МЕТОДОЛОГИЯ И НОВЫЕ ПОДХОДЫ В ГУМАНИТАРНЫХ НАУКАХ

METHODOLOGY AND NOVEL APPROACHES IN HUMANITIES

DOI: 10.12731/3033-5981-2025-17-4-540 **EDN: CQGUZX**
УДК 811.112:81'13

Научная статья

ЭЛЕКТРОННЫЕ КОРПУСЫ НЕМЕЦКОГО ЯЗЫКА КАК ОСНОВА ЛИНГВИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

Т.К. Иванова, Э.В. Гафиятова

Аннотация

Обоснование. В настоящее время лингвистические корпусы являются не только источником эмпирического материала для исследований, но базой для формирования некоторых лингвистических теорий. В статье рассматриваются виды и специфика наиболее известных немецких лингвистических корпусов как инструмента изучения немецкого языка. Описание корпусов производится с целью информирования научной общественности о представляемых ими возможностях. В задачи исследования входят: описание структуры электронных порталов, где размещены данные немецкие корпусы, представление их текстовой базы и объема, а также их структуры. Отдельно оговариваются условия и возможности лингвистического поиска в данных ресурсах. Вначале упоминается история появления первого корпуса, дается его современное определение, делается обзор научной литературы по вопросам использования корпусных данных в лингвистике и смежных науках.

Цель – описание объема, структуры и особенностей немецкоязычных корпусов, а также возможностей автоматизации процесса извлечения материала с учетом запросов исследователей.

Материалы и методы. Материалом исследования послужили немецкоязычные электронные ресурсы, два из которых рассматриваются де-

тально: электронный словарь немецкого языка – Digitales Wörterbuch der deutschen Sprache (DWDS) и корпусы Института Немецкого языка – IDS-Korpora: Corpora of Written Language, (LIMAS), проект DeReKo и COSMAS II, Datenbank für Gesprochenes Deutsch (DGD). В работе также упомянуты проект специального корпуса университета Заарбрюккена Федеральной Земли Заарланд – NEGRA и проект корпусного словаря Лейпцигского университета –– Deutscher Wortschatz (German Vocabulary). Главный метод исследования – структурно-описательный. Количественные показатели использовались как элемент описания и представления данных.

Результаты. Параллельно с описанием корпусов авторы представляют рекомендации по сферам применения того или другого немецкоязычного ресурса и имеющихся на его портале инструментов автоматизированного поиска и обработки запрашиваемой информации. В ходе описания упоминаются объем, структура и особенности лингвистической разметки, поддерживаемой корпусами и определяющей специфику извлекаемого материала. Авторы подчеркивают перспективность освоения корпусных инструментов лингвистами для проверки лингвистических гипотез и обработки эмпирического материала.

Ключевые слова: корпус; немецкий язык; лингвистическая разметка; объем данных; верификация; перспективы использования

Для цитирования. Иванова, Т. К., & Гафиятова, Э. В. (2025). Электронные корпусы немецкого языка как основа лингвистического исследования. *Russian Social and Humanitarian Studies / Российские социогуманитарные исследования*, *17*(4), 137–160. https://doi.org/10.12731/3033-5981-2025-17-4-540

Original article

ELECTRONIC CORPORA OF THE GERMAN LANGUAGE AS A BASIS FOR LINGUISTIC RESEARCH

T.K. Ivanova, E.V. Gafiyatova

Abstract

Background. This article examines the types and specific features of the most well-known German linguistic corpora as tools for studying the German

language. The corpora are described to inform the scientific community about the opportunities they offer. The objectives of the study include: description of the electronic portals, their structure, where these German corpora are hosted; presentation of the data – text base and volume, as well as their structure. The conditions and capabilities of linguistic search in these resources are also discussed. The history of the appearance of the first corpus is mentioned, its modern definition is given. There are a review of the scientific literature about the use of corpus data in linguistics and related sciences too.

Purpose. The description of the volume, structure, and features of German-language corpora, as well as the possibilities of automating the process of extracting material taking into account the needs of researchers.

Materials and methods. The material for the study are German-language electronic resources, two of which are examined in detail: the electronic dictionary of the German language – Digitales Wörterbuch der deutschen Sprache (DWDS) and the corpora of the Institute of the German Language – IDS-Korpora: Corpora of Written Language, (LIMAS), the DeReKo project and COSMAS II, Datenbank für Gesprochenes Deutsch (DGD). The paper also references the NEGRA project, a special corpus of the University of Saarbrücken in the Federal State of Saarland, and the Deutscher Wortschatz (German Vocabulary) – corpus dictionary project of the University of Leipzig. The primary research method is structural and descriptive. Quantitative indicators were used to describe and present the data.

Results. Along with the corpora descriptions, the authors present recommendations on the application areas of each German-language resource and the automated search and processing tools available on its portal. The descriptions also mention the volume, structure, and features of the linguistic markup supported by the corpora, which determines the specifics of the extracted material. The authors emphasize the potential of using corpus tools for linguists to test their hypotheses and process empirical data.

Keywords: corpus; German language; linguistic tagging; data volume; verification; prospects for use

For citation. Ivanova, T. K., & Gafiyatova, E. V. (2025). Electronic corpora of the German language as a basis for linguistic research. *Russian Social and Humanitarian Studies*, *17*(4), 137–160. https://doi.org/10.12731/3033-5981-2025-17-4-540

Введение

Одним из требований современных лингвистических исследований является верификация и подверженность научных выводов, сделанных тем или иным лингвистом в ходе исследования. Поэтому эмпирический материал лингвистического исследования собирается на протяжении долгого времени, требует трудоемкой систематизации и длительной обработки, что ранее существенно удлиняло сроки и усложняло процесс выполнения научного исследования. Ученые-лингвисты, как правило, создавали некие «коллекции» языковых фактов или данных – прообразов современных корпусов. Поэтому и понятие ‘корпуса’ как совокупности определенным образом подобранных текстов, например, одного автора, известно достаточно давно.

Однако сама наука о корпусных данных – корпусная лингвистика – сложилась и обрела самостоятельность вместе с развитием технической мысли и компьютерных технологий. Брауновский корпус или электронный корпус Брауновского университета, созданный У.Н. Френсисом и Г. Кучерой, появился в начале 60-х годов прошлого столетия и включал данные английского языка. С его возникновением получила развитие и новая наука – корпусная лингвистика. Вслед за развитием информационных систем и повсеместным внедрением цифровых технологий она приобретала все большую популярность.

С появлением корпусной лингвистики в лексикон языковеда прочно вошли такие понятия как ‘*тегирование*’, ‘*лемматизация*’, ‘*конкорданс*’, ‘*хеджирование*, ‘*коллокация*’ и пр., а обозначение ‘*корпус*’ стало научным определением: «корпус – это представленный в электронном виде, как правило, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, представляющих как можно большее количество «вариантов» языка» [6, с. 87]. Проблема разметки необработанных данных описывается либо как техническое решение программистами, либо как сопутствующая в статьях о корпусах [6; 10; 24; 26].

Сегодня *корпус* – это совокупность текстов с лингвистической разметкой, которая позволяет извлекать из нее различного рода лингвистическую информацию или языковые факты и, которая вне зависимости от своей первоначальной формы существования (устная или письменная) хранится на электронном носителе, что существенным образом облегчает ее дальнейшую обработку. «С появлением электронных корпусов многообразие форм существования языка стало более наглядным, возможности исследования языковых данных расширились. Современный лингвистический корпус содержит сотни миллионов словоупотреблений, а то, что с помощью электронного корпуса результаты примеров словоупотреблений можно получить за считанные доли секунд, существенно упрощает задачу лингвистам» [6, с. 87].

Корпус исследователя-лингвиста может быть построен и на ограниченном материале, непосредственно под задачи исследования, как например, работа Г.К. Гималетдиновой и А.Р. Алимовой по исследованию субъязыка рекламы, в котором рассматриваются особенности англоязычной рекламы и отмечается, что она – «сложный тип коммуникации, в котором сочетаются черты различных функциональных стилей» [3, с. 108]. Статья интересна и тем, что авторы рассматривают языковые данные исходя из выполняемых рекламой функций, но с учетом уровней языковой системы и квантитативных характеристик, полученных в ходе обработки ограниченного массива данных.

В этой связи были рассмотрены особенности структуры и организации немецких электронных корпусов с целью выявления их потенциала для лингвистических исследований. В статье описаны структура четырех широко известных корпусных массива немецкого языка и перспективы их использования, как в германистике, так и в сопоставительной лингвистике.

Возможности использования корпусов в лингвистике

Возможности использования корпусов и языковые проблемы, в решении которых они могут помочь, достаточно широки и раз-

нообразны. Еще в 2006 г. В. А. Плунгян отмечал, что с помощью корпусных методов языковые исследования могут проводиться не только лишь с позиции выявления обобщенных тенденций и усредненных фактов, а перспективны для выявления территориальных, социальных, гендерных и иных языковых особенностей [11, с. 76–77]. Синхронные и диахронные научные работы в области языкознания также получают новое звучание, равно как и разработки специальных корпусов [10, с. 74].

Текстовый материал для корпуса формируется на основании определенных критериев, к которым, например, относятся: цели и задачи его использования, размер, жанровые особенности текстов, устная или письменная форма предоставления информации, стили речи, общая или профессиональная направленность, уровень сложности, экстралингвистические факторы. В зависимости от представленности и важности данных критериев для систематизации можно провести классификацию корпусов, как например, в работе Н.В. Козловой [6, с. 83-86] или В.П. Захарова [5, с. 56-60]. Кроме того, дополнительными факторами, влияющими на использование корпусов в научной практике, являются такие экстралингвистические факторы, как удобство интерфейса, цифровой дизайн, скорость ответа на запрос и репрезентативность информации.

Большинство развитых национальных языков сегодня имеют электронные корпусы, которые в силу своей специфики (быстрота пополнения, увеличение по мере развития количества параметров автоматизированного поиска, репрезентативность данных, развитие технологий обработки и хранения информации и др.) становятся все более популярным механизмом отражения языковых феноменов конкретного языкового коллектива [13, с. 1-2]. У каждого национального корпуса, как и языка, есть своя специфика, которая делает его уникальным инструментом познания языка.

Так, в 2005 г. была опубликована работа Ангелики Шторрер о перспективах использования корпусных данных немецкими исследователями, в которой подчеркивается разнообразие характеристик, которые лингвисты могут получить из корпусных

данных, а также способы и формы предоставления данной информации и которая охватывает морфологические, семантические, синтаксические и специализированные параметры: «von der Rückführung flektierter Formen auf Grundformen („Lemmatisierung“ genannt) über die Morphemzerlegung und die Zuordnung von Wörtern zu syntaktischen Kategorien („Part-of-Spech-Tagging“ genannt) bis zur partiellen oder vollständigen Analyse syntaktischer Strukturen von Sätzen, die in der Form sog. Baumbanken (Treebanks) über spezialisierte Such- und Recherchewerkzeuge zugänglich gemacht werden» (*от установления исходной формы путем отслеживания флектированных форм (процесс называемый ‘лемматизацией’), разложения на морфемы и соотнесения с синтаксическими категориями (‘частями речи’) до частичного или полного анализа синтаксических конструкций предложений, в форме деревьев путем использования специальных поисковых инструментов и анализаторов’ – перевод авторов*) [27, p.145].

В 2009 г. появилась работа Уве Квастхофа (Uwe Quashoff) по проблемам электронных словарей на корпусной основе, в которой он описывал особенности процесса поиска и обработки данных в трех видах словарей: тезаурусе, словаре неологизмов и словаре коллокаций [26]. Аналогичные работы были опубликованы и о русскоязычных корпусах [2].

За последнее время в отечественном и зарубежном языкознании появилось также большое количество работ, посвященных использованию корпусных технологий в учебных целях, например, в обучении иностранным языкам (Горина, Царакова, 2021; Didakowski, Radtke, 2020; Imo, Weidner, 2018; Flinz, 2020, 2021). О подобных перспективах для немецких корпусов в русскоязычной лингводидактике достаточно подробно пишет в своем исследовании А.Ф. Мухамадьярова [9]. Она перечисляет ряд отечественных и зарубежных работ, в которых описано применение корпусных технологий в преподавании иностранных языков для формирования лексических навыков, письменной речи, правильных предложно-падежных конструкций, расширения стратегий коммуникации [9, с. 249-250].

Корпус может быть создан и для изучения детской речи, особенностей употребления ими отдельных словоформ и лексем, а также построения предложений. Подобное исследование может выявить закономерности формирования предложений и целенаправленного обучения определенным приемам обогащения речи у детей. Так, например, были выявлены особенности употребления анафорических средств детьми в спонтанном высказывании, что отражено в работе А.А. Айсановой и А.С. Морозовой в виде обобщенной диаграммы [1, с. 94].

Применение корпусов в переводческой деятельности также является широко распространенным явлением [20, с. 64], что отражает специфику перевода как языкового посредничества на разных уровнях. Кроме того, в арсенал инструментов переводчика национальные корпусы должны быть включены и как механизм, обеспечивающий качественность перевода при передаче иноязычного текста на родной язык, и с родного на иностранный язык. При этом учитываются не только факты совпадения или расхождения значения отдельных языковых единиц, но и стилистические и синтаксические уровни языка.

Немецкий язык насчитывает порядка десяти известных цифровых проектов по созданию корпусов. У каждого из них есть своя специфика, призванная сделать их использование более удобным и экономичным с точки зрения отраженных в них языковых данных.

Корпусы немецкого языка

Наиболее широко известны следующие корпусы немецкого языка: корпус Берлинской академии наук Федеральной Земли Бранденбург – *Digitales Wörterbuch der deutschen Sprache* (DWDS); корпусы Института немецкого языка – *IDS-Korpora*: письменной корпус *Corpora of Written Language* (LIMAS, проект исследовательской группы Боннского и Регенсбургского университетов), проект корпуса современного немецкого языка свободного доступа – *Deutsches Referenzkorpus* (DeReKo) и его он-лайн вариант с регистрацией пользователей *COSMAS II*, устный корпус немецкого языка – *Datenbank für Gesprochenes Deutsch* (DGD), морфологиче-

ский глоссарий с корпусной поддержкой – *Korpusbasierte Grund-/ Wortformenlisten* (DeReWo) и другие копрусные словари; проект специального корпуса университета Заарбрюккена Федеральной Земли Заарланд – для специалистов в сфере компьютерной лингвистики и информационных наук (NEGRA); проект корпусного словаря Лейпцигского университета -– *Deutscher Wortschatz* (German Vocabulary). Рассмотрим более подробно их структуру и перспективы использования в лингвистическом исследовании.

Корпус Берлинской Академии наук *Digitales Wörterbuch der deutschen Sprache* (DWDS) насчитывает более 100 миллионов словоупотреблений и состоит из основного и дополнительных корпусов. В состав основного корпуса с морфологической разметкой включено около 270 000 слов. Он является одним из первых корпусов немецкого языка XX века, в котором выполнена синтаксическая и семантическая разметка, а также приводятся экстралингвистические факты. В основе DWDS лежат различные типы текстов:

1) подкорпусы газетной периодики (немецкие газеты и журналы: *Zeit*, *Tagesspiegel*, *Berliner Zeitung, Potsdamer Neusten Nachrichten, Spiegel)*, включающие современную интернет-публицистику. Общий объем материалов постоянно пополняется текстами электронных изданий;
2) подкорпус Еврейской периодики (Judicher Periodika), содержащий тексты научной сферы 19-20 вв.;
3) подкорпус немецких художественных текстов эпохи существования Германской Демократической Республики (DDR-Corpus) 1949-1990 гг.;
4) подкорпус немецкой разговорной речи (Gesprochene Sprache), основанный на выступлениях публичных деятелей и включающий, например, речи кайзера Вильгельма, партийного деятеля Вальтера Ульбрихта[1], радиовыступления 1929-1944 гг., а также выдержки из протоколов заседаний парламента Австрии и немецкого Бундестага XX века [17].

[1] Первый секретарь ЦК Социалистической единой партии Германии в 1950-1971 годах. Руководитель ГДР.

На стартовой странице проекта (Рис. 1) после диалогового окна поиска перечислены возможности использования данного ресурса, имеющего более 600 000 вхождений в словари, 75 млрд. ссылок в различных исторических и современных коллекциях текстов. Корпус имеет поддержку немецких авторитетных словарных источников, как например, словаря современного немецкого языка Немецкой Академии наук (*das Wörterbuch der deutschen Gegenwartssprache – WDG*), с недавнего времени словаря WARIG (*WAHRIG Deutsches Wörterbuch – WDW*), а также коллаборационного проекта *OpenThesaurus*, который позволяет при запросе наряду с морфологическими данными и произношением получать при извлечении слова семантические данные: синонимический ряд с указанием гипо- и гиперонима, ассоциативной цепочке, антонимичные отношения, стилистическую принадлежность, ареальные и другие виды ограничений, например, наличие полной или краткой формы слова, сферы его употребления и пр.

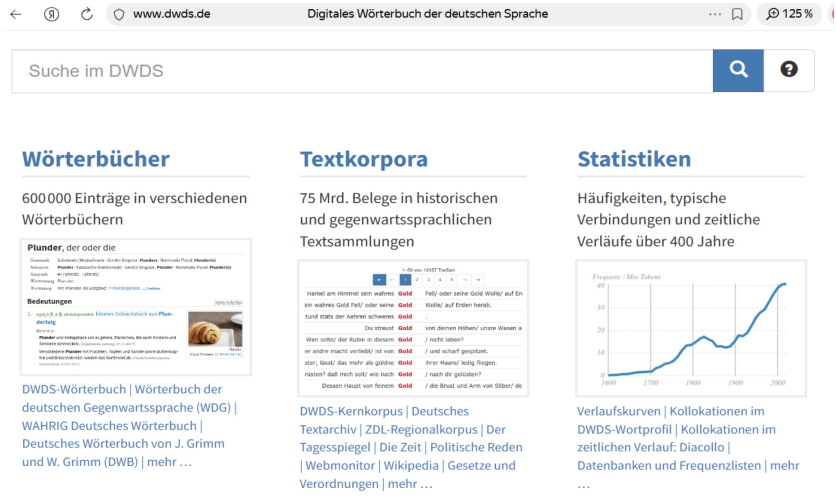

Рис. 1. Стартовая страница DWDS [17]

К проекту подключены несколько диахронных словарей немецкого языка, что обеспечивает при запросе поддержку этимологической справки: немецкий словарь братьев Якоба и Виль-

гельма Гримма (*das Deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm – DWB)* в новом и старом издании; проект «История слова диджитал» Нижнесаксонской Академии наук Геттингена (*Wortgeschichte digital – WGd*), в задачи которого входит описание семантических преобразований в немецком языке с ранне-верхненемецкого периода до современности (словарь построен по принципу объединения тематических групп); этимологического словаря немецкого языка Вольфганга Пфайфера (*das Etymologische Wörterbuch des Deutschen von Wolfgang Pfeifer*), а также немецкий словарь иностранных слов (*das Deutsche Fremdwörterbuch – DFWB*).

Дополнительно на стартовой странице можно запросить данные по обновлению корпуса; позволяющее использовать DWDS на мобильном устройстве без рекламы приложение; увидеть запросы пользователей, выполняемые в режиме реального времени; актуальные темы, связанные с языковыми изменениями и отражающими современное состояние языковой системы, которые призваны привлечь широкий круг пользователей – от специалистов до школьников: языковые интерактивные игры, статьи, тематические блоги. В этом случае можно утверждать, что для популяризации ресурса использована не только академическая репутация, но и современные маркетинговые подходы, что делает данный корпус достаточно востребованным и популярным.

С помощью дополнительных функций исследователь может получить визуализацию различных запросов, например, частотности использования отдельных слов. Так, был сгенерирован график, отражающий сравнение частотности использования двух синонимов немецких понятий, обозначающих ‘союз, объединение, связь или коллаборацию’ – *Union* и *Verein* за период 1946 – 2025 гг. (Рис. 2.).

Отдельно следует упомянуть визуализацию диахронного анализа коллокаций определенных понятий на основе подкорпуса политических речей, входящих в DWDS, – *Kollokationsanalyse in diachroner Perspektive* (DiaCollo) [18]. Анализ коллокаций, как и автоматизация создания лингвистического профиля отдельных слов, – это встроенные в DWDS инструменты. Они очень наглядны, но с технической имеют ряд ограничений по доступности.

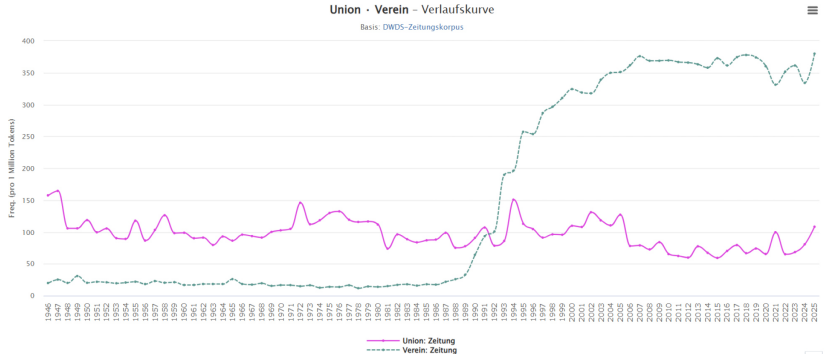

Рис. 2. Сравнение использования понятий Union и Verein в немецком газетном дискурсе (сгенерировано при помощи ресурса DWDS) [21]

Еще одним поставщиком специальных цифровых ресурсов является Институт немецкого языка им. Лейбница – крупнейший разработчик в области цифровой лингвистики. В структуру данного учреждения входят научные отделы [15], занимающиеся популяризацией, разработкой и поддержанием специализированных лингвистических ресурсов, объединенных под названием ‘корпусы Института Немецкого языка’ – *IDS-Korpora*: письменной корпус *Corpora of Written Language* (LIMAS, проект исследовательской группы Боннского и Регенсбургского университетов), научный проект корпуса современного немецкого языка свободного доступа – *Deutsches Referenzkorpus* (DeReKo) и его он-лайн вариант с регистрацией пользователей *COSMAS II*, устный корпус немецкого языка – *Datenbank für Gesprochenes Deutsch* (DGD), морфологический глоссарий с корпусной поддержкой – *Korpusbasierte Grund-/ Wortformenlisten* (DeReWo) и другие корпусные словари.

Из раздела ‘Он-лайн предложения’ на сайте Института можно перейти, не только в корпусы, но и он-лайн-словари с корпусной поддержкой: словарь языковых коннекторов – *grammis,* где можно найти информацию о значении союзов и союзных слов, примеры синтаксических структур и пр.; портал для изучения немецкого языка; проект *OWID,* где находятся специальные словари немецкого языка – пословиц, неологизмов, иностранных слов, а также от-

раслевые словари, носящие названия ‘дискурсориентированных’. В этом же разделе можно обнаружить переход в словарь валентности *E-VALBU*.

Разработчики лингвистического портала Института немецкого языка разместили также ссылки на информационные системы по особенностям региональных вариантов стандартизированного немецкого языка – *AADG* (языковой атлас); грамматическую информационную систему по немецкому языку – *grammis*; мультилингвальную платформу поддержки лексических и лексикографических данных, где можно найти сведения по качественным оценкам лексического состава, например, по изменениям словарного состава в интерактивном режиме – *OWID*^{plus}; портал для исследователей зарубежных национальных вариантов немецкого языка, в частности, русского варианта немецкого языка – *Portal “Russlanddeutsch”*; работающий с 2016 года проект по изучению влияния эмиграционных и интеграционных процессов на вариативность нормативного немецкого языка, языковой адаптации беженцев в Германии – *Deutsch im Beruf*; он-лайн справочник по вариативности грамматических норм немецкого языка – *Variantengrammatik,* а также тематических библиографических данных.

Остановимся лишь на одном из представленных инструментов подробнее. Корпус современного немецкого языка свободного доступа – *Deutsches Referenzkorpus* (DeReKo) и его он-лайн вариант с регистрацией пользователей *COSMAS II* включают на 2025 год 61,5 млрд. слов и являются самой большой в мировом масштабе электронной системой корпусов письменных немецких текстов современности и ближайшего прошлого. Данные для пользователей предоставляются на безвозмездной основе. Корпусы построены на беллетристике, научных и научно-популярных текстах, а также публицистических текстах различных жанров. Их разметка позволяет выполнять разнообразные исследовательские задачи. Материал, размещенный в корпусах, многократно проверен и защищен с точки зрения использования авторского права. Он может быть использован не только в качестве эмпирического материала для линг-

вистических научных работ, но и представляет ценность для междисциплинарных исследований в области психологии, неврологии, когнитивных наук, медиевистики, по проблемам коммуникации или статистических данных. Пользователь получает доступ согласно сформулированному им запросу, то есть ему доступен не весь массив корпусных данных, а лишь его часть. Операционная система поиска достаточно проста в обращении. Для доступа к данным с расширенными функциями требуется простая регистрация (Рис. 3).

Рис. 3. Портал доступа к проекту DeReCo

Таким образом, корпусные ресурсы Института немецкого языка Маннгейма могут быть использованы для решения комплексных лингвистических и междисциплинарных задач, связанных с представлением и обработкой большого массива языковых данных: от простого запроса морфологической структуры слова, до лингвистических гипотез по ассимиляционным процессам, протекающим, например, в словообразовательных моделях немецкого языка.

Корпус NEGRA [7] университета Заарбрюкен Федеральной земли Заарланд в своей второй версии представляет интерес для более узкого круга пользователей – специалистов в сфере компьютерной лингвистики и информационных наук. По сравнению с вышеупомянутыми проектами он небольшой и насчитывает порядка 355 000 токенов. В нем произведена морфолого-синтаксическая разметка, что позволяет проводить специальные языковые исследования в сфере трансформационной или порождающей грамматики, а также зани-

маться построением так называемых ‘синтаксических деревьев’, исследовать лингвистические коллокации в немецком языке. Корпус одноязычен, построен на текстах зафиксированной живой речи.

Последним в данной статье будет представлен многоязычный корпусный проект Лейпцигского университета — *Wortschatz Leipzig* (German Vocabulary) [8]. Основу всех его приложений и сервисов образуют механизмы автоматизированной обработки больших объемов текста, при которых обработка и сканирование веб-страниц осуществляются с помощью методов интеллектуального анализа и собственного программного обеспечения. Он предоставляет доступ к текстовым корпусам и корпусным словарям на более чем 250 языках (Рис. 4), что обосновывает его научную востребованность и значимость.

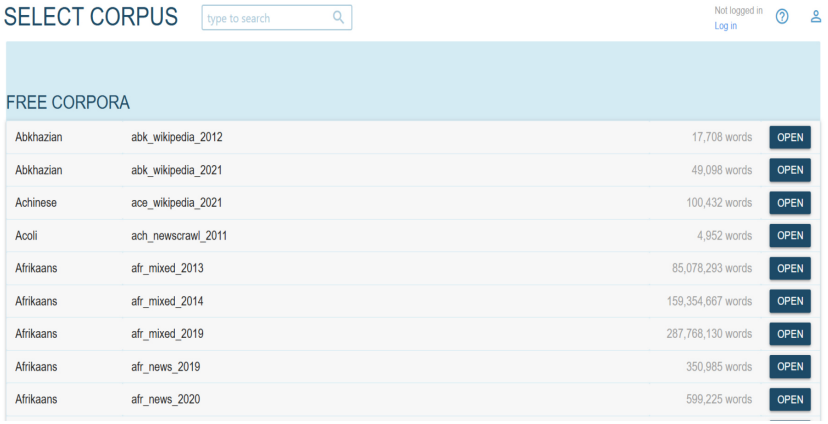

Рис. 4. Доступ к открытым корпусам Лейпцигского проекта (фрагмент) [14]

При этом на портале можно найти информацию по обработке массива сырых оцифрованных текстов с помощью интеллектуального анализа текста, получить результаты автоматизированной обработки любого сгенерированного по запросу пользователя лингвистического корпуса заданной величины на основе свободного доступа к имеющимся на портале многоязычным данным. Лингвистическая статистика по анализу естественных языков может быть выполнена по широкому диапазону параметров в автоматизиро-

ванном режиме на основе алгоритмов интеллектуального поиска собственной разработки Лейпцигского университета, что имеет высокую ценность.

Преимущества и недостатки использования немецких лингвистических корпусов в научных целях

Сегодня в распоряжении языковедов имеется значительное количество ресурсов, позволяющих ускорить проведение научного исследования и, одновременно, подтвердить теоретические выводы о языковой системе на практике. Однако, любой корпусный массив данных не появляется автоматически – появлению нового корпуса непосредственно под задачи исследования предшествует кропотливая работа по оцифровке и разметке лингвистической информации, оформлению автоматизированной системы вызова информации по заданным параметрам и пр. В немецкой языковой традиции, как и в науке вообще, принято детально рассматривать все факторы, влияющие на качество предоставляемой информации. Этим объясняется разнообразие подходов и организации корпусных данных, будь то морфологическая, семантическая или синтаксическая разметка, но, чаще всего, все-таки любая их комбинация.

На корпусные данные влияют как принципы подготовки текстовых данных, базирующиеся на морфологической, синтаксической и семантической разметке или выборе комбинированного ее типа, с одной стороны, так и объем самой выборки, с другой. Ученому-лингвисту необходимо хорошо ориентироваться в принципах моделирования и автоматизации текстовой информации для осуществления оптимального подбора корпуса и создания запроса по извлечению лингвистической информации из него.

Суммируя все вышеизложенное, можно утверждать, что возможности применения корпусных методов в лингвистике далеко не исчерпаны и обнаруживают большое количество преимуществ по сравнению с традиционными подходами, но не заменяют, а дополняют их. Их применение обуславливает необходимость включения в программу подготовки лингвистов основ программирования и

понимания принципов интеллектуального анализа лингвистических данных.

При корпусных исследованиях языка сегодня ученые больше не задаются вопросом, что же представляет из себя корпус, гораздо чаще исследователь должен ответить себе на вопрос: какие параметры и способы хранения языковой информации заложены в его основу, чтобы наиболее полно и глубоко продемонстрировать тот или иной языковой факт? какими существенными характеристиками обладает тот или иной язык и как они изменяются с течением времени? как взаимосвязаны и взаимозависимы естественные и искусственно созданные языки, чтобы усовершенствовать модель общения между человеком и машиной в автоматизированной среде. Межпредметные исследования в этой связи имеют в лингвистике большие перспективы.

Список литературы

1. Айсанова, А. А., & Морозова, А. С. (2025). Построение детского монологического дискурса: особенности употребления анафорических средств в спонтанной детской речи. *Современные исследования социальных проблем, 17*(2), 81–102. https://doi.org/10.12731/2077-1770-2025-17-2-493. EDN: https://elibrary.ru/UFIMMM
2. Ганиева, И. Ф. (2007). Об использовании корпусов в лингвистических исследованиях. *Вестник Башкирского университета. Раздел: Филология и искусствоведение, 12*(4), 104–106.
3. Гималетдинова, Г. К., & Алимова, А. Р. (2025). Лингвостилистические особенности современной англоязычной рекламной коммуникации. *Современные исследования социальных проблем, 17*(2), 103–123. https://doi.org/10.12731/2077-1770-2025-17-2-496. EDN: https://elibrary.ru/VRZVPL
4. Горина, О. Г., & Царакова, Н. С. (2021). Корпусные инструменты, маршруты и эксперименты в современной лингводидактике. *Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация, 19*(2), 36–53. https://doi.org/10.25205/1818-7935-2021-19-2-36-53. EDN: https://elibrary.ru/JWRKME

5. Захаров, В. П., & Богданова, С. Ю. (2020). *Корпусная лингвистика: учебник* (3-е изд., перераб.). Санкт-Петербург: Издательство Санкт-Петербургского университета, 234 с.
6. Козлова, Н. В. (2023). Лингвистические корпуса: определение основных понятий и типология. *Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация, 11*(1), 79–88.
7. *Корпус NEGRA университета Заарбрюкен Федеральной земли Заарланд* [Электронный ресурс]. Получено 20.11.2025, из: https://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5303689618
8. *Многоязычный проект Лейпцигского университета* [Электронный ресурс]. Получено 20.11.2025, из: https://wortschatz.uni-leipzig.de/de
9. Мухамадьярова, А. Ф. (2021). Применение корпусных технологий при формировании лексико-грамматических навыков на немецком языке. *Перспективы науки и образования*, (5), 247–259. https://doi.org/10.32744/pse.2021.5.17. EDN: https://elibrary.ru/EAGTBI
10. Палийчук, Д. А. (2022). Корпусные технологии в лингвистических исследованиях. *Гуманитарные исследования. История и филология*, (6), 72–79. https://doi.org/10.24412/2713-0231-2022-6-72-79. EDN: https://elibrary.ru/VLMGJT
11. Плунгян, В. А. (2006). «Интегрум» и Национальный корпус русского языка в лингвистических исследованиях. В кн.: *Integrum: точные методы и гуманитарные науки* (с. 76–84). Москва. EDN: https://elibrary.ru/PXFBEL
12. *Портал Лейпцигского университета* [Электронный ресурс]. Получено 20.11.2025, из: https://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5303689618
13. Рюкова, А. Р. (2024). Корпусно-ориентированные исследования языка: краткий обзор достижений и трудностей. *Russian Linguistic Bulletin*, (1), 49. Получено 21.11.2025, из: https://rulb.org/archive/1-49-2024-january/. https://doi.org/10.18454/RULB.2024.49.17. EDN: https://elibrary.ru/CEQMAT
14. *Стартовая страница поиска корпуса на портале проекта Лейпцигского университета* [Электронный ресурс]. Получено 20.11.2025, из: https://text.wortschatz-leipzig.de/#open

15. *Структура Института Немецкого языка в виде схемы* [Электронный ресурс]. Получено 20.11.2025, из: https://www.ids-mannheim.de/org/orga/
16. *Электронный корпусный словарь немецкого языка Digitales Wörterbuch der deutschen Sprache (DWDS): стартовая страница* [Электронный ресурс]. Получено 21.11.2025, из: https://www.dwds.de/
17. *Электронный корпусный словарь немецкого языка Digitales Wörterbuch der deutschen Sprache (DWDS): стартовая страница (раздел «Wörterbücher»)* [Электронный ресурс]. Получено 21.11.2025, из: https://www.dwds.de/d/woerterbuecher
18. *Электронный ресурс, скооперированный с DWDS* [Электронный ресурс]. Получено 20.11.2025, из: https://ddc.dwds.de/dstar/politische_reden/diacollo/?query=&format=cloud&corpus=politische_reden
19. Didakowski, J., & Radtke, N. (2020). Verwendung der deutschen Stützverbgefüge mit Adjektiven und ihre Ermittlung mithilfe des DWDS-Wortprofils. In *Funktionsverbgefüge im Fokus* (pp. 101–134). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110697353-005
20. Disanto, G. A. (2009). Korpusbasierte Translationswissenschaft. Eine Untersuchung am Beispiel des JRC-Acquis Parallelkorpus Deutsch-Italienisch. *Corpus-based Translation Studies. A Study of the "JRC-Acquis" Parallel Corpus German-Italian — Abstract. trans-kom, 2*(1), 63–91.
21. *DWDS-Verlaufskurve für „Union / Verein"*, erstellt durch das Digitale Wörterbuch der deutschen Sprache. Получено 20.11.2025, из: https://www.dwds.de/r/plot/?view=1&corpus=zeitungenxl&norm=date%2Bclass&smooth=spline&genres=0&grand=1&slice=1&prune=0&window=0&wbase=0&logavg=0&logscale=0&xrange=1946-%3A2025&q1=Union&q2=Verein
22. Flinz, C. (2021). KORPORA in DaF und DaZ: Theorie und Praxis. *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 26*(1). Получено 24 июля 2021, из: https://ojs.tujournals.ulb.tu-darmstadt.de/index.php/zif/article/view/1112/1108
23. Flinz, C. (2020). Vergleichbare Spezialkorpora für den Tourismus: eine Chance für den Fachsprachenunterricht. In *Sprachvergleich in der mehrsprachig orientierten DaF-Didaktik. Theorie und Praxis* (pp. 133–151). Roma: Istituto Italiano di Studi Germanici.

24. Frank, A. (2001). Treebank Conversion: Converting the NEGRA Corpus to an LTAG Grammar. In *Proceedings of the EUROLAN Workshop on Multi-layer Corpus-based Analysis* (pp. 29–43). Saarbrücken: DFKI GmbH.
25. Imo, W., & Weidner, B. (2018). Mündliche Korpora im DaF und DaZ Unterricht. In *Korpuslinguistik* (pp. 231–252). Berlin, Boston: De Gruyter. Получено 24 июля 2021, из: https://www.degruyter.com/document/doi/10.1515/9783110538649-011/html
26. Quasthoff, U. (2009). Korpusbasierte Wörterbucharbeit mit den Daten des Projekts Deutscher Wortschatz. *Linguistik online, 39*(3), 151–162. Universität Bern, Bern, Schweiz.
27. Storrer, A. (2005). Online-Corpora zur linguistischen Analyse der deutschen Gegenwartssprache [A Survey of Online Corpora für Investigating Contemporary German]. *Zeitschrift für Germanistische Linguistik, 33*(1), 145–150. https://doi.org/10.1515/zfgl.2005.33.1.145

References

1. Aisanova, A. A., & Morozova, A. S. (2025). Constructing children's monologic discourse: features of anaphoric means usage in spontaneous children's speech. *Modern Studies of Social Problems*, *17*(2), 81–102. https://doi.org/10.12731/2077-1770-2025-17-2-493. EDN: https://elibrary.ru/UFIMMM
2. Ganieva, I. F. (2007). On the use of corpora in linguistic research. *Bulletin of Bashkir University. Section: Philology and Art Studies*, *12*(4), 104–106.
3. Gimaletdinova, G. K., & Alimova, A. R. (2025). Linguostylistic features of modern English-language advertising communication. *Modern Studies of Social Problems*, *17*(2), 103–123. https://doi.org/10.12731/2077-1770-2025-17-2-496. EDN: https://elibrary.ru/VRZVPL
4. Gorina, O. G., & Tsarakova, N. S. (2021). Corpus tools, routes, and experiments in modern linguodidactics. *NSU Bulletin. Series: Linguistics and Intercultural Communication*, *19*(2), 36–53. https://doi.org/10.25205/1818-7935-2021-19-2-36-53. EDN: https://elibrary.ru/JWRKME

5. Zakharov, V. P., & Bogdanova, S. Yu. (2020). *Corpus linguistics: textbook* (3rd ed., revised). Saint Petersburg: Saint Petersburg University Publishing House, 234 p.
6. Kozlova, N. V. (2023). Linguistic corpora: defining key concepts and typology. *NSU Bulletin. Series: Linguistics and Intercultural Communication*, *11*(1), 79–88.
7. NEGRA Corpus of Saarland University [Electronic resource]. Retrieved on November 20, 2025, from: https://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5303689618
8. Multilingual project of Leipzig University [Electronic resource]. Retrieved on November 20, 2025, from: https://wortschatz.uni-leipzig.de/de
9. Mukhamadyarova, A. F. (2021). Applying corpus technologies to develop lexico-grammatical skills in German. *Prospects of Science and Education*, *(5)*, 247–259. https://doi.org/10.32744/pse.2021.5.17. EDN: https://elibrary.ru/EAGTBI
10. Paliychuk, D. A. (2022). Corpus technologies in linguistic research. *Humanities Research. History and Philology*, *(6)*, 72–79. https://doi.org/10.24412/2713-0231-2022-6-72-79. EDN: https://elibrary.ru/VLMGJT
11. Plungyan, V. A. (2006). "Integrum" and the Russian National Corpus in linguistic research. In: *Integrum: precise methods and humanities* (pp. 76–84). Moscow. EDN: https://elibrary.ru/PXFBEL
12. Leipzig University Portal [Electronic resource]. Retrieved on November 20, 2025, from: https://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5303689618
13. Ryukova, A. R. (2024). Corpus-oriented language studies: a brief overview of achievements and challenges. *Russian Linguistic Bulletin*, *(1)*, 49. Retrieved on November 21, 2025, from: https://rulb.org/archive/1-49-2024-january/. https://doi.org/10.18454/RULB.2024.49.17. EDN: https://elibrary.ru/CEQMAT
14. Starting page for corpus search on the Leipzig University project portal [Electronic resource]. Retrieved on November 20, 2025, from: https://text.wortschatz-leipzig.de/#open

15. Structure of the Institute for the German Language as a diagram [Electronic resource]. Retrieved on November 20, 2025, from: https://www.ids-mannheim.de/org/orga/
16. Digitales Wörterbuch der deutschen Sprache (DWDS): electronic corpus dictionary of the German language — home page [Electronic resource]. Retrieved on November 21, 2025, from: https://www.dwds.de/
17. Digitales Wörterbuch der deutschen Sprache (DWDS): electronic corpus dictionary of the German language — home page (section "Wörterbücher") [Electronic resource]. Retrieved on November 21, 2025, from: https://www.dwds.de/d/woerterbuecher
18. Electronic resource cooperated with DWDS [Electronic resource]. Retrieved on November 20, 2025, from: https://ddc.dwds.de/dstar/politische_reden/diacollo/?query=&format=cloud&corpus=politische_reden
19. Didakowski, J., & Radtke, N. (2020). Verwendung der deutschen Stützverbgefüge mit Adjektiven und ihre Ermittlung mithilfe des DWDS Wortprofils. In *Funktionsverbgefüge im Fokus* (pp. 101–134). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110697353-005
20. Disanto, G. A. (2009). Korpusbasierte Translationswissenschaft. Eine Untersuchung am Beispiel des JRC Acquis Parallelkorpus Deutsch-Italienisch. *Corpus based Translation Studies. A Study of the "JRC Acquis" Parallel Corpus German-Italian — Abstract*. *trans kom*, *2*(1), 63–91.
21. DWDS time-course diagram for "Union / Verein", created by the Digital Dictionary of the German Language. Retrieved on November 20, 2025, from: https://www.dwds.de/r/plot/?view=1&corpus=zeitungenxl&norm=date%2Bclass&smooth=spline&genres=0&grand=1&slice=1&prune=0&window=0&wbase=0&logavg=0&logscale=0&xrange=1946%3A2025&q1=Union&q2=Verein
22. Flinz, C. (2021). KORPORA in DaF und DaZ: Theorie und Praxis. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *26*(1). Retrieved on July 24, 2021, from: https://ojs.tujournals.ulb.tu-darmstadt.de/index.php/zif/article/view/1112/1108
23. Flinz, C. (2020). Vergleichbare Spezialkorpora für den Tourismus: eine Chance für den Fachsprachenunterricht. In *Sprachvergleich in der mehrsprachig orientierten DaF-Didaktik. Theorie und Praxis* (pp. 133–151). Roma: Istituto Italiano di Studi Germanici.

24. Frank, A. (2001). Treebank conversion: converting the NEGRA corpus to an LTAG grammar. In *Proceedings of the EUROLAN Workshop on Multi-layer Corpus-based Analysis* (pp. 29–43). Saarbrücken: DFKI GmbH.
25. Imo, W., & Weidner, B. (2018). Mündliche Korpora im DaF und DaZ Unterricht. In *Korpuslinguistik* (pp. 231–252). Berlin, Boston: De Gruyter. Retrieved on July 24, 2021, from: https://www.degruyter.com/document/doi/10.1515/9783110538649-011/html
26. Quasthoff, U. (2009). Korpusbasierte Wörterbucharbeit mit den Daten des Projekts Deutscher Wortschatz. *Linguistik online*, *39*(3), 151–162. Universität Bern, Bern, Schweiz.
27. Storrer, A. (2005). Online Corpora zur linguistischen Analyse der deutschen Gegenwartssprache [A survey of online corpora for investigating contemporary German]. *Zeitschrift für Germanistische Linguistik*, *33*(1), 145–150. https://doi.org/10.1515/zfgl.2005.33.1.145

ДАННЫЕ ОБ АВТОРАХ

Иванова Татьяна Константиновна, доктор филологических наук, доцент, профессор кафедры теории и практики преподавания иностранных языков
Казанский федеральный университет
ул. Кремлевская, 18, г. Казань, 420008, Российская Федерация
Tatiana.ivanova@kpfu.ru

Гафиятова Эльзара Василовна, доктор филологических наук, профессор кафедры теории и практики преподавания иностранных языков
Казанский федеральный университет
ул. Кремлевская, 18, г. Казань, 420008, Российская Федерация
Elzara.Gafiyatova@kpfu.ru

DATA ABOUT THE AUTHORS

Tatiana K. Ivanova, Doctor of Sciences (Philology), Associate Professor, Department of Theory and Praxis in Teaching Foreign Languages

Kazan Federal University
18, Kremlevskaya Str., Kazan, 420008, Russian Federation
Tatiana.ivanova@kpfu.ru
SPIN-code: 4785-5424
ORCID: https://orcid.org/0000-0001-5355-6430
ResearcherID: V-8495-2017
Scopus Author ID: 57221643922
Google Scholar: https://scholar.google.ru/citations?hl=ru&user=xQIBFQEAAAAJ

Elsara V. Gafiyatova, Doctor of Sciences (Philology), Professor, Department of Theory and Praxis in Teaching Foreign Languages
Kazan Federal University
18, Kremlevskaya Str., Kazan, 420008, Russian Federation
Elzara.Gafiyatova@kpfu.ru
SPIN-code: 4765-9021
ORCID: https://orcid.org/0000-0003-3190-4566
ResearcherID: I-9556-2014
Scopus AuthorID: 56716561200
Google Scholar: https://scholar.google.ru/citations?user=clUUlt8AAAAJ&hl=ru
ResearchGate: https://www.researchgate.net/profile/Elzara_Gizzatullina-Gafiatova

Поступила 12.11.2025
После рецензирования 04.12.2025
Принята 20.12.2025

Received 12.11.2025
Revised 04.12.2025
Accepted 20.12.2025