УХОД В МОНАСТЫРЬ
Решение уйти в монастырь
Дорога
Молитва
Знамение

В данном случае порядок следования компонентов фиксирован, как и в других блоках. Различия между конкретными реализациями выражаются лишь в том, что в тексте могут быть представлены не все компоненты. Например, могут присутствовать только элементы «Решение уйти в монастырь» и «Дорога» (с описанием пути в монастырь). Таким образом, блок МИРСКАЯ ЖИЗНЬ является единственным блоком, в котором порядок следования компонентов может изменяться.

В итоге уточненный вариант схемы первой части сюжета выглядит следующим образом:

ВСТУПЛЕНИЕ
РОДИТЕЛИ
РОЖДЕНИЕ И МЛАДЕНЧЕСТВО

УЧЕНИЕ
ВОЗДЕРЖАНИЕ
МИРСКАЯ ЖИЗНЬ
УХОД В МОНАСТЫРЬ

Итак, расширение корпуса текстов во многом подтверждает правильность выявленной схемы сюжета житий, но при этом позволяет уточнять детали и вводить более подробную разметку. Эта работа в свою очередь позволит выполнять более быстрый и удобный поиск и более точный анализ особенностей текста. А подтверждение универсальности схемы дает основание продолжить работу над поиском возможностей для автоматизации разметки элементов содержания.

Литература

Азарова И. В., Алексеева Е. Л. Санкт-Петербургский корпус агиографических текстов (СКАТ): формат XML-представления лингвистической информации и организация поиска данных на сайте // Материалы международной научной конференции «Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. El'Manuscript-08». Казань, 2008. С. 3–6.

Алексеева Е. Л., Азарова И. В. Особенности морфо-синтаксической разметки древнерусских агиографических текстов // Труды международной конференции «Корпусная лингвистика — 2013». СПб., 2013. С. 157–164.

Панченко О. В. Поэтика уподоблений (к вопросу о «типологическом» методе в древнерусской агиографии, эпидейктике, гимнографии) // Труды Отдела древнерусской литературы / Российская академия наук. Институт русской литературы (Пушкинский Дом). СПб., 2003. Т. 54. С. 491–534.

Рогозина Е. А. Корпус агиографических текстов СКАТ: сюжетная схема житий, XML-разметка и создание оглавлений // Структурная и прикладная лингвистика: Межвуз. сб. / под ред. А. С. Герда. СПб., 2010.. Вып. 8. С. 243–249.

М. В. Хохлова

БОЛЬШИЕ КОРПУСЫ И ЧАСТОТНЫЕ СУЩЕСТВИТЕЛЬНЫЕ: ПРЕДВАРИТЕЛЬНЫЕ НАБЛЮДЕНИЯ

Аннотация. Статья представляет результаты исследования частотных существительных русского языка на материале корпусов разных объемов. Анализ показывает, что данные, приведенные в частотном словаре и полученные на корпусной основе, отличаются. В статье также дается обзор русских корпусов большого объема.

Ключевые слова. Корпусы, русский язык, статистика, большие данные, частотные существительные.

Maria V. Khokhlova

BIG CORPORA AND HIGH-FREQUENCY NOUNS: PRELIMINARY OBSERVATIONS

Abstract. The paper discusses the results of the study frequency properties for the high-frequency Russian nouns on corpora of various sizes. The analysis shows the data presented in the frequency dictionary of Russian differs from the corpus data. The paper also gives a survey of Russian big corpora.

Keywords. Corpora, Russian language, statistics, big data, high-frequency nouns.

Введение

Задача создания корпусов, содержащих большие данные, ставилась многократно. Исследователей привлекала возможность проверить свои выводы на количественно новом материале. Но только с появлением широких технических возможностей решение данной задачи развилось в отдельное направление. В многочисленных работах обсуждаются подходы к автоматическому сбору материала из Интернета ([Kehoe, Renouf; Kilgarriff, Grefenstette; Беликов,

Селегей, Шаров] и др.) и созданию корпусов текстов больших объемов. Порог в 1 млн словоупотреблений или даже в 100 млн словоупотреблений давно пройден и уже не является эталоном. Вместе с появлением больших корпусов возникают новые вопросы: что дают большие данные, зачем они нужны и как меняются результаты, полученные на их основе.

Целью нашего исследования является сравнение результатов, выдаваемых системой Sketch Engine при работе с тремя корпусами разных объемов. Нами будут рассмотрены частотные характеристики ряда русских существительных.

Большие корпусы русского языка

Для экспериментов были отобраны следующие корпусы русского языка: интернет-корпус Russian Web Corpus (147 млн токенов) и гигакорпус ruTenTen (18,28 млрд токенов). Первый был создан С. А. Шаровым [Sharoff] на базе автоматически загруженных текстов из Интернета по технологии, описанной в [Baroni, Bernardini]. На основе списка из 500 наиболее частотных слов для данного языка, которые не являются служебными или характерными для некоторой предметной области, последовательно создаются многочисленные запросы (от 5000 до 8000), позволяющие получить ссылки на сайты. В каждом запросе участвует по четыре слова из списка. Далее отбираются 10 топ-адресов и скачиваются соответствующие им тексты. После проводится распознавание кодировки, удаление многочисленных версий одних и тех же страниц или текстов на других языках (например, в случае русскоязычного корпуса требуется удаление текстов на славянских языках, использующих кириллицу). Семейство TenTen [The TenTen Corpus Family …], к которому принадлежит второй корпус, включает в себя корпусы разных языков, объем каждого из которых превышает 1 млрд слов. Отличие данного подхода заключается в том, что скачиваются не все тексты из всех возможных доменов, соответствующих рассматриваемому языку. В качестве инструмента используется специальная поисковая программа (англ. crawler), позволяющая загружать только те тексты, в которых содержатся полные предложения (а не страницы с техническими данными). При этом уделяется внимание тому, чтобы новые тексты не были дублями старых, таким образом сокращается

процесс постобработки. Корпус ruTenTen имеет подкорпус объемом 998 млн слов, который также будет нами использован.

Рассмотренные технологии являются привлекательными, так как позволяют создавать корпусы для разных языков, не требуя предварительного затратного по времени и силам сбора текстов (хотя это достоинство можно поставить под сомнение, вспомнив о таком неотъемлемом свойстве корпусов и выборок, как сбалансированность, или репрезентативность).

Материал для исследования

В основном тексты, входящие в состав интернет-корпусов русского языка, представляют собой материалы новостных ресурсов, блогов, рекламных сайтов, групп социальных сетей и др. Художественные тексты представлены не так широко, поэтому было решено обратиться к спискам частотной лексики, которые отражают именно данные функциональные стили. Нами были сформированы два списка слов. В первый список (табл. 1, 2) попали существительные, являющиеся наиболее частотными по словарю [Ляшевская, Шаров] для текстов публицистики и другой нехудожественной литературы (эти две группы представлены в словаре отдельно).

Как видно из табл. 1 и 2, списки слов в них пересекаются, поэтому в окончательный список вошли 14 существительных (индексами отмечены слова, присутствующие в обеих таблицах): *год*[1,2], *время*[1,2],

Таблица 1. **Частотные существительные из словаря другой нехудожественной литературы**

№	Лемма	Частота (ipm)
1	год	4624,2
2	время	2080,5
3	человек	1945,3
4	система	1798,0
5	работа	1766,4
6	статья	1363,0
7	дело	1339,5
8	случай	1259,0
9	процесс	1221,8
10	вопрос	1180,9

Таблица 2. **Частотные существительные из словаря публицистики**

№	Лемма	Частота (ipm)
1	год	5589,5
2	человек	2950,1
3	время	2364,6
4	жизнь	1548,4
5	дело	1482
6	день	1397,8
7	работа	1272,4
8	страна	1203,9
9	вопрос	992
10	слово	989,7

человек[1,2], *система*, *работа*[1,2], *статья*, *дело*[1,2], *случай*, *процесс*, *вопрос*[1,2], *жизнь*, *день*, *страна* и *слово*.

Второй список состоит из существительных, относящихся к так называемой значимой лексике (то есть наиболее характерной) [Ляшевская, Шаров] для нехудожественных текстов и публицистики (табл. 3, 4).

Таблица 3. **Частотные существительные из словаря значимой лексики другой нехудожественной литературы**

№	Лемма	Частота (ipm)		LL-score[1]
		Корпус	Подкорпус	
1	статья	395,0	1363,0	10512
2	система	617,8	1798,0	9943
3	федерация	258,9	1003,1	9329
4	процесс	371,7	1221,8	8639
5	рисунок	179,2	776,2	8451
6	вирус	106,5	584,1	8388
7	исследование	200,5	799,6	7762
8	использование	190,3	757,9	7342
9	суд	371,1	1153,9	7334
10	метод	197,0	772,3	7312

[1] Коэффициент логарифмического правдоподобия — статистическая мера, которая использовалась авторами словаря для выявления значимой лексики. Результаты в табл. 3 и 4 упорядочены по данному параметру.

Таблица 4. **Частотные существительные из словаря значимой газетно-новостной лексики**

№	Лемма	Частота (ipm)		LL-score
		Корпус	Подкорпус	
1	президент	311,0	634,6	2186
2	театр	305,3	611,9	1944
3	год	3727,5	5589,5	1435
4	спектакль	164,7	350,0	1429
5	правительство	277,7	531,2	1341
6	компания	392,7	699,0	1149
7	страна	725,7	1203,9	1135
8	фильм	196,8	380,1	1009
9	реформа	133,1	273,0	963
10	выборы	117,7	243,4	889

Результаты исследования

В ходе исследования мы также рассмотрели 10 наиболее частотных существительных в трех корпусах. Список, полученный на материале Russian Web Corpus, оказался практически идентичным списку в [Ляшевская, Шаров]. Результаты, полученные для двух других корпусов, не столь тривиальны. Так, в ruTenTen наиболее частотными оказались: *год*, *работа*, *время*, *человек*, *компания*, *система*, *сайт*, *день*, *место* и *Россия*. Леммы *система* и *компания* имеют ранги 26 и 59 в частотном списке существительных, а *сайт* и *Россия* вообще в него не входят. Среди результатов из выборки ruTenTen не встречается слово *сайт*, при этом есть лексема *ребенок*, имеющая в [Ляшевская, Шаров] ранг 22.

Нами были изучены частоты слов из вышеприведенных списков (см. табл. 1, 2) на материале трех корпусов, результаты представлены в табл. 5 и на рис. 1, а также в табл. 6 и на рис. 2.

В табл. 5 и на рис. 1 приведены данные для существительных из табл. 1. Можно заметить, что графики для корпуса ruTenTen и выборки из ruTenTen совпадают, то есть данные слова распределены одинаково. Частоты, проиллюстрированные в словаре, оказались максимальными за исключением частоты для леммы *человек*, которая достигает своего наибольшего значения в корпусе Russian Web Corpus. Все три корпуса показывают некоторое расхождение со

Таблица 5. **Частоты существительных из словаря другой нехудожественной литературы в трех корпусах**

№	Лемма	Частота (ipm)			
		Частотный словарь другой нехудожественной литературы	Russian Web Corpus	ruTenTen	
				Корпус	Подкорпус
1	год	4624,2	2220,74	3078,97	3076,99
2	время	2080,5	1761,06	1790,84	1793,41
3	человек	1945,3	2343,79	1955,40	1950,79
4	система	1798,0	527,61	998,41	1006,66
5	работа	1766,4	885,02	1509,37	1510,41
6	статья	1363,0	257,55	293,72	292,09
7	дело	1339,5	1037,09	813,12	809,29
8	случай	1259,0	632,16	750,61	752,11
9	процесс	1221,8	294,37	473,94	478,05
10	вопрос	1180,9	853,94	866,03	869,27

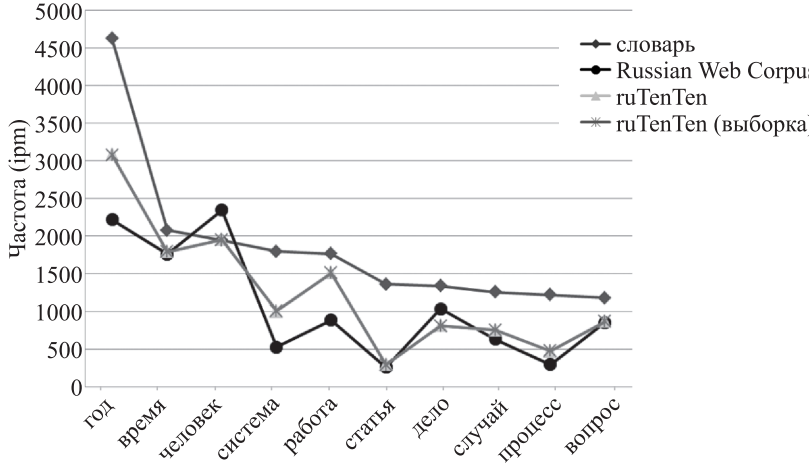

Рис. 1. Распределение частот существительных из словаря другой нехудожественной литературы в трех корпусах

Таблица 6. **Частоты существительных из словаря публицистики в трех корпусах**

№	Лемма	Частота (ipm)			
		Словарь публицистики	Russian Web Corpus	ruTenTen	
				Корпус	Подкорпус
1	год	5589,50	2220,74	3078,97	3076,99
2	человек	2950,10	2343,79	1955,40	1950,79
3	время	2364,60	1761,06	1790,84	1793,41
4	жизнь	1548,40	1054,70	864,40	862,25
5	дело	1482,00	1037,09	813,12	809,29
6	день	1397,80	1052,35	1089,16	1088,18
7	работа	1272,40	885,02	1509,37	1510,41
8	страна	1203,90	576,81	662,05	664,03
9	вопрос	992,00	853,94	866,03	869,27
10	слово	989,70	807,83	633,83	631,81

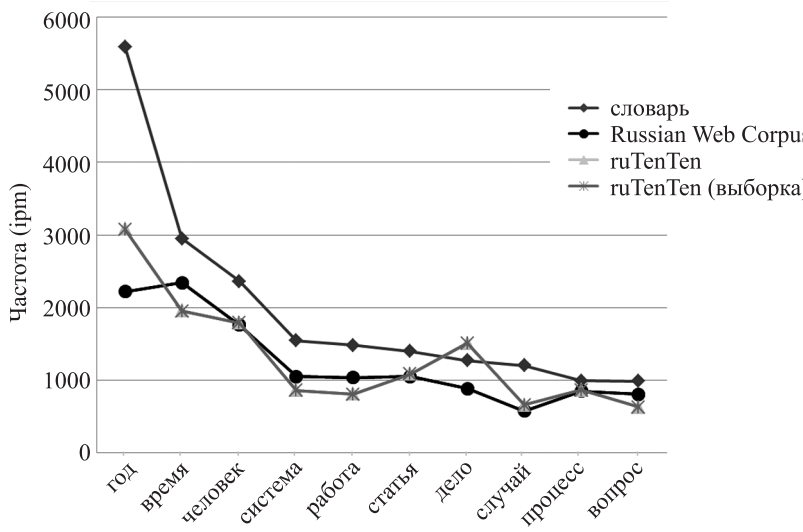

Рис. 2. Распределение частот существительных из словаря публицистики в трех корпусах

словарными данными в рангах для слов: в Russian Web Corpus два существительных имеют те же ранги, а в ruTenTen (и в выборке) — четыре. При этом в обоих корпусах одинаковые ранги имеют четыре слова: *время*, *вопрос*, *процесс* и *статья*. Коэффициент корреляции Спирмена между ранговыми распределениями по словарю и корпусу

Russian Web Corpus равен 0,61, а по словарю и корпусу ruTenTen — 0,78, что говорит о более тесной связи в последнем случае.

На рис. 2 приведены данные для существительных из табл. 2: как и на рис. 1, результаты для корпуса ruTenTen и выборки из ruTenTen одинаковы. Частота для слова *работа* в корпусе ruTenTen оказалась выше частоты по словарю, в котором в остальных случаях отражены максимальные значения. Коэффициент корреляции Спирмена между ранговыми распределениями по словарю и корпусу Russian Web Corpus весьма высок и равен 0,94, что может свидетельствовать о том, что данный корпус по своему составу ближе к газетным текстам. В корпусах Russian Web Corpus и ruTenTen одинаковые ранги имеют только два слова: *время* и *день*.

На следующем этапе эксперимента нами были рассмотрены частоты существительных из словарей значимой лексики (табл. 3, 4). Результаты для первой группы существительных представлены в табл. 7 и на рис. 3, для второй группы — в табл. 8 и на рис. 4.

Наибольшие частоты (в единицах ipm), как видно из рис. 3 и 4, были получены для словарного подкорпуса, что имеет свое

Таблица 7. **Частоты существительных из словаря значимой лексики другой нехудожественной литературы в трех корпусах**

№	Лемма	Частота (ipm)				
		Словарь значимой лексики другой нехудожественной литературы		Russian Web Corpus	ruTenTen	
		Корпус	Подкорпус		Корпус	Подкорпус
1	система	617,8	1798,0	527,61	998,41	1006,66
2	статья	395,0	1363,0	257,55	293,7	292,09
3	процесс	371,7	1221,8	294,37	473,94	478,05
4	суд	371,1	1153,9	197,00	302,98	301,73
5	федерация	258,9	1003,1	88,03	198,31	197,06
6	исследование	200,5	799,6	154,44	265,11	266,86
7	метод	197,0	772,3	153,51	263,74	265,91
8	использование	190,3	757,9	146,83	346,74	350,04
9	рисунок	179,2	776,2	45,19	77,04	76,84
10	вирус	106,5	584,1	21,01	36,90	36,75

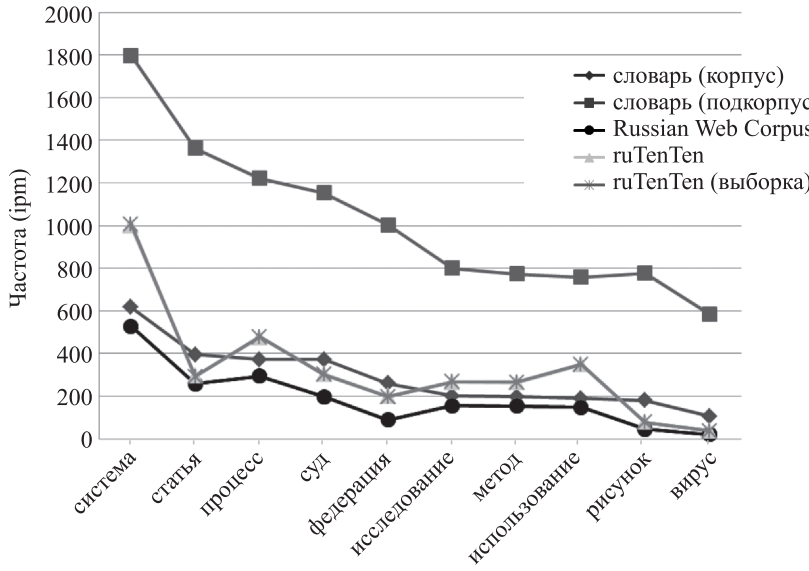

Рис. 3. Распределение частот существительных из словаря значимой лексики другой нехудожественной литературы в трех корпусах

Таблица 8. **Частоты существительных из словаря значимой газетно-новостной лексики в трех корпусах**

№	Лемма	Частота (ipm)				
		Словарь значимой газетно-новостной лексики		Russian Web Corpus	ruTenTen	
		Корпус	Подкорпус		Корпус	Подкорпус
1	год	3727,5	5589,5	2220,74	3078,97	3076,99
2	страна	725,7	1203,9	576,81	662,05	664,03
3	компания	392,7	699,0	390,72	970,15	979,11
4	президент	311,0	634,6	185,6	215,07	213,81
5	театр	305,3	611,9	91,08	102,09	99,28
6	правительство	277,7	531,2	183,25	225,28	224,83
7	фильм	196,8	380,1	172,15	214,16	213,10
8	спектакль	164,7	350,0	37,09	44,42	42,78
9	реформа	133,1	273,0	58,48	47,16	47,74
10	выборы	117,7	243,4	—	62,34	63,20

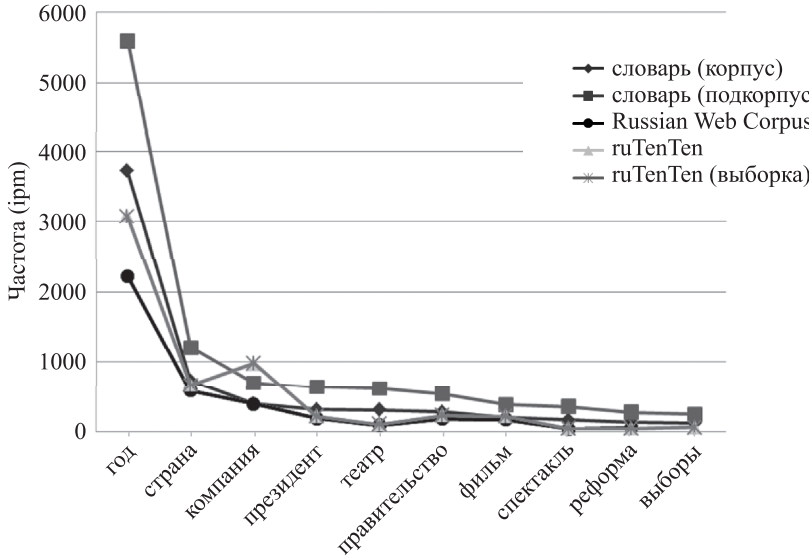

Рис. 4. Распределение частот существительных из словаря значимой газетно-новостной лексики в трех корпусах

объяснение. Несмотря на то что слова по значимости для определенных функциональных стилей отбирались не на основе абсолютных частот, а при помощи показателя LL, тем не менее эта мера включает число единиц в подкорпусе, что и приводит к тому, что лексемы, частоты которых превышают общекорпусные значения, оказываются маркированными как значимые. Согласованность в рангах на материале корпусов Russian Web Corpus и ruTenTen демонстрируют четыре существительных: *система*, *процесс*, *суд*, *федерация*, *рисунок и вирус*. Это максимальное совпадение, которое наблюдалось среди корпусных данных. Коэффициент корреляции Спирмена между ранговыми распределениями по словарю и корпусу Russian Web Corpus равен 0,92 и может указывать на тот факт, что данные словаря значимой нехудожественной литературы и корпуса совпадают в значительной степени (то же характерно, хотя и в меньшей степени, для данных в корпусе ruTenTen, коэффициент корреляции Спирмена между ним и корпусом словаря тоже высок и равен 0,73).

Обращает на себя внимание начальное максимальное значение на диаграмме, соответствующее частоте слова *год* и прослеживаемое

на всех графиках. Это объясняется тем, что данное существительное входит в тройку самых частотных во всех корпусах. Для лексемы *выборы* не были получены результаты в корпусе Russian Web Corpus, так как в его морфологической разметке словоупотребления сведены к лемме *выбор*. Одинаковые ранги имеют первые четыре существительных в Russian Web Corpus и в словаре (в основном корпусе и подкорпусе): *год*, *страна*, *компания* и *президент*. Также для словаря и корпуса Russian Web Corpus коэффициент ранговой корреляции Спирмена максимален и составляет 0,95.

Заключение

Общий вывод, который можно сделать на основе полученных данных, свидетельствует о том, что тексты больших корпусов отражают язык Сети. Результаты, приведенные в частотном словаре, были основаны на Национальном корпусе русского языка, что объясняет их сбалансированность.

Существительные, которые оказались наиболее частотными в корпусе ruTenTen и в его миллионной выборке и не отраженные в списке результатов в частотном словаре (*сайт*, *система*, *компания* и *Россия*), отражают специфику текстов, взятых из сети Интернет, во-первых, из-за большого количества новостных ресурсов, во-вторых, ввиду направленности на описание содержания вебстраниц. Корпус Russian Web Corpus показывает бóльшую согласованность данных с частотным словарем [Ляшевская, Шаров], чем корпус ruTenTen.

Видится дальнейшее развитие описанной методологии. Необходимо изучить результаты для низкочастотных слов, так как именно для них большие корпусы могут дать совершенно иные количественные показатели. Так, предварительные результаты проведенных нами экспериментов по сочетаемости показали, что не всегда увеличение абсолютной частоты приводит к увеличению числа отношений, в которые вступает данная единица (хотя увеличивает число синтагматических партнеров, характерных для каждого отношения).

Литература

Беликов В. И., Селегей В. П., Шаров С. А. Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ) 2012 // Компьютер-

ная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18): в 2 т. М., 2012. Т. 1. С. 37–49.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М., 2009. URL: http://dict.ruslang.ru/freq.php (дата обращения: 24.12.2015).

Baroni M., Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web // Proceedings of LREC 2004. Lisbon: ELDA, 2004. P. 1313–1316.

Kehoe A., Renouf A. WebCorp: Applying the Web to Linguistics and Linguistics to the Web // WWW2002 Conference Proceedings. Honolulu, Hawaii, 2002. URL: http://WWW2002.org/CDROM/poster/67/ (дата обращения: 15.12.2015).

Kilgarriff A., Grefenstette G. Introduction to the Special Issue on Web as Corpus // Computational Linguistics, 29 (3). P. 333–347.

Sharoff S. Creating General-Purpose Corpora Using Automated Search Engine Queries // WaCky! Working papers on the Web as Corpus. Bologna, 2006. P. 63–98.

The TenTen Corpus Family / M. Jakubíček [et al.] // Proceedings of the 7th International Corpus Linguistics Conference. Lancaster, 2013. P. 125–127.

М. Н. Бабарико, С. В. Чебанов

РУССКАЯ ПАРЕМИОЛОГИЧЕСКАЯ АРИФМОЛОГИЯ XIX–XXI ВЕКОВ

Аннотация. В собраниях паремий Даля и Мокиенко — Никитиной фигурирует 71 число от 1 до 1 000 000 (45 и 62 соответственно; 36 общих), выраженных именами числительными и описательными числовыми концептами (ОЧК). Самым частым числительным является 1, далее с заметным отрывом — 2, 3, 7. При учете ОЧК 100 и 24, они занимают третье и четвертое места. Увеличение доли числа 1 в собрании Мокиенко — Никитиной рассматривается как следствие монополизации советского периода, а увеличение количества чисел при уменьшении их частот — как результат размывания структуры идеализированной когнитивной модели (ИКМ) чисел русской языковой картины мира вследствие перехода от счета дюжинами к счету десятками и принятия метрической системы единиц.

Ключевые слова. Числа, арифмология, паремии, собрание паремий Даля, собрание паремий Мокиенко — Никитиной, русская языковая картина мира, имена числительные, описательные числовые концепты.

Maxim N. Babariko, Sergey V. Chebanov

RUSSIAN PAREMIOLOGICAL ARITHMOLOGY OF THE 19th–21th CENTURIES

Abstract. 71 numbers between 1 and 1 000 000 expressed by numerals and descriptive concepts are featured in the collections of proverbs by Dal and Mokienko & Nikitina (45 and 62 respectively; 36 overlapped). The most common is the numeral 1, proceeded with a noticeable margin by the 2, 3, 7. Taking into consideration the descriptive concepts the 100 and the 24, they take the third and the fourth places. The increase in the proportion of the 1 in the collection of proverbs by Mokienko & Nikitina is seen as a consequence of monopolization of the Soviet period, and the increase of the numbers with the decrease of their frequency — as a result of the structure's erosion of an idealized cognitive model (ICM) of the numbers of Russian language worldview owing to the transition from counting in the dozens to counting in the tens and adoption of the metric system.

Keywords. Numbers, arithmology, paroemias (proverbs), Dal's paroemia collection, Mokienko & Nikitina's paroemia collection, Russian linguistic worldview, numerals, number descriptive concepts.

1. Введение

Предыдущая работа авторов [Бабарико, Чебанов], посвященная числовым концептам паремий собрания В. И. Даля [Даль], показала, что с точки зрения структуры идеализированных когнитивных моделей [Лакофф], их арифмологии, культурная среда, в которой бытовали эти пословицы, была по сути глубоко монархической. Такой вывод является очень важным для лингвосоциологии [Никольский; Швейцер, Никольский], с точки зрения которой в русской культуре трудно реализовать, скажем, принцип разделения властей, так что дисфункции законодательной (парламентской) и судебной властей представляются вполне закономерными.

Для проверки этого предположения, в частности методами лингвосоциологии посредством изучения числовых концептов, как и планировалось [Бабарико, Чебанов], были получены результаты на другом материале, который также представляется репрезентативным. Речь идет об исследовании в указанном аспекте собрания поговорок В. М. Мокиенко и Т. Г. Никитиной [Мокиенко, Никитина]; обработка данных велась по электронной версии[1].

Основания, методологические принципы, методика изучения были абсолютно такими же, какие описаны в указанной работе [Бабарико, Чебанов].

При этом надо иметь в виду, что собрание поговорок В. М. Мокиенко и Т. Г. Никитиной (далее СМН) по сравнению с собранием пословиц Даля (далее СД) отражает временные сдвиги использования чисел: в СМН собрано свыше 40 000 русских поговорок, отражающих литературную и народную речь XIX–XXI веков, а СД представляет народную речь XIX и предыдущих веков.

В СМН, так же как и в СД, были рассмотрены все выявленные лексемы числительных (Ч) и обороты, которые выражают описательные числовые концепты (ОЧК), обозначающие числа от единицы (**1**) до миллиона (**1 000 000**).

[1] URL: http://www.twirpx.com/file/222426/.

2. Числа в собрании поговорок В. М. Мокиенко и Т. Г. Никитиной

2.1. Общая характеристика использования чисел

В табл. 1 представлены данные о Ч и соответствующих им ОЧК. Количество употреблений чисел (Ч + ОЧК) не совпадает с количеством поговорок (в одной поговорке может фигурировать больше одного числа).

Данные табл. 1 и иллюстрирующие ее рис. 1 и 2 позволяют сделать заключения о характере использования числовых концептов в поговорках XIX–XXI веков.

Самыми употребляемыми Ч всех обнаруженных (см. далее) разрядов в СМН являются **1** (551 употреблений — 39,9% {34,9%}[2] от общего количества Ч), **2** (280 — 20,3% {22,6%}, накопленная доля (н. д.) 60,2%), **3** (165 — 11,9% {12,3%}, н. д. 69,8%), **7** (86 — 6,2% {4,4%}, н. д. 78,3%), **5** (43 — 3,1% {3,8%}, н. д. 82,1%), **100** (40 — 2,9%, {3,8%}, н. д. 85%), доля же каждого из остальных составляет не более 2,5%.

Последовательность частот ОЧК в СМН оказывается несколько иной, чем в СД: **100** (247 — 25,9% {22,0%}), **24** (147 — 15,4% {24,9%}, н. д. 41,3%), **365** (146 — 15,2% {11,7%}, н. д. 56,5%), **1** (129 — 13,5% {10,8%}, н. д. 70%), **2** (98 — 10,2% {9,1%}, н. д. 80,2%), **60** (71 — 7,4% {6,2%}, н. д. 87,6%), **7** (33 — 3,4% {5,5%}, н. д. 91%), **10** (28 — 2,9% {такой же процент занимала **5** в СД}, н. д. 93,9%); доли остальных не превышают 1,6%. Эта последовательность определена тем, что день интерпретируется как сутки (24 часа), рубль — как 100 копеек (что соответствует реалиям рассматриваемого периода), год — как 365 дней, час — как 60 минут, а минута — как 60 секунд.

Распределение употребления Ч в СМН характеризуется, так же как в СД, очень высокой концентрацией словоупотреблений в самом начале (первые три члена **1**–**2**–**3** выбирают практически 70% всего объема). Примечательно также, что разница между Ч **1** и **2** в СМН оказывается еще больше, чем в СД. Процент употребления Ч **3** в поговорках СМН уменьшается, хотя и незначительно. В то же время почти на 2% увеличивается употребление Ч **7**, а Ч **5** и **100** поменялись местами.

[2] Здесь и далее в фигурных скобках указан соответствующий процент в СД.

Таблица 1. **Числовые концепты в СМН, выраженные Ч и ОЧК**

Число	Выражение числа								
	Ч			ОЧК			Ч + ОЧК		
	Форм	Употреблений		Форм	Употреблений		Форм	Употреблений	
		Абс.	%		Абс.	%		Абс.	%
1	24	551	39,9	22	129	13,5	46	680	29,1
2	22	280	20,3	14	98	10,2	36	378	16,2
3	14	165	11,9	6	8	0,8	20	173	7,4
4	3	26	1,9	3	10	1,0	6	36	1,5
5	9	43	3,1	8	15	1,6	17	58	2,5
6	6	13	0,9	1	1	0,1	7	14	0,6
7	12	86	6,2	7	33	3,4	19	119	5,1
8	5	9	0,7	3	4	0,4	8	13	0,6
9	7	33	2,4	2	9	0,9	9	42	1,8
10	5	30	2,2	8	28	2,9	13	58	2,5
11	2	6	0,4	0	0	0	2	6	0,3
12	3	8	0,6	0	0	0	3	8	0,3
13	1	1	0,1	1	1	0,1	2	2	0,1
14	1	1	0,1	0	0	0	1	1	0,0
15	1	4	0,3	0	0	0	1	4	0,2
16	3	5	0,4	0	0	0	3	5	0,2
17	1	2	0,1	0	0	0	1	2	0,1
20	3	11	0,8	0	0	0	3	11	0,5
21	1	1	0,1	0	0	0	1	1	0,0
22	1	1	0,1	0	0	0	1	1	0,0
23	2	2	0,1	0	0	0	2	2	0,1
24	1	2	0,1	11	147	15,4	12	149	6,4
30	1	3	0,2	1	6	0,6	2	9	0,4
31	1	1	0,1	0	0	0	1	1	0,0
32	2	2	0,1	0	0	0	2	2	0,1
33	2	2	0,1	0	0	0	2	2	0,1
36	1	2	0,1	0	0	0	1	2	0,1
38	1	2	0,1	0	0	0	1	2	0,1
40	2	10	0,7	0	0	0	2	10	0,4
41	1	1	0,1	0	0	0	1	1	0,0
44	1	1	0,1	0	0	0	1	1	0,0

Продолжение табл. 1

Число	Выражение числа								
	Ч			ОЧК			Ч + ОЧК		
	Форм	Употреблений		Форм	Употреблений		Форм	Употреблений	
		Абс.	%		Абс.	%		Абс.	%
45	1	2	0,1	0	0	0	1	2	0,1
46	1	1	0,1	0	0	0	1	1	0,0
47	1	1	0,1	0	0	0	1	1	0,0
50	0	0	0	1	1	0,1	1	1	0,0
52	1	1	0,1	0	0	0	1	1	0,0
55	1	1	0,1	0	0	0	1	1	0,0
60	0	0	0	13	71	7,4	13	71	3,0
69	1	1	0,1	0	0	0	1	1	0,0
70	1	1	0,1	0	0	0	1	1	0,0
95	1	1	0,1	0	0	0	1	1	0,0
99	1	1	0,1	0	0	0	1	1	0,0
100	2	40	2,9	19	247	25,9	21	287	12,3
101	2	4	0,3	0	0	0	2	4	0,2
105	1	1	0,1	0	0	0	1	1	0,0
108	1	1	0,1	0	0	0	1	1	0,0
120	1	1	0,1	0	0	0	1	1	0,0
180	1	1	0,1	0	0	0	1	1	0,0
200	1	3	0,2	0	0	0	1	3	0,1
220	1	1	0,1	0	0	0	1	1	0,0
250	1	1	0,1	0	0	0	1	1	0,0
300	1	1	0,1	0	0	0	1	1	0,0
365	0	0	0	9	146	15,2	9	146	6,2
500	1	1	0,1	0	0	0	1	1	0,0
506	1	1	0,1	0	0	0	1	1	0,0
911	1	1	0,1	0	0	0	1	1	0,0
1 000	3	5	0,4	0	0	0	3	5	0,2

Окончание табл. 1

Число	Выражение числа								
	Ч			ОЧК			Ч + ОЧК		
	Форм	Употреблений		Форм	Употреблений		Форм	Употреблений	
		Абс.	%		Абс.	%		Абс.	%
30 000	1	1	0,1	0	0	0	1	1	0,0
35 000	1	1	0,1	0	0	0	1	1	0,0
40 000	1	1	0,1	0	0	0	1	1	0,0
10 0000	1	1	0,1	0	0	0	1	1	0,0
1 000 000	1	2	0,1	2	2	0,2	3	4	0,2
Итого	157	1382	100,0	80	956	100,0	237	2338	100,0

П р и м е ч а н и е. Здесь и далее округление процентов проведено по общим правилам округления до десятых долей, что допускает некоторое отклонение суммы от 100%.

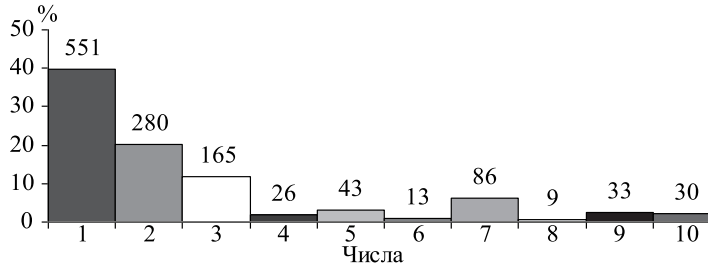

Рис. 1. Числовые концепты от **1** до **10** в СМН, выраженные Ч

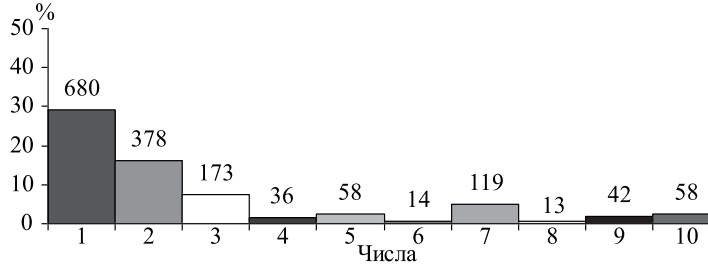

Рис. 2. Числовые концепты от **1** до **10** в СМН, выраженные Ч и ОЧК

Последовательность наиболее высокочастотных ОЧК чисел в СМН отличается от последовательности таковых в СД. Во-первых, самым частотным числом становится **100** (происходит резкий скачок вверх за счет ОЧК «век»), а процент употребления числа **24** резко падает с 24,9 до 15,4%. Также при суммарном равенстве употребления чисел **5** и **10** частоты Ч (43 и 30 употреблений соответственно) и ОЧК (15 и 28) меняются местами.

Суммарное распределение чисел (Ч + ОЧК) частично восстанавливает последовательность, характерную для чисел в СД (но с меньшими долями при возрастании абсолютного количества), — самыми частыми оказываются **1** (680 — 29,1% {25,2%}) и **2** (378 — 16,2% {17,2%}, н. д. 45,3%), далее **100** (287 — 12,3% {11,2%}), **3** (173 — 7,4% {8,4%}), **24** (149 — 6,4% {10,0%}), **365** (146 — 6,2% {4,7%}) и **7** (119 — 5,1% {6,4%}) — **100**, **24** и **365** в основном за счет ОЧК. Высокую долю чисел **100** и **24** (как и менее высокочастотных **60** и **365**) можно связать с явлением, аналогичным лексикализации устойчивых словосочетаний.

В данном распределении можно заметить, что число **1** становится в СМН более употребляемым, чем в СД: частота его употребления возросла с 25,2% (СД) до 29,1% (СМН). Напротив, число **2** стало менее употребляемым: 17,2% (СД), 16,2% (СМН). Частота употребления числа **100** увеличилась на 1%. Примечательно, что числа **3** и **24** меняются местами — число **3** в СМН употребляется чаще, чем **24**. Также меняются местами числа **7** и **365**. В СМН (за счет ОЧК «год») доля числа **365** больше, чем на 1%, превышает долю **7**.

Так же как в СД, в поговорках СМН число **3** не оказывается наиболее частым ни среди употребляемых Ч (четвертое по частоте), ни среди ОЧК (двенадцатое по частоте), занимая в сумме четвертое место (в СД — пятое), причем число **2** остается в СМН на том же месте. **2** имеет большую частоту по сравнению с СД как среди Ч, так и среди ОЧК (пятое по частоте), занимая в сумме второе место.

В СМН числа первого десятка охватывают большую часть высокочастотных концептов (рис. 1 и 2). И для Ч, и для Ч + ОЧК имеет место почти один и тот же характер распределения: от **1** до **4** идет монотонное падение частотности (резко выраженное), на **5** приходится локальный подъем, затем на **6 —** спад, а на **7** — довольно острый локальный максимум (правда, несколько меньший частоты **3**), на

8 — абсолютный минимум (!), а далее доля Ч **9** становится больше, чем у **10**. По сумме Ч и ОЧК доля **9** снова меньше, чем у **10** (этому способствует ОЧК «десятка»), после **7** идет незначительный рост. Для **8** в СМН прибавляется ОЧК «восьмерка» и для **9** — ОЧК «девятка», а для **6** — только одно употребление ОЧК («шестерка»).

Распределение чисел от **11** до **1 000 000** в СМН во многом похоже на их распределение в СД, хотя и имеет характерные отличия.

Во-первых, в поговорках СМН появляются такие числа, как **16**, **17**, **21**, **22**, **23**, **31**, **32**, **38**, **41**, **44**, **45**, **46**, **47**, **55**, **69**, **95**, **99**, **105**, **108**, **120**, **180**, **220**, **250**, **506**, **911**, **30 000**, **35 000**, **40 000**, **100 000** (однако только по одному, реже по два словоупотребления), но не употребляются числа **18**, **25**, **26**, **42**, **72**, **77**, **80**, **90**, **700**, **2000**, **5000**, которые есть в пословицах СД. Таким образом, в целом количество чисел в поговорках увеличивается — в СМН их 62, вместо 45 в СД.

Во-вторых, в поговорках СМН числа **21** (**7** × **3**), **45** (**5** × **3** × **3**), **99** (**11** × **3** × **3**), **108** (**2** × **2** × **3** × **3** × **3** = 12 × **3** × **3**) можно прибавить к ряду чисел со следами двенадцатеричной системы счисления (**3** = **12** × 1 / 4). Тогда числа **80**, **90**, **700**, **2000**, **5000** относятся к десятичной системе счисления. При этом в СМН также прослеживается низкая частота **6** (половины дюжины) и **5** (половины десятка).

Среди числительных в СМН выделяются более высокой частотой **12**, **20** и **40**, а **100** употребляется значительно меньше, чем в СД.

Учет же ОЧК выводит в СМН на первое место (с большим отрывом от других) **100** (за счет «рубля» — здесь, поскольку речь идет о современных реалиях, это не вызывает сомнения — и «века»), позиции **24** (количество часов в сутках — 2 в СД и 149 в СМН) и **365** (количество дней в году — нет в СД и 146 в СМН) почти одинаковы, далее идет **60** (число минут в часе и секунд в минуте), которого нет в СД. За исключением этих четырех концептов суммарное распределение Ч и ОЧК оказывается более равномерным, чем распределение только Ч.

2.2. Числа, выражающие количество

Данные об использовании чисел для передачи количеств в СМН, выражаемых как количественными Ч, так и ОЧК, представлены в табл. 2 и на рис. 3 и 4.

Таблица 2. **Количественные концепты в СМН, выраженные количественными Ч и ОЧК**

Число	Выражение количества								
	Ч			ОЧК			Ч + ОЧК		
	Форм	Употреблений		Форм	Употреблений		Форм	Употреблений	
		Абс.	%		Абс.	%		Абс.	%
1	12	446	43,7	10	70	9,6	22	516	29,4
2	5	192	18,8	0	0	0	5	192	11,0
3	2	123	12,0	2	4	0,5	4	127	7,2
4	2	25	2,4	0	0	0	2	25	1,4
5	2	18	1,8	2	2	0,3	4	20	1,1
6	2	7	0,7	0	0	0	2	7	0,4
7	3	62	6,1	5	31	4,2	8	93	5,3
8	1	4	0,4	0	0	0	1	4	0,2
9	1	5	0,5	0	0	0	1	5	0,3
10	1	15	1,5	4	4	0,5	5	19	1,1
11	1	3	0,3	0	0	0	1	3	0,2
12	2	7	0,7	0	0	0	2	7	0,4
13	0	0	0	1	1	0,1	1	1	0,1
14	1	1	0,1	0	0	0	1	1	0,1
15	1	4	0,4	0	0	0	1	4	0,2
16	1	3	0,3	0	0	0	1	3	0,2
17	1	2	0,2	0	0	0	1	2	0,1
20	1	7	0,7	0	0	0	1	7	0.4
22	1	1	0,1	0	0	0	1	1	0,1
23	1	1	0,1	0	0	0	1	1	0,1
24	1	2	0,2	11	147	20,1	12	149	8,5
30	1	3	0,3	1	6	0,8	2	9	0,5
32	1	1	0,1	0	0	0	1	1	0,1
33	1	1	0,1	0	0	0	1	1	0,1
36	1	2	0,2	0	0	0	1	2	0,1
38	1	2	0,2	0	0	0	1	2	0,1
40	2	10	1,0	0	0	0	2	10	0,6
41	1	1	0,1	0	0	0	1	1	0,1

Окончание табл. 2

Число	Выражение количества								
	Ч			ОЧК			Ч + ОЧК		
	Форм	Употреблений		Форм	Употреблений		Форм	Употреблений	
		Абс.	%		Абс.	%		Абс.	%
44	1	1	0,1	0	0	0	1	1	0,1
45	1	2	0,2	0	0	0	1	2	0,1
46	1	1	0,1	0	0	0	1	1	0,1
47	1	1	0,1	0	0	0	1	1	0,1
50	0	0	0	1	1	0,1	1	1	0,1
52	1	1	0,1	0	0	0	1	1	0,1
55	1	1	0,1	0	0	0	1	1	0,1
60	0	0	0	13	71	9,7	13	71	4,1
69	1	1	0,1	0	0	0	1	1	0,1
70	1	1	0,1	0	0	0	1	1	0,1
95	1	1	0,1	0	0	0	1	1	0,1
99	1	1	0,1	0	0	0	1	1	0,1
100	2	40	3,9	19	247	33,7	21	287	16,4
120	1	1	0,1	0	0	0	1	1	0,1
180	1	1	0,1	0	0	0	1	1	0,1
200	1	3	0,3	0	0	0	1	3	0,2
220	1	1	0,1	0	0	0	1	1	0,1
250	1	1	0,1	0	0	0	1	1	0,1
300	1	1	0,1	0	0	0	1	1	0,1
365	0	0	0	9	146	19,9	9	146	8,3
500	1	1	0,1	0	0	0	1	1	0,1
506	1	1	0,1	0	0	0	1	1	0,1
911	1	1	0,1	0	0	0	1	1	0,1
1 000	3	5	0,5	0	0	0	3	5	0,3
30 000	1	1	0,1	0	0	0	1	1	0,1
35 000	1	1	0,1	0	0	0	1	1	0,1
40 000	1	1	0,1	0	0	0	1	1	0,1
100 000	1	1	0,1	0	0	0	1	1	0,1
1 000 000	1	2	0,2	2	2	0,3	3	4	0,2
Итого	79	1021	100,0	80	732	100,0	159	1753	100,0

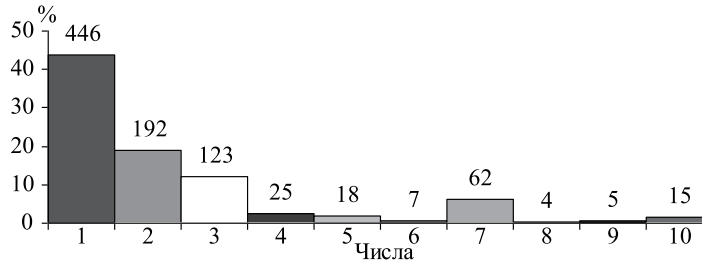

Рис. 3. Количественные концепты от **1** до **10** в СМН, выраженные количественными Ч

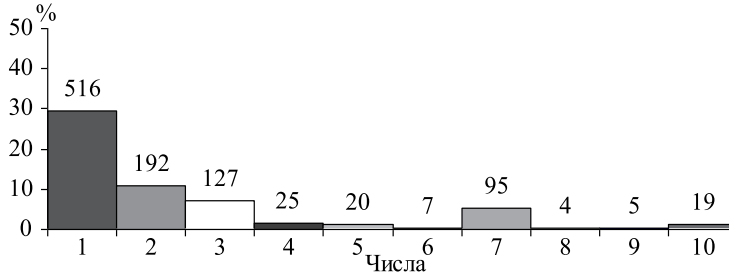

Рис. 4. Количественные концепты от **1** до **10** в СМН, выраженные количественными Ч и ОЧК

Качественно картина не отличается от описанной выше. Самое главное отличие в том, что Ч **1** заметно взлетело по отношению к **2** и **3**.

Числа, используемые для обозначения количества, которое передается с помощью ОЧК, приведены в табл. 3.

Как можно видеть из табл. 3, ОЧК представлены существительными, обозначающими некоторое количество (сотня) или величину, представляемую как количество (год, выраженный в количестве дней, минуты — в количестве секунд, и т. д.). За счет ОЧК меняются количества чисел **24** (149 вместо 2), **60** (71 вместо 0), **100** (287 вместо 40), **365** (146 вместо 0).

Вместе с тем это существительные, которые обозначают самые обычные реалии, что делает такие существительные высокочастотными и долго сохраняющимися, если их директивно выводят из употребления (переход к счету десятками от счета дюжинами в связи

Таблица 3. **Количественные концепты в СМН, выраженные количественными ОЧК**

Число	ОЧК количества	Употреблений	
		Абс.	%
1	копейка, копеечка	70	9,6
3	алтын	4	0,5
5	пятак	2	0,3
7	неделя	31	4,2
10	гривна, деканка	4	0,5
13	чертова дюжина	1	0,1
24	день, сутки, односутки	147	20,0
30	месяц	6	0,8
50	полтина	1	0,1
60	минута, час	71	9,7
100	сотня, век, рубль	247	33,7
365	год	146	20,0
1 000 000	лимон	2	0,3
Итого		732	100,0

с переводом России на метрическую систему мер и весов постановлениями правительства в 1917 и 1925 гг.). Подобное положение дел имеет место и в СД.

2.3. Числа, выражающие порядок

Следующая группа Ч и ОЧК связана с использованием чисел для выражения порядка (табл. 4, рис. 5 и 6).

Приведенные в табл. 4 данные показывают, что для обозначения порядка в поговорках СМН используются все числа, которые есть в СД (**1**–**7**, **9**, **10**, **12**, **13**), а также добавляются **8**, **11**, **16**, **20**, **21**, **23**, **31**–**33**, **101**, **105**, **108**. Характерным отличием распределений порядковых Ч в СД является отсутствие провала на **2**, но сохранение провала на **4** (только для Ч). Для суммарного распределения Ч + ОЧК характерно почти монотонное убывание частоты с ростом числа (с локальными отклонениями для **5**, **7**, **9** и **10**). Следующие за ними числа употребляются в поговорках не более четырех раз. При

Таблица 4. **Порядковые концепты в СМН, выраженные порядковыми Ч и ОЧК**

Число	Выражение порядка								
	Ч			ОЧК			Ч + ОЧК		
	Форм	Употреблений		Форм	Употреблений		Форм	Употреблений	
		Абс.	%		Абс.	%		Абс.	%
1	12	105	33,7	12	59	36,4	24	164	34,6
2	11	54	17,3	10	79	48,8	21	133	28,1
3	10	39	12,5	0	0	0	10	39	8,2
4	0	0	0	3	10	6,2	3	10	2,1
5	6	24	7,7	5	14	8,6	11	38	8,0
6	4	6	1,9	0	0	0	4	6	1,3
7	5	13	4,2	0	0	0	5	13	2,7
8	4	5	1,6	0	0	0	4	5	1,1
9	6	28	9,0	0	0	0	6	28	5,9
10	4	15	4,8	0	0	0	4	15	3,2
11	1	3	1,0	0	0	0	1	3	0,6
12	1	1	0,3	0	0	0	1	1	0,2
13	1	1	0,3	0	0	0	1	1	0,2
16	2	2	0,6	0	0	0	2	2	0,4
20	2	4	1,3	0	0	0	2	4	0,8
21	1	1	0,3	0	0	0	1	1	0,2
23	1	1	0,3	0	0	0	1	1	0,2
31	1	2	0,6	0	0	0	1	2	0,4
32	1	1	0,3	0	0	0	1	1	0,2
33	1	1	0,3	0	0	0	1	1	0,2
101	2	4	1,3	0	0	0	2	4	0,8
105	1	1	0,3	0	0	0	1	1	0,2
108	1	1	0,3	0	0	0	1	1	0,2
Итого	78	312	100,0	30	162	100,0	108	474	100,0

этом **2** почти сравнялось с **1** за счет использования слова «другой» в значении «второй» (так же в СД). Рост частоты концепта «первый» за счет использования «один» в значении «первый» в СМН — незначительный. Частота концепта «пятый» за счет пятницы как пятого дня недели (см. табл. 5) увеличивается в полтора раза. ОЧК

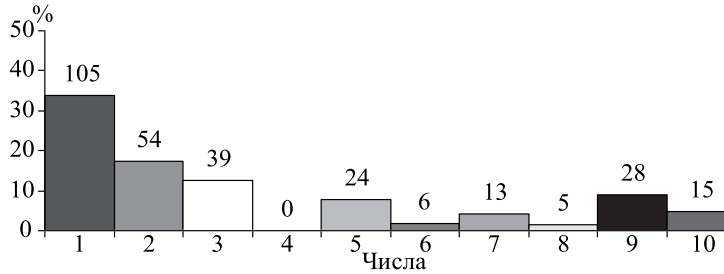

Рис. 5. Порядковые концепты от **1** до **10** в СМН, выраженные порядковыми-Ч

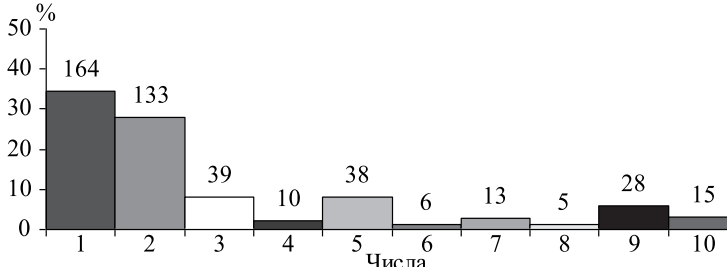

Рис. 6. Порядковые концепты от **1** до **10** в СМН, выраженные порядковыми Ч и ОЧК

Таблица 5. **Порядковые концепты в СМН, выраженные порядковыми ОЧК**

Порядковый концепт	ОЧК	Употреблений	
		Абс.	%
Первый	один (в значении «первый»), понедельник	59	36,4
Второй	другой, вторник	79	48,8
Четвертый	четверг	10	6,2
Пятый	пятница	14	8,6
Итого		162	100,0

в этой группе представлены словами «один» и «другой» в значениях «первый» и «второй» и названиями дней недели.

2.4. Числа, выражающие собирательные числовые концепты

Последняя группа чисел, подвергнутых специальному анализу, — собирательные Ч и ОЧК (табл. 6, рис. 7). Данные представлены для чисел от **2** до **10**. При этом основная их доля приходится на числа **2** (единственный тип числовых концептов, для которых максимум смещен на число 2, но в этой группе это минимальное возможное

Таблица 6. **Собирательные концепты в СМН, выраженные собирательными Ч и ОЧК**

Число	Всего форм	Выражения собирательного концепта							
		Ч			ОЧК			Ч + ОЧК	
		Формы	Употреблений		Формы	Употреблений		Абс.	%
			Абс.	%		Абс.	%		
2	10	оба, обе, обеих, обеими, двое, двоих	34	68,0	пара, паре, двойку, двойке	19	35,2	53	51,0
3	6	трое, троих	3	6,0	тройка, тройке, троица, троицу	4	7,4	7	6,7
4	1	четверо	1	2,0	—	0	0	1	1,0
5	2	пятеро	1	2,0	пятерней	1	1,9	2	1,9
6	1	—	0	0	шестерка	1	1,9	1	1,0
7	6	семеро, семером, семерым, семерых	11	22,0	семерка, семерки	2	3,7	13	12,5
8	3	—	0	0	восьмерка, восьмерку, восьмерки	4	7,4	4	3,8
9	2	—	0	0	девятка, девятку	9	16,7	9	8,7
10	4	—	0	0	десятка, десятку, десяток, десятки	14	25,9	14	13,5
Итого	35		50	100		54	100,0	104	100,0

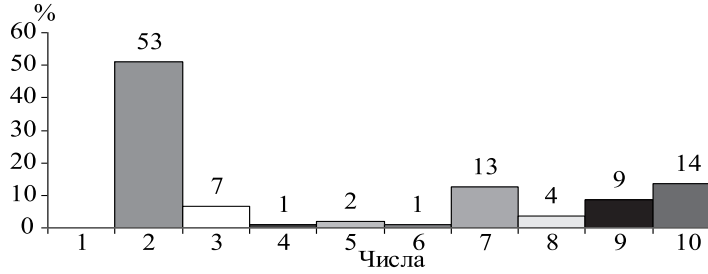

Рис. 7. Собирательные концепты в СМН, выраженные собирательными Ч и ОЧК

число), **10** (за счет ОЧК «десятка») и **7** (как и в большинстве предыдущих распределений).

Следует отметить, что по отношению к трактовке чисел в собирательном значении в поговорках СМН справедливы те же затруднения, которые подробно обсуждены по отношению к ним в СД [Бабарико, Чебанов], а именно — что по крайней мере часть ОЧК («год», «сутки», «час» и т. д.) может интерпретироваться в собирательном смысле. Некоторая неопределенность здесь связана с неразработанностью категории собирательности как в лингвистике, так и в логике [Чебанов], причем, как показано А. В. Степуковой, степень этой неразработанности оказалась выше всех возможных предположений [Степукова, 2014а; Степукова, 2014б; Степукова, 2016; Степукова, Чебанов].

2.5. Сравнение использования чисел в собраниях пословиц и поговорок XIX и XXI веков

Проведенный анализ поговорок СМН показывает некоторые отличия распределения числовых концептов в поговорках СМН от такового для СД.

На первом месте по частоте использования в обоих случаях находится Ч **1**, но его обособленность от других Ч в СМН значительно выше, чем в СД, что может являться показателем увеличения монархичности коллективного сознания. Далее с заметным отрывом идет Ч **2**. Затем идет Ч **3** (тоже с заметным отрывом), далее Ч **7**. Подобное распределение существует и для СД.

Если же учитываются Ч + ОЧК, то картина оказывается более сложной: на первом и втором местах находятся **1** и **2**, третье и четвертое принадлежат ОЧК **100** и Ч **3** (в СД на третьем месте — ОЧК **24**, а Ч **3** — также на четвертом), ОЧК **24** находится на пятом месте и на шестом — ОЧК **365**. Ч **7** в поговорках СМН занимает седьмое место. Однако в пределах первого десятка на число **7** (и для Ч, и для ОЧК) приходится резко выраженный локальный максимум (за исключением порядковых концептов) и вместе с тем мало используются числа **6**, **8** и **9**.

За пределами первого десятка маркированным является Ч **100** (без учета ОЧК). Подобное наблюдается и в пословицах СД. Маркированность чисел **12**, **40** в поговорках СМН, напротив, становится менее выраженной. С учетом ОЧК к числам **12** и **40** прибавляются жестко маркированные **24**, **60**, **100** (маркированность которого резко возрастает) и **365**, если учитывать ОЧК (за которыми стоят значимые реалии). Точно такая же картина имеет место в пословицах СД. Как и в СД, в поговорках СМН использование Ч для обозначения **100** и **1 000 000** определяется желанием не передать точное количество, а обозначить обозримое и необозримое множество предметов. Можно говорить и о том, что предпочтение использования одних чисел диктуется ориентацией на двенадцатеричную систему счисления, а других — на десятичную.

Таким образом, с лингвосоциологической точки зрения структура ИКМ чисел в поговорках СМН может характеризоваться как еще более жестко монархическая (что, по-видимому, является следствием отказа в СССР от принципа разделения властей, распространением лозунгов типа «Вся власть Советам!», максим вроде «Народ и партия едины», «Нерушимый блок коммунистов и беспартийных» и т. п.), чем в СД, а с лингвокультурологической [Воробьев; Маслова] — как ориентированная на предельную целостность [Чебанов, Мартыненко]. Вместе с тем структура числовых концептов СМН оказывается более многокомпонентной, что может трактоваться как показатель ее большего богатства, размытости (разрушения) или переходного состояния.

3. Числовые концепты в русской языковой картине мира (по паремиологическим данным)

Теперь можно рассмотреть общее распределение числовых концептов по обоим паремиологическим сборникам, которые, как представляется, дают репрезентативный образ числовой составляющей русской языковой картины мира.

3.1. Общая характеристика использования чисел

Данные табл. 7 и иллюстрирующие ее рис. 8 и 9 позволяют сделать заключения о характере использования числовых концептов в объединенном корпусе СМН и СД.

Самыми употребляемыми Ч (всех разрядов) являются **1** (1575 употреблений — 36,6% от общего количества Ч), **2** (945 — 21,9%, н. д. 58,5%), **3** (527 — 12,2%, н. д. 70,7%), **7** (294 — 6,8%, н. д. 77,5%), **4** и **100** (по 154 — 3,6%, н. д. 84,7%), **5** (134 — 3,1%, н. д. 87,5%), каждое же из остальных Ч составляет не более 1,4%.

Для ОЧК последовательность оказывается несколько иной: **100** (681 — 23,2%), **24** (640 — 21,8%, н. д. 45%), **365** (378 — 12,9%, н. д. 57,9%), **1** (342 — 11,6%, н. д. 69,5%), **2** (277 — 9,4%, н. д. 78,9%), **60** (194 — 6,6%, н. д. 85,5%), **7** (141 — 4,8%, н. д. 90,3%), **5** (72 — 2,4%, н. д. 92,7%) и далее числа, имеющие частоты ниже 2,4%.

Сравнение динамики этих частот показывает, что распределение употребления чисел отличается очень высокой концентрацией словоупотреблений в самом начале распределения (первые три Ч **1**–**2**–**3** выбирают больше 70% всего объема), при том что первые пять Ч и ОЧК отсекают практически равную долю словоупотреблений для обоих распределений (81,1% и 78,9% соответственно), а для шести первых членов сумма их частот для ОЧК оказывается уже выше, чем для Ч (85,5% и 84,7% соответственно). Последнее обстоятельство связано с тем, что разнообразие обнаруженных ОЧК, оказывается меньше, чем разнообразие Ч.

В последовательность чисел наиболее высокочастотных ОЧК «вклиниваются» в таком же порядке, как у Ч, числа **1** и **2** и, после нескольких пропусков, **7** и **5**.

Примечательно, что совместное распределение Ч в СМН и СД совпадает с распределением Ч в СД.

Таблица 7. **Числовые концепты в СД и СМН, выраженные Ч и ОЧК**

Число	Выражение числа					
	Ч		ОЧК		Ч + ОЧК	
	Употреблений		Употреблений		Употреблений	
	Абс.	%	Абс.	%	Абс.	%
1	1575	36,6	342	11,6	1917	26,5
2	945	21,9	277	9,4	1222	16,9
3	527	12,2	57	1,9	584	8,1
4	154	3,6	50	1,7	204	2,8
5	134	3,1	72	2,4	206	2,8
6	54	1,3	3	0,1	57	0,8
7	294	6,8	141	4,8	435	6,0
8	20	0,5	4	0,1	24	0,3
9	62	1,4	9	0,3	71	1,0
10	71	1,6	40	1,3	111	1,5
11	8	0,2	0	0	8	0,1
12	34	0,8	4	0,1	38	0,5
13	8	0,2	1	0,0	9	0,1
14	3	0,1	0	0	3	0,0
15	14	0,3	0	0	14	0,2
16	5	0,1	0	0	5	0,1
17	2	0,0	0	0	2	0,0
18	2	0,0	0	0	2	0,0
20	22	0,5	0	0	22	0,3
21	1	0,0	0	0	1	0,0
22	1	0,0	0	0	1	0,0
23	2	0,0	0	0	2	0,0
24	2	0,0	640	21,8	642	8,9
25	2	0,0	0	0	2	0,0
26	1	0,0	0	0	1	0,0
30	23	0,5	26	0,9	49	0,7
31	1	0,0	0	0	1	0,0
32	2	0,0	0	0	2	0,0
33	7	0,2	0	0	7	0,1
36	4	0,1	0	0	4	0,1
38	2	0,0	0	0	2	0,0
40	52	1,2	0	0	52	0,7
41	1	0,0	0	0	1	0,0
42	4	0,1	0	0	4	0,1
44	1	0,0	0	0	1	0,0
45	2	0,0	0	0	2	0,0

Окончание табл. 7

46	1	0,0	0	0	1	0,0
47	2	0,0	0	0	2	0,0
50	10	0,2	11	0,4	21	0,3
52	3	0,1	0	0	3	0,0
55	1	0,0	0	0	1	0,0
60	2	0,0	194	6,6	196	2,7
69	1	0,0	0	0	1	0,0
70	9	0,2	0	0	9	0,1
72	2	0,0	0	0	2	0,0
77	4	0,1	0	0	4	0,1
80	1	0,0	0	0	1	0,0
90	11	0,3	0	0	11	0,2
95	1	0,0	0	0	1	0,0
99	1	0,0	0	0	1	0,0
100	154	3,6	681	23,2	835	11,5
101	6	0,1	0	0	6	0,1
105	1	0,0	0	0	1	0,0
108	1	0,0	0	0	1	0,0
120	1	0,0	0	0	1	0,0
180	1	0,0	0	0	1	0,0
200	13	0,3	0	0	13	0,2
220	1	0,0	0	0	1	0,0
250	1	0,0	0	0	1	0,0
300	7	0,2	0	0	7	0,1
365	1	0,0	378	12,9	379	5,2
500	9	0,2	0	0	9	0,1
506	1	0,0	0	0	1	0,0
700	4	0,1	0	0	4	0,1
911	1	0,0	0	0	1	0,0
1000	5	0,1	0	0	5	0,1
2000	1	0,0	0	0	1	0,0
5000	1	0,0	0	0	1	0,0
30 000	1	0,0	0	0	1	0,0
35 000	1	0,0	0	0	1	0,0
40 000	1	0,0	0	0	1	0,0
100 000	1	0,0	0	0	1	0,0
1 000 000	4	0,1	2	0,0	6	0,1
Итого	4308	100,0	2932	100,0	7240	100,0

Примечание. Подчеркиванием выделены числа, которые есть в СМН, но отсутствуют в СД. Курсивом выделены числа, которые есть в СД, но отсутствуют в СМН.

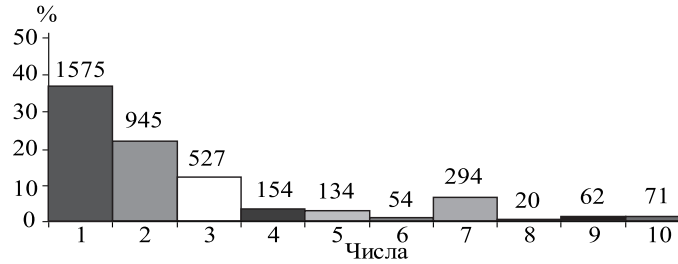

Рис. 8. Числовые концепты от **1** до **10** в СД и СМН, выраженные Ч

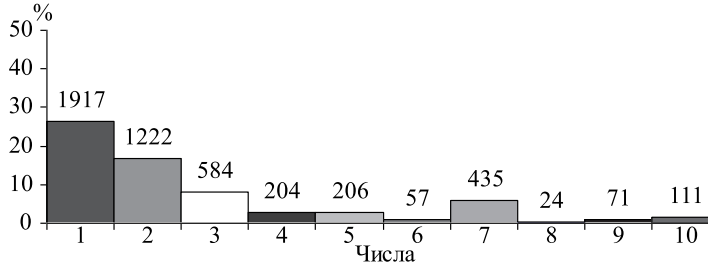

Рис. 9. Числовые концепты от **1** до **10** в СД и СМН, выраженные Ч и ОЧК

Суммарное распределение Ч + ОЧК по объединенным СД и СМН частично (а в самом начале и полностью) восстанавливает последовательность, характерную для чисел (но с меньшими долями при возрастании абсолютного количества) в СД, — самыми частыми оказываются **1** (1917 — 26,5%) и **2** (1222 — 16,9%, н. д. 43,4%), далее **100** (835 — 11,5%, н. д. 54,9%) и **24** (642 — 8,9%, н. д. 63,8%) из числа ОЧК, затем **3** (584 — 8,1%, н. д. 71,9%) и **7** (435 — 6,0%, н. д. 77,9%), так же как и в ряду числительных, далее **365** (379 — 5,2%) из ОЧК. Высокую долю описательных числовых концептов **100** и **24** (как и менее высокочастотных **60** и **365**) можно связать с явлением, аналогичным лексикализации устойчивых словосочетаний.

Крайне важным и неожиданным результатом, полученным на этом материале, является то, что число **3** не оказывается наиболее частым ни среди употребляемых Ч (третье по частоте), ни среди ОЧК (девятое по частоте), занимая в сумме (Ч + ОЧК) пятое место, причем число **2** имеет бóльшую частоту как среди Ч, так и среди ОЧК (пятое по частоте), занимая в сумме второе место. При этом

в XX веке имеет место незначительное увеличение активности числа **3**, что связывается как с распространением тройственной риторической формулы, представленной в низовой народной культуре (например, в трехходовой структуре волшебной сказки [Пропп]), так и с возрастанием потребности в троичных построениях в науке [Баранцев].

Количественные концепты первого десятка охватывают большую часть всех высокочастотных концептов (рис. 8 и 9). И для Ч, и для ОЧК имеет место один и тот же характер распределения: от **1** до **6** имеет место монотонное падение частотности (резко выраженное для первых четырех чисел, от **1** до **4**), на **7** приходится довольно острый локальный максимум (правда, несколько меньший частоты **3**), на **8** — абсолютный минимум (!), а далее на **9** и **10** идет незначительный рост, более выраженный для ОЧК. Для суммы Ч + ОЧК имеет место такое же распределение, за исключением **4** (употребляется незначительно реже **5**).

Распределение числовых концептов от **11** до **1 000 000** имеет более сложный характер.

Всего выявлено 61 число от **11** до **1 000 000**.

Среди них можно выделить два ряда чисел (выраженных как Ч, так и ОЧК). Первый ряд составляют числа **12**, **24**, **30**, **36**, **72**, **90**, **365** (= **360** — отождествление, выявленное Н. Коперником [Коперник]), которые представляют собой следы двенадцатеричной системы счисления (**30** — половина от **60** = **12** × **5**; **90** = **60** + **30**), второй ряд — числа **20**, (**30**), **40**, **50**, (**60**), **80**, (**90**), **100**, **200**, **300**, **500**, **700**, **1000**, **2000**, **5000**, **100 000**, являющиеся маркированными представителями десятичной системы счисления (в скобках указаны числа, допускающие двойную трактовку). Обращает на себя внимание низкая частота **6** (половины дюжины) и **5** (половины десятка). При этом числа **24**, **60** и **365** имеют значительно большую частоту среди ОЧК, что опять же может рассматриваться как проявление важного и более древнего концепта в лексикализированной форме.

Среди числительных выделяются высокой частотой **100** и **40**. Число **100** при этом обозначает скорее большое количество, которое можно зрительно представить, а **1 000 000** — необозримо большое число, за которым не стоит указание конкретного количества.

Учет же ОЧК выводит на первые два места (с большим отрывом от других) **100** (за счет слов «рубль» и «век») и **24** (количество

часов в сутках), за ними идут **365** (количество дней в году) и **60** (число минут в часе и секунд в минуте). За исключением этих четырех концептов суммарное распределение Ч + ОЧК оказывается более равномерным, чем распределение числительных.

В целом выявляется очень четкая и устойчивая во времени структура числовых концептов русской языковой картины мира.

Во-первых, в ней выделяется ядро, представленное числами (выраженными Ч разных разрядов и ОЧК) **1**, **2**, **3**, **7**.

Во-вторых, к ядру примыкают и ОЧК **24**, **60**, **100**, **365**.

В-третьих, некоторую периферию составляют числа **4**, **5**, **10**, **40**.

Остальные числа встречаются с низкой частотой, причем большинство из них — однократно.

Теперь более детально рассмотрим отдельные группы числовых концептов в общем массиве СД и СМН.

3.2. Числа, выражающие количество

Распределение Ч и ОЧК, связанных с использованием чисел для обозначения количества, представлено в табл. 8 и на рис. 10 и 11.

Таблица 8. **Количественные концепты в СД и СМН, выраженные количественными Ч и ОЧК**

Число	Выражение количества					
	Ч		ОЧК		Ч + ОЧК	
	Употреблений		Употреблений		Употреблений	
	Абс.	%	Абс.	%	Абс.	%
1	1270	39,1	126	5,5	1396	25,3
2	700	21,5	0	0	700	12,7
3	381	11,7	46	2,0	427	7,7
4	42	1,3	0	0	42	0,8
5	95	2,9	13	0,5	108	2,0
6	39	1,2	0	0	49	0,7
7	209	6,4	139	6,1	348	6,3
8	15	0,5	0	0	15	0,3
9	20	0,6	0	0	20	0,4
10	46	1,4	16	0,7	62	1,1

Продолжение табл. 8

Число	Выражение количества					
	Ч		ОЧК		Ч + ОЧК	
	Употреблений		Употреблений		Употреблений	
	Абс.	%	Абс.	%	Абс.	%
11	5	0,2	0	0	5	0,1
12	28	0,9	4	0,2	32	0,6
13	5	0,2	1	0,0	6	0,1
14	3	0,1	0	0	3	0,1
15	3	0,1	0	0	3	0,1
16	3	0,1	0	0	3	0,1
17	2	0,1	0	0	2	0,0
18	2	0,1	0	0	2	0,0
20	18	0,6	0	0	18	0,3
22	1	0,0	0	0	1	0,0
23	1	0,0	0	0	1	0,0
24	2	0,1	640	28,1	642	11,6
25	2	0,1	0	0	2	0,0
26	1	0,0	0	0	1	0,0
30	23	0,7	26	1,1	49	0,9
32	1	0,0	0	0	1	0,0
33	6	0,2	0	0	6	0,1
36	4	0,1	0	0	4	0,1
38	2	0,1	0	0	2	0,0
40	52	1,6	0	0	52	0,9
41	1	0,0	0	0	1	0,0
42	4	0,1	0	0	4	0,1
44	1	0,0	0	0	1	0,0
45	2	0,1	0	0	2	0,0
46	1	0,0	0	0	1	0,0
47	2	0,1	0	0	2	0,0
50	10	0,3	11	0,5	21	0,4
52	3	0,1	0	0	3	0,1
55	1	0,0	0	0	1	0,0
60	2	0,1	194	8,5	196	3,5

Окончание табл. 8

Число	Выражение количества					
	Ч		ОЧК		Ч + ОЧК	
	Употреблений		Употреблений		Употреблений	
	Абс.	%	Абс.	%	Абс.	%
69	1	0,0	0	0	1	0,0
70	9	0,3	0	0	9	0,2
72	2	0,1	0	0	2	0,0
77	4	0,1	0	0	4	0,1
80	1	0,0	0	0	1	0,0
90	11	0,3	0	0	11	0,2
95	1	0,0	0	0	1	0,0
99	1	0,0	0	0	1	0,0
100	154	4,7	681	30,0	835	15,1
101	2	0,1	0	0	2	0,0
105	0	0,0	0	0	0	0,0
108	0	0,0	0	0	0	0,0
120	1	0,0	0	0	1	0,0
180	1	0,0	0	0	1	0,0
200	3	0,1	0	0	3	0,1
220	1	0,0	0	0	1	0,0
250	1	0,0	0	0	1	0,0
300	7	0,2	0	0	7	0,1
365	1	0,0	378	16,6	379	6,9
500	9	0,3	0	0	9	0,2
506	1	0,0	0	0	1	0,0
700	4	0,1	0	0	4	0,1
911	1	0,0	0	0	1	0,0
1000	15	0,5	0	0	15	0,3
2000	1	0,0	0	0	1	0,0
5000	1	0,0	0	0	1	0,0
30 000	1	0,0	0	0	1	0,0
35 000	1	0,0	0	0	1	0,0
40 000	1	0,0	0	0	1	0,0
100 000	1	0,0	0	0	1	0,0
1 000 000	4	0,1	2	0,1	6	0,1
Итого	3249	100,0	2277	100	5526	100,0

Примечание. Подчеркиванием выделены числа, которые есть в СМН и отсутствуют в СД. Курсивом выделены числа, которые есть в СД и отсутствуют в СМН.

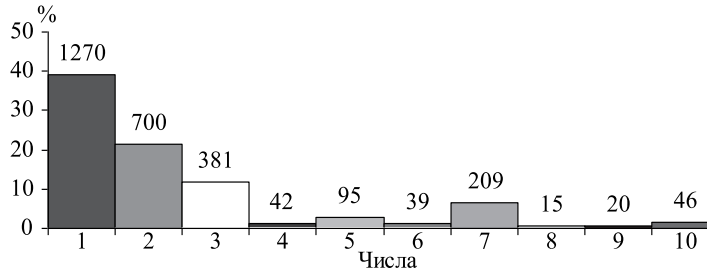

Рис. 10. Количественные концепты от **1** до **10** в СД и СМН, выраженные количественными Ч

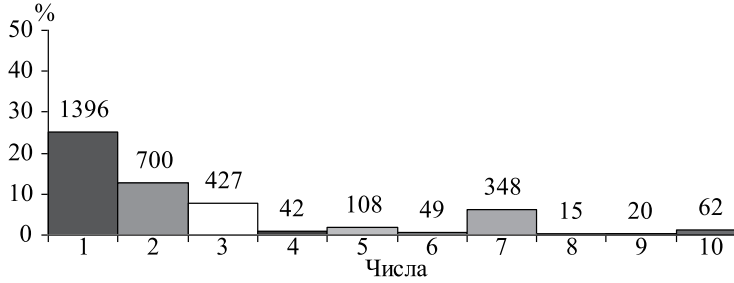

Рис. 11. Количественные концепты от **1** до **10** в СД и СМН, выраженные количественными Ч и ОЧК

Качественно картина распределения количественных концептов, выраженых как количественными числительными, так и ОЧК, передающими количества, не отличается от описанной в разделе 3.1, поскольку числа в пословицах используются в первую очередь именно для передачи количества. Самое большое отличие в том, что частота числа **5** для распределений как Ч, так и ОЧК становится больше частоты числа **4**. Некоторые отличия имеются также в соотношении частот отдельных выражений тех или иных чисел. Частота числа **7**, выраженного Ч и ОЧК, приближается к частоте числа **3**. Обращает на себя внимание также отсутсвие ОЧК для числа **4**. Частоты чисел больше **10** могут незначительно отличаться от суммарного распределения в области низкочастотных чисел.

Как можно видеть из табл. 9, ОЧК количеств объединенного СД и СМН сохраняют их характеристики, приведенные в разделе 2.2.

Таблица 9. **Количественные концепты в СД и СМН, выраженные количественными ОЧК**

Число	ОЧК	Употреблений	
		Абс.	%
1	копейка, копеечка	126	5,5
3	алтын, алтынник	46	2,0
5	пятак	13	0,6
7	неделя, седмица	139	6,1
10	гривна, деканка	16	0,7
12	дюжина	4	0,2
13	чертова дюжина	1	0,0
24	день, сутки, односутки	640	28,1
30	месяц	26	1,1
50	полтина	11	0,4
60	минута, час	194	8,5
100	сотня, век, рубль	681	29,9
365	год	378	16,6
1 000 000	лимон	2	0,1
Итого		2277	100,0

3.3. Числа, выражающие порядок

Следующая группа Ч и ОЧК связана с использованием чисел для выражения порядка (табл. 10 и 11, рис. 12 и 13).

Таблица 10. **Порядковые концепты в СД и СМН, выраженные порядковыми Ч и ОЧК**

Число	Выражение порядка					
	Ч		ОЧК		Ч + ОЧК	
	Употреблений		Употреблений		Употреблений	
	Абс.	%	Абс.	%	Абс.	%
1	305	44,1	216	39,5	521	41,2
2	86	12,4	249	45,6	335	26,5
3	129	18,7	0	0	129	10,2
4	7	1,0	50	9,1	57	4,5

Окончание табл. 10

Число	Выражение порядка					
	Ч		ОЧК		Ч + ОЧК	
	Употреблений		Употреблений		Употреблений	
	Абс.	%	Абс.	%	Абс.	%
5	34	4,9	58	5,6	92	7,3
6	15	2,2	0	0	15	1,2
7	23	3,3	0	0	23	1,8
8	5	0,7	0	0	5	0,4
9	42	6,1	0	0	42	3,3
10	20	2,9	0	0	20	1,6
11	3	0,4	0	0	3	0,2
12	1	0,1	0	0	1	0,1
13	3	0,4	0	0	3	0,2
16	2	0,3	0	0	2	0,2
20	4	0,6	0	0	4	0,3
21	1	0,1	0	0	1	0,1
23	1	0,1	0	0	1	0,1
31	2	0,3	0	0	2	0,2
32	1	0,1	0	0	1	0,1
33	1	0,1	0	0	1	0,1
101	4	0,6	0	0	4	0,3
105	1	0,1	0	0	1	0,1
108	1	0,1	0	0	1	0,1
Итого	691	100,0	573	100,0	1264	100,0

Приведенные данные показывают, что для обозначения порядка в пословицах и поговорках обоих собраний используются числа **1**–**13**, а также добавляются **16**, **20**, **21**, **23**, **31**–**33**, **101**, **105**, **108**. Характерное отличие от ранее рассмотренных распределений Ч — провал на **2** и **4**. Для Ч и ОЧК (табл. 10) характерно почти монотонное убывание частоты с ростом числа (с локальными отклонениями для **5**, **7**, **9** и **10**). Последующие числа употребляются в обоих собраниях не более четырех раз. При этом **2** почти сравнивается с **1** за счет использования слова «другой» в значении «второй» (в основном за счет вклада СД). Увеличение же частоты «первого» за счет

Таблица 11. **Порядковые концепты в СД и СМН, выраженные порядковыми ОЧК**

Порядковый концепт	ОЧК	Употреблений	
		Абс	%
Первый	один (в значении «первый»), понедельник	216	39,5
Второй	другой, вторник	249	45,6
Четвертый	четверг	50	9,1
Пятый	пятница	58	5,6
Итого		573	100,0

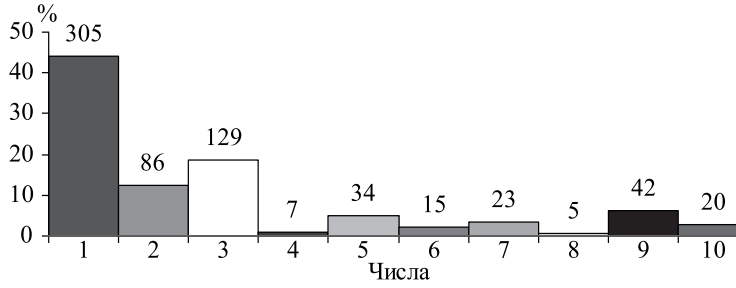

Рис. 12. Порядковые концепты от **1** до **10** в СД и СМН, выраженные порядковыми Ч

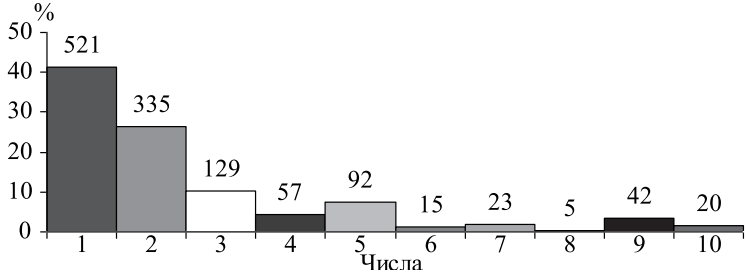

Рис. 13. Порядковые концепты от **1** до **10** в СД и СМН, выраженные порядковыми Ч и ОЧК

использования «один» в значении «первый» в обоих собраниях незначительно. «Пятый» за счет пятницы как пятого дня недели увеличивается в полтора раза.

Таким образом, ОЧК, выражающие порядок, в этом случае представлены словами «один» и «другой» в значениях «первый»

и «второй» и названиями дней недели, причем они относятся только к первым пяти числам.

3.4. Числа, выражающие собирательные числовые концепты

Последняя группа чисел, подвергнутых специальному анализу, — собирательные числовые концепты (табл. 12, рис. 14). Они представлены для чисел от **2** до **10**. При этом основная их доля приходится на **2** (единственный тип чисел, для которых максимум смещен на

*Таблица 1*2. **Собирательные концепты в СД и СМН, выраженные собирательными Ч и ОЧК**

Число	Всего форм	Выражения собирательного концепта							
		Ч			ОЧК			Ч + ОЧК	
		Формы	Употреблений		Формы	Употреблений		Употреблений	
			Абс.	%		Абс.	%	Абс.	%
2	13	оба, обоих, обоим, обе, обеих, обеими, двое, двоих, двоим	159	62,8	пара, паре, двойку, двойке	28	38,4	187	57,4
3	8	трое, троих	17	6,7	тройка, тройке, троица, троицу, троице, троицы	11	15,1	28	8,6
4	1	четверо	5	2,0	—	—	-	5	1,5
5	5	пятеро, пятерых, пятерым	5	2,0	пятерней, пятернею	3	4,1	8	2,5
6	1	—	0	0	шестерка	2	2,7	2	0,6
7	6	семеро, всемером, семерым, семерых	62	24,5	семерка, семерки	2	2,7	64	19,6

Окончабние табл. 12

Число	Всего форм	Выражения собирательного концепта							
		Ч			ОЧК			Ч + ОЧК	
		Формы	Употреблений		Формы	Употреблений		Употреблений	
			Абс.	%		Абс.	%	Абс.	%
8	3	—	0	0	восьмерка, восьмерку, восьмерки	4	5,5	4	1,2
9	2	—	0	0	девятка, девятку	9	12,3	9	2,8
10	6	десятеро, десятерых	5	2,0	десятка, десятку, десяток, десятки	14	19,2	19	5,8
Итого	35		253	100		73	100	326	100

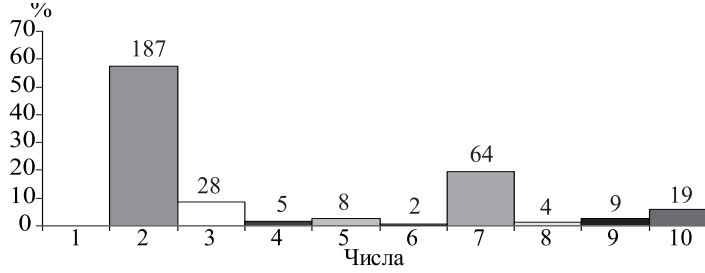

Рис. 14. Собирательные концепты в СД и СМН, выраженные собирательными Ч и ОЧК

число **2**; но оно здесь является минимально допустимым) и **7** (как и в большинстве предыдущих распределений) при несопоставимо меньших частотах остальных концептов.

4. Выводы. Числа в русских паремиях как часть языковой картины мира

Проведенный анализ показывает довольно определенную картину распределения числовых концептов в СМН и СД.

На первом месте по частоте использования чисел находится Ч **1**. Далее с заметным отрывом идет Ч **2** и затем Ч **3** (тоже с заметным отрывом), далее Ч **7**.

Если же учитываются числа, выраженные и Ч, и ОЧК, то картина оказывается более сложной: на первом и втором местах остаются Ч **1** и Ч **2**, третье и четвертое занимают ОЧК **100** и **24**, и только на пятом месте находятся Ч **3**, и на шестом — Ч **7**, причем в пределах первого десятка числу **7** соответствует резко выраженный локальный максимум (за исключением порядковых ОЧК). При этом из первого десятка мало используются выраженные любым способом числа **6, 8** и **9**. Показательно то, что самым частым числом, выраженным любыми средствами (кроме собирательных концептов), является **1**, а для собирательных — **2**, то есть в любом случае минимально возможное в своем разряде. Такое положение дел может интерпретироваться как проявление экономии мышления, своего рода бытовое выражение принципа «бритвы Оккама». В итоге последовательность наиболее частых чисел в суммарном распределении Ч + ОЧК оказывается такой: **1**, **2**, **100**, **24**, **3**, **7**, **365**, **5**, **4**, **60**, **10**.

За пределами первого десятка маркированными являются выраженные Ч **12**, **40** и **100** (без учета ОЧК), к которым прибавляются жестко маркированные ОЧК (за ними стоят значимые реалии) **24**, **60**, **100** (чья маркированность резко возрастает в СМН) и **365**. При этом использование Ч для обозначения **100** и **1 000 000** определяется желанием не передать точное количество, а обозначить обозримое и необозримое множество предметов. Кроме того, можно говорить о том, что предпочтение использования одних чисел диктуется ориентацией на двенадцатеричную систему счисления, а других — на десятичную.

Для уточнения сформулированных выводов о числовой структуре национального языкового сознания могут быть использованы другие фольклорные источники (песни, сказки), что крайне важно для понимания того, какие социальные структуры возможны в русской культуре [Найшуль, Чебанов].

Другим направлением исследований могут быть исследования структуры числовых ИКМ в языковом сознании народов с развитой правовой культурой — римлян, французов, англичан и американцев — и многовековыми юридическими традициями, например китайцев, где значительно выше, чем в русском, частота чисел **10**, **100**, **8**, **9** [Babariko, Jinfeng, Chebanov]. В структуре числовых ИКМ таких языковых культур можно ожидать наличие предпосылок для эффективной реализации принципа разделения властей.

Литература

Бабарико М. Н., Чебанов С. В. Арифмология русских пословиц и поговорок // Структурная и прикладная лингвистика: Межвуз. сб. СПб., 2014. Вып. 10. С. 70–91.

Баранцев Р. Г. Вешки интереса. М.-Ижевск, 2010.

Воробьев В. В. Лингвокультурология: теория и методы. М., 1997.

Даль В. И. Пословицы русского народа. М, 1957.

Коперник Н. О вращениях небесных сфер. Малый комментарий. Послание против Вернера. Упсальская запись. М., 1964.

Лакофф Дж. Женщины, огонь и опасные вещи. Что категории языка говорят нам о мышлении? М., 2004.

Маслова В. А. Лингвокультурология. М., 2001.

Мокиенко В. М., Никитина Т. Г. Большой словарь русских поговорок. М., 2007.

Найшуль В., Чебанов С. Социальная метадисциплина — формальная институционалистика: Программа исследований // Русский журнал. 17.09.09. URL: http://www.russ.ru/pole/Social-naya-metadisciplina-formal-naya-institucionalistika (дата обращения: 26.01.2016).

Никольский Л. Б. О предмете социолингвистики // Вопросы языкознания. 1974. №1. С. 60–67.

Пропп В. Я. Морфология волшебной сказки. М., 2001.

Степукова А. В. Монотипические структуры категорий разного типа // Структурная и прикладная лингвистика: Межвуз. сб. СПб., 2014а. Вып. 10. С. 101–105.

Степукова А. В. Собирательность, разделительность и смежные категории в лингвистике и логике // Актуальные проблемы современной когнитивной науки. Материалы седьмой всероссийской научно-практической конференции с международным участием (16–18 октября 2014 г.). Иваново, 2014б. С. 101.

Степукова А. В. Собирательность в логике и лингвистике // Научный диалог. Екатеринбург, 2016 № 2 (50) (в печати).

Степукова А. В., Чебанов С. В. К формальной характеристике типов категорий // Актуальные проблемы современной когнитивной науки. Материалы шестой всероссийской научно-практической конференции с международным участием (17–19 октября 2013 г.). Иваново, 2013. С. 280–281.

Чебанов С. В. Четырехчленные схемы различения // Актуальные проблемы современной когнитивной науки. Материалы пятой всероссийской научно-практической конференции с международным участием (18–20 октября 2012 г.). Иваново, 2012. С. 204–224.

Чебанов С. В., Мартыненко Г. Я. Синергетика и холистические образы языка // Международная научно-практическая конференция «Рериховское наследие». Т. IV: Охрана культурных ценностей: петербургские традиции. СПб., 2009. С. 404–410.

Швейцер А. Д., Никольский Л. Б. Введение в социолингвистику. М., 1978.

Babariko M., Jinfeng L., Chebanov S. Idealized cognitive model (ICM) of numbers in the Chinese (C) and Russian (R) linguistic world picture (LWP) as a basis of conceptual mapping // 3th International Congress of Numanities (ICoN 2016). Processes, maps, narratives. Program and abstracts. Kaunas, 2016 (in press).

Е. О. Косарева, Г. Я. Мартыненко

ОТНОШЕНИЕ ТЕКСТ — СЛОВАРЬ В ПОВСЕДНЕВНОЙ УСТНОЙ РЕЧИ[1]

Аннотация. В статье обсуждается зависимость размера словаря повседневной речи от числа расшифрованных звукозаписей (объема выборки) на материале корпуса «Один речевой день». Построены две динамические аналитические модели: одна в лингвистике является традиционной (функция Вейбулла), вторая — замедленная логистическая (функция Хауштайна) используется в научно-техническом прогнозировании. Установлено, что зависимость текст — словарь более адекватно описывает функция Хауштайна. Модель показала, что потенциальный словарь коллективной устной речи по своему объему не уступает индивидуальному словарю выдающихся русских писателей — словарю А. П. Чехова, Л. Н. Андреева, А. И. Куприна. Весь объем словаря вычерпывается при объеме выборки, равном 10 млн словоупотреблений.

Ключевые слова. Отношение текст — словарь, объем словаря, лексическое богатство, аналитическое моделирование, функция Вейбулла, функция Хауштайна, аппроксимация, асимптотический уровень, прогнозирование, повседневная речь, русский язык, корпус «Один речевой день».

Ekaterina O. Kosareva, Gregory Ya. Martynenko

THE TYPE-TOKEN RATIO IN EVERYDAY SPOKEN RUSSIAN

Abstract. The paper focuses on the dependence of the vocabulary size of everyday Russian speech on the number of transcribed recordings (sample size). The study is made on the material of the „One Day of Speech" corpus known as the ORD corpus. Two dynamic analytical models are analyzed: the Weibull function, which is traditionally used in linguistics, and the slow logistics Haustein function, which is usually used in scientific and technological forecasting. It was revealed that the Haustein function describes this dependence more adequately. The model showed that the potential of collective vocabulary

[1] Исследование выполнено при поддержке гранта РНФ № 14-18-02070 «Русский язык повседневного общения: особенности функционирования в разных социальных группах».

size of everyday speech is comparable with that of prominent Russian writers — individual vocabularies of Anton Chekhov, Leonid Andreev, Alexander Kuprin. It is expected that the potential vocabulary size may be reached on the sample size of 10 million words.

Keywords. Type-token ratio, vocabulary size, lexical richness, analytical modeling, Weibull function, Haustein function, approximation, the asymptotic level, forecasting, everyday speech, the Russian language, the ORD corpus.

Введение

В последней четверти XX века в отечественной квантитативной лингвистике и смежных областях наблюдался значительный интерес к влиянию размера выборки (или текста) на объем словаря [Нешитой; Тулдава; Горькова]. В западной литературе эта зависимость получила название *type-token ratio*. Это соотношение ввел в научный оборот, по-видимому, Г. Хердан, посвятив ему одну из своих книг — «Type-token Mathematics» [Herdan]. Позднее эта зависимость исследовалась в рамках оценки лексического богатства текста [Wimmer]. Любопытно, что лингвистический интерес исследователей концентрировался только на отношении текст — словарь, хотя в динамике частотного словаря много других структурных характеристик, не менее интересных. Это, видимо, связано с предельной простотой, понятностью объема словаря как характеристики текста или корпуса.

В данной статье ставится задача исследования роста объема словаря в зависимости от числа расшифрованных звукозаписей (объема выборки) на материале корпуса «Один речевой день». Аналогичные исследования нам не известны.

1. Об аналитической зависимости объема словаря от объема выборки

Вопрос о статистической состоятельности стилистических характеристик частотного словаря был поставлен в работе [Мартыненко]. Автор пытался экспериментально проверить гипотезу о сходимости статистик частотных списков (в том числе и их объема) к предельным величинам при увеличении объема выборки. Эта гипотеза была подтверждена на материале частотного словаря по радиоэлектронике, миллионного словаря английского языка и частотного словаря слов-поэтизмов на «без» с помощью функции Вейбулла в облике функции распределения. Годом позднее эта гипотеза была доказана

математически через обращение к предельным теоремам теории вероятностей [Мартыненко, Фомин].

Функция Вейбулла — функция экспоненциального типа. Поэтому она достаточно быстро устремляется к своей асимптоте, то есть к предельному объему словаря в каждом конкретном случае. Такой подход позволил попутно предсказать объем словаря при любом объеме выборки.

Позднее на представительной совокупности рассказов А. П. Чехова связь между объемом словаря и объемом текста была исследована в работе [Гребенников]. В качестве аппроксимирующей модели также была использована функция Вейбулла. Результаты были обнадеживающими. Модель показала удовлетворительную эмпирическую сходимость. Но использовался только словарь словоформ.

В дальнейшем были построены частотные словари рассказов А. П. Чехова, Л. Н. Андреева и А. И. Куприна [Частотный словарь … , 1999; Частотный словарь … , 2003; Частотный словарь … , 2006] и одновременно исследована зависимость объема словаря от объема текста. Для трех словарей были получены следующие асимптотические уровни: для Чехова — 17 тыс., для Куприна — около 39 тыс., для Андреева — около 42 тыс. слов. Это, конечно, не реально наблюдаемые объемы, а теоретические, математические объемы, предсказываемые моделью при допущении, что творческая биография писателя бесконечна. Настораживает, правда, бедность чеховского потенциального словаря на фоне словарей Андреева и Куприна. Но эта бедность иллюзорна, так как зависимость текст — словарь для Чехова была построена с учетом эволюции творчества писателя, в которой отражен переход от кратких пестрых юморесок к глубоким психологическим полотнам, то есть в исследуемой зависимости незримо присутствуют как бы два писателя: Чехонте и Чехов. Это приводит к резкому замедлению роста словаря во второй половине творчества писателя.

Спустя несколько лет в работе [Мартыненко, Мартинович] эта методика была использована для описания статистической организации ассоциативных словарей, построенных по результатам ассоциативного эксперимента.

Предложенный экспериментальный материал был очень компактным, обозримым. Это не громада частотного словаря писателя, это набор сравнительно «коротких» словарей, обладающих, тем не

менее, всеми свойствами классических словарей, кроме невероятной громоздкости последних. Словарь ассоциатов создавал благоприятные условия для усовершенствования методики. Соавторы решили не ограничиваться законом Вейбулла, а в сложившихся благоприятных условиях поэкспериментировать с другими аналитическими функциями. Функций-претендентов было много. Среди них были отобраны те, которые наилучшим образом согласовывались с опытными данными. Таких функций оказалось только две: ранее упомянутая функция Вейбулла и степенной вариант логистической функции (функция Хауштайна), заимствованный из работы по прогнозированию [Haustein]. Результаты моделирования оказались весьма любопытными. Обе функции обнаружили практически полное совпадение результатов моделирования. Они начинали «разбегаться» далеко за пределами данных опыта, устремляясь к своим индивидуальным асимптотам. Итог исследования оказался интересным, но основной вопрос «повис»: какой предел ближе к истине и какая функция более адекватна исследуемому материалу в широком смысле, то есть не только в словаре ассоциатов? Ситуация осложнялась еще и тем, что выбор коэффициентов прогностической функции зависит от объема реально исследованного материала, то есть от прогностической базы: чем больше выборка, тем точнее прогноз. Это верно. Но не только это. Оказалось, что чем больше прогностическая база, тем выше идет прогностическая кривая и тем выше асимптотический уровень.

Итак, были получены две конкурирующие функции. При этом, однако, было не совсем ясно, как они будут вести себя при работе с большими словарями.

Но здесь на помощь поспешили москвичи. Подсказка пришла от корпусников.

На последней конференции по корпусной лингвистике в Санкт-Петербурге прозвучал интересный доклад, посвященный результатам построения в лаборатории А. А. Поликарпова корпуса художественных произведений А. П. Чехова. Объем корпуса (рассказы, повести, драмы) составил 1 381 000 словоупотреблений, которому соответствовало после лемматизации 36 153 лексем [Частотный грамматико-семантический словарь…; Потемкин]. Такое число лексических единиц вступало в противоречие с результатами нашего моделирования с помощью функции Вейбулла. Оказалось, что

у Чехова довольно богатый словарь. Мы воспользовались этим «подарком» и с помощью логистической функции сосчитали объем словаря при таком объеме наблюдения (1 381 000 словоупотреблений). Он оказался равным 33 642, совсем ненамного уступая объему корпуса Поликарпова. Если из объема корпуса вычесть пьесы и повести, то, скорее всего, получился бы примерно вычисленный нами объем. Это означает, что наше предсказание с помощью функции Вейбулла в этом случае не получило экспериментального подтверждения: реальный чеховский объем оказался существенно больше предсказанного этой функцией. Функция Хауштайна, предсказывающая более высокий асимптотический уровень, обеспечивает в данной ситуации более правдоподобный прогноз.

2. Моделирование зависимости выборка — словарь на материале корпуса «Один речевой день»

Существует довольно распространенное мнение о словарной бедности и даже скудости повседневной речи. Так ли это на самом деле? Как выглядит такой словарь на фоне письменной речи: деловой, научной, публицистической, художественной?

Это мы попытались выяснить посредством аналитического моделирования отношения выборка — словарь с помощью двух прогностических функций, рассмотренных выше, на материале корпуса «Один речевой день» [The ORD speech corpus…; Bogdanova-Beglarian, Martynenko, Sherstinova]. Словарь формировался путем присоединения порций расшифрованных записей объемом 10 тыс. словоупотреблений. При построении словаря предварительно осуществлялась лемматизация.

При построении аналитических кривых мы пользовались методикой, рассмотренной в работах [Мартыненко; Гребенников; Мартыненко, Мартинович].

Эта методика включает следующие этапы.

1. Осуществлена линеаризация функций, чтобы упростить задачу моделирования с помощью метода наименьших квадратов. Обе функции приводились к линейному виду путем логарифмирования.

Так, функция Вейбулла $y = K - Ke^{-cx^d}$ преобразовывалась следующим образом.

$$\frac{K}{K - y} = e^{cx^d},$$

$$\ln \frac{K}{K - y} = cx^d,$$

$$\ln \ln \frac{K}{K - y} = \ln c + d \ln x.$$

Переход к линейному варианту функции Хаушстайна выглядит так:

$$y = \frac{Kx^\gamma}{x^\gamma + q},$$

$$\frac{y}{K} = \frac{1}{1 + \frac{q}{x^\gamma}},$$

$$\frac{K}{y} - 1 = \frac{q}{x^\gamma},$$

$$\ln \frac{K - y}{y} = \ln q - \gamma \ln x.$$

Однако асимптота для обеих функций входит в состав зависимой переменной. Поэтому были предприняты следующие дополнительные операции.

2. Осуществлена фиксация следующих в порядке возрастания асимптотических уровней, начиная с последнего максимального эмпирического, и предпринято построение эмпирических моделей для каждого уровня.

3. Осуществлен выбор среди эмпирических моделей той, которая наилучшим образом согласуется с эмпирическими данными. Это сделано путем подсчета суммы отклонений по абсолютной величине. Было выбрано то значение асимптотического уровня, которому соответствует минимальная сумма отклонений.

При моделировании с помощью функций Вейбулла и Хаштайна были получены следующие теоретические функции и соответствующие им значения при нарастающем объеме выборки (см. таблицу).

Результаты аппроксимации роста словаря русской устной повседневной речи

	Опыт	Теория	Теория
Объем выборки (тыс. словоупотреблений)		Вейбулл $y = 45000 - 45000e^{-0,000047x^{0,875}}$	Хаштайн $y = \frac{Kx^{0,696}}{x^{0,696} + 13017}$
10	2057	1 761	2 008
20	3179	3 042	3 166
30	4042	4 106	4 104
40	4837	5 016	4 915
50	5596	5 806	5 637
60	6225	6 499	6 293
70	6869	7 111	6 897
80	7354	7 653	7 457
90	8017	8 136	7 981
100	8774	8 567	8 474
ПРОГНОЗ			
200		11 099	12 292
500		12 616	18 702
1 000		12 747	24 090
10 000		12 750	44 853

Из таблицы видно, что на эмпирическом интервале теоретические кривые довольно близки друг к другу, согласование с эмпирическими данными для функции Хаштайна существенно лучше. Теоретические кривые близки друг к другу вплоть до объема 200 тыс. словоупотреблений. Резкие расхождения начинаются только в районе 500 тыс. словоупотреблений.

Таким образом, в качестве рабочей аналитической модели нужно принять функцию Хаштайна, для которой асимптотический уровень приблизительно равен 45 тыс. лексем. Этот уровень прибли-

зительно соответствует асимптотическому уровню, характерному для выдающихся русских прозаиков, отличающихся исключительным богатством словаря, например А. И. Куприна и Л. Н. Андреева. Это означает, что распространенное представление о том, что коллективный словарь устной речи исключительно беден, не соответствует действительности. Однако он полностью вычерпывается при выборке порядка 10 млн словоупотреблений.

Заключение

Данное исследование имеет предварительный характер. Для повышения надежности прогноза необходим больший объем выборки. Более того, необходимо исследовать зависимость между объемом прогностической базы и точностью прогноза. Необходимо также дополнить данное исследование системой разнообразных переменных, интегрально отражающих структуру словаря и его перестройку по мере увеличения числа расшифрованных звукозаписей.

Литература

Горькова В. И. Информетрия (количественные методы в научно-технической информации) // Итоги науки и техники. Серия «Информатика». М., 1988. Т. 10.

Гребенников А. О. О состоятельности статистик частотного словаря художественной прозы // Структурная и прикладная лингвистика: межвуз. сб. СПб., 1998. Вып. 5. С. 110–123.

Мартыненко Г. Я. Основы стилеметрии. Л., 1988.

Мартыненко Г. Я., Мартинович Г. А. Многопараметрический анализ результатов ассоциативного эксперимента. СПб., 2003.

Мартыненко Г. Я., Фомин С. В. Ранговые моменты // Научно-техническая информация. Серия 2. 1989. № 8. С. 23–29.

Нешитой В. В. Длина текста и объем словаря: показатели лексического богатства текста // Методы изучения лексики. Минск, 1975. С. 110–118.

Потемкин С. Б. Авторский корпус и словарь языка Антона Чехова // Труды Международной конференции «Корпусная лингвистика — 2015». СПб., 2015. С. 382–389.

Тулдава Ю. А. К вопросу об аналитическом выражении связи между объемом словаря и объемом текста // Лингвостатистика и квантитативные закономерности текста. Уч. зап. Тартуского ун-та. Тарту, 1980. Вып. 549. С. 113–124.

Частотный грамматико-семантический словарь языка художественных произведений А. П. Чехова (с электронным приложением) / О. В. Кукушкина [и др.]; под ред. проф. А. А. Поликарпова. М., 2012.

Частотный словарь рассказов А. И. Куприна / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 2006.

Частотный словарь рассказов А. П. Чехова / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 1999.

Частотный словарь рассказов Л. Н. Андреева / авт.-сост. А. О. Гребенников; под ред. Г. Я. Мартыненко. СПб., 2003.

Bogdanova-Beglarian N., Martynenko G., Sherstinova T. The „One Day of Speech" Corpus: Phonetic and Syntactic Studies of Everyday Spoken Russian / eds A. Ronzhin [et al.]. SPECOM 2015, LNAI. Springer, Switzerland, 2015. Vol. 9319. P. 429–437.

Haustein H.-D. Prognoseferfahren in den sozialistischen Wiertschaft. Berlin, 1970.

Herdan G. Type-token Mathematics. The Hague, 1962.

The ORD Speech Corpus of Russian Everyday Communication „One Speaker's Day": Creation Principles and Annotation / eds A. Asinovsky [et al.]; V. Matoušek, P. Mautner. TSD 2009, LNAI. Berlin; Heidelberg, 2009. Vol. 5729. P. 250–257.

Wimmer G. The type-token-relation // Quantitative Linguistik — Quantitative Linguistics. Ein internationales Handbuch / Hrsg. R. Köhler, G. Altmann, R. G. Piotrowski. Berlin; New York: de Gruyter, 2005. P. 325–348.

М. С. Поддубных

ЛЕКСИКО-ГРАММАТИЧЕСКИЕ ПАРАМЕТРЫ ВЫЯВЛЕНИЯ КОНЦЕПТА ФОРМЫ В КОНТЕКСТАХ КОРПУСА

Аннотация. В статье рассматривается многообразие значений концепта формы и его реализации в тексте. В качестве исходной, когнитивно значимой характеристики представления формы рассматривается физический облик материального объекта, который задается лексико-грамматическими параметрами в алгоритме поиска по корпусу текстов.

Ключевые слова. Концепт формы, корпус текстов, контекст, синтаксические конструкции, лексические ограничения, шаблон, поисковый алгоритм.

Maria S. Poddubnykh

LEXICO-GRAMMATICAL FEATURES OF THE „FORM“ CONCEPT IN RUSSIAN CORPUS TEXTS

Abstract. In this paper diverse meanings of the concept of form and its realizations in Russian written texts are investigated. Physical traits of a material object are considered as a basic cognitively important representation of the concept. Detection of these attributes in written texts is carried out with the help of lexico-grammatical features formalized as queries for a search algorithm.

Keywords. Concept of form, lexical corpus, context, syntactic structures, lexical restrictions, pattern, search algorithm.

Вступление

В семантических исследованиях понятие «форма» чаще всего встречается как форма слова, высказывания — форма лингвистическая. Именно в таком ключе появляется в основном это понятие в трудах Ю. Д. Апресяна, Д. Н. Шмелева и других известных

лингвистов (см., например, [Апресян; Бондарко; Кобозева; Шмелев]). Внимание форме как свойству объекта нелингвистического уделяется гораздо реже. Однако она может пониматься, к примеру, как очертания и характеристики физического объекта, и в этом случае исследуется участие формы в процессах категоризации и перцепции.

Важную роль форма материального объекта и ее изменения играют в том, как люди категоризируют окружающие их предметы. У. Лабов в одном из своих исследований определял, как изменится восприятие человеком бытовых емкостей, если изменять их высоту, диаметр и прочие параметры. Так, если увеличить высоту кружки, она постепенно перестанет считаться кружкой и будет рассматриваться человеком уже как ваза [Labov].

Много внимания форме уделяется в работе Е. В. Рахилиной «Когнитивный анализ предметных имен: семантика и сочетаемость» [Рахилина]. Согласно ей, в языковой картине мира, на которую опирается сочетаемость, «прежде всего обращает на себя внимание связь формы и размера. В самом деле: охарактеризовать, например, как *глубокий*, мы можем только объект определенной формы; тем самым, определяя размер, мы, кроме того, одновременно описываем и форму предмета. То же относится и к другим размерам — в частности, к высоте» [Там же. С. 120].

Сочетаемость слов в языке оказывается связанной с формой описываемого объекта и его размерами, которые в свою очередь связаны с тем, как человек их использует и что он принимает за норму.

Е. В. Рахилина также отмечает, что в русском языке мало прилагательных, которые бы описывали форму (например, *круглый*, *овальный*, *квадратный* и пр.). Имена прилагательные могут описывать и «если не саму форму предмета, то такие отклонения от нормы, которые … позволяют эту форму реконструировать» [Там же. С. 152].

О форме и ее роли в категоризации и восприятии действительности писали также Дж. Тейлор [Taylor] и Дж. Лакофф [Lakoff]. Так, Тейлор цитирует Е. Рош [Rosch]: «Существуют свидетельства того, что определенные геометрические формы (хорошие формы гештальтпсихологии, например, *круг*, *квадрат*, *треугольник*) и определенные пространственные ориентиры (например, *вертикальный* и *горизонтальный*, а не *косой*), как и фокальные цвета, перцептивно более заметны, нежели отклонения от этих форм, и следовательно также приобретают статус прототипа» [Taylor, p. 52].

Объектом данного исследования является письменный текст в том виде, в котором он представлен в корпусах русского языка.

Цель данного исследования — выявление реализаций концепта формы в тексте на основании лексико-грамматических параметров контекста. Чтобы оценить эффективность этих параметров, было принято решение формализовать их как систему шаблонов для поискового алгоритма. Форма в данной работе понимается как физический облик некоторого материального объекта, способ протекания некоторого процесса или явления.

Языковые реализации концептов собираются в лексико-семантические базы данных, которые впоследствии могут получить широкое применение в прикладной лингвистике, например для семантического анализа текста или для машинного перевода.

Данная работа проводится в рамках проекта RussNet — разработки на кафедре математической лингвистики СПбГУ компьютерного тезауруса в качестве электронного лексикона для текстовых анализаторов русского языка. Он принадлежит к большой группе wordnet-словарей, составляемых для языков разных типов (на данный момент создано более 50 wordnet-словарей для различных языков мира).

В ходе исследования необходимо выполнить следующие задачи: 1) определение ядра лексико-семантического поля формы, 2) выявление параметров представления концепта формы в текстах, и 3) оценка эффективности выявленных параметров.

Для выполнения первой задачи используется комплекс методов, применяемый для пополнения RussNet [Материалы к компьютерному тезаурусу …].

Подготовка материала к представлению в формате базы данных RussNet осуществляется при помощи комплексной процедуры структурно-семантического анализа. Эта процедура проводится на основе лексикографических данных (словарных дефиниций) и контекстов употребления слов и включает в себя три этапа:

- дефиниционный анализ;
- деривационный анализ;
- контекстный анализ.

На этапе *дефиниционного анализа* устанавливаются семантические признаки, существенные для конкретной лексико-семантической группы. Осуществляется это путем сопоставления значений

слов на основании дефиниций толковых словарей. Для дефиниционного анализа оптимальными оказались данные, представленные в четырехтомном словаре под редакцией А. П. Евгеньевой [Словарь русского языка]. Некоторая часть словарных определений строится по принципу «гипероним + характерные отличительные признаки понятия», что облегчает построение иерархической структуры (родо-видовых отношений) понятий в базах данных типа WordNet и RussNet.

К *деривационному анализу* прибегают, если в лексико-семантической группе имеется разветвленная система словообразовательных связей. В этих случаях анализа словарных дефиниций часто оказывается недостаточно для выделения понятий и их элементов.

Контекстный анализ позволяет глубже заглянуть во внутреннюю организацию лексического значения слов. Во-первых, контекстное окружение позволяет уточнить значение и снять неоднозначность. Во-вторых, благодаря контекстам можно определить внутреннюю иерархию компонентов значения. К примеру, слово *форма* имеет несколько значений, среди которых *внешние очертания предмета* и *единая по цвету, покрою и другим признакам одежда, установленная для лиц определенных категорий.* Контекстный анализ позволяет понять, какое значение встречается чаще и в каком лексическом окружении.

Аналогичный подход к контекстам используется в группе Рахилиной и Кустовой при снятии семантической неоднозначности [Многозначность…; Семантическая разметка…; Семантические фильтры…].

В дальнейшем для исследования количественных данных, полученных при изучении письменного текста, также применяется *статистический анализ*. При помощи соответствующих формул и методов можно получить представление о частотности того или иного языкового элемента, оценить важность и устойчивость некоторого явления, проверить эффективность использования правил и запросов.

На этапе выявления параметров представления концепта формы в письменных текстах основными методами являются контекстный анализ и статистический анализ данных. В частности, вычисляется частота реализации того или иного варианта представления формы в лексическом описании.

При оценке эффективности выявляемых параметров, речь о которых пойдет ниже, применяются некоторые методы информационного поиска, а именно формирование поискового запроса сообразно применяемой в используемом поисковом алгоритме нотации и оценка полноты (отношение найденных релевантных контекстов к общему числу релевантных контекстов) и точности (отношение найденных релевантных контекстов к общему числу найденных контекстов), полученных в результате выполнения запроса данных.

Выявление основных вариантов и параметров представления концепта формы

Первый этап исследования — выявление основных вариантов реализации концепта формы в лексическом описании. На этом этапе использовались следующие источники лингвистической информации:

- частотный список слов русского языка;
- корпус кафедры математической лингвистики (ККМЛ) объемом 21 млн словоупотреблений.

Корпус ККМЛ составлен из небольших фрагментов текстов, а не из цельных произведений, что не позволяет доминировать тематическим словам отдельного произведения или текста. Это дает возможность получить более точное представление о частотности употребления слов в письменной речи. Кроме того, с помощью корпус-менеджера Bonito, которым оснащен ККМЛ, можно получать случайные выборки контекстов по заданному параметру.

Для описания понятия формы в первую очередь из частотного списка слов было отобрано 42 слова, которые потенциально могли отображать концепт формы. Для этих слов были отобраны случайным образом и проанализированы контексты из корпуса ККМЛ. В результате список сократился до 16 имен существительных (*форма*, *образ*, *круг*, *шар*, *пятно*, *кольцо*, *сфера*, *купол*, *облако, диск*, *волна*, *крест*, *фигура*, *угол*, *сеть*, *сетка*), поскольку остальные слова либо не были представлены в корпусе, либо не встречались в выборках в нужном значении. В дальнейшем список был дополнен еще 11 именами существительными со значением формы (*квадрат*, *конус*, *пирамида*, *пирамидка*, *овал*, *клякса*, *призма*, *цилиндр*, *треугольник*, *куб*, *капля*).

Значение формы несут также имена прилагательные, образованные от существительных, обозначающих геометрические фигуры (*круглый*, *овальный*, *прямоугольный*). Доля искомых значений по отношению к общему числу словоупотреблений таких прилагательных стабильно высока и может превышать 90%. Оставшиеся словоупотребления приходятся в основном на наименования (*озеро Овальное*) и устойчивые сочетания (*круглый дурак*).

При контекстном анализе перечисленных слов были выделены синтаксические конструкции и деривационные модели, потенциально обозначающие форму. Для этих конструкций также был проведен контекстный анализ, результаты которого представлены в табл. 1.

Таблица 1

Конструкция	ipm	Всего контекстов в ККМЛ	Рассмотренных контекстов	Обозначений формы	Доля в рассмотренном, %
в виде	62,53	1520	200	181	90,5
похожий на	52,36	1309	200	176	88
напоминать	32,93	1178	200	123	61,5
*образный	25,76	2699	200	42	21
*видный	14,13	1594	200	39	19,5
в форме	14,03	486	200	127	63,5
подобный	5,98	478	200	55	27,5
*подобный	4	290	290	88	30,3

Примечание. ipm (instances per million words) — частота (количество словоупотреблений) на миллион словоформ.

На основании полученной информации и планировалось выявить и формализовать параметры выделения искомого концепта из фрагментов письменных текстов. Следует отметить, что значение формы, особенно формы физического объекта, далеко не самое частотное в просмотренном корпусе и в письменной речи вообще.

Далее был реализован поисковый алгоритм на языке программирования C[1], позволяющий выполнять поиск контекстов по заданным шаблонам с возможностью введения дополнительных фильтров и пользовательских словарей-списков. Программа осуществляет поиск конструкций, соответствующих заданным параметрам, в морфологически размеченном корпусе, представленном в формате XML-документа.

Оценка эффективности выявленных параметров

В результате анализа полученных в ходе исследования данных был выявлен ряд грамматических, лексических и синтаксических параметров, при помощи которых можно выделять из текстов конкретные реализации концепта формы. Эти параметры были формализованы как запросы для поискового алгоритма. В запросе необходимо указать корневой элемент, и есть возможность указать число элементов контекста, место каждого элемента в искомом контексте, конкретную словоформу или лемму, часть леммы, грамматические показатели, согласование между выбранным элементом и корневым элементом запроса.

При помощи составленных запросов был выполнен поиск реализаций концепта формы в корпусе объемом около 1 млн словоупотреблений. Было получено 426 контекстов, из которых 120 содержали искомое значение формы. Точность поиска составила 28%. Параметры и результаты поиска представлены в табл. 2.

Среди отрицательных реализаций для некоторых параметров можно выделить определенные закономерности. Так, для существительного *форма* следует исключить значение форменной одежды, а значит, можно поставить лексическое ограничение на сочетания со словами, обозначающими военных и гражданских служащих. Для конструкции *похожий на* следует исключить имена собственные, так как в таких случаях говорится не о форме или схожести формы, а о сходстве одного человека с другой конкретной личностью. Также вполне четко вырисовываются ограничения для трех словообразовательных параметров.

[1] Поисковый алгоритм реализован А. М. Поповым, аспирантом кафедры прикладной и математической лингвистики СПбГУ.

Таблица **2**

Параметры поиска	Искомые реализации	Отрицательные реализации
Сочетания имен прилагательных и существительных, где имя прилагательное сложное (первый корень произвольный, второй корень — «образ»)	*взрывообразный*, *газообразный*, *волнообразный*, *единообразный*, *женообразный*, *звездообразный*, *человекообразный*, *ведьмообразный*, *крестообразный*, *дискообразный*	*своеобразный*, *разнообразный*, *многообразный*, *однообразный*, *целесообразный*, *несообразный*, *нецелесообразный*
Сочетания имен прилагательных и существительных, где имя прилагательное сложное (первый корень произвольный, второй корень — «вид»)	*пулевидный*, *реповидный*, *древовидный*, *кубовидный*, *змеевидный*, *яйцевидный*, *веретеновидный*, *шаровидный*, *прутовидный*, *кустовидный*, *воронковидный*, *грушевидный*, *ромбовидный*	*очевидный*, *недальновидный*, *ликвидный*, *высоколиквидный*, *завидный*, *незавидный*, *миловидный*, *стекловидный*, *невидный*
Сочетания имен прилагательных и существительных, где имя прилагательное сложное (первый корень произвольный, второй корень — «подоб»)	*речеподобный*, *человекоподобный*, *листоподобный*, *ромашкоподобный*, *языкоподобный*	*преподобный*, *правдоподобный*, *неправдоподобный*
Конструкции типа «похожий на + *X*», где *X* — имя существительное в винительном падеже или именное словосочетание	*лист (агавы)*, *кактус*, *бутылка*, *домик*, *дельфин*, *русалка*, *курок*, *точка*, *заклинание*, *орден*, *шиповник*, *стрела*, *вымя*, *созвездие (Близнецов)*, *корона*, *Баба-яга*, *утка*, *укроп*, *стремена*, *павлин*	*новобранец*, *ад*, *правда*, *вироид*, *уважение*, *балаган*, *Иоанн Златоуст*, *сюжет*, *Марчело Мастрояни*, *аббревиатура*, *(Александр) Лившиц*, *Гусман*, *сочетание*, *сорт*, *господин*, *Лукашенко*, *мачо*, *состояние*, *шарж*, *культ*, *оригинал*, *настоящая любовь*, *приторная Массандра*, *спортивные часы*, *Тосиро Мифунэ*, *грустное предчувствие*, *маленький Дон-Кихот*

Окончание табл. 2

Параметры поиска	Искомые реализации	Отрицательные реализации
Конструкции типа «напоминать + *X*», где *X* — имя существительное в винительном падеже или именное словосочетание	*клубок змей*, *замок*, *круг*, *дыня*, *камера-обскура*, *бутон*, *деревце*, *сумка*, *высокая буква П*, *корыто*, *вензель*	*сон*, *рислинг*, *произведение*, *оплеуха*, *потемкинские деревни*, *Владислав Дворжецкий*, *формат*, *пустырь*, *судейство*, *положение*, *калейдоскоп*
Конструкции типа «в форме + *X*», где *X* — имя существительное в родительном падеже или именное словосочетание	*сердечко*, *ответ*, *диалог*, *курья нога*	*майор*, *офицер*, *генерал*
Сочетания «имя прилагательное в родительном падеже + „форма“ в родительном падеже»	*сферический*, *капсулированный*, *повелительный*, *эллипсоидный*	*подобный*, *клинический*, *внутренний*, *привлекательный*, *этот*, *малый*, *необходимый*, *стандартизированный*, *высший*, *разный*, *некий*
Конструкции типа «имя существительное + „в виде“ + *X*», где *X* — имя существительное в родительном падеже или именное словосочетание	*хронограф*, *жидкость*, *рогалик*, *пузырек*, *амфора*, *квадрат*, *прямоугольник*, *круг*, *ананас*, *частица*, *ямка*, *куст*, *кривая*, *столбик*, *птичка*	*лишение свободы*, *документ*, *форма*, *продажа*, *энцефалит*, *факт*, *база данных*, *ион*, *массив*, *последовательность*, *объект*, *конференция*, *граф*, *пятерка*, *эпиграмма*

Лексические ограничения было решено реализовать в виде стоп-списков. После применения лексических ограничений точность возросла до 42%.

Для более точной оценки эффективности лексико-грамматических и синтаксических параметров выявления концепта формы в лексическом описании было решено отобрать тестовую совокупность.

Результаты

Эффективность шаблонов поискового алгоритма была проверена на тестовой совокупности, состоявшей из 275 контекстов (5015 словоупотреблений). Поскольку слова и конструкции,

относящиеся к концепту формы, как показывает предварительное исследование базовой структуры, были нечастотными, то случайная выборка контекстов была бы не показательна для тестирования. В «насыщенную» тестовую совокупность были отобраны 79 контекстов с искомым значением формы, которые были дополнены случайными контекстами из морфологически размеченного фрагмента Национального корпуса русского языка со снятой неоднозначностью. Программа выделила 81 контекст как содержащий форму. Из них 67 принадлежали «насыщенной» тестовой совокупности. Таким образом, точность выделения концепта формы в текстах составила 83%, при полноте — 85%.

Заключение

Форма — многогранное понятие, в том или ином виде встречающееся и в философии, и в лингвистике, и в математике, и в биологии, и в печатном и литейном деле. В лингвистике в первую очередь форма понимается как форма лингвистическая, то есть форма слова. Однако не последнюю роль играет и форма материального объекта — она и ее изменения оказывают серьезное влияние на то, как люди категоризируют предметы и окружающую реальность в целом.

В данной работе рассматривается понимание формы как физического облика некоторого материального объекта, способа протекания некоторого процесса или явления.

В исследовании было рассмотрено представление понятия формы в письменных текстах и выявлены некоторые его лексические, грамматические и синтаксические параметры. Чтобы оценить эффективность этих параметров, был реализован поисковый инструмент для извлечения искомых реализаций концепта. Дальнейшее уточнение и расширение набора параметров и лексических ограничений позволит осуществлять автоматизированное выделение из текстов фрагментов, содержащих понятие формы, с большей полнотой и точностью.

Литература

Апресян Ю. Д. Лексическая семантика: Синонимические средства языка. М., 1974.

Бондарко А. В. О грамматике функционально-семантических полей // Известия Академии наук СССР. Серия литературы и языка. 1984. Т. 43, № 6. С. 492–503.

Кобозева И. М. Лингвистическая семантика. М., 2002.

Материалы к компьютерному тезаурусу лексики русского языка / сост. И. В. Азарова, О. А. Митрофанова. СПб., 2002.

Многозначность как прикладная проблема: лексико-семантическая разметка в Национальном корпусе русского языка / Е. В. Рахилина [и др.] // Компьютерная лингвистика и интеллектуальные технологии: по материалам Международной конференции «Диалог 2006». М., 2006.

Рахилина Е. В. Когнитивный анализ предметных имен: семантика и сочетаемость. М., 2008.

Русская грамматика. Т. 2. Синтаксис / Н. Ю. Шведова [и др.]. М., 1980.

Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы / Г. И. Кустова [и др.] // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005.

Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные / О. Ю. Шеманаева [и др.] // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2007 г.). М., 2007.

Словарь русского языка: в 4 т. / РАН, Ин-т лингвистич. исследований; под ред. А. П. Евгеньевой. 4-е изд. М., 1999.

Шмелев Д. Н. Очерки по семасиологии русского языка. 3-е изд. М., 2008.

Labov W. The boundaries of words and their meanings // New ways of analyzing variation in English / eds Ch.-J. N. Bailey, R. W. Shuy. Washington, 1973. P. 340–373.

Lakoff G. Women, fire and dangerous things: What cathegories reveal about the mind. Chicago, 1990.

Rosch E. Natural cathegories // Cognitive Psychology. 1973. Vol. 4, N 3. P. 328–350.

Taylor J. R. Linguistic categorization: Prototypes in linguistic theory. Oxford, 1995.

П. В. Паничева

ЛИНГВИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СТРЕССА, БЛАГОПОЛУЧИЯ И ТЕМНЫХ ЛИЧНОСТНЫХ ХАРАКТЕРИСТИК НА МАТЕРИАЛЕ ТЕКСТОВ РУССКОЯЗЫЧНЫХ ПОЛЬЗОВАТЕЛЕЙ FACEBOOK

Аннотация. В данной статье обсуждается постановка эксперимента по выявлению лингвистических коррелятов стресса, субъективного благополучия, отчуждения моральной ответственности и темных личностных характеристик. Описаны теоретические основы анализа данных явлений в психологической литературе и опыт сходных исследований на материале английского языка. Разработано приложение, содержащее опросники для оценки этих параметров у русскоязычных пользователей Facebook и собирающее тексты соответствующих авторов. Предполагается применение морфологического анализа и статистических алгоритмов для выявления лингвистических переменных, значимых относительно данных психологических измерений, а также для моделирования этих измерений и их автоматической оценки. Результаты экспериментов по автоматической классификации текстов относительно измерений стресса, субъективного благополучия, отчуждения моральной ответственности и темных личностных характеристик будут оценены и сопоставлены с результатами сходных исследований на материале английского языка.

Ключевые слова. Морфологический анализ русского языка, автоматическое определение личностных характеристик, стресс, благополучие, Темная триада, отчуждение моральной ответственности.

Polina V. Panicheva

TOWARDS PROFILING OF STRESS, WELL-BEING AND DARK TRAITS IN FACEBOOK TEXTS BY RUSSIAN AUTHORS

Abstract. The paper describes an experiment setting aimed at finding linguistic correlates of stress, subjective well-being, moral disengagement and dark personality traits. The background for the investigation includes, on the one hand, psychological research

on these phenomena and their measurement instruments, and on the other hand, recent advances in linguistic author profiling. A Facebook application has been designed to put together psychological measures based on completed surveys by Russian Facebook users and texts by corresponding authors. Morphological and statistical analysis will be applied in order to reveal significant linguistic features indicative of the certain psychological conditions and to model these conditions in terms of probable text features. The results of the automatic author profiling experiments will be evaluated and compared to respective advances in English author profiling.

Keywords. Russian morphological parsing, automatic author profiling, stress, subjective well-being, Dark triad, moral disengagement.

1. Введение

Сегодня социальные сети предоставляют уникальную возможность для междисциплинарных исследований поведения человека. Они содержат беспрецедентный объем информации, в том числе текстовой, которая, с одной стороны, представляет собой корпусы текстов, имплицитно содержащих в себе разметку по множеству различных параметров, с другой стороны, является естественным хранилищем данных о поведении человека, незаменимым для исследователей гуманитарных направлений, в первую очередь психологии.

В последние годы особое внимание психологов привлекает исследование стресса, его причин и следствий, взаимосвязи с негативными личностными характеристиками и понятием травмирующего опыта [Ледовая, Боголюбова, Тихонов]. Социальные сети содержат информацию о непосредственном проявлении данных факторов в речевом поведении, давая возможность исследовать его параллельно с применением психодиагностических методик и сопоставлять, таким образом, теоретические диагностические критерии с их практической реализацией, не вмешиваясь в последнюю.

Цель описываемого проекта заключается в выявлении лингвистических коррелятов показателей уровня стресса, темных личностных характеристик и психологического благополучия индивида на материале текстов социальной сети Facebook. В задачи входят: 1) реализация и применение приложения, содержащего опросник с оценкой индивидуальных значений по данному набору психологических параметров и осуществляющего сбор открытых текстов пользователя; 2) представление данных в виде корпуса текстов, опосредованно размеченных значениями психологических характеристик авторов; 3) статистическая обработка корпуса с целью исследования

лингвистических коррелятов исследуемых феноменов; 4) построение и оценка моделей речевого поведения, соответствующих показателям стресса, благополучия и темных личностных характеристик.

2. Психологические основания исследования

В качестве исследуемых психологических феноменов применяются следующие шкалы: отчуждения моральной ответственности, скрининга посттравматических симптомов, темных черт личности, индекс хорошего самочувствия ВОЗ-5.

Отчуждение моральной ответственности

Понятие отчуждения моральной ответственности разработано в теории А. Бандуры [Mechanisms of Moral Disengagement …]. Индивид, обладающий достаточно стабильной системой моральных ценностей, тем не менее способен совершать поступки, не всегда с ней согласующиеся, избегая при этом когнитивного диссонанса. Это говорит об избирательном применении моральных норм, когда по отношению к различным поступкам индивидом может применяться или не применяться собственная шкала моральных ценностей. Согласно теории Бандуры, это происходит с помощью механизмов отчуждения моральной ответственности. Исследователь описывает восемь механизмов отчуждения моральной ответственности, разделяя их на три группы: 1) относящиеся к конструированию поведения как такового, 2) искажающие связь между поведением и его вредными для других людей последствиями, 3) искажающие образ пострадавших в результате деструктивного поведения.

В данном исследовании используется краткий опросник по отчуждению моральной ответственности из [Why employees do bad things …] (восемь вопросов), переведенный на русский язык и адаптированный авторами [Ледовая, Боголюбова, Тихонов].

Скрининг посттравматических симптомов

Опросник представлен в работе [The Primary Care …] как процедура первичной оценки выраженности посттравматической симптоматики в рамках медицинского обследования, адаптирован для русского языка авторами [Ледовая, Боголюбова, Тихонов].

Темные черты личности

«Темная триада», состоящая из неклинических черт: психопатии, нарциссизма и макиавеллизма, — исследуется в качестве отдельного синдрома более десяти лет [Paulhus, Williams]. С одной стороны, это три различные черты, включающие каждая свою группу характеристик, с другой — были продемонстрированы многочисленные сходства в поведенческих проявлениях и корреляции между изолированными оценками по трем чертам. Опросник включает в себя 27 вопросов, переведенных и адаптированных в работе [Егорова, Ситникова].

Индекс хорошего самочувствия ВОЗ-5

Конструкт психологического благополучия исследуется уже около 50 лет. Он был определен в работе [Bradburn] как соотношение между позитивным и негативным аффектом, или как состояние счастья. В дальнейшем это понятие дополнялось и расширялось, и в современных исследованиях [Diener] подчеркиваются три аспекта психологического благополучия: оно 1) субъективно; 2) неизбежно включает в себя позитивный аффект, причем однозначного вывода о соотношении позитивного и негативного аффектов в рамках данного конструкта нет; 3) связано с глобальной оценкой всех аспектов жизни индивида.

В данной работе в качестве шкалы психологического благополучия используется индекс хорошего самочувствия ВОЗ-5, переведенный на русский язык Всемирной организацией здравоохранения (ВОЗ) и находящийся в открытом доступе[1].

3. Моделирование личностных характеристик в лингвистике

В последние несколько десятилетий автоматическое моделирование и выявление индивидуальных характеристик авторов на основе речевого поведения приобрело широкую популярность. Во многом такие исследования опираются на опыт лингвистической экспертизы по выявлению угроз в текстах [Баранов], развитие технологий по

[1] URL: https://www.psykiatri-regionh.dk/who-5/Documents/WHO5_Russian.pdf

распознаванию речи с целью психодиагностики на основе речевого поведения [Журавлева, Коваль], а также на лингвистические разработки в направлении создания автоматизированного рабочего места медицинского психолога [Сидоров, Кастро-Санчес].

С одной стороны, индивидуальные характеристики трактуются как объективная информация об авторе, а именно возраст, пол [Effects of Age …], профессиональная принадлежность [Panicheva, Cardiff, Rosso]. С другой стороны, под индивидуальными характеристиками понимаются личностные черты. Они представляют собой более сложную классификацию для исследования, так как, в отличие от объективных характеристик, информация о них не содержится напрямую в социальных сетях и соответствующих корпусах текстов. Лингвистическое моделирование личностных черт требует дополнительного этапа при создании корпуса текстов, содержащего применение методик психологического тестирования, результаты которого далее используются при разметке эталонного корпуса для анализа и обучения статистических алгоритмов.

В качестве лингвистических параметров для текстового моделирования личностных черт используются различные статистические метрики встречаемости слов и конструкций. Все они основаны на предположении о том, что высокая частотность определенных классов языковых единиц в текстах автора отражает высокую вероятность некоторых его личностных особенностей, в то время как низкая частотность данных единиц говорит о других, вероятно противоположных чертах. Так, в работе [Predicting Dark Triad …], посвященной моделированию черт Темной триады на основе текстов англоязычного Твиттера, высокие показатели по шкале психопатии имеют значительную положительную корреляцию с лексическими маркерами гнева и обсценной лексикой и значительную отрицательную корреляцию с лексикой, относящейся к позитивным эмоциям и местоимениям первого лица множественного числа.

Личностные характеристики в компьютерной лингвистике

Лингвистическое моделирование личностных особенностей на основе текстов на материале английского языка приобрело популярность более десятилетия назад. Авторы [Pennebaker, Francis, Booth] представляют лингвостатистический инструмент Linguistic Inquiry

and Word Count (LIWC), предназначенный для многопараметрического анализа встречаемости слов в текстах и использующий множество параметров, начиная от длины слов, сложности предложений, грамматических показателей и заканчивая классами слов, соответствующих психологическим процессам — эмоциональным, мотивационным, когнитивным. В более поздней работе [Pennebaker, Mehl, Niederhoffer] описаны применение данного инструмента и состояние других исследований в этом направлении. Особое значение уделяется частицам, словам с преимущественно грамматическим значением и эмоционально маркированной лексике. Так, слишком высокая встречаемость частиц относительно значимых частей речи может свидетельствовать об органических поражениях мозга в зоне Вернике, в то время как полное их отсутствие — о поражениях в зоне Брока. Частотность местоимения первого лица единственного числа «I» в тексте коррелирована с информацией о возрасте, поле, депрессии, болезни автора; причем подчеркивается важность различения падежных форм местоимений: несмотря на их небогатый арсенал в английском языке, местоимения «I» и «me» следует учитывать в качестве отдельных переменных.

Одной из наиболее популярных психологических категорий с точки зрения лингвостатистического исследования является «Большая пятерка» личностных характеристик [John, Srivastava], включающая экстраверсию, доброжелательность, добросовестность, невротизм, открытость опыту. Большая часть работ направлены на исследование проявлений различных комбинаций данных пяти характеристик в текстах [Lexical Predictors…; Nowson]. Полученные модели подчеркивают значимость с точки зрения этих характеристик таких лингвистических параметров, как оценочные слова и выражения (например, слова, выражающие непосредственно эмоциональную оценку, связаны с высокими показателями по шкале невротизма, в противоположность оценочной лексике объективных качеств, присущих самому объекту); некоторые классы служебных слов (например, вспомогательный глагол первого лица единственного числа «am», возвратное местоимение «yourself» позитивно коррелированы с высокими значениями экстраверсии).

В последнее время растет число успешных попыток лингвистического моделирования и автоматической идентификации более сложных и специфических психологических явлений, среди которых

клиническая психопатия на материале интервью о себе [Hancock, Woodworth, Porter], неклиническая психопатия и проблемы психического здоровья на материале текстов в Твиттере [Predicting Dark Triad …; Harman, Coppersmith, Dredze].

Исследования личностных характеристик на материале русского языка

На материале русского языка большая часть исследований в области анализа личностных характеристик ведется в рамках клинического описательного подхода. Авторы [Экспериментальное исследование…] описывают применение автоматизированного диагностического комплекса, разработанного с опорой на отечественные традиции исследования речи при психической патологии, для выявления параметров, существенных для речевой деятельности мужчин, больных шизофренией. В работе [Завитаев] применяются ручная обработка текстов пациентов с диагнозом шизофрении, компонентный анализ и семантический анализ; в результате успешно выявляются некоторые тематические, прагматические и семантические особенности речи, характеризующие данную патологию.

Основная часть работ по применению автоматизированных методик к моделированию личностных характеристик до сих пор велась на материале английского и других европейских языков, для которых на базовом уровне считается решенной проблема морфосинтаксического анализа. Русский язык представляет собой более сложную задачу с точки зрения морфологического анализа, включающего нормализацию словоформ и их разметку морфологическими тегами, и последующего синтаксического анализа. Тем не менее эти задачи успешно решаются в недавних работах отечественных лингвистов [RU-EVAL-2012…; Разработка лингвистического комплекса…], что говорит о наличии достаточных технических возможностей для применения лингвостатистического подхода к моделированию психологических категорий.

4. Постановка экспериментов по моделированию стресса, благополучия и темных личностных характеристик

Задачи эксперимента

Ведется разработка приложения на русском языке в социальной сети Facebook, в котором предлагается заполнить опросник по четырем категориям, описанным в разд. 2. Для пользователей, заполнивших опросник и давших согласие на исследование, приложение автоматически скачивает все публично доступные текстовые записи на русском языке в социальной сети. Таким образом, каждый текст получает разметку, представляющую собой значения по каждой из шкал стресса, благополучия, склонности к отчуждению моральной ответственности и темных черт личности автора. Предполагается сформулировать вероятность высокой или низкой оценки по той или иной шкале в терминах распределения в тексте слов, словосочетаний, грамматических параметров. В дальнейшем результаты исследования предполагается применить для автоматической классификации / регрессии текстов по отношению к классам, представленным в разметке, в терминах лингвистических параметров.

Лингвистические параметры моделей стресса, благополучия и темных черт личности

В качестве текстовых параметров будут выступать лексические и семантические единицы и N-граммы (значимые и служебные слова, показатели их лексико-семантической принадлежности), синтаксические, оценочные характеристики (сентиментные слова), показатели отрицания («не», «без», «никакой» и др.), а также структурные характеристики, отражающие среднюю длину и сложность слов и предложений. Особое внимание уделяется морфологическим средствам русского языка. Морфологический анализ будет включать методы, основанные на алгоритме [Разработка лингвистического комплекса …]. В результате в набор параметров войдут: 1) нормальные формы слов; 2) значения грамматических категорий (категории лица, числа для местоимений и глаголов; времени, вида, залога для глаголов; рода, числа, падежа для имен существительных и прилагательных). Для работы со сравнительно небольшими наборами

данных также потребуется применение семантического анализа, лингвистических онтологий и тезаурусов [RussNet … ; Бабенко].

Статистические методы выбора значимых параметров (метод относительной энтропии, Relief measure, рекурсивное удаление [Orange: Data Mining …]), алгоритмы статистического вывода (корреляционный анализ, инструмент ANOVA), снижения размерности пространства (факторный анализ, латентный семантический анализ), а также машинного обучения (Naive Bayes, метод опорных векторов, нейронные сети) будут использованы прежде всего для выявления наиболее значимых лингвистических признаков относительно исследуемых психологических измерений, а также для их текстового моделирования и автоматической оценки.

5. Ожидаемые результаты

В соответствии с результатами, полученными в [Predicting Dark Triad … ; Hancock, Woodworth, Porter], ожидается выявление набора лексических и грамматических параметров, характеризующих исследуемые психологические характеристики. Точнее, мы ожидаем, во-первых, что некоторые лексико-семантические единицы и классы, соответствующие мотивационным, аффективным, когнитивным и тематическим областям, будут преобладать в текстах авторов, имеющих различные личностные характеристики. Во-вторых, предполагается особая роль служебных слов, а также морфологических параметров значимых слов. На материале исследований английского языка значимыми по отношению к шкалам темных личностных характеристик оказываются индикаторы первого лица, множественного и единственного числа. В английском языке такую функцию несут местоимения, в то время как русский язык выражает эти категории более богатым набором средств, а именно формами лица и числа глагола. Предполагается подтвердить значимость этих категорий, а также достигнуть сравнимых или более высоких результатов автоматической классификации относительно зарубежных исследований.

6. Заключение

В статье представлены предпосылки и описан предполагаемый ход эксперимента по выявлению лингвистических коррелятов стресса, субъективного благополучия и темных черт личности. Описаны

данные психологические категории и опросники, применяемые для их измерения. На основе методов, предложенных в зарубежной компьютерной лингвистике для решения сходных задач, сформулирован корпусный подход к лингвостатистическому моделированию этих категорий. Предполагается развитие лингвистической составляющей анализа с учетом особенностей морфологии русского языка. В качестве материала используются открытые тексты пользователей социальной сети Facebook. С помощью методик статистического анализа предполагается выявить значимые для данных психологических феноменов лингвистические переменные, на основе которых будут построены и оценены формальные модели для создания автоматизированных инструментов психодиагностики.

Литература

Бабенко Л. Г. Толковый словарь русских глаголов. Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы. М., 1999.

Баранов А. Н. Семантика угрозы в лингвистической экспертизе текста // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). М., 2013. Вып. 12 (19).

Егорова М. С., Ситникова М. А. Темная триада // Психологические исследования. 2014. Т. 7, № 38. С. 12. URL: http://psystudy.ru (дата обращения: 30.09.2015).

Журавлева А. А., Коваль С. Л. Диагностика психологических качеств диктора по устной речи // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2007 г.). М., 2007. С. 183.

Завитаев П. Ю. Аутизм: клинико-семантическое и экспериментально-психологическое исследование // Российский психиатрический журнал. М., 2007. Т. 5. С. 44.

Ледовая Я. А., Боголюбова О. Н., Тихонов Р. В. Стресс, благополучие и Темная триада // Психологические исследования. 2015. Т. 8, № 43. С. 5. URL: http://psystudy.ru (дата обращения: 30.09.2015).

Разработка лингвистического комплекса для морфологического анализа русскоязычных корпусов текстов на основе Pymorphy и Nltk / П. В. Паничева [и др.] // Труды Международной научной конференции «Корпусная лингвистика — 2015» (СПб., 22–26 июня 2015 г.). СПб., 2015. С. 361.

Сидоров Г. О., Кастро-Санчес Н. Система для лингвистической оценки психологических профилей // Труды международной конференции по

компьютерной лингвистике и интеллектуальным технологиям «Диалог 2006». М., 2006. С. 464.

Экспериментальное исследование особенностей речевой деятельности у мужчин, больных шизофренией / Н. В. Зверева [и др.] // Медицинская психология в России: электрон. науч. журн. 2011. № 4. URL: http://medpsy.ru (дата обращения: 30.09.2015).

Bradburn N. M. The Structure of Psychological Well-Being. Chicago, 1969.

Diener E. The Science of Well-Being: The Collected Works of Ed Diener // Springer Science & Business Media, 2009. Vol. 37.

Effects of Age and Gender on Blogging / J. Schler [et al.] // AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006. Vol. 6. P. 199–205.

Hancock J. T., Woodworth M. T., Porter S. Hungry Like the Wolf: A Word-pattern Analysis of the Language of Psychopaths // Legal and Criminological Psychology. 2013. Vol. 18. P. 102–114.

Harman G., Coppersmith M., Dredze C. Quantifying Mental Health Signals in Twitter // ACL 2014. 2014. P. 51.

John O. P., Srivastava S. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives // Handbook of Personality: Theory and Research. 1999. Vol. 2, N 1999. P. 102–138.

Lexical Predictors of Personality Type / S. Argamon [et al.] // Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America. 2005.

Mechanisms of Moral Disengagement in the Exercise of Moral Agency / A. Bandura [et al.] // Journal of Personality and Social Psychology. 1996. Vol. 71, N 2. P. 364–374.

Nowson S. Identifying more bloggers: Towards large scale personality classification of personal weblogs // Proceedings of the International Conference on Weblogs and Social. CiteSeerX. 2007.

Orange: Data Mining Toolbox in Python / J. Demšar [et al.] // The Journal of Machine Learning Research. 2013. Vol. 14, N 1. P. 2349–2353.

Panicheva P., Cardiff J., Rosso P. Identifying Writers' Background by Comparing Personal Sense Thesauri // Natural Language Processing and Information Systems. Berlin, Heidelberg, 2010. P. 288–295.

Paulhus D. L., Williams K. M. The Dark Triad of personality: Narcissism, Machiavellianism, and Psychopathy // Journal of Research in Personality. 2002. Vol. 36, N 6. P. 556–563.

Pennebaker J. W., Francis M. E., Booth R. J. Linguistic Inquiry and Word Count: LIWC2001. Mahwah, 2001. Vol. 71.

Pennebaker J. W., Mehl M. R., Niederhoffer K. G. Psychological Aspects of Natural Language Use: Our Words, Our Selves // Annual Review of Psychology. 2003. Vol. 54, N 1. P. 547–577.

Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets / C. Sumner [et al.] // Proceedings of the IEEE 11th International Conference on Machine Learning and Applications (ICMLA 2012). Vol. 2. 2012. P. 386–393.

RU-EVAL-2012: Evaluating dependency parsers for Russian / A. Gareyshina [et al.] // Proceedings of COLING 2012: Posters. Mumbai, 2012. P. 349–360.

RussNet: Building a Lexical Database for the Russian Language / I. Azarova [et al.] // Proceedings of Workshop on Wordnet Structures and Standardisation and how these affect Wordnet Applications and Evaluation. Las Palmas, 2002. P. 60–64.

The Primary Care PTSD Screen (PC-PTSD): Development and operating characteristics / A. Prins [et al.] // Primary Care Psychiatry. 2004. Vol. 9, N 4. P. 151.

Why employees do bad things: Moral disengagement and unethical organizational behavior / C. Moore [et al.] // Personnel Psychology. 2012. Vol. 65, N 1. P. 1–48.

Г. Т. Букия, Е. В. Протопопова, О. А. Митрофанова

КОРПУСНАЯ ОЦЕНКА СТЕПЕНИ БЛИЗОСТИ ЕДИНИЦ В ЛЕКСИЧЕСКИХ КОНСТРУКЦИЯХ

Аннотация. В статье предлагается метод оценки связи между парой слов для конструкций, не наблюдаемых в корпусе. Представленный метод прост в реализации и не требует больших вычислительных мощностей. Основная идея заключается в том, что контекстно близкие слова несут в себе информацию о сочетаемости друг друга. Для оценки алгоритма предлагается ряд стандартных экспериментов.

Ключевые слова. Меры ассоциации, грамматика конструкций, извлечение коллокаций, сочетаемостные предпочтения.

Grigorii T. Bukia, Ekaterina V. Protopopova, Olga A. Mitrofanova

A CORPUS-DRIVEN ESTIMATION OF ASSOCIATION STRENGTH IN LEXICAL CONSTRUCTIONS

Abstract. The paper presents a method for estimating the association strength of constructions which are not observed in a corpus. The model is flexible, computationally light and easy to implement. The core idea is to aggregate "similar" target words and propagate their selectional preferences among others. In order to describe this idea statistically, two association measures are proposed: confusion probability measured on the observed collocates only and a final association measure derived from individual counts for all possible "similar" words. The quantitative analysis is presented as well as several examples and error discussion.

Keywords. Association measures, construction grammar, collocation extraction, selectional preferences.

1. Введение

В то время как традиционные лингвистические исследования описывают элементарные единицы, составляющие разные уровни языка, современные теории уделяют больше внимания извлечению и описанию воспроизводимых языковых структур, встречающихся в текстах: слов, неоднословных выражений, коллокаций, фразеологизмов и т. п., которые допускают объяснение с точки зрения грамматики конструкций. Вслед за создателями грамматики конструкций [Fillmore; Goldberg; Tomasello] лексической конструкцией будем называть сложную единицу языка, состоящую из фиксированного компонента (целевого слова, ядра) и переменных ячеек (слотов), которые заполняются различными элементами: лексемами, грамматическими и семантическими признаками. Таким образом, лексическую конструкцию можно рассматривать как шаблон, реализуемый в корпусе текстов в лексикализованном виде. Считается, что целевое слово в определенном значении накладывает ограничения на структуру контекста, причем эта структура описывается классами конструкций. Следовательно, основная функция конструкций в тексте — задавать регулярные сочетания слов с целевым словом в определенном значении.

Грамматика конструкций утверждает, что лексические конструкции представляют собой единство формы и содержания. Форма конструкции фиксируется, с одной стороны, за счет заданных элементов, а с другой — с помощью сочетаемостных ограничений (грамматических, лексико-семантических, пропозициональных), накладываемых на заполнение ячеек. Обычно значение конструкции не выводится из значений составляющих элементов, однако допускаются их различные варианты: от свободных сочетаний лексических единиц до фразеологизированных структур.

В данной работе конструкции рассматриваются как многоуровневые структуры, позволяющие описывать сочетаемость целевого слова в данном значении. Конструкции будут задаваться как лексическими единицами (леммами и словоформами), так и грамматическими и лексико-семантическими классами. Такой подход отражает идею о взаимодействии различных языковых уровней и помогает описывать языковые выражения как многоярусные сущности.

Автоматический анализ лексических связей и оценка сочетаемости слов — важная задача прикладной лингвистики. Сгенерированный текст (например, при машинном переводе) может быть грамматически корректен, но лишен при этом всякого смысла. Встречаются фразы, причины семантической несогласованности которых далеко не очевидны, а подчас и необъяснимы. Трудно объяснить, к примеру, разницу в оттенках значений слов *жаркий* и *горячий* [Апресян, с. 510]. Хотя, на первый взгляд, эти прилагательные можно назвать синонимами, существительные, с которыми они сочетаются, у каждого свои. Возникает вопрос: почему слова в парах *горячая вода* и *горячий чайник* сочетаются, тогда как в парах *горячая погода*, *горячее лето* — уже нет? Универсальный ответ для любого словосочетания найти невозможно.

Подобные проблемы встречаются и в других, смежных задачах прикладной лингвистики, поэтому метод, предложенный в данной статье, может быть использован при исследовании сочетаемостных предпочтений глаголов, выделении устойчивых коллокаций и фразеологизмов или при автоматическом выделении семантических классов.

В чистом виде оценка лексической сочетаемости имеет широкий спектр применения. Ее используют для проверки «гладкости» машинного перевода или выбора наиболее вероятного варианта, при построении формальных моделей языка, при генерации примеров употребления слов в машинных словарях. Для снятия лексической омонимии из предложения выделяются контекстные элементы, определяющие значение целевого слова. Исследование контекстов употребления позволяет автоматически выделить синонимы, гиперонимы или гипонимы в зависимости от выбора метрики.

Поскольку лексическая конструкция определяется как словосочетание, которое встречается в речи, естественно предположить, что если некоторая фраза встретилась в достаточно большом корпусе, ее можно считать лексической конструкцией. Здесь, впрочем, возникают определенные проблемы. На практике далеко не каждая биграмма является лексической конструкцией — это может быть описка, нарочное искажение, неверно снятая омонимия, наконец оба слова могут относиться к разным лексическим конструкциям. Поэтому важно определить именно степень сочетаемости слов. Ясно, что совместная частота встречаемости не вполне отражает то, насколько

данные слова характерны друг для друга, ведь одно из них может быть частотным само по себе. Например, фраза *хороший цвет* имеет большую частоту, чем фраза *пурпурный цвет*, при этом сочетаемость последней должна быть выше.

А. Стефанович и С. Грис [Stefanowitsch, Gries] рассмотрели в качестве степени сочетаемости вероятность того, что события «встретилось слово *x*» и «встретилось слово *y*» зависимы. Для этого строилась статистическая модель, называемая таблицей сопряженности, а зависимость признаков оценивалась точным критерием Фишера. При таком подходе наибольшее значение получали устойчивые лексические конструкции.

Однако если словосочетание отсутствует в корпусе, это вовсе не значит, что оно не является лексической конструкцией. Каким бы большим ни был корпус, нельзя ожидать, что он вместит в себя все мыслимые коллокации. Так, например, слово *краска* может встречаться с набором слов *красный*, *белый*, *зеленый*, но не встречаться со словами *бежевый* или *фиолетовый*. Возникает задача — как извлечь информацию о сочетаемости слов, если они не встречались в корпусе.

Можно предположить, что если два слова в некотором смысле близки, то они должны быть взаимозаменяемы, в частности этим свойством нередко обладают синонимы, гиперонимы и гипонимы. Тем не менее едва заметная разница в лексическом значении слов *горячий* и *жаркий* почти во всех случаях не позволяет заменить одно на другое. Поэтому взаимозаменяемыми словами следует считать лишь те, которые имеют значительное пересечение в контекстах употребления.

Сформулировать задачу можно следующим образом. Фиксируется конструкция, состоящая из двух ячеек с заданными морфологическими ограничениями (например, *прилагательное + существительное в именительном падеже*), для каждой ячейки определен подходящий набор слов. Требуется для всякой пары, заполнившей ячейку, получить оценку степени сочетаемости, используя лишь корпусные данные. Основная идея заключается в том, что если два слова (ядра), заполняющие одну ячейку, встречаются в корпусе с похожими наборами слов (контекстами), заполняющими другую ячейку, то первое ядро с высокой вероятностью сочетается с контекстами второго ядра и наоборот, даже если в корпусе данные пары не встречаются.

Для оценки качества модели нами были проведены стандартные эксперименты.

2. Современные исследования в области сочетаемости

Существуют различные статистические методы оценки сочетаемости слов.

При построении моделей языка определенную проблему вызывает разреженность данных: каким бы большим ни был корпус, он не охватывает и малой доли всех мыслимых биграмм (не говоря уже о более длинной цепочке). В простой статистической модели, которая исходит из оценки вероятности сочетания слов по их частоте в корпусе, вероятность любой новой биграммы будет нулевой. Чтобы это обойти, разрабатывались различные способы сглаживания данных: Additive Smoothing, Good-Turing Estimate, Jelinek-Mercer Smoothing, Discounting, Kneser-Ney Smoothing и др. Подробный обзор основных методик сглаживания приведен в [Chen, Goodman]. Популярные в последнее время модели языка, основанные на нейронных сетях [A Neural Probabilistic Language Model], также позволяют оценить вероятность не встреченной в корпусе последовательности. Этот подход близок к описываемым ниже дистрибутивным моделям, так как основан на получении информации о более широком контексте слова, чем в классических статистических моделях, для предсказания следующего слова в цепочке. Важным недостатком нейросетевых моделей является большая вычислительная сложность и склонность к переобучению. Кроме того, как уже отмечалось в предыдущем разделе, вероятность совместной встречаемости не всегда точно характеризует устойчивую тенденцию в сочетаемости.

Существенная доля работ по лексической сочетаемости посвящена глагольным конструкциям, для которых вводится понятие семантического класса — наполнителей ячеек. Часть методов оценки лексической сочетаемости слов, образующих не встречающуюся в корпусе фразу, используют дополнительные лингвистические ресурсы, например, вслед за Ф. Резником [Resnik], можно опираться на WordNet, чтобы определить, с какими семантическими классами употребляется данный глагол. Основанные на WordNet подходы обычно имеют низкую полноту и оказываются хуже статистических корпусных методов.

Большинство альтернативных подходов основаны на дистрибутивной модели семантики. Некоторые исследователи [Abney, Light] применяют кластеризацию для создания подобия семантических классов WordNet'a, в более сложном варианте вместо семантических классов вводятся скрытые переменные [Seaghdha]. Авторы [Dagan, Lee, Pereira] предлагают модель, основанную на сходстве новых последовательностей слов с наблюдаемыми в корпусе. В работе [Erk, Pad S., Pad U.] применяется схожий подход уже к конкретной задаче определения сочетаемостных ограничений. Предлагаются различные способы оценки значимости контекстных признаков и ограничения их количества [Пекар]. В недавней работе [Tian, Xian, Zheng] для выделения новых групп слов, сочетающихся с данным глаголом, используется случайное блуждание по графу предикатов и возможных аргументов. Эта модель реализует схожий с представленным здесь принцип — чем больше пересечение семантических классов у двух глаголов, тем больше вероятность их взаимозаменяемости.

Предложенный в настоящей статье метод, с одной стороны, обладает достаточно высокой точностью, а с другой — не требует больших вычислительных мощностей и громоздких моделей, легко реализуем и применим для практических целей.

3. Предлагаемая модель

3.1. Ядро и его контекст

Рассмотрим два непересекающихся множества слов X и Y. Про некоторые пары слов $x \in X$, $y \in Y$ известно, сколько раз они встречаются в корпусе.

Требуется построить функцию $F: X \times Y \to \mathbb{R}$, характеризующую степень сочетаемости любой пары слов.

Для некоторого слова $x \in X$ рассмотрим множество всех слов $c(x) \subset Y$ с которыми слово x встречается в корпусе. Пару $[x, c(x)]$ будем называть ядро и его контекст. Аналогично определяется ядро $y \in Y$ и его контекст $c(y) \subset X$.

Требуется ввести меру на контекстах данного ядра, характеризующую сочетаемость контекста и ядра, а также расширить самое множество контекстов.

3.2. Базовая мера

Базовой мерой назовем оценку степени сочетаемости для конструкций, встретившихся в корпусе. Вслед за А. Стефановичем и С. Грисом [Stefanowitsch, Gries] будем оценивать степень корреляции двух случайных величин — «в первой ячейке слово *x*» и «во второй ячейке слово *y*» — с помощью удобной статистической модели — таблиц сопряженности (табл. 1).

Таблица 1. **Общий вид таблицы сопряженности**

	X	$\overline{X}$
Y	a	b
$\overline{Y}$	c	d

Здесь
$a = (x\,y)$ — частота совместной встречаемости фразы *xy*;
$b = (x\,\overline{y})$ — частота встречаемости *x* с другими контекстами;
$c = (\overline{x}\,y)$ — частота встречаемости *y* с другими контекстами;
$d = (\overline{x}\,\overline{y})$ — общая частота встречаемости других сочетаний.

3.3. Меры связи в таблицах сопряженности 2 × 2

Для оценки степени связи можно использовать один из перечисленных критериев, который будем обозначать $f(x,y)$.

Коэффициент ассоциации. Статистика критерия:

$$Q = \frac{|ad - bc|}{ad + bc}.$$

Коэффициент ассоциации принимает значения от 0, если слова не связаны, до 1 в случае полной тождественности.

Коэффициент коллигации Юла. Статистика критерия:

$$K = \frac{|\sqrt{ad} - \sqrt{bc}|}{\sqrt{ad} + \sqrt{bc}}.$$

Между K и Q существует связь:

$$Q = \frac{2K}{1 + 2K}.$$

Коэффициент контингенции χ^2. Статистика критерия:

$$\chi^2 = \frac{\left(n|ad - bc| - \frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Значения статистики равны 0, если слова не связаны, и возрастают при наличии связности.

Точный критерий Фишера. Статистика критерия:

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{(a+b+c+d)!} \sum_{j=0}^{a} \frac{1}{(a+b-j)!\,(a+c-j)!\,(a+d-j)!}.$$

С достоверностью $1-p$ проверяемые события коррелируют, значит в качестве степени сочетаемости здесь нужно положить $f(x,y) := 1-p$.

Быстрый критерий *z*. Статистика критерия:

$$z = \frac{a - b + \frac{(a+c-b-d)(a+b)}{a+b+c+d}}{\sqrt{a+b}}.$$

***G*-критерий Вульфа:**

$$G = 2\left\{\left(a+\frac{1}{2}\right)\ln\left(a+\frac{1}{2}\right) + \left(b-\frac{1}{2}\right)\ln\left(b-\frac{1}{2}\right) + \left(c+\frac{1}{2}\right)\ln\left(c+\frac{1}{2}\right) + \right.$$
$$+ \left(d-\frac{1}{2}\right)\ln\left(d-\frac{1}{2}\right) - (a+b)\ln(a+b) - (c+d)\ln(c+d) -$$
$$- (a+c)\ln(a+c) - (b+d)\ln(b+d) +$$
$$\left. + (a+b+c+d)\ln(a+b+c+d)\right\}.$$

Взаимная информация:

$$m = \frac{a(a+b+c+d)}{(a+b)(a+c)}.$$

3.4. Взаимозаменяемость ядер

Чтобы расширить контексты слов, требуется ввести понятие взаимозаменяемости ядер. Два ядра будем называть *взаимозаменяемыми* и писать $x_1 \sim x_2$, если с первым ядром сочетаются контексты второго и наоборот. Можно получить апостериорную оценку этой величины. Если $y_1 \in c(x)_1$ — случайный контекст первого ядра, а $y_2 \in c(x)_2$ — случайный контекст второго ядра, то

$$\mathbb{P}\{x_1 \sim x_2\} = \mathbb{P}\{y_1 \in c(x_2) \,|\, y_1 \in c(x_1)\}\,\mathbb{P}\{y_2 \in c(x_1) \,|\, y_2 \in c(x_2)\}.$$

Эту величину можно оценить по корпусу. Действительно, согласно формуле Байеса,

$$\mathbb{P}\{x_1 \sim x_2\} = \frac{|c(x_1) \cap c(x_2)|^2}{|c(x_1)|\,|c(x_2)|}.$$

Чтобы сократить дисперсию оценки, будем считать, что в корпусе, помимо существующих, есть и дополнительные, «фиктивные» контексты, которые при необходимости дополняют количество контекстов каждого ядра до десяти. Если слово встречается в корпусе крайне редко, то вероятность того, что оно взаимозаменяемо с другим, тем самым пропорционально занижается. Необходимость такой надбавки получена эмпирически.

Отметим, что $\mathbb{P}(x_1 \sim x_1) = 1$, если у слова x_1 более 10 контекстов. Итак, вероятность взаимозаменяемости двух ядер вычисляется по формуле

$$g(x_i, x_j) = \frac{|c(x_i) \cap c(x_j)|^2}{\max(|c(x_i)|, 10)\max(|c(x_j)|, 10)}.$$

3.5. Оценка сочетаемости

Теперь можно оценить сочетаемость $F(x, y)$ для пары слов $x \in X$, $y \in Y$. Первая имеющаяся оценка — это $f(x, y)$, то есть некоторая базовая мера. Ей сопоставим единичный вес. Для некоторого слова $y_i \in c(x)$ из контекста x в качестве оценки $F(x, y)$ рассмотрим $f(x, y_i)$.

Эта оценка хороша, когда $y_i \sim y$, поэтому ей сопоставим вес $g(y_i, y)$ — вероятность того, что y_i и y взаимозаменяемы. Аналогично поступим со словами из $c(y)$. Тем самым имеется три группы оценок:

- $f(x, y)$ с весом 1;
- $f(x, y_i)$ для всех $y \in c(x)$ с весом $g(y, y_i)$;
- $f(x_i, y)$ для всех $x \in c(y)$ с весом $g(x, x_i)$.

В качестве окончательной оценки рассмотрим взвешенное среднее:

$$F(x, y) = \frac{f(x, y) + \Sigma_{y_j \in c(x)} f(x, y_i) g(y_i, y) + \Sigma_{x_i \in c(y)} f(x_i, y) g(x_i, x)}{1 + \Sigma_{y_i \in c(x)} g(y_i, y) + \Sigma_{x_i \in c(y)} g(x_i, x)}.$$

4. Эксперименты

Для оценки качества были проведены стандартные эксперименты — алгоритму предстояло из двух пар слов выбрать более сочетающуюся. Этот эксперимент в англоязычной литературе называется псевдодизамбигуация. Проанализированы ошибки алгоритма и исследована возможность представленных мер дополнять контексты ядер. В заключительной части эксперимента описывается перечень взаимозаменяемых слов данного ядра, то есть тех слов, которые получили наибольший вес.

4.1. Данные

В работе использовался корпус художественных текстов объемом 350 тыс. предложений (источник — Электронная библиотека М. Мошкова[1]), предобработанный морфоанализатором pymorphy2[2]. Из этого корпуса (далее корпус А) были выделены пары *прилагательное + существительное*, их оказалось около 157 тыс., среди которых более 80 тыс. уникальные. В одном эксперименте использовался дополнительный корпус художественной литературы объемом 11 млн предложений (из того же источника, далее корпус Б).

[1] URL: http://www.lib.ru.

[2] URL: http://pymorphy2.readthedocs.org/en/latest/.

Для первого эксперимента из корпуса A были извлечены следующие слова с частотой больше 5:

- 100 случайных существительных $N = \{n_i\}$;
- для каждого существительного — случайное прилагательное из его контекста $A = \{a_i\}$;
- для каждого выбранного прилагательного — следующее за ним прилагательное $X = \{x_i\}$ по частоте в корпусе;
- 100 случайных прилагательных $Y = \{y_i\}$.

Все пары (a_i, n_i) удаляются из корпуса. Задача алгоритма — их восстановить. Для каждой базовой меры считаются следующие оценки:

1. Сколько раз алгоритм предпочел a_i против x_i в паре с n_i:

$$S(X) = \#(F(a_i, n_i) > F(x_i, n_i));$$

2. Сколько раз алгоритм предпочел a_i против y_i в паре с n_i:

$$S(Y_1) = \#(F(a_i, n_i) > F(y_i, n_i));$$

3. К значению $S(Y_1)$ прибавляются те случаи, когда пара (y_i, n_i) действительно сочетается лучше, чем (a_i, n_i) (размечено вручную), полученное значение — $S(Y_2)$.

4. Если пара (y_i, n_i) также сочетается, то она удаляется из выборки, полученное значение — $S(Y_3)$ (размечено вручную).

Данный эксперимент проводился пять раз. В табл. 2 и 3 представлены математическое ожидание и среднеквадратическое отклонение оценок.

Таблица 2. **Математическое ожидание оценок $S(X)$, $S(Y_1)$, $S(Y_2)$, $S(Y_3)$**

$\mathbb{E}$	$S(X)$	$S(Y_1)$	$S(Y_2)$	$S(Y_3)$
Q	74,4	63,4	75,8	68,9
K	74,6	63,4	75,8	68,9
M	74,8	64,6	77,2	70,7
χ^2	67,6	60,6	70,6	62,4
P	74,4	63,4	75,8	68,9
Z	41,6	39,6	48,0	33,5
G	72,6	62,0	73,6	66,1

Таблица 3. **Среднеквадратическое отклонение оценок $S(X)$, $S(Y_1)$, $S(Y_2)$, $S(Y_3)$**

σ	$S(X)$	$S(Y_1)$	$S(Y_2)$	$S(Y_3)$
Q	3,2	3,5	1,4	2,0
K	3,0	3,7	1,7	2,6
M	2,6	3,4	1,7	2,6
χ^2	1,9	1,7	3,3	3,5
p	3,6	3,0	2,0	2,5
z	5,5	5,3	6,3	5,1
G	0,8	2,6	1,6	2,4

Затем аналогичный эксперимент был проведен для прилагательных. Из корпуса выделялись:

- 100 случайных прилагательных $A = \{a_i\}$;
- для каждого прилагательного — случайное существительное из его контекста $N = \{n_i\}$;
- для каждого выбранного существительного — следующее за ним существительное $X = \{x_i\}$ по частоте в корпусе;
- 100 случайных существительных $Y = \{y_i\}$.

Результаты представлены в табл. 4 и 5 (лучшие результаты выделены).

Таблица 4. **Математическое ожидание оценок $S(X)$, $S(Y_1)$, $S(Y_2)$, $S(Y_3)$**

E	$S(X)$	$S(Y_1)$	$S(Y_2)$	$S(Y_3)$
Q	76,8	63,4	66,2	64,0
K	**77,4**	64,8	67,6	65,5
m	77,2	65,4	68,2	66,2
χ^2	70,2	63,6	66,2	64,0
p	76,2	62,8	65,6	63,4
z	25,8	38,2	42,0	38,1
G	75,6	**66,6**	**69,4**	**67,4**

Таблица 5. **Среднеквадратическое отклонение $S(X)$, $S(Y_1)$, $S(Y_2)$, $S(Y_3)$**

σ	$S(X)$	$S(Y_1)$	$S(Y_2)$	$S(Y_3)$
Q	3,3	3,9	4,4	4,1
K	3,6	4,0	4,3	4,1
m	3,6	3,4	4,2	3,9
χ^2	3,9	**1,7**	2,4	**2,1**
p	3,7	4,7	5,2	4,9
z	**2,9**	3,1	**2,0**	3,0
G	4,9	2,6	3,5	3,3

Анализ ошибок алгоритма показал, что среди них можно выделить два класса.

Во-первых, при удалении пары (a_i, n_i) из корпуса вероятность того, что контексты a_i взаимозаменяемы с n_i, равна нулю. Это бывает, когда (a_i, n_i) употребляется в переносном значении, например *ходячий катехизис*, *дремучее равновесие*, *дикое зверство*.

Второй вариант ошибок — очень редкое сочетание слов, обнаружить которое статистически очень трудно. В представленном корпусе среди таких сочетаний *пылкое бегство*, *тонкая восприимчивость*, *сухая конвульсия* и пр.

4.3. Сортировка

Цель данного эксперимента — проверить возможность алгоритма извлекать новые сочетания для данного слова. Для этого потребуется корпус Б, состоящий из 11 млн предложений. Был составлен следующий план эксперимента.

1. Рассматриваются два существительных и два прилагательных: *дом*, *человек*, *красный*, *красивый*.

2. Для существительного ранжируются все прилагательные из корпуса А, не входящие в его контекст, согласно представленной оценке сочетаемости, для прилагательных соответственно ранжируются все существительные. Таким образом, составляется список потенциальных новых конструкций.

3. Если полученное словосочетание встретилось в корпусе Б, то оно помечается как хорошее.

В табл. 6 и 7 представлен процент хороших сочетаний среди 100 и 500 соответственно, получивших наибольшее значение меры сочетаемости (выделены лучшие результаты).

Таблица 6. **Процент новых сочетаний среди 100 лучших слов**

100	Q	K	m	χ^2	p	z	G
дом	**83**	79	80	16	80	3	53
человек	89	84	91	24	**94**	10	78
красный	79	78	**80**	43	77	3	77
красивый	**71**	69	70	29	73	1	65

Таблица 7. **Процент новых сочетаний среди 500 лучших слов**

500	Q	K	m	χ^2	p	z	G
дом	57	58	**59**	19	55	2	51
человек	73	73	**74**	27	73	8	70
красный	57	58	59	49	**61**	3	**61**
красивый	51	51	**52**	33	51	2	49

На рис. 1–4 представлены диаграммы, характеризующие распределение значений «хороших» слов. В качестве базовой меры ука-

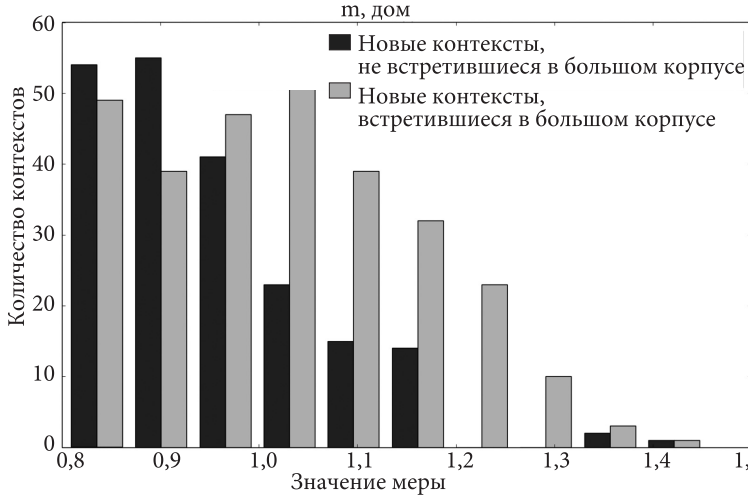

Рис. 1. Диаграмма значений меры на базе взаимной информации для слова *дом*

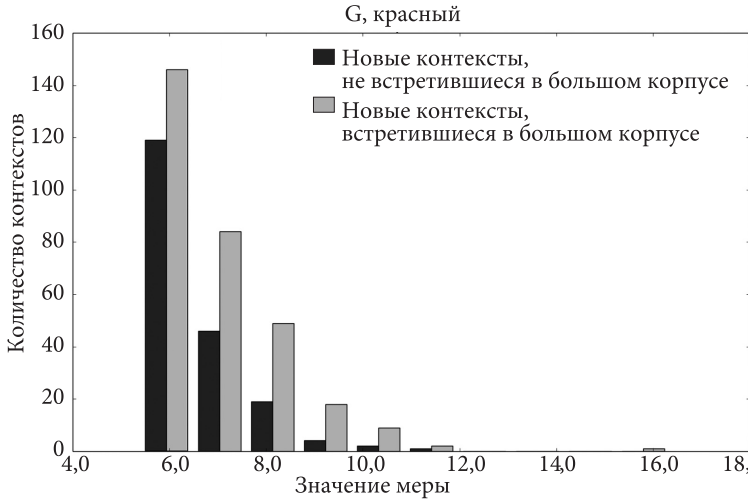

Рис. 2. Диаграмма значений меры на базе критерия Вульфа для слова *красный*

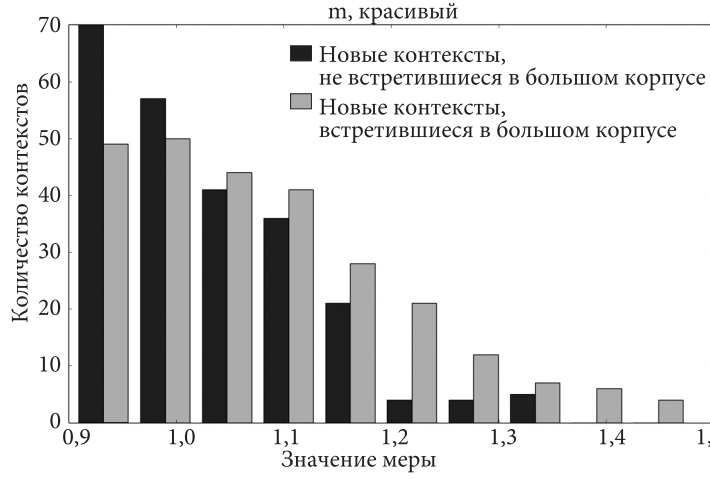

Рис. 3. Диаграмма значений меры на базе взаимной информации для слова *красивый*

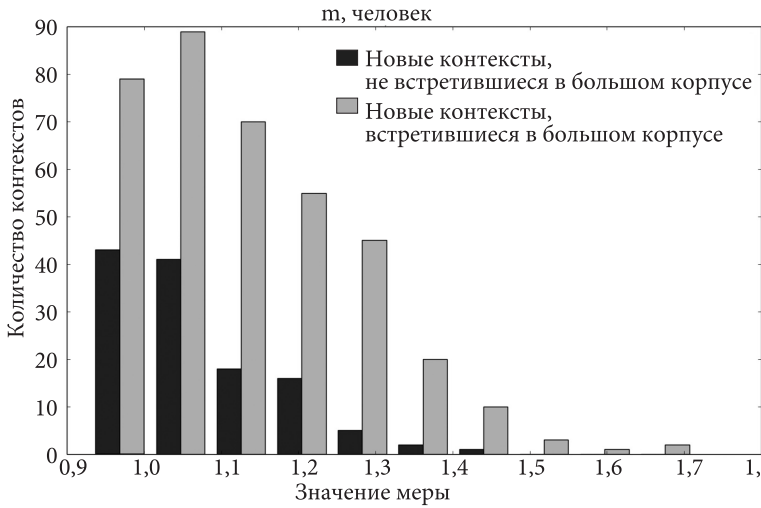

Рис. 4. Диаграмма значений меры на базе взаимной информации для слова *человек*

заны те, которые показали лучшие результаты в экспериментах по псевдодизамбигуации. Стоит отметить, что «хорошими» помечены далеко не все слова, сочетающиеся с ядром, ведь корпус Б не может содержать в себе все мыслимые конструкции. Некоторые слова, попавшие в сотню, хотя и не сочетаются с ядром, но косвенно связаны с его значением. Например, для существительного *человек* это слова *миловидный*, *хорошенький* и пр., которые попали в список через слово *девушка*. Эта псевдосвязь может оказаться полезной при снятии лексической омонимии.

4.5. Взаимозаменяемость

Чаще всего взаимозаменяемыми оказываются следующие группы слов: синонимы, гиперонимы, меронимы, антонимы, гипонимы или ассоциации. В табл. 8–11 для слов *дом*, *человек*, *красный*, *красивый* представлены списки из 10 наиболее близких по мере взаимозаменяемости слов. Наши результаты соответствуют результатам предшествующих исследований (см. например, [Weeds, Weir, McCarthy]), однако отличительной особенностью предложенной оценки взаимозаменяемости является возможность извлечения различных

Таблица 8. **Близкие по значению меры взаимозаменяемости пары для слова *человек***

человек	***g***	**связь**
женщина	0,124	гипон.
мужчина	0,073	гипон.
место	0,072	ассоц.
лицо	0,070	мерон.
вещь	0,069	ассоц.
девушка	0,069	гипон.
парень	0,066	гипон.
молодая	0,065	гипон.
жизнь	0,062	ассоц.
рука	0,054	мерон.

Таблица 9. **Близкие по значению меры взаимозаменяемости пары для слова *дом***

дом	***g***	**связь**
здание	0,095	синон.
квартира	0,079	мерон.
комната	0,056	мерон.
библиотека	0,055	гипон.
особняк	0,054	гипон.
гостиница	0,049	гипон.
семья	0,048	ассоц.
город	0,046	ассоц.
стена	0,045	мерон.
деревня	0,044	ассоц.

Таблица 10. **Близкие по значению меры взаимозаменяемости пары для слова *красный***

красный	***g***	**связь**
синий	0,070	согипон.
желтый	0,063	согипон.
белый	0,055	согипон.
яркий	0,047	ассоц.
розовый	0,041	согипон.
бледный	0,040	ассоц.
золотой	0,036	согипон.
зеленый	0,036	согипон.
коричневый	0,033	согипон.
тусклый	0,028	ассоц.

Таблица 11. **Близкие по значению меры взаимозаменяемости пары для слова *красивый***

красивый	***g***	**связь**
загорелый	0,064	ассоц.
прекрасный	0,063	синон.
худой	0,047	ассоц.
смуглый	0,046	ассоц.
хрупкий	0,045	ассоц.
странный	0,040	нет
обнажённый	0,038	ассоц.
усталый	0,038	нет
чужой	0,037	нет
миловидный	0,035	синон.

синтагматических отношений между лексическими единицами, что может оказаться полезным при дальнейшем анализе.

5. Заключение

В работе представлен новый метод оценки степени сочетаемости для пар слов, не встречающихся в корпусе совместно. Простота алгоритма и почти мгновенная скорость вычисления при относительно высоком качестве оценки делают его привлекательным для прикладных задач. Оценка производится в два этапа: сначала для некоторых пар слов в окрестности искомой пары вычисляется стандартная мера связи, основанная на таблице сопряженности, затем полученные оценки усредняются. Вопреки сложившемуся мнению, которое восходит еще к оригинальной статье А. Стефановича и С. Гриса, точный тест Фишера показал не самые лучшие результаты. В ряду наиболее надежных мер оказался коэффициент взаимной информации.

Полученные нами результаты могут получить развитие в нескольких направлениях, среди которых более подробный анализ конструкций, модификация мер, исследование других типов сочетаний, в частности *существительное + глагол*.

Литература

Апресян Ю. Д. Проспект активного словаря русского языка. М., 2010.

Пекар В. И. Дистрибутивная модель сочетаемостных ограничений глаголов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М., 2004.

A Neural Probabilistic Language Model / Y. Bengio [et al.] // Journal of Machine Learning Research. 2003. Vol. 3. P. 1137–1155.

Abney S., Light M. Hiding a semantic class hierarchy in a Markov model // Proceedings of the ACL-99 Workshop on Unsupervised Learning in NLP. 1999. P. 1–8.

Chen S. F., Goodman J. An empirical study of smoothing techniques for language modeling // Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

Dagan I., Lee L., Pereira F. C. Similarity-Based Models of Word Cooccurrence Probabilities // Machine Learning. Boston, 1999. Vol. 34. P. 43–69.

Erk K., Pad S., Pad U. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences // Computational Linguistics. 2010. Vol. 36, N 4. P. 723–763.

Fillmore Ch. J. The Mechanisms of Construction Grammar // Proceedings of the Berkeley Linguistic Society. 1988. Vol. 14.

Goldberg. A. E. Constructions: A Construction Grammar Approach to Argument Structure. Chicago; London, 1995.

Resnik P. Selectional preference and sense disambiguation // Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington, 1997. P. 52–57.

Seaghdha D. O. Latent variable models of selectional preference // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Morristown, 2010. P. 435–444.

Stefanowitsch A., Gries S. T. Collostructions: Investigating the interaction of words and constructions // International Journal of Corpus Linguistics. 2003. Vol. 8, N 2. P. 209–243.

Tian Zh., Xiang H., Zheng Q. A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation // Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics. Sofia, 2013. P. 1169–1179.

Tomasello M. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, Massachusetts, 2003.

Weeds J., Weir D., McCarthy D. Characterising measures of lexical distributional similarity // Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). Geneva, 2004. P. 1015–1021.

Д. Б. Тискин

К ФОРМАЛЬНОЙ СЕМАНТИКЕ ВЫСКАЗЫВАНИЙ О РАСПРЕДЕЛЁННЫХ ПРОПОЗИЦИОНАЛЬНЫХ УСТАНОВКАХ[1]

Аннотация. В статье рассматривается проблема, возникающая в семантике высказываний о пропозициональных установках. В случаях, когда субъект матричного глагола выражен сочинённой ИГ (*A* и *B*), а одна из ИГ во вложенной клаузе содержит численный или пропорциональный квантор *QX*, предложение в целом может быть высказыванием о распределённой пропозициональной установке. Это означает, что каждый из референтов *A* и *B* имеет соответствующую установку *de re* лишь относительно подмножества *X*, хотя вместе их установки охватывают *X* в целом. Предлагается описание таких случаев с помощью понятия распределённого убеждения, позаимствованного из эпистемической логики, однако такого объяснения недостаточно: требуется объяснить нечувствительность высказываний о распределённых установках к возможным расхождениям в убеждениях между *A* и *B*. Для этого предложен механизм, имеющий прагматическую природу.

Ключевые слова. Логический анализ языка, пропозициональные установки, *de re*, дистрибутивность, «специфические непрозрачные» чтения, группа детерминатора, распределённые убеждения.

Daniel B. Tiskin

TOWARDS A FORMAL SEMANTICS OF DISTRIBUTED ATTITUDE REPORTS

Abstract. The paper deals with a problem in the semantics of propositional attitude reports. The issue is that when the matrix subject is a conjoined NP (such as *A* and *B*) and one of the NPs in the attitude clause is headed by a numerical or a proportional determiner *QX*, the sentence as a whole may be read as a distributed attitude report. This means that each of *A* and *B* has a *de re* attitude towards a proper subset of *X*, although together *A*'s and *B*'s attitudes cover the whole of *X*. I suggest that such cases be described in terms of

[1] Статья подготовлена при поддержке гранта РГНФ, проект № 14-03-00650.

distributed belief, a notion taken from epistemic logic; this is however insufficient since one has to explain why the semantics is insensitive to possible disagreement in *A*'s and *B*'s beliefs. I offer a pragmatic mechanism to take care of that.

Keywords. Propositional attitudes, *de re*, distributivity, "specific opaque" readings, determiner phrase, distributed belief.

Условимся говорить, что пропозициональная установка (ПУ) является *распределённой*, если она принадлежит группе агентов *A* и является *de re* для группы индивидов *G* таким образом, что хотя бы один из членов *A* имеет соответствующую установку лишь относительно собственного подмножества *G*. В настоящей статье на примере чтений (1c) и (3) для предложения (1) показано, что такие чтения представляют проблему для анализа; этому посвящён § 2. В § 3 обсуждается возможность решить названную проблему путём введения нового дистрибутивного оператора. Наконец, § 4 мотивирует, а § 5 развивает прагматически ориентированный анализ, сближающий исследуемую проблему с одним из типов «прозрачных» чтений.

1. Суммативные чтения

В недавней работе З. Г. Сабо [Szabó, 2010] привёл примеры высказываний о пропозициональных установках (ВПУ), подобные (1).

1. Alex and Bob believe that twenty aliens live in their neighbourhood.

(a) 'Для двадцати инопланетян *x* верно[2]: Алекс и Боб думают, что *x* живёт по соседству' (*de re*).

(b) 'Алекс и Боб думают: двадцать инопланетян живут по соседству' (*de dicto*).

(c) 'Относительно двадцати объектов [например, обычных людей] Алекс и Боб думают, что они инопланетяне и живут по соседству' («специфическое непрозрачное» чтение).

Длительная традиция изучения ВПУ, начиная с Б. Рассела [Russell], различает у них два возможных чтения, проиллюстрированных на примере (1) как (1a) и (1b).

[2] Переменные, обозначаемые строчными буквами, пробегают по атомарным индивидам (индивидам в классическом смысле). Далее заглавной *X* мы будем обозначать переменную, пробегающую по множественным индивидам (включая атомарные).

Интерес для Сабо, однако, представлял ещё один тип чтений, впервые отмеченный Дж. Фодор [Fodor] и не признаваемый многими авторами: (1c)[3]. Согласно Сабо, такое значение (1) может иметь, к примеру, в ситуации, когда Алекс и Боб рассматривают фотографии разных людей, отвечая о каждом на два вопроса: является ли это существо инопланетянином и живёт ли по соседству. Впрочем, нас в данной статье «специфические непрозрачные» чтения как таковые занимать не будут[4].

Для нас важно, что (1c) является, в терминологии Сабо, **суммативным** высказыванием (summative report). Данного до сих пор описания ситуации, вообще говоря, недостаточно для того, чтобы исключить, что (1c) является частным случаем чтения *de dicto*. Для этого требуется показать, что *twenty* не является элементом пропозициональной установки Алекса и Боба. Поэтому Сабо добавляет, что из 20 объектов, о которых идёт речь, Алекс назвал инопланетянами, скажем, 11, а его сосед Боб — ещё девять. (Таким образом, вполне возможно, что нет ни одного объекта, о котором было бы доподлинно известно, что его считают инопланетянином и Алекс, и Боб, но (1) всё равно может быть истинно.) Поскольку ни Алекс, ни Боб не знают, что в сумме речь идёт о двадцати объектах, *twenty* не является элементом ПУ как таковой, а привносится в (1) говорящим, который производит суммирование; поэтому (1c) не является случаем *de dicto*[5].

[3] Своим наименованием этот тип чтений обязан следующим соображениям. Их называют специфическими, поскольку речь идёт об объектах, которые могут быть указаны и перечислены, в отличие от (1b), где ничего не говорится о том, кто именно инопланетянин. Непрозрачными такие чтения называют потому, что они, в отличие от обычных *de re*, не допускают замены термина alien на экстенсионально эквивалентный ему.

[4] О них, кроме работ [Szabó, 2010; Szabó, 2011] (последняя касается только одноагентного случая; см. сноску 5), см. критику тезиса о существовании таких чтений в [Ben-Yami] и критику подхода Сабо (а также, по-видимому, пока не опубликованного подхода И. Франсеза), основанную на специфике детерминаторов типа *most* ‘большинство’, в [Santorio]. Некоторые критические положения Бен-Ями независимо повторены в статье [Tiskin], где в сноске впервые упомянута и проблема, которой посвящена данная статья.

[5] Согласно Сабо, существуют и такие суммативные высказывания, в которых носителем ПУ является только один агент; к примеру, если «инопланетян» выявлял один только Алекс, не считая при этом своих подозреваемых, то число 11 может принадлежать только говорящему, а предикат ‘инопланетянин’ — только Алексу, так что (2) также будет выступать в «специфическом непрозрачном» чтении.

Далее мы рассматриваем (1) только в суммативном чтении *de re*[6]:

3. ‘Для двадцати инопланетян *x* верно: Алекс и Боб «совместно» думают, что *x* живёт по соседству’.

2. Постановка проблемы

Заметим теперь, что помимо собственно проблемы «специфических непрозрачных» чтений, (1) в чтении (1c) представляет интерес и с другой точки зрения. На первый взгляд оно кажется примером **кумулятивного** чтения [Scha], подобным (4) или (5).

4. Alex and Bob (together) ate twenty cookies.
   Алекс и Боб (вместе) съели двадцать штук печенья.
5. Alex and Bob (together) know [~ can identify] twenty terrorists.
   Алекс и Боб (вместе) знают [то есть могут назвать] двадцать террористов.

Эти чтения характеризуются тем, что основной предикат предложения выполняется **множественным индивидом** (иначе группой), обозначаемым подлежащим, но не обязательно какой-либо собственной частью этого индивида (соответственно, собственным подмножеством группы). Так, вполне возможно, что Алекс съел 11 штук, а Боб — другие девять. (Здесь хорошо видно, что пересечение этих двух множеств, скорее всего, пусто. Для (5), как и для (1c), пересечение выглядит вполне допустимым, если только **объединение** насчитывает двадцать объектов.)

В чём же состоит отличие (3) от (4)–(5)? Согласно традиционной трактовке глаголов ПУ, они являются кванторами, пробегающими по множествам возможных миров, заданным тем или иным отношением достижимости (в случае глагола *think* ‘думать’ это множество **доксастических** альтернатив — миров *v*, относительно которых носитель ПУ не может на основании своих убеждений исключить, что живёт в *v*)[7]:

---

2. Alex believes that eleven aliens live in his neighbourhood.
   Алекс считает, что одиннадцать инопланетян живут по соседству.

[6] Сабо [Szabó, 2010, p. 36] косвенно указывает на возможность суммативных чтений только для ВПУ de re.

[7] dox(*x*,*w*) — множество миров, доксастически достижимых для индивида *x* в мире *w*.

6. $[\![\text{think}]\!]^{w,g} = \lambda p_{\langle s,t\rangle}.\ \lambda x_e.\ \forall v \in \text{dox}(g(x), w)\colon (p(v) \equiv \mathbf{1})$.

В таком случае при попытке определить, что означает для **множественного** индивида выполнять предикат ‘думает, что *p*’, возникает необходимость указать, как формируется множество миров, по которым пробегает квантор. Если считать множества доксастических альтернатив **атомарных** индивидов (то есть, собственно говоря, отдельных людей) данными, достаточно очевидным выглядит вариант

7. $[\![\text{think}]\!]^{w,g} = \lambda p_{\langle s,t\rangle}.\ \lambda X_e.\ \forall v \in \bigcap_{ai \in g(X)} \text{dox}(a_i, w)\colon (p(v) \equiv \mathbf{1})$.

Вариант (7) сводится к тому, что убеждениями группы считаются пропозиции, истинные на пересечении доксастических множеств всех участников группы. Таково и стандартное определение *распределённого* знания [Ditmarsch, Hoek, Kooi; Roelofsen], хотя случай знания имеет существенное отличие от случая убеждения: поскольку знание — *фактивная* ПУ, то есть знать можно только то, что истинно, — пересечение эпистемических множеств участников любой группы непусто, так как в нём есть как минимум один мир (действительный). Убеждения же могут быть ложными, поэтому в пределах группы убеждения одного участника могут в чём-то противоречить убеждениям другого; тогда пересечение доксастических множеств участников группы пусто. Обычно считается, что квантор всеобщности на множестве возможных миров, как и вообще кванторы всеобщности, обращает любое подкванторное выражение в истину при пустоте домена. Поэтому в случае, когда, к примеру, Алекс верит в Деда Мороза, а Боб нет, формализация (3) как случая кумулятивной предикации (8)[8] будет истинна тривиальным образом (при условии существования 20 инопланетян), вне зависимости от того, кого Алекс и Боб считают своими соседями.

8. [∃ [20 aliens]] λX [$\bigwedge_{\{\text{Alex;Bob}\}}$ $\text{think}_{(7)}$ [*X* lives in the neighbourhood]].

Таким образом, пока что в нашем распоряжении нет ни одного варианта формализации (3).

[8] Заметим, что *twenty aliens* ‘двадцать инопланетян’ мы рассматриваем как квантор существования по множественным индивидам [∃ [20 aliens]] ‘существует такой множественный индивид, состоящий из 20 инопланетян, что…’, а не как численный квантор [$\exists_{20}$ aliens] ‘множество инопланетян, таких что…, непусто и насчитывает 20 элементов’.

3. Нестандартная дистрибутивность?

Значение (3) удалось бы выразить, если бы в семантике существовал оператор ∂', подобный **дистрибутивному** оператору [Lasersohn, p. 1150; восходит к Link]:

9. $\partial = \lambda P.\ \lambda X.\ \forall x \sqsubseteq X(P(x))$,

но отличающийся от него заменой квантора всеобщности на квантор существования, так что интерпретацией (1) в чтении (3) было бы:

10. $[\![[\exists\ [20\ \text{aliens}]]\ \lambda X\ [\Lambda_{\{\text{Alex, Bob}\}}\ \partial'(\text{think})\ [X\ \text{lives in the neighbourhood}]]]\!]^{w,g} = \exists X\colon |X| = 20.\ \iota Y\colon Y = \text{Алекс} \oplus \text{Боб}.\ \exists x \sqsubseteq Y.\ \forall v \in \text{dox}(g(x),w)$ (**живёт_по_соседству**(X)).

Такой оператор позволил бы объяснить, почему замена в (1) *Alex and Bob* ‘Алекс и Боб’ на *Alex or Bob* ‘Алекс или Боб’ исключает суммативное чтение: дизъюнкция не обозначает множественного индивида (обозначенного в (10) как Y). Трудность здесь состоит в том, что введение в семантику оператора ∂' недостаточно мотивировано. Так, например, ничто не свидетельствует о возможности его появления в монологе (11).

11. Мэри и Джон могут решить эту задачу. ? Если предложить её им, то либо Мэри решит её, либо Джон.
12. Мэри и Джон могут решить любую задачу. ?$^{\text{OK}}$Если предложить её им, то либо Мэри решит её, либо Джон.

Аномальность второго предложения в (11) свидетельствует о том, что интерпретация первого предложения, вызываемая присутствием ∂', недоступна (в отличие от обычной дистрибутивной, а также кумулятивной интерпретации; последнюю можно выявить, если за первым предложением следует: *Если предложить её им, то они вместе найдут решение*). Впрочем, замена *эту задачу* на *любую задачу*, по-видимому, приводит к тому, что необычное дистрибутивное чтение возникает. Это заставляет предположить, что ∂', если он существует, возможен только в случаях, когда выбор атомарного индивида в составе множественного **зависит** от какого-то иного выбора; в (11) такого выбора нет, но в (12) он вводится группой детерминатора *любую задачу*. Если попытаться предположить, что является причиной такого ограничения, то речь может идти о праг-

матических факторах: первое предложение в (11) не только недостаточно информативно, поскольку не сообщает о том, кто именно может решить задачу (хотя есть по крайней мере один человек, который решит её в одиночку), но и многозначно (так как имеет ещё дистрибутивную и кумулятивную интерпретации); при этом существует способ избежать омонимии, использовав дизъюнкцию *Мэри или Джон*. Напротив, первое предложение в (12) имеет чтение, при котором для каждой задачи существует (как минимум) один человек из двух, который может её решить в одиночку, но при этом нет никого, кто мог бы в одиночку решить все задачи; поэтому произносящего (12) нельзя упрекнуть хотя бы в недостаточной информативности.

4. Утверждение по умолчанию?

Частным случаем расхождения во мнениях, упомянутого в конце § 2, является и тот, при котором относительно некоторого инопланетянина Алекс полагает, что тот живёт по соседству, а Боб — что вдалеке. Этот случай отличается от того, где Алекс полагает, что инопланетянин живёт по соседству, а Боб **не имеет мнения** по этому вопросу, то есть не исключает ни тот, ни другой вариант. С точки зрения семантики возможных миров, в случае уверенности Боба пересечение доксастических множеств Алекса и Боба пусто, а в случае неуверенности Боба оно непусто (если нет других разногласий) и включает те из альтернатив Боба, в которых инопланетянин живёт по соседству (то есть те, которые являются альтернативами Алекса). Таким образом, если бы вариант (7) был верен, следовало бы ожидать неприемлемости (1) в чтении (3) при наличии хотя бы одного инопланетянина, о котором Алекс думает, что он живёт по соседству, а Боб — что вдалеке. (Повторим, что это не означает неприемлемости (3) в случае, когда некоторые или даже каждый из инопланетян характеризуется как сосед одним из двух агентов и никак не характеризуется другим.)

Аналогичное рассуждение для одноагентного случая (см. сноску 5) мы находим у Х. Бен-Ями [Ben-Yami, с. 180], который предполагает, что (2) может утверждаться, если нет сведений о том, что Алекс считает иначе (например, он не считал своих подозреваемых, но убедился бы, что их 11, если бы сосчитал). Если же Алекс судит по

фотографиям и **ошибочно** полагает, что две фотографии изображают одного и того же человека, то он сам выразил бы своё убеждение как

13. Ten aliens live in the neighbourhood.
По соседству живут десять инопланетян.

В этом случае, как полагают Бен-Ями и его информанты, (2) утверждаться не может. Поскольку при таком изменении сценария число действительных объектов, к которым относится убеждение Алекса, не изменилось, рассуждение Бен-Ями может служить аргументом в пользу того, чтобы считать суммативное чтение (для одноагентных ВПУ) частным случаем *de dicto*.

В то время как в рассуждении Бен-Ями (вполне определённое) представление Алекса о числе изображённых, 10, отличается от их действительного числа, 11, в нашем случае возможно, что различаются представление Алекса о том, кто из изображённых на фотоснимках живёт по соседству, и представление Боба о том же. Как уже сказано, если насчёт того или иного изображённого (например, Тома) Алекс составил твёрдое мнение о том, что он живёт по соседству, а Боб лишь сомневается, никакого затруднения не возникает, и на пересечении доксастических множеств Алекса и Боба Том будет соседом. Если же Боб твёрдо решил, что Том не сосед, пересечение будет пусто.

Если рассуждение Бен-Ями верно, неприемлемость одноагентных «специфических непрозрачных» чтений для некоторых носителей может быть связана с их неготовностью расширить условия употребимости (2) в чтении *de dicto* на случаи, когда Алекс не имеет никакого убеждения относительно численности тех, кого считает инопланетянами. Впрочем, остаётся неясным, как трактовать ВПУ, где суммативность, по крайней мере на первый взгляд, сочетается с обыкновенным *de re*, то есть как раз таки обсуждаемые здесь случаи наподобие (1) в чтении (3).

Хотя для того, чтобы выяснить, насколько приемлемо суммативное чтение в тех многоагентных случаях, где имеет место расхождение во мнениях, требуется отдельное исследование, можно предположить, что при интерпретации суммативных чтений **«по умолчанию»**, то есть пока не стало известно обратное, принимается, что расхождения нет и пересечение доксастических множеств

непротиворечиво. Как только такое предположение сделано, вариант анализа (7) позволяет получить суммативные чтения *de re*.

Напомним, что в нашем распоряжении имеется весомый аргумент в пользу того, что обращение к умолчаниям происходит по крайней мере тогда, когда расхождения во мнениях относительно содержания **данной** ПУ у агентов нет. Действительно, как уже было сказано, для пустоты пересечения доксастических множеств достаточно, чтобы Алекс и Боб не соглашались хотя бы в чём-то, то есть чтобы среди убеждений Алекса существовала пропозиция (например, ‘Дед Мороз существует’), **отрицание** которой входит в число убеждений Боба в действительном мире @:

14. $\exists p\ (\mathrm{dox}(\mathrm{Alex},@) \subseteq [\![p]\!]^s \wedge \mathrm{dox}(\mathrm{Bob},@) \subseteq \overline{[\![p]\!]^s})$.

Поскольку в речевой практике нет никакой возможности удостовериться, что убеждения одного агента ни в чём не противоречат убеждениям другого, самая возможность суммативного чтения для (1) указывает на наличие какой-то иной стратегии приписывания убеждений[9].

5. Умолчания и «прозрачные» чтения

Какой же формальный механизм может стоять за обращением к умолчаниям? В основе наброска его описания, который мы предлагаем, лежит идея М. Швагер (Кауфман) [Schwager][10], касающаяся ещё одного типа чтений для ВПУ, открытого Дж. Фодор, — так называемых «неспецифических прозрачных». Для примера (1) такое чтение выглядело бы как

15. ‘Существует некоторое свойство *P* [например, ‘быть маленьким зелёным человечком’], которое, по мнению говорящего, является достаточным условием принадлежности к классу инопланетян и для

[9] В дальнейшем мы будем предполагать, что расхождение во мнениях по **релевантному** вопросу (например, по вопросу о том, кто из изображённых живёт по соседству, в сценарии для (3)) приводит к неприемлемости суммативных чтений. Как указано выше, этот вопрос нуждается в эмпирическом исследовании, подобном проведённому Бен-Ями.

[10] Ср. сходные рассуждения в статье [Sudo]. На русском языке см. [Куслий, с. 157–159].

которого также верно: Алекс и Боб считают, что 20 обладателей *P* живут по соседству'[11].

Мы не будем подробно останавливаться на том, как Швагер мотивирует свой анализ, и попросту кратко опишем его. Для начала вводится отношение близости между возможными мирами по мере их сходства с @: $w \leq_{@} v$ означает, что *w* не больше отличается от @, чем от него отличается *v*. Далее утверждается, что чтения типа (15) допустимы при следующих условиях: во всех наиболее похожих на @ мирах, где существуют инопланетяне, 1) существуют маленькие зелёные человечки, 2) все маленькие зелёные человечки являются инопланетянами и 3) вариант (1) с заменой *aliens* 'инопланетяне' на *little green men* 'маленькие зелёные человечки' истинен *de dicto*. В этой формулировке для нас важно, что существование инопланетян и/или маленьких зелёных человечков в **самом** @ для истинности (15) не требуется.

Распространим теперь этот способ рассуждения на (3). Предположим, что не слишком строгий интерпретатор готов рассматривать (1) в чтении (3) как высказывание о таких **наиболее похожих на** @ мирах, в которых существует хотя бы одна непротиворечивая пропозиция, являющаяся предметом распределённого убеждения Алекса и Боба. Требование непротиворечивости вводится, чтобы обеспечить непустоту пересечения доксастических множеств Алекса и Боба. Тогда, если пересечение этих множеств непусто уже в самом мире @, (1) в чтении (3) также будет истинным: для @ верно, что

$\neg\exists w \neq @ : w \leq_{@} @.$

Необходимость проверки пересечения доксастических множеств на непустоту можно указать в семантике глагола ПУ в качестве условия, при котором определена соответствующая ему функция:

17. $[\![\text{think}]\!]^{w,g} = \lambda p_{\langle s,t\rangle}.\ \lambda X_{e}\ [\bigcap_{ai \in g(X)} \text{dox}(a_i,w) \neq \emptyset]$: $\forall v \in \bigcap_{ai \in g(X)} \text{dox}(a_i,w)$: $(p(v) \equiv \mathbf{1})$ (ср. (7)).

---

[11] Этот тип чтений более заметен на примерах типа

16. Charley wants to buy a coat like Bill's. [Fodor, p. 226]
'Чарли хочет купить куртку некоторого типа *P*, не зная, что у Билла такая же, но говорящий это знает'.

В заключение заметим, что суммативное чтение в одноагентном случае (2) может быть объяснено аналогичным образом, в полном соответствии с данной выше семантикой для (15) (при условии, что численные кванторы анализируются, как показано в (8)). Для этого следует допустить, что свойство ‘быть множественным индивидом, состоящим из инопланетян, которых Алекс видел на фотоснимках’ влечёт свойство ‘быть множественным индивидом, состоящим из 11 инопланетян’ везде, где последнее непусто (то есть где есть релевантные 11 инопланетян)[12].

Литература

Куслий П. С. Проблема третьего прочтения и семантика сообщений о верованиях // Философия языка и формальная семантика. М., 2013. С. 129–160.

Ben-Yami H. Bare quantifiers? // Pacific Philosophical Quarterly. 2014. Vol. 95, N 2. P. 175–188.

Ditmarsch H. van, Hoek W. van der, Kooi B. Dynamic epistemic logic // Synthese Library. Dordrecht, 2007. Vol. 337.

*Fodor J. D. T*he Linguistic Description of Opaque Contexts. Garland, 1979.

*Lasersohn P. M*ass Nouns and Plurals // Semantics: An International Handbook of Natural Language Meaning / eds C. Maienborn, K. von Heusinger, P. Portner. De Gruyter Mouton, 2011. Vol. II.

Link G. The logical analysis of plurals and mass terms: A lattice-theoretical approach // Meaning, use and interpretation of language / eds R. Bäuerle, C. Schwarze, A. von Stechow. de Gruyter, 1983. P. 303–323.

Roelofsen F. Distributed knowledge // Journal of Applied Non-Classical Logics. 2007. Vol. 17, N 2. P. 255–273.

Russell B. On denoting // Mind. 1905. Vol. 14, N 56. P. 479–493.

Santorio P. Descriptions as variables // Philosophical Studies. 2013. Vol. 164, N 1. P. 41–59.

Scha R. Distributive, collective and cumulative quantification // Truth, Interpretation, and Information: Selected Papers from the Third Amsterdam Colloquium. De Gruyter, 1984. P. 131–158.

Schwager M. Speaking of qualities // Proceedings of SALT. 2009. Vol. 19. P. 395–412.

Sudo Y. On de re predicates // Proceedings of WCCFL 31. 2014.

[12] См. [Tiskin], где мы предположили, что «специфические непрозрачные» чтения могут на поверку оказаться разновидностью «неспецифических прозрачных».

Szabó Z. G. Bare quantifiers // Philosophical Review. 2011. Vol. 120, N 2. P. 247–283.

Szabó Z. G. Specific, yet opaque // Logic, Language and Meaning. Springer, 2010. P. 32–41.

Tiskin D. Specific Opaque Readings and Proportional Determiners. Semantics Archive draft, 2014. URL: http://semanticsarchive.net/Archive/jFhODZmZ/ (accessed: 12.01.2016).

Т. Г. Скребцова

К ПРОБЛЕМЕ ВЫДЕЛЕНИЯ РЕГУЛЯРНЫХ СЕМАНТИЧЕСКИХ КОМПОНЕНТОВ В РУССКОЙ ГЛАГОЛЬНОЙ ЛЕКСИКЕ

Аннотация. В статье описаны некоторые регулярные семантические компоненты, представленные в значениях русских глаголов. После обсуждения ряда дискуссионных вопросов, связанных с методикой их выделения, автор постулирует пять категориальных сем: 'действие', 'движение', 'состояние', 'свойство' и 'отношение', при помощи которых все множество глагольных значений делится на непересекающиеся классы. Каждая из данных сем может иметь при себе дополнительную конкретизацию, обусловленную характером обозначаемой ситуации (физической, ментальной, речевой и т. д.) и/или типом субъекта. На категориальные семы наслаиваются прочие разнообразные семантические компоненты, выражающие фазовые значения или обозначающие качественные, количественные, пространственные или временные аспекты обозначаемой ситуации, а также отражающие некоторые другие ее характеристики.

Ключевые слова. Лексическая семантика, русский язык, глагол, компонентный анализ, сема.

Tatiana G. Skrebtsova

ON THE PROBLEM OF IDENTIFYING REGULAR SEMANTIC COMPONENTS OF RUSSIAN VERBS

Abstract. The paper aims to identify regular semantic components of Russian verbs, unfolding with the discussion of a few controversial issues related to the problem. The author posits five basic semantic components, namely 'action', 'motion', 'state', 'property' and 'relation', which break the whole set of Russian verbs' meanings into five non-overlapping classes. Each of the semantic components can be further specified depending on the nature of the situation (physical, mental, communicative, etc.) and/or the type of the agent. This scaffolding structure is fleshed out with an array of different semantic components,

including phase characteristics and manifold qualitative, quantitative, locative or temporal specifications as well as other aspects of the situation at hand.

Keywords. Lexical semantics, Russian language, verb, componential analysis, semantic component.

Задача, вынесенная в заглавие статьи, встала перед автором в контексте практической работы, связанной с автоматической обработкой текста[1]: известно, что наличие в значении глагола определенных семантических компонентов способно «предсказывать» (хотя и не «гарантировать») появление в предложении соответствующих синтаксических актантов или сирконстантов (с особой наглядностью это можно видеть в случае некоторых способов глагольного действия). Этот вопрос достаточно широко изучался в зарубежной, преимущественно англоязычной, лингвистике, а в последние годы появились соответствующие публикации и на материале русского языка. В настоящей статье, однако, нас интересует несколько иной ракурс этой проблемы, исторически восходящий к исследованиям в области компонентного анализа и теории семантического поля. В идеале нам хотелось бы выделить набор сем, которые регулярно встречаются в значении русских глаголов, и затем исчислить их возможные комбинации, то есть создать целостное, системное, национально-специфичное описание глагольной семантики русского языка. Эта формулировка нуждается в ряде пояснений.

Прежде всего следует отметить, что речь не идет о компонентном анализе в его классическом варианте. Мы не стремимся разлагать значения глаголов на семы до конца, без остатка, поскольку опыт показал, что их число будет огромным и, более того, потенциально бесконечным из-за принципиальной открытости лексической системы. Используя терминологию Каца и Фодора [Katz, Fodor], можно сказать, что нас интересуют только семантические маркеры, но не семантические различители. Иными словами, задача заключается в том, чтобы выявить в глагольных значениях только регулярные семантические компоненты, причем регулярность понимается в весьма «слабом» смысле — как наличие одинаковой семы хотя бы

[1] Работа финансировалась из средств гранта «Проведение фундаментальных исследований по приоритетным направлениям Программы развития СПбГУ», шифр проекта 31.37.105.2011.

у двух значений разных глаголов (ср. аналогичную интерпретацию данного понятия применительно к полисемии в [Апресян]).

Как выделяются семантические компоненты в значении глагола? Важно подчеркнуть, что мы не опираемся исключительно на словарную дефиницию — метод, взятый за основу при составлении таких лексикографических трудов, как [Караулов; Русский семантический словарь …] (см. также [Лукин]). Определение значения в толковом словаре — всего лишь один из источников, причем далеко не главный. Семы определяются исходя из принадлежности значения глагола к тому или иному семантическому полю (подполю, группе, подгруппе) и на основе анализа его места в нем (ней). При этом за основу берется полное, подробное и хорошо структурированное описание глагольной лексики в IV томе Русского семантического словаря под редакцией Н. Ю Шведовой [Русский семантический словарь … , 2007]; также используется словарь [Лексико-семантические группы …]. Необходимо специально оговорить тот факт, что выделенные нами семы лишь отчасти связаны с названиями лексико-семантических групп в том или другом словаре.

Деление целостной семантики глагола на дискретные значения — еще один дискуссионный вопрос, практическое решение которого обусловлено выбором конкретного толкового словаря. Мы остановились на Большом толковом словаре русского языка под редакцией С. А. Кузнецова [Большой толковый словарь …], в силу того что он наиболее адекватно отражает современное словоупотребление и имеет разумный с точки зрения нашей задачи объем. Отдельно рассматривались не только собственно значения, но и их оттенки (которые приравнивались к самостоятельным значениям): таким образом, корректнее было бы говорить о лексико-семантических вариантах слова, но мы все же будем придерживаться более традиционного и привычного термина «значение».

Наконец, стоит упомянуть об определенной условности, связанной с выделением в значении слова тех или иных семантических компонентов. Здесь может быть уместно провести аналогию с усилиями Анны Вежбицкой по разработке естественного семантического метаязыка. Как известно, в качестве словаря этого метаязыка используются так называемые семантические примитивы — слова, выражающие некие первичные, элементарные смыслы, при помощи которых автор пытается описывать значения других, семантически

более сложных слов. Но поскольку в языке объективно не существует таких понятий, как семантически простое или семантически сложное слово, при выборе лексической единицы на роль семантического примитива немаловажным оказывается такой прагматический критерий, как ее способность участвовать в разложении значений большого числа слов (это позволяет описывать лексическую систему языка емко и экономно). Таким образом, понятие примитива у Вежбицкой является не абсолютным, а относительным.

Нечто похожее имеет место и в нашем исследовании. Разумеется, у нас иная задача, и выделенные семы не претендуют на роль примитивов. Но хочется подчеркнуть принципиальную неокончательность списка регулярных сем, его возможную вариативность, обусловленную недискретностью лексического значения. Есть определенная степень свободы в том, чтобы поделить семантическое пространство так или иначе, вычленить некоторый его фрагмент в качестве самостоятельной семы или ограничиться выделением более крупного семантического компонента. К примеру, нужно ли разделять состояния ментальные (*мечтать*) и эмоциональные (*любить*)? На этот вопрос невозможно дать единственно правильный ответ, поскольку есть промежуточные случаи, которые можно отнести как к одной, так и к другой категории, ср. *успокаиваться* (‘чувствовать полное удовлетворение достигнутым, сделанным’), *перемениться* (‘изменить свое отношение к кому-л.’), *замирать* (‘терять на время способность совершать какие-л. действия под влиянием страха, волнения и т. п.’). Так как подобных примеров, попадающих между различными категориями, множество, появление критических замечаний неизбежно, вне зависимости от принятого решения. В любом случае, перечень регулярных сем формируется эмпирически, исходя из языкового материала, а не подгоняется под априорную логическую схему. Именно такова, кстати, современная практика выделения лексико-семантических полей — в отличие от ранних опытов соответствующих описаний у неогумбольдтианцев (ср., например, идеографический словарь [Hallig, Wartburg], реализующий концепцию универсальной понятийной сетки, наброшенной на мир слов).

Обозначив теоретические предпосылки настоящего исследования, перейдем к характеристике основных регулярных семантических компонентов в значениях русских глаголов. Заметим, что мы анализируем исключительно глаголы несовершенного вида. Начнем

с категориальных сем, позволяющих осуществить наиболее общее разбиение глагольных значений на крупные непересекающиеся классы. Таковы, с нашей точки зрения, семы ‘действие’, ‘состояние’, ‘движение’, ‘отношение’ и ‘свойство’. Уже этот первый шаг наглядно демонстрирует, что наше описание не повторяет структуру существующих семантических словарей. Не имея, таким образом, возможности сослаться на авторитетный источник, попытаемся обосновать свою позицию.

Самый многочисленный класс образуют глаголы с категориальной семой ‘действие’, что вполне закономерно. Мы не разграничиваем понятия действия, деятельности и процесса, так как границы между ними весьма расплывчаты. В самом общем смысле ‘действие’ противопоставляется всем остальным семам (кроме ‘движения’) по признаку динамичности/статичности. Действия далее подразделяются на физические, ментальные, информационные, речевые, социальные, физиологические, природные и некоторые другие. Соответственно в большинстве случаев выделяется не общая сема ‘действие’, а более частная сема, ср: ‘действие: соц.’ или ‘действие: реч.’ и т. п. В то же время семантика некоторых глаголов носит столь общий характер, что конкретизация невозможна: ярким примером могут служить фазовые глаголы, ср. *завершать*, *возобновлять* и пр.

Сема ‘движение’ фиксируется у глаголов, обозначающих субъектное, объектное и субъектно-объектное перемещение, а также телодвижения, мимику и жесты (в этой части своего описания мы следуем за Русским семантическим словарем [Русский семантический словарь … , 2007]). Лексико-семантическое поле глаголов движения, как известно, отличается высокой структурированностью, что позволяет выделить целый ряд подчиненных сем, которые либо конкретизируют характер движения (‘направленное’, ‘ненаправленное’, ‘вверх’, ‘вниз’ и некоторые другие), либо представляют его с ориентацией на исходную, промежуточную или конечную точку.

Еще один большой класс образуют глаголы с семой ‘состояние’. Среди состояний выделяются физические, ментальные, эмоциональные, информационные, социальные, материальные, функциональные, физиологические, природные. Из многих названий состояний следует тип субъекта, ср. ‘состояние: физиол.’ или ‘состояние: прир.’. Наиболее заметное исключение составляет класс физических состояний: так, сема ‘состояние: физ.’ выделяется у глаголов позы,

где субъектом является человек (или животное), а также у многочисленных глаголов, характеризующих актуальное свойство неодушевленного объекта, ср. *преть, прилипать, морщить*. Заметим, что аналогично обстоит дело и с физическими действиями, субъектом которых может быть не только человек, но и животное, природная сила и даже неодушевленный объект, ср. *кондиционер охлаждает воздух*, *упавшее дерево порвало провода*, *рюкзак режет плечи*.

Здесь уместно обратиться к проблеме разграничения сем ‘состояние’, ‘отношение’ и ‘свойство’, которые все так или иначе выражают идею статичности. Опираясь на энциклопедические источники [Новая философская энциклопедия; Философия …], мы принимаем следующие рабочие определения этих понятий. Состояние — это текущая характеристика какого-либо объекта в определенный интервал времени, ср.: *бедствовать* (‘испытывать нужду и лишения, жить в нищете’), *брезжить* (‘распространять слабый свет при наступлении утра’), *пламенеть* (‘быть охваченным каким-л. сильным чувством, пылать какой-л. страстью’), *циркулировать* (‘передаваться от одного к другому (об идеях, мыслях, слухах и т. п.)’), *отдыхать* (‘проводить где-л. свободное от работы время, быть в отпуске’), *обладать* (‘иметь что-л. своей собственностью, в своем распоряжении; владеть кем-л., чем-л.’).

Свойство отличается от состояния тем, что является более устойчивой, если не постоянной, характеристикой предмета, например: *равняться* (‘представляя собою какое-л. число, какую-л. величину, быть равным чему-л.’), *хромать* (‘иметь недостатки, быть неполноценным в каком-л. отношении’), *датироваться* (‘относиться к какому-л. времени, отмечаться какой-л. датой’), *носить* (‘иметь какое-л. имя, фамилию, звание’), *умещаться* (‘помещаться, входить куда-л. полностью или целиком’), *петь* (‘уметь петь, иметь пригодный для пения голос’), *бодаться* (‘обладать способностью или склонностью бодать’), *курчавиться* (‘виться мелкими кудрями (о волосах)’), *почковаться* (‘размножаться почкованием’), *вытекать* (‘брать начало (о реке, ручье)’). Из примеров видно, что свойство, как и состояние, может относиться к объектам разной природы, соответственно можно выделить такие семы, как ‘свойство: человека’, ‘свойство: животного’, ‘свойство: предмета’, хотя в некоторых случаях конкретизация невозможна (ср. *отличаться*, *называться* и др.).

Наконец, отношение, в отличие от двух предыдущих понятий, характеризует не отдельный предмет, а пару, тройку и т. д. предметов. Мы выделяем два вида отношений: логические (причинно-следственные, зависимости, сопряженности, принадлежности, соответствия и др.) и социальные, ср.: *базироваться* (‘основываться на чем-л.’), *совпадать* (‘соответствовать чему-л., согласовываться с чем-л.’), *превосходить* (‘быть, оказываться больше, сильнее кого-л., чего-л. по количеству, размерам, силе и т. п.’), *включать* (‘иметь что-л. в своем составе’); *враждовать* (‘быть в состоянии вражды с кем-л., чем-л.’), *общаться* (‘поддерживать связь, общение, взаимные отношения с кем-л., чем-л.’).

Таковы, на наш взгляд, базовые, категориальные семантические компоненты, характеризующие любое глагольное значение. На эту основу могут наслаиваться многочисленные семы различной природы. Прежде всего нужно упомянуть те из них, что связаны с фазовыми значениями. Заметим, что в литературе нет единой точки зрения на набор фазовых значений у русских глаголов (ср. [Богданов; Храковский]), и здесь мы также сформировали свой эмпирический список, отталкиваясь от материала. Вот он[2]: ‘предбытие’ (*близиться*, *грозить*, *веять*, *надвигаться*), ‘начало’ (*начинать(ся)*, *рождаться*, *всходить*, *наставать*, *проступать*, *воцаряться*, *стартовать*), ‘возобновление’ (*возобновлять(ся)*, *воскресать*, *возрождаться*, *возвращаться*), ‘становление’ (*делаться*, *созревать*, *складываться*, *вырастать*, *устанавливаться*, *уверяться*, *укрепляться*), ‘продолжение’ (*продолжать(ся)*, *выживать*, *оставаться*, *подхватывать*), ‘переход’ (*переходить*, *превращаться*, *преображаться*, *перевоплощаться*, *материализоваться*), ‘остановка’ (*останавливать(ся)*, *прекращать(ся)*, *переставать*, *замирать*, *стопорить*, *замораживать*, *вставать*, *перекуривать*), ‘конец’ (*завершать(ся)*, *заканчивать(ся)*, *увенчиваться*, *разрывать*, *дотягивать*, *умирать*, *высыхать*, *догорать*, *исчезать*, *умолкать*).

Помимо фазовых сем, в значении глагола могут присутствовать смысловые компоненты, характеризующие действие (движение, со-

[2] В скобках приводятся примеры глаголов, в значении которых (хотя бы в одном — в случае многозначных лексем) присутствует фазовый компонент. Чтобы не загромождать описание, мы опускаем словарные дефиниции, надеясь, что читатель сможет самостоятельно определить, какое значение полисемичного глагола имеется в виду.

стояние и пр.) с качественной или количественной стороны либо указывающие на его локализацию в пространстве или времени. В том, что касается качественных характеристик, мы отчасти следуем за Русским семантическим словарем [Шведова], авторы которого выделяют у так называемых бытийных глаголов семантические компоненты ‘внезапно, резко, стремительно’, ‘интенсивно, полно’, ‘вольно, плавно, торжественно’, ‘слабо, неясно, вяло’, ‘неустойчиво, беспорядочно’, ‘тайно, скрытно’ и др. Однако, как выясняется, подобного рода семы присутствуют у гораздо более широкого круга глаголов. Например, сема ‘резко, внезапно’ выявляется у таких глагольных значений, как *обрывать* (‘резко пресекать, прекращать (ход, течение чего-л.)’), *запахивать* (‘резким движением закрывать, захлопывать (дверь, окно)’), *шарахаться* (‘внезапно и резко менять свои убеждения’), *выпаливать* (‘говорить громко, быстро (обычно неожиданно, заранее не обдумав)’). Оказывается, что семы качественной характеристики действия играют гораздо более заметную роль в обеспечении системных связей глагольной лексики, чем можно было бы подумать, исходя из содержания Русского семантического словаря [Там же].

Мы также несколько расширили список качественных характеристик, добавив в него, в частности, такие семы, как ‘снова’ и ‘иначе, заново’, ср.: *перечитывать* (‘читать снова, еще раз’), *передумывать* (‘изменять свое решение, намерение’). Эти два разных значения не всегда четко дифференцируются (например, *перестилать* (‘стлать заново или иначе’) или *пересортировывать* (‘сортировать заново или иначе’)), что позволяет грамматистам объединять их в рамках повторительного способа действия. Поводом для их разграничения в нашем описании является то, что в ряде случаев соответствующая дифференциация налицо. Так, семантика некоторых глаголов допускает лишь один из указанных компонентов, например: *пересматривать* (‘рассматривая, изучая заново, переоценивать’), *перепоручать* (‘поручать выполнение чего-л. другому лицу’), *воссоздавать* (‘создавать заново; восстанавливать, воспроизводить’). Кроме того, есть глаголы, у которых данные семы являются по сути различительными признаками двух разных значений, ср.: *переизбирать* ‘избирать на какой-л. пост, на какую-л. должность кого-л. другого вместо ранее избранного’ vs. ‘избирать кого-л. вновь на какой-л. пост’.

Семы, связанные с количественными и темпоральными характеристиками действия, в известной мере перекликаются со способами глагольного действия. Так, сема ‘некоторое время’ соответствует ограничительному и длительно-ограничительному способам (*пережидать*, *просиживать*), сема ‘время от времени’ — прерывисто-смягчительному способу (*потрескивать*, *попыхивать*, *подумывать*), сема ‘все или многие’ (‘всё или многое’) — дистрибутивному (*перепиваться*, *перекидывать*). Однако круг глаголов, имеющих соответствующие семы, гораздо шире, чем можно предполагать, исходя из морфологических показателей того или иного способа действия. Другими словами, далеко не всегда наличие семы маркируется характерным формантом. Если обратиться к уже упомянутым способам действия, то, скажем, сема ‘некоторое время’ обнаруживается не только у глаголов с приставками *по-* и *про-*, но и у таких глаголов, как *отстаивать* (‘долго стоять в очереди’), *призадумываться* (‘задумываться на некоторое время, впадать в некоторое раздумье’), *выхаживать* (‘ходить какое-л. время или на какое-л. расстояние’). Таким образом, автоматизация здесь (как и вообще в семантике) мало что способна дать: только мысленная реконструкция обозначаемой глаголом ситуации вкупе с тщательным анализом словарной дефиниции способны выявить присутствие в лексическом значении тех или иных сем.

Тем более это верно в отношении оценочных компонентов значения. Их выделение осложняется комплексом проблем, связанных с известной размытостью понятия оценки, ее неоднозначными отношениями с эмоциональным и экспрессивным компонентами коннотации, а также предметно-понятийным ядром значения слова. В меньшей степени сказанное относится к количественной оценке — соответствующая сема обнаруживается у таких глаголов, как *превышать*, *зашкаливать*, *двоиться* (‘казаться двойным, представляться разделенным надвое’) и некоторых других.

Что касается качественной оценки, мы связываем ее с понятием субъективности как обозначения ситуации с позиции говорящего, своеобразного «присутствия» говорящего в ней (ср. [Traugott; Langacker]). Если встать на эту точку зрения, вопрос об эмоциональной нейтральности или окрашенности слова перестает быть релевантным. Выделяются положительная и отрицательная качественная оценка, ср. *затмевать* (‘отодвигать на второй план, превосходя

в чем-л.'), *блистать* ('отличаться особой выразительностью, красочностью'), *улыбаться* ('предстоять в недалеком будущем (о чем-л. приятном, благоприятном)'), *благоденствовать* ('успешно развиваться, благополучно существовать') vs. *страдать* ('иметь какой-л. недостаток, отличаться каким-л. отрицательным свойством'), *заживаться* ('жить на свете слишком долго'), *маячить* ('надоедливо возникать перед глазами'), *совать* ('стараться дать, продать что-л. ненужное или недоброкачественное, негодное и т. п.; навязывать'), *хапать* ('с жадностью брать что-л.'). Отрицательная оценка, в частности, характерна для глаголов, обозначающих поведение, поступки (ср. *влипать, жеманиться*, *проворонивать*, *кобениться*, *обезьянничать* и т. д.). В целом, как известно, отрицательная оценка представлена в языке гораздо шире, чем положительная, и наше исследование это полностью подтверждает.

Примечательно, что оппозиция положительной и отрицательной оценок в значении ряда глаголов нейтрализуется, например: *сходить* ('оказываться принимаемым за кого-л., что-л., равноценным кому-л., чему-л. в каком-л. отношении'), *считаться* ('расцениваться каким-л. образом, признаваться кем-л., чем-л. или каким-л.'), *красоваться* ('находиться на видном месте, привлекать к себе внимание, бросаться в глаза'), *выглядеть* ('иметь какой-л. вид, казаться, представляться кем-л., каким-л.'). Это характерно также для глаголов, обозначающих достижение определенного физического состояния, ср.: *докисать*, *докаливаться* и пр. Более того, можно наблюдать даже нейтрализацию оппозиции качественной и количественной оценок, ср. *кусаться* ('быть слишком дорогим, недоступным по цене').

Содержательно близкой к семе качественной оценки является сема 'наблюдатель' (ср. [Скребцова, с. 47–49]), также связанная с идеей субъективности, передачи ситуации «с позиции говорящего». Мы выделяем данную сему у перцептивных глаголов, прежде всего у всех значений, связанных с восприятием: обонянием, осязанием и вкусом, поскольку эти чувства, по мнению некоторых исследователей [Sweetser], имеют более выраженную субъективность, чем зрение и слух, ср. *ощущаться* ('чувствоваться, отмечаться (о запахе, вкусе)'), *горчить* ('иметь горьковатый вкус, привкус'), *благоухать* ('распространять благоухание'). Кроме того, эта сема отмечается у ряда глаголов зрительного (и в меньшей степени слухового) восприятия, значение которых связано с появлением или исчезновением некото-

рого перцепта (тем самым предполагается наличие поля восприятия некоторого субъекта — его-то мы и называем наблюдателем), например: *выныривать* ('вдруг становиться видимым, оказываться перед глазами'), *врываться* ('неожиданно проникать, заполнять собою что-л. (о ветре, звуках, запахах и т. п.)'), *ударять* ('появляться, возникать внезапно, с силой, начинать чувствоваться (о свете, запахе)'), *проступать* ('обнаруживаться, становиться заметным, явным; проявляться'), *открываться* ('появляясь из-за какой-л. преграды, оказываться в поле зрения, становиться видимым'), *скрадываться* ('становиться менее заметным, менее отчетливым, более тихим, приглушенным'). Здесь же следует упомянуть значения, выражающие идею того, что Р. Лангакер [Langacker, p. 88–90] называет субъективным, или ментальным, движением (в нашем описании ему соответствует сема 'движение: фиктивное'), ср.: *мелькать* ('показываться на короткое время и скрываться'), *бегать* ('о светящихся точках: быть, находиться, то появляясь, то исчезая'), *проплывать* ('плавно двигаться (о кажущемся движении неподвижных предметов, мимо которых что-л. движется)'), *удаляться* ('скрываться из вида, пропадать вдали (о неподвижных предметах по отношению к движущимся)'). Наконец, сема 'наблюдатель' может присутствовать в некоторых значениях, не обязательно связанных напрямую с сенсорным восприятием действительности, как, скажем, у глагола *сквозить* в значении 'обнаруживаться в небольшой степени'.

Есть целый ряд других сем, которые не только удовлетворяют нашему критерию регулярности (как было сказано выше, довольно слабому), но и встречаются у глаголов разных семантических групп и с этой точки зрения вносят свой вклад в системность лексики. Например, сема 'взаим.' присутствует в значении глаголов эмоционального состояния (*разделять*, *сопереживать*, *сочувствовать*), речевой деятельности (*беседовать*, *браниться*), социальных действий (*контактировать*, *водиться*), телодвижений (*обниматься*, *ласкаться*), информационных действий (*перестукиваться*, *аукаться*), профессиональной деятельности (*срабатываться*, *сыгрываться*), субъектного и объектного перемещения (*сходиться*, *разъезжаться*, *перебрасываться*), физических действий (*драться*, *брызгаться*), перцептивной деятельности (*переглядываться*). Сема 'возвр.', связанная с возвращением прежнего положения дел, фиксируется у столь различных по семантике глаголов, как *вернуть*,

водворять, *отрезвлять*, *очухиваться*, *реабилитировать(ся), отвоевывать, отыгрывать*. Широко представлена и идея направленности действия на себя (сема ‘себя’), ср. *ублажаться*, *защищаться*, *хвалиться*, *баллотироваться*, *прихорашиваться*, *бриться*, *пачкаться*, *ушибаться*, *глядеться*.

Узкие рамки настоящей статьи не дают возможности описать все выделенные нами регулярные семы в значениях русских глаголов и тем более — специфику их сочетаемости. Последнее представляет собой особенно интересную задачу, так как способно внести вклад в изучение отраженной в русском языке национальной картины мира.

Литература

Апресян Ю. Д. Лексическая семантика. Синонимические средства языка. М., 1974.

Богданов В. В. Фазисность и фазисные конструкции // Типология конструкций с предикатными актантами. Л., 1985.

Большой толковый словарь русского языка / сост. и гл. ред. С. А. Кузнецов. СПб., 1998.

Караулов Ю. Н. Частотный словарь семантических множителей русского языка. М., 1980.

Лексико-семантические группы русских глаголов: учеб. слов.-справ. / под общ. ред. Т. В. Матвеевой. Свердловск, 1988.

Лукин В. А. Семантические примитивы русского языка. Основы теории. М., 1990.

Новая философская энциклопедия: в 4 т. / под ред. В. С. Стёпина. М., 2000–2001.

Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / общ. ред. Н. Ю. Шведова. М., 2007. Т. 4.

Русский семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову / Ю. Н. Караулов [и др.]; отв. ред. С. Г. Бархударов. М., 1983.

Скребцова Т. Г. Семантика глаголов физического действия в русском языке: автореф. дис. … канд. филол. наук. СПб., 1996.

Философия: Энциклопедический словарь / под ред. А. А. Ивина. М., 2004.

Храковский В. С. Семантика фазовости и средства ее выражения // Теория функциональной грамматики. Введение. Аспектуальность. Временная локализованность. Таксис. Л., 1987.

Hallig, R., Wartburg W. von. Begriffsystem als Grundlage für die Lexikographie: Versuch eines Ordnungsshemas. Berlin, 1952.

Katz J. J., Fodor J. A. The Structure of a Semantic Theory // Language. 1963. Vol. 39, N 2. P. 170–210.

Langacker R. W. A view of linguistic semantics // Topics in Cognitive Linguistic / ed. by B. Rudzka-Ostyn. Amsterdam; Philadelphia, 1988. P. 49–90.

Sweetser E. From Etymology to Pragmatics. Cambridge, 1990.

Traugott E. C. From polysemy to internal semantic reconstruction // Proc. of the 12th Annual Meeting of the Berkeley Linguistics Society. Berkeley, 1986. P. 539–550.

СВЕДЕНИЯ ОБ АВТОРАХ

Азарова Ирина Владимировна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: ivazarova@gmail.com

Алексеева Елена Леонидовна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: el.alexeeva@gmail.com

Ачкасов Андрей Валентинович — доктор филологических наук, профессор. Кафедра английской филологии и перевода, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия.
E-mail: a_v_achkasov@mail.ru

Бабарико Максим Николаевич — переводчик. Санкт-Петербург, Россия. E-mail: s.chebanov@gmail.com

Букия Григорий Теймуразович — магистрант. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: gregorybookia@yandex.ru

Волков Сергей Святославович — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: sergejvolkov2006@yandex.ru

Герд Александр Сергеевич — доктор филологических наук, профессор. Заведующий кафедрой математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия.
E-mail: aheard@yandex.ru

Гребенников Александр Олегович — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет; кафедра иностранных языков, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия. E-mail: a.grebennikov@spbu.ru

Добров Алексей Владимирович — кандидат филологических наук, ассистент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: adobrov2010@gmail.com

Захаров Виктор Павлович — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: v.zakharov@spbu.ru

Зубкова Татьяна Ивановна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: tzoobkova@gmail.com

Косарева Екатерина Олеговна — аспирант. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: misskeo@mail.ru

Мартыненко Григорий Яковлевич — доктор филологических наук, профессор. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия.
E-mail: g.martynenko@gmail.com

Миронова Дина Марковна — старший преподаватель. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: 235dina@gmail.com

Митрофанова Ольга Александровна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия.
E-mail: oa-mitrofanova@yandex.ru

Николаев Илья Сергеевич — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: ilya.nikolaev@gmail.com

Паничева Полина Вадимовна — аспирант, кафедра математической лингвистики; инженер-исследователь, кафедра психологии, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия.
E-mail: ppolin86@gmail.com

Поддубных Мария Сергеевна — аспирант. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: maria2850@gmail.com

Протопопова Екатерина Владимировна — магистрант. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: protoev@yandex.ru

Рогозина Елена Андреевна — старший преподаватель. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: renehorn.r@gmail.com

Скребцова Татьяна Георгиевна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: t.skrebtcova@gmail.com

Тискин Даниил Борисович — аспирант. Институт философии Санкт-Петербургского государственного университета, Санкт-Петербург, Россия. E-mail: daniel.tiskin@gmail.com

Хохлова Мария Владимировна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: m.khokhlova@spbu.ru

Цумарев Алексей Эдуардович — кандидат филологических наук, старший научный сотрудник. Институт русского языка им. В. В. Виноградова РАН, Москва, Россия. E-mail: zumarew@yandex.ru

Чебанов Сергей Викторович — доктор филологических наук, профессор. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: s.chebanov@gmail.com

Шадрин Виктор Иванович — доктор филологических наук, профессор. Заведующий кафедрой английской филологии и перевода, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: shadrinvik@yandex.ru

Шелов Сергей Дмитриевич — доктор филологических наук, главный научный сотрудник, руководитель Терминологического центра. Институт русского языка им. В. В. Виноградова РАН, Москва, Россия. E-mail: volehs@mail.ru

INFORMATION ABOUT AUTHORS

Achkasov Andrei V. — Doctor of Philology, Professor at the Department of English Translation Studies, St. Petersburg State University, Russia.
E-mail: a_v_achkasov@mail.ru

Alexeeva Elena L. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: el.alexeeva@gmail.com

Azarova Irina V. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: ivazarova@gmail.com

Babariko Maxim N. — Freelancer. Saint Petersburg, Russia.
E-mail: s.chebanov@gmail.com

Bukia Grigorii T. — Master Student. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: gregorybookia@yandex.ru

Chebanov Sergey V. — Doctor of Philology, Professor at the Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: s.chebanov@gmail.com

Dobrov Alexey V. — Candidate of Philology, Assistant Lecturer. Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: adobrov2010@gmail.com

Gerd Alexander S. — Doctor of Philology, Professor at the Department of Mathematical Linguistics (Head of Department), St. Petersburg State University, Russia. E-mail: aheard@yandex.ru

Grebennikov Alexander O. — Candidate of Philology, Assistant Professor. Department of Mathematical Linguistics, St. Petersburg State University; Department of Foreign Languages, ITMO University, Russia. E-mail: a.grebennikov@spbu.ru

Khokhlova Maria V. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: m.khokhlova@spbu.ru

Kosareva Ekaterina O. — Postgraduate Student. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: misskeo@mail.ru

Martynenko Gregory Ya. — Doctor of Philology, Professor at the Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: g.martynenko@gmail.com

Mironova Dina M. — Assistant Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: 235dina@gmail.com

Mitrofanova Olga A. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: oa-mitrofanova@yandex.ru

Nikolaev Ilya S. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia.
E-mail: ilya.nikolaev@gmail.com

Panicheva Polina V. — Postgraduate Student at the Department of Mathematical Linguistics; Research Engineer at the Department of Psychology, St. Petersburg State University, Russia. E-mail: ppolin86@gmail.com

Poddubnykh Maria S. — Postgraduate Student. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: maria2850@gmail.com
Protopopova Ekaterina V. — Master Student. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: protoev@yandex.ru
Rogozina Elena A. — Assistant Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: renehorn.r@gmail.com
Shadrin Victor I. — Doctor of Philology, Professor at the Department of English Translation Studies (Head of Department), St. Petersburg State University, Russia. E-mail: shadrinvik@yandex.ru
Shelov Serguey D. — Doctor of Philology, Chief Researcher, Head of the Terminology Centre. V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia. E-mail: volehs@mail.ru
Skrebtsova Tatiana G. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: t.skrebtcova@gmail.com
Tiskin Daniel B. — Postgraduate Student. Institute of Philosophy, St. Petersburg State University, Russia. E-mail: daniel.tiskin@gmail.com
Tsumarev Aleksei E. — Candidate of Philology, Senior Researcher. V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia. E-mail: zumarew@yandex.ru
Volkov Sergey S. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: sergejvolkov2006@yandex.ru
Zakharov Victor P. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: v.zakharov@spbu.ru
Zoobkova Tatiana I. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, St. Petersburg State University, Russia. E-mail: tzoobkova@gmail.com

СОДЕРЖАНИЕ

CONTENTS

Научное издание

СТРУКТУРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

Межвузовский сборник

Выпуск 11

Редактор *Т. А. Темкина*
Корректор *Л. С. Козлова*
Компьютерная верстка *Е. М. Воронковой*

Подписано в печать 01.04.2016. Формат 60×84 $^{1}/_{16}$.
Усл. печ. л. 17,6. Тираж 100 экз. (1-й завод). Заказ № 67.
Издательство Санкт-Петербургского университета.
199004, С.-Петербург, В.О., 6-я линия, 11.
Тел./факс +7(812)328-44-22
E-mail: publishing@spbu.ru publishing.spbu.ru

Типография Издательства СПбГУ.
199061, С.-Петербург, Средний пр., 41.

Книги Издательства СПбГУ можно приобрести
в Доме университетской книги
Менделеевская линия, д. 5
тел.: +7(812) 329 24 71
часы работы 10.00–20.00 пн. — сб.,
а также в интернет-магазине OZON.ru