САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

СТРУКТУРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

Межвузовский сборник

Издается с 1978 года

Выпуск 13

Под редакцией И. С. Николаева

ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА

УДК 80+618.31
ББК 81.1
С83

Рекомендовано к публикации научной комиссией
в области наук о языках и литературе
Санкт-Петербургского государственного университета

Структурная и прикладная лингвистика: межвуз. сб.
С83 Вып. 13 / отв. ред. И. С. Николаев. — СПб.: Изд-во С.-Петерб. ун-та, 2019. — 178 с.

Сборник содержит статьи по широкому кругу проблем теоретической и прикладной лингвистики, по использованию математических и компьютерных методов в языкознании.

Предназначен для специалистов по теории языка, прикладной и компьютерной лингвистике.

УДК 80+618.31
ББК 81.1

ПРЕДИСЛОВИЕ

Очередной, тринадцатый выпуск сборника «Структурная и прикладная лингвистика» был подготовлен с участием обновленного и расширенного состава редакционной коллегии, в которую вошли многие известные ученые из российских и зарубежных научных организаций и университетов. В сборнике впервые появились статьи на английском языке зарубежных авторов. Процедура отбора статей стала более сложной и тщательной, статьи просматриваются редакторами и отдельно рецензируются другими специалистами по теоретической, структурной и прикладной лингвистике. Повысились стандарты издательской подготовки сборника.

При этом основные принципы сборника «Структурная и прикладная лингвистика» продолжают идеи, заложенные его основателем профессором Александром Сергеевичем Гердом (1936–2016): максимальная широта обсуждаемых лингвистических проблем, связанных с прикладной и математической лингвистикой, междисциплинарный подход, привлечение ученых из разных университетов России и других стран, приглашение молодых ученых и аспирантов.

Во время подготовки этого выпуска скончался наш учитель, коллега, друг и бессменный автор сборника профессор Григорий Яковлевич Мартыненко (1936–2019), работы которого всегда отличались глубиной идей, новаторством, высоким научным уровнем. Его статьи в предыдущих трех выпусках «Структурной и прикладной лингвистики» задавали тон всему сборнику. И на этот

раз его последняя работа «Междисциплинарные аспекты корпусометрии», публикуемая в настоящем выпуске, заставляет по-новому взглянуть на многие из идей, которые представлены далее в статьях других авторов.

Мы посвящаем этот сборник светлой памяти наших учителей и коллег профессоров А. С. Герда — основателя, автора и редактора сборника «Структурная и прикладная лингвистика» и Г. Я. Мартыненко — постоянного участника издания в течение многих лет, которые способствовали повышению научного уровня сборника и привлечению к нему новых читателей и авторов.

Г. Я. Мартыненко

МЕЖДИСЦИПЛИНАРНЫЕ АСПЕКТЫ КОРПУСОМЕТРИИ*

Аннотация. В статье обсуждается место корпусометрии в системе гуманитарных дисциплин, занятых измерениями на больших массивах текстов: стилеметрии, клиометрики, лексикометрии, наукометрии, библиометрии, социометрии, информметрии, медиаметрии, гедонометрии и др. Определенная связь, особенно в методическом плане, существует и между корпусометрией и описательными дисциплинами естественно-научного толка: технетикой, биометрией, науками о земле. Рассматривается также взаимодействие корпусометрии с теорией сообществ, теорией совокупности, теорией систем. В филологии следует различать измерения на литературоведческих и лингвистических корпусах в связи с различием задач двух ветвей словесности. Однако у них есть «общая территория», на которой решаются литературоведческие задачи лингвистическими методами. В основе построения таких корпусов лежат системные идеи выдающегося русского ученого, писателя и литературоведа Ю. Н. Тынянова.

Ключевые слова. Корпусная лингвистика, корпусометрия, стилеметрия, большие данные, междисциплинарный подход, лингвистика, литературоведение, литературное наследие.

Gregory Ya. Martynenko

INTERDISCIPLINARY ASPECTS OF CORPORAMETRICS

Abstract. The article discusses the place of corporametrics in the system of other humanitarian disciplines engaged in measurements made on large text collections: stylometrics, cliometrics, lexicometrics, scientometrics, bibliometrics, sociometry, informetrics, mediametrics, hedonometrics, etc. A certain connection, especially in methodological terms, exists between corporametrics and the particular descriptive natural sciences —

* Исследование выполнено при поддержке гранта РФФИ № 17-29-09173 «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

technetics, biometrics, and earth sciences. The interaction of corporametrics with the theory of communities and the theory of systems is also considered. In philology, it is necessary to distinguish different approaches to text measurement in linguistics and literary science. However, these approaches do have a 'common territory', where literary problems are being solved by means of linguistic methods. The construction of such literary corpora is based on the systemic ideas first proposed by the famous Russian scholar, literary critic and writer Yury N. Tynyanov.

Keywords. Corpus linguistics, corporametrics, stylometrics, big data, interdisciplinary approach, linguistics, literary studies, literary heritage.

1. Измерительные дисциплины в гуманитарных науках

С XIX века началось бурное вторжение математических методов в гуманитарные науки. Родился длинный перечень измеряющих дисциплин: антропометрия (Адольф Кетле, Альфонс Бертильон), психометрика (Адольф Цейзинг, Густав Фехнер), стилеметрия (Вильгельм Диттенбергер), биометрия (Фрэнсис Гальтон, Карл Пирсон), эконометрия (Вильфредо Парето) [Мартыненко, 2014]. Первопроходцем этого процесса был выдающийся бельгийский ученый А. Кетле (1796–1874). Он явился основателем математической статистики, сделав измерения человеческих масс центром статистического мировоззрения, основанного на теории вероятностей. В более специальном смысле бельгийца можно считать родоначальником антропометрии, стержнем которой он считал синтетический образ среднего человека. Средний человек, по Кетле, — это обобщенный индивидуум среднего роста, веса, силы, средней емкости легких, средней полноты или худобы, средней остроты зрения, слуха, интеллектуальных способностей и моральных качеств.

Более того, эту идею Кетле сделал столь универсальной, что включил в нее эстетическую составляющую бытия человека. Так, ему принадлежит утверждение, согласно которому обычный эстетический тип — это средний человек, в котором находятся в равновесии антропологические, социально-психологические, моральные, языковые и эстетические черты человека конкретной эпохи. Все это позволяет считать бельгийского ученого предвестником искусствометрии, поскольку для Кетле средний человек — предел статистического обобщения, идеальный образец конкретной эпохи, в некотором смысле — эстетический идеал. А создание образа типичного героя является одной из основных задач художественной литературы.

В филологической науке благодаря усилиям немецкого филолога В. Диттенбергера (1840–1896) [Dittenberger, 1881] возникла стилеметрия. Задача этой дисциплины была откровенно текстологической, она состояла в решении проблемы авторства и датировки фрагментов диалогов Платона [Мартыненко, 1988] с помощью лингвостатистических методов. Можно также упомянуть и лексикометрию, первый кирпич в здание которой был заложен В. Н. Куницким (1857–1916) — составителем частотного словаря комедии Грибоедова «Горе от ума» [Куницкий, 1894]. Этот словарь явился первым документом такого рода и вполне отвечает требованиям современной статистической лексикографии. Несколькими годами позднее был опубликован частотный словарь немецкого языка [Käding, 1897–1898]. К этой «числовой компании» можно присоединить и фонометрию [Förstemann, 1852].

Процесс внедрения в обществоведение математических идей в XX веке приобрел лавинообразную форму. К перечисленным измерительным дисциплинам добавились социометрия, наукометрия, библиометрия, клиометрика, искусствометрия, информметрия, медиаметрия и многие другие. Среди новейших дисциплин измерительного толка можно назвать также экзотическую гедонометрию [Reagan et al., 2016], изучающую эмоциональную динамику нарратива на основе методик больших данных (big data) путем измерения эмоционального уровня частей текста, следующих друг за другом.

Преобразилась и стилеметрия. Из науки, занимающейся исключительно атрибуцией, она постепенно превратилась в дисциплину с более широким охватом решаемых задач. Содержание стилеметрии было очерчено так: «Стилеметрия — прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик с целью упорядочивания и систематизации (атрибуции, датировки, диагностики, типологии и т. п.) текстов и их частей» [Мартыненко, 1988, с. 54–55].

Обратим внимание на то, что стилеметрия — прикладная лингвистическая дисциплина, решающая литературоведческие задачи методами математической лингвистики. Но, войдя в пространство литературоведения, она в значительной мере превращается в теоретическую дисциплину, сталкиваясь с фундаментальными проблемами словесности: проблемой жанра и жанровой дифференциации текстов, проблемой эволюции литературно-художественных систем,

атрибуцией текстов, типологией сюжетов и др. Но в любом случае обращение к текстам художественной литературы в языкознании является элементом повседневной работы словесника.

2. Измерительные дисциплины и теория сообществ

Практически все измеряющие дисциплины родились явно или неявно в контексте теории сообществ (ценозов), основоположниками которой были немецкие ученые Густав Рюмелин (1815–1889) и Карл Август Мёбиус (1825–1908).

Первому принадлежит идея социальной группы (социальной массы). Такие группы рассматривались Рюмелином как собирательные понятия, т. е. как целостные единичности, элементы которых не тождественны друг другу. Этим они отличаются от разделительных понятий, которым соответствуют однородные классы единиц. Логика таких совокупностей, по Рюмелину, не зависит от их качественной природы. Такой совокупностью может быть и биологическое сообщество, и население какого-нибудь города, и совокупность слов какого-нибудь текста. Это означает, что уже на этапе возникновения теории сообществ в ней содержалась предметная **междисциплинарность**.

Рюмелин отмечал: «…в области естественных наук… господствуют родовые понятия и постоянные признаки конкретных случаев… Про род нельзя сказать ничего, что не относилось бы вместе с тем и к каждому его члену; родовое понятие есть понятие о типичной особи или конкретном случае… В собирательном понятии, напротив того, соединены в группу, на основании какого-нибудь общего признака, предметы весьма разнообразные. Интерес сосредоточивается на том, что можно сказать о группе как целом, а не о признаках отдельных членов» [Rümelin, 1875], цит. по: [Дружинин, 1979, с. 49]. Это была первая четкая привязка статистики к собирательным понятиям.

Несколько позднее (1877) основатель экологии К. А. Мёбиус выдвинул идею биоценоза. Биоценоз, по Мёбиусу, — это совокупность (сообщество) организмов, совместно населяющих участок суши или водоема. Впоследствии этот термин получил распространение главным образом в немецком и русском языках. В англоязычных странах используется в том же смысле термин «сообщество» (community, population) или «экосистема» (ecosystem).

Через некоторое время эту идею более детально разработал русский статистик А. А. Чупров. Разделяя идеи Рюмелина, Чупров выдвинул идею статистической совокупности, являющейся сообществом единиц, не обязательно однородных и представляющих собой групповое (собирательное) понятие. Чупров ввел также понятие реальной совокупности, под которым он понимал целостное образование, локализованное в конкретных рамках времени и пространства [Чупров, 1909].

С точки зрения теории статистики каждый текст может рассматриваться как реальная совокупность. Это не текст вообще, это всегда конкретный текст — текст, созданный конкретным автором, в конкретное время, в конкретной ситуации. Основным признаком таких реальных совокупностей А. А. Чупров считал их устойчивость во времени: способность в течение более или менее длительного периода сохранять свой состав и характерные черты [Чупров, 1909]. С этой точки зрения текст сверхустойчив: ни одно слово, ни одна фраза из текста после того, как он подписан к печати, удалена быть не может: что написано пером, того не вырубишь топором. «Вторжение» в текст или какие-либо манипуляции с ним допустимы лишь в процессе специально организованной языковой игры или перцептивного эксперимента. Примером такой «забавы», впрочем весьма эффективной в учебном процессе, является игра «Толстой или компьютер» [Орехов, 2015].

В качестве реальной совокупности могут выступать не только тексты, но и текстовые корпусы. Важнейшим фактором, позволяющим считать корпус реальной совокупностью, является фактор целостности. Он формируется принадлежностью корпуса к определенному языку, жанру, стилю, автору или группе авторов в определенную историческую эпоху.

Таким корпусом можно считать, например, множество рассказов А. П. Чехова вместе с реализованными в них фонетическими, лексическими или синтаксическими единицами. Целостность здесь обеспечивается единством жанрового стиля, устойчивостью индивидуальной манеры письма, принадлежностью автора к определенной школе, литературному течению и т. п. Локализация корпуса в определенных рамках времени и пространства в данном случае определяется местом писателя в эволюции русской литературы как определенной литературно-художественной суперсистемы.

О корпусе как совокупности можно говорить и тогда, когда исследуется собрание произведений какой-либо национальной литературы (русской, немецкой, чешской и др.), относящихся в рамках одного жанра к определенной литературной эпохе, например к началу XX века. «Дух эпохи» в самом широком смысле этого слова цементирует целостность, собирательность такого собрания произведений, позволяет, несмотря на их разнородность, видеть в них единую систему.

Стилистическое единство всей литературы данной эпохи осознается не только филологами и литературными критиками. Его остро ощущают и сами художники слова, даже те, кто не без оснований может претендовать на собственную стилистическую исключительность. Так, П. Б. Шелли в предисловии к одной из своих поэм пишет: «…между всеми писателями какой-либо данной эпохи должно быть известное сходство, не зависящее от их собственной воли. Они не могут уклониться от подчинения общему влиянию, проистекающему от бесконечного сочетания обстоятельств, относящихся к эпохе, в которую они живут, хотя каждый из них до известной степени является создателем того самого влияния, которым проникнуто все его существо… Это именно то влияние, от которого не властен ускользнуть ни самый ничтожный писака, ни самый возвышенный гений…» [Шелли, 1904, с. 51].

Следует, однако, иметь в виду, что при переходе от конкретного текста к группе текстов данного автора, и далее к группе текстов в пределах данного жанра, а затем — к многожанровым корпусам вплоть до корпуса данного национального языка в конкретный исторический период происходит постепенное ослабление фактора целостности и однородности совокупности. Расширяя поле наблюдения, мы превращаем совокупность текстов и реализованных в них единиц в конгломерат, теряющий свойство целостности.

3. О системном анализе языка и стиля художественной литературы

В своей книге «Архаисты и новаторы» Ю. Н. Тынянов говорит о синхронических и диахронических литературно-художественных системах. Под синхроническими системами он понимает совокупность произведений данной литературной эпохи, а под диахрониче-

скими — последовательность сменяющих друг друга синхронических систем. При этом он сетует на то, что усилия большей части литературоведов устремлены на изучение произведений выдающихся писателей, тогда как периферия литературы и даже ее «центр» остаются за бортом исследовательского интереса [Тынянов, 1929]. Это означает, что для Тынянова крайне важным был вопрос максимальной представленности авторов в той или иной системе литературы и представительности корпуса текстов этих авторов. Любой текст Тынянов рассматривал как литературный факт, который должен приниматься во внимание независимо от масштабов дарования автора и его роли в литературном процессе.

Системный подход Тынянова рано или поздно будет реализован. Однако для этого необходимы огромные усилия в формировании максимально полных электронных ресурсов художественной литературы, включающих произведения не только крупных, но и второстепенных, периферийных писателей. Ведь оценки специалистов переменчивы: ярлык крупности (великости, известности) очень часто навешивается не только за литературные заслуги. Зачастую крупные писатели заносятся в список второстепенных, а некоторые писатели вообще предаются анафеме или забвению, хотя ни для кого не является секретом, что новые литературно-стилистические веяния часто рождаются именно на периферии литературы, в так называемом «литературном быту» [Тынянов, 1929]. Однако в литературную эпоху, следующую за данной, эти приемы нередко перемещаются на авансцену, будучи освоенными «крупными» писателями.

Проблема системного анализа литературы тесно переплетается с проблемой возрождения и сохранения литературного наследия, существенная часть которого до недавнего времени была вычеркнута из памяти народа. В 90-е годы прошлого века (в большей степени) и в начале XXI века (в меньшей степени) сделано очень много для возвращения народу художественных произведений ушедших эпох. Были переизданы произведения многих авторов, сыгравших выдающуюся роль в национальном литературном движении (в частности, произведения поэтов-символистов, труды выдающихся русских философов), были обнародованы многочисленные энциклопедии, антологии, словари русских писателей и поэтов, написаны теоретические труды, посвященные творчеству выдающихся писателей начала XX века, произведения которых в советской России не переиздавались вообще

или переиздавались в мизерном объеме. Речь идет о произведениях Леонида Андреева, Евгения Чирикова, Зинаиды Гиппиус, Бориса Зайцева, Михаила Кузмина, Федора Сологуба, Владимира Ропшина и многих других авторов. Однако бросается в глаза то, что переиздавались исключительно произведения писателей с именем, писателей заведомо значительных. При этом, однако, за бортом книгоиздательского и филологического внимания остался легион практически забытых, но в свое время весьма популярных властителей читательских дум, хотя очевидно, что нужно прежде всего обращать внимание на наследие тех писателей, которые были значительны именно в контексте своей эпохи, а не с позиции переменчивого взгляда представителей последующих эпох.

И тем не менее следует признать, что в последние годы работа по обеспечению доступа к литературному наследию была проведена достаточно масштабная. При этом учитывались, с одной стороны, интересы массового читателя, а с другой — специфические информационные потребности филологов-профессионалов, литературных критиков, искусствоведов — специалистов, для которых текст и корпус текстов является объектом рефлексии: лингвистической, литературоведческой, перцептивно-эстетической, герменевтической, культурологической и др.

Профессиональная текстовая рефлексия предполагает обращение к информационным ресурсам, несоизмеримым с теми, с помощью которых удовлетворяются интересы массового читателя. По существу, для сообщества исследователей современной формации необходим доступ ко всей литературной продукции, созданной в ту или иную историко-литературную эпоху.

В связи с проектом Тынянова остановимся еще на одном впечатляющем проекте. Речь идет о проекте Андрея Белого, направленном на массовое построение словарей русских писателей [Белый, 1934]. Андрея Белого можно понять. Ведь он всегда тяготел к исследованию больших текстовых коллекций. Но при тех технических средствах, которые существовали в начале XX века, такой проект можно было осуществить только на метро-ритмическом уровне. Ритмические фигуры отличаются большой повторяемостью, а это не требует обращения к большим коллекциям текстов. Иную картину мы наблюдаем на лексическом уровне, где повторяемость элементов чудовищно неравномерна. Поэтому проект Белого, как и проект Тынянова, остался

только декларацией и стал частично осуществляться только в самое последнее время.

В одном из таких проектов речь идет о крупномасштабном исследовании русской художественная прозы конца XIX — начала XX века, т. е. конкретной синхронической системы тыняновского типа, но с одним существенным ограничением — это русская художественная проза, представленная рассказом (или новеллой), причем этот жанр интересовал разработчиков исключительно с синтаксической точки зрения [Мартыненко, 1988]. Почему только рассказ? Причины здесь три. Первая — его чрезвычайная распространенность и популярность в беллетристической среде. Это обеспечивает включение в орбиту исследования максимального числа авторов. Вторая причина состоит в том, что рассказ выполняет функцию «разведчика» — в нем, по сравнению с более крупными прозаическими жанрами (романом, повестью), с опережением рождаются новые стилистические приемы и отмирают старые, т. е. рассказ — это жанр «быстрого реагирования» на стремительно меняющуюся ситуацию в литературном процессе. Определенное значение имеет и то обстоятельство, что изучение структуры рассказа занимает центральное место в нарративистике, а также при обсуждении проблемы стилистической краткости (стиля «короткой строки»).

Еще один крупномасштабный проект, технологически более продвинутый, осуществлен в Великобритании путем корпусного исследования классической английской литературы XIX века: Диккенса, Уайльда, Бронте и др. — CLiC Dickens project[1]. Корпус представляет собой информационную систему, выполняющую кроме информационно-поисковых и классификационных также и исследовательские задачи. В системе, например, предусмотрено автоматическое членение текста на речь автора, речь персонажей и авторские ремарки. Важной функцией системы является определение сходства между текстами разных авторов, а также их отличия от нейтрального усредненного жанрового фона.

В заключение обратимся к проблеме формирования корпуса, которую в корпусометрии можно считать центральной.

[1] https://www.nottingham.ac.uk/research/groups/cral/projects/clic.aspx (дата обращения: 20.02.2018).

Важнейшей проблемой корпусной лингвистики является **представительность выборки**. В понимании этого феномена перекрещиваются лингвистические, литературоведческие и теоретико-статистические представления, которые не всегда согласуются друг с другом.

Для литературоведа обычны персоналистский и антологический подходы (см. табл. 1), корпус для него — это прежде всего собрание текстов, принадлежащих наиболее типичным, образцовым авторам данной эпохи, чаще всего выдающимся; при этом синхроническая или даже ахроническая «великость», «значимость», «авторитетность» писателя часто подвергается ревизии в пестрой динамике постоянно меняющейся социально-политической ситуации. С наибольшей откровенностью антологический подход реализуется в учебном процессе, в котором школьники и студенты знакомятся с лучшими образцами национальной и мировой литературы.

Для лингвиста характерен суммативный подход (см. табл. 2) — стремление включить в корпус максимальное число текстов с целью предельного вычерпывания ресурсов языка; для лингвиста представительность — это в первую очередь размер корпуса. При этом лингвист явно или неявно тяготеет к созданию гиперкорпусов, отражающих лингвистические ресурсы конкретного национального языка, т. е. лингвист вольно или невольно стремится к большим данным (big data). Не секрет, что, будучи воспитанными на классических образцах художественной литературы, лингвисты при формировании гиперкорпусов обычно делают крен в пользу именно таких текстов. Это представляется естественным, так как «[л]итература является ярким примером использования языка; никакой систематический подход не может претендовать на описание языка, если он не охватывает также литературу; при этом она должна рассматриваться не как некое причудливое образование, но как естественное составляющее в системе языка» [Sinclair, 2004, p. 51].

Однако следует признать, что в лингвистике, как и в литературоведении, прочные позиции занимает, как мы отмечали выше, персоналистский подход. Известны также жанровые корпусы [Мартыненко и др., 2000]. Отметим также, что при создании многомиллионных гиперкорпусов крайне трудно выдержать стерильность научного подхода. При отборе текстов здесь всегда будут сочетаться принцип практической целесообразности и принцип случайности, о чем говорит,

Таблица 1. **Подходы к формированию корпуса в литературоведении**

	Библиографический	**Персоналистский**	**Антологический**	**Учебно-консервативный**
Подходы	• Авторский • Жанровый • Хронологический		• Хронологический • Жанровый • Тематический • Концептуальный • Сериальный	
Тексты	– Библиографии – Лексиконы писателей – Литературные энциклопедии и т. п.	– Собрания сочинений: полные неполные – Избранное – Изборники и т. п.	– Антологии – Сборники – Литературные серии и т. п.	– Хрестоматии – Адаптированные тексты – Дайджесты и т. п.

Таблица 2. **Подходы к формированию корпусов в языкознании (лингвистике)**

Подходы	**Глобалистский**	**Персоналистский**	**Иллюстративно-дидактический**	**Информационно-технологический**	**Конструктивно-синтезирующий**
Корпусы	• Национальные корпусы • Жанровые корпусы	• Корпусы выдающихся авторов	• Корпусы учебных текстов • Корпусы параллельных текстов	• Корпусы как речевой материал для создания и тестирования информационных систем	• Лингвистически представительные корпусы

в частности, опыт разработки таких корпусов за рубежом [Фрэнсис, 1983]. Об этом свидетельствует и практика создания больших корпусов и частотных словарей русского языка: «Частотный словарь русского языка» под редакцией Засориной [Частотный словарь…, 1977], «Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)» [Ляшевская, Шаров, 2009], Корпус повседневной устной речи «Один речевой день» [Bogdanova-Beglarian et al., 2016] и др. При определении репрезентативного объема словаря стремление к его сбалансированности (тематической, авторской, жанровой) в идеале должно сочетаться с требованием состоятельности, т. е. сходимости объема словаря и других статистик к предельным величинам (уровню насыщения) [Мартыненко, 1988], что удается далеко не всегда. Для филологии это является большой проблемой, и перспективы ее решения пока туманны.

В статистике традиционно различаются большие и малые выборки. Большими считаются выборки, включающие обычно более сотни единиц, малыми — объемом не более 30 единиц. Каждая из двух типов выборки обрабатывается с помощью своей техники. В частности, для малых выборок вводится поправка на дисперсию, а в качестве теоретического закона, на котором основываются заключения по малой выборке, выступает не нормальный закон, а распределение Стьюдента.

4. Big data и корпусная лингвистика

В классической математической статистике, а вслед за ней и в отраслевых статистиках сложилась устойчивая традиция организации выборочного наблюдения и измерения ошибок выборки. Но все эти методы относятся к малой и большой выборкам.

В последнее время усиленно разрабатываются способы работы с очень большими массивами данных, так называемыми big data. Широкое распространение больших данных связано прежде всего с их экспансией в сети Интернет. Большие данные через YouTube, Facebook, «ВКонтакте» и другие социальные сети и интернет-сайты вошли в жизнь почти каждого человека, населяющего нашу планету. Число пользователей этих сетей достигает миллиардов, а число сообщений — сотен миллиардов. Тотальное воцарение больших данных подкрепляется также массовой оцифровкой печатной продукции, ко-

торая десятилетиями и столетиями ждала своего часа. В настоящее время «спящие» информационные потоки начали новую жизнь и стали объектом интереса для многочисленных исследователей. Начинает сбываться мечта многих поколений ученых-гуманитариев, получивших доступ к огромным массивам информации. Это привело к возникновению цифровой гуманитаристики, которая семимильными шагами развивается в последние годы. Особенно широкое проникновение больших данных характерно для многочисленных ответвлений филологической науки, для которой важной задачей стало построение национальных корпусов и сверхбольших совокупностей текстов с их последующей статистической обработкой.

Однако статус таких сверхбольших массивов в теории выборки не определен, не выяснено также отношение таких массивов к сплошной выборке. Работы в этом направлении обычно осуществляются стихийно, без учета многолетней статистической практики. Эта область современной науки представляет большой интерес для корпусной лингвистики и корпусометрии. При этом надо иметь в виду, что большие данные — это не только объем (volume), но и скорость (velocity) работы с данными, обладающими большим разнообразием (variety) [Канаракус, 2011]. По мере развития теории и практики больших данных кроме перечисленных «трех V» на авансцену вышли и другие (вплоть до семи V): сначала достоверность — veracity, потом изменчивость — variability, ценность — value и визуализация — visualisation. Появление все более новых V связано со сложностью данных и сложностью и многоаспектностью работы с ними [McNulty, 2014]. Часть этих V согласуется со статистическими категориями (достоверность, изменчивость), часть — с общенаучными (разнообразие, ценность), другая имеет сугубо технический или организационный характер, который в традиционной лингвистике и классической статистике во внимание не принимался (скорость, визуализация и др.), но в теории больших данных играет решающую роль.

5. Выводы

В статье рассмотрено место корпусометрии в системе измеряющих дисциплин, ее методическое единство с теми дисциплинами, которые имеют дело с текстом и собранием текстов: стилеметрией, кли-

ометрикой, наукометрией, социометрией, гедонометрией. Выявлено различие в отношении к корпусам лингвистов и литературоведов. Обсуждена тесная связь корпусометрии с теорией сообществ и теорией систем.

Ближайшей перспективой корпусостроения в словесности является разработка системы переменных, позволяющих осуществлять многоаспектное статистическое описание корпусов в синхронии и диахронии.

Источники

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

Частотный словарь русского языка / под ред. Л. Н. Засориной. М.: Наука, 1977.

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016: International Conference on Speech and Computer / ed. by A. Ronzhin, R. Potapova, G. Németh. Heidelberg: Springer, 2016. P. 659–666. (LNCS (LNAI). Vol. 9811).

Förstemann E. Numerische lautverhältnisse im griechischen, lateinischen und deutschen // Germanische Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen / hrsg. von Th. Aufrecht, A. Kuhn. Bd. 1. Göttingen: Ferd. Dümmler's Verlagsbuchhadlung, 1852. S. 159–163.

Käding F. W. Häufigkeitswörterbuch der deutsche Sprachen: Festgestellt durch einen Arbeitsausschuss der deutschen Stenographiesysteme. Steglitz bei Berlin: Selbstverlag des Herausgebers; E. S. Mittler & Sohn, 1898.

Литература

Белый А. Мастерство Гоголя. Л.: ОГИЗ, 1934.

Дружинин Н. К. Развитие основных идей статистической науки. М.: Статистика, 1979.

Канаракус К. Машина Больших Данных // Сети = Network World. 2011. No. 04. https://www.osp.ru/nets/2011/04/13010802/ (дата обращения: 01.11.2018).

Куницкий В. Н. Язык и слог комедии Грибоедова «Горе от ума». (С приложением словаря комедии). Киев: [б. и.], 1894.

Мартыненко Г. Я. Основы стилеметрии. Л.: Изд-во ЛГУ, 1988.

Мартыненко Г. Я. Стилеметрия: возникновение и становление в контексте междисциплинарного взаимодействия. Ч. 1: Первые шаги: XIX век // Структурная и прикладная лингвистика: межвуз. сб. Вып. 10 / под ред. А. С. Герда. СПб.: Изд-во СПбГУ, 2014. С. 3–23.

Мартыненко Г. Я., Гринбаум О. Н., Гребенников А. О. Автоматическая антология русского рассказа как речевой материал для лексикометрических исследований // Материалы XXIX межвуз. науч-метод. конф. преподавателей и аспирантов. Вып. 11: Секция лексикологии. СПб.: Филол. ф-т СПбГУ, 2000. С. 20–21.

Орехов Б. В. История литературы как автопортрет // Третье литературоведение: учеб. записи филол.-методол. семинара (2008–2009) / науч. ред., сост. Б. В. Орехов, С. С. Шаулов, Е. В. Лукьянов. Биробиджан: Приамур. гос. ун-т им. Шолом-Алейхема, 2015. С. 167–174.

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

Фрэнсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике. Вып. 14: Проблемы и методы лексикографии. М.: Мир, 1983. С. 301–334.

Чупров А. А. Очерки по теории статистики. СПб.: Тип. М. М. Стасюлевича, 1909.

Шелли П. Б. Полн. собр. соч.: в 3 т. / пер., [предисл.] К. Д. Бальмонта. Новое изд., перераб. Т. 2. СПб.: Знание, 1904.

Dittenberger W. Sprachliche Kriterien für die Chronologie der Platonischen Dialoge // Hermes. 1881. Vol. 16. No. 3. S. 321–345.

McNulty E. Understanding Big Data: The Seven V's // Dataconomy. Дата публикации: 22.05.2014. http://dataconomy.com/2014/05/seven-vs-big-data/ (дата обращения: 01.11.2018).

Reagan A. J., Mitchell L., Kiley D., Danforth C. M., Dodds P. S. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes // Arxiv.org. Дата публикации: 27.09.2016. https://arxiv.org/pdf/1606.07772.pdf/ (дата обращения: 01.11.2018).

Rümelin G. Zum Theorie der Statistik, Reden und Aufsätze. Osnabrück: Kramer & Haugen Gmb., 1875.

Sinclair J. Trust the Text: Language, Corpus and Discourse. London: Routledge, 2004.

Sources

Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. 2016. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. *SPECOM 2016. International Conference on Speech and Computer*, A. Ronzhin, R. Potapova, G. Németh (eds). Heidelberg, Springer, pp. 659–666. (LNCS (LNAI), vol. 9811).

Förstemann E. 1852. Numerische lautverhältnisse im griechischen, lateinischen und deutschen. *Germanische Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen*, Th. Aufrecht, A. Kuhn (Hrsg.), Bd. 1. Göttingen, Ferd. Dümmler's Verlagsbuchhadlung, SS. 159–163.

Frequency Dictionary of the Russian Language 1977, N. A. Zasorina (ed.). Moscow, Nauka Publ. (In Russ.)

Käding F. W. 1898. *Häufigkeitswörterbuch der deutsche Sprachen: Festgestellt durch einen Arbeitsausschuss der deutschen Stenographiesysteme*. Steglitz bei Berlin, Selbstverlag des Herausgebers; E. S. Mittler & Sohn, 1898.

Liashevskaia O. N., Sharov S. A. 2009. *Russian Language Frequency Dictionary (Based on the Materials of the National Corpus of the Russian Language)*. Moscow, Azbukovnik Publ. (In Russ.)

References

Bely A. 1934. *The Mastery of Gogol'*. Leningrad, OGIZ Publ. (In Russ.)

Chuprov A. A. 1909. *Essays on the theory of statistics*. Saint Petersburg, Tip. M. M. Stasiulevicha Publ. (In Russ.)

Dittenberger W. 1881. Sprachliche Kriterien für die Chronologie der Platonischen Dialoge. *Hermes*, vol. 16, no. 3, pp. 321–345.

Druzhinin N. K. 1979. *The Development of the Basic Ideas of Statistical Science*. Moscow, Statistika Publ. (In Russ.)

Frances W. N. 1983. The Problems of Formation and Machine Representation of Large Text Corpora. *Novoe v zarubezhnoi lingvistike*, issue 14. Problemy i metody leksikografii. Rus. Ed. Moscow, Mir Publ., pp. 301–334. (In Russ.)

Kanarakus Ch. 2011. Big Data Machine. *Networks = Networks World*, no. 4. Rus. Ed. https://www.osp.ru/nets/2011/04/13010802/ (accessed date: 11/01/2018). (In Russ.)

Kunitskii V. N. 1894. Language and Style in Griboyedov's Comedy "Woe from Wit". (With the Appendix of the Dictionary of Comedy). Kiev, [s. p.]. (In Russ.)

Martynenko G. Ya. 1988. *Foundations of Stylometrics*. Leningrad, Izd-vo LGU Publ. (In Russ.)

Martynenko G. Ya. 2014. Stylometry. Emergency and Evolution in Context of Interdisciplinary Interaction, part 1. The 19th Century: the Beginning. *Strukturnaia i prikladnaia lingvistika. Mezhvuz. sb.*, issue 10, A. S. Gerd (ed.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 3–23. (In Russ.)

Martynenko G. Ya., Grinbaum O. N., Grebennikov A. O. 2000. Automatic Anthology of Russian Short Stories as a Material for Lexicometric Studies. *Materialy XXIX mezhvuz. nauch.-metod. konf. prepodavatelei i aspirantov*, issue 11. Section of lexicology. Saint Petersburg, Filol. f-t SPbGU Publ., pp. 20–21. (In Russ.)

McNulty E. 2014. Understanding Big Data: The Seven V's. *Dataconomy*, publication date: 01.11.2018. http://dataconomy.com/2014/05/seven-vs-big-data/ (accessed date: 01.11.2018).

Orekhov B. V. 2015. History of Literature as a Self-portrait. *Tret'e pokolenie literaturovedov: ucheb. zapisi filol.-metodol. seminara (2008–2009)*, B. V. Orekhov, S. S. Shaulov, E. V. Luk'ianov (sc. eds., comp.). Birobidzhan, Priamur. gos. un-t im. Sholom-Aleikhema Publ., pp. 167–174. (In Russ.)

Reagan A. J., Mitchell L., Kiley D., Danforth C. M., Dodds P. S. 2016. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *Arxiv.org*, publication date: 27.09.2016. https://arxiv.org/pdf/1606.07772.pdf/ (accessed date: 01.11.2018).

Rümelin G. 1875. *Zum Theorie der Statistik, Reden und Aufsätze*. Osnabrück, Kramer & Haugen Gmb.

Shelley P. B. 1904. *Complete Works*, in 3 vols., K. D. Bal'mont (transl., [preface]), new ed., rev., vol. 2. Rus. Ed. Saint Petersburg, Znanie Publ. (In Russ.)

Sinclair J. 2004. *Trust the Text: Language, Corpus and Discourse*. London, Routledge.

Tynyanov Yu. N. 1929. *Archaists and Innovators*. Moscow, Priboi Publ. (In Russ.)

С. В. Чебанов

СУДЬБА МАТЕМАТИЧЕСКОЙ ЛИНГВИСТИКИ В ЭПОХУ ВТОРОЙ КОГНИТИВНОЙ РЕВОЛЮЦИИ

Аннотация. Зародившись в начале XX века, математическая лингвистика (МЛ) обрела статус самостоятельной дисциплины в 1950–1960-е годы. Возникнув в результате взаимодействия полярных типов знания, МЛ была представлена небольшим (десятки — первые сотни) числом исследователей, что определяло ее статус как малой науки со специфической методологией (очерченной С. В. Мейеном), отличной от методологии больших наук (рассмотренных Т. Куном). В эпоху первой когнитивной революции востребованность МЛ определялась запросами интенсивно развивавшихся теории информации, теории автоматов, кибернетики, появлением компьютерной техники и возникшими задачами программирования, управления космической и ядерной техникой, потребностями разработки автоматизированных и автоматических технологий и пр. Реализация таких проектов предполагала устранение из сферы функционирования техники человека, замененного алгоритмами его активности. Математические модели языка мыслились как класс таких алгоритмов. Осознание нереализуемости лингвистического автомата, действующего независимо от человека, в рамках второй когнитивной революции (в понимании Р. Харре) приводит к осознанию фундаментальности человеко-машинного комплекса как структурной единицы всех компьютерно-зависимых технологий. При этом проблематика МЛ (формулирование и доказательство теорем, в частности теорем существования, аксиоматизация, исследование предельных случаев и пр.) теряет свою актуальность, замещаясь проблемами вычислительной (компьютерной) лингвистики, с ее практически значимыми задачами эффективности, экономичности, скорости обработки текста. Вытесняемая на периферию МЛ вынуждена при этом самоопределяться в качестве принципиально малой науки, растворяться в прикладной лингвистике или искать свое место среди таких предметных областей, как единая когнитивная наука или НБИКС-технологии.

Ключевые слова. Математическая лингвистика, первая когнитивная революция, вторая когнитивная революция, малая наука, когнитивная наука, НБИКС-технологии, прикладная лингвистика.

Sergey V. Chebanov

THE FATE OF MATHEMATICAL LINGUISTICS IN THE ERA OF THE SECOND COGNITIVE REVOLUTION

Abstract. Born in the early 20th century, mathematical linguistics (ML) acquired the status of independent discipline in the 1950s–1960s. Originating in the interaction of the polar types of knowledge, ML has been shown a small by the number of researchers (tens — first hundred) that determined its status as a small science with a specific methodology (outlined S. V. Meyen) that differs from the methodology of big science (reviewed by T. Kuhn). In the era of the first cognitive revolution, the demand for ML was determined by the demands of intensively developing information theory, the theory of automata, cybernetics, the emergence of computer technology and the emerging problems of programming, control of space and nuclear technology, the needs of the development of automated and automatic technologies, etc. The implementation of such projects involved the elimination of the human out from the sphere of technology, replaced by algorithms of its activity. Mathematical models of the language were thought of as a class of such algorithms. The realization of the unrealizability of a linguistic automaton acting independently of a human being within the framework of the second cognitive revolution (in the understanding by R. Harre) leads to the realization of the fundamental nature of the human-machine complex as a structural unit of all computer-dependent technologies. At the same time, ML problems (formulation and proof of theorems, in particular, existence theorems, axiomatization, study of limiting cases, etc.) lose their relevance, being replaced by problems of computational (computer) linguistics, with its practically significant tasks of efficiency, productive and economic efficience, speed of text processing. The ML displaced to the periphery is forced to self-identify as a fundamentally small science, to dissolve in applied linguistics or to seek its place among the unified cognitive science or NBICS technologies.

Keywords. Mathematical linguistics, first cognitive revolution, second cognitive revolution, small science, cognitive science, NBICS-convergence, applied linguistics.

1. Математическая лингвистика как область знания

Рассматривая сформулированную тему, с тем чтобы не подменить предмет обсуждения обсуждением того, что такое математическая лингвистика, можно воспользоваться трактовкой математической лингвистики, данной А. В. Гладким: «…математическая дисциплина, предметом к<ото>рой является разработка и изучение понятий, образующих основу формального аппарата для описания строения естественных языков (т. е. метаязыка лингвистики). …В М. л. широко используются методы теории алгоритмов, теории автоматов и алгебры. …М. л. постоянно эволюционирует по пути превращения в теоретическую математич<ескую> дисциплину, являющуюся по сути дела одним из ответвлений математич<еской> логики. В то же время круг приложений М. л. расширился — ее методы нашли применение

в теории программирования» [Гладкий, 1982, с. 565–566], дополнив очерченную область лингвостатистикой.

Зародившись в начале — первой половине XX века (работы А. А. Маркова [Марков, 1913], А. Н. Колмогорова (см. о них: [Успенский, 1955; 1957]), С. Г. Чебанова[1] [Чебанов, 1947; Чебанов, Кузнецова, 2012], Л. Теньера 1934 и 1938 годов [Теньер, 1988][2] и др.), математическая лингвистика обрела статус самостоятельной дисциплины в 1950-е годы [Гладкий, 1982].

Сам факт возникновения математической лингвистики представлялся чем-то странным или даже невозможным, поскольку в ней соединяются два типа знания, которые рассматриваются ныне как полярные. В сознании широкого круга людей это связано с противопоставлением математического и гуманитарного знания, что нашло отражение в развернувшейся в СССР в 1960–1970-е годы дискуссии физиков и лириков. В более рафинированной форме это отражено в представлениях В. Виндельбанда о номотетических и идиографических науках [Виндельбанд, 1995] и Г. Риккерта о номотетических и идиографических компонентах каждой науки [Риккерт, 1911], что коррелирует с различением наук о природе и наук о духе В. Дильтея [Дильтей, 2000].

Вообще говоря, эта ситуация если не удивительна, то, по крайней мере, крайне примечательна в историко-культурном отношении. Дело в том, что сближение математики и языка присутствовало в пифагорейских союзах, представлено в семантике исключительно важного для философии и науки греческого слова λόγος, осуществлено в семи свободных искусствах (например, в трактовке Исидора Севильского) как арифметика и геометрия квадривиума с грамматикой и риторикой тривиума [Адо, 2002].

Так или иначе, но в силу сложившейся в последние столетия полярности математики и лингвистики как типов знания круг лиц, занимающихся математической лингвистикой, всегда был невелик (десятки — первые сотни человек), что определило статус математической лингвистики как малой науки.

---

[1] В настоящее время обнаружены документы, свидетельствующие об обучении С. Г. Чебанова в 1920-е годы у Ю. А. Круткова, впервые в России читавшего курсы по квантовым статистикам, и Л. В. Щербы [Чебанов и др., 2019].

[2] Здесь и далее даются в основном ссылки на легко доступные ныне издания, а не на первые публикации.

2. Математическая лингвистика как малая наука

На факт существования малых наук как чего-то радикально отличающегося от наук «больших» (таких как физика, биология, химия и т. п.), которые рассматривались Т. Куном как материал для формулирования представлений о научной парадигме и научных революциях [Кун, 2003], обратил внимание С. В. Мейен [Мейен, 1985]. Малые науки характеризуются небольшим числом соединяемых ими специалистов (от нескольких десятков до нескольких сотен в мире), персональным знакомством всех или почти всех специалистов друг с другом, наличием между ними профессионально значимых личных отношений, значимостью каждой персоны для интеллектуального состояния всего сообщества, зависимостью интеллектуального настроя сообщества в целом и каждой персоны в частности от интеллектуального, психического и физического состояния лидеров, наличием не только плохо осознаваемых парадигмальных противостояний и сближений, но и вполне артикулируемых различий и сходств позиций отдельных исследователей, окрашенных личными симпатиями и антипатиями, сменой доминирующих точек зрения, определяемой не сменой поколений, а особенностями жизненного пути лидеров данной науки, что может как сокращать (причем резко), так и значительно продлевать жизнь той или иной идеи.

Математическая лингвистика сложилась из единичных разрозненных исследований в малую науку в результате того, что получило после публикаций Р. Харре о второй когнитивной революции [Харре, 1996] название первой когнитивной революции.

3. Первая когнитивная революция

Первая когнитивная революция имела место с середины 1940-х по начало 1960-х годов и включала в себя такие события, как формулирование Н. Винером в 1948 году основных идей кибернетики [Винер, 1968], основание К. Шенноном в том же году теории информации [Shannon, 1948; Шеннон, 1963], создание Дж. фон Нейманом теории автоматов [Нейман, 1971], появление первых компьютеров (ЭВМ), включающих в себя лингвистический процессор, и создание языков программирования для написания управляющих ими программ. Совокупность этих событий сделало очевидной идею о том,

что развитие средств автоматического управления сложными техническими системами (транспортные средства и сети, включая космические, в том числе обеспечивающие доставку ядерных боеприпасов, атомные электростанции и автономные энергетические установки, автоматизированные производства и т. д.) немыслимо без математического обеспечения их функционирования. Последнее включает формальные языки представления данных и языки программирования. При этом человек представлялся как сложная ЭВМ (компьютер) или их комплекс. Так получило распространение сопоставление левого полушария с цифровой, а правого — с аналоговой ЭВМ [Иванов, 1978].

4. Математическая лингвистика и первая когнитивная революция

В описанной ситуации математическая лингвистика и получаемые ею результаты оказались востребованными, в результате чего стало возможным ее конституирование как самостоятельной области знания. При этом в ней были получены совершенно разные по характеру результаты, сложились разные направления исследований, итоги которых стали в разной мере известны за пределами математической лингвистики.

Так, одним из самых известных достижений математической лингвистики является формирование стандарта составления частотных словарей, ставших общеупотребительными как в разных областях лингвистики, так и за ее пределами. Столь же употребительным стало использование графов для представления синтаксических отношений членов предложения в соответствии с представлениями Л. Теньера [Теньер, 1988]. Указанные разработки стали основой многих других моделей языка и компьютерных программ обработки текста (в том числе реализуемых ныне с использованием совершенно других вычислительных ресурсов).

Многие разработки раннего периода математической лингвистики отличались остроумием, простотой и изяществом — алгоритмы дешифровки Б. В. Сухотина [Сухотин, 1976] и Ю. В. Кнорозова [Кнорозов, 1963], выделения морфем по З. Харрису [Harris, 1955] или А. Жюйяну [Juilland, 1961], метод фильтров И. Лесерфа [Лесерф, 1963] и Л. Н. Иорданской [Иорданская, 1967]. При этом такие раз-

работки были направлены и на сокращение вычислений. Последнее обстоятельство, с одной стороны, полностью соответствует духу математики, которую можно понимать как средство сокращения вычислений, а с другой стороны, оно оказалось важным, когда стало возможным осуществлять обработку текста и расчеты с использованием компьютера (ЭВМ), поскольку можно было оптимизировать использование ЭВМ определенной конструкции для реализации алгоритмов, предполагающих преимущественное использование процессора, постоянной или оперативной памяти. С развитием вычислительной техники указанные аспекты рассматриваемых разработок утратили свою актуальность, снизив тем самым требования к изяществу создаваемых моделей.

Другой тип результатов этого периода имел чисто идейное значение, тем более что их можно было получить и численными методами. Прежде всего речь идет о доказательстве специфически матлингвистических теорем типа теорем С. Я. Фитиалова об эквивалентности грамматик зависимостей и грамматик непосредственных составляющих [Фитиалов, 1968].

К этому же ряду работ относится введение Л. Заде понятия лингвистической переменной и основанного на нем представления о размытых (нечетких) множествах [Заде, 1976]. В сочетании с представлением об условных вероятностях этот подход привел к созданию вероятностной модели понимания смысла В. В. Налимова [Налимов, 1974]. На время введения в оборот эти подходы почти не имели никакого расчетного значения (из-за отсутствия как пригодных для этого эмпирических данных, так и программных средств) и были важны как математические идеи, позволяющие по-новому взглянуть на природу языка. Эти и подобные им случаи дали В. В. Налимову основание говорить о метафорической роли математики для развития разных дисциплин (лингвистики, биологии, философии, психологии и т. д. [Налимов, 1981]). Однако позже, с середины 1980-х, такие модели стали использоваться в моделях когнитивистов (Дж. Лакофф, У. Лабов, их сотрудники и ученики).

Иная судьба у моделей языка типа порождающих грамматик Н. Хомского [Хомский, 1962], аппликативной порождающей модели С. К. Шаумяна [Шаумян, Соболева, 1963], модели И. А. Мельчука «Смысл ⇔ Текст» [Мельчук, 1974]. Начинаясь с логически прозрачных утверждений о языке, со временем они превратились в разветвленные

модели, включающие многие сотни правил. Это позволило, с одной стороны, довольно детально представлять свойства языка (языков), а с другой — развивать их за счет обогащения и усложнения правил, что несколько заслоняло логику моделей, но делало органичной их реализацию с помощью современной компьютерной техники в совершенно новых исторических условиях.

Так или иначе, в эпоху первой когнитивной революции востребованность математической лингвистики и возможность ее развития определялись указанными выше запросами интенсивно развивавшихся теории информации, теории автоматов, кибернетики, появлением компьютерной техники и возникшими задачами программирования, управления космической и ядерной техникой, потребностями разработки автоматизированных и автоматических технологий и пр. Реализация таких проектов предполагала устранение из сферы функционирования техники человека, замененного алгоритмами его активности. Математические модели языка мыслились как класс таких алгоритмов.

Развитие математической лингвистики, с одной стороны, предполагало привлечение вычислительных средств для обеспечения решения ее расчетных задач (прежде всего принимая во внимание лингвостатистику), а с другой стороны, способствовало совершенствованию этих вычислительных средств, в первую очередь влияя на совершенствование языков программирования и концепции автоматизированной обработки данных.

В результате в качестве смежной с математической лингвистикой развивалась вычислительная, или компьютерная, лингвистика, размеры которой (выражаемые в количестве вовлеченных специалистов, количестве публикаций, запрашиваемых и получаемых объемах финансирования, числе профильных учреждений и их подразделений и т. д.) по сравнению с размерами математической лингвистики взрывообразно разрастались [Прикладная…, 2017; Dörnye, 2007]. При этом в компьютерной лингвистике сложилось представление о лингвистическом автомате — программном комплексе, способном к автоматической обработке текста на естественном языке (опознание языка, индексирование текста, его реферирование, информационный поиск, перевод, распознавание и синтез устной речи и т. д. [Беляева, 2001]).

Важнейшим концептуальным событием этого периода была дискуссия на тему «Может ли машина мыслить?», которая привела к рас-

колу как профессионального сообщества, так и кругов специалистов в смежных областях и широкой общественности, включая философов и политиков [Тьюринг, 1960].

Позиции участников дискуссии разделились по следующим несколько различным, но идейно-практически рассматриваемым как эквивалентные оппозициям:

Машина может мыслить.	—	Машина не может мыслить.
Машина заменяет человека.	—	Машина дополняет человека.
Машина работает автономно.	—	Машина работает во взаимодействии с человеком.
Язык полностью описывается математикой.	—	Язык частично описывается математикой.
Универсальный лингвистический автомат возможен.	—	Универсальный лингвистический автомат невозможен.
Универсальный лингвистический автомат строится как реализация математических функций.	—	Программы обработки текста строятся как реализации математических функций, практических ограничений и эвристических максим.

Подавляющее число специалистов, работавших в области искусственного интеллекта, как и многие специалисты смежных областей, а также заметная часть образованных людей, не связанных профессионально с этой областью, полагали, что возможно создание машины, способной полностью заменить интеллект человека. Напротив, значительная часть образованной публики, не связанной с этими разработками (в особенности художники, музыканты, гуманитарии разных специальностей), а также часть специалистов, работавших в этой области, напротив, полагали, что создание такой машины принципиально невозможно. К их числу относился и Ю. А. Шрейдер [Шрейдер, 1975].

Примерно четверть века эти две позиции существовали как сопоставимо обоснованные, пока ситуации не стала меняться в связи с теми событиями в области когнитивных разработок, которые в итоге и позволили Р. Харре говорить о второй когнитивной революции.

5. Вторая когнитивная революция

Суть этих событий заключается в том, что при решении как исследовательских, так и инженерных задач был снят вопрос о стремлении к получению и формулированию универсальных результатов. Вместо этого ставится цель охватить критически значимое число подлежащих изучению представителей, случаев, вариантов ситуаций и/или предложить решения задачи для практически приемлемой доли возникающих запросов. Примеры таких разработок представлены в публикациях весьма различных типов и интеллектуальных стилей — Б. Берлина и П. Кея [Berlin, Kay, 1969], Дж. Лакоффа [Лакофф, 2004], У. Лабова [Лабов, 1983], Э. Рош [Mervis, Rosch, 1981], Р. Шенка [Шенк, 1980], Т. Винограда [Виноград, 1976], У. Л. Чейфа [Чейф, 1982], М. Минского [Минский, 1979], Ч. Филлмора [Филлмор, 1988] и др.

Содержанием поворота от первой ко второй когнитивной революции является то, что если идеология первой когнитивной революции заключалась в том, что человек (точнее, его мозг) уподоблялся машине (комплексу машин), работа которой (которых) подчиняется строгим математическим алгоритмам, то идеология второй когнитивной революции состоит в том, чтобы создать компьютер, подобный человеку как существу не только чисто рациональному [Харре, 1996]. Последнее проявляется в нескольких аспектах.

1. Знание понимается не просто как совокупность некоторых данных, фактов, концепций, теорий и т. д., но как форма **представления** данных, причем представление данных оказывается не менее ценным, чем сами данные [Шрейдер, 1986].
2. Язык во всех его проявлениях (включая все неголосовые компоненты речи — мимику, жесты, кашель, чихание, икание и т. д.) признается основным и самым совершенным средством представления данных.
3. Компьютер уподобляется человеку — как морфологически (антропоморфные роботы), так и поведенчески (интерфейс на естественном языке [Розалиев, Заболеева-Зотова, 2010]; яндексовский голосовой помощник «Алиса»).
4. Представление данных осуществляется не на жесткой логической основе, а на основе языковой категоризации (с ее идеализированными когнитивными моделями — ИКМ [Лакофф, 2004]).

5. Идеалом представления ситуации является не его универсальность и всеохватность, а эффективность и наглядность. Поэтому описание, кратко и наглядно представляющее ядро представляемой предметной области, ценнее универсального, но длинного, громоздкого [Мусхелишвили, Шрейдер, 1997], многокомпонентного описания той же области. В силу этого вполне удовлетворяющим является такое описание, которое охватывает критически важную долю материала, допускающую автоматическую обработку, и дает оценку частоты ситуаций, которые не охватываются этим описанием и требуют «ручной» обработки (как это имеет место в когнитивистике, например в когнитивной лингвистике).
6. Лингвистический автомат мыслится не как заместитель человека, а как его партнер, так что центральной проблемой оказывается создание оптимально организованных человеко-машинных комплексов, с гостеприимным для пользователя антропоморфным интерфейсом. Последнее обстоятельство выдвигает на первый план проблему представления данных, в качестве наилучшего средства которого выступает естественный язык.

Таким образом, по сути дела была принята вторая, менее популярная среди профессионалов периода первой когнитивной революции, точка зрения о том, что создание автономно действующего лингвистического автомата невозможно.

При этом вторая когнитивная революция разворачивается на фоне смены возможностей компьютерных технологий (а отчасти и включает их) — увеличения производительности компьютеров, невиданного облегчения доступа к ним, перехода от тезаурусного к фреймовому представлению данных, появления Интернета и развития интернет-технологий (включая технологии рекреации) и т. д.

6. Математическая лингвистика и вторая когнитивная революция

Процессы, характерные для второй когнитивной революции, радикально меняют статус математической лингвистики, место ее среди других дисциплин, а отчасти и ее содержание.

Прежде всего изменилась сама математика. Распространение и совершенствование компьютерной техники привели к ничем не

сдерживаемому развитию численных методов, открылись новые возможности решения задач на перебор и комбинаторику, появились машинные методы вывода и работы с графами. Такое положение дел изменило отношение к доказательствам теорем, ценности доказывания теорем как вида математической деятельности.

С другой стороны, появилась возможность работы с большими объемами данных, сложными стандартизованными пакетами прикладных программ, автоматизированным базами лингвистических данных, полнотекстовыми базами данных, языковыми корпусами и т. д. [Захаров, Богданова, 2013]. Все это открыло новые возможности для решения прикладных лингвистических задач (автоматическое реферирование и индексирование, автоматический и автоматизированный перевод, автоматическая и автоматизированная разметка, автоматический поиск, автоматический сбор лингвостатистических данных и т. д.).

Это сильно сместило интересы сложившегося сообщества математических лингвистов в область компьютерной лингвистики с ее вычислительными возможностями. Ценность умения находить изящные математические решения, позволяющие сокращать объем необходимой для работы компьютерной памяти, оказалась утраченной.

При этом далеко не каждый компьютерный лингвист способен восстановить алгоритмы, положенные в основу стандартных пакетов прикладных программ, а тем более реконструировать допущения и ограничения, принятые при создании этих алгоритмов. В результате собственно математическая (не вычислительная) составляющая компьютерной лингвистики оказывается не проясненной, что свидетельствует об утрате культуры именно математической работы, в то время как эта культура критически важна при понимании математической лингвистики как отрасли математики (см. выше). В итоге происходит возврат математической лингвистики к статусу малой науки.

Однако и в статусе малой науки положение математической лингвистики неустойчиво. Это связано с тем, что ныне, по крайней мере в Российской Федерации, нет предпосылок для воспроизводства кадров матлингвистов. Дело в том, что пятилетний учебный план позволил составить приемлемую модель сочетания математической и лингвистической подготовки. Дополнить это еще и подготовкой в области компьютерных дисциплин за тот же учебный период (не

говоря о сроках бакалавриата или магистратуры) невозможно. Поэтому необходимо либо создание 5,5–6-летних курсов подготовки специалистов-матлингвистов (как это было сделано с физиками в 1950-е годы), либо осознанное признание того, что матлингвисты будут формироваться «штучно», проходя уникальные индивидуальные образовательные маршруты, т. е. их образование не будет институализировано.

7. Математическая лингвистика в составе разных областей деятельности

Более реалистичной является возможность исчезновения математической лингвистики как самостоятельно институализированной области деятельности (что не исключает наличие некоторого числа матлингвистов как единичных изолированных автономных профессиональных агентов) и либо интегрирование ее в одну из сфер — в компьютерную лингвистику, единую когнитивную науку или в НБИКС-технологии, либо какое-то перераспределение ее проблематики по двум или трем из указанных областей.

Растворение математической лингвистики в компьютерной кажется вполне возможным. Однако, как представляется, это будет взаимно вредно для обеих областей, хотя величины этого вреда не будут очень большими. Вред для компьютерной лингвистики заключается в том, что у компьютерных лингвистов будет поддерживается иллюзия того, что свою деятельность, которая по сути является **техне**, они могут хотя бы частично основывать на результатах, полученных в математической лингвистике, которая по сути является **гнозисом**. При этом как у матлингвистов будут возникать желания (обычно иллюзорные, но иногда и вполне обоснованные) предлагать свои результаты в качестве основы для работ компьютерных лингвистов, так и у компьютерных лингвистов будут формироваться свои представления о том, чем должны заниматься матлингвисты, что может быть основой выдвижения компьютерными лингвистами тех или иных требований к типу и качеству продуктов деятельности матлингвистов, что, в свою очередь, может категорически исказить характер и смысл деятельности последних. Сохранение математической лингвистики как самостоятельной области деятельности возможно в такой ситуации в том случае, если математическая лингвистика найдет для себя

способ существования, рассматриваемый В. А. Лефевром [Лефевр, 1973] в качестве системы, нарисованной на системе — компьютерной лингвистике.

При этом существование математической лингвистики «в теле» компьютерной лингвистики может быть весьма продуктивным, так как использование автоматических и автоматизированных методов обработки языкового материала может как на многие порядки увеличить объем этого материала, выступающего в качестве базы индукции при обнаружении новых закономерностей, так и проводить принципиально новые типы его математической обработки, обрабатывать один и тот же материал многими разными методами, испытывать большое число новых методов и принципов обработки, решать новые классы задач и т. д. Все это может обогатить математическую лингвистику определенным классом новых результатов и идей, как это имеет место, например, в корпусной лингвистике, основанной на использовании технологий работы с big data, позволяющими оперировать с полнотекстовыми базами данных[3]. При этом получаемые таким образом осязаемые конкретные результаты оттесняют на периферию, скажем, проблемы подчинения предельным теоремам лингвостатистических данных.

Однако при этом возникает и новый класс трудностей, подрывающих основания математической лингвистики как самостоятельной дисциплины.

Во-первых, подавляющее число лиц, уже сейчас занимающихся компьютерной лингвистикой, не в состоянии сформулировать: 1) какие содержательные лингвистические допущения заложены в алгоритмы обработки данных и 2) какие статистические допущения приняты (например, принимается ли справедливость центральной предельной теоремы, каковы изменения дисперсии при росте объема обрабатываемой выборки и т. д.). При этом многие специалисты не знают, что такое значащие цифры, какое их число является достоверным при приближенных вычислениях, которыми являются все стати-

[3] Указанное обстоятельство заслуживает особого внимания. Дело в том, что в 1960–1980-е годы был популярен тезис И. Е. Тамма о том, что будущий (XXI) век будет веком биологии, как XX век был веком физики, который пока не оправдывается — XXI век оказывается веком компьютерных наук. Правда, ныне в науке ожидается третий прорыв — на стыке биологии и информатики — в биоинформатике.

ческие расчеты при фиксированном числе значащих цифр. Ситуацию усугубляет использование программы Excel, стандартные настройки которой выдают результаты с нарушением правил округлений. Не осознана проблема зависимости статистической устойчивости (робастности) решения от точности расчетов, учета разного количества значащих цифр и т. д.

Во-вторых, далеко не все работающие специалисты знают, какие именно алгоритмы положены в основу создания стандартных пакетов прикладных программ и каковы допущения, обеспечивающие корректность интерпретации результатов, получаемых с помощью этих программ. При этом практически не принимается во внимание возможность получения незначительно, но систематически смещенных оценок (даже при высокой дисперсии [Немцева-Плахотя, Чебанов, 2010]), что может быть практически важным при решении задач лингвостилистики и атрибуции текстов.

В-третьих, возникают колоссальные психологические проблемы при необходимости проведения лингвостатистических исследований, предполагающих ручные ввод и разметку текста (исследования по лексической семантике, изучение текстов, написанных редкими и нестандартными шрифтами, рукописных текстов и т. д.). В результате подобные исследования откладываются на неопределенное время, а соответствующие проблемы оттесняются на периферию матлингвистических исследований.

В-четвертых, как кажется на основании наблюдений за существующими тенденциями развития обсуждаемых областей, даже в рамках компьютерной лингвистики не происходит генерализации категорий языка, так чтобы она в равной степени и в полной мере охватывала исторически сложившиеся естественные языки, искусственные языки, языки программирования, геномику — изучение генов, представляющих собой определенным образом структурированные последовательности оснований ДНК [Тищенко, 2004], и протеомику, занимающуюся изучением белков как аминокислотных последовательностей, синтезированных на матрицах этих ДНК [Принципы…, 2015]. Последние две области развиваются ныне в составе биоинформатики [Pevsner, 2015]; их логика, основы расчетов, принципы создания компьютерных программ и т. д. весьма сходны как в своих сильных, так и слабых (например, недоступность пользователям алгоритмов, которые положены в основу этих программ) сторонах. Поэтому,

например, выпускница кафедры математической лингвистики СПбГУ [Жернакова, 2011] смогла сразу же перейти к занятиям биоинформатикой.

Возможен и вариант «встраивания» математической лингвистики в когнитивистику — единую когнитивную науку, объединяющую теорию познания, когнитивную психологию, нейрофизиологию, когнитивную лингвистику, невербальную коммуникацию и теорию искусственного интеллекта [Меркулов, 2009; Шрейдер, 1986]. При этом, конечно, произойдет перераспределение тем, актуальных для разработки, причем как за счет того, что какие-то математические (в том числе расчетные) задачи станут актуальными и для них так или иначе будут найдены стандартные решения, так и за счет того, что другие, наоборот, станут чем-то экзотическим (например, задачи доказательства теорем, в особенности теорем существования или эквивалентности), чем будут заниматься одиночки, верные традициям основателей математической лингвистики. При таком положении дел, однако, вряд ли будет возможным говорить о какой-либо структуре математической лингвистики, которая распадется на множество очень мелких областей, складывающихся вокруг относительно небольших групп сходных задач. В таком случае (как об этом можно судить по опыту развития других дисциплин) будет сильно выражена тенденция подгонки решения новых задач под решения старых.

Наконец, если принять во внимание возможность НБИКС-конвергенции (нано-, био-, инфо-, когни-, соционаук и технологий [Алексеева и др., 2013; Ковальчук, 2011; Ковальчук и др., 2013]), то безотносительно к реалистичности этой конвергенции (в том числе в трактовке авторов концепции [Bainbridge, Roco, 2005]) допустимо рассматривать возможность растворения математической лингвистики как самостоятельной дисциплины в процессах такой конвергенции. Принципиально судьба математической лингвистики при этом будет такой же, как и при встраивании в когнитивистику, которая, в свою очередь, включается в процессы НБИКС-конвергенции. При этом трудно сказать, каковы будут реальные результаты такой конвергенции (как познавательной, так и технологической). Однако известные автору профильные институции и образовательные программы (которые по понятным причинам указываться не будут) представляются несколько поверхностными. Тем не менее складывается впечатление, что за пределами рассмотрения вопросов математической лингви-

стики в контексте стандартизованной обработки больших массивов данных (что органически связано с утратой лингвистической специфики) интересная для традиционных матлингвистов проблематика может присутствовать в таком случае только в качестве тех или иных казуистических случаев, не складываясь в целостную тематическую область.

8. Некоторые итоги

Приведенные соображения, как представляется, убедительно свидетельствуют о том, что математическая лингвистика в том виде, как она складывалась в первой половине XX столетия и как она существовала во второй его половине, далее существовать не будет. При этом основной тенденцией изменений математической лингвистики представляется некоторое ее огрубление и утилитаризация как в математическом (в первую очередь), так и в лингвистическом отношении. При этом открываются необозримые возможности для развития компьютерной лингвистики (в отношении как вводимых в оборот новых типов и массивов лингвистических данных, так и способов их обработки) в разных переплетениях последней с различными областями науки и технологий. Однако культура работы, сложившаяся в математической лингвистике, видимо, будет необратимо утрачена с точки зрения как математической строгости, так и лингвистической тонкости интерпретации получаемых результатов. Тем не менее открывающиеся возможности позволяют надеяться на получение более широких обобщений, справедливых для более разнообразного языкового материала. Осознание же этих обобщений и их содержательная интерпретация, видимо, станет достоянием исследователей другого склада.

При этом проблематика математической лингвистики (формулирование и доказательство теорем, в частности теорем существования, аксиоматизация, исследование предельных случаев и пр.) теряет свою актуальность, замещаясь проблемами вычислительной (компьютерной) лингвистики, с ее практически значимыми задачами эффективности, экономичности, скорости обработки текста. Наряду с этим осознаётся фундаментальность человеко-машинного комплекса как структурной единицы всех компьютерно-зависимых технологий.

В заключение автор считает своим приятным долгом принести благодарность С. А. Гашкову, М. А. Гашковой, В. Л. Каганскому, Л. Ю. Ковригиной, Г. Я. Мартыненко, Т. Г. Петрову и А. С. Щекину за их вклад в работу над текстом.

Литература

Адо И. Свободные искусства и философия в античной мысли. М.: Греколатинский кабинет Ю. А. Шичалина, 2002.

Алексеева И. Ю., Аршинов В. И., Чеклецов В. В. «Технолюди» против «постлюдей»: НБИКС-революция и будущее человека // Вопросы философии. 2013. № 3. С. 12–21.

Беляева Л. Н. Лингвистические автоматы в современных информационных технологиях. СПб.: РГПУ им. А. И. Герцена, 2001.

Виндельбанд В. Избранное: Дух и история. М.: Юрист, 1995.

Винер Н. Кибернетика, или Управление и связь в животном и машине. М.: Советское радио, 1968.

Виноград Т. Программа, понимающая естественный язык. М.: Мир, 1976.

Гладкий А. В. Математическая лингвистика // Математическая энциклопедия: в 5 т. / гл. ред. И. М. Виноградов. Т. 3. М.: Советская энциклопедия, 1982. С. 565–568.

Дильтей В. Собр. соч.: в 6 т. Т. 1: Введение в науки о духе: Опыт полагания основ для изучения общества и истории. М.: Дом интеллектуальной книги, 2000.

Жернакова Д. В. Разработка системы классификации коротких текстов с использованием онтологий: магист. дис. / СПбГУ. СПб., 2011.

Заде Л. А. Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976.

Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб.: Филол. ф-т СПбГУ, 2013.

Иванов В. В. Чет и нечет: асимметрия мозга и знаковых систем. М.: Советское радио, 1978.

Иорданская Л. Н. Межсегментный синтаксический анализ // Автоматический синтаксический анализ: в 2 т. / под общ. ред. А. А. Ляпунова, О. С. Кулагиной. Т. 2. Новосибирск: Наука. Сиб. отделение, 1967.

Кнорозов Ю. В. Письменность индейцев майя. М.; Л.: Изд-во АН СССР, 1963.

Ковальчук М. В. Конвергенция наук и технологий — прорыв в будущее // Российские нанотехнологии. 2011. Т. 6. № 1–2. С. 1–16.

Ковальчук М. В., Нарайкин О. С., Яцишина Е. Б. Конвергенция наук и технологий — новый этап научно-технического развития // Вопросы философии. 2013. № 3. С. 3–11.

Кун Т. Структура научных революций. М.: АСТ, 2003.

Лабов У. Структура денотативных значений // Новое в зарубежной лингвистике. Вып. 14: Проблемы и методы лексикографии. М.: Прогресс, 1983. С. 133–176.

Лакофф Дж. Женщины, огонь и опасные вещи: Что категории языка говорят нам о мышлении. М.: Языки славянской культуры, 2004.

Лефевр В. А. Конфликтующие структуры. М.: Советское радио, 1973.

Лесерф И. Применение программы и модели конфликтной ситуации к автоматическому синтаксическому анализу естественных языков // Научно-техническая информация. 1963. № 10. С. 42–50.

Марков А. А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Императорской Академии наук. Серия 6. 1913. Т. 7. Вып. 3. С. 153–162.

Мейен С. В. М. Ф. Нейбург — 40 лет служения «малой» науке // Страницы из истории московской геологической школы. Вып. 22: Очерки по истории геологических знаний. М.: Наука, 1985. С. 62–79.

Мельчук И. А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М.: Наука, 1974.

Меркулов И. П. Когнитивная наука // Энциклопедия эпистемологии и философии науки / под ред. И. Т. Касавина. М.: Канон+; РООИ «Реабилитация», 2009. С. 364–365.

Минский М. Фреймы для представления знаний. М.: Энергия, 1979.

Мусхелишвили Н. Л., Шрейдер Ю. А. Значение текста как внутренний образ // Вопросы психологии. 1997. № 3. С. 79–91.

Налимов В. В. Вероятностная модель языка: О соотношении естественных и искусственных языков. М.: Наука, 1974.

Налимов В. В. О возможности метафорического использования математических моделей в психологии // Психологический журнал. 1981. Т. 2. № 3. С. 39–47.

Нейман Дж. фон. Теория самовоспроизводящихся автоматов. М.: Мир, 1971.

Немцева-Плахотя В. В., Чебанов С. В. Лингвостатистические последствия орфографической реформы 1918 г. // Структурная и прикладная лингвистика: межвуз. сб. Вып. 8 / под ред. А. С. Герда. СПб.: Изд-во СПбГУ, 2010. С. 60–90.

Прикладная и компьютерная лингвистика / ред. И. С. Николаев, О. В. Митренина, Т. М. Ландо. М.: Ленанд, 2017.

Принципы и методы биохимии и молекулярной биологии / ред. К. Уилсон, Дж. Уолкер. М.: Бином. Лаборатория знаний, 2015.

Риккерт Г. Науки о природе и науки о культуре. СПб.: Образование, 1911.

Розалиев В. Л., Заболеева-Зотова А. В. Моделирование эмоционального состояния человека на основе гибридных методов // Программные продукты и системы: междунар. науч.-практ. журнал. 2010. Вып. 2. С. 141–146.

Сухотин Б. В. Оптимизационные методы исследования языка. М.: Наука, 1976.

Теньер Л. Основы структурного синтаксиса. М.: Прогресс, 1988.

Тищенко П. Геномика: новый тип науки в новой культурной ситуации // BioMediale: Современное общество и геномная культура: [антология] / сост., общ. ред. Д. Булатова. Калининград: Янтарный сказ, 2004. С. 60–72.

Тьюринг А. Может ли машина мыслить? М.: Наука, 1960.

Успенский В. А. Системы перечислимых множеств и их нумерации // Доклады АН СССР. 1955. Т. 105. № 6. С. 1155–1158.

Успенский В. А. К определению падежа по Колмогорову // Бюллетень Объединения по проблемам машинного перевода. 1957. № 5. С. 1–18.

Филлмор Ч. Дж. Фреймы и семантика понимания // Новое в зарубежной лингвистике. Вып. 23: Когнитивные аспекты языка. М.: Прогресс, 1988. С. 52–92.

Фитиалов С. Я. Об эквивалентности грамматик НС и грамматик зависимостей // Проблемы структурной лингвистики 1967. М.: Наука, 1968. С. 71–102.

Харре Р. Вторая когнитивная революция // Психологический журнал. 1996. № 2. С. 3–15.

Хомский Н. Синтаксические структуры // Новое в лингвистике. Вып. 2 / сост. В. А. Звегинцев. М.: Изд-во иностр. литературы, 1962. С. 412–527.

Чебанов С. Г. О подчинении речевых укладов «индоевропейской» группы закону Пуассона // Доклады АН СССР. 1947. Т. 55. № 2. С. 103–106.

Чебанов С. В., Буянова Д. С., Писарев А. С. С. Г. Чебанов, его архив и распределение слов по числу слогов // Исследования языка и современное гуманитарное знание. 2019. Т. 1. № 1. С. 67–78.

Чебанов С. В., Кузнецова Д. С. С. Г. Чебанов и его архив // Структурная и прикладная лингвистика. Вып. 9 / под ред. А. С. Герда. СПб.: Изд-во СПбГУ, 2012. С. 330–339.

Чейф У. Л. Данное, контрастивность, определенность, подлежащее, топики и точка зрения // Новое в зарубежной лингвистике. Вып. 11: Современные синтаксические теории в американской лингвистике. М.: Прогресс, 1982. С. 277–316.

Шаумян С. К., Соболева П. А. Аппликативная порождающая модель и исчисление трансформаций в русском языке. М.: Изд-во АН СССР, 1963.

Шенк Р. Обработка концептуальной информации. М.: Энергия, 1980.

Шеннон К. Работы по теории информации и кибернетике. М.: Изд-во иностр. литературы, 1963.

Шрейдер Ю. А. Присущ ли машине разум? // Вопросы философии. 1975. № 2. С. 82–89.

Шрейдер Ю. А. ЭВМ как средство представления знаний // Природа. 1986. № 10. С. 14–22.

Bainbridge M. S., Roco M. C. Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society. New York: Springer, 2005.

Berlin B., Kay P. Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press, 1969.

Dörnye Z. Research Methods in Applied Linguistics. Oxford: Oxford University Press, 2007.

Harris Z. S. From Phoneme to Morpheme // Language. 1955. Vol. 31. No. 2. P. 190–222.

Juilland A. Outline of a General Theory of Structural Relations. The Hague: Mouton, 1961. (Janua Linguatum. No. 15).

Mervis C. B., Rosch E. Categorization of Natural Objects // Annual Review of Psychology. 1981. Vol. 32. P. 89–115.

Pevsner J. Bioinformatics and Functional Genomics. 3rd ed. Chichester; Hoboken (NJ): Wiley-Blackwell, 2015.

Shannon C. E. A Mathematical Theory of Communication // Bell System Technical Journal. 1948. Vol. 27. P. 379–423.

References

Ado I. 2002. *Liberal Arts and Philosophy in the Antique Idea*. Moscow, Greko-Latinskii kabinet Iu. A. Shichalina Publ. (In Russ.)

Alekseeva I. Iu., Arshinov V. I., Chekletsov V. V. 2013. "Technohuman" versus "Posthuman": NBICS-revolution and Future of Humanity. *Voprosy filosofii*, no. 3, pp. 12–21. (In Russ.)

Applied and Computer Linguistics 2017, I. S. Nikolaev, O. V. Mitrenina, T. M. Lando (eds.). Moscow, Lenand Publ. (In Russ.)

Bainbridge M. S., Roco M. C. 2005. *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society.* New York, Springer.

Beliaeva L. N. 2001. *Linguistic Automaton in the Modern Data Tecnologies*. Saint Petersburg, RGPU im. A. I. Gertsena Publ. (In Russ.)

Berlin B., Kay P. 1969. *Basic Color Terms: Their Universality and Evolution.* Berkeley, University of California Press.

Chafe W. L. 1982. Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View. *Novoe v zarubezhnoi lingvistike*, vol. 11. Sovremennye sintaksicheskie teorii v amerikanskoi lingvistike. Rus. Ed. Moscow, Progress Publ., pp. 277–316 (In Russ.)

Chebanov S. G. 1947. On Conformity of Language Structures within the Indo-European Family to Poisson's Law. *Doklady AN SSSR*, vol. 55, no. 2, pp. 103–106. (In Russ.)

Chebanov S. V., Buianova D. S., Pisarev A. S. 2019. S. G. Chebanov, His Archive and Distribution of Words by the Number of Sullables. *Issledovaniia iazyka i sovremennoe gumanitarnoe znanie*, vol. 1, no. 1, pp. 67–78. (In Russ.)

Chebanov S. V., Kuznetsova D. S. 2012. S. G. Chebanov and His Archive. *Strukturnaia i prikladnaia lingvistika*, vol. 9, A. S. Gerd (ed.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 330–339. (In Russ.)

Chomsky N. 1962. Syntactic Structures. *Novoe v lingvistike*, vol. 2, V. A. Zvegintsev (comp.). Rus. Ed. Moscow, Izd-vo inostr. literatury Publ., pp. 412–527. (In Russ.)

Dilthey W. 2000. *Works*, in 6 vols., vol. 1. Introduction to the Human Sciences: An Attempt to Lay a Foundation for the Study of Society and History. Rus. Ed. Moscow, Dom intellektual'noi knigi Publ. (In Russ.)

Dörnye Z. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Fillmore Ch. J. 1988. Frames and the Semantics of Understanding. *Novoe v zarubezhnoi lingvistike*, vol. 23. Kognitivnye aspekty iazyka. Rus. Ed. Moscow, Progress Publ., pp. 52–92. (In Russ.)

Fitialov S. Ia. 1968. About Equivalence of Immediate Constituent Grammar and Dependency Grammar. *Problemy strukturnoi lingvistiki 1967*. Moscow, Nauka Publ., pp. 71–102. (In Russ.)

Gladkii A. V. 1982. Mathematical Linguistics. *Matematicheskaia entsiklopediia*, in 5 vols., I. M. Vinogradov (ed. in chief), vol. 3. Moscow, Sovetskaia entsiklopediia Publ., pp. 565–568. (In Russ.)

Harré R. 1996. Second Cognitive Revolution. Rus. Ed. *Psikhologicheskii zhurnal*, no. 2, pp. 3–15. (In Russ.)

Harris Z. S. 1955. From Phoneme to Morpheme. *Language*, vol. 31, no. 2, pp. 190–222.

Iordanskaia L. N. 1967. Intersegmental Syntactical Analysis. *Avtomaticheskii sintaksicheskii analiz*, in 2 vols., A. A. Liapunov, O. S. Kulagina (eds.), vol. 2. Novosibirsk: Nauka. Sib. otdelenie Publ. (In Russ.)

Ivanov V. V. 1978. *Odd and Even: Asimmetry of Brain and System of Signs*. Moscow, Sovetskoe radio Publ. (In Russ.)

Juilland A. 1961. *Outline of a General Theory of Structural Relations*. The Hague, Mouton. (Janua Linguatum, no. 15).

Knorozov Iu. V. 1963. *Written Language of the Ancient Maya*. Moscow; Leningrad, Izd-vo AN SSSR Publ. (In Russ.)

Koval'chuk M. V. 2011. Convergence of Sciences and Technologies — Breakthrough to the Future. *Rossiiskie nanotekhnologii*, vol. 6, nos. 1–2, pp. 1–16. (In Russ.)

Koval'chuk M. V., Naraikin O. S., Iatsishina E. B. 2013. Convergence of Sciences and Technologies — New Stage of Scientific and Technical Development. *Voprosy filosofii*, no. 3, pp. 3–11. (In Russ.)

Kuhn T. 2003. *The Structure of Scientific Revolutions*. Rus. Ed. Moscow, AST Publ. (In Russ.)

Labov W. 1983. Denotational Structure. *Novoe v zarubezhnoi lingvistike*, vol. 14. Problemy i metody leksikografii. Rus. Ed. Moscow, Progress Publ., pp. 133–176. (In Russ.)

Lakoff G. 2004. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Rus. Ed. Moscow, Iazyki slavianskoi kul'tury Publ. (In Russ.)

Lecerf J. 1963. Program of Conflicts, Model of Conflicts. Rus. Ed. *Nauchno-tekhnicheskaia informatsiia*, no. 10, pp. 42–50. (In Russ.)

Lefebvre V. A. 1973. *Conflicting Structures*. Rus. Ed. Moscow, Sovetskoe radio Publ. (In Russ.)

Markov A. A. 1913. Example of Statistical Research of the Text of the Poem "Eugene Onegin". *Izvestiia Imperatorskoi Akademii nauk. Series 6*, vol. 7, issue 3, pp. 153–162. (In Russ.)

Meien S. V. 1985. M. F. Neuburg — 40 Years of Service to the "Small" Science. S*tranitsy iz istorii moskovskoi geologicheskoi shkoly*, issue 22. Ocherki po istorii geologicheskikh znanii. Moscow, Nauka Publ., pp. 62–79. (In Russ.)

Mel'chuk I. A. 1974. *The Experience of the Theory of Linguistic Models: Meaning ⇔ Text*. Moscow, Nauka Publ. (In Russ.)

Merkulov I. P. 2009. Cognitive Science. *Entsiklopediia epistemologii i filosofii nauki*, I. T. Kasavin (ed.). Moscow, Kanon+ Publ.; ROOI "Reabilitatsiia" Publ., pp. 364–365. (In Russ.)

Mervis C. B., Rosch E. 1981. Categorization of Natural Objects. *Annual Review of Psychology*, vol. 32, pp. 89–115.

Minskii M. 1979. *Framework for Representing Knowledge*. Moscow, Energiia Publ. (In Russ.)

Muskhelishvili N. L., Shreider Iu. A. 1997. Meaning of Text as Internal Image. *Voprosy psikhologii*, no. 3, pp. 79–91. (In Russ.)

Nalimov V. V. 1974. *Probabilistic Model of Language. About Correlation of Natural and Artificial Languages*. Moscow, Nauka Publ. (In Russ.)

Nalimov V. V. 1981. About Possibility of the Metaphorical Use of Mathematical Patterns in Psychology. *Psikhologicheskii zhurnal*, vol. 2, no. 3, pp. 39–47. (In Russ.)

Nemtseva-Plakhotia V. V., Chebanov S. V. 2010. Linguostatistical Consequences of Spelling Reform of 1918. *Strukturnaia i prikladnaia lingvistika*, vol. 8, A. S. Gerd (ed.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 60–90. (In Russ.)

Neumann J. von. 1971. *Theory of Self-reproducing Automata*. Rus. Ed. Moscow, Mir Publ. (In Russ.)

Pevsner J. 2015. *Bioinformatics and Functional Genomics*, 3rd ed. Chichester; Hoboken (NJ), Wiley-Blackwell.

Principles and Techniques of Biochemistry and Molecular Biology 2015, K. Wilson, J. Walker (eds.). Rus. Ed. Moscow, Binom. Laboratoriia znanii Publ. (In Russ.)

Rickert H. 1911. *Science and History: A Critique of Positivist Epistemology*. Rus. Ed. Saint Petersburg, Obrazovanie Publ. (In Russ.)

Rozaliev V. L., Zaboleeva-Zotova A. V. 2010. Modelling an Emotional Condition of the Person on the Basis of Hybrid Methods. *Programmnye produkty i sistemy: mezhdunar. nauch.-prakt. zhurnal*, issue 2, pp. 141–146. (In Russ.)

Schank R. 1980. *Conceptual Information Processing*. Rus. Ed. Moscow, Energiia Publ. (In Russ.)

Schreider Iu. A. 1975. Is Inherent Intellect in Machine? *Voprosy filosofii*, no. 2, pp. 82–89. (In Russ.).

Schreider Iu. A. 1986. Computer as Means of Presentation of Knowledge. *Priroda*, no. 10, pp. 14–22. (In Russ.)

Shannon C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, vol. 27, pp. 379–423.

Shannon C. E. 1963. Works on Information Theory and Cybernetics. Rus. Ed. Moscow, Izd-vo inostr. literatury Publ. (In Russ.)

Shaumian S. K., Soboleva P. A. 1963. *Applicational Generative Model and Transformational Calculus of Russian*. Moscow, Izd-vo AN SSSR Publ. (In Russ.)

Sukhotin B. V. 1976. *Optimization Methods of the Research of Language*. Moscow, Nauka Publ. (In Russ.)

Tesnière L. 1988. *Elements of Structural Syntax*. Rus. Ed. Moscow, Progress Publ. (In Russ.)

Tishchenko P. 2004. Genomics: New Type of Science in New Cultural Situation. *BioMediale. Sovremennoe obshchestvo i genomnaia kultura. Antologiia*, D. Bulatov (comp., ed.). Kaliningrad: Iantarnyi skaz Publ., pp. 60–72. (In Russ.)

Turing A. 1960. *Can the Machine Think?* Rus. Ed. Moscow, Nauka Publ. (In Russ.)

Uspenskii V. A. 1955. Systems of Enumerable Set and Their Numeration. *Doklady AN SSSR*, vol. 105, no. 6, pp. 1155–1158. (In Russ.)

Uspenskii V. A. 1957. To the Definition of Case According to Kolmogorov. *Biulleten' Ob"edineniia po problemam mashinnogo perevoda*, no. 5, pp. 1–18. (In Russ.)

Wiener N. 1968. *Cybernetics or Control and Communication in the Animal and the Machine*. Rus. Ed. Moscow, Sovetskoe radio Publ. (In Russ.)

Windelband W. 1995. *Selected Works. Spirit and History*. Rus. Ed. Moscow, Iurist Publ. (In Russ.)

Winograd T. 1976. U*nderstanding Natural Language*. Rus. Ed. Moscow, Mir Publ. (In Russ.)

Zadeh L. A. 1976. *The Concept of a Lingustic Variable and its Application to Approximate Reasoning*. Rus. Ed. Moscow, Mir Publ. (In Russ.)

Zakharov V. P., Bogdanova S. Iu. 2013. *Corpus Linguistics*. Saint Petersburg: Filol. f-t SPbGU Publ. (In Russ.)

Zhernakova D. V. 2011. *Development of the System of Classification of Short Texts with the Use of Ontology. Master's thesis*, Saint Petersburg State University. Saint Petersburg. (In Russ.)

Е. Л. Алексеева

К ВОПРОСУ О КЛАСТЕРНОМ АНАЛИЗЕ В ТЕКСТОЛОГИИ (на примере славянских переводов евангелия)

Аннотация. Критическое издание славянского перевода Евангелия от Матфея по 28 рукописям (2005) основано на исследовании более пятисот древнеславянских рукописей X–XVI веков. В качестве материала для кластерного анализа был выбран фрагмент объемом 300 словоупотреблений, встречающийся во всех типах текста Евангелия: четьем, служебном и толковом. В результате анализа было выделено 6 типов евангельского текста. Нами для кластерного анализа был выбран другой фрагмент Евангелия, особенность которого заключается в том, что в служебных евангелиях он используется дважды. Проведенное исследование показывает, что результат классификации евангельских рукописей по типу текста зависит от выбранного в качестве основы фрагмента текста, и демонстрирует актуальность создания корпуса славянских переводов евангелия для обеспечения возможности автоматической кластеризации рукописей по любому фрагменту текста.

Ключевые слова. Славянский перевод, Евангелие, критическое издание, текстология, кластерный анализ.

E. L. Alekseeva

CLUSTER ANALYSIS AND TEXTUAL CRITICISM (On the Example of Slavic Translations of the Gospel)

Abstract. The crirtical edition of the Slavic version of the Gospel according to Matthew using 28 manuscripts (2005) was preceded by the study of over 500 old Slavonic MSS from the 10th–16th centuries. The cluster analysis was carried out on the text fragment of 300 tokens occurring in tetraevangelia, lectionaries and Theophylact's commentaries on the Gospels, producing as a result 6 text-types. We have chosen a different text fragment for the cluster analysis, the one occurring twice in lectionaries. Our experiment shows that the results of cluster analysis of Gospels MSS vary as we choose a different passage for comparison, which emphasises the importance of building a corpus of the Slavic Gospels, opening way for automatic clustering of MSS for any given fragment of the text.

Keywords. Slavic version, Gospels, critical edition, textual criticism, cluster analysis.

1. Вводные замечания

В 2005 году было издано критическое издание церковнославянского текста Евангелия от Матфея: в основу издания было положено Мариинское евангелие XI века, а в аппарате полностью приведены разночтения по 27 другим рукописям [Евангелие…, 2005]. Изданию предшествовало исследование рукописной традиции Евангелия: на основании коллаций 532 рукописей в объеме фрагмента Мф 14.14–34 была осуществлена их классификация методом кластерного анализа, позволившая выделить 6 типов текста и установить для каждого типа лучших представителей, вошедших в издание [Миронова, 2005].

В издание включены все разновидности евангельских текстов: служебные (апракосы), четьи (четвероевангелия, или тетры) и толковые. Если в четвероевангелии одно за другим следуют четыре канонических евангелия (от Матфея, Марка, Луки, Иоанна), в апракосе евангельский текст разбит на отрывки, которые приводятся в том порядке, в каком они читаются во время богослужения в течение года.

В полных апракосах содержатся чтения на все дни года, за исключением шести недель Великого поста, когда литургия служится только в субботу и воскресенье; в кратких апракосах представлены только субботние и воскресные чтения, за исключением Страстной недели и семи недель от Пасхи до Пятидесятницы — для этого периода объем чтений полного и краткого апракоса совпадает. Поскольку некоторые перикопы читаются в течение года два-три раза, в апракосе либо текст воспроизводится соответствующее число раз, либо используются отсылки к ранее встретившемуся чтению.

Во время сверки рукописей при подготовке издания было замечено, что в главе 21 Евангелия от Матфея имеются значительные расхождения между повторяющимися перикопами в составе полных апракосов, поэтому было решено провести кластерный анализ включенных в издание рукописей на материале фрагмента Мф 21.18–32 с учетом встречающихся дважды апракосных чтений.

2. Материал исследования

Рукописи, использованные в издании, перечислены в разделе «Сокращения» в конце статьи. Распределение этих рукописей по типу текста и функциональной разновидности представлено в табл. 1.

Таблица 1. **Распределение рукописей по типу текста и функциональной разновидности**

Тип текста / Функциональная разновидность	Древний текст	Развитие древнего текста	Преславский текст	Периферия древнего и преславского текстов	Поздний текст	Чудовский Новый Завет
Тетр	Gl ME Tp Zg	Lc	—	Bn Fl	TL A B OB	Cd Pg
Краткий апракос	As OE	SK Or	Ar	—	—	—
Полный апракос	Mr	Kr	Dl Gf Ju Tr Vl	Vk	—	—
Воскресный апракос	—	Uv	—	—	—	—
Толковое евангелие	—	—	—	—	Th	—

Текст Мф 21.18–32 в течение года читается во время богослужения дважды: во-первых, в составе большой перикопы (обычно стихи 18–43) на Страстной неделе на утрене Великого понедельника, а во-вторых, разбитый на три отрывка, на литургии с понедельника по среду десятой недели по Пятидесятнице. Соответственно, этот текст в апракосах, в отличие от тетров и толкового евангелия, встречается в неодинаковом объеме. Он полностью отсутствует в кратких апракосах Ar, As, OE и SK, и только в Or есть чтение утрени Великого понедельника (в воскресном апракосе Uv от этого чтения из-за дефекта рукописи до нас дошли только первые три стиха). Из восьми полных апракосов в пяти — Dl, Gf, Kr, Tr, Vk — есть и чтение утрени (будем отмечать его *m* при сигле рукописи), и чтения десятой недели (будем отмечать их *s* при сигле рукописи); в Mr они представлены в неполном объеме — стихи 18–28 и 18–27 соответственно; в апракосах Ju и Vl имеются только чтения десятой недели.

3. Результаты кластерного анализа

Для проведения кластерного анализа на интересующем нас отрезке евангелия мы использовали следующие тексты: 12 тетров (без Zg), толковое евангелие, чтения утрени из апракосов: Orm, Dlm, Gfm, Krm, Mrm, Trm, Vkm, чтения десятой недели из апракосов: Dls, Gfs, Jus, Krs, Mrs, Trs, Vks и Vls. Кроме того, мы привлекли к сравнению еще один полный апракос — Мстиславово евангелие (ГИМ, Син. 1203, нач. XII в.), в котором есть чтения и утрени, и десятой недели (Msm и Mss). На этом отрезке текста было получено 117 узлов разночтений, чего вполне достаточно для проведения кластерного анализа. Итоги разбиения рукописей на кластеры приведены в табл. 2.

Мы видим, что выделилось 7 кластеров:

1) B, OB, A, TL — поздний текст, при этом TL занимает промежуточное положение между этим кластером и следующим, находясь на периферии обоих;
2) Gl, ME, Tp, Vls, Lc — древний текст, на периферии этого кластера находятся Mrm и Orm;
3) Trs, Mss, Jus, Dls — преславский текст; сюда попали чтения десятой недели по Пятидесятнице;
4) Krs, Krm, Bn — развитие древнего текста;
5) Gfs, Vks, Vkm — преславский текст, сербские апракосы;
6) Gfm, Msm, Trm, Dlm — преславский текст; сюда вошли чтения утрени Великого понедельника; на периферии — Fl, Mrs;
7) Cd, Pg — группа Чудовского Нового Завета с толковым евангелием (Th) на периферии.

Кластерный анализ подтвердил предположение о несовпадении в составе некоторых полных апракосов текстов, читаемых на утрене Великого понедельника (кластер 6) и на литургии в понедельник, вторник и среду десятой недели по Пятидесятнице (кластер 3).

При этом, если кластер 3 имеет примерно 65–70 % сходства с соседними кластерами (1, 2, 4, 5), то кластер 6, как и кластер Чудовского Нового Завета, занимает в таблице 2 периферийную позицию.

4. Интерпретация полученных результатов

Интересно проанализировать характер разночтений, выделяющих кластеры 3 и 6. Поскольку самый ранний представитель этих

кластеров — Мстиславово евангелие — датируется началом XII века (не позднее 1117 года), мы не будем учитывать более поздние группы рукописей, относящиеся к 1-му и 7-му кластерам, поскольку они не имеют отношения к истории исследуемого текста.

1. В то время как для полных апракосов, относящихся к преславскому типу текста, характерен перевод на славянский язык греческой церковной терминологии, и именно такая терминология многократно употребляется на протяжении всего текста апракоса, в отрывке, читаемом на утрене, сохранены грецизмы: 21.21 и 21.31 *аминь* вместо *право*, 21.23 *архиереи* вместо *стареишины жьрьчьскы*.
2. Другие лексические противопоставления:
   - 21.19–20: *смокы* (только кластер 3) — *смоковьница* (остальные кластеры);
   - 21.21: *и* — *ти* (только кластер 6);
   - 21.21: *смоковьное (-ничьное, -ничьское)* — *смокъвие* (только кластер 6);
   - 21.24: *речете* — *повѣсте* (только кластер 6);
   - 21.25, 32: *не ѣсте ему вѣры* — *не вѣроваете ему*;
   - 21.32: *вѣровати ему* — *ѣти ему вѣры*;
   - 21.28: *сыну* (только кластер 3) — *чѧдо*;
   - 21.31: *волѣ отьчѣ* — *волѣ отьцу* (только кластер 6);
   - 21.31–32: *любодѣица* — *блѣдьница* (только кластер 6).
3. Грамматические различия:
   - 21.26: *имѣтъ* — *имѣахѫ* (только кластер 6);
   - 21.30: *пристѫпль* — *пристѫпивъ* (только кластер 6);
   - 21.31: *вы* — *васъ* (только кластер 6);
   - 21.32: *и* — *его* (только кластер 6).
4. Добавления:
   - 21.23: *старьци людьстии* — *с. л. и стареишины жьрьць* (только кластер 6);
   - 21.24: *речете* — *хощѫ да повѣсте* (только кластер 6);
   - 21.32: *приде бо* — *приде бо къ вамъ*.
5. Омиссии:
   - 21.21: *двигни сѧ и въвръзи сѧ* — *въвръзи сѧ* (только кластер 6);

Таблица 2. **Итоговая матрица кластерного**

№	Текст	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	B	0	98	88	78	79	77	78	72	77	70	62	63	64	68
2	OB	98	0	88	79	80	78	79	73	76	69	63	64	65	69
3	A	88	88	0	73	79	80	77	70	80	73	63	62	61	65
4	TL	78	79	73	0	77	71	77	68	67	67	65	63	62	67
5	Gl	79	80	79	77	0	91	89	81	80	71	70	72	68	81
6	ME	77	78	80	71	91	0	88	81	81	71	69	70	65	76
7	Tp	78	79	77	77	89	88	0	88	78	69	70	71	65	78
8	Vls	72	73	70	68	81	81	88	0	71	63	69	69	66	74
9	Lc	77	76	80	67	80	81	78	71	0	72	66	64	62	68
10	Mrm	70	69	73	67	71	71	69	63	72	0	67	59	58	62
11	Orm	62	63	63	65	70	69	70	69	66	67	0	56	62	60
12	Trs	63	64	62	63	72	70	71	69	64	59	56	0	89	87
13	Mss	64	65	61	62	68	65	65	66	62	58	62	89	0	83
14	Jus	68	69	65	67	81	76	78	74	68	62	60	87	83	0
15	Dls	63	63	59	62	65	62	63	69	59	56	57	84	87	81
16	Krs	69	68	70	61	72	71	71	74	77	73	68	61	66	67
17	Krm	68	68	70	66	77	76	74	70	75	69	65	62	58	66
18	Bn	70	71	73	66	74	74	73	71	72	66	68	52	59	61
19	Gfs	62	63	58	65	66	62	66	68	60	56	63	67	71	68
20	Vks	67	66	63	61	72	69	70	74	65	70	61	75	73	73
21	Vkm	69	71	68	70	74	72	74	72	65	70	64	58	58	64
22	Gfm	60	60	59	66	55	53	63	55	53	49	49	53	52	52
23	Msm	63	63	61	69	59	56	66	58	55	51	53	57	55	56
24	Trm	62	60	57	65	54	52	62	55	56	51	54	57	55	52
25	Dlm	59	59	55	59	52	53	57	54	50	45	45	51	48	48
26	Fl	60	61	56	70	58	55	62	57	53	52	53	56	52	56
27	Mrs	62	63	64	67	63	65	68	68	61	67	62	64	59	63
28	Cd	63	63	64	57	51	53	51	46	56	45	41	44	45	47
29	Pg	66	66	67	60	56	54	57	54	59	52	44	47	50	50
30	Th	63	64	66	63	59	59	60	52	58	58	49	56	53	58

15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
63	69	68	70	62	67	69	60	63	62	59	60	62	63	66	63
63	68	68	71	63	66	71	60	63	60	59	61	63	63	66	64
59	70	70	73	58	63	68	59	61	57	55	56	64	64	67	66
62	61	66	66	65	61	70	66	69	65	59	70	67	57	60	63
65	72	77	74	66	72	74	55	59	54	52	58	63	51	56	59
62	71	76	74	62	69	72	53	56	52	53	55	65	53	54	59
63	71	74	73	66	70	74	63	66	62	57	62	68	51	57	60
69	74	70	71	68	74	72	55	58	55	54	57	68	46	54	52
59	77	75	72	60	65	65	53	55	56	50	53	61	56	59	58
56	73	69	66	56	70	70	49	51	51	45	52	67	45	52	58
57	68	65	68	63	61	64	49	53	54	45	53	62	41	44	49
84	61	62	52	67	75	58	53	57	57	51	56	64	44	47	56
87	66	58	59	71	73	58	52	55	55	48	52	59	45	50	53
81	67	66	61	68	73	64	52	56	52	48	56	63	47	50	58
0	63	57	55	69	71	60	48	52	53	54	59	56	45	51	54
63	0	81	75	62	74	70	53	55	54	53	55	63	45	52	49
57	81	0	71	59	66	68	52	54	50	52	51	55	49	53	55
55	75	71	0	63	64	72	58	57	54	58	56	63	49	55	53
69	62	59	63	0	80	73	56	57	52	54	54	58	44	48	50
71	74	66	64	80	0	78	53	55	55	51	52	63	41	49	47
60	70	68	72	73	78	0	58	60	55	56	55	64	49	56	54
48	53	52	58	56	53	58	0	94	84	84	71	71	46	50	54
52	55	54	57	57	55	60	94	0	87	85	72	70	48	53	56
53	54	50	54	52	55	55	84	87	0	79	74	68	50	54	55
54	53	52	58	54	51	56	84	85	79	0	71	63	51	52	54
59	55	51	56	54	52	55	71	72	74	71	0	70	44	50	52
56	63	55	63	58	63	64	71	70	68	63	70	0	47	52	53
45	45	49	49	44	41	49	46	48	50	51	44	47	0	89	73
51	52	53	55	48	49	56	50	53	54	52	50	52	89	0	69
54	49	55	53	50	47	54	54	56	55	54	52	53	73	69	0

- 21.30: *идѫ господи — господи*;
- 21.32: *иоанъ крьститель — иоанъ*.

6. Перестановка:
   - 21.19: *абие усъше* (только кластер 3) — *усъше абие* (только кластер 6).

Судя по грамматическим различиям, можно сказать, что текст утрени более поздний, чем текст, читаемый на десятой неделе. К сожалению, отсутствие критического издания греческого лекционария не позволяет проверить происхождение добавлений и некоторых лексических замен.

Интересно чтение *вольѫ отьцу* вместо *вольѫ отьчѧ*: возможно, оно объясняется новгородским цоканьем, и в таком случае можно предположить, что источником для других апракосов с таким чтением на утрене послужило Мстиславово евангелие. Такой вывод согласуется с мнением крупнейшего специалиста в области изучения славянского евангелия Л. П. Жуковской, которая доказывала, что славянский полный апракос как особый тип текста появился в Древней Руси [Жуковская, 1959].

С другой стороны, в той же статье Жуковская пишет, что в полных апракосах части текста, общие с краткими апракосами (сюда, вообще говоря, относится и чтение утрени Великого понедельника), близки кратким апракосам и древним тетрам, в то время как тексты, дополняющие краткий апракос до полного (чтения для будних дней периода от Пятидесятницы до Великого поста), представляют собой особый перевод [Жуковская, 1959, с. 96]. Кластерный анализ показывает несколько иную картину: чтения будних дней образуют периферию древнего типа текста, представленного здесь тетрами, а чтения утрени стоят особняком.

Есть и еще один момент, требующий дополнительного исследования. Сербский полный апракос Gf в чтениях десятой недели образует кластер с другим сербским памятником, Vk, имея с новгородским Мстиславовым евангелием всего 70 % сходства; в чтении же утрени он совпадает с Ms на 94 %. Известно, что Ms переплетали в Царьграде [Сводный каталог…, 1984, с. 90] — следует ли отсюда, что рукопись была там скопирована и копия попала к южным славянам?

5. Выводы

Вышеприведенный опыт изучения фрагмента текста по ряду источников показывает вариативность разбиения рукописей на кластеры и демонстрирует актуальность проекта по трансформации критического издания в корпус славянских переводов евангелия [Азарова, Алексеева, 2015], что позволит проводить «скользящий» кластерный анализ по всему объему евангельского текста, выявляя подчас меняющийся характер связей между рукописями на всем его протяжении.

Сокращения

ГИМ — Государственный исторический музей.
НБКМ — Национальная библиотека им. св. Кирилла и Мефодия.
РГАДА — Российский государственный архив древних актов.
РГБ — Российская государственная библиотека.
РНБ — Российская национальная библиотека.
ФИРИ РАН — Санкт-Петербургский филиал Института российской истории Российской академии наук.

A — Афонский текст, редакция А (по рукописи: РНБ, F. п.I. 109, XIV–XV вв., тетр).
Ar — Архангельское евангелие, 1092 г., краткий апракос (РГБ, ф. 178, 1666).
As — Ассеманиево евангелие, XI в., краткий глаголический апракос (Ватикан, Cod. Slav. 3).
B — Афонский текст, редакция В (по рукописи: РНБ, Q. I. 13, XV в., тетр).
Bn — Баницкое евангелие, конец XIII — начало XIV в., тетр (София, НБКМ, 847).
Cd — Чудовский Новый Завет святителя Алексея, XIV в. (по фототипическому изданию [Новый Завѣтъ…, 1892]).
Dl — Добрилово евангелие, 1164 г., Галиция, полный апракос (РГБ, ф. 256, 103).
Fl — XIV в., тетр (РНБ, F. п.I. 14).
Gf — ок. 1284 г., полный апракос (РНБ, собр. Гильфердинга 1).
Gl — Галицкое евангелие, 1144 г., тетр (ГИМ, Син. 404).
Ju — Юрьевское евангелие, 1119–1128 гг., Новгород, полный апракос (ГИМ, Син. 1003).
Kr — Карпинское евангелие, конец XIII в., полный апракос (ГИМ, собр. Хлудова 28).

Lc — Евангелие Н. П. Лихачева, конец XIII — начало XIV в., тетр (ФИРИ РАН, ф. 238, оп. 1, № 223).

ME — Мариинское евангелие, XI в., тетр (РГБ, Григ. 6).

Mr — Мирославово евангелие, конец XII в., сербское, полный апракос (Белград, Народный музей, 1538).

OB — Острожская библия, 1581 г., текст редакции B (печатное издание [Библия…, 1581]).

OE — Остромирово евангелие, 1056/1057 г., Новгород, краткий апракос (РНБ, F. п.1.5).

Or — Орбельское евангелие, XIII–XIV вв., краткий апракос (РНБ, Q. п.I. 43).

Pg — вторая рукопись Чудовской группы, 2-я половина XIV в., тетр (РНБ, собр. Погодина, 21).

SK — Саввина книга, X в., краткий апракос (РГАДА, ф. 381, 14).

Th — Толковое евангелие Феофилакта Болгарского (по рукописи: РНБ, собр. Погодина, 174, XVI в.).

Tp — Типографское евангелие, XII в., тетр (РГАДА, ф. 381, 1).

TL — Новый литургический тетр (по рукописи: РНБ, собр. Гильфердинга, 2, XIV в.).

Tr — конец XII — начало XIII в., полный апракос (собр. Третьяковской галереи, К 5348).

Uv — начало XIV в., воскресный апракос (ГИМ, Увар. 379).

Vl — конец XIII в., полный апракос (РГБ, ф. 113 (Волоколамск.), 1).

Zg — Зографское евангелие, XI–XII вв., глаголический тетр (РНБ, Глаг. 1).

Vk — Вуканово евангелие, XIII в., полный апракос (РНБ, F. п.I. 82).

Источники

Библия: сиреч книгы Ветхаго и Новаго Завета, по языку словенску…. Острог: [Иван Федоров], 1581.

Евангелие от Матфея в славянской традиции / изд. подгот. А. А. Алексеев, И. В. Азарова, Е. Л. Алексеева и др.; ред. А. А. Алексеев. СПб.: Синод. б-ка Моск. патриархата, 2005.

Новый Заветъ Господа Нашего Иисуса Христа / [пер. с др.-греч.] Алексия, митр. Московскаго и всея Руси. М.: Фототип. изд. Леонтия, митр. Московского, 1892.

Сводный каталог славяно-русских рукописных книг, хранящихся в СССР: XI–XIII вв. / отв. ред. Л. П. Жуковская. М.: Наука, 1984.

Литература

Азарова И. В., Алексеева Е. Л. Использование аппарата критического издания Четвероевангелия для создания корпуса славянских переводов Еванге-

лия // Структурная и прикладная лингвистика: межвуз. сб. Вып. 11 / под ред. А. С. Герда, И. С. Николаева. СПб.: Изд-во СПбГУ, 2015. С. 75–85.

Жуковская Л. П. О переводах Евангелия на славянский язык и о «древнерусской редакции» славянского Евангелия // Славянское языкознание: сб. статей / отв. ред. В. В. Виноградов. М.: Изд-во АН СССР, 1959. С. 86–97.

Миронова Д. М. Классификация славянских рукописей Евангелия от Матфея // Евангелие от Матфея в славянской традиции / изд. подгот. А. А. Алексеев, И. В. Азарова, Е. Л. Алексеева и др.; ред. А. А. Алексеев. СПб.: Синод. б-ка Моск. патриархата, 2005. С. 163–168.

Sources

The Bible, or Books of the Old and New Testament, in Slavic Language… 1581. Ostrog, [Ivan Fedorov Publ.]. (In Russ.)

The Gospel According to Matthew in Slavonic Tradition 2005, A. A. Alekseev, I. V. Azarova, E. L. Alekseeva (preps. publ.), A. A. Alekseev (ed.). Saint Petersburg, Sinod. b-ka Mosk. patriarkhata Publ. (In Russ.)

The New Testament Of Our Lord Jesus Christ 1892, Alexy, Metropolitan of Moscow and all Russia (transl. from ancient Greek). Moscow, Fototip. izd. Leontiia, mitr. Moskovskogo Publ. (In Russ.)

The Comprehensive Catalogue of Slavic-Russian Manuscript Books, Stored the USSR. 11th–13th cc. 1984, L. P. Zhukovskaia (ed.). Moscow, Nauka Publ. (In Russ.)

References

Azarova I. V., Alekseeva E. L. 2015. Apparatus of the Critical Edition of the Slavic Tetraevangelion as the Basis for the Corpus of the Slavic Versions of the Gospels. *Strukturnaia i prikladnaia lingvistika. Mezhvuz. sb.*, vol. 11, A. S. Gerd, I. S. Nikolaev (eds.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 75–85. (In Russ.)

Zhukovskaia L. P. 1959. On Slavonic Translations of the Gospels and the "Old Russian redaction" of the Slavonic Gospels. *Slavianskoe iazykoznanie: sb. statei*, V. V. Vinogradov (ed.). Moscow, Izd-vo AN SSSR Publ., pp. 86–97. (In Russ.)

Mironova D. M. 2005. Classification of the Slavonic Manuscripts of the Gospel According to Matthew. *Evangelie ot Matfeia v slavanskoi traditsii*, A. A. Alekseev, I. V. Azarova, E. L. Alekseeva (preps. publ.), A. A. Alekseev (ed.). Saint Petersburg, Sinod. b-ka Mosk. patriarkhata Publ., pp. 163–168. (In Russ.)

В. П. Захаров

МЕТОДЫ АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ СЕМАНТИЧЕСКИХ ПОЛЕЙ*

Аннотация. Статья посвящена теме формирования семантических полей. В качестве объекта выбрано поле «империя» в русском, английском и чешском языках. Понятие семантического поля описано в литературе под разными названиями, сам термин интуитивно понятен, но формальных методов выявления наполнения полей не так много. Трудность состоит в том, что словарный состав полей представляет собой сложную систему отношений и оппозиций, как лингвистических, так и экстралингвистических. Возможны разные подходы к решению этой задачи. В статье описывается дистрибутивно-статистический подход на базе корпусов текстов. Конечным результатом должен стать тезаурус с количественными параметрами связанности лексических единиц. На последнем этапе должна быть произведена лингвистическая и культурно-историческая корреляция наполнения тезаурусов для трех языков. В статье представлены текущие результаты автоматизированного формирования семантического поля «империя» на основе данных корпусов и дистрибутивно-статистического метода. Идея состоит в том, что из данных об окружении одних лексических единиц можно извлечь другие, семантически связанные с первыми парадигматическими отношениями различной силы. Наличие больших корпусов и сложных алгоритмов анализа предоставляет такую возможность и позволяет достичь разумных результатов. В статье описываются инструменты (система Sketch Engine) и методология наполнения семантических полей на основе морфологически размеченных корпусов и специальной скетч-грамматики. Исходный текстовый материал был представлен корпусами собственной генерации на основе русских текстов XVIII–XX веков. В ходе работы было сформировано семантическое поле для русского языка для концепта «империя» и оценены полученные результаты. Также были проведены эксперименты по формированию семантического поля для концепта «империя» для чешского языка. В заключении определены дальнейшие шаги по улучшению качества полученных результатов.

Ключевые слова. Семантическое поле, семантические отношения, понятие империи, тезаурус, русский язык, концепт, корпус, дистрибутивно-статистический анализ.

---

* Исследование поддержано грантом РФФИ № 18-012-00474 «Семантическое поле „империя“ в русском, английском и чешском языках» и частично грантом РФФИ № 17-04-00552-ОГН-А «Параметрическое моделирование лексической системы современного русского литературного языка».

Victor P. Zakharov

WAYS OF THE AUTOMATIC CONSTRUCTION OF SEMANTIC FIELDS

Abstract. The paper is dealing with subject of forming semantic fields. The semantic field "empire" in 3 languages (Russian, English and Czech) was chosen as an object of investigation. The concept of "semantic field" is used in linguistics to denote a set of linguistic units united by a common semantic feature (or features), that is, having a common component(s) of meaning. Such lexical units are words and phrases, both common and proper names. The concept of semantic field has been described many times, the term itself is intuitive, but nevertheless there are not so many formal methods of identification of fields filling. The difficulty lies in the fact that the vocabulary of fields includes a complex system of relations and oppositions, both linguistic and extralinguistic. There are different approaches to this task. This paper describes a descriptive and statistical method based on linguistic corpora. The task is a specific lexicographic product (thesauri) with the quantitative characteristics of the connectedness of lexical units and examples from corpora. At the last stage linguistic and cultural-historical correlation of the content of these three thesauri is necessary. The paper presents ongoing results of automatic creation of a semantic field of "empire" based on distribution and statistical method using corpus data. The idea is to extract from data on syntagmatic collocability a set lexical units connected by semantic paradigmatic relations of various strength using distributional analyses techniques. Nowadays the presence of big corpora and sophisticated algorithms give the possibility and hope to reach a reasonable results. The paper describes tools of the Sketch Engine corpus system and methodology to fill semantic fields by lexical units on the basis of morphologically tagged corpora and special sketch grammar. Text material was represented by own topical Russian corpora created from Russian texts of 18th–20th centuries. In the course of work we have formed the semantic field for Russian for the concept of "empire" and evaluated results obtained. The similar experiments were fulfilled for the Czech language. At conclusion further steps were identified to clarify the perspective areas of work and to improve the results obtained.

Keywords. Semantic field, semantic relations, concept of the empire, thesaurus, Russian, concept, corpus, distributive and statistical analysis.

Введение

Работа многих автоматизированных систем обработки текстовой информации базируется на словарях. Как «уровневые» анализаторы, так и прикладные системы используют различные словари: словари сочетаемости, терминологические словари, фактографические справочники, словари валентностей и т. п. Среди них выделяют также семантические словари.

Существуют различные типы семантических словарей. Иногда они, будучи изоморфными по сути, отличаются наименованиями, иногда имеют существенные отличия с точки зрения наполнения и использования. Вообще можно сказать, что почти любой словарь

несет семантическую информацию о своих единицах и может рассматриваться как семантический. Самый простой тип семантических словарей — это толковые словари.

К семантически простым словарям можно отнести также и семантические поля. Именно они и являются предметом данного исследования.

1. Семантическое поле в лингвистике

Понятие «семантическое поле» применяется в лингвистике для обозначения совокупности языковых единиц, объединенных каким-то общим семантическим признаком, имеющих некоторый общий компонент значения (см. энциклопедию «Кругосвет»[1]). В роли таких лексических единиц выступают слова и словосочетания, как нарицательные, так и имена собственные. В основе теории семантических полей лежит представление о существовании в языке групп лексики, словарный состав которых объединен различными отношениями, как лингвистическими, так и экстралингвистическими, которые их связывают и которые одновременно представляют собой сложную систему оппозиций.

Термин этот имеет различные модификации (иногда их можно считать синонимами, иногда это сущностно различающиеся понятия), как то: поле, семантическое поле, лексическое поле, лексико-семантическое поле, функционально-семантическое поле, кластер, тезаурус, онтология. Сюда же можно добавить понятие терминосистемы. Каждый из этих терминов по-своему задает тип языковых единиц, входящих в поле, и/или тип связи между ними.

Приведем определение О. С. Ахмановой: «Поле — совокупность содержательных единиц, покрывающая определенную область человеческого опыта и образующая более или менее автономную микросистему» [Ахманова, 1966, с. 366].

Это понятие начало активно употребляться после выхода в свет работ неогумбольдтианцев Й. Трира и Г. Ипсена. Сам термин «семантическое поле» впервые был введен Г. Ипсеном [Ipsen, 1924]. Первые попытки выделения семантических полей были предпри-

[1] https://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/SEMANTICHESKOE_POLE.html (дата обращения: 30.10.2018).

няты при создании идеографических словарей, или тезаурусов, — например, тезауруса П. Роже. Семантический признак, лежащий в основе семантического поля, может также рассматриваться как некоторая понятийная категория [Бондарко, 1984; Васильев, 1990; Кобозева, 2000]. В трактовке В. Г. Адмони поле характеризуется наличием инвентаря элементов, связанных системными отношениями. 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, притом могут иметь и признаки, присущие соседним полям [Адмони, 1973]. 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 или 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. Теории семантического поля в лингвистике посвящено большое число работ [Уфимцева, 1961; Щур, 1974; Аскольдов, 1980; Семантико-функциональные поля…, 1990; Фрумкина, 1992; Апресян, 1995; Вежбицкая, 2001].

2. Семантическое поле в когнитивистике

Еще одно направление исследований по выявлению и описанию семантических полей связано с понятиями «языковая картина мира» и «языковое сознание». Здесь нужно отметить психолингвистические исследования, которые в нашей стране связаны с именами Н. И. Жинкина [Жинкин, 1982], А. А. Леонтьева [Леонтьев, 1988], Ю. Ф. Тарасова [Тарасов, 1996], А. А. Залевской [Залевская, 1977], Н. В. Уфимцевой [Уфимцева, 2011] и др. Психолингвисты старались преодолеть функциональную ограниченность традиционных значения и смысла и разработать теорию, в которой органически слились бы логико-психологические и языковедческие категории. Языковое сознание понимается как совокупность структур сознания, в формировании которых были использованы социальные знания, связанные с языковыми знаками, или как образы сознания, овнешняемые языковыми

средствами: отдельными лексемами, словосочетаниями, фразеологизмами, текстами, ассоциативными полями и ассоциативными тезаурусами как совокупностью этих полей.

Исследования, осуществляемые в московской психолингвистической школе в последние 20 лет на материале «Русского ассоциативного словаря» [Караулов и др., 1994–1998] и «Ассоциативного тезауруса английского языка» (The Associative Thesaurus of English [Kiss et al., 1972]), показали, что ассоциативный тезаурус является моделью сознания человека. Эта знаковая модель качественно отличается по презентации образов сознания от предметных представлений образов, включая в себя вербальные и невербальные значения.

Психолингвистическое направление фактически слилось с когнитивной лингвистикой. На сегодняшний день очень активно себя ведет термин «концепт», по частоте употребления значительно опередивший все прочие протерминологические новообразования. Необходимо отметить, что понятие концепта является достаточно разработанным в культурологии и лингвистике. Термин «концепт» покрывает предметные области нескольких научных направлений: прежде всего когнитивной психологии и когнитивной лингвистики, занимающихся проблемами мышления и познания, хранения и переработки информации, а также лингвокультурологии, определяясь и уточняясь в границах теории, образуемой их постулатами и базовыми категориями.

3. Постановка задачи

Ставится задача автоматического выявления синтагматических и парадигматических связей, а именно автоматическое наполнение лексико-семантических полей. В качестве исследовательской площадки» мы избрали семантическое поле «империя», можно сказать, мини-поле. Чтобы усложнить себе задачу и сделать ее поинтереснее, мы будем строить это поле для трех языков: русского, английского и чешского. Выбор наш можно объяснить тем, что в этих языках понятие «империя» сильно связано с исторической памятью народа и что оно живо в языковом сознании носителей языка. Также нам интересно исследовать разные языки, с одной стороны, принадлежащие к одной языковой семье (русский, чешский), с другой — к разным (английский).

Формирование семантического поля «империя» мы основываем на комбинации трех подходов: анализа лексикографических материалов, анализа ассоциативных словарей и дистрибутивно-статистического анализа на основе корпусов. Первый и второй подходы предполагают ручной сбор и анализ информации, по крайней мере на начальном этапе. Корпусный анализ, наоборот, решает поставленную задачу автоматически с последующим привлечением экспертных знаний. Данная статья посвящена именно этому третьему подходу.

Конечной целью нашего исследования является создание указанного лингвистического ресурса (точнее, трех, по числу языков) в виде тезауруса с количественными характеристиками связанности лексических единиц и примерами употреблений из корпусов.

4. Методология исследования

Методология заявленного исследования — это корпусно-ориентированный анализ парадигматики и синтагматики лексических единиц с использованием дистрибутивно-статистических методов, учитывающий семантические связи разного типа. Материал и инструмент исследования — существующие и специально создаваемые корпусы с лингвистической разметкой и корпусные лингвистические процессоры. При этом в анализ по мере необходимости будут включены и существующие лексикографические ресурсы.

В настоящее время технологическое состояние работы с лингвистическими данными достигло того состояния, когда стала возможной реализация крупномасштабных автоматизированных проектов. В частности, появились большие и сверхбольшие корпусы, появилась возможность формирования своих корпусов, ориентированных на цели исследования, и инструменты для их статистического и контекстно-смыслового анализа. Это позволяет решать задачи на новом уровне, в том числе с привлечением строгого статистического аппарата.

Словари, как правило, отражают в себе два аспекта функционирования языкового знака — синтагматику и парадигматику. Синтагматические связи между лексическими единицами нашли свое отражение, пусть и неполно, в традиционных словарях разного типа. Корпусная лингвистика также научилась «вычислять» разные типы

сочетаемости, которые обычно объединяются под понятием «многословные единицы» (MWE — multiword expressions). Для их автоматического выделения из текста разработаны и разрабатываются различные автоматизированные методы извлечения коллокаций (collocation extraction), распознавания именованных сущностей (named entity recognition) и др.

Сложнее обстоит дело со словарями, которые отражают парадигматику. Под парадигматикой мы здесь 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 представляет 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 исходя в первую очередь из линейной последовательности, 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.

Построение семантического поля — это задача моделирования понятийной подсистемы языка. Ее можно разбить на две части: выявление системы понятий и выявление отношений между ними. Первая задача может решаться «вручную» путем экспликации и формализации профессионального знания, накопленного в системе 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 [Рубашкин, 2012], когда на каждом этапе решается одна задача; в данной работе это выявление множества основных взаимосвязанных понятий вокруг выбранного ядерного элемента (ключевого слова).

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, 1964; Пиотровский, 1975; Караулов, 1981; Войскунский и др., 1983].

Принцип перехода от изучения текстуальных (синтагматических) связей к системным (парадигматическим) лежит в основе различных дистрибутивно-статистических методик [Шайкевич, 1976; 1982; Smrž, Rychlý, 2001; Pekar, 2004]. Считается, что два элемента связаны парадигматически, если оба они текстуально систематически связаны с какими-то третьими элементами. 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 любых статистических распределений, нужны очень большие массивы данных. Таковые появились только с развитием веба и созданием больших корпусов текстов. Одновременно стали появляться и соответствующие программные средства [Kilgarriff et al., 2004; Sharoff, 2006; Сидорова, 2008; Blancafort et al., 2010].

Исследователи обращают внимание на то, что важно также учитывать наличие синтаксической связи между контекстно близкими эле-

ментами текста [Gamallo et al., 2001; Pazienza et al., 2005]. Наш подход предполагает описание сочетаемости с помощью лексико-синтаксических шаблонов (иногда их называют лексико-грамматическими или морфологическими шаблонами). В нашем понимании лексико-синтаксический шаблон — это модель (структурный образец) языковой конструкции, в котором указываются существенные грамматические характеристики множества лексем, которые входят в языковые выражения, принадлежащие данному классу, и синтаксические условия употребления языкового выражения, построенного в соответствии с шаблоном (например, правила согласования морфологических признаков лексем). Дистрибутивно-статистический анализ в нашем исследовании базируется на грамматике лексико-синтаксических шаблонов для русского языка, разработанной М. В. Хохловой для системы Sketch Engine [Хохлова, 2010].

5. Инструменты исследования

Наиболее широко в своей работе мы используем систему Sketch Engine[2], которая представляет собой корпусный менеджер, работающий с морфологически размеченным корпусом. Также мы используем корпусы, работающие под управлением других корпусных менеджеров, но главное, что все они реализуют примерно тот же набор функций, что и Sketch Engine. Может быть, главное преимущество последней заключается в том, что она позволяет создавать свои исследовательские корпусы.

Sketch Engine в числе прочих функций выдает частотные списки лексических единиц, входящих в корпус, и эти частоты, безусловно, характеризуют лексический состав анализируемого корпуса. Далее, есть возможность контрастивного анализа, когда данные исследуемого корпуса сравниваются с нейтральным фоновым. При этом относительные частоты в текстах исследуемого корпуса должны существенно превосходить частоту этих слов в некотором фоновом неспециализированном корпусе. Главная ценность использования этой системы для целей проекта — наличие в ней специальных средств, реализующих методику дистрибутивно-статистического анализа, к которым относятся «Тезаурус» (построение тезауруса, другими

[2] https://old.sketchengine.co.uk/auth/corpora/ (дата обращения: 19.06.2018).

словами — лексико-семантического поля), «Кластеризация» (группировка единиц тезауруса в кластеры — лексико-семантические группы), «Дифференциация» (выявление сходства и разницы в сочетаемости для пар слов) и «Лексические шаблоны» (выявление коллигаций — коллокаций в рамках синтаксических моделей). Все они, разными способами, выявляют парадигматические (т. е. семантические) связи между терминами с количественным указанием силы этой связи.

Тезаурус в системе Sketch Engine (или, как его называют, дистрибутивный тезаурус) позволяет увидеть, какие слова имеют схожую дистрибуцию с заданным словом, что, как правило, является следствием их семантической близости, т. е. фактически этот инструмент формирует семантическое поле из унитермов. Схожесть дистрибуции слов высчитывается статистически на основе меры ассоциации logDice [Rychlý, 2008] и с учетом лексико-синтаксических шаблонов [Kilgarriff, Rychly, 2007; Хохлова, 2010].

6. Материал исследования

Планировалось, в первую очередь, использовать Национальный корпус русского языка, Британский национальный корпус, Чешский национальный корпус (ЧНК) и ряд других. Проведенные эксперименты показали недостатки некоторых корпусов и корпусных менеджеров. Они, эти недостатки, двуплановы. Во-первых, недостаточна частотность слов ядра семантического поля, отсюда вытекает невозможность или малая вероятность выявить периферийную, но специальную лексику. Во-вторых, разноплановость и нередко недостаточные функциональные возможности корпусных менеджеров не позволяют решать задачи в полном объеме и по единой методологии.

В результате было принято решение основываться на близких по функционалу системах Sketch Engine[3] и Corpus.Byu.Edu (корпусный менеджер и система корпусов, разработанные Марком Дэвисом в Университете Бригама Янга)[4], а также на немецкой системе Deutscher Wortschatz / Leipzig Corpora Collection[5]. Причины обращения к ней

[3] https://app.sketchengine.eu/ (дата обращения: 19.06.2018).
[4] http://corpus.byu.edu (дата обращения: 19.06.2018).
[5] http://wortschatz.uni-leipzig.de/de (дата обращения: 19.06.2018).

следующие: функциональность, близкая к двум названным выше, наличие корпусов на разных языках и единообразная технология для всех языков. Вторая характерная особенность принятой стратегии — создание своих репрезентативных корпусов по теме исследования. Под их репрезентативностью мы понимаем достаточный объем и насыщенность специальными текстами по теме проекта.

В начале работы основной упор делался на подбор материала и создание русскоязычных корпусов. Была изучена литература, затрагивающая тему империи в русской культуре, составлен библиографический перечень, на данный момент насчитывающий 217 наименований, и сформирована электронная библиотека, в состав которой вошли 319 файлов по теме проекта, полнотекстовых или фрагментарных, начиная с XVIII по середину XX века.

Основной материал исследования — это созданный нами совместно с М. В. Хохловой на основе этих 319 файлов корпус по теме «империя». Корпус делится на 4 подкорпуса по хронологическому принципу: XVIII век (идентификатор подкорпуса — XVIII), первая половина XIX века (XIX-1), вторая половина XIX века (XIX-2) и XX век (XX). Подкорпусы загружены в систему Sketch Engine. Их суммарный объем составляет 10,25 млн слов, характеристика представлена в приложении. Граничные даты подкорпусов выбраны как своего рода вехи в осознании понятия империи в развитии русской общественной мысли. Жанрово-тематическое наполнение — история, литература, публицистика, философия. Кроме того, был создан еще ряд корпусов, главный из которых составлен на базе текстов из веба по технологии WaCky. Его объем — около 1300 файлов и 24 млн слов (31,5 млн токенов).

Корпусов английского и чешского языков по традиционной технологии мы пока не создавали. Экспериментальные работы ведутся на данных ЧНК и Корпуса современного американского английского языка (COCA — Corpus of Contemporary American English), на англоязычных корпусах Sketch Engine и семейства Aranea Corpora, а также в системе Wortschatz. Однако, поскольку ни один из них не является специализированным с точки зрения нашей темы, были созданы веб-корпусы по нашей тематике на базе чешского интернета (342 млн токенов), английского интернета на основе сайтов с доменом *.uk (103 млн токенов) и на основе английской «Википедии» (136 млн токенов). Из всех веб-корпусов последний является наиболее насыщенным лексикой, относящейся к семантическому полю «империя».

7. Корпусный анализ парадигматических и синтагматических связей и формирование дистрибутивных тезаурусов для понятия «империя»

7.1. Технология исследования

Эта работа велась преимущественно на базе русского и чешского языков. Была разработана и опробована технология формирования семантического поля на основе нескольких подходов (подробно о ней см.: [Zakharov, 2018]).

Суть ее состоит в следующем.

Используются разные подходы к выявлению лексических единиц, предположительно относящихся к семантическому полю «империя», а именно: формирование дистрибутивного тезауруса (можно сказать, семантического поля в чистом виде или мини-поля для заданного термина), формирование списка коллокаций, формирование списка коллигаций — коллокаций с учетом структурных синтаксических формул, формирование статистически значимой лексики методом контрастивного анализа и др.

Эти методы мы реализуем на разных подкорпусах — в описываемом эксперименте это четыре хронологических подкорпуса, упомянутые выше. В дальнейшем планируется создание также жанровых подкорпусов.

Затем данные в каждом из четырех массивов-результатов, а именно лексические единицы, ранжируются по соответствующей для каждого метода оценочной мере, а затем все они объединяются в один массив. При этом каждой лексической единице (термину или словосочетанию) в объединенном массиве присваиваются коэффициенты в зависимости от того, в каком количестве массивов та или иная единица встретилась и какие она там получила ранги. В конечном счете для всего сводного массива по простой формуле вычисляются средний и нормированный ранги, выражающие силу семантической связи соответствующей лексемы с заглавным словом, т. е., можно сказать, коэффициент «пригодности» данной лексемы для данного семантического поля. Эта схема выполняется для всех методов, вычисляющих лексические единицы, отбираемые в качестве кандидатов в семантическое поле.

7.2. Предварительные результаты

Вот некоторые результаты исследования на базе русских корпусов.

Был создан объединенный список терминов по теме «империя», представляющий собой сумму дистрибутивных тезаурусов по данным четырех подкорпусов. Каждый тезаурус (мини-тезаурус) был ограничен объемом в 40 слов. В результате было установлено, что из 160 терминов 79 слов появляются единожды, т. е. лишь в каком-то одном из мини-тезаурусов, при этом распределение по подкорпусам этих уникальных слов следующее: подкорпус XVIII — 32 слова, XIX-1 — 16, XIX-2 — 14, XX — 17.

Оставшиеся вхождения (81) появляются в двух, трех или всех четырех мини-тезаурусах, при этом распределение по подкорпусам следующее: 8 представляли XVIII век, 24 — первую половину XIX века, 26 — вторую половину XIX века и 23 слова относились к XX веку. В итоге разных слов среди них насчиталось 33 (естественно, с учетом повторов получившееся число меньше суммы всех вхождений). Эти 33 слова мы называем ядром семантического поля.

Представляем ядро семантического поля «империя», полученное в результате реализации предложенной технологии; список ранжирован по разным основаниям:

а) по алфавиту: *Англия, государственность, государство, держава, Европа, император, искусство, история, культура, литература, мир, монархия, наука, нация, общество, община, политика, правительство, просвещение, революция, религия, Рим, Россия, союз, страна, традиция, учреждение, философия, Франция, христианство, царство, церковь, цивилизация*;

б) по нормированному рангу: *государство, император, держава, Европа, царство, церковь, Рим, Франция, христианство, монархия, правительство, страна, общество, философия, революция, культура, нация, Россия, литература, государственность, просвещение, религия, мир, искусство, община, политика, история, учреждение, Англия, союз, цивилизация, традиция, наука*;

в) по коэффициенту семантической близости (score): *держава, государство, общество, союз, государственность, нация, император, политика, культура, страна, община, церковь, царство, христианство, религия, мир, просвещение, правитель-*

ство, монархия, Европа, цивилизация, философия, Рим, литература, искусство, учреждение, традиция, Англия, Франция, история, Россия, революция, наука;

г) по относительной частоте (ipm): *Россия, общество, церковь, мир, история, государство, наука, просвещение, правительство, держава, политика, царство, литература, революция, философия, союз, страна, Европа, община, культура, император, цивилизация, искусство, христианство, нация, учреждение, Англия, религия, Рим, Франция, государственность, традиция, монархия.*

Приведем еще результаты формирования списка биграммных коллокаций —кандидатов в семантическое поле «империя» (с упорядочением по мере logDice).

Всего в сумме было выделено 115 биграмм, в подавляющем большинстве это биграммы типа Adj + *империя*, *империя* + Ngen, N + *империи*. Еще одна группа слов — термины из парадигматического ряда, уже выявленные инструментом «Тезаурус». Количественные характеристики следующие: 78 биграмм характерны лишь для одного из подкорпусов, 13 — для двух, 10 — для трех и 4 — для четырех.

Ядро синтагматических коллокаций составляют 24 словосочетания: *Российская империя, Византийская империя, империя германской нации, Восточная империя, Священная империя, падение империи, Австрийская империя, Великая империя, пределы империи, Турецкая империя, столица империи, Западная империя, могущество империи, Оттоманская империя, империя Карла, существование империи, восстановление империи, Латинская империя, область империи, империя Рима, империя Наполеона, разрушать империю, эпоха империи.*

На основе полученных результатов можно отметить, что по разным параметрам понятие «империя» в разные периоды времени в русской культуре имеет разные коннотации. Так, бросается в глаза существенное отличие текстов XVIII века. Это видно по составу лексики: так, из 79 слов тезауруса, «уникальных» только для одного периода, 32 относятся к XVIII веку. Отличие проявляется и в именах собственных, вошедших в периферию поля. И, наверное, можно сформулировать осторожный вывод, что, несмотря на присутствие империи в России в XVIII веке в реальности, сам концепт империи в русской культуре в XVIII веке еще не сложился.

Эксперименты с чешским языком проводились на базе синхронного Национального корпуса чешского языка (ЧНК), где доля специфической лексики, связанной с империей, естественно, намного меньше. Поиски проводились по корпусу SYN2015. Реализовывать технологию, описанную в разделе 7.1 (первоначально поиски по отдельным подкорпусам), не имело смысла, так как остальные подкорпусы ЧНК (SYN2000, SYN2005, SYN2010) представляют фактически один и тот же период развития языка (с конца XX века и до 2015 года). Тем не менее можно сказать, что инструменты Sketch Engine работают одинаково хорошо и на материале чешского языка. Во избежание попадания в результирующее поле периферийной лексики здесь объем выдачи дистрибутивного тезауруса был ограничен числом 30.

Вот несколько примеров для чешского языка для слова *říše* ('империя') (приводится начало списков):

а) по нормированному рангу: *království, civilizace, Británie, Rusko, společenství, vesmír, Řím, impérium*…;

б) по коэффициенту семантической близости (мера score в дистрибутивном тезаурусе): *civilizace, království, země, impérium, Británie, Amerika, armáda, lidstvo, monarchie*…

Начало списка коллокаций выглядит так: *říše: Třetí, Římská, Osmanská, Německá, svatá, vládce říše, Velkomoravská, zánik říše*…

Знающий чешский язык согласится, что все приведенные термины безусловно имеют с «империей» сильную семантическую связь.

Заключение

Мы видим, что использование корпуса текстов и «умных» корпусных инструментов позволяет выявлять в автоматизированном режиме синтагматические и парадигматические связи и создавать более адекватное наполнение терминосистемы. Были получены списки слов и словосочетаний, значительно расширяющие имеющиеся лексикографические пособия. Однако это статистическое расширение иногда получается чрезмерно широким (в периферии сводного дистрибутивного тезауруса мы видим такие слова, как *посол, отечество, воевода, воин* и т. д., которые вряд можно отнести к полю «империя»). Анализ всех полученных списков побуждает нас повторить эксперименты с более жесткими параметрами корпусных инструментов.

Иногда встает также вопрос, правомерно ли включать в поле «империя» авторов, пишущих о ней (Герцен, Киреевский, Тютчев и др.), или акторов — действующих субъектов в империях. Правомерно ли включать в поле «империя» названия народов, населявших империи, и, если да, то каких. Совершенно ясно, что в этих и многих других случаях нужно прибегать к помощи экспертов, как при настройке системы, так и при решении конкретных лингвистических вопросов.

Запланированы и уже начаты подбор текстов и создание параллельных англо-русского, чешско-русского и англо-чешского корпусов. Объемы, по-видимому, будут небольшими из-за проблем с подбором параллельных текстов, но сделать это нам представляется важным, ибо в этих текстах элементы поля окажутся в одной и той же временно́й и исторической парадигме. Также интересно посмотреть, какие слова (и почему) будут превалировать при переводе одного и того же понятия; например, чешское *říše* по-русски может звучать как *империя, королевство, царство, рейх, Германия*. Русское *империя* на чешский переводится как *impérium, říše, císařství, država* и др. То же самое касается и других терминов и других пар языков.

Отдельная работа — формирование поля «империя» для английского языка по той же технологии, которая применялась для русского и чешского языков.

Далее мы планируем построить поле «империя» для трех языков на материале корпусов системы Wortschatz и соотнести с ранее построенными полями.

Но и работа с русскими корпусами не заканчивается. В числе первоочередных задач — создание единого «ядерного» корпуса русских текстов со сбалансированными временны́ми периодами, создание подкорпуса русских текстов после 1917 года и проведение соответствующих экспериментов. Также планируется повторить эксперименты с измененными параметрами инструментов «Тезаурус» и «Лексические шаблоны» (в частности, уменьшить количество терминов, включаемых в дистрибутивный тезаурус, и увеличить размер окна выявления коллокаций). Будет проведена работа по выявлению элементов семантического поля (дистрибутивного тезауруса) для терминов, вошедших в ядро поля «империя», т. е. ставится задача создать тезаурусы (поля) второго уровня и сформировать объединенный список, по возможности в виде семантической сети.

В заключение можно констатировать, что в задаче построения одного небольшого по объему семантического поля «империя» как в капле воды отражаются все особенности лексико-семантической системы языка и возможности и препятствия на пути автоматизации семантических процессов.

Источники

Ахманова О. С. Словарь лингвистических терминов. М.: Советская энциклопедия, 1966.

Караулов Ю. Н., Сорокин Ю. С., Тарасов Е. Ф., Уфимцева Н. В., Черкасова Г. Л. Русский ассоциативный словарь: в 6 кн. М.: Помовский и партнеры; Ин-т рус. языка РАН, 1994–1998.

Литература

Адмони В. Г. Синтаксис современного немецкого языка: система отношений и система построения. Л.: Наука, 1973.

Апресян Ю. Д. Образ человека по данным языка: попытка системного описания // Вопросы языкознания. 1995. № 1. С. 37–67.

Арапов М. В. Некоторые принципы построения словаря типа «тезаурус» // Научно-техническая информация. Серия 2. 1964. № 4. С. 40–46.

Аскольдов С. А. Концепт и слово // Русская словесность: От теории словесности к структуре текста: антология / под общ. ред. В. П. Нерознака. М.: Academia, 1980. С. 267–279.

Бондарко А. В. Функциональная грамматика. Л.: Наука. Ленингр. отделение, 1984.

Васильев Л. М. Современная лингвистическая семантика. М.: Высшая школа, 1990.

Вежбицкая А. Понимание культур через посредство ключевых слов. М.: Языки славянской культуры, 2001.

Войскунский В. Г., Захаров В. П., Мордовченко П. Г., Сороколетова Л. И. О некоторых лексико-семантических проблемах в «бестезаурусных» ИПС // Структурная и прикладная лингвистика: межвуз. сб. Вып. 2 / под ред. В. В. Богданова. Л.: Изд-во ЛГУ, 1983. С. 170–177.

Жинкин Н. И. Речь как проводник информации. М.: Наука, 1982.

Залевская А. А. Проблемы организации внутреннего лексикона человека. Калинин: Калинин. гос. ун-т, 1977.

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

Кобозева И. М. Лингвистическая семантика. М.: Эдиториал УРСС, 2000.

Леонтьев А. А. Языковое сознание и образ мира // Тезисы IX Всесоюз. симпозиума по психолингвистике и теории коммуникации «Языковое сознание». М.: Ин-т языкознания РАН, 1988. С. 105–106.

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

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

Семантико-функциональные поля в лексике и грамматике / отв. ред. В. М. Аринштейн. Л.: ЛГПИ им. А. И. Герцена, 1990.

Сидорова Е. А. Подход к построению предметных словарей по корпусу текстов // Труды междунар. конф. «Корпусная лингвистика — 2008» / отв. ред. А. С. Герд, В. П. Захаров, О. А. Митрофанова. СПб.: Изд-во СПбГУ, 2008. С. 365–372.

Тарасов Е. Ф. Межкультурное общение — новая онтология анализа языкового сознания // Этнокультурная специфика языкового сознания / отв. ред. Н. В. Уфимцева. 2-е изд., испр. и доп. М.: Эйдос, 1996. С. 7–22.

Уфимцева А. А. Теории «семантического поля» и возможности их применения при изучении словарного состава языка // Вопросы теории языка в современной зарубежной лингвистике / отв. ред. Р. М. Будагов, М. М. Гухман. М.: Изд-во АН СССР, 1961. С. 30–63.

Уфимцева Н. В. Языковое сознание: динамика и вариативность. М.: Ин-т языкознания РАН, 2011.

Фрумкина Р. М. Концепт, категория, прототип // Лингвистическая и экстралингвистическая семантика: сб. обзоров. М.: ИНИОН РАН, 1992. С. 33–41.

Хохлова М. В. Разработка грамматического модуля русского языка для специализированной системы обработки корпусных данных // Вестник СПбГУ. Серия 9: Филология, востоковедение, журналистика. 2010. Вып. 2. С. 162–169.

Шайкевич А. Я. Распределение слов в тексте и выделение семантических полей // Иностранные языки в высшей школе. Вып. 2 / гл. ред. Н. С. Чемоданов. М.: Русвузиздат, 1963. С. 14–26.

Шайкевич А. Я. Дистрибутивно-статистический анализ в семантике // Принципы и методы семантических исследований / отв. ред. В. Н. Ярцева. М.: Наука, 1976. С. 353–378.

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

Щур Г. С. Теория поля в лингвистике. М.: Наука, 1974.

Blancafort H., Daille B., Gornostay T., Heid U., Méchoulam C., Sharoff S. TTC: Terminology Extraction, Translation Tools and Comparable Corpora // Proceedings of the 14th Euralex International Congress / ed. by A. Dykstra, T. Schoonheim. Leeuwarden: Fryske Akademy, 2010. P. 263–268.

Gamallo P., Gasperin C., Augustini A., Lopes G. P. Syntactic-based Methods for Measuring Word Similarity // Text, Speech and Dialogue: 4th International Conference (TSD-2001) / ed. by V. Matoušek, P. Mautner, R. Mouček, K. Tauser. Heidelberg: Springer, 2001. P. 116–125. (LNCS (LNAI). Vol. 2166).

Ipsen G. Der alte Orient und die Indogermanen // Stand und Aufgaben der Sprachwissenschaft: Festschrift für W. Streiberg. Heidelberg: C. Winter, 1924. S. 30–45.

Kilgarriff A., Rychly 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. Czech Republic, June 2007 / conference chair S. Ananiadou. Prague: Association for Computational Linguistics, 2007. P. 41–44.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress / ed. by G. Williams, S. Vessier. Lorient: Universite de Bretagne-Sud, 2004. P. 105–116.

Kiss G., Armstrong C., Milroy R. The Associative Thesaurus of English. Edinburgh: Edinburgh University Press, 1972.

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: Springer, 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 / ed. by M. Zock, P. S. Dizier. Geneva: University of Geneva, 2004. P. 15–21.

Rychlý P. A Lexicographer-friendly Association Score // RASLAN 2008: Recent Advances in Slavonic Natural Language Processing / ed. by P. Sojka, A. Horák. Brno: Masaryk University Press, 2008. P. 6–9.

Sharoff S. Open-source Corpora: Using the Net to Fish for Linguistic Data // International Journal of Corpus Linguistics. 2006. Vol. 11. No. 4. P. 435–462.

Smrž P., Rychlý P. Finding Semantically Related Words in Large Corpora // Text, Speech and Dialogue: 4th International Conference (TSD-2001) / ed. by V. Matoušek, P. Mautner, R. Mouček, K. Tauser. Heidelberg: Springer, 2001. P. 108–115. (LNCS (LNAI). Vol. 2166).

Zakharov V. The Distributive and Statistical Analysis as a Tool to Automate the Formation of Semantic Fields (on the Example of the Linguocultural Concept of "Empire") // CMLS 2018: Computational Models in Language and Speech: Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) Co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL 2018). Kazan, Russia, November 1, 2018 / ed. by A. Elizarov, N. Loukachevitch. [Kazan: s. p., 2018]. P. 144–162. http://ceur-ws.org/Vol-2303/ (дата обращения: 30.10.2018).

Sources

Akhmanova O. S. 1966. *Dictionary of Linguistic Terms*. Moscow, Sovetskaia entsiklopediia Publ. (In Russ.)

Karaulov Iu. N., Sorokin Iu. S., Tarasov E. F., Ufimtseva N. B., Cherkasova T. L. 1994–1998. *The Associative Thesaurus of Russian*, in 6 books. Moscow, Pomovskii i partnery Publ.; In-t rus. iazyka RAN Publ. (In Russ.)

References

Admoni V. G. 1973. *Syntax of Modern German: the System of the Relations and the System of Construction*. Leningrad, Nauka Publ. (In Russ.)

Apresian Iu. D. 1995. The Image of a Person According to the Language: An Attempt of the System Description. *Voprosy iazykoznaniia*, no. 1, pp. 37–67. (In Russ.)

Arapov M. V. 1964. Some Principles of Creation of the "Thesaurus" Dictionary. *Nauchno-tekhnicheskaia informatsiia. Seriia 2*, no. 4, pp. 40–46. (In Russ.)

Askol'dov S. A. 1980. Concept and Word. *Russkaia slovesnost'. Ot teorii slovesnosti k strukture teksta. Antologiia*, V. P. Neroznak (ed.). Moscow, Academia Publ., pp. 267–279. (In Russ.)

Blancafort H., Daille B., Gornostay T., Heid U., Méchoulam C., Sharoff S. 2010. TTC: Terminology Extraction, Translation Tools and Comparable Corpora. *Proceedings of the 14th Euralex International Congress*, A. Dykstra, T. Schoonheim (eds.). Leeuwarden, Fryske Akademy, pp. 263–268.

Bondarko A. V. 1984. *Functional Grammar*. Leningrad, Nauka. Leningr. otdelenie Publ. (In Russ.)

Frumkina R. M. 1992. Concept, Category, Prototype. *Lingvisticheskaia i ekstralingvisticheskaia semantika. Sb. obzorov*. Moscow, INION RAN, pp. 33–41. (In Russ.)

Gamallo P., Gasperin C., Augustini A., Lopes G. P. 2001. Syntactic-based Methods for Measuring Word Similarity. *Text, Speech and Dialogue. 4th International Conference TSD-2001*, V. Matoušek, P. Mautner, R. Mouček, K. Tauser (eds.). Heidelberg, Springer, pp. 116–125. (LNCS (LNAI). Vol. 2166).

Ipsen G. 1924. Der alte Orient und die Indogermanen. *Stand und Aufgaben der Sprachwissenschaft. Festschrift für W. Streiberg*. Heidelberg, C. Winter, SS. 30–45.

Karaulov Iu. N. 1981. *Linguistic Designing and Thesaurus of the Literary Language*. Moscow, Nauka Publ. (In Russ.)

Khokhlova M. V. 2010. Development of the Grammatical Module of Russian for the Specialized System of Processing of Corpus Data. *Vestnik SPbGU. Seriia 9. Filologiia, vostokovedenie, zurnalistika*, issue 2, pp. 162–169. (In Russ.)

Kilgarriff A., Rychly P. 2007. 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.*

Czech Republic, June 2007, S. Ananiadou (conference chair). Prague, Association for Computational Linguistics, pp. 41–44.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. 2004. The Sketch Engine. *Proceedings of the XI^th^ Euralex International Congress*, G. Williams, S. Vessier (eds.). Lorient, Universite de Bretagne-Sud, pp. 105–116.

Kiss G., Armstrong C., Milroy R. 1972. *The Associative Thesaurus of English*. Edinburgh, Edinburgh University Press.

Kobozeva I. M. 2000. *Linguistic Semantics*. Moscow, Editorial URSS Publ. (In Russ.)

Leont'ev A. A. 1988. Linguistic Consciousness and Image of the World. *Tezisy IX Vsesoiuz. simpoziuma po psikholingvistike i teorii kommunikatsii «Iazykovoe soznanie»*. Moscow, In-t iazykoznaniia RAN Publ., pp. 105–106. (In Russ.)

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

Pekar V. 2004. Linguistic Preprocessing for Distributional Classification of Words. *Proceedings of the COLING-04 Workshop on Enhancing and Using Electronic Dictionaries*, M. Zock, P. S. Dizier (eds.). Geneva, University of Geneva, pp. 15–21.

Piotrovsky R. G. 1975. *Text, Computer, Human*. Leningrad, Nauka Publ. (In Russ.)

Rubashkin V. Sh. 2012. *Ontologic Semantics. Knowledge. Ontologies. Ontologically Oriented Methods of Information Analysis of Texts.* Moscow, Fizmatlit Publ. (In Russ.)

Rychlý P. 2008. A Lexicographer-friendly Association Score. *RASLAN 2008. Recent Advances in Slavonic Natural Language Processing*, P. Sojka, A. Horák (eds.). Brno, Masaryk University Press, pp. 6–9.

Semantic and Functional Fields in Vocabulary and Grammar 1988, V. M. Arinshtein (ed.). Leningrad, LGPI im. A. I. Gertsena Publ. (In Russ.)

Shaikevich A. Ia. 1963. Distribution of Words in the Text and Allocation of Semantic Fields. *Inostrannye iazyki v vysshei shkole*, issue 2, N. S. Chemodanov (ed. in chief). Moscow, Rusvuzizdat Publ., pp. 14–26. (In Russ.)

Shaikevich A. Ia. 1976. The Distributive and Statistical Analysis in Semantics. Printsipy i metody semanticheskikh issledovanii, V. N. Iartsev (ed.). Moscow, Nauka Publ., pp. 353–378. (In Russ.)

Shaikevich A. Ia. 1982. *Distributive and Statistical Analysis of Texts. PhD Thesis.* Leningrad. (In Russ.)

Sharoff S. 2006. Open-source Corpora. Using the Net to Fish for Linguistic Data. *International Journal of Corpus Linguistics*, vol. 11, no. 4, pp. 435–462.

Shchur G. S. 1974. *Field Theory in Linguistics*. Moscow, Nauka Publ. (In Russ.)

Sidorova E. A. 2008. The Approach to the Construction of Subject Dictionaries by the Body of Texts. *Trudy mezhdunar. konf. «Korpusnaia lingvistika — 2008»*,

A. S. Gerd, V. P. Zakharov, O. A. Mitrofanova (eds.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 365–372. (In Russ.)

Smrž P., Rychlý P. 2001. Finding Semantically Related Words in Large Corpora. *Text, Speech and Dialogue. 4th International Conference (TSD–2001)*, V. Matoušek, P. Mautner, R. Mouček, K. Tauser (eds.). Heidelberg, Springer, pp. 108–115. (LNCS (LNAI). Vol. 2166).

Tarasov E. F. 1996. Intercultural Communication Is a New Ontology of Analysis of Linguistic Consciousness. *Etnokul'turnaia spetsifika iazykovogo soznaniia*, N. B. Ufimtseva (ed.), 2nd ed., rev. and exp. Moscow, Eidos Publ., pp. 7–22. (In Russ.)

Ufimtseva A. A. 1961. The Theory of "Semantic Field" and the Possibility of Their Application in the Study of the Vocabulary of the Language. *Voprosy teorii iazyka v sovremennoi zarubezhnoi lingvistike*, R. M. Budagov, M. M. Guhman (eds.). Moscow, Izd-vo AN SSSR Publ., pp. 30–63. (In Russ.)

Ufimtseva N. V. 2011. *Linguistic Consciousness: Dynamics and Variability*. Moscow, In-t iazykoznaniia RAN Publ. (In Russ.)

Vasil'ev L. M. 1990. *Modern Linguistic Semantics*. Moscow, Vysshaia shkola Publ. (In Russ.)

Vežbicka A. 2001. Understanding of Cultures Through Keywords. Rus. Ed. Moscow, Iazyki slavanskoi kul'tury. (In Russ.)

Voiskunskii V. G., Zakharov V. P., Mordovchenko P. G., Sorokoletova L. I. 1983. About Some Lexico-semantic Problems in "Thesauruless" IRS. *Strukturnaia i prikladnaia lingvistika. Mezhvuz. sb.*, issue 2, V. V. Bogdanov (ed.). Saint Petersburg, Izd-vo LGU Publ., pp. 170–177. (In Russ.)

Zakharov V. 2018. The Distributive and Statistical Analysis as a Tool to Automate the Formation of Semantic Fields (On the Example of the Linguocultural Concept of "Empire"). *CMLS 2018. Computational Models in Language and Speech. Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018) Co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics (TEL 2018). Kazan, Russia, November 1, 2018*, A. Elizarov, N. Loukachevitch (eds.). [Kazan, s. p., 2018], pp. 144–162. http://ceur-ws.org/Vol-2303/ (accessed date: 30.10.2018).

Zalevskaia A. A. 1977. *Human Internal Lexicon Organization.* Kalinin, Kalinin. gos. un-t Publ. (In Russ.)

Zhinkin N. I. 1982. *Speech as an Information Conductor*. Moscow, Nauka Publ. (In Russ.)

ПРИЛОЖЕНИЕ

Справочные данные о корпусах русского языка, созданных по тематике «империя» на сайте системы Sketch Engine

XVIII

Counts	
Tokens	1,411,991
Words	1,142,151
Sentences	46,829
Documents	44

General info	
Language	Russian
Encoding	UTF-8
Compiled	03/24/2014 12:08:15
Tagset	Description
Word sketch grammar	Definition

Lexicon sizes	
word	117,120
tag	528
lempos	59,600
lemma	58,430
lemma_lc	58,306
lc	116,929

Tags legend	
noun	N.*
verb	V.*
adjective	A.*
pronoun	P.*
adverb	R.*
adposition	S.*
conjunction	C.*
numeral	M.*
particle	Q.*

Lempos suffixes	
noun	-n
verb	-v
adjective	-a
pronoun	-p
adverb	-r
adposition	-s
conjunction	-c
numeral	-m
particle	-q

XIX-1

Counts	
Tokens	2,852,650
Words	1,603,424
Sentences	72,164
Documents	58

General info	
Language	Russian
Encoding	UTF-8
Compiled	03/24/2014 12:08:08
Tagset	Description
Word sketch grammar	Definition

Lexicon sizes	
word	146,050
tag	559
lempos	66,926
lemma	65,725
lemma_lc	65,296
lc	145,530

Tags legend	
noun	N.*
verb	V.*
adjective	A.*
pronoun	P.*
adverb	R.*
adposition	S.*
conjunction	C.*
numeral	M.*
particle	Q.*

Lempos suffixes	
noun	-n
verb	-v
adjective	-a
pronoun	-p
adverb	-r
adposition	-s
conjunction	-c
numeral	-m
particle	-q

XIX-2

Counts	
Tokens	4,701,567
Words	3,691,662
Sentences	166,420
Documents	75

General info	
Language	Russian
Encoding	UTF-8
Compiled	03/24/2014 12:08:13
Tagset	Description
Word sketch grammar	Definition

Lexicon sizes	
word	264,638
tag	579
lempos	136,967
lemma	134,361
lemma_lc	133,465
lc	263,619

Tags legend	
noun	N.*
verb	V.*
adjective	A.*
pronoun	P.*
adverb	R.*
adposition	S.*
conjunction	C.*
numeral	M.*
particle	Q.*

Lempos suffixes	
noun	-n
verb	-v
adjective	-a
pronoun	-p
adverb	-r
adposition	-s
conjunction	-c
numeral	-m
particle	-q

XX

Counts	
Tokens	3,181,077
Words	2,474,143
Sentences	144,424
Documents	79

General info	
Language	Russian
Encoding	UTF-8
Compiled	03/24/2014 12:08:18
Tagset	Description
Word sketch grammar	Definition

Lexicon sizes	
word	232,274
tag	574
lempos	107,730
lemma	105,851
lemma_lc	105,528
lc	231,873

Tags legend	
noun	N.*
verb	V.*
adjective	A.*
pronoun	P.*
adverb	R.*
adposition	S.*
conjunction	C.*
numeral	M.*
particle	Q.*

Lempos suffixes	
noun	-n
verb	-v
adjective	-a
pronoun	-p
adverb	-r
adposition	-s
conjunction	-c
numeral	-m
particle	-q

И. С. Николаев

МОДЕЛИРОВАНИЕ ТОПОНИМИЧЕСКИХ СИСТЕМ: МЕТОДЫ И ПРЕДЕЛЫ ИХ ВОЗМОЖНОСТЕЙ

Аннотация. В статье обсуждаются общие теоретические проблемы моделирования топонимии, рассматриваются методы, позволяющие полно и достоверно ее описать, а также пределы возможностей этих методов. В основу моделирования положено понятие системности топонимии, которая выявляется в процессе изучения географии данной территории, исторических и демографических факторов, а также языковых особенностей топонимов. Анализируя небольшой фрагмент топонимической системы Ингерманландии, мы демонстрируем, как могут быть выявлены ее основные характеристики и как элементы системы связаны между собой и с топонимической системой всего региона в целом. Пределы возможностей методов топонимических исследований, прежде всего, зависят от качества собранного первичного топонимического материала. Однако можно расширить представление о топонимической системе, используя статистико-комбинаторные и синтетические методы. Важным фактором является также использование данных о топонимических системах сопредельных территорий.

Ключевые слова. Топонимия, топонимика, топонимическая система, моделирование, методы топонимических исследований, Ингерманландия.

Ilya S. Nikolaev

MODELING OF TOPONYMIC SYSTEMS: METHODS AND THEIR LIMITS

Abstract. In the article general theoretical problems of toponymy modeling are discussed, methods for full and reliable description of toponymy, as well as limits of these methods are considered. As a basis of modelling we underline a notion of toponymy systematics, which is revealed in the process of the study of geography of the given territory, of historical and demographic factors: as well as of linguistic peculiarities of toponyms. Analysing a small fragment of toponymic system of Ingermanland, we demonstrate how its main characteristics can be revealed and how the elements of the system are connected with each other and with the toponymic system of the whole region in general. The limits of possibilities of toponymic studies first of all depend upon the quality of primary toponymic data. However we can widen

the representation of the toponymic system, using statistical and combinatorial methods as well as synthetic methods. One more important factor is the use of data about toponymic systems of neighbourgh territories.

Keywords. Toponymy, toponymics, toponymic system, modeling, toponymic study methods, Ingermanland.

В этой статье мы рассмотрим одну из наиболее актуальных проблем в современной топонимике — проблему моделирования топонимических систем. Описание топонимов той или иной местности независимо от целей такого описания приводит к появлению некоторого представления о том, как эти топонимы взаимосвязаны, как они появились и видоизменялись. Такое представление в общем виде можно назвать моделью топонимической системы. Далее мы рассмотрим существенные принципы, которые необходимо учесть при моделировании топонимии, и методы, позволяющие наиболее полно и достоверно описать топонимическую систему.

Научное изучение географических названий (топонимии) основывается на понятии системности. Топонимия, которая складывается на данной территории в ходе исторических, политических, социальных, демографических, культурных и языковых процессов, отражает освоение различными этносами конкретных географических ареалов в определенные исторические эпохи и образует систему. В системе географических названий все эти процессы определенным образом отражаются на языковом облике топонимов. Системность топонимии в конечном итоге предполагает, что в основе любой топонимической системы лежат некоторые общие принципы топонимообразования, действующие или действовавшие в определенное время на данной территории, на которые местные жители или колонизаторы опираются для именования конкретных географических объектов.

Задача топонимики как науки — выявить эти принципы, исходя из анализа устройства современных или исторических топонимических систем. Иначе говоря, если имеется некий топонимический текст, то есть возможность установить язык, на котором он написан, т. е. «открыть» его грамматику и словарь, а именно установить набор единиц, лежащий в его основе, и правила их сочетания.

На практике, однако, все выглядит намного сложнее. Поскольку принципы топонимообразования базируются на большом числе разнообразных факторов, не всегда возможно выявить их все и сразу. Тем не менее многолетний опыт топонимической работы показывает,

что ничего случайного в топонимике не бывает и что рано или поздно находится объяснение даже, на первый взгляд, случайным фактам.

Моделирование топонимических систем позволяет на основе известных факторов, участвующих в топонимообразовании, объяснять происхождение, функционирование и развитие географических названий. В ходе моделирования прежде всего описывается совокупность географических объектов, подлежащих именованию. Это могут быть, например, только населенные пункты или только водные объекты на определенной территории. Но могут описываться и все географические объекты, крупные и мелкие, регионально и локально значимые.

Выбор объектов приводит, во-первых, к набору географических терминов, которые используются для типизации этих объектов, в который может входить как общеязыковая географическая, так и местная или диалектная терминология.

Географическая терминология позволяет определить некоторый базовый набор апеллятивов — имен нарицательных, лежащих в основе имен собственных.

Далее выбранная совокупность географических объектов позволяет перейти к выявлению географических особенностей этих объектов, которые являются значимыми для людей в процессе хозяйственного освоения ландшафта и которые тоже входят в состав топонимов либо как апеллятивы, либо как определения в сложных названиях.

Одновременно определяются языки и диалекты, которые использовались в образовании топонимов. Для этого используются как собственно лингвистическая информация на основе языкового облика названий, так и данные истории и этнографии.

Исторические сведения, а именно данные летописей, исторических карт, списков населенных мест и другая архивная информация, используются на всех этапах моделирования топонимической системы для подтверждения гипотез о происхождении и развития топонимии.

Когда установлено, какие этносы и языки участвовали в топонимообразовании, становится актуальным установление прозвищ, личных и родовых имен, культурных и хозяйственных реалий, которые могли использоваться для создания географических названий.

Все полученные сведения сопоставляются с данными географических карт, в том числе и исторических, и, по возможности, проверя-

ются на местности для выявления особенностей ландшафта и характеристик географических объектов, которые могли иметь существенное значение для внутренней формы имени собственного.

На этапе структурного моделирования топонимии исследуются топонимические модели — наиболее продуктивные в языках и диалектах образцы, по которым создаются географические названия. В топонимах выделяют топоосновы и топоформанты — главные составляющие, которые представляют собой наиболее частотные единицы, используемые носителями рассматриваемых языков и диалектов для создания топонимов по существующим образцам.

Затем может следовать этап картографирования ареалов распространения топооснов и топоформантов, а также топонимических моделей.

Рассматривая этапы моделирования топонимических систем, мы увидели, что исследование топонимики представляет собой довольно сложный процесс, требующий привлечения данных из разных источников на всех стадиях работы. Топонимические исследования в этом аспекте очень похожи на процесс описания структуры неизвестного языка, когда данных недостаточно и возможности работы с информантами ограничены.

При изучении топонимической системы, как было продемонстрировано, мы используем разнообразные методы, и эти методы не только лингвистические, но и географические, исторические, этнографические. Кроме того, говоря о технической стороне топонимических исследований, необходимо упомянуть методы полевой лингвистики, лингвистической географии, лексикографии, информационного поиска и картографии.

Для полного и достоверного моделирования топонимической системы, таким образом, необходимо не просто пройти все указанные этапы, но пройти их неоднократно, с каждым новым разом уточняя и дополняя описание.

Необходимо сказать еще об одном аспекте в моделировании, связанном с использованием аналитических и синтетических моделей. Традиционные топонимические исследования используют аналитическую методику: на основании анализа топонимических материалов делается вывод о языках, топонимических моделях, топоосновах, апеллятивах, типах топонимов и о существенных особенностях географических объектов. Это методика от общего к частному.

Есть и другая возможность — синтетическая методика, от частного к общему. Одним из примеров использования такого подхода является метод анализа лингвистической структуры субстратных топонимов и их этимологизации на основе моделирования компонентов топонимических систем, предложенный А. К. Матвеевым [Матвеев, 2006]. Этот метод предполагает конструирование потенциально возможных субстратных топонимов на основе апеллятивов исходного языка с целью последующей этимологизации изучаемых топонимов. Для исследования топонимической системы Ингерманландии мы предложили расширить этот подход статистико-комбинаторными методами анализа для количественной оценки полученных результатов.

В качестве примера моделирования топонимической системы рассмотрим фрагмент топонимической системы муниципального образования «Куземкинское сельское поселение» Кингисеппского района Ленинградской области (см. таблицу) по материалам топонимической картотеки кафедры математической лингвистики СПбГУ.

Историко-демографический аспект. В XIX веке деревни этого района относились к Наровской волости с правлением в деревне Венекюля. В них проживали преимущественно ижорцы, водь, эстонцы и финны. В начале XX века волость была преобразована в Куземкинскую волость с центром в деревне Большое Куземкино. Часть волости в 1920 году по условиям Тартуского договора отошла Эстонии, а после 1940 года волость снова воссоединилась. После Второй мировой войны большая часть местного населения не смогла вернуться в родные места. В 1946 году был организован совхоз «Ударник-Ропша», в который приехали на работу жители других областей России. Состав населения изменился: сократилось количество коренных народов и увеличилось количество русских. Число постоянных жителей в настоящее время составляет 1366 человек.

Географический аспект. Территория Куземкинского муниципального образования находится на побережье Финского залива и по берегам рек Луга и Россонь, которая связывает Лугу с рекой Наровой и Финским заливом. Река Мертвица соединяет Лугу и Россонь. Ближайшие озера находятся на севере на Курголовском полуострове. Имеется некоторое количество небольших болот и ручьев. Местность преимущественно ровная, лесистая, встречаются холмы. С юга на север проходит шоссейная дорога от Таллинского шоссе до порта Усть-Луга.

Таблица. **Топонимическая система деревень Кирьямо и Струппово**

Объект	Топоним		Топооснова, детерминанты	Перевод
	Русский	Прибалтийско-финский		
Деревня	*Кирьямо*	*Кирьямо*	*kirjamo*	*пестрый*
Лес	*Кирьямский*	—	*кирьям*	—
Деревня	*Хамолово*	*Хамоланкюля*	*hamala*	*конец*
Озеро	*Хамоловское*	*Хамолонярви*	*hamala*	*конец*
Покос	*Корпи*	*Корпи*	*korpi*	*пустошь*
Озеро	*Липовское*	*Веняярви*	*venä*	*русский*
Холм	—	*Киркkомяки*	*kirkko*	*церковь*
Хутор	*Пустоши*	—	*пустошь*	—
Болото	*Большое*	—	*большой*	—
Болото	—	*Мустаметсянсуо*	*musta metsä*	*черный лес*
Озеро	*Белое*	*Валкиярви*	*valki*	*белый*
Лес	*Белоозерский*	—	*белое озеро*	—
Болото	—	*Киркkосуо*	*kirkko*	*церковь*
Ручей	—	*Рянни*	*ränni*	*желоб*
Заводь	—	*Алая*	*alaja*	*нижний*
Покос	—	*Санникко*	*sannikko*	*папоротник*
Холм	—	*Меримяки*	*meri*	*море*
Деревня	*Струппово*	*Струуппа*	*struuppa*	*?*
Дорога	*Струпповская*	—	*струппов*	—
Покос	*Струпповский*	—	*струппов*	—
Холмы	*Струпповские*	—	*струппов*	—

Окончание табл.

Объект	Топоним		Топооснова, детерминанты	Перевод
	Русский	Прибалтийско-финский		
Лес	*Струпповский*	—	*струппов*	—
Ручей	*Глубокий*	—	*глубокий*	—
Болото	*Домрыбаковское*	—	*дом рыбаков*	—
Место	*Лесопильный завод*	—	*лесопильный завод*	—
Речка	*Заводская*	—	*завод*	—
Поляна	*Юркин хутор*	—	*юркин*	—
Водоем	*Голубка*	—	*голубка*	—
Остров	*Камышовый*	—	*камыш*	—
Бухта	*Якорь*	—	*якорь*	—
Место	*Липа*	—	*липа*	—

Исходя из этого, мы можем установить базовую систему географических объектов: деревня, хутор, покос, дорога, холм, лес, озеро, болото, река, ручей и др.

Лингвистический аспект. Русский язык является доминирующим, многие географические названия получили русские эквиваленты, а некоторые новые русские топонимы не переводились на местные прибалтийско-финские языки. В муниципальном образовании до сих пор проживают носители финского и ижорского языков старшего поколения.

Базовая система языков: прибалтийско-финские (ижорский, финский) и русский.

Топонимическую систему данного района мы моделируем, выделяя базовые географические объекты (первый столбец таблицы), основные языки топонимообразования (второй и третий столбцы), основные топоноосновы и детерминанты (четвертый столбец). Для

прибалтийско-финских топооснов и детерминантов предусмотрен перевод на русский язык (пятый столбец).

Как видно из таблицы, большая часть базовых географических объектов имеет прибалтийско-финские названия с дублетами на русском языке. Некоторые микротопонимы были записаны только на ижорском или финском языках. Часть объектов имеет только русские названия (нижняя часть таблицы). Это прежде всего новые микротопонимы.

Основные населенные пункты (*Кирьямо, Хамолово, Струппово*) становятся основой для вторичных топонимов (*Кирьямский лес, Хамоловское озеро, Струпповская дорога*). Топоосновы отражают апеллятивы, характеризующие географические особенности данного района (*море, пустошь, лес, папоротник, церковь*). Детерминанты характерны для топонимии Северо-Запада России (*русский, белый, черный, большой, нижний*).

Проанализировав небольшой фрагмент данной топонимической системы, мы уже можем выделить те базовые характеристики, которые могут быть положены в основу синтетической модели топонимической системы Ингерманландии.

В заключение можно сказать, что пределы возможностей рассмотренных методов исследования зависят, главным образом, от количества и качества топонимического материала. Очевидно, что такое большое количество факторов, которое необходимо учесть при моделировании топонимической системы, довольно сложно учесть и проанализировать исчерпывающим образом, особенно для больших регионов. Обычно наиболее успешно моделируются топонимические системы небольших населенных пунктов и административных районов (в основном микротопонимия). Достаточно успешно удается моделировать и некоторые фрагменты топонимических систем крупных ареалов, например гидронимию [Агеева, 1989] или ойконимию. Пределы возможностей топонимических методов исследования прежде всего зависят от объема, полноты и качества топонимических данных, собранных в топонимических экспедициях и добытых в ходе архивных исследований, которые, в свою очередь, определяются наличием достаточно постоянного коллектива научных исследователей и долговременным характером топонимических исследований. Например, достаточно серьезные успехи в моделировании топонимических систем Ленинградской области

и Республики Карелия стали возможны благодаря коллективам исследователей кафедры математической лингвистики СПбГУ под руководством проф. А. С. Герда [Герд и др., 2012; Nikolaev, Stolyarov, 2014] и Карельского научного центра РАН под руководством проф. И. И. Муллонен [Муллонен, 2002]. Моделирование субстратной топонимии Русского Севера проводилось под руководством проф. А. А. Матвеева в Уральском университете (Екатеринбург) [Матвеев, 2006]. Другим важным фактором является наличие и доступность географических и исторических архивных материалов, что зависит от степени изученности истории и географии данного региона. Еще одним фактором является количество используемых методов, их согласованность и интеграция. Неоднократно отмечалось, что только один метод исследования не дает необходимой глубины и качества исследования. Наконец, одним из главных факторов оказывается широта охвата территорий в топонимических исследованиях: чем больше ареал исследований, тем больше закономерностей удается выявить как в синхронии, так и в диахронии. Возможности топонимических методов, таким образом, достаточно велики, и наиболее успешно их удается продемонстрировать в комплексных долговременных топонимических исследованиях больших топонимических ареалов, проводимых большими научными коллективами.

Литература

Агеева Р. А. Гидронимия Русского Северо-Запада как источник культурно-исторической информации. М.: Наука, 1989.

Герд А. С., Дмитриев А. В., Николаев И. С., Столяров Д. А. Научно-образовательный веб-ресурс «Топонимия Ингерманландии (Ленинградская область)»: перспективы исследования // Структурная и прикладная лингвистика: межвуз. сб. Вып. 9 / под ред. А. С. Герда. СПб.: Изд-во СПбГУ, 2012. С. 148–158.

Матвеев А. К. Ономатология. М.: Наука, 2006.

Муллонен И. И. Топонимия Присвирья: Проблемы этноязыкового контактирования. Петрозаводск: Изд-во Петрозаводск. гос. ун-та, 2002.

Nikolaev I., Stolyarov D. Linguistic Information System of Multicultural Russian-Fennic Region of Ingermanland // SGEM 2014: Proceedings of International Multidisciplinary Scientific Conferences on Social Sciences and Arts. Albena: SGEM, 2014. P. 131–137.

References

Ageeva R. A. 1989. *Hydronymics of the Russian North-West as a Source of Cultural and Historical Information*. Moscow, Nauka Publ. (In Russ.)

Gerd A. S., Dmitriev A. V., Nikolaev I. S., Stoliarov D. A. 2012. Research and Education Web-resource "Toponymy of Ingermanland (Leningrad region)". Perspectives of Investigation. *Strukturnaia i prikladnaia lingvistika. Mezhvuz. sb*., issue 9, A. S. Gerd (ed.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 148–158. (In Russ.)

Matveev A. K. 2006. *Onomatology*. Moscow, Nauka Publ. (In Russ.)

Mullonen I. I. 2002. *Toponymy of Prisvir'ye. Problems of Ethno- and Language Contacts*. Petrozavodsk, Izd-vo Petrozavodsk. gos. un-ta Publ. (In Russ.)

Nikolaev I., Stolyarov D. 2014. Linguistic Information System of Multicultural Russian-Fennic Region of Ingermanland. *SGEM 2014. Proceedings of International Multidisciplinary Scientific Conferences on Social Sciences and Arts)*. Albena, SGEM, pp. 131–137. (In Russ.)

Д. Б. Тискин

ЕЩЕ О РАЗДЕЛЕНИИ СЕМАНТИЧЕСКОГО ТРУДА*

Аннотация. Статья посвящена проблеме выбора, стоящей перед лингвистической теорией при анализе конкретных явлений в семантике естественных языков: приписать тот или иной семантический эффект значению некоторой лексической единицы или синтаксической конструкции или же рассматривать его как производное «наивной онтологии» — модели мира, служащей основой интерпретации языковых выражений. Основные примеры, рассматриваемые в статье, — отношение совокупного сходства между возможными мирами как способ решения проблем в семантике условных высказываний у Д. Льюиса; представление об индивидах действительного мира как онтологически приоритетных по отношению к некоторым лишь возможным индивидам у У. Зауэрланда; структура шкал, включая шкалу «логической силы» высказываний в классической и неклассических логиках.

Ключевые слова. Семантика, наивная онтология, возможные миры, грамматикализованные импликатуры, шкалы.

Daniil B. Tiskin

SOME CONSIDERATIONS REGARDING THE DIVISION OF LABOUR IN SEMANTICS

Abstract. The paper points out the two choice options potentially available to a semanticist dealing with the meanings of a particular class of expressions within a given language. The choice is between attributing the pertinent semantic effect to a device hidden in the denotation of a lexical item or a syntactic configuration, on the one hand, and tracing it to the structures within the "naïve ontology" underlying the language as the source of any possible denotations of its expressions. The crucial case studies include David Lewis's overall similarity ordering on possible worlds and its role in the semantics of conditionals; Uli Sauerland's ontological priority of actual individuals w. r. t. individuals in agents' doxastic alternatives; and the structure of scales, in particular the structure formed by the relations of logical consequence in classical and various non-classical logics.

Keywords. Semantics, naïve ontology, possible worlds, grammaticalised implicatures, scales.

---

* Исследование выполнено при поддержке РФФИ, проект № 18-011-00895.

Введение

При построении формальной теории языка приходится так или иначе распределять объяснительную нагрузку между различными модулями теории. Здесь возможны такие вопросы, как «синтаксис или семантика?» (например, при выборе между синтаксически не мотивированными нулями наподобие нулевых артиклей и чисто семантическими операциями сдвига типа описанных Б. Парти [Partee, 1987]), «синтаксис или лексикон?» (например, при выборе между синтаксически не мотивированными нулями и лексической омонимией), «семантика или прагматика?» (при выборе между неартикулированными компонентами (unarticulated constituents) [Recanati, 2002] и прямым влиянием контекста, а также между прагматическими и грамматикализованными импликатурами). Решения, принимаемые конкретными учеными по этим вопросам, могут быть более или менее принципиальными или приуроченными к конкретному случаю, однако, как кажется, сама возможность выбора во всех этих случаях осознаётся относительно легко — может быть, потому, что любое решение оставляет теоретика в пределах проблемного поля лингвистики (при достаточно широком ее понимании, включающем, по крайней мере, исследование стратегий коммуникации). В последнем случае различие и вовсе носит до известной степени терминологический (а потому номинальный) характер.

Настоящая статья посвящена еще одному вопросу такого рода, самая постановка которого, как мы полагаем, может показаться достаточно новой (но ср.: [Bach, 1986, p. 575]), хотя нам и предстоит рассмотреть несколько примеров его практических решений. Отличие его от названных ранее состоит в том, что один из вариантов выбора уводит теорию за рамки лингвистики в собственном смысле и тем самым ставит ее в методологически «неудобное» положение: независимо обосновать сделанное предположение, по крайней мере имеющимися в распоряжении филолога методами, оказывается практически невозможно. Речь идет о вопросе «семантика, понимаемая как модуль, ответственный за оперирование уже определенными объектами, или (наивная) онтология, понимаемая как опись таких объектов?». (Новизна относится именно к проблеме выбора, а не к проблеме подлежащей языку онтологии как таковой; о последней см. обзор Ф. Мольтманн [Moltmann, 2018].) Заметим сразу, что речь **не** идет об использо-

вании языкового материала как «царского пути» в картину мира говорящих; признание некоторых обнаруживаемых при семантическом исследовании фактов фактами не семантики, но подлежащей языку онтологии скорее облегчает бремя самой семантики, освобождая ее от необходимости объяснять и каждый раз снова воспроизводить эффект, который может быть отнесен на счет менее частных и, возможно, менее формальных механизмов.

Статья не содержит новых семантических проблем или новых подходов к известным проблемам. Задача ее в том, чтобы представить известные идеи в семантике и философском анализе языка как примеры решений поставленного выше вопроса. При этом мы не проводим четкой границы между работами, конечной целью которых был ответ на онтологический вопрос, и работами, которые отвечают на тот или иной семантический вопрос; и те, и другие представлены как опыты использования онтологии для нужд семантики. В разделе 1 приводятся некоторые случаи, когда анализ семантики высказываний на естественном языке требует допущений относительно наивной картины мира говорящих, без которых ассерция этих высказываний была бы аномальна. В разделе 2 мы переходим к случаям, когда анализ языка приводит к постулированию в картине мира говорящих уже не просто тех или иных сущностей и отношений между ними, а тех или иных **категорий** сущностей, т. е. переходит из сферы наивной физики в сферу наивной онтологии. В наиболее пространном разделе 3 речь идет об аналитических решениях, использующих фундаментальные (для той или иной онтологии) отношения между сущностями для объяснения ограничений, эмпирически наблюдаемых в семантике естественно-языковых высказываний. Итоги подведены в заключении.

1. Семантика и наивная картина мира

Необходимость различать научную и наивную картину мира (не говоря о реальном положении дел), включая инвентарь вещей[1], признаваемых существующими, может быть легко продемонстрирована на таких примерах, как способность утверждения быть истинным, несмотря на очевидную ложность той наивной астрономии, которая лежит в основе интерпретации высказываний типа *Солнце садится*:

[1] Здесь речь идет о конкретных вещах, а не о категориях, как в наивной онтологии.

(1) Вот уже вечер, **солнце садится**, галки медленно летят над деревьями вдоль Страстного бульвара… (В. А. Каверин)[2].

В несколько иной форме, без указаний на то, что одна из картин мира более верна, чем остальные, идея модели как концептуализации высказывается С. Лауэром и А. Джалали [Lauer, Djalali, 2014][3]. (Среди своих предшественников они называют М. Крифку, приводя цитату со сходными мыслями из его работы [Krifka, 1998, p. 198].)

Этот вопрос оказывается менее ясным там, где разграничение провести сложнее или где релевантный фрагмент научной картины мира недостаточно развит или неизвестен широкой публике. Так, весьма вероятно, что различным типам ментальных состояний, таких как убеждение, желание, опасение и пр., в действительности соответствуют достаточно сложные и образующие континуум (или сеть сходств) комплексы событий в головном мозге; тем не менее наивно-психологическая картина мира, в которой существуют различные «содержания» сознания, от этого не теряет статус ключа к пониманию таких высказываний, как:

(2) …я боялся, что по литературе у меня в году будет «плохо» (В. А. Каверин).

Точка зрения, согласно которой если наивная психология в значительной мере ложна, то и наилучшей экспликацией значения примера 2 будет сложное нейрофизиологическое описание, не имеет под собой достаточных оснований как раз в силу наличия между действительностью и ее описанием в языке того промежуточного уровня концептуализации, о котором говорят С. Лауэр и А. Джалали. Усиливая их тезис, мы можем добавить, что если адекватный анализ значения примера 2 требует признания истинными тех или иных пресуппозиций (например, о природе ментальных состояний), которые **достоверно ложны**, то и это должно быть условно признано для целей семантического моделирования.

---

[2] Здесь и далее цитаты с автором, приведенным в круглых скобках, взяты из Национального корпуса русского языка (НКРЯ): https://www/ruscorpora.ru (дата обращения: 22.06.2018).

[3] В числе их примеров и противопоставление исчисляемых и неисчисляемых существительных [Chierchia, 2010], которое проводится разными языками различно, но отчасти ограничено свойствами объектов в мире.

2. Выявление наивной онтологии

Иногда естественный язык выступает своего рода арбитром при разрешении онтологических споров между философами. Так, один из доводов С. Шиффера [Schiffer, 1996] в пользу «существования» пропозиций — абстрактных объектов, носителей истинности и квантов информации — состоит в том, что в различных языках осмысленны предложения типа:

(3) Катя думает то же, что и Кирен;
(4) Katya believes something Kiren believes.

С другой стороны, это *то же* как минимум иногда нельзя эксплицитно **назвать** пропозицией, поскольку, к примеру, «бояться того же, что и Кирен» не означает «бояться той же пропозиции, что и Кирен», потому что Кирен, скорее всего, боится не пропозиции, а чего-то другого (ср.: [Moltmann, 2003, p. 82]).

Аналогичные рассуждения приводят к выводам о категоризации действительности в том или ином отдельном языке. Так, ранее мы (уже в лингвистических целях) предположили, что в русском языке (точнее, в «модели» мира, относительно которой носители русского языка интерпретируют предложения на нем) сортовые индивиды, соответствующие естественным классам объектов (например, *доктор* или *горный козёл*), помещены в тот же домен, что и обычные индивиды [Тискин, 2015]. Основанием для такого предположения были примеры типа:

(5) Нет, говорю, я не вор и никогда **им** не был, кабы не любовь к женщине (К. Г. Паустовский);
(6) Несмотря на то что оно выглядело птицей, оно **ею** не было (цит. по: [Тискин, 2015, с. 334]).

В отличие, к примеру, от английского языка, где в позиции предиката часто используются особые анафорические слова *so, that, one*, в русском языке в таких случаях свободно употребляются личные местоимения 3-го лица. При этом на антецедент местоимения налагаются некоторые ограничения; так, он должен иметь субстантивную вершину[4] (ср. неприемлемое предложение 7), а при наличии адъек-

[4] Хотя ср. с адъективным антецедентом: *Предстоящая человечеству жизнь все более, полнее и очевиднее станет промышленною, хотя начальная вовсе **ею** не была* (Д. И. Менделеев); *Моментами я очень жалею, что я не богат и никогда **им** не был* (А. Н. Бенуа).

тивных распространителей значением антецедентной именной группы должен быть естественный класс (пример 8):

(7) *Иван Павлыч был добрым, а Николай Антоныч **им** не был;
(8) Саня был {[ОК]военным, [??]храбрым} лётчиком, а Ромашка **им** не был.

Такие ограничения могут свидетельствовать о том, что антецедент предикатного местоимения должен обозначать естественный класс; а поскольку форма местоимения в подобных случаях не отличается от формы обычного анафорического местоимения, следует считать, что как минимум в некоторых случаях естественные классы рассматриваются в русском языке как индивиды.

3. Ограничения на модели в интересах семантики

Помимо того, какие категории сущностей необходимо признать в наивной онтологии и какие «факты» о них следует условно признать, чтобы интерпретация высказываний на данном языке была успешной, остается еще как минимум один вопрос. Какие бы категории объектов и конкретные объекты ни были признаны, наряду с их «номенклатурой» данным языком может предполагаться еще некоторая составляемая ими структура — базовые отношения и операции, связывающие их. Мы не претендуем отличить базовое от «надстроечного»; по своему характеру (и, вероятно, по сложности) эта проблема напоминает различение логического и фактического. Будем, однако, считать, что отношения, определенные на целых доменах (как алгебраические операции могут быть определены на целом домене, которым занимается данная математическая теория, или как логические отношения наподобие следования могут быть определены на всем домене высказываний, или как мереологические отношения могут быть определены для любых concreta), являются, по-видимому, базовыми, а отношения, имеющие смысл только для ограниченного подмножества данного домена (как ‘отец’ или ‘нравиться на вкус’), в число базовых не входят.

Итак, различные модели могут различаться не только инвентарем объектов (и не только инвентарем **типов** объектов), но и тем, какие отношения между ними считаются допустимыми.

3.1. Отношение сходства между мирами

Д. Льюис [Lewis, 1973] предложил анализ высказываний о причинной связи типа *Если* (*бы*) *A, то B*, основанный на семантике возможных миров, несколько измененной по сравнению с традиционной семантикой Крипке. В этой семантике вместо бинарного отношения достижимости определено тернарное отношение совокупного сходства. Так, $w \leq_{@} v$ означает, что мир *w* отличается от действительного мира @ не больше, чем мир *v*, по всей сумме обстоятельств в этих мирах. (Как именно рассчитывается сумма, заранее не указано, что Льюис считает достоинством своего подхода: ведь и в употреблении условных высказываний есть некоторая неопределенность.) По Льюису, *Если* (*бы*) *A, то B* означает:

(9) Во всех ближайших к @ *A*-мирах (т. е. мирах, где имеет место *A*) имеет место *B*.

Причинная связь между событиями e_1 и e_2 определяется как конъюнкция:

(10) Если происходит e_1, то происходит e_2, и если не происходит e_1, то не происходит e_2.

Сам Льюис перечисляет некоторые проблемы, с которыми сталкивается такая трактовка. В частности, (а) если e_2 вызывает e_1, но не наоборот (т. е. причинные отношения обратны описанным выше), и почти ничто в мире не может этому помешать, высказывание в примере 10 тоже оказывается истинно; по определению получается, что причинная связь направлена от e_2 к e_1, что противоречит условию «но не наоборот». Кроме того, (б) если e_3 вызывает e_2, но также и некоторое не связанное прямо с последним e_1 («эпифеномен» события e_3), то названное высказывание также истинно, откуда по определению можно утверждать, что между эпифеноменом e_1 и e_2 имеется причинная связь, что противоречит условию.

Чтобы справиться с этими проблемами, Льюис не предлагает никакой модификации семантики в примере 9 или релевантной модификации анализа причинных высказываний в примере 10. Фактически он выражает надежду на то, что **само отношение сходства** $\leq$ справится с ними: в обоих случаях (а и б) Льюис отрицает истинность *Если не происходит* e_1, *то не происходит* e_2. Он заявляет, что миры, где происходит событие-причина (e_2), но обстоятельства и законы природы

отличаются от имеющих место в действительности, а потому не происходит событие-следствие (e_1), **ближе к действительному миру** @, чем миры, где не происходят ни e_1, ни e_2. Поэтому (а) если не происходит e_1, то — в ближайших к @ мирах без e_1 — e_2 все-таки происходит, просто не вызывает своего нормального следствия e_1. Поэтому же (б) если не происходит эпифеномен e_1, то e_2 все-таки происходит, просто e_3 не вызывает своего обычного эпифеномена e_1.

Решение Льюиса существенным образом полагается на то, что «верным» способом расчета сходства между мирами окажется один из таких, которые решают названные проблемы, и таким образом налагает ограничения на модели — множества возможных миров с определенными на них отношениями сходства.

3.2. Онтологически приоритетные индивиды

Другой пример того, как ограничения на возможные модели помогают решать конкретные семантические проблемы, представлен в работе У. Зауэрланда [Sauerland, 2014]. Он обсуждает проблему, описанную в ряде предшествующих работ (в частности, О. Перкуса [Percus, 2000]): хотя в предложениях типа представленного в примере 11 подлежащее придаточного предложения может интерпретироваться de re, это невозможно для сказуемого.

(11) **Он** думает, что **ты** живешь в Ташкенте (В. А. Каверин).

Иными словами, это высказывание может быть истинно как если «он» думает, глядя на того, кто назван *ты* (т. е., согласно сюжету, на Петьку), что этот человек живет в Ташкенте, так и если «он» думает что-то вроде «единственный рыжий художник родом из Энска живет в Ташкенте», тогда как **на самом деле** этому описанию удовлетворяет как раз Петька. Такой свободы, однако, нет для *живешь в Ташкенте*: высказывание в примере 11 не будет истинно, если «он», глядя на базу данных с перечнем жителей Ташкента, но не зная, что за перечень перед ним, подумал: «Петька должен быть в списке, который сейчас передо мной». По какой-то причине замена эквивалентных в одних частях придаточного предложения оказывается возможна, а в других — невозможна (или, самое меньшее, гораздо сложнее).

Чтобы объяснить этот факт, Зауэрланд предполагает следующее. Допустим, *думать, что A* означает находиться в некотором отношении к множеству возможных миров, в которых верно *A*. Если в по-

ложении дел *A* участвуют какие-либо индивиды, в разных мирах это не может быть один и тот же индивид: индивид существует только в конкретном мире; в других мирах могут быть только «двойники» данного индивида (как в исходной семантике двойников Д. Льюиса [Lewis, 1968]). Предположим теперь, что можно восстановить «двойника» реального индивида в каком-то ином возможном мире, но не «двойника» воображаемого индивида в реальном мире, поскольку возможность иметь убеждения относительно реальных индивидов причинно связана с реальным существованием этих индивидов, тогда как это существование не зависит от того, кто и что о них думает [Sauerland, 2014, p. 78]. (Здесь рассуждения Зауэрланда совершенно явно покидают собственно лингвистическую сферу, хотя мотивация асимметрии, которая здесь вводится, могла бы быть и совершенно иной без ущерба для ее семантической полезности.)

Таким образом, Зауэрланд принимает некоторое ограничение на возможные отношения (‘быть двойником’) между индивидами в зависимости от свойств и отношений миров. Если принять «онтологию» Зауэрланда с приоритетом действительных индивидов над лишь возможными, то окажется, что найти «двойника» Петьки, соответствующего описанию «единственный рыжий художник родом из Энска», в мирах, где выполняются убеждения человека, названного *он* в примере 11, возможно; он и будет выполнять там свойство «живет в Ташкенте». С другой стороны, взять «двойника» Петьки и выяснить, живет ли он в Ташкенте в **действительности**, невозможно, а как раз это требовалось бы, чтобы перейти от описания «список, который сейчас передо мной» к описанию «список людей, живущих в Ташкенте».

Чтобы адекватно оценить предложенное Зауэрландом решение, следует учитывать, что ему приходится соперничать в изяществе и объяснительной силе с решениями «внутрилингвистического» характера, такими как решение Дж. Ромоли и Я. Судо, основанное на механизме проекции для пресуппозиций [Romoli, Sudo, 2009], и решение Д. Шулера, использующее альтернативный синтаксический механизм связывания переменных [Schueler, 2011].

3.3. Логическая структура и импликатуры

Еще один пример, приводимый С. Лауэром и А. Джалали [Lauer, Djalali, 2014], связан со структурой шкал, ассоциированных с градуируемыми признаками (типа ‘высокий’ или ‘добрый’).

Известно, что у шкалы часто имеется маркированный конец — тот, термин для которого используется также и как термин для данного параметра вообще (*высота*, но не **низкость*; в НКРЯ сочетания типа *короче другого* встречаются почти втрое чаще, чем *длиннее другого*, но *выше другого* — более чем вчетверо чаще, чем *ниже другого*). Ни логика, ни физика не определяют для всех случаев, какому из концов быть маркированным (хотя наличие на том или ином конце предельно возможного значения играет здесь некоторую роль).

Как известно, понятие шкалы лежит в основе теории скалярных импликатур — фактически одного из компонентов потенциальной «логики естественного языка», т. е. описания всех возможных отношений следования, существующих между множествами высказываний некоторого естественного языка.

Не все шкалы подобны шкале измерительного прибора; так, некоторыми семантистами признаются шкалы, связанные с логической силой утверждения. Своего рода тип структур, из которого наивная картина мира может принимать одни и отвергать другие, составляют логические структуры — рассуждения и лежащие в их основе логические отношения, в первую очередь отношение следования. Вполне допустим вопрос: является ли классическая логика (в языке какой бы то ни было выразительной силы) той логикой, которой пользуются в своих рассуждениях носители (данного) естественного языка, или таковой является какая-то более слабая логика? Последнее представляется вероятным как минимум в том отношении, что некоторые рассуждения, верные с точки зрения классических логик, но отвергаемые более слабыми (релевантными), на естественном языке выглядят если и не ошибочными, то странными. Более однозначный, почти классический аргумент в пользу семантически более сложной логики, чем классическая, основан на недостаточности материальной импликации для анализа условных высказываний на естественном языке. Так, для истинности нижеследующего высказывания недостаточно, чтобы слушающий имел слабую волю, тогда как $(A \rightarrow B)$ истинно при ложности A вне зависимости от какой-либо связи между A и B:

(12) Если бы у тебя была **сильная** воля, ты бы хорошо учился (В. А. Каверин).

Отношение этой проблемы к вопросу о шкалах состоит в том, что в классической логике высказывания A и $(A \wedge B)$ упорядочены друг

относительно друга на некоторой шкале: второе «сильнее» первого, т. е. первое логически следует из второго. В шкале, соответствующей релевантной или немонотонной логике, эти объекты несравнимы (т. е. ни один из них в общем случае — для произвольных *A* и *B* — не «сильнее» другого и не влечет его).

Одна из тенденций в современной семантике (см.: [Chierchia, 2013]) состоит в том, чтобы в ряде случаев рассматривать скалярные отношения как грамматикализованные (а потому принадлежащие к семантике, а не к прагматике) и тем объяснять дистрибуцию и фактически наблюдаемое значение чувствительных к полярности единиц (таких как *ни*, англ. *any*, фр. *ou… ou* и т. д.). При этом в некоторых случаях шкалу образуют не различные по логической «силе» высказывания, а поддомены данного домена. Так, альтернативами домену {Саня, Катя, Ромашка} будут {Саня, Катя}, {Катя, Ромашка}, {Саня, Ромашка}, а также {Саня}, {Катя} и {Ромашка}; чтобы объяснить, почему англ. *any* возможно при отрицании, как в примере 13, но невозможно без отрицания, допускают, что в структуре примера 13[5] имеется еще нулевой элемент O_D, необходимый для появления *any*:

(13) a. I do not see anybody.
b. O_D [not [I see anybody]].

Этот оператор иногда называют «немым *только*»: его семантический эффект состоит в том, чтобы добавить к утверждению *p*, что все поддоменные альтернативы *p*, не следующие логически из самого *p*, ложны. Если истинно сказанное в примере 13, т. е. если говорящий не видел никого из множества {Саня, Катя, Ромашка}, то, разумеется, он не видел никого из {Саня, Катя} и никого из прочих перечисленных подмножеств; поэтому в приведенных высказываниях не возникает противоречия. Попытка же утверждать **I saw anybody* наталкивается на противоречие: если говорящий видел кого-то из {Саня, Катя, Ромашка}, отсюда логически не следует для каждого из собственных подмножеств этого множества, что он видел кого-либо из данного подмножества. Поэтому O_D отрицает все такие альтернативы, откуда и противоречие: ‘Я видел кого-то из {Саня, Катя, Ромашка}, но для каждого подмножества *D′* этого множества не видел никого из *D′*’.

[5] Мы рассматриваем пример на английском языке, чтобы не вникать в особенности русского *ни-*. Это предмет отдельного рассмотрения в работах данного и других авторов.

Можно утверждать, что, как только вообще дан домен индивидов, тем самым в некотором смысле даны и все его поддомены, тем более что именно они служат экстенсионалами предикатов. В случае, когда поддоменными альтернативами выступают пропозиции (как ‘Я видел Саню’ и ‘Я видел Катю’ для *I haven't seen either Sanya or Katya*, где *either… or* ведет себя подобно *any*), альтернативы данной пропозиции могут быть вычислены исходя из отношений логического следования. Тогда, если классическую логику считать уже данной, перед нами не столько новый элемент «наивной **онтологии**», сколько новое **семантическое** решение, не известным ранее способом использующее уже готовые отношения в модели.

Заключение

Рассмотрим еще раз решение, предложенное нами для примеров 5–6. В чем его достоинство с точки зрения семантики? Гипотеза относительно онтологии, допускающая «реальное» в том или ином релевантном смысле включение сортовых индивидов в пробег переменных, соответствующих местоимениям, усложнила (или, во всяком случае, сделала более определенной) постулируемую наивную онтологию, но позволила не усложнять лексикон, постулируя в нем отдельную категорию сортовых местоимений, к тому же омонимичных индивидным, не усложнять синтаксис за счет постулирования нулевых операторов, доопределяющих значение местоимения до сортового (а также — в других случаях — до индивидного), и не усложнять семантику в узком смысле за счет введения в нее операций, меняющих тип с индивидного (или нейтрального) на сортовой.

Вероятно, существует трактовка данной гипотезы, при которой она оказывается если не эмпирически проверяемым положением, то, по крайней мере, утверждением о **языке**, а не о лежащей в основе его использования картине мира: переменные некоторого типа в русском языке таковы, что их пробег включает как классические, так и сортовые индивиды. С другой стороны, интуитивно считать такое положение дел «грубым фактом», т. е. фактом, объяснение которого невозможно, позволительно разве что в рамках семантической теории; за ее пределами вопрос об объяснении возможно как минимум поставить. Ответ, даваемый онтологией, может рассматриваться как объяснение этого факта.

В заключение предложим некоторые предварительные критерии, которыми можно руководствоваться при «разделении труда» между семантикой и онтологией.

I. Вопрос об «элиминируемости» сущностей некоторых типов (свойств и отношений, абстрактных объектов, множеств, мереологических сумм) активно обсуждается в философии. Свою актуальность он имеет и в семантике, которая вообще в некоторых случаях тесно смыкается с философским анализом языка. Вопрос о том, как «на самом деле» воспринимают значение носители, ставится далеко не всегда, а относительно слабая эквивалентность между условиями истинности высказывания и той или иной их формализацией часто не разрешает вопроса «онтологических обязательств» (в смысле, вкладывавшемся в этот термин У. Куайном [Quine, 1948]). Так или иначе, если некоторое высказывание хотя бы на первый взгляд требует для того, чтобы быть понятым (а не только для своей истинности), признания некоторого домена объектов, допустимо принять его в постулируемой для данного языка наивной онтологии.

II. Если некоторый рисунок отношений между элементами модели делает более компактным семантический анализ хотя бы одного типа контекстов и **не опровергается** другими их типами, допустимо постулировать этот рисунок в наивной онтологии как единственно возможный, пока не будет найден контрпример. Так, например, существенно легче раз и навсегда включить домен сортов в домен индивидов, чем ввести для них различные анафорические средства или прописывать квантификацию по объединенному домену в условиях истинности конкретных предложений (хотя в других фрагментах семантики ввиду этого может потребоваться эксплицитно **ограничивать** пробег переменных только вещественными индивидами). Аналогичным образом, легче ввести ограничение на допустимые структуры возможных миров и определенное на индивидах отношение ‘быть двойником’, чем бороться с ошибочными семантическими предсказаниями для отдельных типов предложений.

III. Если некоторый рисунок отношений упрощает собственно семантический компонент анализа в ряде случаев *S*, но (неизбежно) приводит к некорректному анализу как минимум одного

типа контекстов *T*, этот рисунок не может быть постулирован в онтологии и должен считаться продуктом дальнейшего ограничения онтологических возможностей собственно семантикой контекстов, принадлежащих *S*, каковое по той или иной причине не происходит в *T*. Вероятно, примером здесь могут служить различные типы ограничения домена: так, в примере 14 *каждый* квантифицирует по ученикам интерната, а не по всем людям (тем более — индивидам) в мире, однако нет общего правила, которое запрещало бы неограниченную квантификацию по всем индивидам; поэтому нецелесообразно считать, что домен индивидов не существует как что-либо единое, а представляет собой конгломерат отдельных, «ситуативных» доменов.

(14) Выстраивалась очередь, и **каждый**, без различия формы, возраста и происхождения, получал по ложке ещё горячей каши, дьявольски вкусной, с лопающимися пузырьками (В. А. Каверин).

Аналогично, по-видимому, нет существенных синтаксических различий между *Пригласи кого угодно* и *Назови кого угодно*, однако первое из них, произнесенное в разговоре между организаторами конференции, не включает в пробег своего квантора Оккама или Соссюра, тогда как второе, произнесенное как совет сдавшего экзамен студента своему товарищу, вполне может включить и их. Поэтому ограничивать пробег квантора **ныне существующими** индивидами может быть полезно в семантике конкретных типов высказываний (или некоторых коммуникативных ситуаций), но не в общем очерке отношений между семантикой языка и его наивной онтологией.

Литература

Тискин Д. Б. Анафора к предикату и характеристики русской именной группы // Типология морфосинтаксических параметров — 2015: материалы междунар. конф. / отв. ред. Е. А. Лютикова, А. В. Циммерлинг, М. Б. Коношенко. М.: Моск. пед. гос. ун-т, 2015. С. 334–352.

Bach E. Natural Language Metaphysics // Logic, Methodology and Philosophy of Science VII: Proceedings of the 7th International Congress of Logic, Methodology and Philosophy of Science, Salzburg, 1983. Amsterdam; New York: Elsevier Science, 1986. P. 573–595.

Chierchia G. Mass Nouns, Vagueness and Semantic Variation // Synthese. 2010. Vol. 174. No. 1. P. 99–149.

Chierchia G. Logic in Grammar: Polarity, Free Choice, and Intervention. Oxford: Oxford University Press, 2013.

Krifka M. The Origins of Telicity // Events and Grammar. Dordrecht: Springer, 1998. P. 197–235.

Lauer S., Djalali A. A Conceptual-epistemic Perspective on Model Theory: [Workshop Presentation] // Stefan Kaufmann: [сайт]. Дата публикации: [20.08.2014]. http://stefan-kaufmann.uconn.edu/Models_ESSLLI2014/Slides/LauerDjalali.pdf (дата обращения: 22.06.2018).

Lewis D. Counterpart Theory and Quantified Modal Logic // Journal of Philosophy. 1968. Vol. 65. No. 5. P. 113–126.

Lewis D. Causation // Journal of Philosophy. 1973. Vol. 70. No. 17. P. 556–567.

Moltmann F. Propositional Attitudes Without Propositions // Synthese. 2003. Vol. 135. No. 1. P. 77–118.

Moltmann F. Natural Language and Its Ontology // Metaphysics and Cognitive Science / ed. by A. I. Goldman, B. P. McLaughlin. Oxford: Oxford University Press, 2019. P. 206–232.

Partee B. Noun Phrase Interpretation and Type-shifting Principles // Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers / ed. by J. Groenendijk, D. de Jongh, M. Stokhof. Dordrecht: Foris, 1987. P. 115–143.

Percus O. Constraints on Some Other Variables in Syntax // Natural Language Semantics. 2000. Vol. 8. Issue 3. P. 173–229.

Quine W. V. O. On What There Is // The Review of Metaphysics. 1948. Vol. 2. No. 5. P. 21–38.

Recanati F. Unarticulated Constituents // Linguistics and Philosophy. 2002. Vol. 25. No. 3. P. 299–345.

Romoli J., Sudo Y. De Re / *De Dicto* Ambiguity and Presupposition Projection // Proceedings of *Sinn und Bedeutung* — 13 / ed. by A. Riester, T. Solstad. Stuttgart: University of Stuttgart, 2009. P. 425–438.

Sauerland U. Counterparts Block Some 'De Re' Readings // The Art and Craft of Semantics: A Festschrift for Irene Heim: in 2 vols. / ed. by L. Crnič, U. Sauerland. Vol. 2. Cambridge (MA): MIT Press, 2014. P. 65–85.

Schiffer S. Language-created Language-independent Entities // Philosophical Topics. 1996. Vol. 24. No. 1. P. 149–167.

Schueler D. World-variable Binding and Beta-binding // Journal of Semantics. 2011. Vol. 28. No. 2. P. 241–266.

References

Bach E. 1986. Natural Language Metaphysics. *Logic, Methodology and Philosophy of Science VII. Proceedings of the 7th International Congress of Logic, Methodology and Philosophy of Science, Salzburg, 1983*. Amsterdam; New York, Elsevier Science, pp. 573–595.

Chierchia G. 2010. Mass Nouns, Vagueness and Semantic Variation. *Synthese*, vol. 174, no. 1, pp. 99–149.

Chierchia G. 2013. *Logic in Grammar. Polarity, Free Choice, and Intervention*. Oxford, Oxford University Press.

Krifka M. 1998. The Origins of Telicity. *Events and Grammar.* Dordrecht, Springer, pp. 197–235.

Lauer S., Djalali A. 2014. A Conceptual-epistemic Perspective on Model Theory. [Workshop Presentation]. *Stefan Kaufmann*, [site]. Publication date: [20.08.2014]. http://stefan-kaufmann.uconn.edu/Models_ESSLLI2014/Slides/LauerDjalali.pdf (accessed date: 22.06.2018).

Lewis D. 1968. Counterpart Theory and Quantified Modal Logic. *Journal of Philosophy*, vol. 65, no. 5, pp. 113–126.

Lewis D. 1973. Causation. *Journal of Philosophy*, vol. 70, no. 17, pp. 556–567.

Moltmann F. 2003. Propositional Attitudes Without Propositions. *Synthese*, vol. 135, no. 1. pp. 77–118.

Moltmann F. 2019. Natural Language and Its Ontology. *Metaphysics and Cognitive Science*, A. I. Goldman, B. P. McLaughlin (eds.). Oxford, Oxford University Press, pp. 206–232.

Partee B. 1987. Noun Phrase Interpretation and Type-shifting Principles. *Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers*, J. Groenendijk, D. de Jongh, M. Stokhof (eds.). Dordrecht, Foris, pp. 115–143.

Percus O. 2000. Constraints on Some Other Variables in Syntax. *Natural Language Semantics*, vol. 8, issue 3, pp. 173–229.

Quine W. V. O. 1948. On What There Is. *Review of Metaphysics*, vol. 2, no. 5, pp. 21–38.

Recanati F. 2002. Unarticulated Constituents. *Linguistics and Philosophy*, vol. 25, no. 3, pp. 299–345.

Romoli J., Sudo Y. 2009. *De Re* / *De Dicto* Ambiguity and Presupposition Projection. *Proceedings of Sinn und Bedeutung —13*, A. Riester, T. Solstad (eds.). Stuttgart, University of Stuttgart, pp. 425–438.

Sauerland U. 2014. Counterparts Block Some 'De Re' Readings. *The Art and Craft of Semantics. A Festschrift for Irene Heim*, in 2 vols., L. Crnič, U. Sauerland (eds.), vol. 2. Cambridge (MA), MIT Press, pp. 65–85.

Schiffer S. 1996. Language-created Language-independent Entities. *Philosophical Topics*, vol. 24, no. 1, pp. 149–167.

Schueler D. 2011. World-variable Binding and Beta-binding. *Journal of Semantics*, vol. 28, no. 2, pp. 241–266.

Tiskin D. B. 2015. Anaphora to Predicates and the Characteristics of Russian Noun Phrases. *Tipologiia morfosintaksicheskikh parametrov — 2015. Materialy mezhdunar. konf.*, E. A. Liutikova, A. V. Tsimmerling, M. B. Konoshenko (eds.). Moscow, Mosk. ped. gos. un-t Publ., pp. 334–352. (In Russ.)

М. В. Хохлова

СТАТИСТИЧЕСКИЙ ПОДХОД ПРИМЕНИТЕЛЬНО К ИССЛЕДОВАНИЮ СОЧЕТАЕМОСТИ: ОТ МЕР АССОЦИАЦИИ К МАШИННОМУ ОБУЧЕНИЮ*

Аннотация. Статистические методы активно используются в лингвистике на протяжении долгого времени. С развитием технологий, позволяющих обрабатывать большие текстовые данные, появились новые методы, которые активно применяются в междисциплинарных исследованиях. К ним относится машинное обучение, которое используется при решении разных лингвистических задач. В статье рассматривается эволюция статистических методов применительно к задаче автоматического выявления словосочетаний, обсуждаются как традиционные подходы, связанные с применением мер ассоциации, так и разнообразные статистические алгоритмы.

Ключевые слова. Сочетаемость, статистические методы, машинное обучение, коллокации, корпус текстов.

Maria V. Khokhlova

STATISTICAL APPROACH TO COLLOCATION EXTRACTION: FROM ASSOCIATION MEASURES TO MACHINE LEARNING

Abstract. Statistical algorithms are actively applied in linguistic studies and have a long tradition. The development of new technologies that can be implemented in big data processing gave birth to new methods. They include machine learning that is useful for solving linguistic tasks. The paper describes the evaluation of statistical methods applied to the task of collocation extraction and describes both traditional approaches that involve association measures and various statistical algorithms.

Keywords. Collocability, statistical measures, machine learning, collocations, text corpus.

---

* Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-2513.2018.6 «Исследование методов автоматического извлечения лексических конструкций на основе машинного обучения».

В России первые серьезные попытки применить статистические методы на языковом материале были предприняты в начале XX века исследователями Н. А. Морозовым и А. А. Марковым. Н. А. Морозов рассматривал вопросы, связанные с изучением языка разных авторов, и стремился вывести общие стилеметрические законы [Морозов, 1915]. Он выделял служебные частицы как маркеры особенностей стиля писателя, предлагая вычислять частоту той или иной единицы в первой тысяче слов изучаемого текста. Данный метод был проиллюстрирован на примерах произведений А. С. Пушкина, Н. В. Гоголя, И. С. Тургенева, Л. Н. Толстого, Н. М. Карамзина и М. Н. Загоскина. Таким образом частотные распределения, или спектры, были использованы для различения подлинных текстов авторов и текстов, которые могли быть им ошибочно приписаны. Результаты экспериментов были подвергнуты критике в работе видного русского математика А. А. Маркова [Марков, 1916], который указывал, что найденные закономерности могут относиться только к тем отрывками, на материале которых проводилось исследование, и могут быть характерны не для всех текстов, а только для текстов рассмотренных писателей. Исследователь использовал материал поэмы «Евгений Онегин» для демонстрации случайных процессов, получивших название «марковских». Позднее академик В. В. Виноградов указывал на необходимость исследовать частоты употребления разных типов слов в текстах разных стилей и тем самым выявлять их различия [Виноградов, 1938, с. 155–156].

Направление, посвященное сопряжению статистических методов и традиционной лингвистики, получило новый импульс в связи с появлением технических возможностей, позволивших привлечь вычислительные средства к решению ряда задач, в том числе получить новые данные и проверить гипотезы на большом материале. Можно назвать ряд работ, которые появились в этот период и позволили по-новому взглянуть на языковой материал и дать ему количественную оценку (например: [Андреев, 1967; Пиотровский, 1968; Головин, 1970; Арапов, 1988]). Статистические методы нашли активное применение в задачах, связанных с атрибуцией текстов [Марусенко, 1990], с определением стилистических характеристик текстов [Мартыненко, 1988; Мартыненко, Чебанов, 1996], при составлении словарей языка авторов [Шайкевич и др., 2003] и частотных словарей [Ляшевская, Шаров, 2009]. Вычисляются разнообразные статистические показатели, кото-

рые могут использоваться для кластеризации лексики, определения тематически значимых слов и др.

Корпусная лингвистика стала еще одним направлением в прикладном языкознании, где активно стали использоваться компьютерные технологии. Результаты, выдаваемые системами в ответ на разнообразные запросы, снабжены статистической информацией. Она может включать данные о частотах слов или грамматических категорий, об оценке устойчивости выделенных словосочетаний, о продуктивности словообразовательных моделей или синтаксических конструкций и многое другое. Современные лингвистические теории уделяют внимание выделению и описанию воспроизводимых в тексте языковых конструкций, которые определяются как закономерностями сочетаемости на уровнях лексики, семантики и синтаксиса, так и статистическими характеристиками. Определение подобных языковых структур и их последующий анализ весьма важны для прикладных задач.

В последнее время в связи с возросшей потребностью в автоматизированных системах большое внимание уделяется вопросу, связанному с исследованием коллокаций, — словосочетаний, которые обладают определенной воспроизводимостью в речи. Ряд мер получил название мер ассоциации, или ассоциативных мер. Они позволяют вычислять силу связи между элементами словосочетаний и основываются на частотах данных словосочетаний и входящих в них отдельных слов. Таким образом, может быть вычислена некоторая устойчивость, присущая лексическим единицам, позволяющая расположить их на шкале от свободных сочетаний до фразеологизированных структур. Данное направление выросло из тех задач, с которыми сталкивались (и до сих пор сталкиваются) лексикографы и терминологи. Речь шла о выявлении значимых слов и словосочетаний, а также уменьшении ручной работы, заключавшейся в просмотре текстов и последующем отборе единиц, которые могли претендовать на лексикографическое или терминологическое описание.

Автоматическое извлечение данных из текстового материала в целом можно свести к двум подходам: подход, основанный на правилах, и подход, основанный на статистических данных. Изначально для нахождения значимых цепочек слов в текстах использовалась лингвистическая информация. Впоследствии стала рассматриваться статистическая неслучайность совместного появления разных лек-

сических единиц, т. е. давалась количественная оценка воспроизводимости.

Первой работой, в которой обсуждались статистический аппарат и его использование в оценке устойчивости словосочетаний, стали труды Симпозиума по методам измерения статистической ассоциации применительно к документам, представленным в машинном виде [Proceedings of the Symposium…, 1965]. В статье Дж. Берри-Рогге впервые применяется данный аппарат к извлечению словосочетаний [Berry-Rogghe, 1973]. В работе Ч. Мэннинга и Х. Шульце было дано описание нескольких мер, оценивающих связанность [Manning, Schütze, 1999]. В исследовании С. Эверта приводятся данные о статистическом аппарате, который используется для вычисления силы связанности компонентов словосочетаний [Evert, 2004]. Чаще всего рассматриваются двухкомпонентные словосочетания, или биграммы, реже трехкомпонентные (триграммы). В монографии П. Печина описаны 82 ассоциативные меры, позволяющие оценивать силу связанности единиц внутри биграмм [Pecina, 2009]. На сегодняшний день это наиболее полное описание статистического инструментария. Автор делит меры на три группы: 1) меры, учитывающие наблюдаемые и ожидаемые частоты слов, образующих словосочетание; 2) меры, оценивающие энтропию между словами; 3) меры, учитывающие контекст. Несмотря на имеющийся столь обширный список разработанных мер, наиболее часто в литературе, исследующей понятие коллокации и вычисление силы связанности, упоминается только несколько: коэффициенты взаимной информации, логарифмического правдоподобия, Дайса, t-score, хи-квадрат и др. При этом другие меры ассоциации могут выдавать иные результаты, но они почти не описаны и используются весьма редко.

Также необходимо отметить тот факт, что аппарат мер ассоциации часто применяется к английскому языку, в то время как другие языки рассматриваются намного реже. Среди работ на ином материале можно упомянуть исследования французского [Dias et al., 1999], немецкого [Evert, Krenn, 2001], чешского [Pecina, 2009], русского [Хохлова, 2008; Kormacheva, Pivovarova, Kopotev, 2018], латышского и литовского [Mandravickaite, Krilavičius, 2017] языков.

В исследованиях уделяется меньше внимания словосочетаниям большей длины (триграммам, четырех- и пятиграммам и т. д.) ввиду усложнения вычислений и увеличения количества возможных син-

таксических моделей, которые могли бы использоваться в качестве фильтров. В работе К. Рамиша и соавторов представлены формулы для вычисления связанности в цепочках слов разной длины. Всего приводятся четыре статистические меры, которые могут применяться не только к биграммам: коэффициенты взаимной информации, Дайса, t-score и оценка максимальной вероятности [Ramisch et al., 2010].

Традиционные методы извлечения словосочетаний используют лексико-грамматические шаблоны для поиска возможных примеров конструкций. В качестве механизма для фильтрации применяются пороговые значения мер или частот, списки стоп-слов и др. Полученные списки словосочетаний затем ранжируются в соответствии с одной из статистических мер, отражающих степень связи между двумя признаками.

Значения мер могут показывать разное ранжирование, целью которого является получение списков «хороших» коллокаций, сконцентрированных в верхней части выдаваемых списков. Для проверки полученных словосочетаний должны привлекаться дополнительные данные — это могут быть словари, оценки информантов или экспертов.

При проведении экспериментов с привлечением автоматических методов важным вопросом является создание золотого стандарта — некоторого эталона, относительно которого будут оцениваться полученные результаты [Khokhlova, 2018]. Информацию о сочетаемости можно извлекать из лексикографических источников и формировать на ее основе списки верифицированных словосочетаний по типу золотого стандарта. К подобным словосочетаниям, которые могут пополнить золотой стандарт, относятся фразеологизмы, идиомы, устойчивые словосочетания и другие конструкции. Они могут содержаться не только в специально обозначенных частях словарных статей, но также в цитатах и речениях. Построение такого эталона является отдельной задачей, которая решается, в частности, при проведении соревнований по автоматическому выявлению неоднословных единиц. В простом случае разделение выделенных словосочетаний носит характер бинарного (коллокации и неколлокации, т. е., иными словами, присутствующие или отсутствующие в золотом стандарте). Таким образом вычисляется точность (отношение коллокаций к общему числу выданных словосочетаний). При этом весьма

важным является объем экспертных данных, так как количество словосочетаний, выдаваемых на материале больших интернет-корпусов, может превысить их на несколько порядков, что создаст определенные трудности при оценке эффективности статистических методов. Ограниченная сочетаемость может быть по-разному представлена в словарях. Например, на более чем 83 тыс. словарных статей в так называемом Малом академическом словаре (МАС) [Словарь русского языка, 1981–1984] приходится свыше 10 тыс. единиц в заромбовой части. В «Большом толковом словаре русского языка» (БТС) [Большой толковый словарь…, 1998] описано 8 тыс. фразеологизмов при общем количестве словарных статей, равном 80 тыс. единиц. В зарубежной лексикографии придерживаются иного подхода. В Оксфордском словаре коллокаций [Oxford Collocation Dictionary…, 2009] приводится около 250 тыс. словосочетаний, которые представляют сочетаемость 9 тыс. существительных, глаголов и прилагательных. Словарь немецких коллокаций [Häcki Buhofer et al., 2014] описывает 95 тыс. немецкоязычных словосочетаний для 2 тыс. существительных, глаголов и прилагательных.

Можно сказать, что корпусные данные благодаря статистическим методам проверяются экспериментально. Наряду с тем, что словари могут использоваться в качестве эталона при проверке результатов автоматической обработки, также данные самих экспериментов могут служить основой для словарей и грамматик нового типа, привлекающих корпусный материал[1].

Среди выдаваемых словосочетаний могут присутствовать единицы разных типов — свободные и устойчивые словосочетания, термины, фразеологизмы, имена собственные и др. В связи с этим может стоять задача их автоматического разбиения на данные классы с учетом значений статистических мер. При последующей оценке списков коллокаций и, как следствие, работы алгоритмов могут привлекаться данные, полученные от разных экспертов. Также может вычисляться их согласованность между собой. При этом подобная работа может быть довольно субъективной, на что указывают исследователи С. Эверт и Б. Кренн [Evert, Krenn, 2001].

[1] Здесь можно упомянуть грамматику русского языка (http://rusgram.ru, дата обращения: 01.10.2018), разрабатываемую на корпусном материале (Национальный корпус русского языка — НКРЯ).

Наряду с рассмотрением значений только одной меры (или по отдельности нескольких мер) используется подход, который подразумевает объединение разных мер. При этом речь может идти об использовании как значений мер, так и рангов, полученных одними и теми же словосочетаниями в разных списках [Klyshinsky, Khokhlova, 2017]. Также при оценке эффективности применения той или иной меры может использоваться сравнение с некоторым базовым показателем (baseline) — например, насколько успешно алгоритм себя проявляет по сравнению с обычной частотой совместной встречаемости.

В 1990-х годах при решении задачи извлечения коллокаций стали использоваться методы машинного обучения. Машинное обучение представляет собой совокупность статистических алгоритмов и в определенной степени является продолжением описанных выше методов. Упомянутые значения статистических мер в том числе могут быть использованы в качестве признаков в методах машинного обучения. В отличие от простого статистического аппарата, машинное обучение показывает лучшие результаты.

П. Печиной были проведены эксперименты по автоматическому выделению словосочетаний, построенных по определенным синтаксическим моделям, на материале немецкого и чешского языка при помощи статистических мер и методов машинного обучения [Pecina, 2008]. Он использовал 55 мер ассоциации. Каждое словосочетание рассматривалось как вектор, который состоит из значений данных мер, а также имеет значение 0 или 1 в зависимости от того, встретилась ли данная биграмма в золотом стандарте. Часть данных использовалась для обучения статистических моделей классификации предсказанию, является ли словосочетание «правильным», т. е. присутствует ли оно в золотом стандарте. Классификация не предполагала разделения по признаку «да/нет», а скорее приписывала определенное значение, которое могло быть использовано для ранжирования результатов. В ходе работы сравнивались статистические меры и методы машинного обучения. Среди последних тестировались следующие модели классификаторов: линейная логистическая регресссия, линейный дискриминантный анализ (linear discriminant analysis — LDA), нейронные сети с одним и пятью скрытыми слоями. Словосочетания, которые являлись эталоном при проведении экспериментов, были размечены профессиональными

лексикографами и поделены на несколько групп, в том числе: коллокации (и иные многословные единицы), частотные словосочетания (могут являться частью коллокационной модели), неидиоматизированные словосочетания, свободные словосочетания и ошибки лематизации. Извлекались немецкоязычные биграммы, построенные по модели Adj-N.

В ходе первого эксперимента словосочетания рассматривались относительно представленности в группах коллокаций и частотных словосочетаний. При этом базовый уровень точности, достигнутый при использовании частоты совместной встречаемости, был довольно высок и составил 42,12 %. Наилучшие результаты были продемонстрированы при помощи меры ассоциации Пятерского — Шапиро (Piatersky — Shapiro coefficient): макроусредненная средняя точность — 62,88 %. В рамках данного эксперимента ни один из классификаторов не превзошел данного результата.

В ходе второго эксперимента словосочетания классифицировались по трем группам (была добавлена группа неидиоматизированных словосочетаний). Базовый уровень точности составил 51,78 %. Классификатор LDA продемонстрировал более успешные результаты, хотя не намного превзошел меру Пятерского — Шапиро. В качестве материала для второй серии экспериментов рассматривались глагольные группы с предлогами на немецком языке, которые были поделены на два класса: обладающие связанностью и не обладающие. Базовая точность оказалась очень низкой и составила 2,91 %, лучшей мерой была признана мера доверия (confidence) с результатом 18,26 %. Среди методов машинного обучения наилучшие результаты были продемонстрированы нейронной сетью с одним скрытым слоем (30,77 %). При использовании частотного порога, равного 30, мера доверия также показала лучшие результаты. Ее превзошла нейронная сеть с пятью слоями.

В ходе третьего эксперимента пуассоновская мера значимости (Poisson significance measure) показала лучший результат (43,97 %), при этом несущественно лучше показала себя LDA. Наряду со словосочетаниями на английском языке также оценивались коллокации на чешском языке, извлеченные из синтаксически размеченного корпуса (Prague Dependency Treebank). Для чешского языка лучший результат демонстрирует мера униграммного подкортежа (Unigram subtuple), ее превосходит нейронная сеть с пятью слоями.

Исследователи приходят к заключению о том, что не существует одной меры, которая демонстрировала бы наилучшие результаты при выявлении словосочетаний разных типов [Pecina, 2008; Ramisch et al., 2010]. Иными словами, невозможно выбрать одну наиболее удачную статистическую метрику, в том числе и при работе с разными языками [Evert et al., 2018].

В работе К. Франци и соавторов отобранные посредством частеречных фильтров словосочетания оцениваются при помощи меры C-value на предмет того, являются ли они многословными терминами [Frantzi et al., 2000].

Алгоритмы обучения с учителем (наивный байесовский классификатор, байесовская сеть, случайный лес, основанный на правилах классификатор OneR) были использованы в работе Ю. Мандравичкайте и Т. Крилавичюса для извлечения многокомпонетных словосочетаний на материале латышского и литовского корпусов [Mandravickaite, Krilavičius, 2017]. Также привлекались статистические меры и лингвистическая разметка. Меры ассоциации показали крайне низкую точность (0,1 и 0,2 % для латышского и литовского соответственно) При использовании алгоритма с фильтрацией точность результатов, выдаваемых методом OneR, увеличилась до 100 %, однако наибольшее значение F-меры было достигнуто при применении случайного леса с дополнительными параметрами (66,7 % для латышского и 85,6 % для литовского).

Ф. Мартинец и его коллеги также рассматривают нейронные сети в качестве методики извлечения многословных единиц в задаче информационного поиска. Результаты показали, что распознавание подобных конструкций улучшает методы информационного поиска. Был использован метод Кохонена для квантования обучающих векторов. Авторский подход сводится к тому, чтобы использовать статистические меры для нейронной сети (хи-квадрат, коэффициент Дайса, коэффициент взаимной информации и др.) [Martinez et al., 2002].

В работах П. Печины, П. Шлезингера, К. Рамиша и др. было продемонстрировано, что использование нескольких мер, а также дополнительной информации в качестве признаков для алгоритмов машинного обучения может показывать более успешные результаты [Pecina, Schlesinger, 2006; Ramisch et al., 2010].

В статье М. Карана, Дж. Шнайдера и Б. Башича [Karan et al., 2012] автоматическое извлечение словосочетаний с их последующей оцен-

кой рассматривается как задача классификации — разбиение на два класса (коллокации и неколлокации). На материале хорватского языка тестируются алгоритмы машинного обучения — метод опорных векторов, нейронные сети (для биграмм), дерево решений, наивный байесовский классификатор и логистическая регрессия. В качестве признаков используются статистические показатели (коэффициенты Дайса и взаимной информации, хи-квадрат), частотные характеристики и частеречные теги. Исследователи также дополнительно оценивают семантическую близость при помощи латентного семантического анализа и применяют полученные значения в алгоритмах. В задачи исследователей входит не только улучшение результатов по автоматическому извлечению словосочетаний, но и подбор оптимальных признаков для каждого из используемых методов машинного обучения. Авторы извлекают именные словосочетания, состоящие из двух и трех слов. В качестве базового алгоритма был выбран коэффициент взаимной информации, который показал более успешные результаты по сравнению с двумя другими мерами ассоциации. Логистическая регрессия показывает наилучшие результаты при выделении биграмм, в то время как дерево решения — для триграмм. В качестве признаков из мер ассоциации коэффициент взаимной информации в наибольшей степени улучшил алгоритмы.

Некоторые работы были нацелены на разработку программных средств для выявления устойчивых словосочетаний, т. е. изначально связаны с решением прикладных задач. В работе Ф. Смадья извлечение данных, основанное на использовании *n*-грамм и коэффициента взаимной информации, проводилось при помощи программы Xtract [Smadja, 1993]. В работе К. Рамиша и соавторов [Ramisch et al., 2010] приводится описание инструмента mwetoolkit для извлечения многословных терминов на материале разных языков при помощи статистического или комбинированного подхода. На начальном этапе конструкции отбираются по определенному шаблону (частеречным тегам, леммам, словоформам) или без шаблона как простые *n*-граммы. Далее для них вычисляются статистические показатели (частоты или меры ассоциации). Особенностью системы является также возможность вычисления метрик не только на материале имеющего пользовательского корпуса, но и на основе веб-данных (используются поисковые системы Google и Yahoo). Если есть предварительно размеченные списки коллокаций, то могут быть использованы модели машин-

ного обучения с учителем. В проведенном эксперименте использовался алгоритм опорных векторов (SVM). Результаты сравнивались с системами Yahoo и Xtract и показали, что метод опорных векторов, использующий в качестве признаков значений данных четырех мер ассоциации, является более удачным. Также дополнительно был использован порог отсечения по частоте взаимной встречаемости, равный 1 и 5. Результаты показали, что при его увеличении растет точность, но убывает полнота.

Векторное представление слов (word embeddings) стало еще одним методом машинного обучения, который нашел применение при автоматическом извлечении словосочетаний. Е. В. Еникеева и О. А. Митрофанова протестировали данный метод на материале русского языка, он показал точность до 0,9 [Enikeeva, Mitrofanova, 2017].

Эксперименты доказывают эффективность подхода, заключающегося в использовании в качестве признаков нескольких статистических мер, а также машинного обучения в целом. Это свидетельствует о том, что применение метрик в совокупности дает лучшие результаты, чем по отдельности. Отметим тот факт, что по-прежнему не существует статистического метода, который автоматически выдавал бы одинаково успешно словосочетания разных синтаксических типов, а также степеней устойчивости.

Литература

Андреев Н. Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л.: Наука, 1967.

Арапов М. В. Квантитативная лингвистика. М.: Наука, 1988.

Виноградов В. В. Современный русский язык: Грамматическое учение о слове. Вып. 1. М.: Гос. учеб.-пед. изд-во Наркомпроса РСФСР, 1938.

Головин Б. Н. Язык и статистика. М.: Просвещение, 1970.

Словарь русского языка: в 4 т. / под ред. А. П. Евгеньевой. 2-е изд., испр. и доп. М.: Русский язык, 1981–1984.

Большой толковый словарь русского языка: справочное издание / под ред. С. А. Кузнецова. СПб.: Норинт, 1998.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. (Электронная версия издания под загл. «Новый частотный словарь русской лексики»: http://dict.ruslang.ru/freq.php (дата обращения: 01.12.2018).)

Марков А. А. Об одном применении статистического метода // Известия Императорской Академии наук. Серия 6. 1916. Т. 10. № 4. С. 239.

Мартыненко Г. Я. Основы стилеметрии. Л.: Изд-во ЛГУ, 1988.

Мартыненко Г. Я., Чебанов С. В. Стилеметрия // Прикладное языкознание: учебник / отв. ред. А. С. Герд. СПб.: Изд-во СПбГУ, 1996.

Марусенко М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд-во ЛГУ, 1990.

Морозов Н. А. Лингвистические спектры: Средство для отличия плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд // Известия Отделения языка и словесности Императорской Академии наук. 1915. Т. 20. Кн. 4. С. 93–134.

Пиотровский Р. Г. Информационные измерения языка. Л.: Наука, 1968.

Хохлова М. В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia. Т. 34: Инструментарий русистики: Корпусные подходы / под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой. Хельсинки: Helsingin yliopisto, 2008. С. 343–357.

Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А. Статистический словарь языка Достоевского. М.: Языки славянских культур, 2003.

Berry-Rogghe G. The Computation of Collocations and Their Relevance in Lexical Studies // The Computer and Literary Studies / ed. by A. J. Aitken et al. Edinburgh: Edinburgh University Press, 1973.

Dias G., Guillore S., Lopes J. Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora // Actes de 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN), July 12–17, Cargèse, France. Cargèse: [s. p.], 1999. P. 333–339.

Enikeeva E., Mitrofanova O. Russian Collocation Extraction Based on Word Embeddings // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегод. междунар. конф. «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): в 2 т. / гл. ред. В. П. Селегей. Т. 1. М.: Изд-во РГГУ, 2017. С. 52–64. http://www.dialog-21.ru/media/3908/enikeevaevmitrofanovaoa.pdf (дата обращения: 01.10.2018).

Evert S. The Statistics of Word Cooccurrences: Word Pairs and Collocations: Dissertation. [Stuttgart]: Institut für maschinelle Sprachverarbeitung. University of Stuttgart, 2004. http://purl.org/stefan.evert/PUB/Evert2004phd.pdf (дата обращения: 01.10.2018).

Evert S., Krenn B. Methods for the Qualitative Evaluation of Lexical Association Measures // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse: [s. p.], 2001. P. 188–195.

Evert S., Uhrig P., Proisl T., Khokhlova M. Contrastive Collocation Analysis — a Comparison of Association Measures Across Three Different Languages Using Dependency-parsed Corpora // Proceedings of the XVIII Euralex International Congress: Lexicography in Global Contexts: Book of Abstracts / ed. by

J. Čibej, V. Gorjanc, I. Kosem, S. Krek. Ljubljana: Ljubljana University Press, 2018. P. 44–46.

Frantzi K., Ananiadou S., Mima H. Automatic Recognition of Multiword Terms: The C-value/NC-value Method // International Journal on Digital Libraries. 2000. Vol. 3. No. 2. P. 115–130.

Häcki Buhofer A., Dräger M., Meier S., Roth T. Feste Wortverbindungen des Deutschen: Kollokationenwörterbuch für den Alltag. Tübingen: Francke, 2014.

Karan M., Šnajder J., Dalbelo Bašić B. Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian // Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012) / ed. by N. Calzolari. Istanbul: European Language Resources Association (ELRA), 2012. P. 657–662.

Khokhlova M. Building a Gold Standard for a Russian Collocations Database // Proceedings of the XVIII Euralex International Congress: Lexicography in Global Contexts: Book of Abstracts / ed. by J. Čibej, V. Gorjanc, I. Kosem, S. Krek. Ljubljana: Ljubljana University Press, 2018. P. 863–869.

Klyshinsky E., Khokhlova M. In Search of Lost Collocations: Combining Measures to Reach the Top Range // Internet and Modern Society: Proceedings of the International Conference IMS-2017 (St. Petersburg; Russian Federation, 21–24 June 2017) / ed. by R. V. Bolgov, N. V. Borisov, L. V. Smorgunov, I. I. Tolstikova, V. P. Zakharov. New York: ACM Press, 2017. P. 160–163. (ACM International Conference Proceeding Series).

Kormacheva D., Pivovarova L., Kopotev M. Evaluation of Collocation Extraction Methods for the Russian Language // Quantitative Approaches to the Russian Language / ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki. S. l.: Routledge, 2018. P. 137–157.

Mandravickaitė J., Krilavičius T. Identification of Multiword Expressions for Latvian and Lithuanian: Hybrid Approach // Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) / ed. by S. Markantonatou, C. Ramisch, A. Savary, V. Vincze. Valencia: Association for Computational Linguistics, 2017. P. 97–101.

Manning Ch., Schütze H. Foundations of Statistical Natural Language Processing. Cambridge (MA): MIT Press, 1999.

Martínez F., Martín M. T., Rivas V. M., Díaz M. C., Ureña L. A. Using Neural Networks for Multiword Recognition in IR // Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge Across Boundaries: Proceedings of the 7th International ISKO Conference (Granada, Spain, July 10–13, 2002) / ed. by M. J. López-Huertas. Würzburg: Ergon, 2002. P. 559–564.

Oxford Collocations Dictionary for Students of English / ed. by C. McIntosh. Oxford: Oxford University Press, 2009.

Pecina P. A Machine Learning Approach to Multiword Expression Extraction // Proceedings of the LREC Workshop Towards a Shared Task for Multiword Ex-

pressions (MWE 2008) / ed. by N. Grégoire, S. Evert, B. Krenn. Marrakech: [s. p.], 2008. P. 54–57.

Pecina P. Lexical Association Measures: Collocation Extraction. [S. l.]: Ústav formální a aplikované lingvistiky, 2009.

Pecina P., Schlesinger P. Combining Association Measures for Collocation Extraction // Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006). Sydney: Association for Computational Linguistics, 2006. P. 651–658.

Ramisch C., Villavicencio A., Boitet Ch. Mwetoolkit: A Framework for Multiword Expression Identification // Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010). Valetta: European Language Resources Association, 2010. P. 662–669.

Smadja F. Retrieving Collocations from Text: Xtract // Computational Linguistics. 1993. Vol. 19. No. 1. P. 143–177.

Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation. Vol. 269 / ed. by M. E. Stevens, V. E. Giuliano, L. B. Heilprin. Washington: National Bureau of Standards Miscellaneous Publication, 1965.

References

Andreev N. D. 1967. *Statistical and Combinatorial Methods in Theoretical and Applied Linguistics*. Leningrad, Nauka Publ. (In Russ.)

Arapov M. V. 1988. *Quantitative Linguistics*. Moscow, Nauka Publ. (In Russ.)

Berry-Rogghe G. 1973. The Computation of Collocations and Their Relevance in Lexical Studies. *The Computer and Literary Studies*, A. J. Aitken et al. (eds.). Edinburgh, Edinburgh University Press.

Dias G., Guillore S., Lopes J. 1999. Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora. *Actes de 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN), July 12–17, Cargèse, France*. Cargèse, [s. p.], pp. 333–339.

Enikeeva E., Mitrofanova O. 2017. Russian Collocation Extraction Based on Word Embeddings. *Komp'iuternaia lingvistika i intellektual'nye tekhnologii. Po materialam ezhegod. mezhdunar. konf. «Dialog» (Moskva, 31 maia — 3 iiunia 2017 g.)*, issue 16 (23), in 2 vols., V. P Selegei (ed. in chief), vol. 1. Moscow, Izd-vo RGGU Publ., pp. 52–64. http://www.dialog-21.ru/media/3908/enikeevaevmitrofanovaoa.pdf (accessed date: 01.10.2018).

Evert S. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation*. [Stuttgart], Institut für maschinelle Sprachverarbeitung. University of Stuttgart. http://purl.org/stefan.evert/PUB/Evert2004phd.pdf (accessed date: 01.10.2018).

Evert S., Krenn B. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, [s. p.], pp. 188–195.

Evert S., Uhrig P., Proisl T., Khokhlova M. 2018. Contrastive Collocation Analysis — a Comparison of Association Measures Across Three Different Languages Using Dependency-parsed Corpora. *Proceedings of the XVIII Euralex International Congress. Lexicography in Global Contexts. Book of Abstracts*, J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.). Ljubljana, Ljubljana University Press, pp. 44–46.

Frantzi K., Ananiadou S., Mima H. 2000. Automatic Recognition of Multiword Terms. The C-value/NC-value Method. *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115–130.

Golovin B. N. 1970. *Language and Statistics*. Moscow, Prosveshchenie Publ. (In Russ.)

Häcki Buhofer A., Dräger M., Meier S., Roth T. 2014. *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen, Francke.

Karan M., Šnajder J., Dalbelo Bašić B. 2012. Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, N. Calzolari (ed.). Istanbul, European Language Resources Association (ELRA), pp. 657–662.

Khokhlova M. 2018. Building a Gold Standard for a Russian Collocations Database. *Proceedings of the XVIII Euralex International Congress. Lexicography in Global Contexts. Book of Abstracts*, J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.). Ljubljana, Ljubljana University Press, pp. 863–869.

Khokhlova M. V. 2008. Experimental Verification of Collocation Methods. *Slavica Helsingiensia*, vol. 34. Instrumentarii rusistiki. Korpusnye podkhody, A. Mustaioki, M. V. Kopotev, L. A. Birulin, E. Iu. Protasova. Helsinki, Helsingin yliopisto, pp. 343–357. (In Russ.)

Klyshinsky E., Khokhlova M. 2017. In Search of Lost Collocations: Combining Measures to Reach the Top Range. *Internet and Modern Society. Proceedings of the International Conference IMS-2017 (St. Petersburg; Russian Federation, 21–24 June 2017)*, R. V. Bolgov, N. V. Borisov, L. V. Smorgunov, I. I. Tolstikova, V. P. Zakharov (eds.). New York, ACM Press, pp. 160–163. (ACM International Conference Proceeding Series).

Kormacheva D., Pivovarova L., Kopotev M. 2018. Evaluation of Collocation Extraction Methods for the Russian Language. *Quantitative Approaches to the Russian Language*, M. Kopotev, O. Lyashevskaya, A. Mustajoki (eds.). S. l.: Routledge, pp. 137–157.

Large Explanatory Dictionary of the Russian Language: Reference Book 1998, S. A. Kuznetsov (ed.). Saint Petersburg, Norint Publ. (In Russ.)

Liashevskaia O. N., Sharov S. A. 2009. *Frequency Dictionary of the Modern Russian Language (Based on the National Corpus of the Russian Language)*. Moscow,

Azbukovnik Publ. (Electronic version under the title "New Frequency Dictionary of Russian Lexicon": http://dict.ruslang.ru/freq.php (accessed date: 01.12.2018).) (In Russ.)

Mandravickaitė J., Krilavičius T. 2017. Identification of Multiword Expressions for Latvian and Lithuanian: Hybrid Approach. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, S. Markantonatou, C. Ramisch, A. Savary, V. Vincze (eds.). Valencia, Association for Computational Linguistics, pp. 97–101.

Manning Ch., Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge (MA), MIT Press.

Markov A. A. On One Application of the Statistical Method. *Izvestiia Imperatorskoi Akademii nauk. Seriia 6*, vol. 10, no. 4, pp. 239. (In Russ.)

Martínez F., Martín M. T., Rivas V. M., Díaz M. C., Ureña L. A. 2002. Using Neural Networks for Multiword Recognition in IR. *Challenges in Knowledge Representation and Organization for the 21st Century. Integration of Knowledge Across Boundaries. Proceedings of the 7th International ISKO Conference (Granada, Spain, July 10–13, 2002)*, M. J. López-Huertas (ed.). Würzburg, Ergon, pp. 559–564.

Martynenko G. Ya. 1988. *Foundations of Stylometrics*. Leningrad, Izd-vo LGU Publ. (In Russ.)

Martynenko G. Ya., Chebanov S. V. 1996. Stylometrics. *Prikladnoe iazykoznanie Uchebnik*, A. S. Gerd (ed.). Saint Petersburg, Izd-vo SPbGU Publ. (In Russ.)

Marusenko M. A. 1990. *Attribution of Anonymous and Pseudonymous Literary Works by Pattern Recognition Methods*. Leningrad, Izd-vo LGU Publ. (In Russ.)

Morozov N. A. 1915. Linguistic Spectra. A Means for Distinguishing Plagiarism from the True Works of One or Another Well-known Author. Stylemetric Etude. *Izvestiia Otdeleniia iazyka i slovesnosti Imperatorskoi Akademii nauk*, vol. 20, book 4, pp. 93–134. (In Russ.)

Oxford Collocations Dictionary for Students of English 2009, C. McIntosh (ed.). Oxford, Oxford University Press.

Pecina P. 2008. A Machine Learning Approach to Multiword Expression Extraction. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, N. Grégoire, S. Evert, B. Krenn (eds.). Marrakech, [s. p.], pp. 54–57.

Pecina P. 2009. *Lexical Association Measures. Collocation Extraction*. [S. l.], Ústav formální a aplikované lingvistiky.

Pecina P., Schlesinger P. 2006. Combining Association Measures for Collocation Extraction. *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*. Sydney, Association for Computational Linguistics, pp. 651–658.

Piotrovskii R. G. 1968. *Informational Measurements of a Language*. Leningrad, Nauka Publ. (In Russ.)

Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation 1965, vol. 269, M. E. Stevens, V. E. Giuliano, L. B. Heilprin (eds.). Washington, National Bureau of Standards Miscellaneous Publication.

Ramisch C., Villavicencio A., Boitet Ch. 2010. Mwetoolkit. A Framework for Multiword Expression Identification. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, European Language Resources Association, pp. 662–669.

Russian Dictionary 1981–1984, in 4 vols., A. P. Evgen'eva (ed.), 2nd ed., rev. and exp. Moscow, Russkii iazyk Publ. (In Russ.)

Shaikevich A. Ia., Andriushchenko V. M., Rebetskaia N. A. 2003. *Statistical Dictionary of Dostoevsky*. Moscow, Iazyki slavianskikh kul'tur Publ. (In Russ.)

Smadja F. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, vol. 19, no. 1, pp. 143–177.

Vinogradov V. V. 1938. *Modern Russian language. Grammatical Studies about a Word*, issue 1. Moscow, Gos. ucheb.-ped. izd-vo Narkomprosa RSFSR Publ. (In Russ.)

Bill Louw

TOWARDS A THEORY OF CORPUS LINGUISTICS: PROOFS BANISH PROSCRIPTION

Abstract. The paper defends the repeated objective of B. Louw that corpora offer inductively "instrumentation for language". This paper advocates induction when relying on the reference corpus data and restricts the use of intuition to the choice of the given line for wildcarding vocabulary items. Subsequent analysis will hinge on the reading of the same line now furnished with the most frequent variables of the wildcarded forms. Any other activity is deemed to be vitiated by intuitive opacity and hence likely to leave the task of text interpretation as work in progress and hence incomplete.

Keywords. Corpus, shared logical form, intuition, subtext, lexical variables, enhanced literacy, induction.

Б. Лоу

К ТЕОРИИ КОРПУСНОЙ ЛИНГВИСТИКИ: АРГУМЕНТЫ ПРОТИВ ЦЕНЗУРЫ

Аннотация. Статья содержит аргументацию в пользу неоднократно высказанного автором утверждения, что референциальные корпусы нужно использовать индуктивно как единственно точные «измерительные приборы» в изучении языка. Автор отстаивает необходимость метода индукции в применении референциальных корпусов к интерпретации текстовых сегментов. Роль интуиции необходимо ограничить выбором текстовых сегментов для анализа. Поскольку интуиция не в состоянии получить сведения о корпусном подтексте, такие методы интерпретации не являются конечными и остаются неполными.

Ключевые слова. Корпус, логическая форма, интуиция, корпусный подтекст, лексические переменные, цифровая грамотность, индукция.

Introduction

A point which is generally agreed is that a new theory is expected, if not required, to *make a difference*. And it is upon this premise that much confusion rests.

Those who seek to protect the *status quo* are likely to attempt to label the arrival of a new theory as a methodology in order to prevent the theory from gaining *followers*. However, even where an alteration of method and its outcomes can be said to 'make a difference,' the observation of this change will always be theory-laden, i. e., that it is "impossible to observe without making any theoretical assumptions" [Mautner, 2005, p. 616]. Often, very substantial changes may be accepted as mere methodology and broadly referred to as 'progress'. *Method* in Greek means a form of change that is more momentous and far-reaching than those who use the term to sell a new theory short may have hoped. It means an 'after-path' (Gk. *meta* + *hodos*) and new methods often replace old methods quicker and more permanently than we expect. For example, after Sinclair produced dictionaries that used COBUILD methods to produce themselves, all dictionaries dropped the old method in favour of the new. This move may not, of itself, have created a new theory but it was at the very least theory-laden or, as Popper might have described it, *theory-soaked* [Magee, 1985, p. 19]. He adds:

> The method of basing general statements on accumulated observations of specific instances is known as *induction*, and is seen as the hallmark of science. In other words, the use of the inductive method is seen as the criterion of demarcation between science and non-science [Magee, 1985, p. 19].

Now, suppose we agree that COBUILD produced a dictionary in ways that were revolutionary and 'made a difference', would that qualify this achievement for the label of a new linguistic theory? If not, what type of *difference* exactly would qualify?

Well, the answer would involve taking all aspects of language to a new standpoint, rather than lexicography alone. Sinclair begins this sort of journey in a book that makes a claim: *Trust the Text* [Sinclair, 2004]. The text is trustworthy and impliedly truthful; but the title of the book is never entirely borne out. Where is there to go? He raises the issues of semantic prosody and intuition and its problem of '20–20 hindsight', but we are not shown precisely what makes text truthful enough to justify our trust. His background in philosophy produced a dictionary that would have satisfied even Wittgenstein, but as a grammarian and a philosopher he does not venture into the method of the Logical Positivists of Vienna who, prompted by Wittgenstein, saw logical form as a means for splitting the atom of language into logical form and its lexical variables as metaphysics. We never hear the suggestion that the text is trustworthy because its logic is *prior*.

And so we move to criteria that justify the use of the term theory because they make a difference. Even as I write this, a senior and respected academic, Tony McEnery, at Lancaster University tweeted:

> Are universities restricting free speech? *BBC Reality Check* looks at whether universities have banned books, speakers or changed courses[1] (@ Tony McEnery, 23 October, 2018) (emphasis added).

1. Intuitive opacity as a leap of faith

It is likely that members of the Vienna Circle had a deeper sense of method in mind than linguists. The reason for this is to be found in the fact that the term ***method*** appears in the most popular slogan of their creed:

> The meaning of a proposition is the **method** of its verification (emphasis added).

This is not a trivial point, but rather the keystone of their philosophy. We are told that as a general rule during their meetings members were asked to call out the letter M every time argument based upon Metaphysics rather than logic had been detected. They yearned for a way of separating the logic of natural language from its vocabulary, and this fact alone singles out the isolation of logical form from the sentences natural language as quasi-propositions as worthy of the title *theory*. What we are talking about here is *transcendental* truth that arises from full empiricism, as Ayer and Carnap both require. And yet, even G. J. Warnock [Warnock, 1958, p. 43] devotes a full chapter to the verification principle, but *never once mentions* the term *method* and its role in unravelling what was meant. Its meaning was, of course, quasi-mathematical. Ayer was forced to alter several versions of his ground-breaking book because he could not verify the slogan discoursally [Ayer, 1956]. All members of the Circle are now deceased. It is a sad irony today that we are now in a position to revive logical positivism by means of corpora and produce a product in the form of contexts of situation as *Sachverhlaten* that Ernest Mach would have approved of. He was the hero and mentor of the *Wienerkreis*. He predicted that verification would be sensory and in the world of our experience. His prediction was met with doubt and scorn.

[1] All of these breaches apply to me as the University of Zimbabwe, I have taught there for all of my working life.

2. Induction and the leap of disciplines: the beginning of academic tyranny

Within the short period of no more than eight years as the members of the Vienna Circle were forced to flee the from Vienna and Berlin, the combination of *collocation* and *induction* cracked the enemy codes at Bletchley Park. The collocation aspect at the Park was under the supervision of Angus McIntosh of Edinburgh University (information from the obituary for McIntosh by John Sinclair). The mathematical aspect of decription was supervised by Alan Turing.

Based upon a rumour that every code (changed at midnight every night by the Enigma machine) contained the words: Heil Hitler, *consilience of induction* [Butts (ed.), 1989, p. 8] occurred and created a short-cut for mathematics. The secrecy that surrounds induction and collocation today is now out of time under the Public Records legislation in the UK. Research in both areas is protected by international treaties for academic freedom. However, our claim for the status of theory for subtext, CPT and SP has never been stronger. At Bletchley the leap was *consilience*. Mautner puts it like this:

> Whewell called this 'consilience of inductions' and claimed that no theory which achieved it had subsequently been found to be false [Mautner, 2005, p. 654].

This strengthens our appeal for the status of Theory, because *collocation* leapt over and above disciplinary boundaries to create a short-cut for mathematics. Collocation is a hard science and akin to physics. We refuse to give up a struggle that has already been won.

3. Texts that share logical form will 'read' one another and bring lies to a CLOZE

This phenomenon makes a difference. Lies like to lurk. The truth's lexical variables are delexical and generalised. The variables in the case of a lie may even be hapaces [Louw, 2010].

'Text reads text' began with a poet's work. Henry Williams was impressed that the logical form of his hapax line:

A	**heart**	made	**bleak**	by	**sacrifice**
[A	*	made	*	by	*]

could be read and falsified by the most frequent form of the same logical form. The woman in the hapax line is a tyrant who runs an old age home with a rod of iron. The line that reads it shows with a bump rather than a jump how she came to power:

> A **situation** made **possible** by…

The search-line was: *a* + * + *made* + * + *by* + *.

Notice how the line behaves like a Socratic teacher administering a CLOZE test.

> Teacher: Brenda, please complete the line: **A situation made possible by…**
>
> Brenda: Umm… **made possible by** the fact that people hate looking after the elderly and would rather put them into a home!
>
> Teacher: Well done!

This method may be being used without acknowledgement by those who are 'Deep-Vetting' migrants. The author finds this repugnant if this is going on. He devised the method for use in literary stylistics and **not** to invade the privacy of people. Collocation, induction and relexicalisation are a sub-class of consilience.

The victimisation of academics who research truth, especially *induction* and *collocation*, would stop if place for these areas were to be put into the curriculum. Even Roman Jakobson was aware of the victimisation of elderly academics and hints at this after the mysterious death of J. R. Firth on 14 December, 1960. Jakobson wrote:

> The scholar who dared to look far ahead and to **defy** the creed of this time was **proscribed** to become **le sauvant maudit** [Jakobson, 1966, p. 251].

It is our duty as scientists to look ahead and defy.

4. Colligation is glue but collocation can jump

The relationship, if any at all, between collocation and colligation has always been mysterious. The two names seem similar. Indeed, Hoey purports to derive his apparently flawed theory from colligation, but the advertisements for his workshops seem to have more to say about the ostensibly banned collocation and seem to conflate both [Hoey, 2005].

There may have been a reason to make them sound as similar as twin ducklings; but Firth allows collocation to appear almost inside the context

of culture at the top of his diagram, along with stylistics. Colligation, according to Firth, languishes near the limits of intuitive opacity, barely above morphology.

However, once we start working with Whewell, we are forced to revise out opinions, because, although colligation remains a form of 'glue', Whewell says its role is to glue states of affairs and situational contexts of a Machian king together! So, far from distracting us from the role of collocation, Whewellian colligation [Butts (ed.), 1989, p. 27] may, to the alarm of all linguists, turn out to be a twin duckling after all. The latter-day remnant of logical positivism that still yearn for Ernst Mach's vision of telos, will, of course, be delighted by the colligation of facts. In other words, Wittgenstein's 'totality of facts, not of things' will be lovingly held together by a freshly revived colligation.

5. Residual issues: monitor corpora; the nine-word window and co-selection

Because all human beings are affected by their own intuitive opacity and its blindness, the idea that intuition can tell a monitor corpus what to monitor now needs to be abandoned. Gradually the issue of obtaining a clearer picture of the results and justification of induction will begin to replace the old notions, all of which were formed under the influence of vocabulary (metaphysics) rather than logic (obscured by opacity).

Old ideas will go to the wall, such as needing a fishing corpus to analyse fishing language. It was Firth who said that we are 'fishing in our own tank' if we used intuition as an instrument. Expansion and/or contraction of the nine-word window will need to uncover the reasons for its size and behaviour. Co-selection will need further investigation of the provenance limits set out in the excellent research of Jeremy [Clear, 1987, p. 41].

Conclusion

More than anything else, we need to abandon the fear that we are ahead of our time as researchers or in breach of the largely rumour-fed notion of what is off-limits in today's creed. Natural language is a system large enough to counteract Kurt Gödel's two theorems of incompleteness. Trust the text. It can prove anything by induction.

Science advances and we advance with it. We must not cease from exploration or we may never know the place for the first or only time.

Systems on any scale are at risk when the knowledge that supports them becomes disintegrated, or restricted to a few privileged knowers. When knowledge becomes a commodity, produced for and distributed by the market, both these things may happen. And if universities, which should be preparing us for the future, choose to emasculate the humanities in the way that has been suggested, they will happen (M. A. K. Halliday, Public Lecture, 2011. Queen's University, Belfast).

References

Ayer A. J. 1956. *The Problem of Knowledge*. Harmondsworth, Pelican Books.

Butts R. E. (ed.) 1989. *William Whewell. Theory of Scientific Method*. Indianapolis, Hackett Press.

Clear J. 1987. Computing. *Looking Up. An Account of the COBUILD Project in Lexical Computing*, J. M. Sinclair (ed.). London, Collins ELT.

Hoey M. 2005. *Lexical Priming. A New Theory of Words and Language*. London, Routledge.

Jakobson R. 1966. Henry Sweet's Paths Towards Phonamics. *In Memory of J. R. Firth*, C. E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins (eds.). London, Longmans, pp. 242–254.

Louw W. E. 2010. Collocation as Instrumentation for Meaning. A Scientific Fact. *Literary Education and Digital Learning. Methods and Technologies for Humanities Studies*, W. Van Peer, S. Zyngier, V. Viana (eds.). Hershey (PA), IGI Global.

Magee B. 1985. *Popper*, 3rd ed. London, Fontana Press.

Mautner T. 2005. *The Penguin Dictionary of Philosophy*. 2nd ed. London, Penguin Group.

Sinclair J. M. 2004. *Trust the Text. Language, Corpus and Discourse*. London, Routledge.

Warnock J. G. 1958. *English Philosophy Since 1900*. Oxford, Oxford University Press.

Marija Milojkovic

CORPUS-DERIVED SUBTEXT AND PROSPECTION IN NOVEL-WRITING: EXAMINING FAULKNER'S *ABSALOM, ABSALOM!* AND DELILLO'S *WHITE NOISE*

Abstract. The paper sheds light on the characteristics of writing styles of Faulkner and DeLillo, whose work is considered representative of modernist vs. postmodernist writing. Applying Louw's Contextual Prosodic Theory, the paper succeeds in showing how corpus-derived subtext contributes to our appreciation of nuances of meaning in the opening passages of Faulkner's *Absalom, Absalom!* and DeLillo's *White Noise*. Moreover, semantic auras of grammar strings are shown to *prospec*t the events in Faulkner's novel throughout his initial paragraph, while in DeLillo's first lines subtext does not depart from text except in the very first grammar string. This is the first discovered instance of corpus-derived prospection in novel writing.

Keywords. Corpus stylistics, Contextual Prosodic Theory, semantic prosody, corpus-derived subtext, prospection, Faulkner, DeLillo.

М. Милойкович

ПОДТЕКСТ И ПРОСПЕКЦИЯ В РОМАНИСТИКЕ ПО ДАННЫМ КОРПУСА: ИССЛЕДОВАНИЕ РОМАНОВ У. ФОЛКНЕРА «АВЕССАЛОМ! АВЕССАЛОМ!» И Д. ДЕЛИЛЛО «БЕЛЫЙ ШУМ»

Аннотация. В статье представлен опыт применения контекстуально-просодической теории Лоу для выявления особенностей авторского стиля Фолкнера и Делилло, считающихся главными представителями американского модернизма vs. постмодернизма. В работе демонстрируется, как выделение корпусного подтекста способствует обнаружению тонких нюансов в содержании начальных строк романов Фолкнера «Авессалом, Авессалом!» и Делилло «Белый шум». Более того, исследование показало, что семантические ауры грамматической организации начального фрагмента предвещают ход событий в романе Фолкнера на протяжении всего первого абзаца произведения. Что касается романа Делилло, только первая строка его начального абзаца имеет корпусный подтекст, который тем не менее также с большой точностью предвосхищает события в романе. Автор впервые предлагает эмпирические данные, которые

свидетельствуют о способности корпусного подтекста предвещать развитие событий в таком жанре, как роман.

Ключевые слова. Корпусная стилистика, контекстуально-просодическая теория, семантические просодии, корпусный подтекст, проспекция, Фолкнер, Делилло.

1. Investigating genre through corpus stylistic analysis

"One of the main stylistic applications of corpus techniques is to set patterns in a literary text against those found in general corpora of the language. This cross-referencing offers clear points of contrast between so-called everyday language and the, perhaps more nuanced, variants that mark a particular writer's craft. [A]ny discussion of foregrounding or deviation in literature is after all contingent upon, and relative to, some quantitative observation about what is 'normal' in language" [Simpson, 2014, p. 48]. Contextual Prosodic Theory (CPT), developed by Bill Louw, adopts this principle [Simpson, 2014, p. 101–102], but involves all levels of language, from the level of collocation to that of the context of culture [Louw, Milojkovic, 2016, p. 44]. Still, CPT has never been employed to research a class of texts, for instance, a canon, or genre, or a particular author, despite its potential to do so, since even though meaning is studied at the level of lexico-grammatical combinations, any text may viewed as a combination of its meanings. An empirical study of this sort would uncover specificities of an author or a genre that are invisible to the naked eye and might help to differentiate genres at the level not only of stylistics, but also, through Firth's context of situation, at the level of content.

This paper proposes to define the characteristics of modernist versus postmodernist writing. The texts chosen as representative of these genres are Faulkner's *Absalom, Absalom!* and DeLillo's *White Noise*. In accordance with the empirical nature of CPT, the opening paragraphs of these novels will be studied 'from scratch', without previous assumptions of any sort. These are the actual texts:

> From a little after two o'clock until almost sundown of the long still hot weary dead September afternoon they sat in what Miss Coldfield still called the office because her father had called it that — a dim hot airless room with the blinds all closed and fastened for forty-three summers because when she was a girl someone had believed that light and moving air carried heat and that dark was always cooler, and which (as the sun shone fuller on that side of the house) became latticed with yellow slashes full of dust motes which Quentin thought of

as being flecks of the dead old dried paint itself blown inward from the scaling blinds as wind must have blown them [Faulkner, 1986, p. 3] (120 words).

The station wagons arrived at noon, a long shining line that coursed through the west campus. In single file they eased around the orange I-beam sculpture and moved towards the dormitories. The roofs of the station wagons were loaded down with carefully secured suitcases full of light and heavy clothing; with boxes of blankets, boots and shoes, stationery and books, sheets, pillows, quilts; with rolled-up rugs and sleeping bags; with bicycles, skies, rucksacks, English and Western saddles, inflated rafts. As cars slowed to a crawl and stopped, students sprang out and raced to the rear doors to begin removing the objects inside; the stereo sets, radios, personal computers; small refrigerators and table ranges; the cartons of phonograph records and cassettes… [DeLillo, 1985, p. 3] (120 words).

Being of the same length and at the same position in the novels, these texts qualify for this kind of analysis, despite the fact that DeLillo's paragraph had to be end-stopped in mid-sentence. The following sections contain a corpus-stylistic analysis of these texts and a comparison of the findings.

2. Contextual Prosodic Theory applied to lexico-grammatical collocation

The paper will apply CPT to the analysis of these texts in order to prove its main assumptions regarding the presence of modernist/postmodernist features at the stylistic/micro-contextual level. Normally the analysis in question is conducted on the level of lexical collocation and that of lexico-grammatical collocation. Lexical collocation is here understood as defined by Sinclair — allowing up to four words to intervene between collocates. In this sense, there are two language mechanisms of importance that, we assume, participate in meaning construal. One is semantic prosody, determined by a word's frequent collocates in the reference corpus [Louw, 1993], and thus influencing our understanding through collocates that are present in the corpus but absent in the actual text. The other, in accordance with Wittgenstein's *Tractatus*, is "states of affairs," that are, according to Sinclair, visible in the corpus as co-selection: two or more words in the vicinity of one another in the author's text are co-selected in the reference corpus in order to see which states of affairs are called up in the corpus as a result.

As for lexico-grammatical collocation, Louw splits sequences in texts into lexis and grammar and studies the most frequent lexical collocates of grammar strings in the corpus. These, named by him quasi-propositional variables (QPVs), determine the string's corpus derived subtext — a deeper layer of meaning — on the basis of that same frequency in the corpus that determines the semantic prosody (SP) of a word. Naturally, lexico-grammatical collocations do not allow for collocates intervening, so Sinclair's definition here does not apply. This paper will focus on how corpus-derived subtext, and not lexical collocation, assists in textual interpretation.

3. Corpus-derived subtext in Faulkner's opening passage

The first stretch of Faulkner's text that can reasonably be investigated for its grammatical subtext is "they sat in what Miss Rosa still called the office." It contains potentially productive search lines: "they * in what" and "still *ed the *". The first consists of grammar words only; the second must begin with a lexical word, otherwise the search line "*ed *" will produce results that are too general and vague owing to its high frequency in the language.

The search line "they * in what" yielded the following quasi-propositional variables in COCA[1] (I accepted verbs in the past tense): *believed* (25), *were* (6), *lived* (3), *engaged* (2), *took* (2), *wrote, surfaced, succeeded, stood, stayed, settled, played, bounced, involved, evidenced, determined, decided, conversed, came*.

When it comes to the QPV *believe* (25 occurrences), the wider contexts in the reference corpus show that, from the narrator's point of view, the belief in question may be correct or otherwise, but it is always fairly strong. As to the QPV *were* (6), in four occurrences the purpose of the noun phrase starting with *what* is hedging, e. g. "they were in what appeared to be the heart of a sun." In the other two cases the sequence is part of a larger structure. Hedging is also the function of *in what* in the contexts of *live* (3) and *engage* (2). The case of *took* (2) should be disregarded altogether, since it is part of the phrasal verb *took in*. The QPVs occurring once, in their majority, appear in contexts of hedging.

It is significant that *believe* appears 25 times, several times more frequently than the other QPVs. If applied to the novel in question, we must

[1] The Corpus of Contemporary American English (2008–present) by Mark Davies (450 million words, 1990–present): http://corpus.byu.edu/coca/ (accessed date: 12.03.2018).

recall that it is contained in practically the first searchable grammar string of its whole text. *Prospection* has been found to operate in the first lines of short poems, where it predicts the development of thought further in the text, as well as in the first lines of students' essays [Louw, Milojkovic, 2016, p. 176–188]. Apparently, this is the first instance of such mechanism found within a novel. *Believe* may be said to foreshadow the whole behavior pattern of Miss Rosa, her daily life, an unhealthy routine strictly ruled by her worldview, her ideas of the present and, more importantly, of the past. Her habitation, as described, is a picture of her vision of what life has turned out to be. It has decayed because it was not allowed to progress naturally, while other people's houses change as life changes. Miss Rosa's whole life, as the novel will show, has been lived according to what she believes, not to what is or was. At the intuitive level it can also be said that 'they sat in what' does allude to a system of beliefs rather than to a factual state of affairs, as hedging generally is a device that is intended to be sensed by the recipient of information. On the whole, this is an instance of prospection, the subtext of the grammar string alluding to events that the reader will be informed of in the course of the novel.

The search line "still *ed the *" underlying the sequence "still called the office" is, strictly speaking, a lexico-grammatical collocation rather than a proper grammar string. On some occasions it helps to narrow down the implications of a grammar string by including a significant lexical collocate, as will be shown below (it is particularly helpful in languages with a developed inflection and flexible word order, such as Russian. This is the frequency list of this search line:

still looked the same (8);
still walked the earth (5);
still covered the ground (3);
still remembered the look (3);
still roamed the earth (3);
still roamed the streets (3);
still supported the war (3);
still walked the streets (3).

The sequence "still looked the same" (8) requires no explanation and perfectly fits its context — it is abundantly clear from the author's description that Miss Rosa's house has looked the same for several decades, without a thing missing or moved, apart from having grown older as a consequence of lack of intervention. "Still walked the earth" (5) is used in the

reference corpus in the meaning of 'still lived' of a species of people who have disappeared but at a time referred to were still present. "Still covered the ground" refers to *frost, water,* and *rubble.* "Still remembered the look" means exactly that and refers to a state of affairs in existence in the past. "Still roamed the earth" (3) is used in the same meaning as 'still walked the earth,' but refers to *dinosaurs* in two contexts and *men* to one. "Still roamed the streets" (3) contains collocates *looters*, *armed bands* and *gangs* and is reminiscent of the sequence "still supported the war" (3). "Still walked the streets" (3) describes desolate cities or districts where they used to be still populated at the time referred to by the narrator. Overall, the semantic aura of this line does not only point to Miss Rosa's not having changed, but also to the time of her life having taken place some considerable time ago.

So far corpus-derived subtext has alluded to Miss Rosa's system of beliefs and to their and herself not having changed, as well as her house, whose unchangingness is practically the subject of the passage under investigation. The next search string, "her father had *ed", again a lexico-grammatical collocation, yielded the following QPVs: ***died* (66)**, *insisted* (10), *worked* (10), *passed* (8), *wanted* (8), *asked* (7), *called* (7), *explained* (7), *lived* (7), *married* (7), *owned* (7), *purchased* (7), *warned* (7), *moved* (6), *planted* (6), *started* (6), *turned* (6), *stopped, talked, tried, used, walked* (4), *founded, forced, decided, betrayed, arrived, believed, allowed, invested, managed, molested, presented, raised, suffered, treated, vanished* (3).

The most frequent QPV, *died* (66), 6.6 times more frequent than the next two, is hugely significant to the passage. In fact, *passed* is followed by *away* in four cases out of eight and therefore the total of the QPVs with the meaning of 'die' is 70, not 66. Miss Rosa's father is indeed deceased, his voluntary starvation in wartime influencing the future of his daughter who then went to live with her sister — upon which, at her house, she endured the offence of her life, and this led her to adopt the life-in-death existence which we witness in the passage. The other QPVs in the corpus show *father* as an active agent, an authoritative figure whose actions in the past have influenced the present of the characters in the given texts. Therefore, this is yet another instance of a grammar string prospecting the content of what follows, in our case, the events in the novel that have to do with the key transitions in Miss Rosa's past.

The sequence 'with the blinds all closed' contains the grammar string "with the *s all *ed". In the four contexts yielded by COCA, four are negative (including a positive reference to tidiness in the context of a recent

death), one is positive, and one neutral because it is merely descriptive. Clearly, the proportion of negative contexts suggests that the grammar string in question carries a negative semantic prosody. It is worth noting that in the two contexts of it in the British National Corpus (BNC)[2] and one in the Google Books US corpus[3] are all negative. The impression is that of an accident, inhibition or frustration (as with collocates *folded* and *bunched* in the concordance). This is in accordance with Faulkner's use, as well as with the whole life experience of Miss Rosa.

The string "someone had *ed" in the sequence "someone had believed" calls up 927 QPVs in COCA. These are the first ten: *called* (48), *tried* (44), *turned* (43), ***died* (40),** *asked* (38), *pulled* (34), *dropped* (25), *placed* (23), *used* (22), *dumped* (18). The most frequent one, *called*, is similar to *believed* in the aspect of making a mental connection. In the remaining contexts there are no surprises. What interests us here is the fourth most frequent one, *died*. After the grammar string contained in "her father had called," death hovers once again in the passage, even if obscurely.

Predictably, the next searched string is "* was always *er" in "dark was always cooler." If we exclude *there* and pronouns (*it, she* etc), and leave only nouns in the first slot, and adjectives in the second slot and not, for instance, expressions with *her* — in other words, if we exclude cases where the grammar string could not syntactically replace the investigated one — what emerges is a positive semantic prosody, i. e. in agreement with the author's use.

As for the sequence "became latticed with yellow slashes," the string under study was "became *ed with." The list of QPVs follows, with those normally associated with difficulty or distress highlighted in bold. Variables appearing once have not been included: ***obsessed* (206)**, *involved* (153), *acquainted* (117), *fascinated* (103), *associated* (86), ***infected* (45)**, *intrigued* (35), ***disillusioned* (33)**, *preoccupied* (27), ***frustrated* (26)**, *filled* (26), *identified* (25), ***disenchanted* (23), *dissatisfied* (23), *bored* (23)**.

Of course, a close corpus study might reveal other variables with negative meaning, or even modify the superficial intuitive impression created by highlighted QPVs. Still, there is no denying that the most frequent variable,

[2] BYU-BNC (2004–present) by Mark Davies (based on the British National Corpus from Oxford University Press): http://corpus.byu.edu/bnc/ (accessed date: 12.03.2018).

[3] Google Books Corpus (2011–present) by Mark Davies (based on Google Books *n*-grams): http://googlebooks.byu.edu/ (accessed date: 12.03.2018).

obsessed (206), fits remarkably well into the context of Miss Rosa's — in fact, obsession is the word.

"Wind must have blown them" contains the grammar string "* must have * them" with two lexical variables. In COCA, one collocation was repeated: "God must have made them" (2). The other strings were too diverse for me to generalize. I resorted to the Google Books US corpus, containing 155 billion words. At least the search will make more sense on a bigger sample. Strings containing lexis other than nouns in both wildcarded slots were excluded. On the frequency list below, all collocations come from the same source, with the exception of one:

1. The **boy must have read them**, as he and the Spirit crossed the threshold (Charles Dickens, 127 times).
2. It seems my **uncle must have left them** here when he went to America… (Autobiography of Benjamin Franklin, 127 times).
3. **Julia must have brought them** for you, Hedda (Henrik Ibsen, 63 times).
4. The retreating **troops must have roused them** (Washington Irving, 52 times).
5. "God must have made them" — the first one on the frequency list to appear from diverse sources, instead of one, 51 times.
6. **Isaak must have uttered them** (Edwin Arlingtom Robinson, from the poem "Isaak and Archibald," cited in various sources, 44 times).

It is clearly line 5 that must have the greatest impact on our interpretation. All other quotes come from the same source, albeit from different editions. They must have a bearing on interpretation simply because a frequently published author, of necessity, must have influenced the language the work was written in. However, what we must rely on most is the collocation featuring the word *God*, as it comes from diverse sources. "God must have made them" is archetypal, primeval and intertextual. While the Bible may be said to underlie the development of Western civilisation, and may therefore be seen as relevant in any context within its scope, here Faulkner's choice of *wind* and *blow* may be deemed significant. God makes; wind blows. God has a plan; wind is arbitrary. Miss Rosa's existence, founded on obsession and resistance to change and resulting in fire and annihilation, has been purposeless and destructive, as nature can be when left to itself.

Faulkner's subtext is undoubtedly powerful and even if we restrict our comment only to the first QPV of every searched string, its results are nothing but telling. The subtext of the sequence "they sat in what" proves to be *believed*, which is the most frequent QPV by far. The author's expression

"still called the office," if we retain the lexical word *still*, also calls up contexts featuring a species of beings that even then were rare and are now altogether history has as its subtext "still looked the same." Another studied phrase, also containing a lexical word, *father*, has *died* as its subtext, the QVP again being by far the most frequent. "Became latticed with" refers to blinds but in the reference corpus the most frequent variable is *obsessed*. Finally, the arbitrariness of Miss Rosa's fate rather than a reasonable divine plan is seen in the sequence "wind must have blown them," whose subtext is convincingly revealed to be "God must have made them." That the influence of belief, father's death and obsession are crucial to Miss Rosa's life story is beyond doubt; that she has conserved her girly essence despite the passage of time is equally obvious. Faulkner's subtext and its powers of prospection may be said to be precise beyond belief.

If we pass the stage of verification of subtext as a valid means of investigating the deeper meaning of a text and go on to summarise its findings, we may say that, besides its accuracy in prospection, in Faulkner it reveals states rather than actions, belief, focus on death and obsession rather than doing.

4. Corpus-derived subtext in DeLillo's opening passage

As for DeLillo's passage, "*s *ed at" in "the station wagons arrived at noon," the opening phrase of the novel, called up the following most frequent combinations, divisible into four types of states of affairs (see table, Milojkovic forthcoming).

Interestingly, the novel is about a college professor who specializes in Hitler and at the end of the book enters into an armed conflict. As for the category of contexts "involving looking," the novel opens with this character observing students' goods being unloaded from estate cars.

Generally, DeLillo's text corpus-derived subtext prevalently shows action. In the first investigated grammar string it emerges as steps taken rather than physical action, although DeLillo's chosen variable does exist in the reference corpus. The next three grammar strings (*ed through the *, they *ed around the *, *ed towards the *) support the idea of physical action, as in DeLillo's text, with variables such as *walked* and *moved*. The next string's ("were *ed down with") subtext corresponds to Delillo's ("loaded"). The next grammar string yields a variety of states of affairs in the corpus; these are not easily generalisable and do not contribute to detecting a deeper meaning. Finally, the last two strings (*s *ed down to a, to * *ing the *s) carry the

Table. **The QPVs of "*s *ed at" in COCA**

No.	Active measures	Involving looking	Higher education and scientific research	Involving armed participants
1	programs aimed at 219	eyes stared at 65	researchers looked at 93	shots fired at 36
2	policies aimed at 112	eyes widened at 35	students enrolled at 78	guns pointed at 32
3	efforts aimed at 96	girls looked at 32	papers presented at 59	officers **arrived** at 28
4	measures aimed at 84	eyes looked at 26	studies conducted at 31	missiles aimed at 27
5	talks aimed at 61	boys looked at 23	samples collected at 25	attacks aimed at 16
6	interventions aimed at 59	Lucas looked at 15	studies aimed at 23	weapons aimed at 13
7	initiatives aimed at 56	eyes glared at 14	levels measured at 22	guns aimed at 12
8	activities aimed at 48	James looked at 13	courses offered at 21	officers stationed at 12
9	strategies aimed at 47	eyes peered at 12	studies looked at 20	soldiers stationed at 12
10	laws aimed at 44	fingers pointed at 12	students scored at 19	agents **arrived** at 11

subtext of reduction and of activity, respectively, and this corresponds to their respective contexts. In every string studied here, with the exception of the first, corpus-derived subtext does not reveal a deeper meaning: the author does not depart from the norm shown by the reference corpus.

5. Comparison of corpus-derived subtext and prospection in Faulkner's and DeLillo's opening passages

If we compare the subtext of Faulkner's opening passage to that of DeLillo's, it is the contrast that makes this comparison worthwhile. Faulkner's text is about mental states and a focus on the past, while DeLillo's description hinges on action, first taken generally, and then as physical movement, the latter in complete agreement with the author's context of situation. Un-

like Faulkner's text, rich in subtext, DeLillo's subtext corresponds to his text except in line 1. Still, both novels' opening lines contain prospection, verifiable by corpus data.

If these two paragraphs are representative of both these novels, Faulkner is concerned with irrealis, with atmosphere, with describing images in one's head. DeLillo, on the other hand, is a reporter, describing events as they occur, mentioning concrete actions and objects. A detailed analysis of both novels will exceed the scope of any paper of this size. Suffice it to say that Faulkner does dwell more on impressions than DeLillo, and that DeLillo does mention more concrete actions and objects than Faulkner. In the final count, Faulkner's text is fraught with deeper meaning, while DeLillo sets out to create a real world, palpable and precise, no matter how skewed and — hopefully — different from the one we live in.

Sources

DeLillo D. 1985. *White Noise*. New York: Penguin Books.
Faulkner W. 1986. *Absalom, Absalom!* New York: Vintage International.

References

Louw B., 1993. Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. *Text and Technology. In Honour of John Sinclair*, M. Baker, G. Francis, E. Tognini-Bonelli (eds.). Amsterdam, John Benjamins, pp. 152–176.

Louw B., Milojkovic M. 2016. *Corpus Stylistics as Contextual Prosodic Theory and Subtext*. Amsterdam, John Benjamins.

Simpson P. 2014. *Stylistics. A Resource Book for Students*, 2nd ed. Abingdon, Routledge.

Д. Андреева, О. А. Митрофанова

ЭКСПЕРИМЕНТЫ ПО КЛАСТЕРИЗАЦИИ РУССКОЯЗЫЧНЫХ НОВОСТНЫХ ТЕКСТОВ НА ОСНОВЕ СПИСКОВ ЛЕКСИЧЕСКИХ КОНСТРУКЦИЙ*

Аннотация. В статье представлены результаты экспериментов по кластеризации новостных сообщений из русскоязычного корпуса текстов. Эксперименты производились на корпусе новостных сообщений, собранных на основе данных из трех целевых рубрик («Политика», «Вести.Наука», «HiTech.Интернет») интернет-портала «Вести.Ru». В ходе исследования была проверена гипотеза о возможности оптимизации процедуры кластеризации путем перехода от кластеризации полных текстов к обработке их представлений в виде списков лексических конструкций. С помощью алгоритма автоматического выделения ключевых выражений RAKE из каждого текста корпуса извлекались ключевые лексические конструкции. Затем были проведены процедуры кластеризации текстов и наборов их лексических конструкций методами K-means++ и агломеративной иерархической кластеризации. Результаты подтверждают возможность использования семантической компрессии документов для их качественной кластеризации, что позволяет сократить объем обрабатываемой текстовой информации и снизить трудоемкость процедуры.

Ключевые слова. Кластеризация текстов, новостные сообщения, русскоязычный корпус, автоматическое выделение лексических конструкций, семантическая компрессия.

Darya Andreyeva, Olga A. Mitrofanova

EXPERIMENTS ON CLUSTERING RUSSIAN NEWS TEXTS BASED ON LISTS OF LEXICAL CONSTRUCTIONS

Abstract. The paper presents results of research work aimed at clustering news articles from a Russian text corpus. Experiments were carried out on the corpus of news messages collected from three target rubrics (“Politics”, “Vesti.Science”, “HiTech.Internet”) of the Inter-

* Эксперименты выполнялись в рамках проектов, поддержанных грантами РФФИ № 16-06-00529 «Разработка лингвистического комплекса для автоматического семантического анализа русскоязычных корпусов текстов с применением статистических методов» и № 17-29-09159 «Квантитативная грамматика русских предложных конструкций».

net portal “Vesti.Ru”. In course of experiments we proved a hypothesis on the possibility of improving clustering procedure by transition from clustering of full texts to processing their compressed representations as lists of lexical constructions. From each text of the corpus we extracted key lexical constructions with the help of RAKE algorithm for automatic keyword extraction. Then we performed clustering of texts and sets of their lexical constructions by means of K-means++ and agglomerative hierarchical clustering. Our results prove the possibility to use semantic compression of documents for their effective clustering, thus, reducing the amount of processed textual data and procedure complexity.

Keywords. Text clustering, news articles, Russian corpus, automatic extraction of lexical constructions, semantic compression.

1. Постановка задачи

Автоматическое определение семантической близости текстов — задача компьютерной лингвистики, решение которой зачастую переводится в плоскость поиска наилучшего алгоритма классификации или кластеризации текстов в корпусе, а также выбора оптимальных параметров анализа текстов. Эта задача важна для процедур информационного поиска, суммаризации, определения спама, в социологических исследованиях, в анализе социальных сетей, в судебной экспертизе и т. д. [Aliguliyev, 2009; Bharambe, Kale, 2011; Bekkerman et al., 2001; Peixin, Cun-Quan, 2011; Nassif, Hruschka, 2013].

Исследования показывают, что универсальные алгоритмы для работы с текстовыми данными отсутствуют, их выбор определяется структурой текстовой коллекции и задачами ее анализа. В связи с этим предпринимаются различные попытки совершенствования существующих подходов: ставится задача улучшения качества кластеризации при помощи дополнительной информации о структуре текстов и его информативно значимых фрагментах (ср.: [Гуляев, Лукашевич, 2013; Киселев и др., 2005; Baker, McCallum, 1998; Barak, 2009] и т. д.).

Так, в работе С. В. Поповой и В. В. Даниловой ставится задача улучшения качества кластеризации при помощи аннотаций и расширенного списка стоп-слов. Материалом для исследования послужили аннотации научных статей, ключевые понятия которых можно описать конструкциями вида «Adj + N». Эксперименты показали, что представление документов с использованием таких последовательностей понижает размерность пространства признаков, что сокращает временные затраты и положительно сказывается на качестве кластеризации [Попова, Данилова, 2014].

А. М. Андреев с коллегами исследует метод синтеза аннотаций для формируемых кластеров новостных текстов. Суть подхода заключается в составлении аннотаций на основе релевантных фрагментов текстов (так называемая экстрактивная суммаризация). Это особенно ценно в информационном поиске, когда пользователь, получая набор кластеров, охарактеризованный лаконичными фразами, может составить первичное представление о предметной области найденных документов [Андреев и др., 2008].

В экспериментах В. Б. Барахнина и Д. А. Ткачева предлагается метод кластеризации документов с учетом ключевых словосочетаний. Преимущество такого подхода по сравнению с простой кластеризацией заключается, в том числе, в сокращении вычислительных затрат [Барахнин, Ткачев, 2010].

В исследовании Г. Т. Букии реализован метод кластеризации и назначения меток кластеров новостных сообщений на основе методов дистрибутивной семантики. Предложенный алгоритм, опирающийся на выделение ключевых слов и ассоциированных с ними биграммных конструкций, допускает полную автоматизацию, не требует машинного обучения и не предполагает ручную разметку данных [Букия, 2019].

Помимо использования ключевых слов и конструкций, для автоматической кластеризации текстов в корпусе могут привлекаться внешние ресурсы, прежде всего формальные онтологии и тезаурусы. Так, в исследовании Э. Хото, П. Кея и М. О'Коннора такой подход позволяет учитывать семантическую близость различных терминов при помощи синсетов WordNet, что положительно сказывается на значениях близости документов, подвергаемых кластеризации [Hotho et al., 2003].

Представленный в статье проект направлен на проверку гипотезы возможности кластеризации документов на основе результатов семантической компрессии. Наше предположение основывается на том, что базовым методом семантической компрессии можно считать работу с ключевыми словами и словосочетаниями [Ягунова, 2008]; в то же время ключевые выражения чаще всего понимают как структурные единицы текста, содержащие наиболее важную информацию о его содержании и обладающие синтаксической оформленностью, семантической связностью и устойчивостью относительно исходного текста, а тем самым соотносимые с лексическими конструкциями

в понимании лингвистики конструкций [Fillmore et al., 1988; Goldberg, 1995; Stefanowitsch, Gries, 2008; Лингвистика конструкций, 2010].

Известно, что в качестве информативных фрагментов, подаваемых на вход алгоритмам кластеризации, могут использоваться аннотации, ключевые слова и словосочетания, подготовленные вручную или сгенерированные автоматически, и т. д. Наша задача состояла в том, чтобы доказать, что в процедуре кластеризации применимо представление текста в виде списка лексических конструкций, автоматически выделенных на основе лексико-грамматических шаблонов.

2. Организация эксперимента

2.1. Корпусные данные

Для эксперимента был подготовлен новостной корпус, автоматически сформированный на основе информационного интернет-портала «Вести.Ru»[1], в котором из раздела «Новости» были выбраны три целевые рубрики: «Политика», «Вести.Наука», «HiTech.Интернет» (в нашем экспериментальном корпусе им соответствуют рубрики Politics, Science, Hi-tech). Хронологические рамки создания текстов ограничены февралем 2018 года. На портале документы уже размечены по рубрикам, данная рубрикация рассматривалась как эталонная и использовалась для оценки результатов. Тексты были собраны при помощи библиотеки для парсинга HTML- и XML-документов — BeautifulSoup[2], реализованной на языке Python.

Ниже приведены примеры заголовков новостных текстов из экспериментального корпуса:

- рубрика Politics: «Против санкций и за Крым: немецкие депутаты не боятся угроз Киева»; «Посол России в США: Нарышкин посетил Вашингтон»;
- рубрика Science: «Своя правда: люди чаще отрицают факты, сказанные не на родном языке»; «Найдены гусеницы, владеющие приемами самообороны против хищников»;

[1] http://www.vesti.ru/ (дата обращения: 20.06.2018).

[2] https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (дата обращения: 20.06.2018).

- рубрика Hi-tech: «СМИ: производство iPhone X сократят вдвое из-за низкого спроса»; «На MWC показали первый по-настоящему безрамочный смартфон».

Первоначальный объем корпуса составил 695 документов (298 957 словоупотреблений). Такие параметры корпуса удовлетворяют условиям эксперимента. Затем была проведена предварительная обработка корпуса, включающая процедуры токенизации, удаления знаков препинания, перевода слов в нижний регистр, лемматизации, автоматического разрешения морфологической неоднозначности. Данные процедуры были выполнены с помощью морфологического анализатора pymorphy2[3]. В результате объем корпуса сократился до 239 000 словоупотреблений. Отступлением от стандартной методики предварительной обработки текстов стало сохранение имен собственных на латинице, например *Google, Instagram, Telegram*, поскольку они являются частотными и информативно значимыми в текстах рубрики Hi-tech.

При работе с корпусом в процессе кластеризации использовался стоп-словарь, импортированный из библиотеки для Python — stop-words[4], в котором представлены наиболее часто используемые стоп-слова для различных языков, включая русский.

2.2. Выбор алгоритмов кластеризации

Для нашего эксперимента мы остановили свой выбор на двух методах, считающихся базовыми в задаче автоматической кластеризации документов: на кластеризации методом *K*-means и агломеративной иерархической кластеризации [Aggarwal, Cheng Xiang, 2012; Manning et al., 2008; Кириченко, Герасимов, 2001; Пархоменко и др., 2017]. Существует ряд реализаций *K*-means на разных языках программирования; в наших экспериментах использовалась стабильная версия алгоритма *K*-means++ из библиотеки Scikit-Learn[5] для языка программирования Python [Arthur, Vassilvitskii, 2007]. Мы выбрали реализацию агломеративной иерархической кластеризации в библи-

[3] https://pymorphy2.readthedocs.io (дата обращения: 20.06.2018).
[4] https://pypi.org/project/stop-words/ description (дата обращения: 20.06.2018).
[5] http://scikit-learn.org (дата обращения: 20.06.2018).

отеке SciPy[6] для языка программирования Python. В качестве метода объединения в кластеры в нашем случае использовался метод Уорда.

Методы кластеризации работают с корпусом текстов, предварительно преобразованным в матрицу TF-IDF. Объем матрицы для полных текстов экспериментального корпуса составляет 19 278 терминов, а для представлений текстов в виде списков лексических конструкций — 18 791 термин. Таким образом, матрица для списков лексических конструкций меньше соответствующей матрицы для корпуса, и это означает, что свертка текстов до лексических конструкций может рассматриваться как лингвистический метод снижения размерности матрицы.

2.3. Выбор алгоритма выделения лексических конструкций

При переходе от полных текстов в корпусе к их представлению в виде списков лексических конструкций было принято решение применить алгоритм извлечения ключевых слов и словосочетаний. Такой подход позволяет не только выделить сложные конструкции, но и сохранить униграммы, которые тоже важны при определении семантической близости документов и должны учитываться в матрице TF-IDF.

Проанализировав существующие методы автоматического извлечения ключевых словосочетаний [Beliga, [2014]; Kaur, Gupta, 2010; Shivastava, Sahami, 2009; Ванюшкин, Гращенко, 2016], мы остановили свой выбор на методе, сочетающем в себе статистический и лингвистический подходы, а именно на RAKE (Rapid Keyword Extraction Algorithm)[7] [Москвина и др., 2017]. Работа алгоритма RAKE основана на предположении о том, что ключевые выражения представляют собой фразы, которые не включают в себя знаки пунктуации, служебные слова и т. д. На первом этапе обработки текста необходимо выделить фразы — кандидаты в ключевые словосочетания. Для этого текст разбивается на фрагменты по знакам препинания и стоп-словарю, содержащему служебные элементы текста. Для каждого слова на основе его общей частоты и средней длины фразы, в которой оно наблюдает-

[6] https://www.scipy.org/ (дата обращения: 20.06.2018).

[7] *Rose S. J., Cowley W. E., Crow V. L., Cramer N. O.* Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. Patent No.: US 8,131,735 B2. Date of Patent: Mar. 6, 2012. Filed: Sep. 9, 2009. http://www.google.co.ve/patents/US8131735 (дата обращения: 20.06.2018).

ся, рассчитывается вес. Вес фразы-кандидата далее определяется как сумма весов включенных в нее слов.

В экспериментах использовалась версия RAKE [Москвина и др., 2017] из библиотеки NLTK[8] для языка программирования Python, адаптированная для работы с русскоязычными корпусами текстов. Предобработка текста предполагает процедуры сегментации текста на условные слова по пробелам и установки границ условных синтаксических групп с помощью набора правил. Группы выделяются в линейных последовательностях слов с определенными грамматическими характеристиками (например, группы Adj + N, N + Adj + N (при наличии согласования), одиночные леммы и т. д.). Правила были составлены на основе данных синтаксического парсера NLTK4RUSSIAN [Москвина и др., 2016]. Специфика метрики, используемой для назначения весов, оценивает выше более сложные словосочетания, поэтому в результирующем списке фразы располагаются в порядке убывания сложности (например, пентаграммы, квадриграммы, триграммы, биграммы, униграммы); ср. фрагмент списка ненормализованных лексических конструкций для одного из новостных текстов: *стратегии роста, представителями вооруженных сил, российскими предпринимателями, парламентариями, стратегии роста, университетского сообщества, пост президента, защите прав, уголовных дел* и т. д. Вышеперечисленные особенности RAKE подтверждают пригодность данного алгоритма для решения поставленной в исследовании задачи.

2.4. Структура эксперимента

Итак, представленное в статье экспериментальное исследование включает в себя три этапа:

1) кластеризация полных текстов в корпусе, сравнение с эталонными рубриками, оценка результатов;
2) выделение ключевых слов и словосочетаний для каждого текста, лингвистическая интерпретация полученных выражений;
3) кластеризация текстов в корпусе на основе наборов ключевых выражений, сравнение с эталонными рубриками и с результатами кластеризации полных текстов, оценка результатов.

[8] http://www.nltk.org/ (дата обращения: 20.06.2018).

3. Результаты экспериментов

Мы провели несколько серий экспериментов с различными параметрами. Наилучшие результаты показали значения параметров по умолчанию (оценивалось вхождение слов в документы для расчета матрицы TF-IDF, при этом оптимальным оказалось присутствие терминов минимум в 10 % документов корпуса). Ниже приведены сравнительные результаты точности кластеризации полных текстов и кластеризации их представлений в виде списков лексических конструкций, полученные в ходе экспериментов с двумя методами, *K*-means++ и агломеративной иерархической кластеризации (см. таблицу). Точность (*P*) рассчитывалась как доля корректно кластеризованных документов по отношению ко всем документам, отнесенным к заданному кластеру. Для экспериментов с методом *K*-means++ потребовалась обработка результатов пяти итераций и расчет средней точности, поскольку данный алгоритм чувствителен к изначальному распределению центроидов кластеров: в случае близкого их распределения наблюдается значительное пересечение кластеров текстов.

Таблица. **Сравнительные результаты точности кластеризации полных текстов и кластеризации их представлений в виде списков лексических конструкций, полученные в ходе экспериментов с двумя методами, *K*-means++ и агломеративной иерархической кластеризации, *P***

Рубрика / Вид кластеризации	Politics	Science	Hi-tech	Среднее значение
Кластеризация методом K-means++				
Кластеризация полных текстов, средние значения для 5 итераций	0,9325	0,9564	0,9197	0,9362
Кластеризация представлений текстов в виде списков лексических конструкций, средние значения для 5 итераций	0,9892	0,9659	0,8926	0,9492
Агломеративная иерархическая кластеризация				
Кластеризация полных текстов	0,9480	0,9579	0,9127	0,9395
Кластеризация представлений текстов в виде списков лексических конструкций	0,9708	0,9735	0,9033	0,9493

Из таблицы видно, что в целом кластеризация документов в корпусе показала очень высокие результаты. Это объясняется аккуратной подготовкой данных для корпуса и тщательным подбором параметров экспериментов. Наилучшие результаты показывает кластеризация новостных сообщений в рубриках Politics и Science, несколько ниже показатели точности для кластеризации текстов в рубрике Hi-tech. Типы и причины ошибок работы алгоритмов обсуждаются ниже в разделе 4. Результаты для полных текстов и представлений текстов в виде списков лексических конструкций для всех рубрик сопоставимы. Точность кластеризации текстов на основе списков лексических конструкций в среднем по всем сериям экспериментов оказалась выше с расхождением на сотые доли. Это превосходство нельзя признать значительным, однако оно указывает на то, что переход от полных текстов к их представлениям в виде списков лексических конструкций не приводит к потере качества кластеризации при выигрыше в объеме обрабатываемых данных.

4. Анализ ошибочных решений

Мы решили проанализировать тексты, ошибочно кластеризованные в результате экспериментов.

Так, например, в рубрике Science и для полных текстов, и для списков лексических конструкций неправильно был интерпретирован следующий документ:

> Талантливые иностранные студенты должны оставаться, убежден **президент**. Создать максимально удобные и привлекательные условия для того, чтобы талантливая молодежь из других стран приезжала учиться в **российские** университеты, а лучшие иностранные выпускники наших вузов оставались здесь работать, призвал **президент РФ Владимир Путин**. Об этом **глава государства** заявил, оглашая ежегодное **Послание Федеральному Собранию**. «Фокус внимания должен быть на тех, кто нужен стране: на молодых, здоровых, хорошо образованных людях. Для них нужно создать упрощенную систему получения **гражданства России**», — отметил **Путин**. Образование в **России** должно быть одинаково доступно для всех граждан, а программы обучения должны учитывать реалии цифровой эпохи, уверен президент. «Равные образовательные возможности — мощный ресурс для **развития страны** и обеспечения **социальной справедливости**», — сказал **Путин**.

Эталонная рубрикация относит данный текст к рубрике Science, тем не менее в этой новостной статье присутствует достаточное количество маркеров, характерных для текстов политической направленности (*президент РФ Владимир Путин, глава государства, Послание Федеральному Собранию* и т. д. — здесь и далее в примерах такие маркеры выделены жирным шрифтом). Это объясняет решение алгоритма в пользу рубрики Politics.

Ниже приведен пример документа из эталонной рубрики Science, ошибочно отнесенного к рубрике Politics при кластеризации на основе списка лексических конструкций:

> **Президент России Владимир Путин** внес в **Госдуму законопроект**, предусматривающий уточнение целей работы Российской академии наук (РАН), ее основных задач, функций и полномочий. К целям деятельности РАН предложено отнести, в частности, прогнозирование основных направлений **научного, научно-технологического и социально-экономического развития России**, передает **РИА Новости**. Также в них предполагается включить научно-методическое руководство научной и научно-технической деятельностью научных организаций и вузов. К задачам РАН предлагается отнести проведение финансируемых из **федерального бюджета** фундаментальных научных исследований и поисковых научных исследований, в том числе в интересах **обороны страны и безопасности государства**.

Так же, как и в предыдущем случае, объяснение подобного результата связано с обилием лексических конструкций — маркеров политических новостей (*Президент России Владимир Путин, федерального бюджета, обороны страны и безопасности государства* и т. д.). Хотя для этой статьи характерна пограничность содержания (возможно, усилившаяся в результате семантической свертки), алгоритм правильно отнес полный текст к рубрике Science.

Подобные случаи ошибочных разборов наблюдаются в рубрике Hi-tech. Так, тексты, описывающие достижения робототехники, были правильно отнесены к рубрике Hi-tech методом, основанным на лексических конструкциях, в то время как кластеризация полных текстов отнесла их к рубрике Science, например:

> **Робототехническая компания Boston Dynamics**, не так давно проданная Alphabet Inc. (владеет Google) **японскому технологическому холдингу SoftBank**, продолжает сеять семена будущего конфликта человечества

с **разумными машинами**. В очередном опубликованном ролике показано, как **робот** пытается выполнить поставленную задачу, а человек ему мешает. В видео на YouTube, набравшем с момента публикации вчера вечером уже более 200 000 просмотров, человек отталкивает **манипулятор робота** SpotMini от ручки двери, которую **машина** пытается открыть. Позволив машине приоткрыть дверь, человек пытается придержать ее полузакрытой, а потом — удержать **робота** за «хвост» — прицепленный к машине сзади ремень. В процессе от **робота** отваливается часть **обшивки**, но в итоге машине позволяют **выполнить программу** и пройти через дверь.

Лексические конструкции — маркеры тематики статьи (*робот, разумная машина, выполнить программу* и т. д.) в данном случае неоднозначны и соотносимы как с новостями науки, так и с новостями из области высоких технологий.

Ниже приведен текст из эталонной рубрики «Hi-tech.Интернет» сайта «Вести.Ru», который был кластеризован вместе с текстами из рубрики «Вести.Наука» как на уровне полного текста, так и на уровне списка лексических конструкций и, таким образом, по ошибке оказался в рубрике Science нашего экспериментального корпуса:

Ученые из дочерней **компании Google Verily** разработали **алгоритм машинного обучения**, который по **сетчатке глаза** способен определить **сердечно-сосудистые заболевания**. Чтобы **обучить нейросеть**, разработчики использовали медицинские данные почти трехсот тысяч пациентов. **Сетчатка глаза** заполнена **кровеносными сосудами**, которые отражают **общее состояние организма**. **Изучая** их внешний вид с помощью **камеры** и **микроскопа**, **врачи** могут определить такие вещи как **давление**, **возраст пациента** и курит ли он. Эти показатели являются важными **предикторами сердечно-сосудистых заболеваний**. Теперь это способна сделать и **нейросеть** после **анализа скана сетчатки**. Как заявляют **ученые**, **алгоритм** от **Google** смог выдать результат с 70 % точностью, в то время как метод SCORE, основанный на **анализе крови**, дает правильные результаты в 72 % случаев, при этом он значительно более трудоемкий.

Выделенные в данном тексте лексические конструкции (*алгоритм машинного обучения, сетчатка глаза, сердечно-сосудистые заболевания, нейросеть, алгоритм Google* и т. д.) являются маркерами одновременно двух наших рубрик, Science и Hi-tech, чем и объясняется результат кластеризации.

В целом основная причина возникновения ошибок состоит в том, что при переходе от полных текстов к их представлениям в виде спи-

сков лексических конструкций: а) сохраняется неоднозначность семантического наполнения новостных текстов, б) в список конструкций не включаются определенные термины, которые влияют на правильное отнесение текста к той или иной рубрике.

5. Итоги и перспективы исследования

В нашем исследовании был предложен и протестирован метод кластеризации текстов на основе лексических конструкций. Были проведены эксперименты по автоматической кластеризации новостных сообщений в русскоязычном корпусе с использованием двух алгоритмов кластеризации (*K*-means++ и агломеративная иерархическая кластеризация) в двух режимах: на основе полных текстов и на основе их представлений в виде списков лексических конструкций, полученных в результате применения гибридного алгоритма выделения ключевых слов и словосочетаний RAKE.

Результаты кластеризации свидетельствуют о сохранении качества кластеризации текстов при переходе от полного текста к его семантической свертке на уровне лексических конструкций и о возможности сокращения трудоемкости процедуры в связи с ограничением объема текстовой информации.

Перспективы исследования связаны с тестированием различных алгоритмов кластеризации и автоматического выделения ключевых выражений, определением их пригодности к решению обсуждаемой задачи, а также с использованием в экспериментах текстов различной стилевой и жанровой принадлежности.

Литература

Андреев А. М., Березкин Д. В., Морозов В. В., Симаков К. В. Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: X Всерос. науч. конф. RCDL'2008: труды конф. Дубна: Объединенный ин-т ядерных исследований, 2008. С. 220–229.

Барахнин В. Б., Ткачев Д. А. Кластеризация текстовых документов на основе составных ключевых термов // Вестник Новосиб. гос. ун-та. Серия: Информационные технологии. 2010. Т. 8. Вып 2. С. 5–14.

Букия Г. Т. Автоматическая кластеризация новостных сообщений с опорой на ключевые слова и биграммные конструкции // Структурная и при-

кладная лингвистика: межвуз. сб. Вып. 12 / под ред. И. С. Николаева. СПб.: Изд-во СПбГУ, 2019. С. 221–233.

Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. Т. 19. С. 85–93.

Гуляев О. В., Лукашевич Н. В. Автоматическая классификация текстов на основе заголовка рубрики // Новые информационные технологии в автоматизированных системах. 2013. Т. 16. С. 238–244.

Кириченко К. М, Герасимов М. Б. Обзор методов кластеризации текстовой информации // Труды Междунар. семинара «Диалог'2001» по компьютерной лингвистике и ее приложениям: в 2 т. / отв. ред. А. С. Нариньяни. Т. 2. М.: Наука, 2001. С. 2–26.

Киселев М. В., Пивоваров В. С., Шмулевич М. М. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики // Интернет-математика 2005: Автоматическая обработка веб-данных / отв. ред. И. Сегалович, М. Маслов, Ю. Зеленков. М.: Яндекс, 2005. С. 412–435. http://elar.urfu.ru/handle/10995/1421 (дата обращения: 20.06.2018).

Лингвистика конструкций / отв. ред. Е. В. Рахилина. М.: Азбуковник, 2010.

Москвина А. Д., Митрофанова О. А., Ерофеева А. Р., Харабет Я. К. Автоматическое выделение ключевых слов и словосочетаний из русскоязычных корпусов текстов с помощью алгоритма RAKE // Труды междунар. конф. «Корпусная лингвистика — 2017» / отв. ред. В. П. Захаров. СПб.: Изд-во СПбГУ, 2017. С. 268–275.

Москвина А. Д., Орлова Д., Паничева П. В., Митрофанова О. А. Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Компьютерная лингвистика и вычислительные онтологии: труды XIX Междунар. объединенной науч. конф. «Интернет и современное общество» / отв. ред. В. П. Захаров. СПб.: Ун-т ИТМО, 2016. С. 44–54.

Пархоменко П. А., Григорьев А. А., Астраханцев Н. А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды Ин-та системного программирования РАН. 2017. Т. 29. № 2. С. 161–200.

Попова С. В., Данилова В. В. Представление документов в задаче кластеризации аннотаций научных текстов // Научно-технический вестник информационных технологий, механики и оптики. 2014. Т. 1. № 89. С. 99–107.

Ягунова Е. В. Набор опорных слов как вид свертки текста (в сопоставлении с набором ключевых слов) // Язык и речевая деятельность. 2008. [На тит. листе и обл.: 2005]. Т. 8. С. 225–235.

Aggarwal C., Zhai C. A Survey of Text Clustering Algorithms // Mining Text Data / ed. by C. Aggarwal, C. Zhai. Boston (MA): Springer, 2012. P. 77–128.

Aliguliyev R. A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization // Expert Systems with Applications. 2009. Vol. 36. No. 4. P. 7764–7772.

Arthur D., Vassilvitskii S. K-means++: The Advantages of Careful Seeding // SODA'07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms / conference chair H. Gabow. Philadelphia (PA): Society for Industrial and Applied Mathematics, 2007. P. 1027–1035.

Baker L. D., McCallum A. K. Distributional Clustering of Words for Text Classification // SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval / ed. by W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel. Melbourne; New York: ACM Press, 1998. P. 96–103.

Barak L., Dagan I., Shnarch E. Text Categorization from Category Name via Lexical Reference // Proceedings of NAACL HLT 2009: Short Papers. S. l.: ACL, 2009. P. 33–36.

Bekkerman R., El-Yaniv R., Winter Y., Tishby N. On Feature Distributional Clustering for Text Categorization // SIGIR'01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval / ed. by D. H. Kraft, W. B. Croft, D. J. Harper, J. Zobel. New York: ACM Press, 2001. P. 146–153.

Beliga S. Keyword Extraction: A Review of Methods and Approaches // Langnet: [site]. [2014]. http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf (дата обращения: 20.20.2018).

Bharambe U., Kale A. Landscape of Web Search Results Clustering Algorithms // Advances in Computing, Communication and Control: International Conference, ICAC3 2011, Mumbai, India, January 28–29, 2011: Proceedings / ed. by S. Unnikrishnan, S. Surve, D. Bhoir. Berlin; Heidelberg: Springer, 2011. P. 95–107.

Fillmore C., Kay P., O'Connor M. Regularity and Idiomaticity in Grammatical Constructions: The Case of "Let Alone" // Language. 1988. Vol. 64. No. 3. P. 501–538.

Goldberg A. Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press, 1995.

Hotho A., Staab S., Stumme G. Ontologies Improve Text Document Clustering // Proceedings of the 3rd IEEE International Conference on Data Mining. Melbourne (FL): IEEE, 2003. P. 541–544.

Kaur J., Gupta V. Effective Approaches for Extraction of Keywords // International Journal of Computer Science Issues. 2010. Vol. 7. Issue 6. P. 144–148. http://www.ijcsi.org/papers/7-6-144-148.pdf (дата обращения: 20.06.2018).

Manning C., Raghavan P., Schutze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

Nassif F., Hruschka E. Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection // Information Forensics and Security. 2013. Vol. 8. No. 1. P. 46–54.

Peixin Z., Cun-Quan Z. A New Clustering Method and Its Application in Social Networks // Pattern Recognition Letters. 2011. Vol. 32. No. 15. P. 2109–2118.

Stefanowitsch A., Gries S. Collostructions: Investigating the Interaction Between Words and Constructions // International Journal of Corpus Linguistics. 2008. Vol. 8. No. 2. P. 209–243.

Text Mining: Classification, Clustering, and Applications / ed. by A. Shivastava, M. Sahami. Boca Raton: CRC Press, 2009.

References

Aggarwal C., Zhai C. 2012. A Survey of Text Clustering Algorithms. *Mining Text Data*, C. Aggarwal, C. Zhai (eds.). Boston (MA), Springer, pp. 77–128.

Aliguliyev R. A 2009. New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization. *Expert Systems with Applications*, vol. 36, no. 4, pp. 7764–7772.

Andreev A. M., Berezkin D. V., Morozov V. V., Simakov K. V. 2008. Method for Clustering Documents in Text Collections and for Cluster Annotation Synthesis. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii. X Vseros. nauch. konf. RCDL'2008. Trudy konf.* Dubna, Ob'edinennyi in-t iadernykh issledovanii, pp. 220–229. (In Russ.)

Arthur D., Vassilvitskii S. 2007. K-means++. The Advantages of Careful Seeding. *SODA'07. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, H. Gabow (conference chair). Philadelphia (PA), Society for Industrial and Applied Mathematics, pp. 1027–1035.

Baker L. D., McCallum A. K. 1998. Distributional Clustering of Words for Text Classification. *SIGIR'98. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel (eds.). Melbourne; New York, ACM Press, pp. 96–103.

Barak L., Dagan I., Shnarch E. 2009. Text Categorization from Category Name via Lexical Reference. *Proceedings of NAACL HLT 2009. Short Papers*. S. l., ACL, pp. 33–36.

Barakhnin V. B., Tkachev D. A. 2010. Clustering of Text Documents Based on Compound Key Terms. *Vestnik Novosib. gos. un-ta. Seriia: Informatsionnye tekhnologii*, vol. 8, issue 2, pp. 5–14. (In Russ.)

Bekkerman R., El-Yaniv R., Winter Y., Tishby N. 2001. On Feature Distributional Clustering for Text Categorization. *SIGIR'01. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, D. H. Kraft, W. B. Croft, D. J. Harper, J. Zobel (eds.). New York, ACM Press, pp. 146–153.

Beliga S. [2014]. Keyword Extraction: A Review of Methods and Approaches. *Langnet*, [site]. http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf (accessed date: 20.06.2018).

Bharambe U., Kale A. 2011. Landscape of Web Search Results Clustering Algorithms. *Advances in Computing, Communication and Control. International Conference, ICAC3 2011, Mumbai, India, January 28–29, 2011. Proceedings*, S. Unnikrishnan, S. Surve, D. Bhoir (eds.). Berlin; Heidelberg, Springer, pp. 95–107.

Bukiia G. T. 2019. Automatic Clustering of News Texts Based on Keywords and Bigram Constructions. *Strukturnaia i prikladnaia lingvistika. Mezhvuz. sb.*, vol. 12, I. S. Nikolaev (ed.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 221–233. (In Russ.)

Construction Linguistics 2010, E. V. Rakhilina (ed.). Moscow, Azbukovnik Publ. (In Russ.)

Fillmore C., Kay P., O'Connor M. 1988. Regularity and Idiomaticity in Grammatical Constructions. The Case of "Let Alone". *Language*, vol. 64, no. 3, pp. 501–538.

Goldberg A. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago, University of Chicago Press.

Guliaev O. V., Lukashevich N. V. 2013. Automatic Text Classification based on Rubric Title. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh*, vol. 16, pp. 238–244. (In Russ.)

Hotho A., Staab S., Stumme G. 2003. Ontologies Improve Text Document Clustering. *Proceedings of the 3rd IEEE International Conference on Data Mining*. Melbourne (FL), IEEE, pp. 541–544.

Iagunova E. V. 2008 [on the title and the cover: 2005]. Basic Word Set as a Variety of Text Compression (Compared with Key Word Set). *Iazyk i rechevaia deiatel'nost'*, vol. 8, pp. 225–235. (In Russ.)

Kaur J., Gupta V. 2010. Effective Approaches for Extraction of Keywords. *International Journal of Computer Science Issues*, vol. 7, issue 6, pp. 144–148. http://www.ijcsi.org/papers/7-6-144-148.pdf (accessed date: 20.06.2018).

Kirichenko K. M., Gerasimov M. B. 2001. Survey of Methods for Clustering Textual Information. *Trudy Mezhdunar. seminara «Dialog'2001» po komp'iuternoi lingvistike i ee prilozheniiam*, in 2 vols., A. S. Narin'iani (ed.), vol. 2. Moscow, Nauka Publ., pp. 2–26. (In Russ.)

Kiselev M. V., Pivovarov V. C., Shmulevich M. M. 2005. Method of Text Clustering Considering Key Terms Co-occurrence, Its Application in News Flow Topic Analysis and Its Dynamics. *Internet-matematika 2005. Avtomaticheskaia obrabotka veb-dannykh*, I. Segalovich, M. Maslov, Iu. Zelenkov (eds.). Moscow, Iandeks Publ. http://elar.urfu.ru/handle/10995/1421 (accessed date: 20.06.2018).

Manning C., Raghavan P., Schutze H. 2008. *Introduction to Information Retrieval*. Cambridge, Cambridge University Press.

Moskvina A. D., Mitrofanova O. A., Erofeeva A. R., Kharabet Ia. K. 2017. Automatic Extraction of Key Words and Phrases from Russian Text Corpora by Means of RAKE Algorithm. *Trudy mezhdunar. konf. «Korpusnaia lingvistika — 2017»*, V. P. Zakharov (ed.). Saint Petersburg, Izd-vo SPbGU Publ., pp. 268–275. (In Russ.)

Moskvina A. D., Orlova D., Panicheva P. V., Mitrofanova O. A. 2016. Development of the Core for Syntactic Parser for Russian based on NLTK Libraries. *Komp'uternaia lingvistika i vychislitel'nye ontologii. Trudy XIX Mezhdunar. ob"edinennoi nauch. konf. «Internet i sovremennoe obshchestvo»*, V. P. Zakharov (ed.). Saint Petersburg, Un-t ITMO Publ., pp. 44–54. (In Russ.)

Nassif F., Hruschka E. 2013. Document Clustering for Forensic Analysis. An Approach for Improving Computer Inspection. *Information Forensics and Security*, vol. 8, no. 1, pp. 46–54.

Parkhomenko P. A., Grigor'ev A. A., Astrakhantsev N. A. 2017. Survey and Experimental Comparison of Text Clustering Methods. *Trudy In-ta sistemnogo programmirovaniia RAN*, vol. 29, no. 2, pp. 161–200. (In Russ.)

Peixin Z., Cun-Quan Z. 2011. A New Clustering Method and Its Application in Social Networks. *Pattern Recognition Letters*, vol. 32, no. 15, pp. 2109–2118.

Popova S. V., Danilova V. V. 2014. Document Representation in the Task of Clustering Research Paper Abstracts. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki*, vol. 1, no. 89, pp. 99–107. (In Russ.)

Stefanowitsch A., Gries S. 2008. Collostructions: Investigating the Interaction Between Words and Constructions. *International Journal of Corpus Linguistics*, vol. 8, no. 2, pp. 209–243.

Text Mining. Classification, Clustering, and Applications 2009, A. Shivastava, M. Sahami (eds.). Boca Raton, CRC Press.

Vaniushkin A. S., Grashchenko L. A. 2016. Methods and Algorithms for Keyword Extraction. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh*, vol. 19, pp. 85–93. (In Russ.)

И. В. Азарова, В. П. Захаров

КОРПУСНОЕ ИССЛЕДОВАНИЕ ЗНАЧЕНИЙ РУССКИХ ПРЕДЛОЖНО-ПАДЕЖНЫХ КОНСТРУКЦИЙ*

Аннотация. В статье описывается метод использования статистических параметров в случайных выборочных совокупностях контекстов, полученных в сбалансированном корпусе современных русских текстов, для ранжирования значений предложно-падежных конструкций. Описываются структуры локативных и темпоральных значений в виде ранжированной совокупности с группировкой предложных значений по параметру корпусной частотности (число единиц на миллион токенов в корпусе). Проверяется гипотеза об обратной зависимости между явно выраженным набором семантических компонентов в значении предложной конструкции и ее корпусной статистикой. В первую очередь ранги соотносятся со значениями первообразных предлогов, для которых далее приводятся синонимичные производные предлоги. Указываются селективные ограничения на употребление синтаксически главных слов («хозяев») и зависимых («слуг») в терминах базовых семантических типов wordnet-словарей.

Ключевые слова. Корпусная статистика, предложные конструкции, лексико-грамматическая характеристика, современный русский язык, основные семантические типы, локативные конструкции, темпоративные конструкции.

Irina V. Azarova, Victor P. Zakharov

CORPUS STATISTICS OF SENSES EXPRESSED IN RUSSIAN PREPOSITIONAL PHRASES

Abstract. The article describes the method of using statistical parameters in random sample context sets obtained from a balanced corpus of modern Russian texts in order to rank the meanings of prepositional constructions. The structures of locative and temporal meanings are described in terms of ranked groups of prepositional meanings calculated according to the corpus frequency (the number of evidences per million tokens in the corpus). The hypothesis of the inverse ratio between the explicit set of semantic components in the

* Исследование поддержано грантом РФФИ, проект № 17-29-09159 офи_м «Квантитативная грамматика русских предложных конструкций».

meaning of the prepositional construction and its corpus statistics is tested. First of all, ranks are assessed as regards the meanings of primary prepositions, then synonymous secondary prepositions are given. The article specifies selective rules for usage of syntactic main words ("governers") and dependent words ("governee") for prepositional constructions in terms of the basic semantic types of wordnet dictionaries.

Keywords. Corpus statistics, prepositional constructions, lexico-grammatical characteristics, modern Russian language, basic semantic types, locative constructions, temporal constructions.

Введение

Служебные слова, и в частности предлоги, остаются в лингвистической науке недоопределенными. В качестве служебных слов предлоги трактуются как слова без лексического значения, однако при сравнении их с артиклями становится ясно, что предлоги безусловно более лексически значимы, чем артикли, — правда, не во всех случаях.

Определение предлога в популярной электронной энциклопедии «Википедия»: «…служебная часть речи, обозначающая **отношение между объектом и субъектом**, выражающая синтаксическую зависимость имен существительных, местоимений, числительных от других слов в словосочетаниях и предложениях»[1] — хорошо передает частеречную принадлежность главного и зависимого слов в конструкции подчинения, которая в определении обозначается как «синтаксическая зависимость», что может быть более широким понятием. В плане содержания сужение значения до «отношения между объектом и субъектом» абсолютно необоснованно, поскольку огромное количество обстоятельственных значений несводимо к этому отношению.

В определении академической «Русской грамматики» 1980 года: «Предлог — это служебная часть речи, оформляющая подчинение одного знаменательного слова другому в словосочетании или в предложении и тем самым выражающая **отношение друг к другу** тех предметов и действий, состояний, признаков, которые этими словами называются» [Русская грамматика, 1980, т. 1, § 1655], — напротив, не указывается, какие слова могут оформляться предлогами: «подчинение одного знаменательного слова другому» предполагает, что **любое** знаменательное слово может быть оформлено предлогом, что несправедливо.

[1] https://ru.wikipedia.org/wiki/Предлог (дата публикации: 08.05.2018, дата обращения: 30.10.2018).

В толковых словарях определение предлога может быть и вовсе расплывчатым. Например: предлог — «служебное слово, выражающее **отношения между грамматически зависящими друг от друга словами** (словом и формой слова)» [Ожегов, Шведова, 1999, с. 2142]. В этом определении также нет указания на то, какие слова могут выражать «грамматическое отношение». Определение в четырехтомном толковом словаре показывает более ясное понимание роли предлогов: «…это служебные слова, которые, сочетаясь с существительными, местоимениями и числительными, **указывают на синтаксические отношения** их к другим словам» [Словарь русского языка, 1999, т. 3, с. 365]. Указание на **сочетание** абсолютно справедливо, но не указано самое важное: что перечисленные части речи стоят в **определенной падежной форме**, для которых предлог как «неизменяемая частица» служит «для более точного определения значения… падежа» [Энциклопедический словарь, 1898, с. 13].

Во всех определениях подчеркивается реляционный характер предлогов в том смысле, что они оформляют отношение обозначаемого предмета, явления, свойства к действию, предмету, свойству и т. п. Таким образом, предлог в сочетании с падежной формой имени передает существенно бóльший набор комбинаций, чем просто падежные формы или предложные конструкции. Поэтому мы будем считать предлогами **стереотипные способы уточнения падежных значений имен (чаще всего имен существительных) при выражении валентных позиций знаменательных слов (чаще всего глаголов и отглагольных дериватов) и/или различных обстоятельственных квалификаторов в предложении**. «Стереотипность» значения будет доказываться высокой частотностью выражения определенного значения в корпусе или посредством синонимии или частичной синонимии предложно-падежных конструкций.

В отношении смысловой квалификации отношений, выражаемых предлогами, отмечают [Русская грамматика, 1980, т. 1, § 1663], что исконные (первообразные) предлоги исходно имеют два типа значений: пространственные [Филипенко, 2000] и/или объектные [Солоницкий, 2003]. Пространственные значения предлогов поддерживаются значениями глагольных приставок, которые могут дублироваться предлогами формально и/или семантически, образуя своеобразную рамочную конструкцию: ***в**ойти **в** дом, **с**ъехать **с** горы, **до**йти **до** подъезда*; ***пере**йти **через** дорогу, **вы**йти **из** дома, **про**йти **по** улице*.

Типичные предложные значения можно ассоциировать с обстоятельственными значениями:

- характеристикой пространственного расположения/направления: *побывать в Питере* и *на Байкале, перепрыгнуть через канаву, приехать в Москву*;
- характеристикой момента/интервала времени, осложненного повторами: *прыгать с утра до вечера, прийти в 10.45, выехать через 3 часа, приходить по пятницам*;
- отношением сравнения и сопоставления: *ростом с меня, что-то вроде вспышки*;
- отношением совместности: *взять с собой, приехать с братом, стоять с зонтиком*;
- различными каузативными отношениями, например причины/следствия: *зарыдать от горя, приготовить к передаче, уйти в отставку, подарить в знак любви, установить для защиты, отдать в стирку*;
- условия/цели: *платье для выхода, еда для праздника, помочь при необходимости, действовать под нажимом, разбить на счастье, синяк от ушиба, слабость от недоедания, брак по расчету*.

Наряду с обстоятельственным значением столь же распространенным является выражение характеристики объектных отношений, как физических, так и интеллектуальных или эмоциональных: *забыть про отдых, говорить о свадьбе, введение в действие, вступление в брак, присоединение к юридическому лицу, соединение с группами, участие в борьбе, признание в преступлении, движение за мир, курс на разрядку, право на труд, победа над врагом, суд над грабителями, отношения между братьями.*

Очевидно, что характер предложных отношений определяется сочетанием интерпретации синтаксических отношений в словосочетании с особенностями синтагматического компонента лексических значений слов — так называемых активных или пассивных «валентностей». Причина такой «разноосновности» лежит в двойственной природе предложного значения: это конкретизация падежной формы имени [Русская грамматика, 1980, т. 1, § 1667] и формализация способа соединения синтаксических единиц для выражения определенных смысловых отношений, т. е. грамматическая конструкция. Мы будем

описывать семантику предлогов, отталкиваясь от этого противопоставления: предлоги не имеют собственных лексических значений, однако выражают смысловые связи в тексте.

1. Описание семантики предлогов в толковых словарях

Описание семантики предлогов в обычных толковых словарях ориентировано на носителей языка. Толковые словари дают лишь некоторую систематизацию способов использования предлога в речи, сопровождая ее примерами регулярного использования предлогов. Например:

В, предлог.

I. с вин. и предл. п. 1. Употр. при обозначении места, направления куда-н. или нахождения где-н. *Положить бумаги в стол. Бумаги лежат в столе. Уехать в Сибирь. Жить в Сибири. Подать заявление в университет. Учиться в университете*. 2. Употр. при обозначении явлений, представляющих собой область деятельности, состояние кого-н. *Вовлечь в работу. Весь день в работе. Впасть в сомнение. Погрузиться в глубокое раздумье*. 3. Употр. при обозначении состояния, формы, вида чего-н. *Растереть в порошок. Лекарство в порошках. Сахар в кусках. Разорвать в клочки. Все пальцы в чернилах*. 4. Употр. при указании на внешний вид кого-чего-н., на оболочку, одежду. *Завернуть в бумагу. Конфеты в обёртке. Одеться в шубу. Ходить в шубе. Нарядиться в новое платье*. 5. Употр. для указания количества каких-н. единиц, из к-рых что-н. состоит. *Комната в двадцать метров. Комедия в трёх актах. Отряд в сто человек*. 6. Употр. при обозначении момента времени. *В ночь на четверг. В один день. В прошлом году. В третьем часу*.

II. с вин. п. 1. Употр. при обозначении соотношений чисел. *В три раза меньше*. 2. Ради, для, в качестве чего-н. *Сделать что-н. в насмешку. Не в обиду будь сказано*. 3. Употр. для указания на семейное сходство с кем-н. *Весь в мать*.

III. с предл. п. 1. Употр. при обозначении расстояния от чего-н., временного отрезка. *В двух шагах от дома. В пяти минутах езды от города*. 2. Употр. при обозначении предметов, лиц, явлений, по отношению к к-рым что-н. происходит, наблюдается. *Недостатки в воспитании. Знаток в литературе. Разбираться в людях. Разница в годах*. 3. Употр. при обозначении субъекта — носителя состояния. *В юноше зреет пианист. В человеке живёт уверенность. В душе радость* [Ожегов, Шведова, 1999, с. 199–200].

В словарных статьях описывается предлог, управляющий той или иной падежной формой, при этом объединение форм (см. в примере выше) объединяет употребление одинаковых зависимых слов, стоящих в предложно-падежной форме, хотя фактически в случае предлогов, сочетающихся более чем с одной падежной формой, мы имеем дело с омонимией предлогов. Описание каждого отдельного значения начинается с «употребляется при…». При этом ни слова не говорится о типе отношений между грамматически связанными словами. Тем же грешат и многочисленные публикации на тему семантики предлогов, которые пытаются более подробно и исчерпывающе инвентаризировать все оттенки смысла, которые способен выразить тот или иной предлог. Можно сказать, что такой способ описания семантики предлогов выглядит неструктурированно, учитывая списки устойчивых словосочетаний с тем или другим предлогом, довольно обширные, но далеко не исчерпывающие. Вероятно, полный перечень выражений с характеристикой значений был бы актуальным, поскольку в задачах информационного анализа деловых текстов таковые либо вообще не используются, либо используются в очень ограниченном объеме и могут быть заданы обозримым списком.

В четырехтомном словаре русского языка указывается семантический тип не только зависимого слова, но и главного:

> **В** и **ВО**, предлог с винительным и предложным падежами.
>
> I. С винительным падежом.
>
> 1. Употребляется при обозначении предмета, места, пространства, внутрь или в пределы которого направлено действие, движение. *Пойти в театр. Внести вещи в комнату*.
>
> <…>
>
> II. С предложным падежом.
>
> 1. Употребляется при обозначении предмета, места, пространства, внутри или в пределах которого кто-, что-л. находится или что-л. происходит. *Он жил в саду во флигеле, а я в старом барском доме*. Чехов, Дом с мезонином. <…> [Словарь русского языка, 1999, т. 1, с. 132].

В первом случае главными будут обозначения действия или движения, во втором — обозначение местоположения или места события.

2. Лексико-грамматические параметры предложных конструкций

Для пространственных употреблений предлогов можно говорить о собственном лексическом значении последних, которое реализуется в определенной конструкции для обозначения противопоставления конфигураций объектов: *часы* [*висят*] *над картиной* и *часы* [*висят*] *под картиной.* Неясно, что первично: лексикализация употребления предлога или регулярное сочетание предлогов с глаголами, определяющими местоположение объектов (*находиться, лежать, стоять, висеть* и т. д.), вследствие чего закрепляется интерпретация предложно-падежного сочетания. Другие пары предлогов формируют антонимические противопоставления для глаголов движения: *войти в дом* vs. *выйти из дома*; *закатиться за угол* vs. *выкатиться из-за угла* и т. п.

Поэтому основное значение предлогов лежит в плоскости смысла синтаксических отношений между знаменательными словами. И семантическая интерпретация предлога должна рассматриваться как часть задачи семантической интерпретации синтаксических связей.

2.1. Синтаксемы в качестве схемы значений предложных конструкций

Предлог в сочетании с падежной формой образует предложно-падежную синтаксему — минимальную и нечленимую единицу синтаксиса. Место предложно-падежной синтаксемы в синтаксической структуре предложения наиболее полно описано в синтаксическом словаре Г. А. Золотовой [Золотова, 2011]. Предложно-падежные синтаксемы могут употребляться в изолированной позиции заголовка, экспозиционного предложения или драматургической ремарки, но сущность их, равно как и беспредложных синтаксем, — реализация синтаксических функций. Синтаксемы характеризуются морфологическим оформлением и служат конструктивно-смысловыми компонентами словосочетаний и предложений. И в этом смысле семантика предлогов неотрывна от категориально-семантического значения предложно-падежных синтаксем.

Подчеркнем, что предложные синтаксемы, как правило, связывают не предлог и какое-то знаменательное слово, а два знаменательных

слова, и эта связь уточняется и конкретизируется предлогом. В ряде же случаев наличие оформляющего синтаксическую связь предлога играет чисто служебную (формальную) роль (*учиться **в** университете, посылать **по** почте, ехать **по** дороге*). Как правило, в подобных случаях сочетание с предлогом можно заменить близко синонимичным беспредложным сочетанием (*посылать почтой, ехать дорогой, получать университетское образование*).

2.2. Лексико-статистические параметры предлогов в корпусе

Будем считать, что значение предлога в тексте определяется парой знаменательных слов: синтаксически главным для предлога («хозяином») и синтаксически подчиненным («слугой»). Тогда основная задача квантитативного описания семантики предлогов — это составление реестра синтаксических пар и выявление (вычисление) статистических параметров таких конструкций в корпусе.

Статистические параметры предложных конструкций мы оценивали по сбалансированному корпусу кафедры математической лингвистики [Azarova, 2008], используя менеджер корпуса Bonito [Rychlý, 2007], который позволяет получать случайные выборочные совокупности контекстов. Частотные конструкции, занимающие определенные доли в выборочной совокупности, можно экстраполировать на генеральную совокупность текстов с заданным балансом жанров.

Актуальная исследовательская задача — переход от списка конструкций к конструкционной грамматике. В нашем случае речь идет о конструкциях с предлогами. В этом случае в центре такой грамматики оказываются предложные фреймы, слоты которых — конкретные лексемы или семантические классы. Соответственно, можно сказать, что значение предлогов есть значение этих фреймов или, по-другому, что значение предлога является функцией от его окружения.

Эта грамматика должна описывать синонимию и вариативность конструкций по какому-либо признаку — например, по абстрактному семантическому значению предлога или по значениям грамматических признаков. Все параметры этой грамматики предложных конструкций должны сопровождаться статистическими данными, причем эти данные должны описывать соответствующие конструкции в разных аспектах: в плане жанрово-тематической принадлежности текстов, в плане хронологического периода и т. п.

3. Группы предложных значений

В нашем проекте предполагается переход от лексически сформулированных значений предлогов, как они даются в толковых словарях, к обобщенным значениям в терминах синтаксем [Золотова, 2011] или «семантических рубрик» [Мустайоки, 2006]. Такая единица — это обобщенное название группы предложных значений, т. е. ряда предлогов, сочетающихся с одной и той же или разными падежными формами. Объединение значений в «класс эквивалентности» не предполагает абсолютного тождества между ними в смысле двусторонней импликации — замены любой предложно-падежной конструкции на другую. Предполагается, что в группе есть высокочастотный «лидер» — доминанта группы, которая определяет максимально возможный набор синтаксических пар «хозяин — слуга». Другие члены группы лишь в некоторой степени могут использоваться для синонимических замен, поэтому синонимия в группе частичная. Периферию группы обычно представляют производные предлоги, мотивированные знаменательными частями речи.

Для характеристики семантического типа слов в синтаксической паре «хозяин — слуга» будем использовать набор базовых семантических типов, которые были получены в рамках проекта по интеграции wordnet-тезаурусов RussNet и YARN[2].

Формируя номенклатуру групп предложных значений, мы будем опираться в первую очередь на терминологию синтаксем [Золотова, 2011], внося там, где требуется, свои коррективы. В рамках данной статьи мы будем рассматривать две группы предложных значений: локативные и темпоральные.

3.1. Группа локативных значений

Исходная гипотеза нашего исследования базируется на идее маркированности грамматических значений [Якобсон, 1985], что отражается в корпусной статистике. Чем больше смысловых параметров явно выражено в значении предложной конструкции, тем менее частотным будет данное значение в корпусе текстов.

При рассмотрении базовых типов локативных значений можно выделить три базовых подтипа.

[2] http://ct05647.tmweb.ru/russnet/?page=synsets (дата обращения: 30.10.2018).

1. Собственно **локализация** (местоположение) объекта при глаголах, обозначающих действие из различных семантических групп (*вздохнуть, остановить, увидеть* и т. д.), глаголах местоположения (*находиться* и т. п.) и неглагольных «хозяевах». Наиболее частотными русскими предлогами являются *в* и *на* в сочетании с предложным падежом зависимого существительного (*в/на столе*); кроме того, для отдельной группы лексем 1-го склонения мужского рода таковой может быть форма «местного» падежа, которая внешне совпадает с формой дательного (*в лесу, на носу*). Предлог *в* имеет наибольшую частоту — около 4000 ipm[3], *на* следует за ним — 2500 ipm. Остальные предлоги, используемые для конкретизации местоположения, уступают им по корпусной статистике в 10–20 раз. Доминирование предлога *в* объясняют тем, что оформляемое существительное является трехмерным объектом и передает дополнительный семантический признак «включение», в то время как предлог *на* — признак «смежность» (contiguity) в сочетании с признаком «опора» (support) [Herskovits, 1985]. В тех случаях, когда оба предлога могут характеризовать местоположение (например, *в/на столе*), это противопоставление очевидно, однако часто только один из предлогов присоединяется к существительному с этим значением, ср.: *в городе*, но *на улице*. При этом в английском сочетаемость с аналогами *in* и *on* может быть иной: *in the street, in the tree*.

Менее частотные предлоги сочетаются с разными падежными формами (чаще всего с родительным падежом) при упорядочении по убыванию частотности: *у, около, вокруг* и *под* — 250 ± 75, *за, над* — 150 ± 50, *перед, от, до* — 50 ± 25 ipm; замыкают перечень производные предлоги, состоящие из нескольких орфографических слов: *рядом с, близко/вблизи от, вдалеке/вдали от, к востоку/западу/северу/югу от, далеко от, недалеко/неподалеку от, справа/слева от, спереди/сзади от, посредине/посреди* и т. д. — 2–10 ipm.

2. Отображение траектории **движения** при соответствующих глаголах, причем в качестве отдельных аспектов траектории частотно представлены: (а) конечная точка / цель движения (по Г. А. Золотовой «директив»), (б) начальная точка / источник движения; (в) пересекаемое пространство (по Г. А. Золотовой «транзитив»). Наиболее

[3] Ipm (instances per million words) — количество употреблений на миллион слов.

частотными являются предлоги *в* и *на*, обозначающие первый аспект, однако такое употребление существенно уступает по корпусной статистике их использованию для обозначения местоположения. В этом значении синтаксемы «директива» предлоги *в* и *на* присоединяют форму винительного падежа (*вышел на улицу / во двор*), аналогично первому типу избирательная сочетаемость с предлогом не всегда имеет логическое основание.

В этой группе есть одна важная особенность: глаголы движения и перемещения объектов регулярно сочетаются с предложными группами, в которых предлог полностью повторяет приставку глагола или передает примерно то же самое пространственное значение: *вбежать в комнату, выбежать из комнаты, наступить на щель, перепрыгнуть через щель и* т. п.

3. Относительно независимыми от семантической характеристики «хозяина» являются оппозиции **направлений ориентации**, которые в какой-то мере проявляются при характеристике значений первой группы (местоположения), но там это можно связать с семантическими типами главных слов. Зачастую эти противопоставления рассматривают как мотивирующее для использования в качестве локатива.

Базовые оппозиции:

- базовое противопоставление «включение — смежность/опора», которое обсуждалось выше: *в* vs. *на*;
- точная локализация в противоположность «пересечению двух точек» [Herskovits, 1985]: *в* и *на* vs. *у*: *в городе, у города*;
- местоположение в противоположность вертикальной локализации без учета «опоры»: *на* vs. *под/над*: *на столе, под столом, над столом*, не подразумевая совпадения точных границ объектов;
- визуальная локализация объекта, как его видит говорящий: *на/в* vs. *за/перед*: *на столе / перед столом*, не подразумевая совпадения точных границ объектов;
- дейктическая оценка близости/удаленности локализации: *близко/вблизи от, вдалеке/вдали от, далеко от, недалеко/неподалеку от* (‘not far from’);
- позиционирование относительно границ некоторой области: *посредине/посреди, на границе*;

- комбинирование системы ориентации, дейктического позиционирования или сторон света: *к востоку/западу/северу/югу от, справа/слева от*.

Характеристика местоположения нескольких объектов по отношению друг к другу используется довольно редко (суммарно 500 ipm). Используются производные предлоги, обозначающие «включение»/«невключение» объектов друг в друга (*внутри, вне*) и их наречные аналоги (*среди, снаружи*). Уточняется пространственное размещение объектов: *напротив*. Указывается пространственная область посредством границ: *от… до, между, с… до*.

3.2. Группа темпоральных значений

Л. Теньер описывал темпоральную характеристику как собственно сирконстантную, которая может быть присоединена к любой пропозиции, т. е. любому семантическому типу действия, состояния и процесса [Теньер, 1988]. Такая «неизбирательность» неизбежно должна привести к повышению корпусной частоты темпоральных значений русских предлогов, однако это не соответствует действительности. А. В. Солоницкий темпоральное значение рассматривает как первичное для первообразных предлогов наряду с локативным [Солоницкий, 2003], что не подтверждается корпусной статистикой.

Наибольшую частотность имеет конструкция предлога *в* с предложным/«местным» падежом для обозначения периода времени или момента совершения некоторого события. Зависимые существительные относятся к семантическому типу моментов/периодов времени [Всеволодова, Потапова, 1975] (*в 1995 году, в мае, в детстве* и т. д.). Корпусная частотность этой предложной конструкции (1850 ipm) в два раза меньше, чем базовая конструкция характеристики местоположения, что, вероятно, подтверждает метафорический характер употребления характеристики времени при помощи тех же самых локативных предлогов.

Следующая по частотности (985 ipm) предложная конструкция — предлог *в* в сочетании с винительным падежом. В принципе она синонимична предыдущему типу, отличием является использование абстрактных существительных в качестве зависимых (*в этот момент, в это время, в настоящее время* и т. п.), а также естественных процессов (*в бурю, в мороз* и т. п.). В качестве низкочастотных аналогов таких

конструкций выступают словосочетания (по частотности предлогов): *во время* + родительный (170), *в период* + родительный (30), *в момент* + родительный (21), *во времена* + родительный (15 ipm). Зависимые существительные обозначают действия, состояния или события: *во время беременности, в период беременности, в момент опасности* и т. п.

Следующая частотная группа темпоративов (850 ipm) указывает на конечную границу периода времени: *до* (*до пятого марта, до зимы, до 1917 года* ('before 1917'), *до поезда, до появления заболеваний* и т. п.). Логическое расширение этой конструкции — использование интервала времени *от… до* (40), *с… до* (12 ipm): *с 12 до 15 лет, с 5 часов утра до 7 часов вечера*. Первая конструкция может модифицироваться составной конструкцией: *на срок от… до* (20), *за период от… до* (10), [*с*] *продолжительностью от… до* (0,5 ipm). Продолжительность интервала времени также может задаваться точно: *за* [*день*] *до* [*встречи*]. Предлог *до* довольно широко используется в составных конструкциях: *задолго до* (10), *незадолго до* (10), *до момента* (2), *примерно до* (1 ipm). В качестве синонима используется предлог *перед* (75 ipm) в сочетании с обозначением события: *перед боем, перед встречей с министром* и т. п.

Следующие по частотности предлоги *при* (675) и *после* (500 ipm) сочетаются с обозначением событий, действий, состояний. Для первого предлога характерно указание на одновременность событий (пересечение временных точек): *при оказании помощи, при обыске на квартире*. Предлог *после* указывает на следование за событием (таксис следования): *после оказания помощи, после обыска на квартире*. Синонимичные производные предлоги: *по окончании* (6) *боевых действий*; *вслед за* (3) *заключением мира*; *по завершении* (1 ipm) *работы*.

В следующую по частотности группу входят *с* (320) и *через* ('after') (300 ipm). Предлог *с* сочетается с родительным падежом существительных, обозначающих начальный момент периода времени: *с прошлого года, с двух часов дня*. Он используется при обозначении интервалов времени, описанных выше. Его производный вариант: *с момента* (10 ipm) *ареста*. Предлог *через* присоединяет форму винительного падежа для обозначения прошедших периодов времени: *через 3 часа, через 10 лет, через несколько минут*.

Следующая группа частотности включает три предлога: *за* (260) и производные *во время* (170) и *в течение* (120 ipm). Предлог *за* со-

четается с формой винительного падежа для обозначений периодов времени: *за годы реформ, за 2 недели*. Производные *во время* и *в течение* могут заменять *за* в некоторых контекстах, при этом опускается указание на период: *в течение реформ*, *во время реформ*, *во время войны*, таким образом зависимое существительное, обозначающее деятельность, используется метонимически.

Следующий ранг частотности имеет предлог *к* (155 ipm), который обозначает момент завершения: *к 9 часам утра, к 1917 году, к вылету самолета*. Производный синоним *к моменту* (5 ipm) присоединяет обозначения событий: *к моменту ареста*.

Темпоральные предлоги с частотой менее 50 ipm обозначают позицию в интервале времени, это производные *в начале, в конце, в ходе, в период, в процессе*, которые присоединяют обозначения интервалов времени или обозначения событий: *в начале года, в начале строительства, в конце дня, в ходе строительства, в процессе строительства*.

Заключение

Предлагаемая методика описания статистических параметров предложно-падежных конструкций в сбалансированном корпусе русских текстов позволяет выстраивать иерархию усложнения подобных значений как в плане сложности семантической структуры, так и в плане селективных ограничений на семантические типы главных и зависимых компонентов.

Семантические типы «хозяев» предложных конструкций зачастую представляют собой объединение нескольких базовых типов в обширные группы. Для семантических типов «слуг» в отдельных случаях требуется дополнительная субкатегоризация по отношению к базовым семантическим типам, что наглядно демонстрирует избирательность оформления при помощи локативных частотных предлогов *в* и *на*, которые имеют национально-специфические характеристики латентной языковой классификации объектов действительности. Эти признаки являются аргументами в пользу взаимодействия лексического и грамматического в сфере оформления и интерпретации предложных конструкций в русском языке. Дополнительный грамматический компонент значения имеет падежная форма существительного, присоединяемого предлогом.

Источники

Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка: 80 000 слов и фразеологических выражений. 4-е изд., доп. М.: Азбуковник, 1999.

Словарь русского языка: в 4 т. 4-е изд., стер. / под ред. А. П. Евгеньевой, Г. А. Разумниковой. М.: Русский язык; Полиграфресурсы, 1999.

Литература

Всеволодова М. В., Потапова Г. В. Способы выражения временных отношений в современном русском языке. М.: Изд-во МГУ, 1975.

Золотова Г. А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. 4-е изд. М.: Едиториал УРСС, 2011.

Мустайоки А. Теория функционального синтаксиса: от семантических структур к языковым средствам. М.: Языки славянской культуры, 2006.

Русская грамматика: в 2 т. / гл. ред. Н. Ю. Шведова. М.: Наука, 1980.

Солоницкий А. В. Проблемы семантики русских первообразных предлогов. Владивосток: Изд-во Дальневост. гос. ун-та, 2003.

Теньер Л. Основы структурного синтаксиса / пер. с фр. И. М. Богуславского и др.; предисл., общ. ред. В. Г. Гака. М.: Прогресс, 1988.

Филипенко М. В. Проблемы описания предлогов в современных лингвистических теориях // Исследования по семантике предлогов: сб. статей / отв. ред. Д. Пайар, О. Н. Селиверстова. М.: Русские словари, 2000.

Энциклопедический словарь: в 86 т. / издатели Ф. А. Брокгауз, И. А. Ефрон. Т. 49. СПб.: Типо-лит. И. А. Ефрона, 1898.

Якобсон Р. О структуре русского глагола // Якобсон Р. Избр. работы. М.: Прогресс, 1985. С. 210–221.

Azarova I. V. RussNet as a Computer Lexicon for Russian // IIS-2008: 16th International Conference on Intelligent Information Systems. Zakopane: [s. p.], 2008. P. 447–456.

Herskovits A. Semantics and Pragmatics of Locative Expressions // Cognitive Science. 1985. Vol. 9. No. 3. P. 341–378.

Rychlý P. Manatee/Bonito — A Modular Corpus Manager // RASLAN 2007: Proceedings of Recent Advances in Slavonic Natural Language Processing / ed. by P. Sojka, A. Horák. Brno: Masaryk University, 2007. P. 65–70.

Sources

Ozhegov S. I., Shvedova N. Iu. 1999. *Explanatory Dictionary of the Russian Language*, 4th ed., suppl. Moscow, Azbukovnik Publ. (In Russ.)

Russian Dictionary 1999, in 4 vols., A. P. Evgen'eva, G. A. Razumnikova (eds.). Moscow, Russkii iazyk Publ.; Poligrafresursy Publ. (In Russ.)

References

Azarova I. V. 2008. RussNet as a Computer Lexicon for Russian. *IIS-2008. 16th International Conference on Intelligent Information Systems*. Zakopane, [s. p.], pp. 447–456.

Encyclopedic Dictionary 1898, in 86 vols., F. A. Brockhaus, I. A. Efron (publs.), vol. 49. Saint Petersburg, Tipo-lit. I. A. Efrona Publ. (In Russ.)

Filipenko M. 2000. Problems of Preposition Description in Modern Linguistic Theories. *Issledovaniia po semantike predlogov. Sb. statei*, D. Paiar, O. N. Seliverstova (eds.). Moscow, Russkie slovari Publ. (In Russ.)

Herskovits A. 1985. Semantics and Pragmatics of Locative Expressions. *Cognitive Science*, vol. 9, no. 3, pp. 341–378.

Jakobson R. 1985. On the Structure of the Russian Verb. *Jakobson R. Selected Works.* Moscow, Progress Publ., pp. 210–221. (In Russ.)

Mustaioki A. 2006. *Theory of Functional Syntax: From Semantic Structures to Linguistic Means*. Rus. Ed. Moscow, Iazyki slavianskoi kul'tury Publ. (In Russ.)

Russian Grammar 1980, in 2 vols., N. Iu. Shvedova (ed. in chief). Moscow, Nauka Publ. (In Russ.)

Rychlý P. 2007. Manatee/Bonito — A Modular Corpus Manager. *RASLAN 2007. Proceedings of Recent Advances in Slavonic Natural Language Processing*, P. Sojka, A. Horák (eds.). Brno, Masaryk University, pp. 65–70.

Solonitsky A. V. 2003. *The Problems of the Semantics of Russian Primitive Prepositions*. Vladivostok, Izd-vo Dal'nevost. gos. un-ta Publ. (In Russ.)

Tesnière L. 1988. *Éléments de syntaxe structurale*, I. M. Boguslavskii et al. (transl. from Fr.), V. G. Gak (introd., ed.). Rus. Ed. Moscow, Progress Publ. (In Russ.)

Vsevolodova M. V., Potapova G. V. 1975. *Ways of Expressing Temporal Relations in Modern Russian*. Moscow, Izd-vo MGU Publ. (In Russ.)

Zolotova G. A. 2011. *Syntactical Dictionary. A Set of Elementary Units of Russian Syntax*, 4th ed. Moscow, Editorial URSS Publ. (In Russ.)

СВЕДЕНИЯ ОБ АВТОРАХ

Азарова Ирина Владимировна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: i.azarova@spbu.ru.

Алексеева Елена Леонидовна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: el.alexeeva@gmail.com.

Андреева Дарья — магистрант, Центр наук о сознании и мозге CIMeC, Университет Тренто, Тренто, Италия. E-mail: darya.andreyeva@studenti.unitn.it.

Захаров Виктор Павлович — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: v.zakharov@spbu.ru.

Лоу Билл — PhD, иностранный приглашенный профессор, Университет Ковентри, Ковентри, Великобритания. E-mail: godelkreiss@gmail.com.

Мартыненко Григорий Яковлевич — доктор филологических наук, профессор. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: g.martynenko@gmail.com.

Милойкович Мария — MA, старший преподаватель. Кафедра английского языка, Белградский университет, Белград, Сербия. E-mail: marija.milojkovic@fil.bg.ac.rs.

Митрофанова Ольга Александровна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: o.mitrofanova@spbu.ru.

Николаев Илья Сергеевич — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: i.s.nikolaev@spbu.ru.

Тискин Даниил Борисович — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: d.tiskin@spbu.ru.

Хохлова Мария Владимировна — кандидат филологических наук, доцент. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: m.khokhlova@spbu.ru.

Чебанов Сергей Викторович — доктор филологических наук, профессор. Кафедра математической лингвистики, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. E-mail: s.chebanov@gmail.com.

INFORMATION ABOUT AUTHORS

Azarova Irina V. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: i.azarova@spbu.ru.

Alekseeva Elena L. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: el.alexeeva@gmail.com.

Andreyeva Darya — M. Sc. Student, CIMeC Center for Mind/Brain Sciences, University of Trento, Trento, Italy. E-mail: darya.andreyeva@studenti.unitn.it.

Chebanov Sergey V. — Doctor of Philology, Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: s.chebanov@gmail.com.

Khokhlova Maria V. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: m.khokhlova@spbu.ru.

Louw Bill — PhD, International Visiting Professor, Coventry University, Coventry, United Kingdom. E-mail: godelkreiss@gmail.com.

Martynenko Gregory Ya. — Doctor of Philology, Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: g.martynenko@gmail.com.

Mitrofanova Olga A. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: o.mitrofanova@spbu.ru.

Milojkovic Marija — MA, Senior Language Instructor, English Department, Faculty of Philology, University of Belgrade, Belgrade, Serbia. E-mail: marija.milojkovic@fil.bg.ac.rs.

Nikolaev Ilya S. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: i.s.nikolaev@spbu.ru.

Tiskin Daniil B. — Candidate of Philology, Assistant Professor. Department of of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: d.tiskin@spbu.ru.

Zakharov Victor P. — Candidate of Philology, Associate Professor. Department of Mathematical Linguistics, Saint Petersburg State University, Saint Petersburg, Russia. E-mail: v.zakharov@spbu.ru.

СОДЕРЖАНИЕ

CONTENTS

Научное издание

СТРУКТУРНАЯ И ПРИКЛАДНАЯ
ЛИНГВИСТИКА

Межвузовский сборник

Выпуск 13

Редактор *О. В. Косенко*
Корректор *Н. Е. Абарникова*
Компьютерная верстка *Е. М. Воронковой*

Подписано в печать 05.11.2019. Формат 60×84 $^{1}/_{16}$.
Усл. печ. л. 10,3. Планируемый тираж 300 экз. (1-й завод — 70 экз.). Заказ №

Издательство Санкт-Петербургского университета.
199004, С.-Петербург, В.О., 6-я линия, д. 11.
Тел./факс +7(812)328-44-22
publishing@spbu.ru

publishing.spbu.ru

Типография Издательства СПбГУ. 199034, С.-Петербург, Менделеевская линия, д. 5.