оригинальная статья
https://elibrary.ru/bueqak

Языковое образование онлайн: принципы создания размеченного корпуса ошибок в специализированном английском языке русскоязычных работников IT-сферы

Виноградова Юлия Сергеевна
НИУ ВШЭ – Санкт-Петербург, Россия, Санкт-Петербург
https://orcid.org/0009-0006-6602-690X
Ильченко Игорь Владимирович
НИУ ВШЭ – Санкт-Петербург, Россия, Санкт-Петербург
https://orcid.org/0009-0002-9685-8900
igvladilchenko@gmail.com
Ширяева Полина Сергеевна
НИУ ВШЭ – Санкт-Петербург, Россия, Санкт-Петербург
https://orcid.org/0009-0006-5300-5325
Горина Мария Сергеевна
ООО «Яндекс», Россия, Москва
https://orcid.org/0009-0008-0246-7807

Аннотация: Сегодня создание учебных корпусов представляет большой интерес для многих лингвистов. В статье рассматривается значимость учебных корпусов в современной лингвистике и педагогике, их важность как инструмента для выявления типичных ошибок в речи на неродном языке, анализа проблем в освоении языка и создания эффективных методик обучения второму языку. В работе приводятся существующие обзоры на учебные корпусы, а также краткий обзор работ, посвященных классификации ошибок. Наше исследование посвящено созданию устного корпуса ошибок русскоязычных студентов, изучающих специализированный английский язык в сфере информационных технологий. Исследование проводится на материале 50 видеозаписей занятий, на которых студенты общаются с англоговорящими IT-специалистами и выполняют задания на английском языке в формате диалога. Цель создания корпуса – выявить основные трудности в использовании английского языка взрослыми, работающими в IT-сфере. В результате для разметки корпуса была разработана классификация самых частотных ошибок носителей русского языка в речи на английском языке и система тегов для них. Все ошибки делятся по уровню языка на морфологические, синтаксические, лексические и фонетические. Сделан вывод, что созданная в рамках работы классификация ошибок может быть использована для аннотации будущих учебных корпусов речи носителей русского языка на английском языке, а также для автоматизации тегирования ошибок.

Ключевые слова: корпус ошибок, учебный корпус, аннотация, тегирование, обнаружение ошибок, английский как иностранный

Цитирование: Виноградова Ю. С., Ильченко И. В., Ширяева П. С., Горина М. С. Языковое образование онлайн: принципы создания размеченного корпуса ошибок в специализированном английском языке русскоязычных работников IT-сферы. *Виртуальная коммуникация и социальные сети.* 2024. Т. 3. № 3. С. 245–254. https://doi.org/10.21603/2782-4799-2024-3-3-245-254

Поступила в редакцию 01.06.2024. Принята после рецензирования 13.08.2024. Принята в печать 19.08.2024.

VIRTUAL COMMUNICATION
& SOCIAL NETWORKS 2024, 3(3)

full article

Online Language Education: Principles of Creating a Marked Corpus of Learners' Mistakes

Iuliia S. Vinogradova
National Research University Higher School of Economics (HSE University), Russia, St. Petersburg
https://orcid.org/0009-0006-6602-690X
Igor V. Ilchenko
National Research University Higher School of Economics (HSE University), Russia, St. Petersburg
https://orcid.org/0009-0002-9685-8900
igvladilchenko@gmail.com
Polina S. Shiriaeva
National Research University Higher School of Economics (HSE University), Russia, St. Petersburg
https://orcid.org/0009-0006-5300-5325
Maria S. Gorina
OOO Yandex, Russia, Moscow
https://orcid.org/0009-0008-0246-7807

Abstract: Today, academic corpora are a matter of great interest to many linguists. The article examines the importance of academic corpora in modern linguistics and pedagogy as a tool for identifying typical errors in the speech of foreign language students. They reveal problems in foreign language acquisition and create new teaching methods. The paper reviews the existing academic corpora and error classifications. The authors developed a corpus of oral speech errors typical of Russian-speaking IT students that study English for IT purposes. The corpus relied on 50 video recordings of in-class activities in which students communicated with English-speaking IT specialists and made up dialogues in English. The corpus made it possible to identify the main difficulties experienced by adult learners of English for IT purposes. It involves a tagged classification of typical errors divided by language level into morphological, syntactic, lexical, and phonetic ones. The research demonstrated good prospects for developing other academic speech corpora with automated error tagging based on the speech of English learners.

Keywords: corpus of errors, learner corpora, abstract, tagging, error detection, English as a foreign language

Citation: Vinogradova Iu. S., Ilchenko I. V., Shiriaeva P. S., Gorina M. S. Online Language Education: Principles of Creating a Marked Corpus of Learners' Mistakes. *Virtual Communication and Social Networks*, 2024, 3(3): 245–254. (In Russ.) https://doi.org/10.21603/2782-4799-2024-3-3-245-254

Received 1 Jun 2024. Accepted after review 13 Aug 2024. Accepted for publication 19 Aug 2024.

Введение

Во многих лингвистических работах появление корпусов сравнивают с новой эпохой в лингвистике [Грудева и др. 2018: 64; Дмитриев и др. 2020]. Действительно, их создание и изучение дали новый импульс лингвистическим исследованиям в XXI в. [Рахилина 2016: 20–21]. За последние 50 лет появилось такое количество корпусов, что для них уже составлена своя типология по различным критериям: языку, жанру текстов, задачам корпуса и другим признакам [Копотев 2014; Khokhlova 2023: 59–61]. В этой статье мы остановимся на одном из существующих типов корпуса – учебном, или корпусе ошибок.

Сегодня создание учебных корпусов представляет большой интерес для лингвистов по нескольким причинам:

1) такие корпусы являются ценным материалом для определения типичных ошибок, совершаемых студентами при изучении неродного языка: фиксируя речь студентов на неродном языке, корпус ошибок дает когнитивным лингвистам и психолингвистам возможность проанализировать не только совершаемые студентами ошибки, но и природу их появления (например, установить, в каких случаях имеет влияние языковая интерференция),

их обусловленность закономерностями организации билингвального языкового сознания.

2) «подсвечивая» проблемные места в речи студентов, учебные корпусы позволяют выявить трудности в освоении языка, с которыми сталкиваются инофоны, что может принести особую пользу преподавателям иностранных языков и методистам при разработке более эффективных методик обучения второму языку [Колмогорова 2019].
3) современные исследования в области второго языка показывают, что корпусы ошибок обладают огромным потенциалом для понимания особенностей освоения языка, речевого онтогенеза [Захарова 2016].

Основная цель создания нашего корпуса ошибок – определить общие ключевые трудности в использовании английского языка у взрослых обучающихся, работающих в IT-сфере. В результате обработки корпуса мы ожидаем получить статистику ошибок, с опорой на которую можно будет сделать выводы о тех аспектах английского языка, которые «проседают» у студентов курса. Стоит отметить, что все студенты находятся на начальных этапах курса, что поможет понять основные потребности клиентов платформы, чтобы затем адаптировать к этим потребностям методическую организацию курса. Для достижения цели исследования необходимо выполнить следующие задачи:

1. Провести обзор имеющихся работ, посвященных учебным корпусам и типам ошибок в речи на английском языке как иностранном.
2. Составить классификацию ошибок и тегов к ним.
3. Расшифровать и разметить записи уроков.
4. Провести статистический анализ ошибок по корпусу.
5. Выявить основные сложности студентов и сформировать рекомендации по развитию курса.

В работе мы последовательно рассмотрим примеры уже существующих учебных корпусов, опишем принципы разработки нашего корпуса, его материал, этапы работы и планируемые результаты.

В теоретическом плане наш проект в основном опирается на работы, посвященные, во-первых, созданию и обработке учебных корпусов, во-вторых, классификации и тегированию ошибок в речи говорящих на неродном языке.

Об учебных корпусах написано множество работ, в том числе и обзорных статей. Так, М. В. Хохлова подробно описывает существующие виды корпусов в зависимости от их задач, родного языка учащихся, уровня владения вторым языком, жанра текста и т. д. [Khokhlova 2023]. Как и другие исследователи учебных корпусов, М. В. Хохлова подчеркивает, что, поскольку устные корпусы требуют больше времени и усилий на запись и расшифровку, большинство корпусов сосредоточены на письменных данных. Действительно, Y. Soyeon утверждает, что по состоянию на февраль 2020 г. в CECL Католического университета Лувена перечислено 177 учебных корпусов, существующих в мире. Более половины из них (105 из 177, или 59,3 %) ориентированы на английский язык, но всего 35 англоязычных учебных корпусов (33,3 %) являются устными или содержат как устные, так и письменные данные [Soyeon 2020: 30–31]. Безусловно, это лишь примерные цифры, особенно на текущий момент, когда появляется все больше новых корпусов, однако данная статистика доказывает, что устных корпусов создается гораздо меньше, чем письменных.

Многие исследователи сходятся во мнении о важности, скорее, не объема корпуса, а его репрезентативности. В первую очередь нужно понимать, для каких задач собирается корпус, насколько глубоко будет проведена с ним работа (например, как именно будет аннотироваться корпус, и будет ли он аннотироваться вообще). Для одних целей не хватит и миллионов знаков, для других – будет достаточно и пяти тысяч [Копотев 2014].

Корпусы ошибок особенно часто используются в педагогических целях [Дмитриев и др. 2020; Павлова 2021], например при обучении русскоговорящих студентов работе с коллокациями в английском языке. Исследование, описанное В. И. Ивановой и Т. И. Кулагиной, показало, что «использование корпусов текстов при изучении иноязычной лексики на занятиях в целом способствовало более успешному усвоению студентами лексических единиц. Студенты стали совершать меньше ошибок, связанных с сочетаемостью слов» [Иванова, Кулагина 2022: 148]. Учителям и преподавателям иностранных языков важно знать о сравнительной характеристике изучаемого и родного для учащегося языков, о потенциальных ошибках, возникающих из-за взаимного влияния этих двух языков, а также о типичных ошибках, которые говорящие на определенном языке могут совершать при изучении другого иностранного языка. Если преподаватель работает с учебными корпусами, то он может постоянно корректировать свою педагогическую деятельность и более эффективно обучать иностранному языку [Грудева и др. 2018].

Крайне важный этап при создании корпуса ошибок – его аннотация. Аннотация обычно включает в себя три этапа: выявление ошибок, их классификацию и исправление. В процессе классификации ошибки группируются по определенным типам (например, лексические, морфологические, синтаксические). После выявления ошибки она подлежит исправлению, в результате чего в аннотации фиксируются оба варианта: исходный (с ошибкой) и исправленный. Таким образом, можно выделить два уровня аннотации ошибок: первый связан с разметкой ошибок по их категориям, а второй – с их исправлением [Khokhlova 2023: 63].

Методической литературы на тему самых распространенных ошибок в английском языке у носителей русского языка крайне мало. Много работ посвящено стратегиям коррекции ошибок [Тишулин 2012: 134–136; Lyster, Ranta 1997: 44–51], существует множество классификаций ошибок в речи на неродном языке с точки зрения причины их возникновения [Богданова 2014: 67–68], степени их грубости [Теренин 2016: 153], уровня языка [Кондрашова 2015: 28–37], однако теоретических работ с подробным описанием типов ошибок внутри их групп по уровням языковой системы мы не обнаружили. С методической точки зрения J. Edge предлагает разделять все ошибки на три группы: оговорки; ошибки, появляющиеся в пройденном материале; ошибки, возникающие в неизученном материале [Edge 1989: 9]. Такая общая классификация не подходит для цели настоящего исследования. В связи с этим мы разрабатываем собственную классификацию ошибок, речь о которой пойдет далее.

Проанализировав имеющуюся литературу по теме исследования, мы пришли к выводу, что существующие в педагогике типологии ошибок либо не опираются на практический материал, либо описывают наблюдения за речью учеников без структурного статистического анализа. В своей работе мы предлагаем классификацию ошибок, основанную на корпусном материале и отражающую реальные трудности в речи русскоязычных студентов на английском языке.

В области корпусных исследований ошибок большинство корпусов посвящены ошибкам в английском языке. Однако наш корпус отличается тем, что представленные в нем учащиеся – взрослые студенты с уверенным знанием английского языка, которые хотят улучшить навыки повседневной устной коммуникации в рамках рабочих задач. Такая специфика позволяет считать наш корпус особенно актуальным и востребованным для преподавателей профессионального английского языка для специалистов IT-сферы.

Новизна исследования определяется его материалом – онлайн-занятия русскоязычных студентов с англоговорящими специалистами в области информационных технологий. На момент написания этой статьи подобных корпусов в открытом доступе не было обнаружено, что подчеркивает уникальность работы. Принципиально новым аспектом исследования также выступает разработка классификации ошибок для корпуса, опирающейся на практический материал и включающей конкретные типы ошибок, распределенные по уровням языка. Более того, четкие задачи исследования от компании-заказчика позволяют назвать представленную работу актуальной не только с точки зрения исследований в областях лингвистики и педагогики, но и с точки зрения применения полученных знаний на практике. Исследование поможет модернизировать опыт онлайн-обучения для будущих студентов в данном сервисе онлайн-образования. Это новый взгляд на обучение, который не наказывает студентов за ошибки, а опережает их возникновение и заранее подготавливает преподавателей к будущим трудностям.

Методы и материалы

Учебные корпусы могут быть составлены на материале письменных или устных текстов, подготовленной или спонтанной речи, текстов разных жанров и языков и т. д. Наше исследование посвящено созданию корпуса ошибок русскоязычных студентов, обучающихся в рамках онлайн-курса по английскому языку для IT-специалистов. Курс является продуктом одного из крупных российских международных сервисов онлайн-образования[1]. Для исследования были взяты 50 видеозаписей занятий студентов с англоговорящими специалистами в области информационных технологий. Общая длительность записей составляет 42 часа. Это один из типов занятий, представленных на курсе: студент занимается не с преподавателем, а с англоговорящим IT-специалистом, коммуникация с которым происходит исключительно на английском языке. Кроме того, все задания выполняются в формате диалога. Таким образом, материалом корпуса является устная речь русскоязычных студентов на английском языке.

[1] Название сервиса и подробности курса находятся под NDA.

 https://doi.org/10.21603/2782-4799-2024-3-3-245-254

Результаты

На сегодняшний день не существует четких требований к объему корпуса, и мы, вслед за М. В. Хохловой, считаем, что ключевая характеристика любого корпуса – это его качество, а не количество материала [Khokhlova 2023: 59]. При выборе объема корпуса учитывалось несколько факторов: задача корпуса, количество разметчиков и выделенное на работу время. Наш корпус является узконаправленным: во-первых, его материал – это устная речь определенной группы студентов, а именно взрослых людей (примерно 20–45 лет), работающих в сфере информационных технологий и владеющих английским языком приблизительно на уровне B1–B2. Во-вторых, корпус состоит из речи студентов, проходящих курс *специализированного* английского языка. Следовательно, задача нашего корпуса заключается в выявлении проблемных аспектов в изучении английского языка в IT-сфере в рамках указанной группы студентов и конкретного курса от российского международного сервиса онлайн-образования. Таким образом, имея достаточно узкую выборку, корпус не требует большого объема материала. Количество исследователей и время на обработку корпуса, к сожалению, также ограничены: сейчас над корпусом работают три человека, и уделить на разметку планируется 5–6 месяцев. В этих условиях обработка именно 50 уроков кажется разумной целью.

Необходимо к тому же описать характер анализируемых занятий. Основная цель всего курса длиной в 7 месяцев – подготовить студентов к работе в международной компании на позиции продакт-менеджера, т. е. специалиста, отвечающего за разработку и запуск продукта или услуги. Для корпуса мы выбрали один из типов представленных на курсе занятий – беседу с англоговорящим IT-специалистом. Это особые занятия, которые проводятся в конце каждого месяца обучения для закрепления пройденного материала и отработки рабочих ситуаций (собеседование, обсуждение проекта в команде и т. п.). Важно вновь подчеркнуть, что роль преподавателя в таких случаях выполняет именно специалист в сфере информационных технологий, который не является педагогом по образованию, поэтому занятия очень приближены к ситуации общения в международной команде. Для удобства далее мы будем называть их преподавателями, а сами занятия – симуляциями, т. к. они симулируют общение студента с будущим коллегой из международной компании. Все общение во время симуляций осуществляется в устной форме: выполнение заданий, перед которыми студент имеет несколько минут на подготовку, и свободное общение с IT-специалистом (рассказ о себе, ответы на вопросы вне заданий).

Итак, материалом корпуса является как подготовленная, так и спонтанная устная речь на английском языке. Стоит отметить, что мы брали только первые или вторые симуляции курса, поскольку одна из целей исследования – определить языковые проблемы студентов, недавно пришедших на курс. Все отобранные занятия проводились в 2023–2024 гг. Каждое онлайн-занятие длится один час. Основные этапы исследования:

1. Создание классификации ошибок носителей русского языка в речи на английском языке.
2. Присваивание тегов всем типам ошибок.
3. Расшифровка 50 видеозаписей занятий и разметка ошибок.
4. Анализ размеченного корпуса и выявление самых частых типов ошибок.

На этапе составления собственной классификации самых распространенных ошибок, совершаемых русскоязычными студентами при изучении английского языка, мы основывались на исследованиях уроков английского языка в русских школах, а также на собственном преподавательском опыте. Так, Е. А. Яновская и А. В. Нескрёба считают, что чаще всего встречаются ошибки на уровне грамматики, особенно пропуск предлогов и артиклей [Яновская, Нескрёба 2020].

На этапе создания тегов мы опирались на уже существующие корпусы с разметкой ошибок, такие как *Russian Learner Corpus* (RLC) [Рахилина 2016], и работы по автоматизации разметки учебных корпусов [Bryant et al. 2017]. Например, из RLC мы взяли теги *Morph*, *Lex*, *WO*, *Tense* и др. (табл.). Помимо этого, некоторые теги были расширены, например, тег *Pronoun* в группе лексических ошибок был добавлен уже в ходе разметки, т. к. мы заметили часто встречающиеся ошибки в употреблении местоимений *other – another* и т. п.

При непосредственной разметке расшифровок онлайн-занятий мы столкнулись с некоторыми трудностями: например, с определением дочерних тегов для ошибок на лексическом уровне. Основной вопрос заключался в том, что считать ошибкой в коллокации, а что – неверным подбором слова для конкретного контекста. Коллокациям посвящено множество работ, однако термин все еще остается размытым

[Палийчук 2022; Черноусова 2019]. В рамках этого исследования мы будем считать ошибкой в коллокации и отмечать тегом *Colloc* те случаи, когда употребленное студентом словосочетание не встречается в речи носителей языка. Так, если студент говорит *do a mobile app* вместо *make a mobile app*. Но в тех случаях, когда само словосочетание возможно в языке, но было неверно употреблено в конкретном контексте, мы ставили общий тег *Lex*. Например:

So I'm already on module two, but I finished the first one. So I ***get {have}[Lex]*** *enough knowledge*.

В данном случае студент, отвечая на вопрос IT-специалиста о его обучении, имел в виду, что он уже прошел первый модуль программы курса, поэтому у него достаточно знаний для занятия. Хотя словосочетание *get knowledge* существует в английском языке, в указанном контексте верным будет вариант *have knowledge*. Проблема возникла не из-за сочетаемости двух слов, а из-за контекста, поэтому мы поставили общий тег *Lex*, а именно без добавления уточняющего тега *Colloc*.

В таблице помещена последняя версия дерева тегов на текущий момент. Оно состоит из родительских тегов, соответствующих уровням языка (*Morph,*

Табл. Теги ошибок
Tab. Error tags

Уровень языка	Тег	Подтег	Тип ошибки
Морфология тег: *Morph*	Plur	–	Неправильная форма числа или выбрано не то число, ex.g.: **advices; this – these / that – those; constraint – constraints*
	WordForm	Verb Adv Noun Pronoun Adj ingForm Num	Неправильное словообразование, ex.g.: *creative – creativity; to speak – speaking; *winned – won*
	SVA	–	Нарушение согласования подлежащего и сказуемого, ex.g.: **he have – he has*
Синтаксис тег: *Synt*	Art	WArt ZeroArt	Неверный артикль или пропуск артикля
	Prep	WPrep ZeroPrep	Неверный предлог или пропуск предлога
	WO	–	Неправильный порядок слов
	WordZero	–	Пропуск слова, ex.g.: пропуск *it*: *It's pretty depressing when always snow*
	Tense	PresS PastS FutureS PresCont PastCont PresPerf PastPerf PresPerfCont PastPerfCont	Неправильный выбор времени глагола
	AgrTense	–	Нарушение согласования времен
	Modal	–	Ошибка в модальных глаголах, ex.g.: пропуск *to* и т. п.
	Conj	–	Ошибка в использовании союза
	Constr	–	Ошибка в конструкции, ex.g.: **if I will*; порядок слов в придаточном
	Link	–	Ошибка в использовании глагола-связки, ex.g.: пропуск, вставка лишнего и т. д.
Лексика тег: Lex	Colloc	–	Ошибка в сочетаемости слов, ex.g.: *do – make, much – many*
	Calq	–	Калька с русского (идиомы / специфичные для русского термины), ex.g.: *you are true {you are right}*
	Pronoun	–	Неправильный выбор местоимения, ex.g.: *this – that, other – another, it – he*
Фонетика тег: *Pron*	–	–	Ошибка в произношении

 https://doi.org/10.21603/2782-4799-2024-3-3-245-254

Synt, Lex, Pron), и дочерних, уточняющих тип ошибки. Таким образом, большинство ошибок имеют минимум 2 тега. Приведем примеры разметки ошибок каждого уровня. По нашим наблюдениям, на данный момент одной из часто встречающихся ошибок на морфологическом уровне является ошибка в образовании формы слова, например:

> *Maybe I need to make the first one shorter,* ***much more shorter {much shorter}[Morph][WordForm] [Adj]****.*

Здесь первый тег указывает на морфологический характер ошибки, второй уточняет, что это ошибка в образовании формы слова, а третий – что это форма прилагательного. Кроме того, для всех ошибок в фигурных скобках мы указываем исправленный вариант.

На синтаксическом уровне студенты достаточно часто ошибаются в выборе времени глагола. В таких случаях дочерним тегом устанавливается то время, которое является правильным:

> *It was in school actually, quite a good* ***preparation {training}[Lex]*** *because we* ***have{had}[Synt][Tense] [PastS]*** *a really nice teacher.*

В этом предложении описывается ситуация, произошедшая в прошлом, о чем свидетельствует начало предложения (*It was…*), поэтому у слова *have* поставлен тег уровня языка *Synt*, тег типа ошибки *Tense* и тег *PastS*, уточняющий конкретное время глагола. В этом примере можно также увидеть лексическую ошибку, помеченную тегом *Lex*. Подобные ошибки мы не помечаем тегом для коллокаций *Colloc*, поскольку само словосочетание *a good preparation* может существовать, т. е. ошибка состоит именно в подборе неверного слова для данного контекста. Приведем пример ошибки в произношении:

> *It's* ***also [Pron]*** *pretty shiny.*

Разметка на фонетическом уровне на представленном этапе исследования служит, скорее, предварительной: мы отмечаем только явные ошибки в произношении слов, такие как неправильная постановка ударения, замена одной фонемы на другую (например, в слове *process* фонему [s] часто заменяют на [ts]) и т. д. Особенности русского акцента (неправильное произношение фонемы [θ] и т. д.) мы не учитываем. Это связано с тем, что для студентов анализируемого курса фонетический аспект языка оказывается наименее важным, т. к. их задача сводится не к достижению уровня носителя языка, а к способности поддерживать коммуникацию в международной команде, где каждый обладает своим акцентом и особенностями произношения.

Третьим этапом выступает расшифровка аудиозаписей. Для этого была использована компьютерная модель *AI Whisper*[2], а именно его маленькая версия, т. к. более глубокая модель Large V2 исправляет некоторые ошибки студентов (например, вставляет пропущенные артикли, меняет форму глагола и т. п.), что мешает нашему исследованию. Безусловно, в автоматических расшифровках встречаются неточности, поэтому перед разметкой они проверяются вручную. Далее в речи студентов мы выделяем ошибки и расставляем теги (речь преподавателя не анализируется, но она будет присутствовать в корпусе для сохранения контекста).

Если правильным вариантом является отсутствие слова (например, когда вставлен лишний предлог), то в исправленный вариант записывается правый и левый контекст в размере одного слова с каждой стороны, для неразборчивых фрагментов ставится знак <inaud>. Наряду с тегами, представленными в таблице, мы ввели тег *miscom* для ситуаций нарушения коммуникации: он позволит в дальнейшем отследить ошибки, препятствующие взаимопониманию между собеседниками. Приведем фрагмент транскрипта, размеченный при помощи разработанной нами системы тегов:

> *And I actually* ***need {needed}[Synt][Tense][PastS]*** *to sell this idea to, actually, to prove that there's* ***only right way {the only right way}[Synt][Art][ZeroArt]*** *in our situation to* ***top {the top}[Synt][Art][ZeroArt]*** *manager of* ***products {product manager}[Lex][Colloc]*** *from* ***other {another}[Lex][Pronoun]*** *department.*

Заключение

Современная лингвистика стремительно развивается, и одним из ключевых инструментов, ставших настоящим прорывом в исследованиях, стали корпусы. В частности, учебные корпусы, также известные как корпусы ошибок, играют важную роль в понимании особенностей освоения иностранного

[2] AI Whisper. URL: https://github.com/openai/whisper (accessed 20 Feb 2024).

языка студентами. Исследования в указанной области позволяют выявить типичные ошибки, проанализировать проблемные аспекты в освоении языка и создать более эффективные методики обучения.

В результате нашей работы на этом этапе выполнены первые три задачи: изучена необходимая литература, разработана классификация ошибок и система тегов к ним, а также размечены 50 записей онлайн-занятий. В ближайшем будущем нам необходимо собрать статистику, выявить самые частотные категории ошибок и сформировать рекомендации по развитию курса.

В итоге выделено несколько перспектив исследования: 1) более тщательное изучение ошибок в произношении и расширение тегов на фонетическом уровне; 2) использование корпуса для создания основанной на технологии машинного обучения модели автоматического тегирования ошибок русскоговорящих студентов на английском языке.

В данной статье мы сфокусировались на принципах разработки нашего учебного корпуса, составленного на материале речи русскоязычных студентов онлайн-курса английского языка для IT-специалистов. Работа с таким корпусом представляет интерес не только для методистов курса, но и для лингвистического сообщества в целом, поскольку изучение ошибок русскоязычных студентов в онлайн-школе английского языка является актуальной задачей в областях лингвистики и педагогики. Это исследование позволит выявить как общие, так и уникальные для онлайн-контекста языковые трудности, открывая новые перспективы для улучшения процесса обучения второму языку. Кроме того, созданная в рамках работы классификация ошибок может быть использована для аннотации будущих учебных корпусов речи носителей русского языка на английском языке, а еще для автоматизации тегирования ошибок.

Конфликт интересов: Авторы заявили об отсутствии потенциальных конфликтов интересов в отношении исследования, авторства и / или публикации данной статьи.

Conflict of interests: The authors declared no potential conflict of interests regarding the research, authorship, and / or publication of this article.

Критерии авторства: Ю. С. Виноградова – менеджмент проекта, формулирование идеи, цели и задач, аннотирование и очистка данных, проведение анализа, написание черновика, редактирование, визуализация.
И. В. Ильченко – формулирование идеи, цели и задач, аннотирование и очистка данных, проведение анализа, написание черновика, редактирование, визуализация.
П. С. Ширяева – формулирование идеи, цели и задач, аннотирование и очистка данных, проведение анализа, написание черновика, редактирование, визуализация.
М. С. Горина – предоставление материалов, супервайзинг, редактирование.
Contribution: Iu. S. Vinogradova supervised the project, developed the research concept, goals, and objectives, drafted and proofread the manuscript, checked the data, provided visualization, and performed the analysis.
I. V. Ilchenko developed the research concept, goals, and objectives, wrote the review, checked the data, performed the analysis, drafted and proofread the manuscript, and provided visualization.
P. S. Shiriaeva developed the research concept, goals, and objectives, wrote the review, checked the data, performed the analysis, drafted and proofread the manuscript, and provided visualization.
M. S. Gorina provided materials, supervised the research, and edited the manuscript.

Благодарности: Авторы выражают благодарность своему научному руководителю Колмогоровой Анастасии Владимировне за ценные советы при планировании исследования и рекомендации по оформлению статьи.
Acknowledgement: The authors would like to express their gratitude to their research advisor, Anastasia V. Kolmogorova, for valuable advice on research planning and recommendations.

Литература / References

Богданова Т. Г. Роль исправления ошибок при обучении иностранному языку в неязыковом вузе. *Научный Вестник Южного института менеджмента.* 2014. № 4. С. 66–69. [Bogdanova T. G. The role of error correction in teaching a foreign languages at business scholls. *Scientific bulletin of Uzhny institute of management*, 2014, (4): 66–69. (In Russ.)] https://elibrary.ru/toecrx

Грудева Е. В., Бучилова И. А., Волкова Н. А. Корпусы ошибок: целевая аудитория, возможная архитектура корпуса. *Вестник Череповецкого государственного университета*. 2018. № 5. С. 63–72. [Grudeva E. V., Buchilova I. A., Volkova N. A. Corpora of Errors: Target audience, a possible architecture of the corpus. *Bulletin of Cherepovets State University*, 2018, (5): 63–72. (In Russ.)] https://doi.org/10.23859/1994-0637-2018-5-86-7

Дмитриев А. В., Коган М. С., Вдовина Е. К. Теоретико-прикладное значение корпусов в компьютерной лингводидактике. *Litera*. 2020. № 1. С. 200–216. [Dmitriev A. V., Kogan M. S., Vdovina E. K. Theoretical-applied significance of corpora in computer linguodidactics. *Litera*, 2020, (1): 200–216. (In Russ.)] https://doi.org/10.25136/2409-8698.2020.1.32219

Захарова Е. А. Применение результатов исследований корпусной лингвистики в обучении грамматике английского языка на продвинутом уровне. *Вестник Российского университета дружбы народов. Серия: Русский и иностранные языки и методика их преподавания*. 2016. № 2. С. 41–49. [Zakharova E. A. Corpus-based studies in English grammar teaching at the advanced level. *Bulletin of Peoples' friendship university of Russia. Series: Russian and foreign languages. Methods of its teaching*, 2016, (2): 41–49. (In Russ.)] https://elibrary.ru/vwnqzr

Иванова В. И., Кулагина Т. И. Использование лингвистических корпусов текстов для формирования иноязычной учебно-познавательной компетенции. *Вестник ПНИПУ. Проблемы языкознания и педагогики*. 2022. № 3. С. 142–152. [Ivanova V. I., Kulagina T. I. Formation of foreign-language educational and cognitive competence by means of linguistic corpora. *PNRPU Linguistics and Pedagogy Bulletin*, 2022, (3): 142–152. (In Russ.)] https://doi.org/10.15593/2224-9389/2022.3.12

Колмогорова А. В. Эмоциональная тональность как значимый субъективный параметр учебного текста при овладении русским языком как иностранным. *Филологический класс*. 2019. № 3. С. 95–101. [Kolmogorova A. V. Emotional tonality as a valuable subjective parameter of study text for Russian as foreign language learners. *Philological class*, 2019, (3): 95–101. (In Russ.)] https://doi.org/10.26170/FK19-03-13

Кондрашова Н. В. Прогнозирование и исправление студенческих ошибок при обучении иностранным языкам. *Научный диалог*. 2015. № 7. С. 27–47. [Kondrashova N. V. Prediction and correction of students' mistakes when teaching of foreign languages. *Scientific dialogu*e, 2015, (7): 27–47. (In Russ.)] https:/elibrary.ru/tzymln

Копотев М. В. Введение в корпусную лингвистику. Praha: Animedia, 2014. 195 с. [Kopotev M. V. *Introduction to corpus linguistics*. Praha: Animedia, 2014, 195. (In Russ.)]

Павлова О. Ю. Использование языковых корпусов в обучении иностранному языку. *Язык и культура*. 2021. № 54. С. 283–298. [Pavlova O. Yu. Linguistic corpora in foreign language teaching. *Language and Culture*, 2021, (54): 283–298. (In Russ.)] https://doi.org/10.17223/19996195/54/16

Палийчук Д. А. Проблема определения понятия *коллокация* в современной лингвистике. *Евразийский гуманитарный журнал*. 2022. № 1. С. 20–25. [Palitchuk D. A. The problem of defining of "collocation" in modern linguistics. *Eurasian Humanitarian Journal*, 2022, (1): 20–25. (In Russ.)] https://elibrary.ru/fnxnkd

Рахилина Е. В. О новых инструментах описания русской грамматики: корпус ошибок. *Русский язык за рубежом*. 2016. № 3. С. 20–25. [Rakhilina E. V. Russian learner corpus as a new tool of grammatical description of Russian. *Russian Language Abroa*d, 2016, (3): 20–25. (In Russ.)] https://elibrary.ru/wffcob

Теренин А. В. Место и роль ошибки в языковом развитии. *Филологические науки. Вопросы теории и практики*. 2016. № 5-3. С. 153–155. [Terenin A. V. The place and role of an error in the language development. *Philological sciences. Issues of theory and practice*, 2016, (5-3): 153–155. (In Russ.)] https://elibrary.ru/vsmgfh

Тишулин П. Б. Виды языковых ошибок и возможности их исправления при обучении иностранному языку. *Известия высших учебных заведений. Поволжский регион. Гуманитарные науки*. 2012. № 1. С. 132–137. [Tishulin P. B. Types of language errors and improving them when teaching a foreign language. *University proceedings. Volga region. Humanities*, 2012, (1): 132–137. (In Russ.)] https://elibrary.ru/oxoqnt

Черноусова А. О. К вопросу о коллокациях. *Вестник Московского государственного областного университета. Серия: Лингвистика*. 2019. № 1. С. 57–64. [Chernousova A. O. On the notion of collocations. *Bulletin of Moscow State Regional University. Series "Linguistics"*, 2019, (1): 57–64. (In Russ.)] https://doi.org/10.18384/2310-712X-2019-1-57-64

Яновская Е. А., Нескрёба А. В. Наиболее типичные ошибки при изучении иностранного языка и некоторые пути их преодоления. *Иностранные языки в контексте межкультурной коммуникации*: XII Всерос. науч.-практ. конф. с Междунар. участием. (Саратов, 25–26 февраля 2020 г.) Саратов: Сарат. ист-к, 2020.

C. 325–330. [Yanovskaya E. A., Neskreba A. V. Typical mistakes in learning a foreign language and some ways to overcome them. *Foreign languages in the context of intercultural communication*: Proc. XII All-Russian Sci.-Prac. Conf. with Intern. Participation, Saratov, 25–26 Feb 2020. Saratov: Sarat. ist-k, 2020, 325–330. (In Russ.)] https://elibrary.ru/vdstmb

Bryant C., Felice M., Briscoe T. Automatic annotation and evaluation of error types for grammatical error correction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, 30 Jul–4 Aug. Vancouver: Association for Computational Linguistics, 2017, 793–805. https://doi.org/10.18653/v1/P17-1074

Edge J. *Mistakes and Corrections*. NY: Longman, 1989, 80.

Khokhlova M. V. Learner corpora: Relevant information and an overview of the existing frameworks. *Terra Linguistica*, 2023, 14(1): 57–69. https://doi.org/10.18721/JHSS.14106

Lyster R., Ranta L. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 1997, 19(1): 37–66. https://doi.org/10.1017/S0272263197001034

Rakhilina E., Vyrenkova A., Mustakimova E., Ladygina A., Smirnov I. Building a learner corpus for Russian. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC*: Proc. Conf., Umeå, 16 Nov 2016. Linköping: LiU Electronic Press, 2016, 66–75.

Soyeon Y. The learner corpora of spoken English: What has been done and what should be done? *Language Research*, 2020, 56(1): 29–51. https://doi.org/10.30961/lr.2020.56.1.29