АПФ&ПЛ Тематический выпуск ПРИКЛАДНАЯ ЛИНГВИСТИКА: СОВРЕМЕННЫЕ РАКУРСЫ И ПЕРСПЕКТИВЫ CIP&PL Thematic issue APPLIED LINGUISTICS: MODERN RESEARCH AREAS AND PERSPECTIVES http://philjournal.ru 2022 No 2 118-130

Оригинальная статья

УДК 81-25+81'44+81'322
DOI: 10.29025/2079-6021-2022-2-118-130

Создание устного учебного корпуса русского как иностранного: первые результаты и перспективы

Е.А. Власова[1]*, Ю.В. Бец[2], Е.В. Каллистратидис[3]
[1]Национальный исследовательский университет «Высшая школа экономики»,
105066, Российская Федерация, Москва, ул. Старая Басманная, д. 21, стр. 1,
[2,3]Южный федеральный университет,
344006, Российская Федерация, Ростов-на-Дону, пер. Университетский, 93;
[1]ORCID ID: 0000-0001-6121-1934; [1]Researcher ID: R-9491-2016;
[2]ORCID ID: 0000-0002-9383-472X;
[3] ORCID ID: 0000-0003-2521-1026;
[2]AuthorID: 554193;
[3]AuthorID: 884502;
*e-mail: *evlasova@hse.ru*

***Резюме**:* Статья посвящена первым результатам проекта по разработке устного учебного корпуса, который представляет собой затранскрибированную и размеченную по аномалиям коллекцию записей спонтанной устной речи иностранных обучающихся, осваивающих русский язык как иностранный. В работе представлен обзор научной литературы, посвященной созданию устных корпусов нестандартных текстов, обсуждаются особенности отбора материала для стимулирования устной спонтанной речи иностранца, описан опыт транскрибирования, классификации и аннотации фонетических и коммуникативных нестандартных явлений, проводится качественный анализ ряда нарушений и аномалий, свойства которых не могут быть исследованы на эмпирическом материале письменной речи. Опытным путем установлено, что использование упрощенной транскрипции является достаточным для исследования фонетических аномалий и просодических свойств нестандартной русской речи. Анализируются следующие просодические явления, характеризующие качество устной речи иностранного студента: паузы, хезитации (заполненные голосом паузы), физиологические паузы, фонетические неточности и самоисправления. Количественный анализ четырех пробных образцов речи показал, что аудиозаписи почти не различаются между собой по количеству физиологических пауз, однако выявлены существенные различия по количеству фонетических неточностей, заполненных пауз и самоисправлений говорящего. На основе соотношения указанных явлений выделено несколько профилей, отражающих уровень коммуникативной компетенции говорящего: а) профиль с большим числом пауз, свидетельствующих о планировании сообщения, но малым числом исправлений и фонетических неточностей; в этом случае речь иностранца медленная, но грамматически и фонетически более точная и связная; б) профиль с большим числом пауз, фонетических неточностей и самоисправлений: говорящий испытывает трудности с планированием высказывания и произношением; в) профиль с небольшим числом пауз и хезитаций, но с большим числом фонетических неточностей: речь довольно быстрая, однако качество произношения невысокое.

Ключевые слова: *русский язык как иностранный, звуковой корпус нестандартной речи, устный корпус русского языка, учебный корпус русского языка, говорение, дата-информированный подход в лингвистических исследованиях, корпусные исследования, освоение второго иностранного языка.*

Благодарности: Работа выполнена в рамках соглашения о научном сотрудничестве № 6.13.1-02/250821-1 по проекту «Конвергенция языковых пластов русского языка в зеркале цифровых реше-

ний» между Южным федеральным университетом (ЮФУ) и Национальным исследовательским университетом «Высшая школа экономики» (НИУ ВШЭ) («Зеркальные лаборатории НИУ ВШЭ»).

Для цитирования: Власова Е.А., Бец Ю.В., Каллистратидис Е.В. Создание устного учебного корпуса русского как иностранного: первые результаты и перспективы. *Актуальные проблемы филологии и педагогической лингвистики.* 2022. №2. С. 118–130.

Original Paper
DOI: 10.29025/2079-6021-2022-2-118-130

Creating an Oral Educational Corpus of the Russian Language for Non-Native Speakers: the Initial Results and Prospects

E.A. Vlasova [1], Y.V. Bets[2], E.V. Kallistratidis[3]
[1]National Research University Higher School of Economics (HSE University),
21, Staraya Basmannaya Str., Moscow, Russian Federation, 105006;
[2,3]Southern Federal University,
93 Universitetskiy Lane, Rostov-on-Don, Russian Federation, 344006;
[1]ORCID ID: 0000-0001-6121-1934; [1]Researcher ID: R-9491-2016;
[2]ORCID ID: 0000-0002-9383-472X;
[3] ORCID ID: 0000-0003-2521-1026;
[2] Scopus Author ID: 554193;
[3] Scopus Author ID: 884502;
*e-mail: evlasova@hse.ru

Abstract: The article presents initial results of the project on design of the oral educational corpus containing a transcribed and annotated collection of spontaneous/unprepared speech recordings of students learning Russian as a foreign language. This article includes a literature review on design of oral corpora; a discussion on how to choose stimuli for production of spontaneous oral speech by non-native speakers; a description of the transcription experience, the classification and the summary of non-standard phonetic and communicative phenomena; the quality examination of a number of deviances which properties are impossible to investigate in the written speech.

The article explores the following prosodic features typical for the oral speech of a foreign student: pauses, hesitations (voiced pauses), physiological pauses, phonetic inaccuracies, and self-corrections. The quantitative examination of the four trial speech pieces revealed the fact that the recordings do not demonstrate any differences in terms of the number of physiological pauses. However, there are significant fluctuations in the number of phonetic inaccuracies, voiced pauses and self-corrections. Comparing the above mentioned observations we identified several profiles, which reflect communicative performance of the speaker: a) the profile with a significant number of pauses indicating planning of the statement, but with a few corrections and phonetic inaccuracies; in this case the foreigner's speech is slow but grammatically and phonetically more accurate and cohesive; b) the profile with a significant number of pauses, phonetic inaccuracies and self-corrections: the speaker has difficulties with statement planning and pronunciation; c) the profile with a few pauses and hesitations, but with a significant number of phonetic inaccuracies: the speech is quite fast, while the pronunciation is rather poor.

***Keywords*:** Russian as a foreign language, sound corpus of non-standard speech, oral corpus of the Russian language, educational corpus of the Russian language, speaking, data-informed approach in linguistics, corpus studies, second language acquisition

***Acknowledgement**:* The study was carried out within the framework of the agreement on scientific cooperation between the Southern Federal University (SFedU) and the National Research University Higher School of Economics (HSE), project number 6.13.1-02/250821-1.

***For citation**:* Vlasova E.A., Bets Y.V., Kallistratidis E.V. Creating an Oral Educational Corpus of the Russian Language for Non-Native Speakers: the Initial Results and Prospects. *Current Issues in Philology and Pedagogical Linguistics.* 2022, no 2, pp. 118–130. (In Russ.).

Введение

Русский язык как иностранный сегодня является не только лингводидактической специальностью, но и самоценным предметом научных разысканий лингвистов, которые при помощи корпусных [1; 2], экспериментальных [3; 4] и описательных [5; 6] методов пытаются найти ответы на вопросы о том, как устроены коммуникативная, речевая и языковая компетенции в условиях ограниченного использования языка, каким образом и в каком порядке происходит становление разных языковых явлений и интерференция из доминантного языка, а также какие механизмы компенсируют лексическую и грамматическую несбалансированность в речи иностранного учащегося. Большое место отводится исследованию ошибок – разнообразным аномалиям и неточностям, наблюдаемым в письменных текстах иностранцев [5]. Современная компьютерная лингвистика предлагает широкие возможности для документирования, открытого доступа и качественно-количественных лингвистических описаний нестандартных явлений. Одним из таких проектов является «Русский учебный корпус» (англ. Russian Learner Corpus)[1] — открытый ресурс, содержащий коллекцию русскоязычных текстов, созданных нестандартными носителями – иностранцами и другими несбалансированными билингвами, с возможностями поиска по лингвистическим параметрам, типам аномалий и способу изучения русского языка. В рамках указанной парадигмы исследователи преимущественно концентрируются на речевых аномалиях и письменных текстах, при этом оставляя без внимания другие важные аспекты становления коммуникативной компетенции – явления устной спонтанной речи на разных уровнях владения русским языком.

Устная речевая продукция студентов, изучающих русский язык как иностранный, представляет интерес по нескольким причинам. Во-первых, устная речь характеризуется особыми психолингвистическими свойствами: спонтанностью, непринужденностью, ситуативностью и сниженными ресурсами грамматического контроля [7], в отличие от письменной продукции, при создании которых у иностранцев больше времени на обдумывание и перепроверку [8; 9]. Во-вторых, по наблюдениям М.Д. Воейковой [10], просодические ресурсы, такие как паузы, хезитации, самоисправления, несут важную информацию о механизмах речепорождения и могут существенным образом дополнить представления становлении речевой компетенции. В-третьих, устная речь обладает акустическими характеристиками и количественными метриками, которые значительно расширяют возможности статистического анализа [10]. Наконец, в прикладной лингвистике до сих пор отсутствуют систематические научные описания того, как устроена устная речевая компетенция иностранцев на разных уровнях и как объективно проводить границу между коммуникативно значимыми и незначимыми нарушениями, грубыми и негрубыми ошибками при оценке части «Говорение» [11].

Данная статья посвящена опыту разработки устного учебного корпуса – коллекции записей устной русской речи иностранцев для последующего транскрибирования и разметки по метаданным, просодическим характеристикам, количественным акустическим метрикам. Актуальность исследования обусловлена общим развитием компьютерной лингвистики, которая предлагает широкий набор корпусных инструментов для создания, транскрибирования и обработки нестандартной устной речи. Новизна состоит в том, что к нестандартной речевой продукции иностранцев применяются корпусные методы анализа устной речи.

Цель исследования

Цель данной статьи – представить первичные результаты проекта, посвященного разработке устного корпуса по русскому языку как иностранному, а также оценить теоретический и прикладной потенциал затранскрибированной устной коллекции. В работе обсуждаются особенности отбора стимульного материала для получения разнообразных по дискурсивным свойствам записей спонтанной русской речи от иностранцев, описан опыт транскрибирования, классификации и аннотации фонетических аномалий, проведен качественный и количественный анализ ряда нарушений, свойства которых невозможно исследовать на основе письменных текстов.

Методы и материал исследования

В основу исследования положены методы корпусной лингвистики, включающие разработку процедуры сбора устных записей и проектирования метаданных, отражающих информацию о носителях (доминантный язык, уровень владения, опыт изучения русского языка, возраст и место сбора). Отдельное внимание уделялось требованию репрезентативности и сбалансированности устной коллекции. На базе

[1] Русский учебный корпус http://web-corpora.net/RLC

Южного федерального университета (Ростов-на-Дону) ведется сбор образцов устной речи, разнообразной по дискурсивным качествам и типам стимульного материала, провоцирующего иностранного говорящего на речепорождение: реплика-вопрос, план рассказа, текст для пересказа, сюжетная картинка, картинка-виммельбух, видео для пересказа. Для качественного лингвистического анализа разработана аннотация аномальных фонетических и коммуникативных явлений с возможностями поиска, выгрузки коллекции примеров с одинаковым типом нарушений и количественного анализа. Ценность и новизна устного материала состоит в том, что исследуемая речевая продукция получена в условиях спонтанной коммуникации, в отличие от письменных текстов или других методов, например, анкетирования или при экспериментах с изолированными контекстами.

Обзор литературы

Одной из самых влиятельных работ, заложивших основы сбора устной спонтанной речи для изучения становления коммуникативной компетенции, стал известный проект [12], в котором детям из пяти стран предлагалось описать последовательность картинок М. Мейера «Где ты, лягушка?» (англ. M. Meyer. Frog, Where Are You?). На картинках от мальчика сбегает лягушка, и он с собакой идет ее искать, расспрашивая о ней разных лесных обитателей и попадая в небольшие приключения. Таким образом исследовалось развитие пространственных маркеров и нарративных способностей детей из разных стран на основе затранскрибированных записей.

Потенциал сюжетной иллюстрации как стимульного материала для развития речи был реализован в формате книг-виммельбухов. Виммельбухи (нем. «мельтешащая книга») – это жанр, объединяющий книги, в которых доминирующее положение занимают детально разработанные автором иллюстрации, а текст почти не используется или играет вспомогательную роль. Картинки, как правило, изображают не выдуманный сказочный мир, а максимально естественные и привычные жизненные и бытовые ситуации городского жителя. Обилие разнообразных предметов и персонажей, вовлеченных в социальное взаимодействие, – часто с ярко выраженными эмоциональными реакциями, – довольно точно моделируют реальную жизнь на одном развороте и изначально разрабатывались детским психологом Куртом Зельманном (нем. Kurt Seelmann) и иллюстратором Али Митгучем (нем. Ali Mitgutsche) как материал для развивающих занятий. Среди первых виммельбухов, получивших награду, была книга Rundherum in meiner Stadt (Around in my city), опубликованная в 1968 году. Жизнь европейского города стала центральной темой многих виммельбухов, однако в более поздних версиях наблюдается расширение тематического ряда и разработка дополнительных материалов, рассчитанных не только на наблюдательность и речевые навыки, но и на решение головоломок и деятельностный подход. Постепенно появляются национальные традиции виммельбухов – европейская, вокруг европейского города и реалий (см. выше), англо-саксонская традиция серии М. Хандфорда «Где Уолли?» (“Where is Wally?” Martin Handford). В 2016 г. издательство “Златоуст” опубликовало российскую версию виммельбуха, адаптированного под национальные реалии и представляющего собой коллаборацию художника Е. Салатова и доктора педагогических наук Протасовой Е.Ю., специалиста по многоязычию и детскому билингвизму [13].

Виммельбухи широко используются как развивающие пособия, в том числе при работе с двуязычными детьми, однако недавно исследователи обратились к этому жанру для получения образцов спонтанной устной речи и оценки коммуникативной компетенции. Уже имеется опыт[2] применения виммельбухов на занятиях по русскому языку как иностранному, который показал, что даже при базовом лексическом запасе студенты-иностранцы способны порождать связные спонтанные устные истории на основе предложенных иллюстраций.

Trinity Lancaster Corpus (TLC) – совместный проект лондонского Тринити-колледжа (Trinity College London) и Центра корпусных исследований в социальных науках Ланкастерского университета (the Centre for Corpus Approaches to Social Science (CASS) at Lancaster University). Указанный корпус представляет собой коллекцию устных речевых произведений, порожденных носителями английского как второго (иностранного) языка. Разработчики проекта утверждают, что в настоящее время корпус включает в общей сложности 4,2 миллиона слов, извлеченных из записей транскрибированной речи 2000 информантов в возрасте от 9 до 72 лет, владеющих английским как вторым языком. География информантов покрывает 9 стран: Италия, Испания, Мексика, Аргентина, Бразилия, Китай, Индия, Шри-Ланка и

[2] Иваненко А.А., Выренкова А.С. Стратегии описания сложных сюжетных элементов русскоязычными монолингвами и билингвами с доминантным немецким. *Проблемы онтолингвистики - 2021: языковая система ребенка в ситуации одно- и многоязычия*. СПб.: ООО «ВВМ», 2021; 128-134.

Россия. Степень сформированности языковой компетенции продуцентов текстов корпуса соответствует трём уровням владения английским языком в следующем диапазоне: A 2.2 – B 1.2 – элементарный уровень, B 2.1. – B 2.3 – средний уровень, C 1.1. – C 2. – продвинутый уровень. Такой крупный объем данных получен в процессе проведения GESE (GESE- Graded Examinations in Spoken English) – экзаменационного тестирования по разговорному английскому языку как второму, которое регулярно проводится центрами тестирования Тринити-колледжа для государственной сертификации, необходимой для оформления и/или пролонгирования семейной и некоторых видов рабочей визы, а также вида на жительство. В качестве стимульного материала, побуждающего информантов к порождению устных речевых произведений, используются экзаменационные задания, которые представляют собой сбалансированное сочетание сценариев, в которых четко обозначены роли говорящих и их коммуникативные цели, однако сам способ взаимодействия между носителем языка и информантом носит динамический характер [14: 142–146]. Экзаменуемые могут заранее планировать, что именно они хотели бы обсудить во время экзамена, и решить, какую точку зрения и с помощью каких вербальных средств им хотелось бы озвучить во время диалога с экзаменатором. На среднем и продвинутом уровне испытуемые могут взять с собой на экзамен свои записи, чтобы восполнить нехватку лексики. Разработчики проекта подчеркивают тот факт, что в процессе диалога экзаменатор имеет возможность с помощью естественных аутентичных реплик и междометий направлять диалог в нужное русло, побуждать информанта к продолжению рассказа, уточнению деталей, озвучиванию дополнительной информации, тем самым помогая ему создать полноценное речевое произведение и наилучшим образом продемонстрировать уровень своей языковой компетенции.

RUEG Corpus – мультилингвальный корпус Исследовательской группы «Эмерджентные грамматики в ситуациях языкового контакта» (Research Unit «Emerging Grammars in Language Contact Situations», сокращенно – RUEG) Берлинского университета Гумбольдта (Humboldt-Universität zu Berlin). Объектом исследования RUEG являются языковые системы и языковые ресурсы «эритажных говорящих» (билингв из семей иммигрантов), реализация этих систем и ресурсов в обоих языках, а также их проявления в разных сочетаниях языков, в различных регистрах речи лиц, относящихся к разным возрастным группам. Этим обусловлена организация корпуса, а также принцип сбора входящих в него речевых материалов. RUEG corpus, развивающийся с 2018 г., в настоящее время состоит из пяти коллекций, каждая из которых соотносится с одним из основных (эритажных) языков информантов-билингв: DE – немецкий субкорпус (260 информантов), EN – английский (287), EL – греческий (167), RU – русский (193), TR – турецкий (188). Были использованы устные и письменные речевые произведения в общей сложности 720 говорящих, из них 349 взрослых (от 20 до 37 лет) и 371 подросток (от 13 до 19 лет) (RUEG Corpus, 2022). Для сбора языкового материала был использован метод языковых ситуаций (LangSit method), суть которого состоит в том, что функцию стимульного материала выполняет статический или динамический визуальный текст: фотография или видеоролик сюжетного характера. Информант знакомится с историей и представляет себя свидетелем произошедшего, а затем действует в заданных коммуникативных ситуациях, предполагающих взаимодействие с разными партнерами.

Значительный опыт создания устных корпусов накоплен и российскими исследователями. При исследовании русско-тюркского двуязычия в Южной Сибири создана коллекция, содержащая интервьюирование, беседы с информантами, самозаписи информантами разных форм обыденной коммуникации [15]. В Санкт-Петербурге активно ведется разработка «Корпуса аннотированной текстотеки» [16], который изначально содержал речь петербуржцев, но постепенно стал развиваться по пути расширения коллекции и пополнения ее записями устной русской речи иностранцев. При создании указанного корпуса использованы стимульные материалы, отражающие несколько коммуникативных сценариев: пересказ, описание картинки, рассказ на заданную тему.

Устные корпуса занимают значительное место в исследованиях детской речи. В исследовании Н. Гагариной [17] разработан метод MAIN, при котором детям предлагается пересказать два рассказа и описать две серии картинок для получения устных высказываний разной дискурсивной организации. В недавней работе[3] детям-носителям русско-немецкого двуязычия были предложены два динамиче-

[3] Иваненко А.А. Возможности применения виммельбуха в методике преподавания русского языка как иностранного. *Филологический аспект*. 2018; 37 (5): 38-43. Доступно по: https://scipress.ru/philology/articles/vozmozhnosti-primeneniya-vimmelbukha-v-metodike-prepodavaniya-russkogo-yazyka-kak-inostrannogo.html. Ссылка активна на 16.05.2022.

ских видеофрагмента из мультфильма «Маша и медведь» длительностью 30 секунд каждый, а также дополненные шестью картинками со стоп-кадрами из просмотренного фрагмента. Исследование показало, что наибольшую трудность у информантов вызывают микрофабулы с предикатами изменения пространственной ориентации, например, *полезть*, *залезть*, *пойти*, *вскарабкаться* и пр.

Приведенный обзор показывает, что при создании устных корпусов, независимо от цели исследования, используется несколько разных видов стимульного материала, который тем не менее важно подбирать с учетом достаточного лексического запаса информанта. Без соблюдения данного условия существуют риски отказа информанта от участия в эксперименте или отсутствия основы для спонтанного речепорождения. На основе приведенного обзора сформирована коллекция стимульного материала, состоящего из наиболее распространенных форматов: вопрос, тематический план, сюжетная картинка, видеофрагмент для пересказа, текст для пересказа, сценарий коммуникативной ситуации.

Результаты и дискуссия

К настоящему моменту в процессе проектирования устного учебного корпус на основе уже существующих разработок отечественных и зарубежных специалистов в области корпусной лингвистики были получены следующие результаты:

- разработана коллекция стимульного материала, побуждающего нестандартных носителей к спонтанному речепорождению и созданию устных (звучащих) речевых произведений, которые могут быть включены в корпусную коллекцию;
- спроектированы метаданные, необходимые для фиксации как социолингвистически значимой информации, так и параметров, имеющих принципиальное значение с точки зрения лингводидактики;
- предложена классификация коммуникативных аномалий;
- сформулированы принципы транскрибирования устных текстов и проведены первичные качественные и количественные обследования.

В процессе обработки первичных аудиозаписей и их анализа с применением корпусных методов были получены следующие результаты:

- составлена классификация произносительных аномалий, регулярно возникающих в результате сбоев в устной спонтанной речи, которые являются универсальными.
- теги, используемые для маркирования аномалий в уже существующих корпусах нестандартной речи, были адаптированы для разметки устных речевых произведений носителей русского как второго (иностранного) языка.
- установлено, что количество физиологических пауз не зависит от сложности стимульного материала и индивидуальных возможностей информанта, что позволило выявить следующие подвергаемые варьированию параметры: пауза, хезитации, исправления, фонетика.
- эмпирическим путем было доказано, что предполагаемая нами система разметки устных текстов нестандартных носителей делает возможным вычисление «профиля» речевой компетенции говорящего с помощью методов корпусной лингвистики, что позволяет автоматизировать, ускорить и сделать более объективным процесс оценки степени сформированности речевой компетенции инофонов.

1.1. Проектирование метаданных

Для исследования становления устной коммуникативной компетенции имеют значение следующие метаданные об информанте:

- шифр информанта, который используется как ключ при обработке таблиц;
- пол;
- возраст;
- основная страна проживания;
- доминантный язык – родной язык носителя для учета интерференции и социокультурных факторов;
- уровень владения письменной речью на русском языке по шкале CERF (A2-C1) на основе саморефлекции информанта;
- уровень владения устной речью на русском языке по шкале CERF (A2-C1) на основе саморефлексии информанта;
- уровень группы, в которой проходит обучение информант: начинающий, продолжающий, совершенствующийся, нет информации.

Для исследования взаимосвязей между коммуникативной компетенцией и условиями обучения имеют значение следующие параметры, актуальные в момент сбора материала:

- местонахождения информанта: Россия / зарубежное государство;
- тип обучения в момент сбора материала: очное, дистанционное, очно-заочное;
- общая продолжительность изучения русского языка (в месяцах);
- общее количество месяцев, проведенных в России;
- имеется ли доступ к русскоязычному окружению и общению на русском языке с носителями: да, нет;
- дата предоставления личных данных;
- место заполнения анкеты.

Сбор указанных данных проводится вместе с получением информированного согласия на участие в эксперименте.

Для анализа дискурсивных свойств устной речевой продукции для каждой полученной аудиозаписи указываются следующие параметры:

- тип стимульного материала: вопрос, тематический план, сюжетная картинка, видеофрагмент для пересказа, текст для пересказа, сценарий коммуникативной ситуации;
- тема: например, «Где ты, лягушка?»;
- тип коммуникативной ситуации: преимущественно интерактивная, то есть предполагающая обмен репликами, преимущественно монологическая, другое;
- дата сбора материала;
- место сбора материала;
- шифр информанта;
- инициатор сбора материала – фамилия преподавателя.

1.2. Аннотация произносительных аномалий

На основании анализа разметки неканонических явлений в других корпусах нестандартной речи [18] выделены универсальные сбои в устной спонтанной речи и составлена классификация произносительных аномалий для разметки явлений: пояснения и примеры из наших записей представлены в Таблице 1.

Таблица 1

Классификация произносительных аномалий

тег	пояснение	пример
/	короткая пауза	да / это время в России называются [зоротая] осень // = *да, это время в России называют золотая осень*
//	конец дискурсивной единицы	да / это время в России называются [зоротая] осень // = *да, это время в России называют золотая осень*
HES	заполненная голосом пауза (хезитация)	летом **HES** есть много / овощей и фруктов / [цветав] // = *летом есть много овощей и фруктов, цветов*
PHYS	физиология речи -физиологическая пауза (дыхание, сглатывание и др.)	У Игоря есть **PHYS** любимое время года / лето = *У Игоря есть любимое время года лето*
[]	фонетика – выделяется слово, содержащее фонетическую аномалию, то есть нарушение в произношении	там **[**хородно**]** а / дома **[**тепро**]** и **[**уёно**]** = *там холодно, а дома тепло и уютно*
{… SLIP...}	исправления – повтор слова, его части или словосочетания с целью самокоррекции	и [здеси] есть и зелёные **{**листь **SLIP** листья **}** // = *и здесь есть и зелёные листья*

1.3. Транскрибирование

В устных корпусах сложилась традиция разбивать записи на элементарные дискурсивные единицы и использовать в качестве основной формы обработки данных упрощенную письменную транскрипцию устных аудиозаписей [19]. В отличие от расширенной фонетической транскрипции, используемой при исследовании литературного языка и его региональных и диалектных вариантов, цель упрощенной транскрипции – зафиксировать наиболее заметные фонематические отклонения, не свойственные носителям русского языка, например, путаницу *р* и *л* или *б* и *в* в речи корейских студентов, вставки гласных звуков внутри консонантных кластеров (*доволийные* вместо *довольные*), пропуски и искажения в произношении слов, путаницы гласных в ударных позициях. Нефонематические аномалии произношения,

такие как разная степень редукции гласных, полумягкость согласных и др. на письме не отображаются, но могут быть при необходимости исследованы отдельно путем выгрузки файлов.

1.4. Результаты обследования первичной тестовой коллекции

Для изучения возможностей обработки материала в Южном федеральном университете (г. Ростов-на-Дону) под руководством Ю.В. Бец и Е.В. Каллистратидис собраны аудиозаписи студентов, изучающих русский язык как иностранный. В качестве стимульного материала четырем китайским студентам уровня В 1 предъявлялся набор иллюстраций «Где ты, лягушка?», а также тезисный план для рассказа об осени.

Абсолютное количество разных типов аномалий представлено в Таблице 2.

Таблица 2

Количество произносительных аномалий

Аудио	паузы	хезитации	физиология	исправления	фонетика	всего единиц
осень1	9	18	3	10	37	24
осень2	20	16	3	10	20	17
лягушка1	36	22	6	10	2	36
лягушка2	8	17	5	8	21	24

Указанные данные, однако, не дают возможности для объективного сопоставления полученных записей по числу наблюдаемых аномалий, поскольку очевидно, что чем больше длина текста, тем больше нарушений в нем ожидается. Для первичного сравнительного анализа полученные величины были нормализованы: каждый из абсолютных показателей разделен на количество элементарных дискурсивных единиц. Таким образом вычисляется коэффициент частотности каждого из аномальных явлений в расчете на элементарную дискурсивную единицу. Нормализованные данные представлены в Таблице 3.

Таблица 3

Количество произносительных аномалий: нормализованные данные

Аудио	паузы	хезитации	физиология	исправления	фонетика
осень1	0.38	0.75	0.13	0.42	1.54
осень2	1.18	0.94	0.18	0.59	1.18
лягушка1	1.00	0.61	0.17	0.28	0.06
лягушка2	0.33	0.71	0.21	0.33	0.88

Ниже на Рисунке 1 представлена диаграмма, составленная на основе нормализованных данных и позволяющая оценить распределение разных видов произносительных аномалий в исследуемых устных высказываниях китайских студентов продолжающего уровня.

Во-первых, заметно, что, независимо от жанра, сложности стимульного материала и индивидуальных особенностей носителя, количество физиологических пауз для перевода дыхания, сглатываний, кашля и пр. оказывается во всех текстах довольно ровным и в целом представляет собой невысокий показатель.

По коэффициенту хезитаций и исправлений наблюдается варьирование, однако разброс значений невелик. Интерес представляют категории «паузы» и «фонетика», так как по этим показателям аудиозаписи значительно различаются. В «тексте 1» и «лягушка 2» количество пауз значительно меньше, чем в других записях. Большой разброс показателей наблюдается в категории «фонетика», которая описывает степень аккуратности произношения: выделяется аудио «осень1» с большим количеством фонематических нарушений и рассказ о лягушке «лягушка1» с крайне низким числом произносительных неточностей. О последней аудиозаписи известно, что информант долгое время проживал в России и активно вовлечен в общение с русскоязычным окружением.

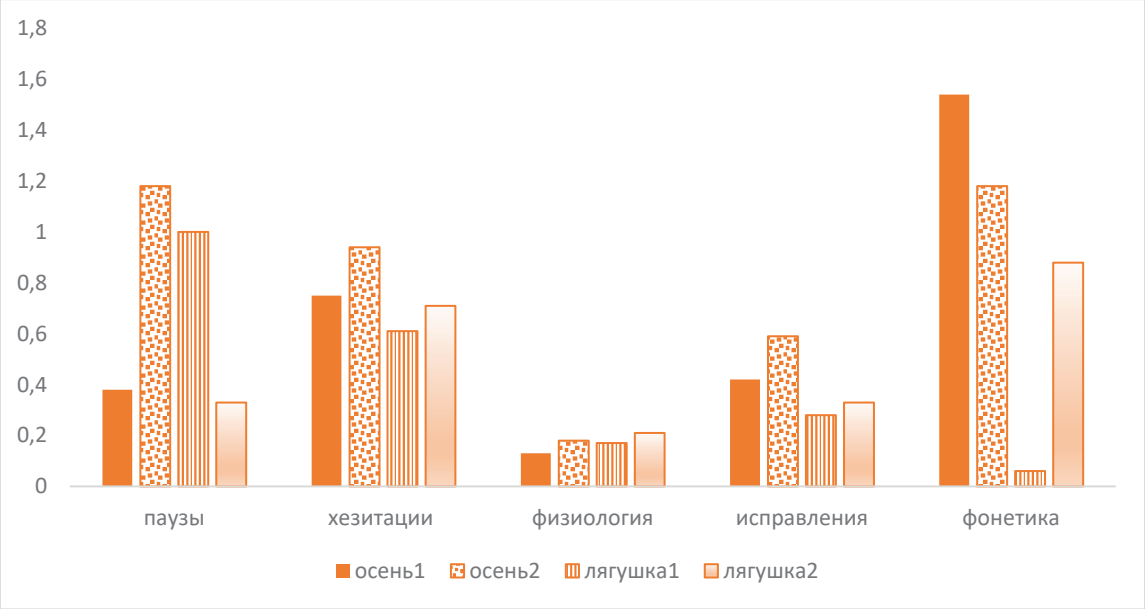

Рисунок 1. Распределение произносительных аномалий

Следующая диаграмма на Рисунке 2 позволяет оценить соотношение разных типов аномалий внутри каждого аудио и определить сходства и различия исследуемых образцов речи. Из диаграммы исключена категория «физиология», так как по этому показателю различий между образцами не выявлено.

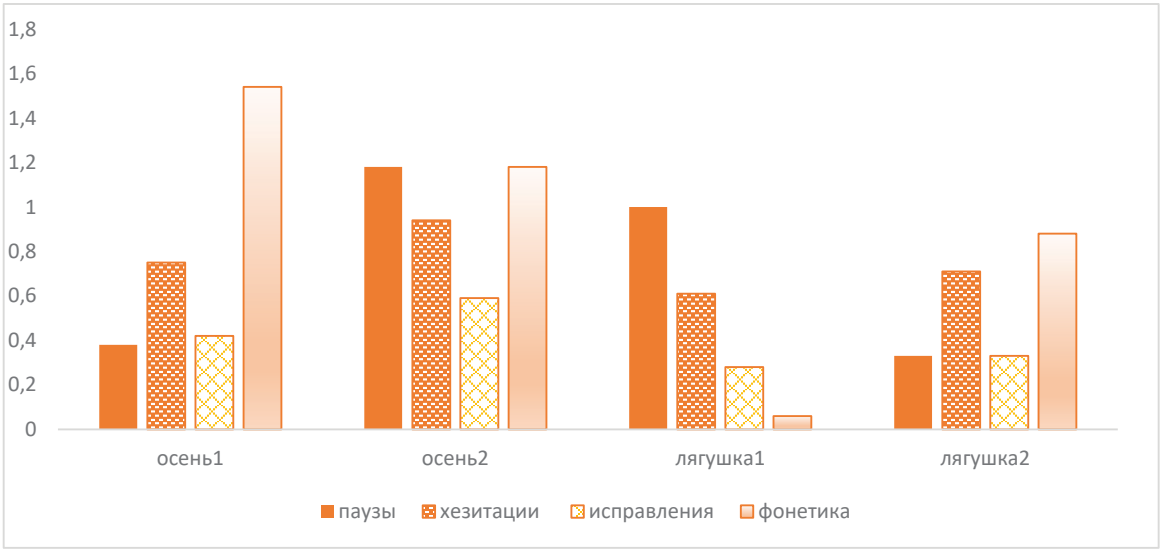

Рисунок 2. Соотношение произносительных аномалий в текстах

В первую очередь обратим внимание на сходство «контура» диаграмм для записей «осень1» и «лягушка2»: оба образца характеризуются большим числом фонетических неточностей и заполненных пауз, при этом количество исправлений и обычных межсловных пауз оказывается на одинаковом уровне. На фоне других записей выделяется «лягушка 1». Для этого образца характерны невысокие показатели фонетических аномалий и исправлений – это две категории, указывающие на высокое качество произношения и грамматическую точность, при этом сохраняются высокие коэффициенты межсловных и заполненных пауз, которые обычно связываются с планированием сообщения. Эти наблюдения указывают на то, что информант использует компенсаторные механизмы замедления речи, такие как паузы и заполненные паузы, при этом сохраняет лингвистическую точность, демонстрируя аккурат-

ное произношение и меньшее число исправлений. Наконец, запись «осень2» в целом характеризуется большим числом пауз, в том числе заполненных, самым высоким коэффициентом самоисправлений и высоким показателем фонетических аномалий: в совокупности указанные признаки свидетельствуют о сложностях в устном спонтанном речепорождении. Говорящему чаще требуется планировать сообщение, при этом он демонстрирует невысокое фонетическое качество речи.

Элементарный количественный анализ показал, что предложенная разметка затранскрибированной аудиозаписи позволяет вычислять «профиль» речевой компетенции и при наличии крупной коллекции дает возможность кластеризовать информантов с учетом количественных корпусных обследований и качественных характеристик, собранных в метаданных.

Заключение

Нами был представлен опыт сбора и обработки записей устной речи иностранных студентов для проектирования устного учебного корпуса по русскому языку как иностранному. Было установлено, что при сборе аудиозаписей для русского устного учебного корпуса особое значение имеет качество стимульного материала: разнообразие его дискурсивных свойств, способность задавать структуру высказыванию и тематический круг, соответствующий уровню владения русским языком информанта для обеспечения спонтанных условий коммуникации. Использование упрощенной транскрипции достаточно для исследования наиболее ярких фонетических аномалий и просодических свойств нестандартной русской речи иностранцев и значительно упрощает процесс обработки материала. Опытным путем доказано, что выделение в затранскрибированных текстах базовых просодических параметров (пауз, хезитаций, самоисправлений и фонетических аномалий), является достаточно информативным для оценки качества устной спонтанной речи нестандартного носителя. По ним возможно осуществлять поиск и создавать крупные подборки примеров для более глубокого лингвистического анализа и каталогизации наиболее типичных нарушений устной речи по уровням владения.

Базовые методы количественного анализа показали, что используемая разметка позволяет достаточно легко извлекать несколько количественных метрик и при большом числе аудиозаписей использовать не только описательные, но и объяснительные и предсказательные статистические методы.

Список литературы

1. Rakhilina EV, Vyrenkova AS, Mustakimova EG, Ladygina AA, Smirnov IY. Building a learner corpus for Russian. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. 2016: 66–75. Available at: https://aclanthology.org/W16-6509.pdf. Accessed April 18, 2022.

2. Glaznieks A, Frey J-C, Stopfner M, Zanasi L, Nicolas L. Leonide: a longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*. 2022; 8 (1): 97 – 120. https://doi.org/10.1075/ijlcr.21004.gla.

3. Slioussar NA, Cherepovskaia NV, Denissenko A. Acquisition of the nominal case system in Russian as a second language. *Second Language Research*. 2020. https://doi.org/10.1177/0267658320988058.

4. Aldaihni MS, Al-Houti KS. Perception of English Fricatives (/v/, /ʒ/) by Undergraduate Kuwaiti Arabic Learners. *International Journal of Linguistics and Communication.* 2021; 9 (2): 1–17. https://doi.org/10.15640/ijlc.v9n2a1.

5. Никунласси А., Протасова Е.Ю. О важности ошибок для исследования многоязычия. Под общей редакцией А. Никунласси и Протасовой Е.Ю. *Slavica Helsingiensia 45. Инструментарий русистики: ошибки и многоязычие.* Хельсинки: Unigrafia; 2014; 5–13. Доступно по https://blogs.helsinki.fi/slavica-helsingiensia/slavica-helsingiensia-45/. Ссылка активна на 15.05.2022.

6. Izutsu MN, Izutsu K. American and Irish English speakers' perceptions of the final particles *so* and *but. World Enleshes.* 2020; 41 (2): 207 – 233. https://doi.org/10.1111/weng.12521

7. Богданова-Бегларян Н.В. Звуковой корпус как материал для анализа русской речи. *Коллективная монография. Ч. 1: Чтение. Пересказ. Описание.* СПб.: Филологический ф-т СПбГУ; 2013. Доступно по https://ruslang.ru/doc/trudy/vol21/6-bogdanova-beglaryan.pdf. Ссылка активна на 15.05.2022.

8. Земская Е.А. Русская разговорная речь: лингвистический анализ и проблемы обучения. 2-е издание, исправленное и дополненное. Москва: Русский язык; 2018. Доступно по: https://biblioclub.ru/index.php?page=book&id=83088. Ссылка активна на 15.05.2022.

9. Plonsky L. Advancing quantitative methods in second language research. New York: Routledge; 2015. https://doi.org/10.4324/9781315870908.

10. Воейкова М.Д. Усвоение первого и второго иностранного языка: сходства и различия. *Путь в язык: Одноязычие и многоязычие*. Под редакцией Цейтлин С.Н., Елисеевой М.Б. М.: Языки славянских культур, 2011: 49-75. Доступно по: https://cdn1.ozone.ru/s3/multimedia-m/6012627634.pdf. Ссылка активна на 16.05.2022.

11. Фомина Н.С. Коммуникативно значимые и незначимые ошибки в письменной речи тестируемых по русскому языку как иностранному. *Концепт*. 2015; 12: 136-140. Доступно по: https://cyberleninka.ru/article/n/kommunikativno-znachimye-i-neznachimye-oshibki-v-pismennoy-rechi-testiruemyh-po-russkomu-yazyku-kak-inostrannomu. Ссылка активна на 16.05.2022.

12. Berman RA, Slobin DI. Relating events in narrative: A cross-linguistic developmental study. Hillsdale, NJ: L. Erlbaum; 1994. Доступно по: https://www.routledge.com/Relating-Events-in-Narrative-A-Crosslinguistic-Developmental-Study/Berman-Slobin/p/book/9781138984912. Ссылка активна на 16.05.2022.

13. Протасова Е.Ю., Салатов И. Карты, карточки, картинки. *Вот как-то так: жизнь в картинках*. СПб: Златоуст, 2016.

14. Gablasova D**,** Brezina V, McEnery A. The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research*. 2019; 5 (2): 126-158. https://doi.org/10.1075/ijlcr.19001.gab.

15. Резанова З.И. Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки. *Вопросы лексикографии*. 2017; 11: 105-118. https://doi.org/10.17223/22274200/11/7.

16. Завадская Ю.О. Случайность или закономерность? Частотность словоформы и другие возможные причины грамматических речевых сбоев в русской спонтанной речи. *Acta Linguistica Petropolitana.* 2021; 17(1): 123-142. https://doi.org/10.30842/alp23065737171123142.

17. Bohnacker U, Gagarina N. Developing Narrative Comprehension. Multilingual Assessment Instrument for Narratives. John Benjamins; 2020. https://doi.org/10.1075/sibil.61.

18. Подлесская В.И., Кибрик А.А. Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи. *Научно-техническая информация*. 2007; 2: 2-23. Доступно по: https://iling-ran.ru/kibrik/Self-repair@NTI_2007.pdf. Ссылка активна на 16.05.2022.

19. Кибрик А.А., Подлесская В.И. Проблема сегментации устного дискурса говорящего и когнитивная система. *Когнитивные исследования*. 2022; 1: 138-158. Доступно по: https://iling-ran.ru/kibrik/Segmentation_discourse@Cognitive_studies_2006.pdf. Ссылка активна на 16.05.2022.

References

1. Rakhilina EV, Vyrenkova AS, Mustakimova EG, Ladygina AA, Smirnov IY. Building a learner corpus for Russian*. Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. 2016: 66–75. Available at: https://aclanthology.org/W16-6509.pdf. Accessed April 18, 2022.

2. Glaznieks A, Frey J-C, Stopfner M, Zanasi L, Nicolas L. Leonide: a longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*. 2022; 8 (1): 97 – 120. https://doi.org/10.1075/ijlcr.21004.gla

3. Slioussar NA, Cherepovskaia NV, Denissenko A. Acquisition of the nominal case system in Russian as a second language. *Second Language Research*. 2020. https://doi.org/10.1177/0267658320988058.

4. Aldaihni MS, Al-Houti KS. Perception of English Fricatives (/v/, /ʒ/) by Undergraduate Kuwaiti Arabic Learners. *International Journal of Linguistics and Communication.* 2021; 9 (2): 1 – 17. https://doi.org/10.15640/ijlc.v9n2a1.

5. Nikunlassi A, Protasova EYu. About the importance of errors for the studies of multilingualism. Edited by Nikunlassi A. and Protasova E. Yu. *Slavica Helsingiensia 45. Instrumentarium of the Russian Linguistics: errors and multilingualism.* Helsinki: Unigrafia; 2014; 5–13. Available at: https://blogs.helsinki.fi/slavica-helsingiensia/slavica-helsingiensia-45/. Accessed 15 May, 2022. (In Russ.).

6. Izutsu MN, Izutsu K. American and Irish English speakers' perceptions of the final particles *so* and *but*. *World Enleshes.* 2020; 41 (2): 207 – 233. https://doi.org/10.1111/weng.12521.

7. Bogdanova-Beglaryan NV. Oral corpus as data for analysis of Russian speech. Part 1: Reading. Retelling. Description. Saint-Petersburg: philology faculty of Saint-Petersburg State University; 2013. (In Russ.).

8. Zemskaya EA. Spoken Russian: linguistic analysis and teaching issues. 2nd edition, revised. Moscow: Russian language, 1987. Available at: https://biblioclub.ru/index.php?page=book&id=83088. Accessed 16 May, 2022. (In Russ.).

9. Plonsky L. Advancing quantitative methods in second language research. New York: Routledge; 2015. https://doi.org/10.4324/9781315870908.

10. Voeikova MD. The first language acquisition and the second language acquisition: similarities and differences. *Path to language: monolingualism and multilingualism*. Edited by Tseitlin S.N. and Eliseeva M.B. Moscow: Yazyki slavyanskikh kul'tur; 2011: 49-75. (In Russ.).

11. Fomina NS. Significant and insignificant errors in written test essays produced by foreign learners of Russian. *Concept*. 2015; 12: 136-140. Available at: https://cyberleninka.ru/article/n/kommunikativno-znachimye-i-neznachimye-oshibki-v-pismennoy-rechi-testiruemyh-po-russkomu-yazyku-kak-inostrannomu. Accessed 16 May, 2022. (In Russ.).

12. Berman RA, Slobin DI. Relating events in narrative: A cross-linguistic developmental study. Hillsdale, NJ: L. Erlbaum; 1994.

13. Protasova EYu, Salatov I. Cards, tiny cards and pictures. Vot kak-to tak! Life in illustrations. Saint-Petersburg: Zlatoust, 2016. (In Russ.).

14. Gablasova D, Brezina V, McEnery A. The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research*. 2019; 5 (2): 126-158. https://doi.org/10.1075/ijlcr.19001.gab.

15. Rezanova ZI. Spoken subcorpus of Russian-Turkic bilinguals from the Southern Siberia. Typologically relevant parameters. *Russian journal of lexicography*. 2017; 11: 105-118. https://doi.org/10.17223/22274200/11/7. (In Russ.).

16. Zavadskaya YuO. Arbitrary or regular? The frequency of a wordform and other underlying factors of spoken disfluencies in Russian spontaneous speech. *Acta Linguistica Petropolitana.* 2021; 17(1): 123-142. Available at: doi10.30842/alp23065737171123142 (In Russ.).

17. Bohnacker U, Gagarina N. Developing Narrative Comprehension. Multilingual Assessment Instrument for Narratives. John Benjamins; 2020. https://doi.org/10.1075/sibil.61.

18. Podlesskaya VI, Kibrik AA. Speaker's self-corrections and other types of speech disfluencies as an object of annotation in spoken corpora. *Automatic Documentation and Mathematical Linguistics.* 2007; 2: 2-23. Available at: https://iling-ran.ru/kibrik/Self-repair@NTI_2007.pdf. Accessed 16 May, 2022. (In Russ.).

19. Kibrik AA, Podlesskaya VI. The segmentation issue of spoken discourse of a speaker and cognitive system. *Cognitive studies.* 2022; 1: 138-158. Available at: https://iling-ran.ru/kibrik/Segmentation_discourse@Cognitive_studies_2006.pdf. Accessed 16 May, 2022. (In Russ.).

История статьи:
Получена: 14.04.2022
Принята: 12.05.2022
Опубликована онлайн: 25.06.2022

Article history:
Received: 14.04.2022
Accepted: 12.05.2022
Published online: 25.06.2022

Сведения об авторах:

Власова Екатерина Александровна, кандидат филологических наук, доцент, Национальный исследовательский университет «Высшая школа экономики», Москва, Российская Федерация; e-mail: *evlasova@hse.ru.*

Бец Юлия Васильевна, кандидат филологических наук, доцент, Южный федеральный университет, Ростов-на-Дону, Российская Федерация; e-mail: *betsju@sfedu.ru.*

Каллистратидис Евгения Владимировна, кандидат филологических наук, доцент, Южный федеральный университет, Ростов-на-Дону, Российская Федерация; e-mail: *evakallas@sfedu.ru.*

Bionotes:

Ekaterina A. Vlasova, PhD in Philology, HSE University (National Research University Higher School of Economics), Moscow, Russian Federation; e-mail: *evlasova@hse.ru.*

Yulia V. Bets, PhD in Philology, Southern Federal University, Rostov-on-Don, Russian Federation; e-mail: *betsju@sfedu.ru.*

Evgenia V. Kallistratidis, PhD in Philology, Southern Federal University, Rostov-on-Don, Russian Federation; e-mail: *evakallas@sfedu.ru.*