

ОРИГИНАЛЬНАЯ СТАТЬЯ

УДК 81'23

doi: 10.26907/2541-7738.2023.3.43-54

ПСИХОЛИНГВИСТИКА, КОРПУСНАЯ ЛИНГВИСТИКА И ТЕКСТОВАЯ АНАЛИТИКА: ВОЗМОЖНОСТИ ВЗАИМОДЕЙСТВИЯ

И.В. Привалова

Казанский (Приволжский) федеральный университет, г. Казань, 420008, Россия

Аннотация

В статье рассмотрены потенциальные возможности взаимодействия психолингвистической науки, корпусной лингвистики и интеллектуального анализа текста. Отмечено, что достижения психолингвистики должны приниматься во внимание при разработке приемов обработки больших данных (например, языковых корпусов). В новых реалиях цифровизации происходит переосмысление текста как объединенных единым смыслом множества языковых единиц. Вносятся уточнения в психолингвистику текста, поскольку при глубинном анализе появляется новый взгляд на роли отправителя и получателя и на первый план выходят способы качественной переработки больших объемов информации. Без исследования механизмов речепорождения и речевосприятия невозможно дальнейшее совершенствование нейронных сетей и электронных переводчиков. Обсуждаются теоретические основания психолингвистики, которые могут быть использованы для подготовки кадров в области текстовой аналитики. В качестве примера рассмотрена встроенность таких дисциплин, как «Психолингвистика» и «Психолингвистические методы изучения языка и текста», в психолого-лингвистический модуль программы обучения магистрантов по профилю «Текстовая аналитика в образовании и науке», которая реализуется на базе Казанского федерального университета. Описана психолингвистическая составляющая в дисциплине «Корпусная лингвистика» для магистрантов.

Ключевые слова: психолингвистика, корпусная лингвистика, текстовая аналитика, большие данные, магистерская программа

Введение

Междисциплинарный подход в современной науке является залогом успешного исследования. Степень интеграции многих дисциплин настолько высока, что решение конкретной задачи невозможно в рамках одной научной области. Между такими дисциплинами, как психолингвистика, корпусная лингвистика и текстовая аналитика, существует неопровержимая связь. С известной степенью упрощения можно определить психолингвистику как науку о механизмах порождения языка, а корпусную лингвистику – как науку о языковых больших данных. Эти научные области сфокусированы на трех основных темах: язык, человек и информация. Текстовая аналитика, как отдельная область научного знания, предлагает новый формат работы с языковым материалом, а именно разработку автоматизированных процессов извлечения определенных показателей из больших баз данных неструктурированных текстов. Текстовая аналитика представляет собой актуальное направление в современной филологии, которое возникло благодаря развитию новейших технологий. Последние десятилетия характеризуются двумя процессами –

беспрецедентным ростом информации и бурным развитием цифровизации. Данные процессы взаимосвязаны, поскольку новые программные продукты и платформы позволяют генерировать большое количество сообщений и разнообразить их виды, что, в свою очередь, порождает потребность в разработке соответствующих механизмов обработки и хранения больших объемов данных.

С появлением больших данных произошло переосмысление понятий «текст» и «обработка языковых данных». К простейшим процедурам текстовой аналитики докомпьютерной эпохи можно отнести картографирование языковых единиц, нацеленное на описание отдельного явления (например, когда вручную составлялась картотека лексем, входящих в одну лексико-семантическую группу или лексико-семантическое поле). Источниками для поиска необходимых единиц выступали художественные произведения, учебные тексты, лекции и другие образцы устной или письменной речи различных жанров. При таком подходе определение текста как совокупности языковых единиц, объединенных общим смыслом, остается актуальным.

С развитием специальных компьютерных программ появилась возможность оцифровывать значительные массивы текстов, такие как, например, произведения одного писателя, и создавать индивидуальные словари отдельной языковой личности. В этом случае по-прежнему оставалось исчерпывающим понимание текстов как сверхфразовых единств или комплексных синтаксических целых. Необходимость переосмысления понятия «текст» отсутствовала и тогда, когда информация появилась на платформах Web 1.0. Способ презентации информации поменялся с печатного на электронный, что не потребовало новых аналитических подходов. Однако анализ текстовых массивов данных, размещенных на платформах Web 2.0, уже ставит под сомнение критерий связанности текста. Дело в том, что самогенерируемый контент начинает жить по своим, особым правилам, и текст уже становится совокупностью информации. Наряду со связанными по смыслу текстами появляются оформленные семиотическим кодом языковых знаков данные, которые подразделяют на структурированную, неструктурированную и квазиструктурированную информацию. Лингвистические корпуса, например Национальный корпус русского языка (НКРЯ), представляют собой структурированную лингвистическую информацию. К материалам этих корпусов обращаются психолингвисты, специалисты в области когнитивной лингвистики, исследующие, например, концепты и фреймы. Структурированность – один из маркеров антропоцентричности больших данных. Другим маркером является то, что практически все большие данные представляют собой продукт человеческой деятельности и связаны с процессами речепорождения и речевосприятия, которые исследуются психолингвистами. Как видим, существуют точки соприкосновения психолингвистики, корпусной лингвистики и текстовой аналитики, позволяющие рассмотреть возможности взаимодействия этих научных направлений. Данные точки взаимодействия являются предметом нашего исследования, объект же – антропоцентрический характер научных дисциплин, которые имеют прямое отношение к языковой деятельности.

Теоретические основания

Взаимодействие психолингвистики, корпусной лингвистики и приемов глубинного анализа текста открывает новые перспективы и позволяет решать научные задачи, которые казались неразрешимыми еще несколько лет тому назад. Теоретические положения психолингвистики как науки о механизмах речепорождения и речевосприятия актуальны для преподавания естественных языков, для перевода текстов, созданных в разных языковых системах. Специфику

порождения речи необходимо учитывать, когда стоит задача перекодирования языкового материала из одной формы существования в другую. Например, с учетом психолингвистических особенностей осуществляется разработка компьютерных программ, которые позволяют эффективно перевести устную речь в письменную. В повышении точности голосового набора текста с помощью программ транскрибирования речи заинтересованы специалисты, работающие во многих отраслях. Например, особой популярностью пользуется подобное программное обеспечение среди профессиональных письменных переводчиков. До сих пор не удается достичь абсолютной корректности в наборе текста через диктовку, так как в работе программ распознавания речи трудно нивелировать индивидуальные особенности говорящего, а также условия, в которых происходит запись текста. А.Р. Биктимиров, Д.Ю. Груздев анализируют основные факторы лингвистического и программно-технического характера, которые оказывают влияние на эффективность преобразования компьютером естественной речи в письменный текст. Отмечается, что повышение продуктивности набора текста через диктовку зависит от точности работы программы распознавания речи [1, с. 73].

Объектом изучения корпусной лингвистики являются языковые корпуса как совокупности Big Data. Цифровизация языка обеспечивает централизацию информации во многих формах, одной из которых и являются языковые корпуса. В них представлен практически весь спектр языкового материала – образцы разговорной речи, литературные произведения, публицистика, отрывки электронной коммуникации в диахронической или синхронической перспективе. Так, НКРЯ предоставляет возможность работать с данными из диалектологического, параллельного, мультимедийного, акцентологического и других подкорпусов. В большинстве отечественных языковедческих исследований отмечается обращение к данным НКРЯ и другим языковым корпусам. Кроме того, значимость языковых корпусов состоит в том, что они выполняют кумулятивную функцию и способствуют сохранению и передаче языкового материала будущим поколениям.

К разным типам корпусов в лингвистике относятся текстовые, речевые, лексические, межязыковые, а также одноязычные, двуязычные и даже трехязычные корпуса, которые широко используются в качестве электронных словарей. Языковые корпуса незаменимы при решении практических задач в процессе изучения иностранного языка, поскольку в них содержатся примеры, иллюстрирующие функционирование отдельной языковой единицы в различных стилевых и жанровых условиях [2–4]. Существуют работы, посвященные, например, применению материалов из языковых корпусов для анализа грамматики [5], стилистических норм [6] или словарных единиц [7]. Оригинальный подход представлен при использовании языковых корпусов в социолингвистике применительно к данным на разных языках, полученным через веб-сайты, сайты социальных сетей и блоги [8].

Отметим важность статистических методов для текстовой аналитики, которые существенно помогают в оценке репрезентативности языкового материала и достоверности полученных результатов. Кроме того, они оказались весьма эффективными при построении статистических моделей и прогнозов. Согласно современным определениям, корпусная лингвистика – это мощная количественная методология, которая в значительной степени опирается на данные частотности и статистические процедуры. Трудно говорить о корпусной лингвистике, не упоминая статистические показатели, основанные на частоте и распределении [9].

Необходимость статистического анализа для обработки больших данных языковых корпусов впервые была обоснована Майклом Оуксом в его книге “Statistics for Corpus Linguistics” [10]. Нельзя не упомянуть издание “Quantitative

Corpus Linguistics with R. A Practical Introduction” [11], в котором описано, как применять язык программирования R для интеллектуального анализа текста. Монография Вацлава Бжезины “Statistics in Corpus Linguistics: A Practical Guide” [12] не только предлагает теоретическое обоснование для статистического анализа, но и объясняет, как на практике применяются некоторые методы исследования. Приводятся примеры решения конкретных задач и материал для самостоятельного изучения.

Монография Шона Уоллиса “Statistics in Corpus Linguistics: A New Approach” [13] представляет новый подход к интеграции статистических методов исследования, текстовой аналитики и больших баз данных. Впервые удачно терминологизированы такие понятия, как «корпус» и «набор данных». Исследователь подчеркивает, что в самом общем смысле корпуса – это просто наборы языковых данных, собранные для того, чтобы сделать их доступными для исследовательских целей. Наборы данных, извлеченные из корпусов, получены не в контролируемых, а в «натуралистических» или «экологических» условиях [13]. Ш. Уоллис обращает внимание на то, что характеристики языкового корпуса напрямую зависят от характеристик текста. Корпуса письменных данных не должны оформляться как корпуса устных данных, в которые необходимо включать транскрипционные записи. Лингвисты, занимающиеся проблемами мультимодальности, также указывают на то, что в речевых корпусах целесообразно использовать особый подход к записи дискурсов, диалогов и ответов. Это регистрация мультимодальной сущности языка с помощью звуковых доказательств и просодии [14].

Ш. Уоллис предлагает выделять в отдельные корпуса данные, собранные в результате психолингвистических экспериментов. Он утверждает, что полученные таким образом языковые данные имеют особый характер, при этом качество самих данных зависит от роли и квалификации экспериментатора [13, р. 114]. Материал языковых корпусов – это продукт человеческой языковой/речевой деятельности. Антропоцентрической характеристикой языкового корпуса является и то, что он находится в постоянном развитии, как и сам языковой код. Языковые данные корпусов – это продукты множественного выбора. Ученый полагает: если язык конструируется путем выбора, количественные языковые изменения следует рассматривать как изменение склонности говорящих и писателей к тому или иному выбору [15, р. 19]. Помимо выбора говорящего, источниками вариаций языковых данных могут быть тема, контекст, жанр, стиль и т. д. По мнению Л. Бауэра, если бы у говорящих не было выбора слов или конструкций, язык был бы неспособен к изменениям. Другим психолингвистическим аспектом является изучение интенций говорящего: исследование выбора в языковом корпусе требует вывода контрфактуального. Л. Бауэр отмечает: наряду с тем, что участники написали или сказали, исследователь должен сделать вывод о том, что они предположительно могли бы написать или сказать вместо этого [13, р. 51].

Языковые корпуса содержат материал для интеллектуального анализа текстов. Например, в современную словарную практику внедряется автоматическая обработка языковой информации. В работе Е.А. Горобец и А.В. Мамонтовой описывается версия морфологического анализатора на языке Python 3.8.2 с использованием библиотеки морфологического анализатора `rumorphy2` версии 0.9.1, которая позволяет выявить нестандартные слова в заголовках. Это дает возможность одновременно использовать несколько словарей из различных изданий. Проблема представленности заголовочных единиц в разных формах, но относящихся к одной лексеме, решалась ранее с помощью ручной, неавтоматической обработки материала. Это был трудоемкий и затратный по времени процесс. Примечательно, что возможности созданного анализатора не ограничиваются работой

исключительно с лексикографическими источниками: «Разработанный алгоритм и программу можно использовать для любого списка слов, из которых необходимо автоматически выделить неначальные словоформы» [16, с. 131].

Однако весь массив данных *Big Data* существует именно в форме неструктурированной информации. Многочисленные проблемы возникают при работе «с неструктурированной информацией, которая поступает из разных источников: социальных сетей, блогов, форумов, новостных сайтов и др. Следует отметить, что именно в формате неструктурированного контента представлено около 80% всей накопленной информации в мире» [17, с. 49]. Приемы текстовой аналитики могут быть эффективны для анализа как структурированной, так и неструктурированной информации, представленной в языковых корпусах. Задаются критерии отбора материала из объемного массива данных – ключевые слова, синтаксические структуры, темы, что позволяет выделять морфемный, лексический, синтаксический и семантический типы анализа данных. Учет правил функционирования национального языка, закономерностей речепроизводства, а также применение статистических методов помогают проводить систематизацию больших баз данных, а извлечение целевой информации происходит с применением инструментов текстовой аналитики.

Интеллектуальный анализ текста находит свое практическое применение в разных сферах. Очевидно, что текстовая аналитика наиболее востребована сейчас в бизнес-сфере: рекламе, менеджменте, работе с кадрами предприятия [18]. Появилось понятие бизнес-аналитики (*business intelligence*), которое описывает прежде всего информационную структуру организации. Эффективность взаимодействия различных информационных потоков в организации предопределяет ее успешность. Анализ информации позволяет выявить слабые места и предвидеть возникновение проблем. Иными словами, текстовая аналитика информационных структур в бизнесе повышает его мобильность и конкурентоспособность. Ведется постоянный поиск новых подходов обработки больших баз данных в режиме реального времени. «При этом основной упор должен делаться на гибридность методов, т. е. возможность анализа информации как в традиционных структурах данных (реляционных базах данных), так и в структурах, в которых нет четко установленной структуры хранения данных (NoSQL Databases), где акцент делается на масштабируемости и доступности, за счет согласованности и атомарности информации, хранящейся в базе данных. Также немаловажным аспектом является возможность новых подходов и методов проводить сложные аналитические операции с неструктурированной и квазиструктурированной информацией» [19, с. 41]. В современных условиях особую актуальность приобретают текстовые анализаторы, стоящие на страже информационной и физической безопасности граждан. Применяется интеллектуальный анализ больших баз данных с целью поиска в них экстремистской информации. Набор ключевых слов облегчает поисковые задачи [20].

Нельзя не упомянуть о речевых анализаторах, которые осуществляют интеллектуальный анализ текстов, представленных только в звуковом виде. Примером таких анализаторов является «Войси речевая аналитика» [21]. Программа речевой аналитики “iqinscript” позволяет работать с записями звучащей речи, как монологической, так и диалогической. Это сложная многоступенчатая система, которая занимается сначала перекодировкой звучащей речи, а потом ее анализом [22]. В данном случае наблюдается совмещение программ распознавания речи (проблемы которого описали А.Р. Биктимиров и Д.Ю. Груздев [1, с. 73]) и текстового анализатора. Реальным примером соединения корпусной лингвистики и текстовой аналитики являются сервисы AntConc. Группа этих сервисов исполь-

зует текстовую аналитику при построении и анализе корпусов различного рода текстов [23, с. 111]. Среди аналогичных отечественных сервисов отметим систему сбора и анализа информации больших объемов «Оптимум пересвет аналитика» (optimum peresvet analytics) [24].

Практическое применение

Рассмотрев практические и теоретические моменты взаимодействия психолингвистики, корпусной лингвистики и текстовой аналитики, проанализируем то, как взаимопересечение данных дисциплин происходит в процессе подготовки будущих кадров. В 2021 г. на базе Института филологии и межкультурной коммуникации Казанского федерального университета была начата подготовка магистров по профилю «Текстовая аналитика в образовании и науке». Данная программа предполагает обучение магистрантов, которые будут заниматься преподавательской и научной деятельностью, связанной с разработкой анализаторов текста, нейронными сетями, искусственным интеллектом, изучением родных и иностранных языков, машинным переводом и другими важными вопросами современной науки. Специалисты этой сферы должны обладать знаниями в различных научных областях. Необходимо иметь навыки программирования и работы с лингвистическими корпусами, текстовыми анализаторами, применять статистические методы.

Магистерская программа «Текстовая аналитика в образовании и науке» разрабатывалась с учетом особенностей контингента слушателей, которые в большинстве своем имеют базовое филологическое или педагогическое образование, но также приходят обучаться студенты, окончившие бакалавриат по гуманитарным, юридическим или техническим специальностям. Все дисциплины магистратуры разбиты на несколько модулей. В базовый модуль основных дисциплин включена «Корпусная лингвистика». В психолого-лингвистический модуль включены дисциплины «Психолингвистика» и «Психолингвистические методы изучения языка и текста».

В рамках дисциплины «Корпусная лингвистика» представлена общая информация о связи корпусной лингвистики со смежными дисциплинами, такими как компьютерная лингвистика, психолингвистика и статистика. Рассматривается наполнение понятий «языковой корпус» и «большие данные», а также обсуждаются такие вопросы, как история возникновения электронных языковых корпусов, типология языковых корпусов, национальный корпус и требования к его созданию. Отдельная лекция посвящена теме «Электронная презентация психолингвистических данных», где рассказывается о том, как был собран материал Ассоциативного тезауруса русского языка [25]. Раскрывается научный потенциал тезауруса, поскольку вся база данных, содержащихся в нем, может предоставить информацию о нескольких тысячах слов-стимулов и языковых реакций к ним. Электронная страница тезауруса представляет собой простейший анализатор, в нем возможна классификация материала по гендерному и профессиональному признаку, а также с учетом частотности зафиксированных языковых единиц. Разновидностью электронного ассоциативного психолингвистического словаря является проект СИБАС [26], с которым также необходимо познакомить будущих специалистов в области интеллектуального анализа текстов. Проект представляет собой русскую региональную ассоциативную базу данных, которая на протяжении уже двенадцати лет формируется сотрудниками Новосибирского государственного университета и сектора русского языка Института филологии СО РАН. Отличием данного психолингвистического языкового корпуса от Ассоциативного тезауруса русского языка является то, что все информанты – жители Сибирского федерального округа и Дальнего Востока, то есть это региональный

ассоциативный языковой корпус. Кроме прямого и обратного словаря, имеется постоянно обновляемая база данных. Представлена уникальная информация, дающая возможность изучения особенностей русского языкового сознания информантов, проживающих за пределами нашей страны. Кроме гендерного и профессионального критерия, присутствует еще и региональный.

Дисциплина «Психолингвистика» (или «Введение в психолингвистику», или «Основы психолингвистики») изучается на всех филологических факультетах, поэтому слушатели, которые поступили в магистратуру после получения языкового или педагогического образования, уже знакомы с азами этой дисциплины. Однако есть и студенты, для которых психолингвистика – новая наука. Это обстоятельство нельзя не принять во внимание при разработке курса «Психолингвистика» для магистрантов). Кроме того, следует учитывать и направление подготовки магистров, а именно их будущую работу с интеллектуальным анализом текстов. Исходя из этого, акцент делается на работах, посвященных взаимосвязи языка и мышления (Л.С. Выготский, А.Р. Лурия, А.Н. и А.А. Леонтьевы, Н.И. Жинкин и др.), взаимосвязи языка, сознания и культуры (Е.Ф. Тарасов, Ю.А. Сорокин и др.). Выпускники программы должны быть готовы работать с различного рода текстами – продуктами речевой деятельности разных языковых личностей, поэтому актуальны для обсуждения вопросы онтолингвистики (динамического развития речи детей). Для работы с билингвальными языковыми корпусами требуются знания в области психолингвистических основ билингвизма. Понимание того, как действуют механизмы порождения речи, как влияет психологическое состояние говорящего на речевую продукцию, необходимо для работы с корпусами, состоящими из записей звучащей речи, и речевыми анализаторами. Наконец, внимание уделяется специфике проявлений индивидуальных, национально-культурных и профессиональных особенностей в языке и речи, то есть психолингвистическим аспектам речевой деятельности национально-культурной языковой личности. Рассмотрение текста как объекта психолингвистики позволяет сфокусироваться на таких его параметрах, как смысл, содержание и семантика. На наш взгляд, следует подчеркнуть целесообразность разграничения таких понятий, как «текст» и «совокупность информации».

Основной задачей курса «Психолингвистические методы изучения языка и текста» является изучение основных психолингвистических методов и способов корреляции материала и методов, применяемых для его анализа. Когда материал и методология не соответствуют целям исследования, полученные результаты вряд ли можно квалифицировать как достоверные. В числе основных тем для обсуждения предлагаются: психолингвистические экспериментальные методы и их отличие от лингвистических и психологических методов; виды ассоциативных экспериментов и способы интерпретации полученной информации; психолингвистические методы исследования текста; методы семантического дифференциала и градуальное шкалирование; виды опросников и математические способы обработки полученных данных; авторские эксперименты и др. Как видим, арсенал психолингвистических методов достаточно обширен и использует наработки других смежных наук – социологии, психологии, лингвистики. Также применяются достижения в таких областях, как неврология, программирование, компьютерное моделирование, корпусная лингвистика, традиционные методы наблюдения за языком и поведением. Иными словами, в век цифровизации, чтобы достичь выдающихся результатов, исследователи должны выходить за рамки одной дисциплины [27, с. 555]. Заметим, что психолингвистика предоставляет экспериментатору определенную свободу действий и является благодатной сферой для разработки авторских методов [28, р. 198].

Заключение

Таким образом, потенциал взаимодействия психолингвистической науки, корпусной лингвистики и интеллектуального анализа текста достаточно высок. Психолингвистика развивается в русле антропоцентрической парадигмы на стыке нескольких научных направлений. Она существует в междисциплинарном пространстве и поэтому имеет большой исследовательский потенциал. Интеллектуальный анализ текста применяется, когда есть необходимость извлекать различного рода информацию из больших баз данных и систематизировать неструктурированные тексты больших объемов. Для текстового анализа применяются специально созданные программные продукты и статистические методы. Арсенал инструментов текстового анализа постоянно расширяется в соответствии с потребностями, связанными с практической деятельностью людей. Текстовые анализаторы активно используются в бизнесе (эффективный маркетинг, работа с клиентскими базами данных, организация электронной документации компании), в обеспечении безопасности жизнедеятельности, в образовании, науке и многих других сферах. В связи с этим спрос на специалистов в области текстовой аналитики будет постоянно расти, что обуславливает необходимость подготовки соответствующих кадров.

Литература

1. Биктимиров А.Р., Груздев Д.Ю. Способы повышения эффективности работы программы транскрибации речи // Научный результат. Вопросы теоретической и прикладной лингвистики. 2022. Т. 8, № 4. С. 72–89. URL: <https://doi.org/10.18413/2313-8912-2022-8-4-0-6>.
2. Friginal E. Corpus Linguistics for English Teachers. Tools, Online Resources, and Classroom Activities. N.Y., Abingdon: Routledge, 2018. 366 p. URL: <https://doi.org/10.4324/9781315649054>.
3. Pérez-Paredes P. Corpus Linguistics for Education. A Guide for Research. N.Y.: Routledge, 2020. 194 p. URL: <https://doi.org/10.4324/9780429243615>.
4. Timmis I. Corpus Linguistics for ELT. Research and Practice. Abingdon, N.Y.: Routledge, 2015. 230 p. URL: <https://doi.org/10.4324/9781315715537>.
5. Jones Ch., Waller D. Corpus Linguistics for Grammar. A Guide for Research. Abingdon, N.Y.: Routledge, 2015. 218 p. URL: <https://doi.org/10.4324/9781315713779>.
6. Semino E., Short M. Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing. London, N.Y.: Routledge, 2011. xii, 256 p. URL: <https://doi.org/10.4324/9780203494073>.
7. Szudarski P. Corpus Linguistics for Vocabulary. A Guide for Research. N.Y.: Routledge, 2017. 238 p. URL: <https://doi.org/10.4324/9781315107769>.
8. Friginal E., Hardy J. Corpus-Based Sociolinguistics. A Guide for Students. N.Y., Abingdon: Routledge, 2014. 328 p. URL: <https://doi.org/10.21283/2376905X.2.32>.
9. Han C. Statistics in Corpus Linguistics: A Practical Guide, by Valclav Brezina, Cambridge, Cambridge University Press, 2018, xix+296 pp., ISBN: 978-1107565241 // J. Quant. Linguist. 2020. V. 27, No 4. P. 379–383. URL: <https://doi.org/10.1080/09296174.2019.1646069>.
10. Oakes M. Statistics for Corpus Linguistics // Ser.: Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh Univ. Press, 1998. 287 p. URL: <https://doi.org/10.1515/9781474471381>.
11. Gries S.Th. Quantitative Corpus Linguistics with R. A Practical Introduction. – N.Y.: Routledge, 2016. 286 p. URL: <https://doi.org/10.4324/9781315746210>.

12. *Brezina V.* Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge Univ. Press, 2018. xx, 296 p. URL: <https://doi.org/10.1017/9781316410899>.
13. *Wallis S.* Statistics in Corpus Linguistics Research: A New Approach. N.Y., Abingdon: Routledge, 2021. 382 p. URL: <https://doi.org/10.4324/9780429491696>.
14. *Svenja A., Carter R.* Spoken Corpus Linguistics. From Monomodal to Multimodal. N.Y., Abingdon: Routledge, 2013. 216 p. URL: <https://doi.org/10.4324/9780203526149>.
15. *Bauer L.* Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century. London: Longman, 1994. 200 p.
16. *Горобец Е.А., Мамонтова А.В.* Алгоритм автоматического поиска нестандартных словарных единиц при создании сводного словаря // Актуальные проблемы филологии и педагогической лингвистики. 2021. № 2. С. 131–142. URL: <https://doi.org/10.29025/2079-6021-2022-2-131-142>.
17. *Борисов А.В.* Современные решения и подходы к обработке массивов неструктурированной текстовой информации в области больших данных // Проблемы современной науки и образования. 2017. № 1 (83). С. 49–52.
18. *Попазова О.А., Шихова Н.Н.* Управление персоналом на основе анализа больших данных: риски и возможности // Изв. СПбГЭУ. 2019. № 3 (117). С. 110–115.
19. *Мизюков Г.С.* Основные подходы в области аналитики больших массивов текстовой неструктурированной и квазиструктурированной информации // Труды Рост. гос. ун-та путей сообщения. 2017. № 2. С. 41–44.
20. *Машечкин И.В., Петровский М.И., Поспелова И.И., Царёв Д.В.* Методы автоматического аннотирования и выделения ключевых слов в задачах обнаружения экстремистской информации в сети Интернет // Современные информационные технологии и ИТ-образование. 2016. Т. 12. № 1. С. 188–198.
21. Войси речевая аналитика. Свидетельство о регистрации программы для ЭВМ RU 2019667520, 24.12.2019. Заявка № 2019666394 от 12.12.2019.
22. *Гончаровский И.А., Полозов В.А.* Речевая аналитика “iqinscript”. Свидетельство о регистрации программы для ЭВМ 2021667213, 26.10.2021. Заявка № 2021666043 от 14.10.2021.
23. *Пиотровская К.Р.* Пакеты текстовой аналитики и компьютерная лингводидактика // Актуальные проблемы теоретической и прикладной лингвистики и оптимизация преподавания иностранных языков. К 80-летию профессора Л.И. Корниловой: сборник материалов. 2018. С. 111–117.
24. *Аз-зари Х.М., Анастасьев А.Н.* Система сбора и анализа информации больших объемов optimum пересвет аналитика (оптимум пересвет аналитика) optimum peresvet analytics. Свидетельство о регистрации программы для ЭВМ RU 2018665392, 04.12.2018. Заявка № 2018661634 от 24.10.2018.
25. Русский ассоциативный словарь: в 4 кн. / Караулов Ю.Н., Сорокин Ю.А., Тарасов Е.Ф., Уфимцева Н.В., Черкасова Г.А. М., 1994. (Ассоциативный тезаурус современного русского языка / РАН. Ин-т рус. яз.). URL: <http://thesaurus.ru/dict/>, свободный.
26. СИБАС Русская региональная ассоциативная база данных (Сибирь и Дальний Восток) 2008–2020 гг. URL: <http://adict.ru.nsu.ru/>, свободный.
27. *Privalova I., Kazachkova M.* Review of Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge. ISBN 9781138589384 ISBN 9780429491696 (eBook) // Russ. J. Linguist. 2022. V. 26, No 2. P. 550–557. URL: <https://doi.org/10.22363/2687-0088-30209>.
28. *Privalova I.V.* Methods of psycholinguistic research as possible cognitive approaches to linguistic data processing // Sukhova, N.V., Dubrovskaya, T., Lobina, Y.A. (Eds) Multimodality, Digitalization and Cognitivity in Communication and Pedagogy. Book Ser.: Humanities – Arts and Humanities in Progress. V. 20. Cham, Switzerland: Springer, 2021. P. 181–201. URL: https://doi.org/10.1007/978-3-030-84071-6_11.

Поступила в редакцию 21.05.2023

Принята к публикации 21.07.2023

Привалова Ирина Владимировна, доктор филологических наук, ведущий научный сотрудник
НИИЛ «Текстовая аналитика»

Казанский (Приволжский) федеральный университет

ул. Кремлёвская, д. 18, г. Казань, 420008, Россия

E-mail: ivprivalova@mail.ru

ISSN 2541-7738 (Print)

ISSN 2500-2171 (Online)

UCHENYE ZAPISKI KAZANSKOGO UNIVERSITETA. SERIYA GUMANITARNYE NAUKI

(Proceedings of Kazan University. Humanities Series)

2023, vol. 165, no. 3, pp. 43–54

ORIGINAL ARTICLE

doi: 10.26907/2541-7738.2023.3.43-54

**Psycholinguistics, Corpus Linguistics, and Text Analytics:
Potential for Interaction**

I.V. Privalova

Kazan Federal University, Kazan, 420008 Russia

E-mail: ivprivalova@mail.ru

Received May 21, 2023; Accepted July 21, 2023

Abstract

This article explores the possibilities of interaction between psycholinguistics, corpus linguistics, and text mining. When it comes to big data processing, the recent advances in psycholinguistics must be taken into account. In our digital reality, we should no longer view text as a mere sum of individual language units that are assembled based on their similarities in meaning. The insights of psycholinguistics into text have been revised, as in-depth text analysis unravels the new roles of the sender and the recipient, bringing to the fore qualitative methods for processing large data arrays. Without studying the mechanisms of speech generation and perception, it is impossible to improve language neural networks and machine translation engines. Here, the potential of psycholinguistics for training specialists in text analytics is discussed. As an example, the advantages of integrating the disciplines “Psycholinguistics” and “Psycholinguistic Methods for Studying Language and Text” into the psychological and linguistic module of the Master’s program “Text Analytics in Education and Science”, which has been launched at Kazan Federal University (Russia), are shown. The psycholinguistic component of the discipline “Corpus Linguistics” for Master’s students is described.

Keywords: psycholinguistics, corpus linguistics, text analytics, big data, Master’s program

References

1. Biktimirov A.R., Gruzdev D.Yu. Boosting speech-to-text software potential. *Nauchnyi Rezul'tat. Voprosy Teoreticheskoi i Prikladnoi Lingvistiki*, 2022, vol. 8, no. 4, pp. 72–89. URL: <https://doi.org/10.18413/2313-8912-2022-8-4-0-6>. (In Russian)
2. Friginal E. *Corpus Linguistics for English Teachers. Tools, Online Resources, and Classroom Activities*. New York, Abingdon, Routledge, 2018. 366 p. URL: <https://doi.org/10.4324/9781315649054>.
3. Pérez-Paredes P. *Corpus Linguistics for Education. A Guide for Research*. New York, Routledge, 2020. 194 p. URL: <https://doi.org/10.4324/9780429243615>.

4. Timmis I. *Corpus Linguistics for ELT. Research and Practice*. Abingdon, New York, Routledge, 2015. 230 p. URL: <https://doi.org/10.4324/9781315715537>.
5. Jones Ch., Waller D. *Corpus Linguistics for Grammar. A Guide for Research*. Abingdon, N.Y., Routledge, 2015. 218 p. URL: <https://doi.org/10.4324/9781315713779>.
6. Semino E., Short M. *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing*. London, New York, Routledge, 2011. xii, 256 p. URL: <https://doi.org/10.4324/9780203494073>.
7. Szudarski P. *Corpus Linguistics for Vocabulary. A Guide for Research*. New York, Routledge, 2017. 238 p. URL: <https://doi.org/10.4324/9781315107769>.
8. Friginal E., Hardy J. *Corpus-Based Sociolinguistics. A Guide for Students*. New York, Abingdon, Routledge, 2014. 328 p. URL: <https://doi.org/10.21283/2376905X.2.32>.
9. Han C. Statistics in Corpus Linguistics: A Practical Guide, by Valclav Brezina, Cambridge, Cambridge University Press, 2018, xix+296 pp., ISBN: 978-1107565241. *Journal of Quantitative Linguistics*, 2020, vol. 27, no. 4, pp. 379–383. URL: <https://doi.org/10.1080/09296174.2019.1646069>.
10. Oakes M. *Statistics for Corpus Linguistics*. Ser.: Edinburgh Textbooks in Empirical Linguistics. Edinburgh, Edinburgh Univ. Press, 1998. 287 p. URL: <https://doi.org/10.1515/9781474471381>.
11. Gries S.Th. *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York, Routledge, 2016. 286 p. URL: <https://doi.org/10.4324/9781315746210>.
12. Brezina V. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge, Cambridge Univ. Press, 2018. xx, 296 p. URL: <https://doi.org/10.1017/9781316410899>.
13. Wallis S. *Statistics in Corpus Linguistics Research: A New Approach*. New York, Abingdon, Routledge, 2021. 382 p. URL: <https://doi.org/10.4324/9780429491696>.
14. Svenja A., Carter R. *Spoken Corpus Linguistics. From Monomodal to Multimodal*. New York, Abingdon, Routledge, 2013. 216 p. URL: <https://doi.org/10.4324/9780203526149>.
15. Bauer L. *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London, Longman, 1994. 200 p.
16. Gorobets E.A., Mamontova A.V. Algorithm of automatic search for non-standard vocabulary units when creating a comprehensive dictionary. *Aktual'nye Problemy Filologii i Pedagogicheskoi Lingvistiki*, 2021, no. 2, pp. 131–142. URL: <https://doi.org/10.29025/2079-6021-2022-2-131-142>. (In Russian)
17. Borisov A.V. Current solutions and approaches to processing arrays of unstructured textual data in the field of big data. *Problemy Sovremennoi Nauki i Obrazovaniya*, 2017, no. 1 (83), pp. 49–52. (In Russian)
18. Popazova O.A., Shikhova N.N. Personnel management based on big data analytics: Risks and opportunities. *Izvestiya SPbGEU*, 2019, no. 3 (117), pp. 110–115. (In Russian)
19. Mizyukov G.S. Main approaches to processing large arrays of unstructured and quasi-structured textual data. *Trudy Rostovskogo Gosudarstvennogo Universiteta Putei Soobshcheniya*, 2017, no. 2, pp. 41–44. (In Russian)
20. Mashechkin I.V., Petrovskii M.I., Pospelova I.I., Tsarev D.V. Automatic summarization and keywords extraction methods for discovering extremist information on the Internet. *Sovremennye Informatsionnye Tekhnologii i IT-Obrazovanie*, 2016, vol. 12, no. 1, pp. 188–198. (In Russian)
21. Voicy speech analytics software. State Software Registration Certificate no. 2019667520, Dec. 24, 2019. Application no. 2019666394, Dec. 12, 2019. (In Russian)
22. Goncharovsky I.A., Polozov V.A. IQInScript speech analytics software. State Software Registration Certificate no. 2021667213, Nov. 26, 2021. Application no. 2021666043, Nov. 14, 2021. (In Russian)
23. Piotrovskaya K.R. Text analytics packages and computer linguodidactics. In: *Aktual'nye problemy teoreticheskoi i prikladnoi lingvistiki i optimizatsiya prepodavaniya inostrannykh yazykov. K 80-letiyu professora L.I. Kornilovoi: sbornik materialov* [Urgent Problems of Theoretical and Applied Linguistics and Optimization of Teaching Foreign Languages. On the 80th Anniversary of the Birthday of Professor L.I. Kornilova: A Collection of Articles], 2018, pp. 111–117. (In Russian)

24. Az-zari Kh.M., Anastas'ev A.N. Optimum Peresvet Analytics system for collecting and analyzing large arrays of data. State Software Registration Certificate no. 2018665392, Dec. 4, 2018. Application no. 2018661634, Nov. 24, 2018. (In Russian)
25. Karaulov Yu.N., Sorokin Yu.A., Tarasov E.F., Ufimtseva N.V., Cherkasova G.A. *Russkii assotsiativnyi slovar'* [Russian Associative Dictionary]. Moscow, 1994. Associative Thesaurus of the Contemporary Russian Language, Russian Academy of Sciences, Russian Language Institute. URL: <http://thesaurus.ru/dict/>. (In Russian)
26. SIBAS – Russian Regional Associative Database (Siberia and Russian Far East), 2008–2020. URL: <http://adict.ru.nsu.ru/>. (In Russian)
27. Privalova I., Kazachkova M. Review of Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York, Oxon, Routledge. ISBN 9781138589384 ISBN 9780429491696 (eBook). *Russian Journal of Linguistics*, 2022, vol. 26, no. 2, pp. 550–557. URL: <https://doi.org/10.22363/2687-0088-30209>.
28. Privalova I.V. Methods of psycholinguistic research as possible cognitive approaches to linguistic data processing. In: Sukhova, N.V., Dubrovskaya, T., Lobina, Y.A. (Eds) *Multimodality, Digitalization and Cognitivity in Communication and Pedagogy*. Book Ser.: Humanities – Arts and Humanities in Progress. Vol. 20. Cham, Switzerland, Springer, 2021, pp. 181–201. URL: https://doi.org/10.1007/978-3-030-84071-6_11.

⟨ **Для цитирования:** Привалова И.В. Психолингвистика, корпусная лингвистика и текстовая аналитика: возможности взаимодействия // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. 2023. Т. 165, кн. 3. С. 43–54. <https://doi.org/10.26907/2541-7738.2023.3.43-54>. ⟩

⟨ **For citation:** Privalova I.V. Psycholinguistics, corpus linguistics, and text analytics: Potential for interaction. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki*, 2023, vol. 165, no. 3, pp. 43–54. <https://doi.org/10.26907/2541-7738.2023.3.43-54>. (In Russian) ⟩