Корпус студенческих текстов на немецком языке как источник данных для образования и науки

Ирина Котюрова, Людмила Щеголева

Статья поступила в редакцию в июле 2022 г.

Котюрова Ирина Аврамовна — кандидат филологических наук, доцент, заведующая кафедрой немецкого и французского языков ФГБОУ ВО «Петрозаводский государственный университет». Адрес: 185910, Петрозаводск, просп. Ленина, 33. E-mail: koturova@petrsu.ru, ORCID: https://orcid.org/0000-0001-6766-0458 (контактное лицо для переписки)

Щеголева Людмила Владимировна — доктор технических наук, доцент, профессор кафедры прикладной математики и кибернетики ФГБОУ ВО «Петрозаводский государственный университет». E-mail: schegoleva@petrsu.ru, ORCID: https://orcid.org/0000-0001-5539-9176

Аннотация

Одним из примеров цифровизации образования служит создание лингвистического корпуса студенческих работ на иностранном языке при учебном заведении с целью использования его для исследований, преподавания и образовательной аналитики. В статье рассматривается опыт формирование такого лингвистического корпуса при ФГБОУ ВО «Петрозаводский государственный университет». Основы Петрозаводского аннотированного корпуса текстов (ПАКТ) заложены в 2020 г., и два года его развития уже показали перспективность внедрения корпуса в работу вуза и разнообразие вариантов его применения. Дано общее описание ПАКТа по состоянию на июнь 2022 г. и приводятся примеры его использования — реализуемые и планируемые к реализации. Основной исследовательский вопрос работы: какие возможности для повышения эффективности обучения и профессионального роста будущих преподавателей иностранного языка дает корпус студенческих текстов на иностранном языке, созданный при вузе? Отвечая на него, авторы обращаются и к опыту другого корпуса студенческих текстов — REALEC, англоязычного аннотированного ученического корпуса, послужившего прототипом ПАКТа. Авторы приходят к выводу, что корпус студенческих текстов на иностранном языке как пример больших данных, собираемых в вузе, имеет широкие перспективы применения в образовательных целях непосредственно на занятиях по иностранному языку, для научно-исследовательской работы студентов и в качестве ресурса для автоматического генератора индивидуальных заданий; в исследовательских целях для специалистов в области лингвистики, лингводидактики, психолингвистики, информационных технологий и искусственного интеллекта; в целях образовательной аналитики вуза.

Ключевые слова

Петрозаводский аннотированный корпус текстов, ПАКТ, корпус студенческих текстов, корпус ученических текстов, иностранный язык, *Data Driven Learning*, большие данные в образовании, анализ образовательных данных.

Для цитирования

Котюрова И.А., Щеголева Л.В. (2022) Корпус студенческих текстов на немецком языке как источник данных для образования и науки. *Вопросы образования / Educational Studies Moscow*, № 4, сс. 322–349. https://doi.org/10.17323/1814-9545-2022-4-322-349

Learner Corpus in German as a Data Source for Education and Science

Irina Kotiurova, Liudmila Shchegoleva

Irina A. Kotiurova — Candidate of Sciences in Philology, Associate Professor, Head of the Department of German and French Languages, Petrozavodsk State University. Address: Lenin Ave., 33, Petrozavodsk 185910, Russian Federation. E-mail: koturova@petrsu.ru, ORCID: https://orcid.org/0000-0001-6766-0458 (corresponding author)
Liudmila V. Shchegoleva — Doctor of Technical Sciences, Associate Professor, Professor of the Department of Applied Mathematics and Cybernetics, Petrozavodsk State University. E-mail: schegoleva@petrsu.ru, ORCID: https://orcid.org/0000-0001-5539-9176

Abstract One example of the digitalization of education is the creation of a linguistic learner corpus of student papers in a foreign language at an educational institution in order to use this corpus for research, teaching and learner analytics. This paper describes the experience of creating such linguistic learner corpus at Petrozavodsk State University. Petrozavodsk Annotated Corpus of Texts (PACT) was founded in 2020, but even 2 years of its development have already shown a wide field for implementing this experience in the work of a university. The article provides a general description of the learner corpus PACT and gives examples of its use – being implemented and planned to be implemented. The key research question is what opportunities for training and professional development of future foreign language teachers offers and what potential for educational data mining and management of the educational process the learner corpus in a foreign language has. Answering this basic question, the authors refer to the experience of another corpus of student texts — REALEC, the English annotated learner corpus, which is the prototype of the original PACT corpus. The authors conclude that the corpus of student texts in a foreign language, as an example of big data collected in a university, has great potential in several directions: for research purposes for specialists in linguistics, linguodidactics, psycholinguistics, information technology and artificial intelligence; for educational purposes directly applied in the foreign language classes, for students research work and as a resource for individual tasks automatic generator; for educational data mining.

Keywords Petrozavodsk annotated corpus of PACT texts, learner corpus, foreign language, Data Driven Learning, Big Data in Education, educational data mining.

For citing Kotiurova I.A., Shchegoleva L.V. (2022) Korpus studencheskikh tekstov na nemetskom yazyke kak istochnik dannykh dlya obrazovaniya I nauki [Learner Corpus in German as a Data Source for Education and Science]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 4, pp. 322–349. https://doi.org/10.17323/1814-9545-2022-4-322-349

Среди основных характеристик современной технологической революции неизменно выделяют «цифровую трансформацию» и «масштабирование процессов цифровизации» во всех сферах жизни, в том числе в образовании [Княгинин и др., 2017. С. 26]. Пути внедрения в образовательный процесс передо-

вых цифровых технологий обсуждаются на многочисленных форумах и становятся предметом исследований, проводимых во всем мире [Другова, Велединская, Журавлева, 2021; Радаев и др., 2018; Дворецкая и др., 2022; Bates, Bates, 2015; Bawa, 2020; Langthaler, Bazafkan, 2020]. Исследователи образования анализируют основные проблемы цифровизации и возможные пути их решения, продвигают передовой опыт использования искусственного интеллекта в образовательном процессе [Уваров, Фрумин, 2019].

Перспективным направлением цифровой трансформации образования является сбор больших данных для управления образованием и автоматизации оценки работы учащихся [Уваров, Фрумин, 2019; Фиофанова, 2020; 2021; Ищенко, 2020; Hou et al., 2019]. В преподавании иностранных языков в образовательных и исследовательских целях используются лингвистические корпуса [Lüdeling, Walter, 2009; Виноградова, 2021; Vinogradova, Viklova, Smilga, 2021; Кузнецова, Шангараева, 2021; Черепанова, 2015; Katinskaia, Nouri, Yangarber, 2017; 2018; Kormacheva, Pivovarova, Kopotev, 2014]. Своеобразной точкой пересечения этих двух направлений цифровизации образования стало создание собственного лингвистического корпуса студенческих работ на иностранном языке при учебном заведении для использования с целью повышения качества образования.

В данной статье описан опыт создания такого лингвистического корпуса при ФГБОУ ВО «Петрозаводский государственный университет». Петрозаводский аннотированный корпус текстов (ПАКТ)[1] содержит тексты на немецком и французском языках. Более активно развивается немецкоязычная часть корпуса, поэтому в статье далее речь пойдет именно о немецкоязычном подкорпусе ПАКТа[2].

Создание ПАКТа было мотивировано успехами англоязычного корпуса — REALEC (*Russian Error Annotated Learner Corpus*), созданного в Научно-учебной лаборатории учебных корпусов Школы лингвистики НИУ ВШЭ (Москва)[3]. Вообще же в настоящее время в мире существует несколько сотен учебных корпусов на 26 языках. Подавляющее большинство таких кол-

[1] Разные эксперты используют разные определения для таких коллекций текстов обучающихся: учебный корпус, ученический корпус, корпус ученических текстов. Поскольку словосочетание «учебный корпус» имеет и другое значение — здание для организации учебного процесса, в статье оно используется только в составе имени собственного: Научно-учебная лаборатория учебных корпусов. Наиболее точным в отношении ПАКТа является определение «корпус студенческих текстов», хотя и остальные указанные варианты применимы.

[2] ПАКТ: https://pact.ai.petrsu.ru/app

[3] Научно-учебная лаборатория учебных корпусов: https://hum.hse.ru/lcl/

лекций сочинений посвящены английскому языку как иностранному — более 100. Согласно сайту Лувенского католического университета[4], на втором месте по частотности указания в качестве целевого языка стоит испанский (22 корпуса) и на третьем — немецкий язык (20 корпусов). Доступны для исследований, в том числе и по запросу, следующие коллекции, полностью или частично построенные на немецкой речи, как письменной, так и устной, неносителей языка: BeMaTaC[5], DaZAF[6], DISKO[7], ESF-Korpus[8], Falko[9], GeWiss[10], KanDel[11], MERLIN[12], MIKO[13], MULTILIT[14], P-Moll-Korpus[15], SWIKO[16].

Обилие ученических корпусов и постоянный рост их количества свидетельствует, с одной стороны, об эффективности их использования в различных целях, а с другой — о специфичности каждого из них. Многие коллекции учебных текстов на немецком языке включают работы носителей английского, испанского, словенского и других языков — а значит, могут служить основой для сравнительных исследований немецкого языка как иностранного, но не могут заменить корпус текстов на немецком языке, написанных русскоязычными студентами в отечественных вузах.

Основы Петрозаводского аннотированного корпуса текстов были заложены в 2020 г., и два года его развития уже доказали перспективность внедрения этого опыта в работу вуза. Основной вопрос, на который мы рассчитываем дать ответ в данной статье: какими возможностями для повышения эффективности обучения и профессионального роста будущих преподавателей иностранного языка располагает корпус студенческих текстов на иностранном языке, созданный при вузе?

В статье дана общая характеристика ПАКТа по состоянию на июнь 2022 г., описаны методы его построения и заполнения, приводятся примеры получаемых на его основе статисти-

[4] Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[5] BeMaTaC: http://u.hu-berlin.de/bematac
[6] DaZAF: https://hdl.handle.net/1839/00-0000-0000-0000-69D7-E
[7] DISKO: https://home.uni-leipzig.de/sprastu/korpora/DISKO/
[8] ESF-Korpus: https://hdl.handle.net/1839/00-0000-0000-0004-CCB0-8
[9] Falko: https://hu-berlin.de/falko
[10] GeWiss: https://gewiss.uni-leipzig.de
[11] KanDel: https://hu-berlin.de/kandel
[12] MERLIN: https://merlin-platform.eu/
[13] MIKO: https://home.uni-leipzig.de/sprastu/korpora/miko/
[14] MULTILIT: https://publishup.uni-potsdam.de/frontdoor/index/index/docId/8039
[15] P-Moll-Korpus: https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A
[16] SWIKO: http://www.institut-mehrsprachigkeit.ch/de/content/schweizer-lernerkorpus-swiko

ческих данных и материалов, а также предлагаются методики использования корпуса и его данных в образовательных и исследовательских целях.

1. Общее описание ПАКТа

Петрозаводский аннотированный корпус текстов расположен по адресу https://pact.ai.petrsu.ru/app, где любой, в том числе неавторизованный, пользователь может задавать условия поиска на языке CQL (*Corpus Query Language*) и просматривать его результаты в виде списков предложений, удовлетворяющих заданным условиям. Доступ к базе данных для исследовательских целей возможен только по прямому обращению к разработчикам корпуса, контакты которых указаны на главной странице корпуса.

По состоянию на июнь 2022 г. корпус включает более 1100 студенческих текстов на немецком языке объемом около 300 тыс. токенов[17]. Тексты представляют собой сочинения-рассуждения, тематические эссе, разного рода описания, пересказы.

Для каждого предложения в текстах выполнена автоматическая разметка по частям речи с помощью программного обеспечения *RFTagger*, поскольку именно этот частеречный разметчик по итогам исследования проявил себя как наиболее подходящий к условиям ученического корпуса [Kotiurova, Trenina, 2022]. Программа разбивает предложение на отдельные токены, включающие слова и знаки пунктуации. В результате обработки предложения каждый токен получает тег, определяющий часть речи (рис. 1). Отображение частеречной разметки в корпусе настраивается вручную и может быть отключено.

Рис. 1. **Образец отображения автоматической частеречной разметки в ПАКТе**

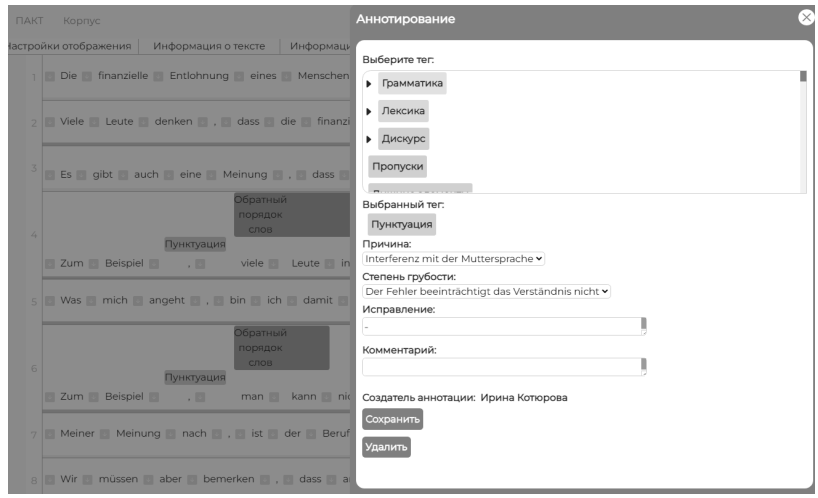
7	NN Familie	.												
8	ADJD Wahrscheinlich	VVFIN gibt	PPER es	PLAT kein	NN Wort	.	PRELS das	ADJD wärmer	KON und	ADJD sanfter	KOKOM als	PDS dieses	VAFIN ist	-

Второй тип разметки по лингвистическим типам ошибок проводится экспертным путем. Ошибка может охватывать несколько токенов. Один и тот же токен может быть включен в несколько ошибок. Для аннотирования ошибок лингвистического типа разработана уникальная иерархическая классификация, включающая 91 наименование (Приложение 1). Прототипом разра-

[17] Токенами в корпусной лингвистике принято называть основные единицы корпуса, т.е. слова, словоупотребления.

ботанной иерархической классификации послужила классификация ошибок англоязычного корпуса REALEC. Однако учет особенностей немецкого языка и типичных ошибок в речи изучающих его русскоговорящих студентов обусловили уникальное наполнение классов и подклассов иерархии. Число пунктов классификации менялось в процессе наполнения корпуса и, возможно, еще изменится, поскольку в ходе работы аннотаторов, вручную выполняющих разметку ошибок в текстах, выявляется необходимость некоторых уточнений или обобщений, в результате которых, хоть и незначительно, меняется количество тегов в классификации. Как видно из таблицы в Приложении 1, ошибки распределяются по семи основным классам: грамматика, лексика, дискурс, пропуски, лишние элементы, орфография и пунктуация. Для каждой ошибки при разметке обязательно указываются ее тип (согласно классификации из Приложения 1), степень грубости (по трем категориям: не влияет на восприятие / затрудняет восприятие / искажает смысл) и исправление. Факультативно, в случае отсутствия сомнений у аннотатора, размечается одна из возможных причин ошибки: интерференция с родным языком, интерференция с английским языком (первым иностранным языком студентов) или опечатка (рис. 2).

Рис. 2. **Пример экспертного аннотирования ошибок в ПАКТе**



Очень важной составляющей корпуса является метаразметка, сопровождающая текст. В нее входят следующие пункты, заполняемые отчасти автоматически, отчасти вручную автором текста: автор (его идентификационный номер в корпусе); номер академической группы; на каком курсе написан текст; дата написания текста; самооценивание текста (по шкале от 1 до 5);

понравилось ли задание (по шкале от 1 до 5); эмоциональное состояние в момент написания текста (подавленность / через силу / усталость / равнодушие / бодрость / приподнятое настроение / другое); тип текста (письмо, эссе, пересказ и т.п.); тип выполнения работы (на занятии / дома, от руки / с использованием текстового редактора).

Для работы с корпусом выделены две роли: преподаватель и студент. Преподаватель имеет права на загрузку, просмотр и редактирование любых текстов корпуса, включая сам текст (в некоторых случаях возникает необходимость в удалении части загруженного текста, например связанной с указанием данных студента, написавшего текст), метаразметку текста, разметку по частям речи в случае некорректных результатов работы автоматического разметчика, разметку по типам ошибок и выставление оценки за выполненное студентом задание. Студент имеет права на загрузку и просмотр своих текстов, заполнение метаинформации по тексту. Перед первым сеансом работы с ученическим корпусом обучающийся подписывает согласие на сбор и обработку индивидуальных метаданных, затем регистрируется в системе и получает возможность самостоятельно загружать свои сочинения, написанные как на занятиях по немецкому языку, так и дома. В момент подписания согласия студентам разъясняют суть проекта, его ценность и важность заполнения метаданных при самостоятельной загрузке текстов в корпус.

Для наполнения и аннотирования корпуса студенческих работ первоначально была использована система BRAT (*Brat Rapid Annotation Tool*)[18]. С целью обеспечить возможность осуществления помимо частеречной разметки еще и разметки ошибок в систему были интегрированы соответствующие классификации тегов. В ходе практической работы выяснилось, что система BRAT имеет ряд критичных недостатков.

1. Минимальная длина аннотированной последовательности равняется одному символу, что повышает риск некорректного указания позиции аннотации при ручной разметке текстов.
2. Информация об аннотации хранится в отдельном файле, создаваемом для каждого текста в корпусе.
3. Текстовый формат хранения *standoff*, основанный на записи соответствия аннотации диапазону символов в тексте, в совокупности с хранением информации об аннотации в отдельном файле, создаваемом для каждого текста, является причиной крайне низкой скорости обработки запросов к корпусу.

[18] BRAT: https://brat.nlplab.org/index.html

4. Отсутствует поддержка баз данных — как направленных на учет пользователей, так и предназначенных для хранения информации об аннотациях.

Поэтому для работы с корпусом было разработано специализированное программное обеспечение. Серверная часть системы реализована на языке программирования *Python 3.10* и фреймворка *Django 4.0.3*. Клиентская часть реализована на гипертекстовом языке разметки HTML, каскадных таблицах стилей CSS и языка программирования *JavaScript* с применением фреймворка *Vue.js* и HTTP-клиента *Axios*.

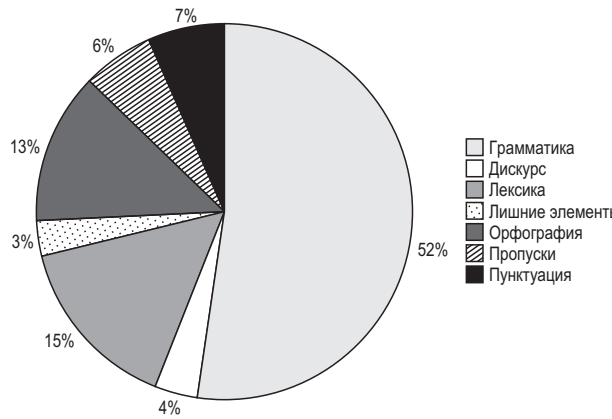
Веб-интерфейс пользователя позволяет проводить авторизацию и регистрацию пользователей, добавлять в корпус тексты, производить их разметку, выполнять проверку правильности разметки, вносить изменения в текст и разметку. Тексты и их разметки хранятся в реляционной базе данных *MySQL*.

2. Информационные возможности корпуса

Основное направление использования корпуса состоит в получении персонализированной или агрегированной информации, которая может применяться в процессе обучения как для корректировки образовательных траекторий, так и для проведения научных, в том числе педагогических, исследований. Использование реляционной базы данных позволяет формировать срезы данных по разным критериям, включая лонгитюдные — относительно отдельного студента, отдельной группы, определенного года изучения языка и т.д., используя любой элемент метаразметки, и строить для них самые разные статистики, позволяющие исследовать различные взаимосвязи.

Так, например, статистика укрупненных групп ошибок показывает, что более половины всех исправлений касаются грамматических ошибок (рис. 3).

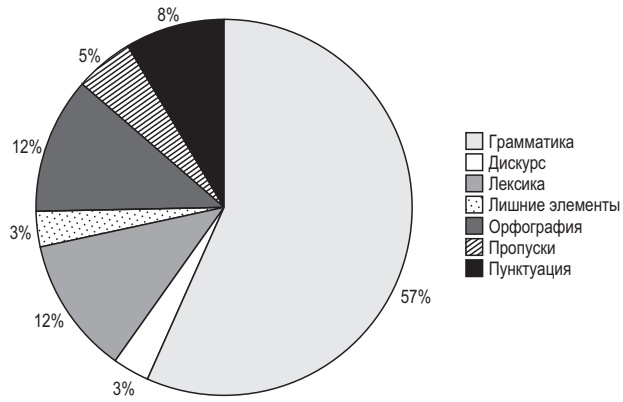
Рис. 3. **Распределение ошибок по корпусу в целом**



Анализ распространенности всех типов ошибок из классификации (91 наименование, см. Приложение 1) позволяет сделать вывод, что самые высокие доли в целом по корпусу имеют ошибки на орфографию, выбор лексемы и на порядок слов, при этом практически нет ошибок на использование превосходной степени прилагательных, управление прилагательных, конъюнктива и пассива. Причина низкой доли ошибок в формах сослагательного наклонения и пассивного залога, возможно, заключается в крайне редком их появлении в сочинениях студентов. Частотность употребления этих форм может стать предметом сравнительного анализа, включающего сопоставление с аналогичными текстами носителей немецкого или других иностранных языков.

Данные в ПАКТе можно собирать не только по корпусу в целом, но и по отдельным категориям, например для отдельного курса (рис. 4).

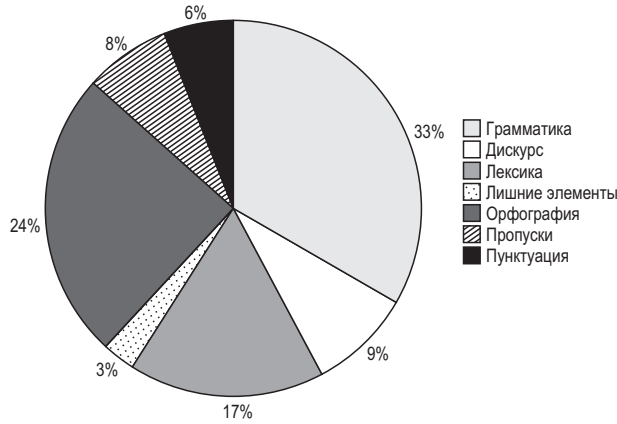
Рис. 4. **Распределение ошибок у студентов на первом году изучения языка**



В образовательных целях могут быть полезны данные по каждому отдельному обучающемуся. Например, в представленном на рис. 5 распределении по основным семи классам ошибок, допущенных студентом, который изучает немецкий язык первый год, очень мало ошибок категории «лишние элементы», но большое количество грамматических и орфографических ошибок.

Сравнение результатов данного студента (рис. 5) с типичным распределением ошибок у студентов на первом году изучения языка (рис. 4) показывает, что доля орфографических ошибок у выбранного студента (24%) существенно превышает «типичную» долю таких ошибок (12%), что может стать основанием для оказания преподавателем адресной поддержки в ча-

Рис. 5. **Распределение ошибок студента X, изучающего немецкий язык первый год**



сти орфографии этому студенту. Так как для всех студентов есть статистика по каждому из пунктов классификации (*N* = 91), можно организовать «точечную» работу над слабыми местами для каждого конкретного обучающегося. Преподаватель может также посмотреть общую статистику по конкретным ошибкам всей академической группы, в которой он ведет занятия, и выстраивать обучение с учетом этих данных с целью повышения эффективности образовательного процесса.

Другой пример использования агрегации данных — оценка распределения ошибок разных типов по категориям грубости (не влияет на восприятие / искажает смысл / затрудняет восприятие). Как видно на рис. 6, пунктуационные ошибки практически никогда не искажают смысла, и даже грамматические ошибки крайне редко оказываются критическими для понимания. Чаще всего искажают смысл ошибки на выбор лексемы, дискурсивные ошибки и пропуски (подробнее см. [Котюрова, Сафонов, 2022. С. 158]).

Благодаря многообразию обязательных и факультативных тегов в корпусе база данных позволяет строить статистические оценки и по другим параметрам. Так, метаразметка ПАКТа включает описания эмоционального состояния обучающегося в момент выполнения работы. Загружая текст в корпус, студент выбирает один из предложенных вариантов: подавленность / через силу / усталость / равнодушие / бодрость / приподнятое настроение.

На рис. 7 и 8 показано, как выглядит «эмоциональная картина» одной из академических групп.

Соотношение количества допущенных студентами ошибок с метаданными по настроению автора в момент написания текста представлено на рис. 9.

Рис. 6. **Распределение частот критичности по типам ошибок**

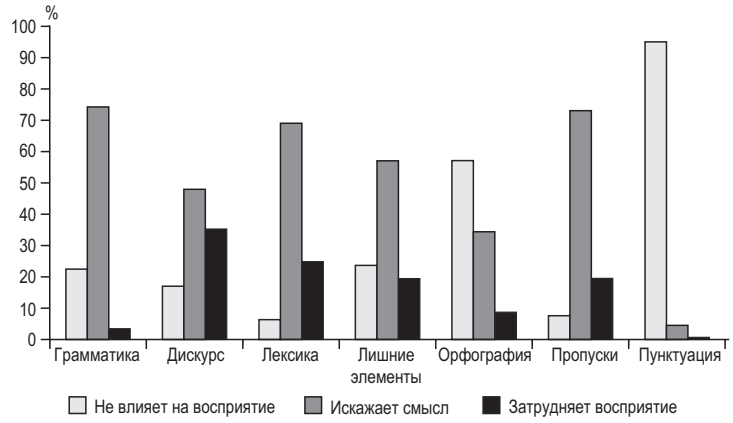
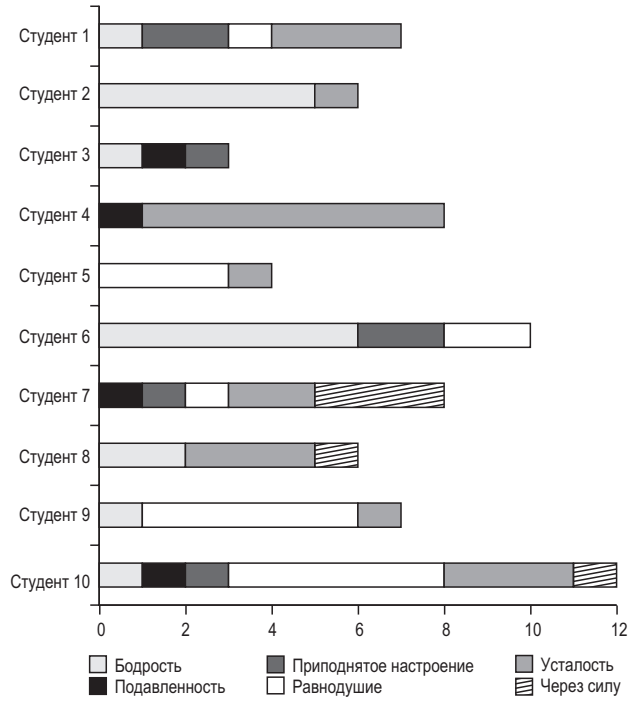


Рис. 7. **Эмоциональное состояние при написании разных работ. Данные одной академической группы**



Такие данные могут служить эмпирической базой для исследования психологического благополучия обучающихся и влияния эмоционального состояния на языковые компетенции.

ПАКТ имеет также тег «Самооценивание». На втором курсе студенты оценивают свои работы выше, чем на первом (рис. 10).

Рис 8. **Процентное соотношение показателей эмоционального состояния по группе в целом**

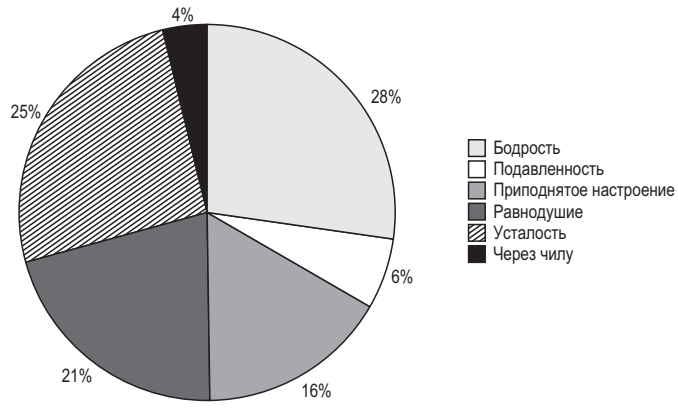
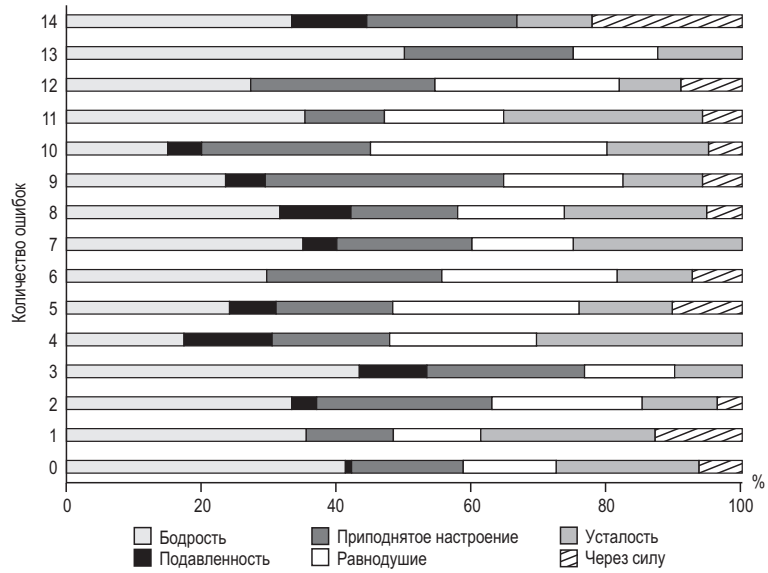


Рис. 9. **Соотношение эмоционального состояния в момент написания текста и количества допущенных ошибок**



Показатели корреляции количества ошибок с оценкой, которую выставляют себе сами студенты (рис. 11), свидетельствуют о том, что прямой зависимости между самооцениванием и количеством допущенных ошибок в общей картине корпуса не наблюдается. Диаметр кружка на диаграмме означает количество соответствующих текстов.

На рис. 12 приведена статистика ошибок, связанных с интерференцией с русским языком, интерференцией с английским

Рис. 10. **Процентное соотношение баллов самооценивания на 1-м и на 2-м курсе**

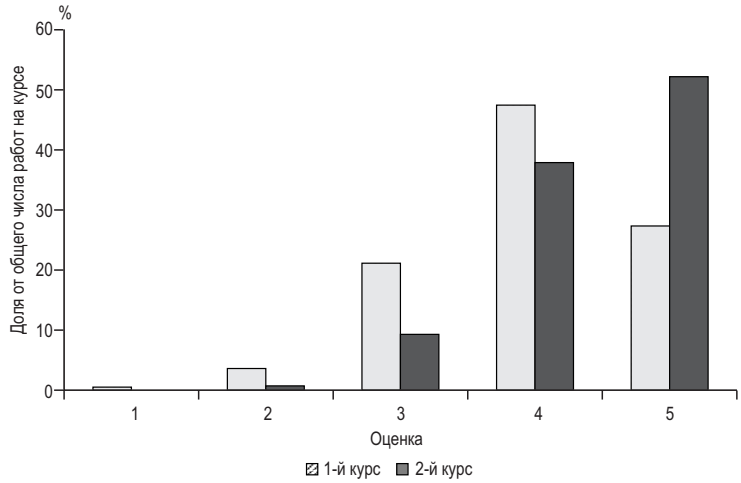
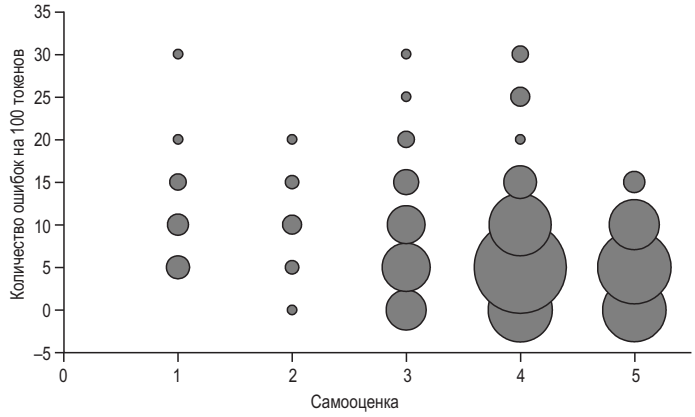


Рис. 11. **Корреляция количества ошибок на каждые 100 токенов с оценкой, которую выставляют себе сами студенты**



языком и опечатками. На диаграмме видно, что интерференция с родным языком вызывает в первую очередь грамматические, лексические и пунктуационные ошибки, интерференция с английским языком становится причиной орфографических ошибок чаще, чем других, но основной причиной орфографических ошибок все же являются опечатки, а не интерференция с английским языком.

Специалист, размечающий текст, ставит соответствующий тег, только если не сомневается, что ошибка вызвана одной из указанных трех причин; поэтому такую статистику нельзя считать полной по корпусу. Тем не менее поиск примеров по данному критерию представляет интерес для лингводидактического анализа, который могут проводить как студенты в ходе практических занятий, так и исследователи (табл. 1–3).

Рис. 12. **Количество ошибок, связанных с интерференцией или опечаткой**

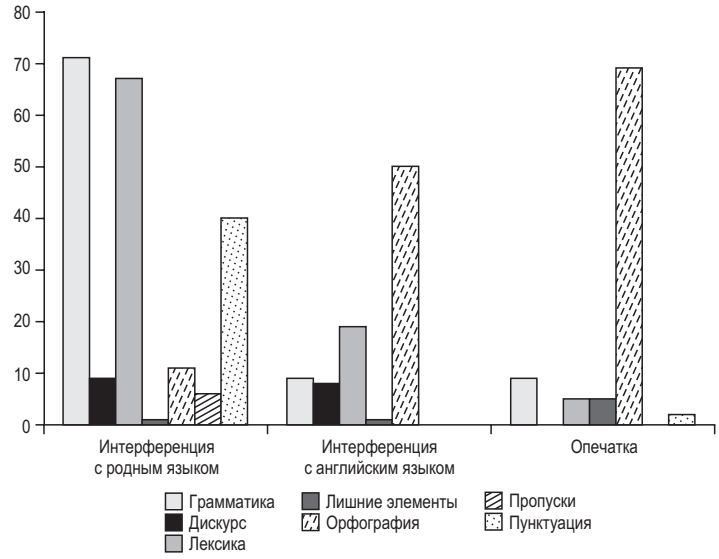


Таблица 1. **Пример выборки из базы данных по тегу «Опечатка»**

Dieser Mozaet sagte ihm, dass er selbst sein Leben in eine wertlose Existenz verwandelte.
Harry besteht nicht aus zwei Wesen, sondern aus hundert, aus lausenden.
Und die Natur muss ungestürt sein.
An die Wände stehen die Regalen für Dukumente und Bücher.
Im Bar gibt es drei Sessel, einen Bartressen und Lauptsprecher mit einem Musikzentrum.
Als Fazit gibt es gibt es Produktivität.

Таблица 2. **Пример выборки из базы данных по тегу «Интерференция с русским языком»**

Einer von Ihnen — Russisch, der andere-Amerikaner.
Ich habe über meine lieblingsskulpturen gesprochen, aber es gibt tatsächlich viel mehr von Ihnen.
Persönlich ich mag viele von ihnen wirklich.
Es ist symbolisch, dass nach der Idee einer der Fischer russisch und der andere amerikanisch ist.
Außerdem, gibt es Trainingsgeräte für den Sport und Spielplätze für Kinder.
Es ist wirklich ein sehr schönes Ort.

Таблица 3. **Пример выборки из базы данных по тегу «Интерференция с английским языком»**

Das ist meiner Mutters Geschenk.
Sculptur Fischer Autor: Rafael Consuegra aus der Partnerstadt Duluth, Minnesota, USA.

Zum Beispiel, die Skulptur Fischers".
Der Kai hat ein deutsches Geschenk auch.
Kinder, die gemobbt werden, finden es schwierig, sich auf Ihr Studium zu konzentrieren.
Das Studentenwohnheim Leben ist also sehr lustig, aber nicht so bequem.

3. Методики использования корпуса и его данных для образовательных и исследовательских целей

Использование лингвистических корпусов в иноязычном образовании — относительно новая и перспективная тема. Термин DDL (*Data Driven Learning*) становится понятным все более широкому кругу преподавателей иностранных языков, хотя пока имеет мало примеров практического использования на занятиях. Дидактический метод DDL предполагает, что обучающиеся в ходе анализа больших лингвистических данных индуктивно извлекают информацию о структурах, контекстах, использовании и функциях языковых элементов — информацию, которая не всегда содержится в учебных пособиях [Flinz, 2021. P. 3]. Несмотря на положительные отзывы о применении этого метода в преподавании иностранного языка, даже в европейских университетах DDL пока используют только отдельные преподаватели-энтузиасты. Причина, по мнению экспертов, состоит в нехватке квалицированных преподавателей, а внедрение нового метода требует специальной подготовки педагогических кадров [Ibid. P. 6].

Когда речь идет о подготовке педагогических кадров по иностранному языку, метод DDL может применяться не только к корпусам аутентичной устной и письменной речи, но и к ученическим корпусам: работая с массивом данных, в котором размечены ошибки и легко осуществлять поиск как по слову, так и по типу ошибки, анализируя типичные примеры неверного использования конкретных лексических и/или грамматических единиц, обучающиеся индуктивно сами формулируют проблемы при обучении языку и предлагают дидактические способы их решения.

Примером использования ученического корпуса на занятиях по иностранному языку является также подключение к работе генераторов заданий на базе работ обучающихся. В Научно-учебной лаборатории учебных корпусов НИУ ВШЭ разработан и активно используется в курсе английского языка *Testmaker* на базе англоязычного ученического корпуса REALEC [Кустова, 2017]. Этот генератор тестов позволяет легко создавать индивидуальные задания для каждого обучающегося как на основе его собственных текстов и допущенных в них ошибок, так и с использованием всего корпуса, в котором отбираются ошибки по определенной теме, вызывающей затруднения у данного студента. Большие перспективы открывает подклю-

чение к такому генератору заданий технологий искусственного интеллекта для выстраивания индивидуальной учебной траектории, что также частично уже реализовано Научно-учебной лабораторией учебных корпусов НИУ ВШЭ[19].

Изучающие иностранный язык студенты педагогического направления могут получить дополнительные преимущества от использования корпуса. ПАКТ создан таким образом, что тексты в него загружают как специалисты, работающие с корпусом, так и сами студенты — на занятии или дома. Выполнение загрузки текстов по определенным правилам и с разметкой некоторых обязательных и факультативных тегов развивает общепрофессиональные компетенции ОПК-9 «Информационно-коммуникационные технологии для профессиональной деятельности» и «Контроль и оценка формирования результатов образования», которые в Федеральном государственном образовательном стандарте высшего образования (бакалавриат) по направлению подготовки 44.03.05 «Педагогическое образование (с двумя профилями подготовки)»[20] определены как способность «понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности» и способность «осуществлять контроль и оценку формирования результатов образования обучающихся, выявлять и корректировать трудности в обучении».

Федеральный государственный образовательный стандарт высшего образования предусматривает также овладение универсальной компетенцией УК-6 «Самоорганизация и саморазвитие», согласно которой обучающийся должен быть «способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни». ПАКТ имеет тег «Самооценивание», заполнение которого развивает эту универсальную компетенцию обучающихся, заставляя их каждый раз при загрузке текста в корпус задать себе самому вопросы о качестве выполненной работы, о возможных причинах недостаточно высокого качества и способах его повышения.

Данные различных корпусов ученических текстов предоставляют богатый материал для лингводидактических исследований, поэтому они уже много лет активно используются зарубежными специалистами в области преподавания иностранных языков [Lüdeling, Walter, 2009; Ellis, 2008; Goscher, Stefanowitsch,

[19] REALEC Testmaker: https://github.com/nicklogin/REALEC-English-Test-Maker

[20] Федеральный государственный образовательный стандарт высшего образования (бакалавриат) по направлению подготовки 44.03.05 «Педагогическое образование (с двумя профилями подготовки)»: https://fgosvo.ru/uploadfiles/FGOS%20VO%203++/Bak/440305_B_3_15062021.pdf

2014]. На стыке психолингвистики и лингводидактики находятся исследования по овладению иностранным языком и его отдельными лексическими и грамматическими элементами [Bordag, Sieradz, 2012; Warditz, 2019]. Они помогают выявить типичные характеристики «иностранной немецкой речи», например недостаточное употребление форм пассива и модальных слов или регулярное позиционирование союза *aber* перед группой «подлежащее + сказуемое», а не после нее. В табл. 4 приводятся результаты сравнительного исследования на базе ученических корпусов: на схеме отражена степень недостаточного или чрезмерного использования отдельных служебных слов в речи на немецком языке носителей датского, английского, французского, польского и русского языков. Интенсивность окрашивания ячеек показывает, насколько сильно отличается интенсивность использования данного служебного слова от его бытования в речи носителей немецкого языка.

Таблица 4. **Сравнение использования некоторых служебных слов на немецком языке как иностранном [Lüdeling, Walter, 2009]**

		Dänisch	Englisch	Franzosisch	Polnisch	Russich
auch	+					
	–					
fur	+					
	–					
sind	+					
	–					
sich	+					
	–					
Ich	+					
	–					

Собранные в ПАКТе работы студентов, изучающих язык с нуля в течение нескольких лет, позволяют отследить типичные для носителей русского языка случаи интерференции и другие психолингвистические особенности, негативно или позитивно отражающиеся на овладении немецким языком как иностранным, и составить на основе таких исследований необходимые методические рекомендации (см., например, табл. 2, 3 и рис. 12).

Так, в исследовании использования заимствований в текстах студентов ПетрГУ, установлено, среди прочего, что изучающие немецкий язык русскоговорящие студенты практически не допускают ошибок в немецкой грамматике и орфографии в заимствованиях из разных языков, но нередко ошибаются, со-

четая эти заимствования с другими лексемами, например *gerechtfertigte Person* или *die finanzielle Person*[21].

Распространенная схема лингводидактических исследований — сопоставление немецкого языка как иностранного и как родного. Интересные возможности в этом отношении предоставляет сравнение ПАКТа Петрозаводского государственного университета с немецким корпусом Falko (*Fehlerannotiertes Lernerkorpus*) Берлинского университета им. Гумбольдта[22], а именно с тем его подкорпусом, где содержатся эссе-рассуждения немецких школьников выпускных классов в возрасте 17–19 лет. Сопоставляемые корпуса можно считать сбалансированными, поскольку в ПАКТе среди прочих собираются эссе по тем же четырем темам, которые представлены в Falko. Таким образом, в распоряжении исследователей оказываются подкорпуса эссе на одинаковые темы, написанные носителями немецкого языка и русскоговорящими студентами, изучающими немецкий язык в вузе 3–4 года. Исследования проводятся, в частности, в рамках курсовых и выпускных квалификационных работ. В 2021/2022 учебном году студенты ПетрГУ направления «Педагогическое образование» выполнили четыре курсовых и два выпускных квалификационных исследования с использованием данных ПАКТа.

Корпус как пример больших данных служит практическим материалом для образовательных и научных целей не только в разных областях лингвистики, но и в сфере информационных технологий. ПАКТ стал базой для исследования степени влияния разного рода ошибок в неаутентичных текстах на результаты работы автоматического частеречного разметчика [Котюрова, Щеголева, 2022]. Такой анализ выявил некоторые закономерности, учет которых позволит более эффективно организовать верификацию автоматической частеречной разметки в ученических корпусах на немецком языке и будет полезен для разработчиков автоматических частеречных разметчиков.

Работа над созданием ПАКТа объединила группу лингвистов и программистов Петрозаводского государственного университета на базе Центра искусственного интеллекта, созданного при вузе в июне 2020 г. Магистранты и преподаватели вуза активно разрабатывают инструментарий для формирования корпуса и осуществления поиска в нем. Корпус представляет собой большие данные, на которых можно не только отрабатывать технологии и строить экспериментальные программы, но и обучать искусственный интеллект. Так, в настоящее время создается от-

[21] Кайзер П. Выпускная квалификационная работа на тему «Сопоставительный анализ использования заимствований носителями немецкого языка и студентами, изучающими немецкий как иностранный».
[22] Falko: https://hu-berlin.de/falko

крытая библиотека для автоматизации оценивания учебных текстов на немецком языке. Для этого разрабатываются архитектуры нейронных сетей, которые затем обучаются на массиве данных корпуса. Корпус REALEC, послуживший прототипом ПАКТа, уже использует разработанные на его базе технологии искусственного интеллекта для автоматической идентификации и разметки ошибок разных типов [Torubarov, 2020], а также проводит другие исследования с привлечением технологий искусственного интеллекта [Vinogradova, Lyashevskaya, Panteleeva, 2017; Lyashevskaya, Panteleeva, Vinogradova, 2021], которые впоследствии планируется перенести и на материал больших данных ПАКТа на немецком языке.

В эпоху цифровой трансформации образования все активнее развиваются технологии искусственного интеллекта для обработки больших данных при решении задач управления образованием и автоматизации оценки работы учащихся [Уваров, Фрумин, 2019. С. 146]. В перспективе планируется сделать ПАКТ частью информационно-аналитической интегрированной системы вуза, что позволит собирать все данные о ходе учебной работы каждого обучающегося и делать их доступными для самих обучающихся, преподавателей и администрации. Разработка же систем (частичной) автоматизации оценивания работ обучающихся и генераторов тестов и упражнений позволит в будущем не только снять часть работы с преподавателя, но и обеспечить интеллектуальное управление учебным процессом.

4. Заключение

Таким образом, корпус студенческих текстов на иностранном языке как пример больших данных, собираемых в вузе, предоставляет очень широкие возможности в нескольких направлениях:

- его статистические данные и выборки на основе поисковых запросов являются базой для исследований лингвистов, психологов и методистов, а также для студентов лингвистических и педагогических направлений при написании курсовых и выпускных квалификационных работ; работа над совершенствованием веб-интерфейса и расширением технических возможностей корпуса предоставляет образовательные возможности и для студентов информационно-технологических направлений. Подключение к функциям ПАКТа технологий искусственного интеллекта (например, для автоматизированного исправления ошибок, индивидуального подбора упражнений и выставления оценок) открывает новые темы для научно-практических исследований и стартапов для магистрантов и аспирантов технических направлений;

- в образовательных целях корпус используется непосредственно на занятиях по иностранному языку или для выполнения домашних заданий, при этом кроме развития языковых компетенций происходит и формирование таких универсальных и общепрофессиональных компетенций, как «Информационно-коммуникационные технологии для профессиональной деятельности», «Самоорганизация и саморазвитие» и «Контроль и оценка формирования результатов образования».

В качестве перспектив развития корпуса студенческих текстов намечены разработка методических рекомендаций по использованию данных при подготовке педагогических кадров; создание автоматического генератора тестов для реализации индивидуальных образовательных траекторий; встраивание базы данных в информационно-аналитическую интегрированную систему вуза в целях образовательной аналитики. При этом особенно эффективной в аналитических целях могла бы стать реализация технологий искусственного интеллекта.

Благодарности

Сбор, аннотирование корпуса и разработка программного обеспечения для его использования проводятся при финансовой поддержке гранта Фонда содействия инновациям (соглашение № 4ГУКодИИС12-D7/72694).

Приложение 1. Теги разметки ошибок в корпусе

Основной класс ошибок	Уровень в иерархии	Название тега	Перевод названия тега на русский язык
1. Грамматика	1	Grammatik	Грамматика
	1.1	Substantive	Существительное
	1.1.1	Geschlecht	Род
	1.1.2	Numerus	Число
	1.1.3	Deklination	Склонение
	1.1.4	Rektion von Substantiven	Управление существительного
	1.2	Artikel	Артикль
	1.2.1	bestimmter Artikel	Определенный артикль
	1.2.2	unbestimmter Artikel	Неопределенный артикль
	1.2.3	Negationsartikel	Отрицательный артикль
	1.2.4	Nullartikel	Нулевой артикль
	1.3	Zahlwörter	Числительное
	1.3.1	Kardinalzahlen	Количественное числительное
	1.3.2	Ordinalzahlen	Порядковое числительное

Основной класс ошибок	Уровень в иерархии	Название тега	Перевод названия тега на русский язык
1. Грамматика	1.3.3	weitere Zahlwörter	Другие указатели на число
	1.4	Pronomen	Местоимение
	1.4.1	Personalpronomen	Личное местоимение
	1.4.2	Possessivpronomen	Притяжательное местоимение
	1.4.3	Demonstrativpronomen	Указательное местоимение
	1.4.4	Interrogativpronomen e	Вопросительное местоимение
	1.4.5	Reflexivpronomen	Возвратное местоимение
	1.4.6	Deklination des Pronomens	Склонение местоимения
	1.5	Verben	Глагол
	1.5.1	Konjugation	Спряжение
	1.5.1.1	Modalverben	Модальный глагол
	1.5.1.2	Starke Verben	Сильный глагол
	1.5.1.3	Trennbare Verben	Глагол с отделяемой приставкой
	1.5.1.4	Untrennbare Verben	Глагол с неотделяемой приставкой
	1.5.2	Zeitform	Временная форма
	1.5.2.1	Wahl der Zeitform	Выбор временной формы
	1.5.2.1.1	Präsens (Zeit)	Презенс (время)
	1.5.2.1.2	Präteritum (Zeit)	Претерит (время)
	1.5.2.1.3	Perfekt (Zeit)	Перфект (время)
	1.5.2.1.4	Plusquamperfekt (Zeit)	Плюсквамперфект (время)
	1.5.2.1.5	Futurum I (Zeit)	Футурум I (время)
	1.5.2.2	Bildung der Zeitform	Образование временной формы
	1.5.2.2.1	Präsens (Bildung)	Презенс (форма)
	1.5.2.2.2	Präteritum (Bildung)	Претерит (форма)
	1.5.2.2.3	Perfekt (Bildung)	Перфект (форма)
	1.5.2.2.4	Plusquamperfekt (Bildung)	Плюсквамперфект (форма)
	1.5.2.2.5	Futurum I (Bildung)	Футурум I (форма)
	1.5.3	Modus	Наклонение
	1.5.3.1	Imperativ	Императив
	1.5.3.2	Konjunktiv	Конъюнктив
	1.5.4	Genus	Залог
	1.5.4.1	Aktiv	Активный залог
	1.5.4.2	Passiv	Пассивный залог
	1.5.4.3	Zustandspassiv	Пассив состояния
	1.5.5	Rektion von Verben	Управление глаголов
	1.6	Partizipien	Причастия
	1.6.1	Partizip I	Причастие I

Основной класс ошибок	Уровень в иерархии	Название тега	Перевод названия тега на русский язык
1. Грамматика	1.6.2	Partizip II	Причастие II
	1.7	Präpositionen	Предлоги
	1.7.1	Wahl der Präposition	Выбор предлога
	1.7.2	Präpositionen mit einem bestimmten Kasus	Предлог с определенным падежом
	1.7.3	Wechselpräpositionen	Предлог, управляющий несколькими падежами
	1.8	Konjunktionen	Союзы
	1.9	Adjektive	Прилагательное
	1.9.1	Deklination von Adjektiven	Склонение прилагательного
	1.9.2	Komparativform	Сравнительная степень
	1.9.3	Superlativform	Превосходная степень
	1.9.4	Rektion von Adjektiven	Управление прилагательного
	1.10	Adverbien	Наречия
	1.11	Wortfolge	Порядок слов
	1.11.1	Verbstellung	Место глагола
	1.11.1.1	direkte Wortfolge	Прямой порядок слов
	1.11.1.2	indirekte Wortfolge	Обратный порядок слов
	1.11.1.3	Satzklammer	Рамочная конструкция
	1.11.1.4	Wortfolge in einer Satzreihe	Порядок слов в придаточном предложении
	1.11.2	Wortstellung	Место второстепенных членов предложения
	1.11.2.1	Wortstellung in der Negation	Порядок слов при отрицании
	1.11.2.2	Thema-Rhema-Gliederung	Тема-рематическое членение предложения
	1.12	Vergleichskonstruktionen	Сравнительные конструкции
	1.13	Infinitivkonstruktionen	Инфинитивные конструкции
	1.13.1	Infinitivkonstruktionen mit zu	Инфинитивные конструкции с *zu*
	1.13.2	Infinitivkonstruktionen ohne zu	Инфинитивные конструкции без *zu*
2. Лексика	2	Lexik	Лексика
	2.1	Lexemauswahl	Выбор лексемы
	2.2	Feste Wendungen	Устойчивые обороты
	2.3	Wortbildung	Словообразование
	2.3.1	Derivative Suffixe	Словообразовательные суффиксы
	2.3.2	Derivative Präffixe	Словообразовательные префиксы
	2.3.3	Zusammengesetzte Wörter	Сложные слова

Основной класс ошибок	Уровень в иерархии	Название тега	Перевод названия тега на русский язык
3. Дискурс	3	Diskurs	Дискурс
	3.1	Logik	Логика
	3.1.1	Konnektoren	Соединительные элементы
	3.2	Sprachstil	Стиль
4. Пропуски	4	Auslassungen	Пропуски
5. Лишние элементы	5	Überflüssige Elemente	Лишние элементы
6. Орфография	6	Orthographie	Орфография
7. Пунктуация	7	Interpunktion	Пунктуация

Литература

1. Виноградова О.И. (2021) Работа с языковыми корпусами в изучении иностранных языков, в обучении им и в их использовании. *Межкультурное пространство: лингвистический и дидактический аспекты. Материалы научно-практической онлайн-конференции. Ч. 1: Пленарное заседание и секция «Межкультурная дидактика»*. Петрозаводск: Изд-во ПетрГУ, сс. 20–29.
2. Дворецкая И.В., Карлов И.А., Кочак Э., Савицкий К.Л. (2022) *Измерение перехода школы к цифровой трансформации образования: опыт, трудности, результаты и возможности.* М.: НИУ ВШЭ.
3. Другова Е.А., Велединская С.Б., Журавлева И.И. (2021) Развивая цифровую педагогику: вклад образовательного дизайна. Рецензия на книгу: Beetham H., Sharpe R. (2020) Rethinking Pedagogy for a Digital Age. *Вопросы образования / Educational Studies Moscow*, № 4, сс. 333–354. https://doi.org/10.17323/1814-9545-2021-4-333-354
4. Ищенко А.А. (2020) Большие данные как информационная база для анализа качества образования. *Научный потенциал*, № 2 (29), сс. 10–14.
5. Княгинин В.Н., Идрисов Г.И., Кузьмина А.С., Рожкова Е.С., Султанов Д.К. (2017) *Новая технологическая революция: вызовы и возможности для России. Экспертно-аналитический доклад*. М.: Центр стратегических разработок.
6. Котюрова И.А., Сафонов Г.Р. (2022) Анализ степени грубости ошибок в студенческих сочинениях (корпусное исследование). *Отражения: язык и культура в синхронии и диахронии. Электронный сборник научных статей*. Петрозаводск: Петрозаводский государственный университет, сс. 154–161.
7. Котюрова И.А., Щеголева Л.В. (2022) Анализ некорректной работы POS-разметчиков в корпусе немецких ученических текстов с лингвистическими ошибками. *Научный результат. Вопросы теоретической и прикладной лингвистики*, т. 8, № 3, сс. 87–99. https://doi.org/10.18413/2313-8912-2022-8-3-0-6
8. Кузнецова Э.К., Шангараева Л.Ф. (2021) Возможности использования лингвистического корпуса в обучении иностранным языкам в неязыковом вузе. *Язык. Культура. Медиакоммуникация*, т. 1, № 2, сс. 45–50.

9. Кустова М.А. (2017) Автоматическая разработка учебных тестов по английскому языку на основе корпуса. *Труды международной научной конференции «Корпусная лингвистика — 2017» (Санкт-Петербург, 27–30 июня 2017 г.)*. СПб.: СПбГУ, ИЛИ РАН, РГПУ им. А.И. Герцена, сс. 236–240.
10. Радаев В.В., Медведев С.А., Талалакина Е.В., Дементьев А.В. (2018) Пять моих главных вызовов в преподавании. Круглый стол. (Москва, НИУ ВШЭ, 8 сентября 2017 г.). *Вопросы образования / Educational Studies Moscow*, № 1, сс. 200–233. https://doi.org/10.17323/1814-9545-2018-1-200-233
11. Уваров А.Ю., Фрумин И.Д. (ред.) (2019) *Трудности и перспективы цифровой трансформации образования*. М.: НИУ ВШЭ.
12. Фиофанова О.А. (ред.) (2021) *Большие данные в образовании: доказательное развитие образования. Сборник научных статей II Международной конференции (Москва, 15 октября 2021 г.)*. М.: Дело.
13. Фиофанова О.А. (2020) *Анализ больших данных в сфере образования: методология и технологии*. М.: Дело.
14. Черепанова А.И. (2015) Корпусные технологии в изучении английского языка. *Иностранный язык в контексте проблем профессиональной коммуникации. Материалы II Международной научной конференции (Томск, 27–29 апреля 2015 г.)*, Томск: Национальный исследовательский Томский политехнический университет, сс. 276–277.
15. Bates A.T., Bates A.W. (2015) *Teaching in a Digital Age. Guidelines for Designing Teaching and Learning*. Vancouver, BC: Tony Bates Associates Ltd.
16. Bawa P. (2020) Learning in the Age of SARS-COV-2: A Quantitative Study of Learners' Performance in the Age of Emergency Remote Teaching. *Computers and Education Open*, vol. 1, Article no 100016. https://doi.org/10.1016/j.caeo.2020.100016
17. Bordag B., Sieradz M. (2012) Erwerb von Perfekt und Passiv bei DaF-Lernern: Eine Korpusstudie. *German as a Foreign Language*, no 1. Available at: http://www.gfl-journal.de/1-2012/bordag-sieradz.pdf (accessed 23 November 2022).
18. Ellis N.C. (2008) Usage-Based and Form-Focused Language Acquisition: The Associative Learning of Constructions, Learned Attention, and the Limited L2 Endstate. *Handbook of Cognitive Linguistics and Second Language Acquisition* (eds P. Robinson, N.C. Ellis), London: Routledge/Taylor & Francis Group, pp. 372–405.
19. Flinz C. (2021) KORPORA in DaF und DaZ: Theorie und Praxis. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, bd. 26, nr 1, ss. 1–43.
20. Goschler J., Stefanowitsch A. (2014) Korpora in der Weitspracherwerbsforschung: Sieben Probleme aus korpuslinguistischer Sicht. *Zweitspracherwerb im Jugendalter* (eds B. Ahrenholz, P. Grommes), Berlin/Boston: Mouton de Gruyter, ss. 341–361. https://doi.org/10.1515/9783110318593
21. Hou J., Koppatz M., Hoya Quecedo J.M., Stoyanova N., Kopotev M., Yangarber R. (2019) Modeling Language Learning Using Specialized Elo Ratings. *Innovative Use of NLP for Building Educational Applications* (eds H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, T. Zesch), Stroudsburg, PA: Association for Computational Linguistics, pp. 494–506. http://dx.doi.org/10.18653/v1/W19-4451
22. Katinskaia A., Nouri J., Yangarber R. (2018) Revita: A Language Learning Platform at the Intersection of ITS and CALL. Proceedings of the *Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan, 2018, 7–12 May)*, pp. 4084–4093.
23. Katinskaia A., Nouri J., Yangarber R. (2017) Revita: A System for Language Learning and Supporting Endangered Languages. Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017. *Linköping Electronic Conference Proceedings*, vol. 134, pp. 27–35.

24. Kormacheva D., Pivovarova L., Kopotev M. (2014) Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams. Proceedings of the *Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (Tübingen, Germany, 2014, 11 August)*, pp. 27–33.
25. Kotiurova I., Trenina P. (2022) Comparative Analysis of Automatic POS Taggers Applied to German Learner Texts. Proceedings of the *31st Conference of Open Innovations Association (FRUCT) (Helsinki, Finland, 2022, 27–29 April)*, pp. 115–124. https://doi.org/10.23919/FRUCT54823.2022.9770886
26. Langthaler M., Bazafkan H. (2020) *Digitalization, Education and Skills Development in the Global South: An Assessment of the Debate with a Focus on Sub-Saharan Africa. ÖFSE Briefing Paper no 28.* Vienna: Austrian Foundation for Development Research.
27. Lüdeling A., Walter M. (2009) Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung. Stark erweiterte Fassung von Lüdeling/Walter (erscheint) Korpuslinguistik. *HSK 19. Deutsch als Fremdsprache* (ed. G. Helbig), Berlin: Mouton de Gruyter.
28. Lyashevskaya O., Panteleeva I., Vinogradova O. (2021) Automated Assessment of Learner Text Complexity. *Assessing Writing*, vol. 49, no 4, Article no 100529. https://doi.org/10.1016/j.asw.2021.100529
29. Torubarov I. (2020) *Automated Error Detection and Correction System for Learner Essays in English Produced by Students with Russian Native Language*. Available at: https://www.hse.ru/en/edu/vkr/368892883 (accessed 24 November 2022).
30. Vinogradova O., Lyashevskaya O., Panteleeva I. (2017) Multi-Level Student Essay Feedback in a Learner Corpus. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017). Iss. 16 in 2 vols. Vol. 1: Computational Linguistics: Practical Applications*, pp. 373–386.
31. Vinogradova O.I., Viklova A., Smilga V. (2021) Punctuation in L2 English: Computational Methods Applied in the Study of L1 Interference. *Emerging Writing Research from the Russian Federation* (ed. L.A. Squires), Fort Collins, CO: WAC Clearinghouse, pp. 211–233. https://doi.org/10.37514/INT-B.2021.1428
32. Warditz V. (2019) Russisch als Migrationssprache in Deutschland: Zur Typologie des Mikrosprachwandels (Eine systemlinguistische Studie). *Handbuch des Russischen in Deutschland. Migration–Mehrsprachigkeit–Spracherwerb* (eds K. Witzlack-Makarevich, N. Wulff), Berlin: Frank & Timme, pp. 283–302.

References

Bates A.T., Bates A.W. (2015) *Teaching in a Digital Age. Guidelines for Designing Teaching and Learning.* Vancouver, BC: Tony Bates Associates Ltd.

Bawa P. (2020) Learning in the Age of SARS-COV-2: A Quantitative Study of Learners' Performance in the Age of Emergency Remote Teaching. *Computers and Education Open*, vol. 1, Article no 100016. https://doi.org/10.1016/j.caeo.2020.100016

Bordag B., Sieradz M. (2012) Erwerb von Perfekt und Passiv bei DaF-Lernern: Eine Korpusstudie. *German as a Foreign Language*, no 1. Available at: http://www.gfl-journal.de/1-2012/bordag-sieradz.pdf (accessed 23 November 2022).

Cherepanova A.I. (2015) Korpusnye tekhnologii v izuchenii angliyskogo yazyka [Corpus Technologies in Learning English]. Proceedings of the *II International Scientific Conference "Foreign Language in the Context of Problems of Professional Communication" (Tomsk, 2015, April 27–29)*, Tomsk: National Research Tomsk Polytechnic University, pp. 276–277.

Drugova E.A., Veledinskaya S.B., Zhuravleva I.I. (2021) Razvivaya tsifrovuyu peda-gogiku: vklad obrazovatel'nogo dizayna. Retsenziya na knigy: Beetham H., Sharpe R. (2020) Rethinking Pedagogy for a Digital Age [The Role of In-

structional Design in Promoting Digital Pedagogy. Review of the book: Beetham H., Sharpe R. (2020) Rethinking Pedagogy for a Digital Age]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 4, pp. 333–354. https://doi.org/10.17323/1814-9545-2021-4-333-354

Dvoretskaya I.V., Karlov I.A., Kochak E., Savitsky K.L. (2022) *Izmerenie perekhoda shkoly k tsifrovoy transformatsii obrazovaniya: opyt, trudnosti, rezul'taty i voznozhnosti* [Measuring the Transition of the School to the Digital Transformation of Education: Experience, Difficulties, Results and Opportunities]. Moscow: HSE.

Ellis N.C. (2008) Usage-Based and Form-Focused Language Acquisition: The Associative Learning of Constructions, Learned Attention, and the Limited L2 Endstate. *Handbook of Cognitive Linguistics and Second Language Acquisition* (eds P. Robinson, N.C. Ellis), London: Routledge/Taylor & Francis Group, pp. 372–405.

Fiofanova O.A. (ed.) (2021) *Bol'shie dannye v obrazovanii: dokazatel'noe razvitie obrazovaniya. Sbornik nauchnykh statey II Mezhdunarodnoy konferentsii (Moscow, 15 oktyabrya 2021 g.)* [Big Data in Education: Evidentiary Development of Education. Proceedings of the II International Conference (Moscow, 2021, October 15)]. Moscow: Delo.

Fiofanova O.A. (2020) *Analiz bol'shikh dannykh v sfere obrazovaniya: metodologiya i tekhnologii* [Big Data Analysis in the Field of Education: Methodology and Technologies. Moscow: Delo.

Flinz C. (2021) KORPORA in DaF und DaZ: Theorie und Praxis. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, bd. 26, nr 1, ss. 1–43.

Goschler J., Stefanowitsch A. (2014) Korpora in der Weitspracherwerbsforschung: Sieben Probleme aus korpuslinguistischer Sicht. *Zweitspracherwerb im Jugendalter* (eds B. Ahrenholz, P. Grommes), Berlin/Boston: Mouton de Gruyter, ss. 341–361. https://doi.org/10.1515/9783110318593

Hou J., Koppatz M., Hoya Quecedo J.M., Stoyanova N., Kopotev M., Yangarber R. (2019) Modeling Language Learning Using Specialized Elo Ratings. *Innovative Use of NLP for Building Educational Applications* (eds H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, T. Zesch), Stroudsburg, PA: Association for Computational Linguistics, pp. 494–506. http://dx.doi.org/10.18653/v1/W19-4451

Ishchenko A.A. (2020) Bol'shie dannye kak informatsionnaya baza dlya analiza kachestva obrazovaniya [Big Data as an Information Base to Analyze the Quality of Education]. *Nauchny potentsial*, no 2 (29), pp. 10–14.

Katinskaia A., Nouri J., Yangarber R. (2018) Revita: A Language Learning Platform at the Intersection of ITS and CALL. Proceedings of the *Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan, 2018, 7–12 May)*, pp. 4084–4093.

Katinskaia A., Nouri J., Yangarber R. (2017) Revita: A System for Language Learning and Supporting Endangered Languages. Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017. *Linköping Electronic Conference Proceedings*, vol. 134, pp. 27–35.

Knyaginin V.N., Idrisov G.I., Kuzmina A.S., Rozhkova E.S., Sultanov D.K. (2017) *Novaya tekhnologicheskaya revolyutsiya: vyzovy i vozmozhnosti dlya Rossii. Ekspertno-analiticheskiy doklad* [New Technological Revolution: Challenges and Opportunities for Russia. Expert-Analytical Report]. Moscow: The Center for Strategic Research.

Kormacheva D., Pivovarova L., Kopotev M. (2014) Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams. Proceedings of the *Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (Tübingen, Germany, 2014, 11 August)*, pp. 27–33.

Kotiurova I.A., Safonov G.R. (2022) Analiz stepeni grubosti oshibok v studencheskikh sochineniyakh (korpusnoe issledovanie) [Analysis of the Degree of Errors in Student Essays (Corpus-Based Study)]. *Otrazheniya: yazyk i kul'tura v sinkhronii i diakhronii. Elektronny sbornik nauchnykh statey* [Reflections: Language and Culture in Synchrony and Diachrony. Electronic Collection of Scientific Articles]. Petrozavodsk: Petrozavodsk State University, pp. 154–161.

Kotiurova I.A., Shchegoleva L.V. (2022) Analiz nekorrektnoy raboty POS-razmetchikov v korpuse nemetskikh uchenicheskikh tekstov s lingvisticheskimi oshibkami [Analysis of Incorrect POS-Tagging in Student Texts with Linguistic Errors in German]. *Research Result. Theoretical and Applied Linguistics*, vol. 8, no 3, pp. 87–99. https://doi.org/10.18413/2313-8912-2022-8-3-0-6

Kotiurova I., Trenina P. (2022) Comparative Analysis of Automatic POS Taggers Applied to German Learner Texts. Proceedings of the *31st Conference of Open Innovations Association (FRUCT) (Helsinki, Finland, 2022, 27–29 April)*, pp. 115–124. https://doi.org/10.23919/FRUCT54823.2022.9770886

Kustova M.A. (2017) Avtomaticheskaya razrabotka uchebnykh testov po angliyskomu yazyku na osnove korpusa [Automated Development of Educational Corpus-Based Tests in English]. Proceedings of the *International Conference "Corpus Linguistics–2017" (Saint-Petersburg, 2017, 27–30 June)*. St. Petersburg: Saint-Petersburg State University, Institute for Linguistic Studies RAS, Herzen University, pp. 236–240.

Kuznetsova E., Shangaraeva L. (2021) Vozmozhnosti ispol'zovaniya lingvisticheskogo korpusa v obuchenii inostrannym yazykam v neyazykovom vuze [Possibilities of Using the Linguistic Corpus in Foreign Language Teaching in a Non-Lingustic University]. *Yazyk. Kul'tura. Mediakommunikatsiya*, vol. 1, no 2, pp. 45–50.

Langthaler M., Bazafkan H. (2020) *Digitalization, Education and Skills Development in the Global South: An Assessment of the Debate with a Focus on Sub-Saharan Africa. ÖFSE Briefing Paper no 28.* Vienna: Austrian Foundation for Development Research.

Lyashevskaya O., Panteleeva I., Vinogradova O. (2021) Automated Assessment of Learner Text Complexity. *Assessing Writing*, vol. 49, no 4, Article no 100529. https://doi.org/10.1016/j.asw.2021.100529

Lüdeling A., Walter M. (2009) Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung. Stark erweiterte Fassung von Lüdeling/Walter (erscheint) Korpuslinguistik. *HSK 19. Deutsch als Fremdsprache* (ed. G. Helbig), Berlin: Mouton de Gruyter.

Radaev V., Medvedev S., Talalakina E., Dementiev A. (2018) Pyat' moikh glavnykh vyzovov v prepodavanii. Krugly stol (Moscow, NIU VShE, 2017, 8 September) [My Five Major Challenges as a Teacher. Discussion (Moscow, HSE, 2017, September 8). *Voprosy obrazovaniya / Educational Studies Moscow*, no 1, pp. 200–233. https://doi.org/10.17323/1814-9545-2018-1-200-233

Torubarov I. (2020) *Automated Error Detection and Correction System for Learner Essays in English Produced by Students with Russian Native Language*. Available at: https://www.hse.ru/en/edu/vkr/368892883 (accessed 24 November 2022).

Uvarov A.Yu., Froumin I.D. (eds) (2019) *Trudnosti i perspektivy tsifrovoy transformatsii obrazovaniya* [Difficulties and Prospects of Digital Transformation of Education]. Moscow: HSE.

Vinogradova O.I. (2021) Rabota s yazykovymi korpusami v izuchenii inostrannykh yazykov, v obuchenii im i v ikh ispol'zovanii [Using Language Corpora in Learning Foreign Languages and Teaching Them]. *Mezhkul'turnoe prostranstvo: lingvisticheskiy i didakticheskiy aspekty. Materialy nauchno-prakticheskoy onlain-konferentsii. Ch. I: Plenarnoe zasedanie i sektsiya "Mezhkul'turnaya didaktika"* [Intercultural Space: linguistic and Didactic Aspects. Materials of the Scientific and Practical Online Conference. Part 1: Plenary Session and Section "Intercultural Didactics"]. Petrozavodsk: Petrozavodsk State University, pp. 20–29.

Vinogradova O.I., Viklova A., Smilga V. (2021) Punctuation in L2 English: Computational Methods Applied in the Study of L1 Interference. *Emerging Writing Research from the Russian Federation* (ed. L.A. Squires), Fort Collins, CO: WAC Clearinghouse, pp. 211–233. https://doi.org/10.37514/INT-B.2021.1428

Vinogradova O., Lyashevskaya O., Panteleeva I. (2017) Multi-Level Student Essay Feedback in a Learner Corpus. Computational Linguistics and Intellectual Technologies. *Papers from the Annual International Conference "Dialogue" (2017). Iss. 16 in 2 vols. Vol. 1: Computational Linguistics: Practical Applications*, pp. 373–386.

Warditz V. (2019) Russisch als Migrationssprache in Deutschland: Zur Typologie des Mikrosprachwandels (Eine systemlinguistische Studie). *Handbuch des Russischen in Deutschland. Migration–Mehrsprachigkeit–Spracherwerb* (eds K. Witzlack-Makarevich, N. Wulff), Berlin: Frank & Timme, pp. 283–302.