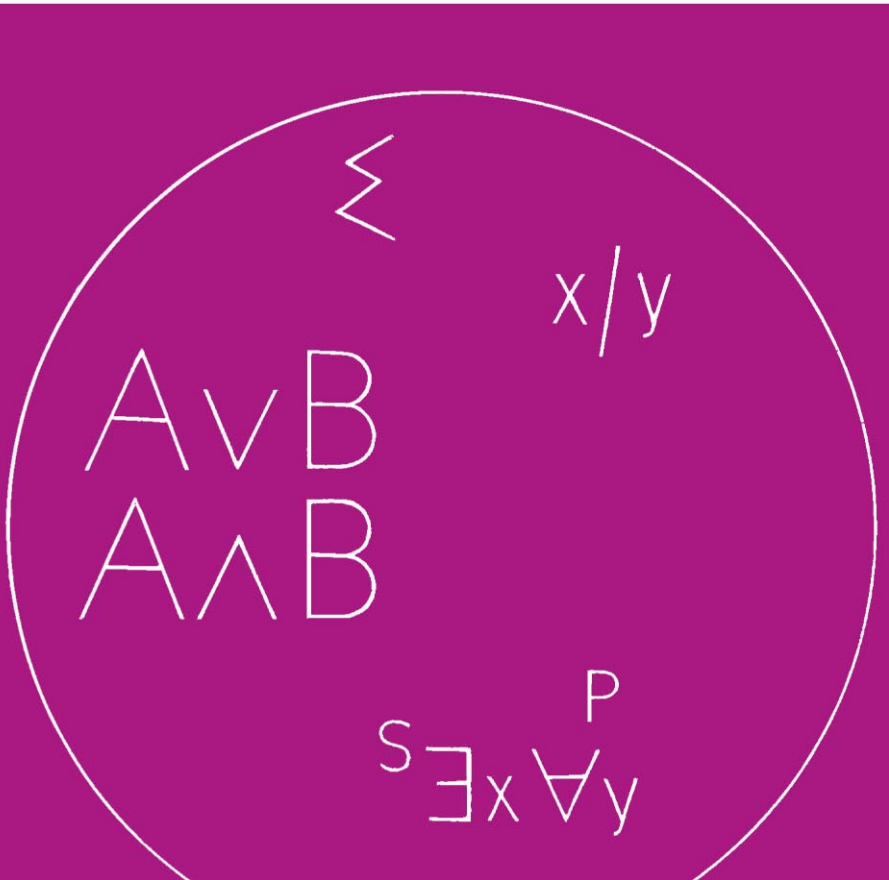
ISSN 0202–2400

Структурная и прикладная лингвистика

13



УДК 80+618.31
ББК 81.1
С83

Редакционная коллегия: д-р филол. наук, проф. *Л. Н. Беляева* (Рос. гос. пед. ун-т им. А. И. Герцена), PhD, науч. сотр. *В. Бенко* (Ин-т языкознания им. Л. Штура Словац. акад. наук), чл. Мекс. акад. наук, PhD, проф. *А. Гельбух* (Нац. политехн. ин-т Мехико), PhD *Дао Хонг Тху* (Вьетн. ассоц. по лингвистике), канд. филол. наук, доц. *В. П. Захаров* (С.-Петерб. гос. ун-т), д-р филол. наук, проф. *А. В. Колмогорова* (Сиб. фед. ун-т), PhD habil., доц. *М. В. Копотев* (Ун-т Хельсинки), д-р техн. наук, проф. *Н. Н. Леонтьева* (Рос. гос. гуманит. ун-т), д-р филол. наук, проф. *Г. Я. Мартыненко* (С.-Петерб. гос. ун-т), д-р филол наук, проф. *М. А. Марусенко* (С.-Петерб. гос. ун-т), канд. филол. наук, доц. *И. С. Николаев* (гл. ред., С.-Петерб. гос. ун-т), Doc. RNDr. канд. наук *В. Петкевич* (Карлов ун-т), PhD, науч. сотр. *О. Скривнер* (Индиан. ун-т), д-р филол. наук, проф. *М. К. Тимофеева* (Новосиб. гос. ун-т), д-р филол. наук, проф. *С. В. Чебанов* (С.-Петерб. гос. ун-т), д-р филол. наук, проф. *А. Я. Шайкевич* (Ин-т рус. языка Рос. акад. наук), канд. физ.-мат. наук, проф. *С. А. Шаров* (Ун-т Лидса), д-р филол наук, гл. науч. сотр. *С. Д. Шелов* (Ин-т рус. языка Рос. акад. наук), д-р филол. наук, проф. *Тао Юань* (Шэньсийк. пед. ун-т)

Рецензенты: д-р филол. наук, проф. *В. И. Шадрин* (С.-Петерб. гос. ун-т), канд. филол. наук, доц. *О. Н. Камшилова* (Рос. гос. пед. ун-т им. А. И. Герцена)

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

© С.-Петербургский государственный университет, 2019

СОДЕРЖАНИЕ

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)».

© Г. Я. Мартыненко

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).