Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2025»

23–25 апреля 2025 г.

“Okie dokie, here’s the no-cap truth!”: Everyday Russian Youth Speech in Corpus Representation (Structure and Application of the ESC Sound Corpus)

Sherstinova T. Yu., Melnik A. G., Petrova I. A., Azarevich K. I., Melkozerova V. I., Chepovetskaya S. V.
Laboratory for Language Convergence, National Research University Higher School of Economics — St. Petersburg
123A Griboedov Canal Embankment, St. Petersburg, 190068, Russia
tsherstinova@hse.ru, melnik-a@esc-corpus.ru, {iapetrova_2, kiazarevich, vimelkozerova, avchepovetskaya}@edu.hse.ru

Abstract

The article focuses on the creation and potential applications of the Everyday Student Conversations Russian speech corpus (ESC corpus), referred to as KURS corpus in Russian. The corpus is being developed based on a modified methodology of the “one day of speech” recording approach, adopted from the ORD Corpus, with an emphasis on capturing contemporary youth speech. The article highlights the key aspects of the corpus creation process, with special attention to its structure, data collection and processing methodology, as well as the functionality of the corpus online demo version. The primary aim of the project is to explore linguistic changes in the youth environment, create a resource for scientific and applied research, and develop a big data collection of everyday speech for machine learning and advanced artificial intelligence systems. For linguists, the corpus offers unique opportunities to study sociolinguistic, pragmatic, and phonetic features of speech, making it a valuable tool for analyzing contemporary discourse.

Keywords: Russian language; everyday speech; youth speech; corpus linguistics; sociolinguistics; pragmatics; speech technologies; oral discourse

DOI: 10.28995/2075-7182-2025-23-XX-XX

«Оки доки, вот вам без рофла!»: Повседневная речь молодежи в корпусном представлении (структура и назначение звукового корпуса КУРС/ESC)

Шерстинова Т. Ю., Мельник А. Г., Петрова И. А., Азаревич К. И., Мелкозерова В. И., Чеповецкая С. В.
Лаборатория языковой конвергенции, Научно-исследовательский университет «Высшая школа экономики» — Санкт-Петербург
Санкт-Петербург, 190068, наб. канала Грибоедова, д.123, лит. А
tsherstinova@hse.ru, melnik-a@esc-corpus.ru, {iapetrova_2, kiazarevich, vimelkozerova, avchepovetskaya}@edu.hse.ru

Аннотация

Статья посвящена принципам создания и возможностям звукового Корпуса устной речи студентов и молодежи КУРС/ESC (Everyday Student Conversations в английском переводе). Корпус разрабатывается на основе модифицированной методики записи «одного речевого дня», заимствованной из Корпуса ОРД, с акцентом на фиксацию современной речи молодежи. В статье описываются ключевые аспекты процесса создания корпуса, при этом основное внимание уделяется структуре корпуса, методологии сбора и обработки данных, возможностям демоверсии сайта корпуса. Основная цель проекта — изучение изменений в языке молодежной среды, а также создание ресурса для научных и прикладных исследований, формирование «больших данных»

повседневной речи для задач машинного обучения и создания систем сильного искусственного интеллекта. Для лингвистов корпус предлагает уникальные возможности для изучения социолингвистических, прагматических и фонетических характеристик речи, что делает его важным инструментом для анализа современного дискурса.

Ключевые слова: русский язык; повседневная речь; речь молодежи; корпусная лингвистика; социолингвистика; прагматика; речевые технологии; устный дискурс

1 Введение

В статье дается описание нового мультимедийного лингвистического ресурса — Корпуса устной речи студентов и молодежи (КУРС/ESC), содержащего реальные звукозаписи речевого общения, записываемые в течение определенного, достаточно длительного промежутка времени (чаще всего в течение суток или «речевого дня»).

История лонгитюдных звукозаписей повседневной речи берет свои истоки в подходе, впервые реализованном в исследованиях языкового существования (gengo seikatsu), инициированных в конце 1940-х гг. Японским национальным институтом исследования японского языка (NLRI) [4]. Разнообразные по методике проекты были направлены на фиксацию речевого поведения информантов «в течение 24 часов» [5] или сбор речевого материала в естественных условиях общения (напр., для японского Корпуса ESP, входящего в крупный проект JST/CREST, посвященный обработке эмоциональной речи) [6]. По похожей методике записывались данные и для Британского национального корпуса [7]. В отечественном языкознании метод записи речи в течение «одного речевого дня» применялся при создании корпуса ОРД, когда волонтеры-информанты получали на сутки профессиональный диктофон, фиксирующий все их общение в течение этого периода времени [8], [9].

Создание этого корпуса методологически основано на подходе, использованном разработчиками ОРД с небольшими модификациями. Необходимость создания нового ресурса повседневной речи, сфокусированного именно на речи молодежи, объясняется тем, что с момента последних записей ОРД прошло уже 7 лет, с тем пор язык претерпел определенные изменения, наблюдаемые главным образом в речи молодежи и студенчества. Создание корпуса направлено на фиксацию текущего языкового состояния с целью выявления новой разговорной лексики, а также новых значений и коннотаций известных языковых единиц.

2 Методология сбора данных

Проект по получению молодежных звукозаписей по методике «Одного речевого дня» и их корпусному представлению был инициирован в 2023 г. Лабораторией языковой конвергенции НИУ ВШЭ — Санкт-Петербург при поддержке Программы фундаментальных исследований НИУ ВШЭ. Информанты записывают свою речь в течение обычного дня на профессиональный диктофон. Основной единицей записи, получаемой от информантов, является их «речевой день» — совокупность аудиофайлов, отражающих их вербальное взаимодействие с момента пробуждения утром до отхода ко сну вечером.

Как и при сборе данных для корпуса ОРД [10], основным условием остается информированное согласие всех основных участников разговора на проведение записи. Информантам предлагается заранее уведомить своих потенциальных собеседников о предстоящей записи за день до ее начала, чтобы тематика разговоров не сводилась преимущественно к обсуждению деталей звукозаписи. В настоящее время информантом для корпуса может стать любой волонтер, чье речевое поведение осуществляется в русскоязычной (или преимущественно в русскоязычной) речевой среде, даже если русский язык не является для него родным.

Для участия в записи добровольцы сначала заполняют первичную анкету, отражающую их намерение[1], после чего с ними связывается представитель разработчиков, чтобы согласовать наиболее удобный день для получения оборудования. Одновременно с передачей диктофона осуществляется инструктаж, который включает в себя как технические моменты использования аппаратуры, так и организационные вопросы, включая заполнение информационных согласий, анкет и ознакомление с правилами осуществления записи и ее использования в корпусе. В день или

[1] Возможность заявить о желании принять участие в звукозаписи скоро появится на сайте корпуса.

сразу после звукозаписи «речевого дня» информанты заполняют дневник, в котором отмечаются основные события, темы, условия и участники разговоров, а также отвечают на вопросы социологической анкеты про себя и своих собеседников, проходят психологическое тестирование и тест на пассивный словарный запас. Записи передаются в корпус в анонимизированном виде — информация о говорящих кодируются, для открытого представления расшифровок на сайте все личные имена заменяются.

Звукозаписи, полученные от информантов, передаются на экспертное прослушивание и сегментирование на макроэпизоды [12], одновременно происходит их аннотирование с точки зрения типа коммуникации, места коммуникации и социальной роли информанта, а также нормализация этих параметров для корпусного представления.

Отсегментированные на макроэпизоды звуковые файлы становятся основной единицей представления и описания данных в корпусе. Они обрабатываются одной из систем автоматического распознавания речи (см. п. 3.3). После этого автоматически распознанные записи снова передаются экспертам, которые проверяют и исправляют ошибки распознавания и диаризации (атрибуции речи по говорящим). Это наиболее трудоемкий этап создания корпуса, так как системы автоматического распознавания речи до сих пор недостаточно хорошо работают на зашумленных повседневных звукозаписях [15], а лексическое своеобразие молодежной коммуникации еще более усложняет эту задачу. Выверенные расшифровки после этого направляются на токенизацию, лемматизацию и грамматическую разметку.

Несмотря на то, что общая методология сбора данных была заимствована из Корпуса ОРД, методику создания Корпуса КУРС/ESC отличает несколько ключевых моментов:

1) В корпус принимаются как записи целых «речевых дней», так и их фрагменты или отдельные разговоры.
2) У респондентов появилась возможность записи «речевого дня» на собственные смартфоны (при условии высокого качества встроенного диктофона).
3) Произошел полный отказ от бумажной документации: все анкеты, инструкции и согласия заполняются в электронном виде онлайн.
4) Транскрибирование аудиозаписей проводится в полуавтоматическом режиме: сначала создаются черновые транскрипты с использованием систем распознавания речи, которые затем проверяются и корректируются экспертами.
5) При первичной расшифровке транскрипты звукозаписей используют стандартную пунктуацию и членение на «предложения».
6) Планируется привлечение краудсорсинга на запись речевого материала.

3 Структура корпуса КУРС/ESC

Структура звукового корпуса КУРС состоит из следующих основных блоков:

1) Звукозаписи, полученные от информантов.
2) Метаинформация к звукозаписям, описывающая как участников разговора, так и условия коммуникации.
3) Расшифровки звукозаписей, выполненные автоматическим методом.
4) Расшифровки звукозаписей, прошедшие экспертные проверку.
5) Лингвистическая аннотация.
6) Информационная система и утилиты обработки данных.

3.1 Звукозаписи

Для работы используются два вида профессиональных звукозаписывающих устройств: Roland R09-HR и Zoom H1n. Формат выходного файла: WAV, стерео, 16 бит, 44 100Гц. Поскольку в корпус принимаются также записи, выполненные на собственную звукозаписывающую технику информантов, включая мобильные телефоны, их оригинальные записи могут отличаться по формату, количеству каналов и частоте дискретизации. В последнем случае осуществляется их конвертация в указанный выше формат.

3.2 Метаинформация

Метаинформация к звукозаписям, описывающая как участников разговора, так и условия коммуникации, содержит: 1) социологическую анкету информанта и коммуникантов, 2) дневник речевого дня, 3) результаты двух тестов, выполненных информантом (психологического теста и на пассивный словарный запас).

Дневник речевого дня описывает основные события в виде следующей таблицы: 1) код информанта, 2) код коммуниканта, 3) кем приходится этот коммуникант информанту, 4) имя коммуниканта (этот параметр подлежит анонимизации), 5) когда (и сколько) происходило общение, 6) где происходило общение, 7) основные темы разговора. На основе информации из «дневника» осуществляется нарезка речевого материала на макроэпизоды.

Социологическая анкета информанта предполагает ответы на следующие вопросы: 1) пол, 2) возраст, 3) место рождения, 4) текущее место жительства, 5) проживание в других местах (где и как долго), 6) родной язык, 7) языки, которыми владеет информант (с указанием уровня), 8) языки, используемые в семейном общении и с друзьями, 9) профессия родителей, 10) учебное учреждение, 11) образовательная программа или специальность, 12) образовательная ступень, 13) курс, 14) место работы (если есть), 15) вопросы и комментарии, 16) согласие на публикацию записей. Подобная информация по возможности заполняется на всех основных коммуникантов.

Психологическое тестирование осуществляется с помощью пятифакторного опросника личности 5PFQ, позволяющего оценить степень выраженности каждого из пяти факторов «большой пятёрки»: экстраверсии, доброжелательности, добросовестности, эмоциональной стабильности и открытости новому опыту [16]. Кроме того, информантам предложен тест на определение пассивного словарного запаса [17].

Данные, полученные из тестов, заносятся в информационную базу данных и могут быть использованы для изучения социолингвистической и психолингвистической вариативности.

3.3 Автоматические расшифровки звукозаписей

В настоящее время для автоматического распознавания звукового материала используются две ASR системы: 1) Whisper от OpenAI [13], обученная на письменных текстах и порождающая «олитературизированные» расшифровки, напоминающие скорректированные тексты интервью, и 2) акустическая модель, разработанная отечественной компанией ООО НТР [14], в которой нет языковой модели, что позволяет получать звукозапись, похожую на упрощенную фонетическую транскрипцию. Для чернового распознавания речи используется Whisper, а акустическая модель находит свое применение при анализе редуцированных форм. Качество итоговой расшифровки в значительной степени зависит от качества исходной звукозаписи, ее фоновой зашумленности и количества одновременно говорящих. В достаточно «чистых» непродолжительных звукозаписях с хорошим уровнем громкости полезного сигнала показатель WER (доля неправильно распознанных слов) может быть небольшим (напр., 10 %), с ухудшением качества сигнала и увеличением количества говорящих этот показатель может возрастать до 90 %. На момент подготовки статьи, ручная коррекция требуется в 35-40 % словоупотреблений. Для повышения качества автоматического распознавания осуществляется поиск более точных моделей, а также проводится их дообучение (fine-tuning) на собранной звуковой выборке.

Транскрипты звукозаписей автоматически членятся на реплики говорящих, которые могут содержать неограниченное количество фраз. Если фразы одного говорящего идут без длительных пауз, то они заносятся в одну реплику. Транскрипты выгружаются в табличном виде, при этом указываются имя эпизода (звукового файла), начало реплики от начала эпизода, результаты диаризации[2] (выгружаемые в виде *speaker_N*, где *N* — порядковый номер говорящего от 0 до общего количества говорящих в данном эпизоде), и собственно транскрипт реплики.

[2] Диаризация речи — это автоматический процесс сегментации аудиозаписи по говорящим, в ходе которого звуковой поток разбивается на фрагменты, соответствующие индивидуальным репликам, и каждому фрагменту присваивается идентификатор условного участника

3.4 Расшифровки звукозаписей, проверенные экспертами

Расшифровки звукозаписей, выполненные автоматическим методом, проходят экспертную корректуру, в процессе которой звукозапись прослушивается, проверяется корректность идентификации говорящих, заполняется поле SCode, содержащее уникальный код информанта в корпусе, и выверяется корректность расшифровки. Итоговый транскрипт выгружается в базу данных в табличном виде (см. рис. 1).

SFName	Time	Speaker	SCode	Text
escAF001_06m	00:00:01	speaker_0	AF001	Привет.
escAF001_06m	00:00:03	speaker_1	XF001	Как дела?
escAF001_06m	00:00:05	speaker_0	AF001	Средненько.
escAF001_06m	00:00:07	speaker_1	XF001	Что такое?
escAF001_06m	00:00:10	speaker_1	XF001	О, Король и Шут.
escAF001_06m	00:00:11	speaker_1	XF001	У меня подруга снималась там в массовке.
escAF001_06m	00:00:13	speaker_0	AF001	Прикольно.

Рисунок 1: Пример выгрузки транскрипта системы распознавания речи (фрагмент)

В случае наложения речи используется символ @, обозначающий смену говорящего, код второго говорящего указывается в дополнительном столбце.

Пример автоматической расшифровки и результат ее коррекции представлен в табл. 1.

ASR расшифровка	Экспертная коррекция
Рассказывала короче Сон, смотри как ты знаешь я читаю так титанов и поэтому мне теперь часто снятся титаны. Ну потому что они с**и страшные сегодня не исключение и я там непосредственно участвовала в событиях, выдавала себя заикаю. Но то же потом. Сначала, конечно, все отодрались, там что то там превращался, еще какие то люди бегали, еще какие то титаны там, короче, просто мясо, заварушка, крошево какое то, огонь это все на каком то острове происходит. Ехали туда какие то пожарные, посмотрели такие ничего не можем сделать, уехали, короче, вот. То есть просто полное мясо из каких то людей. Вот и мне в какой то момент понадобилось, ну, куда то уйти, я не знаю почему. То есть я там не то чтобы я там что то помогала как то, я просто бегала и пыталась не попасть под руку Тим. Вот. а потом, когда я вернулась, уже все закончилось. все почему то сидят за столом, вроде как наши победи я такая ну здорово.	Рассказываю, короче, сон... Свой. Ээ, как ты знаешь, я читаю "Атаку Титанов", и поэтому мне теперь часто снятся Титаны, ну, потому что они, с**и, страшные. Вот, сегодня не исключение, и я там непосредственно участвовала в событиях, выдавала себя за Микасу. Но это уже потом. Сначала, конечно, все подрались. Там что-то Эрен в титана превращался, еще какие-то люди бегали, еще какие-то титаны там, короче, просто мясо, заварушка, крошево какое-то, огонь, это все на каком-то острове происходит. Приехали туда какие-то пожарные, посмотрели такие: "Ничего не можем сделать", уехали, короче, вот. То есть просто полное мясо из каких-то людей. Вот и мне в какой-то момент понадобилось, ну, куда-то уйти, я не знаю почему. То есть я там не то чтобы я там что-то помогала как-то, я просто бегала и пыталась не попасться под руку Титану. Вот, а потом, когда я вернулась, уже все закончилось. Все почему-то сидят за столом, вроде как наши победили, я такая: "Ну, здорово".

Таблица 1: Сопоставление автоматической и экспертной транскрипции (фрагмент монолога о сне, эпизод *escAF001_03*)

3.5 Лингвистическая аннотация

Токенизация выполняется по пробелам, лемматизация и грамматическая разметка выполняются в MyStem [19]. Выбор этого морфологического парсера обусловлен, во-первых, тем фактом, что именно с его помощью выполнена разметка Национального корпуса русского языка, а во-вторых — высокими показателями эффективности разметки по сравнению с другими парсерами для русского языка [26]. Неологизмы и особенности лексики молодежного сленга отмечаются на отдельном уровне аннотации. В поле комментариев указываются внешние неречевые события, которые могут быть полезны для понимания происходящего.

3.6 Информационная система и утилиты обработки данных

Информация из всех модулей корпуса сводится в единую базу данных, обеспечивающую поиск и фильтрацию записей по запросу пользователя. Одновременно разрабатываются программные средства, осуществляющие ряд вспомогательных функций, таких как анонимизация личных имен и персональной информации, цензурирование (замена звездочками непечатной лексики), конвертация форматов аннотации (например, для перевода табличной транскрипционной записи в формат ELAN [18], используемый в корпусе ОРД), а также конвертация нестандартных форматов звуковых файлов, полученных от информантов.

4 Текущая статистика корпуса КУРС/ESC

К началу 2025 г. получены записи от 90 информантов в возрасте от 16 до 27 лет, средний возраст информантов составляет 22 года. В настоящее время информанты представляют студентов 10 вузов Санкт-Петербурга, юношей и девушек преимущественно гуманитарных специальностей. Примерно половина информантов родилась в Санкт-Петербурге. Среди других упомянутых мест рождения встречаются Красноярск (4,26 % информантов), Екатеринбург, Тюмень, Челябинск (по 3,19 %) и т.д. Кроме того, среди волонтёров, записавших свою речь для корпуса, есть граждане государств СНГ (Казахстана, Белоруссии, Кыргызстана и Узбекистана), при этом в большинстве случаев их родной язык — русский.

На настоящий момент суммарная длительность всех исходных записей составляет 1178 часов. Из них примерно 800 часов записано в 2024 году. В среднем от каждого информанта было получено приблизительно по 12 часов записи речи.

Автоматическое распознавание речи выполняется по мере сегментации. Примерно половина полученных записей прошла через процесс ручного сегментирования, благодаря чему было получено 952 макроэпизода. Экспертная проверка и коррекция транскриптов выполнена для 258 макроэпизодов, что соответствует 84 часам звучания. Сбор новых звукозаписей продолжается. Корпус КУРС/ESC планируется к публикации в открытом доступе для тех материалов, которые имеют разрешение на публикацию.

5 Демоверсия сайта

В настоящее время сайт корпуса КУРС/ESC работает в тестовом режиме по адресу https://esc-corpus.ru. Для отработки возможностей сайта отобраны 38 коммуникативных макроэпизодов, содержащих 12,85 часов речевого материала, полученного от 59 говорящих. Расшифровки для этих эпизодов содержат 59 162 словоформы, относящихся к 4601 реплике.

При отборе данных для сайта приоритетно отбирались звукозаписи, для которых есть разрешение для открытой публикации звукового контента, что дает возможность совместить расшифровку звукозаписей с их звуковым оригиналом[3].

Полученная подборка звукозаписей отражает многогранность повседневной жизни петербургских студентов, включая личное, профессиональное и социальное общение. Сюда входят голосовые сообщения друзьям, обсуждение фильмов, шопинг, разговоры с близкими, включая братьев, сестер, партнеров, друзей и подруг, а также взаимодействие с домашними питомцами и даже с самим собой. Представлены эпизоды учебной и профессиональной деятельности, (например, проведения уроков), а также обсуждения бытовых тем, таких как текущие расходы, питание, здоровье и косметика. Включены звукозаписи, сделанные в общественных местах, таких как кафе, магазины, транспорт, а также представлены моменты уединения, например, размышления вслух и игры в одиночестве.

На рис. 2 представлен принтскрин результатов поиска по словоформе/подстроке с возможностью просмотра кратких метаданных об участниках разговора и особенностях коммуникации.

[3] Публикация звуковых материалов запланирована на 3-й квартал 2025 года.

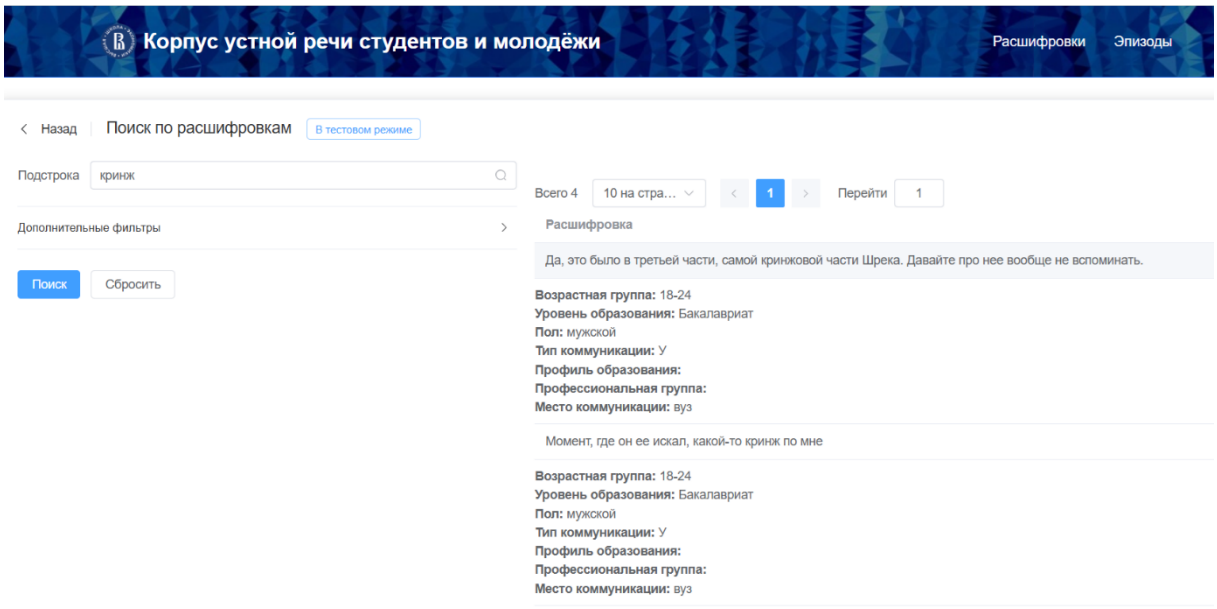

Рисунок 2: Поиск по подстроке/словоформам в расшифровках корпуса КУРС/ESC

На вкладке «Эпизоды» можно использовать фильтры по коммуникативным ситуациям (рис. 3), а также перейти к полным текстам расшифровок для каждого эпизода (рис. 4).

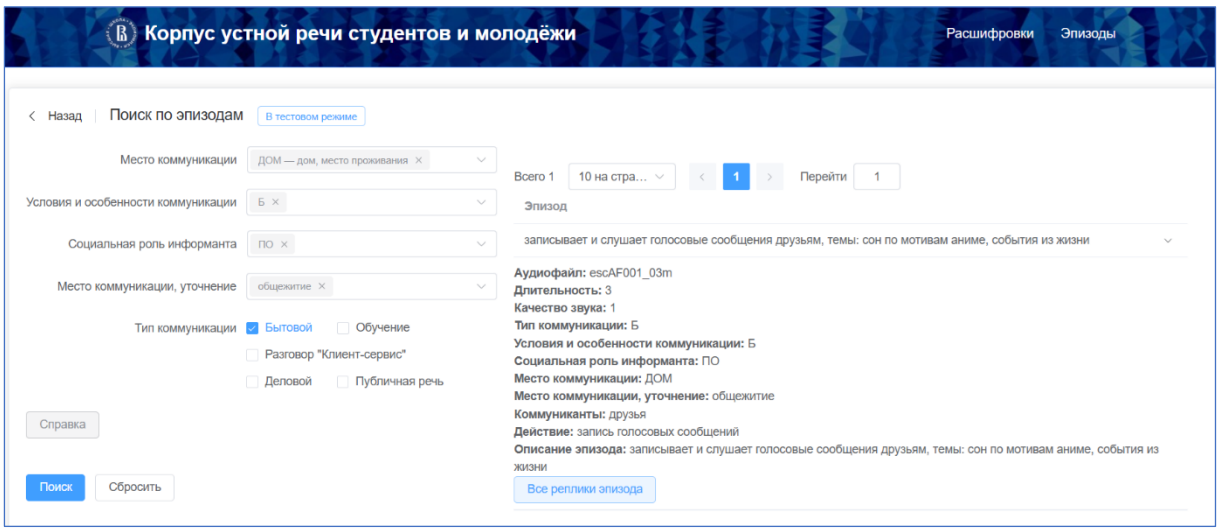

Рисунок 3: Поиск эпизодов по условиям коммуникации в корпусе КУРС/ESC с переходом на полный текст расшифровки (кнопка «Все реплики эпизода»)

Прости меня. Я думал, что я записывал аудио. Смотри, вы до начала кинопоказа лучше пообедайте, потому что столовая работает до семи. Вы подходите, говорите, что на вас заложено питание в рамках день ментального здоровья. У них там листочек в клеточку, список из десяти человек, среди них есть Эмма%, есть Ваня%. И просто расписываетесь, набираете на триста пятьдесят рублей все, что хотите.	›	XM001	00:02:07
Девочки, прикиньте, у меня сегодня питание в столовке на триста пятьдесят рублей, вот это вообще!	›	AF001	00:02:49
И как мне в этом качестве смотреть? Вы с дуба рухнули?	›	AF001	00:02:59

Рисунок 4: Выгрузка полного текста эпизода с указанием кода говорящего и временной метки (фрагмент)

В 2025 году поиск по транскриптам корпуса будет расширен за счет поиска по леммам и грамматическим категориям, а планируется выборочное размещение звукового контента. Количество эпизодов, представленных на сайте, также будет увеличено.

Корпус КУРС/ESC планируется расширять с использованием краудсорсинговых возможностей. Для этого на сайте будет организована возможность «донировать» собственные материалы (звукозаписи и их транскрипты), подобно тому, как это сделано в Американском национальном корпусе [20] или отечественных ресурсах: корпусе дневников и воспоминаний «Прожито» [21] и корпусе открыток «Пишу тебе» [22].

6 Возможности и научная новизна корпуса КУРС/ESC

Материалы корпуса КУРС предназначены для решения следующих актуальных задач:

1) Фиксировать ускользающую звучащую материю, отражающую как текущее состояние языка повседневного общения, так и принятые на данный момент в обществе особенности коммуникативного поведения. Звукозаписи повседневной речи могут стать ценным архивом для сохранения языкового и культурного наследия, документируя изменения в языке и коммуникативном поведении различных групп общества.

2) Служить основой для проведения фундаментальных исследований языка повседневного общения и устного дискурса в целом — изучать особенности фонетики, просодии, лексики, синтаксиса, семантики, прагматики, структуру и законы коммуникации, а также социолингвистическую вариативность этих явлений, гендерные и возрастные особенности, диалекты и жаргоны, а также изменения в языке под влиянием технологий и глобализации.

3) Применяться для решения целого спектра прикладных задач — от уже упомянутого обучения больших языковых моделей, систем распознавания и синтеза речи, обеспечивая модели аутентичными данными, отражающими разнообразие акцентов, интонаций и речевых паттернов, до создания реалистичных учебных материалов, которые помогают учащимся адаптироваться к живой, естественной речи, включая изучение интонации, ритма и типичных ошибок носителей. Данные повседневного общения могут использоваться для создания приближенных к реальности диалогов для чат-ботов и виртуальных сред самого разного назначения, где требуется высокая степень достоверности воссоздания речевых взаимодействий.

4) Использоваться для решения научных и практических задач в ряде дисциплин, изучающих поведение человека в обществе и взаимодействие социальных групп, таких как социология, антропология, культурология, а также психология, лингвистика и межкультурная коммуникация, предоставляя уникальные данные для анализа социальных и культурных процессов.

Научная новизна корпуса КУРС/ESC по сравнению с существующими корпусами русской повседневной речи заключается в следующем:

1) До настоящего времени не существовало специализированного корпуса, посвящённого современной студенческой и молодёжной устной речи, записанной в естественных условиях после 2020 года. С учётом стремительных языковых изменений, вызванных цифровизацией, глобализацией и влиянием соцсетей, КУРС/ESC фиксирует новую языковую реальность, которую невозможно смоделировать по более ранним корпусам (например, ОРД).

2) Совмещение полевых методов и автоматических технологий распознавания речи. Впервые для корпуса русской повседневной речи внедрена полуавтоматическая схема обработки данных, включающая комбинацию двух разных ASR-систем (одна — с языковой моделью, другая — без нее) с последующей экспертной верификацией, что позволяет ускорить процесс транскрипции.

3) Помимо лингвистической разметки и социолингвистической информации о респондентах, в корпус впервые включены результаты теста на пассивный словарь и пятифакторного опросника личности, что открывает новые перспективы для междисциплинарных исследований — в частности, для анализа взаимосвязей между языковым поведением и личностными характеристиками говорящих.

4) Благодаря методологической преемственности с корпусом ОРД, новый корпус позволяет проводить сопоставление данных с временным интервалом более 15 лет, что создает уникальную возможность отслеживания динамики развития повседневного дискурса в русской устной речи.

5) Проект ориентирован на открытое распространение данных, а также на формирование модели участия пользователей в наполнении корпуса. Такая модель ранее не применялась в русскоязычных аудиокорпусах и позволяет значительно расширить репрезентативность собранного материала.

7 Некоторые наблюдения о лексических особенностях речи современных студентов

Несмотря на тот факт, что корпус КУРС/ESC находится в состоянии разработки, полученные данные можно использовать для выявления наиболее характерных особенностей современной молодежной речи. Покажем это на примере лексики.

Материалом для представленного в данном разделе исследования послужили текстовые расшифровки 203 коммуникативных макроэпизодов [12] молодежной студенческой коммуникации, которые содержат в общей сложности 283643 словоупотреблений. Эти коммуникативные эпизоды были сначала автоматически расшифрованы с помощью акустической модели распознавания, разработанной российской компанией ООО НТР [14], после чего была выполнена их экспертная проверка. В выборке представлена речь 212 студентов (156 девушек и 56 юношей). Относительная несбалансированность корпусного материала по половому признаку носит временный характер и объясняется тем, что запись корпуса была начата на гуманитарных факультетах, где большинство студентов традиционно составляют девушки.

Рассмотрим наиболее частотную лексику молодежной коммуникации. В табл. 2 представлен список из 100 самых частотных слов (без лемматизации), которые покрывают 53% всей лексики в проанализированной выборке. Для каждого слова приводятся его абсолютная и относительная частоты, а также накопленная частота (кумулятивный процент). В качестве «реферативного корпуса» для сравнения используется выборка из корпуса ОРД, описанная в статье [24], состоящая из 152 макроэпизодов и содержащая речь 209 человек, записанную в Санкт-Петербурге в 2007 и 2010 гг. (общим объемом подкорпуса в 232370 словоупотреблениях), и соответствующий ей частотный список наиболее употребительных слов повседневной устной речи [там же].

Список самых частотных слов молодежной речи демонстрирует преобладание личных местоимений (*я*, *ты*, *он*, *она*, *мы* и их словоизменительных форм), служебных слов (союзов, предлогов, частиц), дискурсивных и прагматических маркеров, и во многом напоминает список наиболее частотной лексики, полученный на материале корпуса ОРД. Как и для языка повседневного общения в целом, значительную часть этого списка составляют дискурсивные и прагматические маркеры, играющие важную роль в структурировании и управлении дискурсом (такие как *вот*, *ну*, *так*, *короче*, *ладно*). Это говорит о том, что молодежный лексикон не только насыщен субъективными и эмоциональными элементами, но и активно использует средства организации речи, направленные на поддержание контакта и управления коммуникативным процессом.

Также заметно широкое использование частиц и вводных слов, придающих речи оттенки модальности и оценки: *просто*, *типа*, *кажется*, *наверное*, *конечно*. Они выполняют функции уточнения, смягчения высказывания или маркирования неуверенности, что типично для разговорной речи молодых людей, особенно в неформальной обстановке.

Ранг	Слово	Абс. ч.	%	Кумул.%	Ранг	Слово	Абс. ч.	%	Кумул.%
1	я	8350	2,94	2,94	51	тебе	755	0,27	44,60
2	не	7342	2,59	5,53	52	может	743	0,26	44,86
3	что	6618	2,33	7,87	53	за	727	0,26	45,12
4	ну	6144	2,17	10,03	54	можно	724	0,26	45,37
5	да	5725	2,02	12,05	55	тебя	716	0,25	45,63
6	вот	5514	1,94	13,99	56	тут	644	0,23	45,85
7	и	5367	1,89	15,89	57	потом	642	0,23	46,08
8	в	5216	1,84	17,73	58	такой	642	0,23	46,31
9	это	4867	1,72	19,44	59	к	585	0,21	46,51
10	а	4809	1,70	21,14	60	ой	585	0,21	46,72
11	у	3712	1,31	22,45	61	такая	585	0,21	46,93
12	там	3513	1,24	23,68	62	только	583	0,21	47,13
13	на	3281	1,16	24,84	63	его	578	0,20	47,33
14	то	3113	1,10	25,94	64	хорошо	554	0,20	47,53
15	как	3013	1,06	27,00	65	такое	539	0,19	47,72
16	ты	2711	0,96	27,96	66	короче	535	0,19	47,91
17	так	2706	0,95	28,91	67	из	531	0,19	48,10
18	все	2395	0,84	29,75	68	чтобы	527	0,19	48,28
19	с	2373	0,84	30,59	69	раз	518	0,18	48,46
20	мне	2049	0,72	31,31	70	где	515	0,18	48,65
21	он	1899	0,67	31,98	71	для	507	0,18	48,82
22	просто	1837	0,65	32,63	72	до	498	0,18	49,00
23	есть	1823	0,64	33,27	73	даже	487	0,17	49,17
24	нет	1715	0,60	33,88	74	была	473	0,17	49,34
25	меня	1700	0,60	34,48	75	конечно	472	0,17	49,51
26	мы	1625	0,57	35,05	76	что-то	466	0,16	49,67
27	она	1624	0,57	35,62	77	наверное	453	0,16	49,83
28	но	1508	0,53	36,15	78	быть	446	0,16	49,99
29	еще	1478	0,52	36,68	79	почему	441	0,16	50,14
30	типа	1377	0,49	37,16	80	этот	435	0,15	50,30
31	бы	1270	0,45	37,61	81	про	434	0,15	50,45
32	они	1268	0,45	38,06	82	о	424	0,15	50,60
33	угу	1252	0,44	38,50	83	кажется	419	0,15	50,75
34	по	1195	0,42	38,92	84	ничего	418	0,15	50,89
35	же	1174	0,41	39,33	85	блин	414	0,15	51,04
36	очень	1163	0,41	39,74	86	давай	407	0,14	51,18
37	если	1059	0,37	40,12	87	ее	407	0,14	51,33
38	уже	1051	0,37	40,49	88	знаешь	394	0,14	51,46
39	вообще	1030	0,36	40,85	89	был	391	0,14	51,60
40	знаю	1025	0,36	41,21	90	их	383	0,14	51,74
41	потому	1005	0,35	41,56	91	от	382	0,13	51,87
42	было	973	0,34	41,91	92	тогда	379	0,13	52,01
43	тоже	965	0,34	42,25	93	здесь	376	0,13	52,14
44	надо	957	0,34	42,59	94	такие	376	0,13	52,27
45	когда	951	0,34	42,92	95	два	375	0,13	52,40
46	вы	859	0,30	43,22	96	какой	373	0,13	52,53
47	нас	828	0,29	43,52	97	эти	372	0,13	52,67
48	сейчас	800	0,28	43,80	98	ладно	368	0,13	52,80
49	или	766	0,27	44,07	99	вас	361	0,13	52,92
50	будет	758	0,27	44,33	100	нужно	353	0,12	53,05

Таблица 2: Наиболее употребительные слова устной студенческой речи

В таблице 3 представлены списки наиболее частотных слов, по употреблению которых современная студенческая речь больше всего отличается от общего частотного списка [24]. Первые четыре столбца показывают списки слов, ранг или доля которых в молодежной речи существенно больше, чем для общего списка. Любопытно, что в этих группах слов лидируют два — *типа*[4] и *короче*[5], которые в 2016 г. возглавляли список слов, характерных для мужской речи [там же].

Частотная лексика, чаще встречающаяся в студенческой речи				**Частотная лексика, реже встречающаяся в студенческой речи**			
По разности рангов		**По разности долей**		**По разности рангов**		**По разности долей**	
типа	>50	что	5520	значит	-120	вот	-3846
короче	34	типа	>4855	говорит	-86	так	-2690
из	30	я	3832	чего	-49	а	-2511
потому	28	то	3310	здесь	-46	там	-2342
для	22	просто	2710	говорю	-44	да	-2031
такая	19	не	2646	вам	-32	сейчас	-1987
быть	18	очень	1720	ага	-28	угу	-1830
такое	17	была	1668	ли	-23	значит	-1811
чтобы	17	мне	1660	давай	-22	ну	-1608
просто	16	потому	1611	вас	-20	говорит	-1468
очень	15	наверное	1597	сейчас	-20	нет	-1386
даже	15	почему	1555	знаешь	-18	здесь	-1157
можно	14	как	1543	один	-16	говорю	-852
когда	13	про	1530	кто	-11	надо	-727
раз	12	о	1495	эти	-10	чего	-705
до	11	ты	1480	угу	-10	ага	-613
или	10	кажется	1477	ничего	-10	давай	-579
за	10	ее	1435	надо	-10	знаешь	-543

Таблица 3:Динамика встречаемости наиболее частотной лексики (топ-100) в речи современных студентов по сравнению с повседневной речью корпуса ОРД

Также можно отметить увеличение частоты *такая* и *такой* в функции ксенопоказателей (*я такая…* и т. п.), которое закономерно снижает долю *говорю* и *говорит*. В речи студентов уменьшается роль метакоммуникатива *знаешь*, прагматического маркера *значит*, маркеров обратной связи *угу* и *ага*, частиц *давай* и *да*, а также делимитативного маркера *вот*. *Здесь* и *сейчас* также, неожиданно, теряют популярность в молодежной среде.

Возможно, часть отмеченных особенностей касаются не только молодежной речи, но и языковых тенденций в целом. Однако для уточнения этого вопроса необходимо расширение корпуса с привлечением информантов всех возрастных групп.

Характерной для молодежной студенческой речи является употребление английских заимствований. Ниже приводится список наиболее частотных заимствований с указанием абсолютной частоты и ipm (в скобках), их можно условно поделить на 4 группы:

1) лексика, связанная с передачей эмоций и реакций: *окей* — 42 (148), *вау* — 19 (67), *трэш* — 6 (21), *кринж* — 30 (106), *кринжовый* — 5 (18),

2) лексика, относящаяся к обучению, коучингу и бизнесу: *коучи* — 27 (95), *коуч* — 24 (85), *коучинг* — 14 (49), *коучинга* — 7 (25), *коворкинг* — 6 (21), *прокторинг* — 6 (21), *оффер* — 9 (32),

3) лексика, связанная с интернет-активностями и мультимедиа: *видос* — 22 (78), *мем* — 17 (60), *загуглить* — 11 (39), *ватсап* — 10 (35), *рилс* — 7 (25), *сторис* — 7 (25),

4) лексика, связанная с модой и молодежной культурой: *свитшот* — 14 (49), *бейби* — 5 (18), *вайб* — 5 (18).

[4] https://www.ord-multimedia-dict.com/tipa

[5] https://www.ord-multimedia-dict.com/koroche

Следует отметить, что некоторые из этих слов, такие как *окей*, *вау*, *загуглить*, *оффер* и др. в настоящее время используются не только в молодежной студенческой речи, но и в более широком контексте современного общения, выходя за рамки возрастных и социальных групп.

Представленную в данном разделе информацию следует рассматривать как предварительную. По мере увеличения объема расшифрованной части корпуса эти данные будут скорректированы.

8 Заключение

Создание звукового корпуса повседневной речи, собранной «полевым способом», представляет собой достаточно трудоемкий и неизбежно долгий процесс. На настоящем этапе работы в фокусе интереса разработчиков находится в первую очередь лексико-грамматические, социолингвистические и прагматические характеристики молодежного дискурса. Фонетические и просодические особенности устной речи пока остаются за кадром нашей работы. Мультимедийный характер исходного материала позволяет расширять корпус и в этом направлении. Представляется, что для аннотирования речевых данных на фонетическом уровне также уже можно использовать современные речевые технологии — от акустических моделей, записывающих речь «как она слышится», что позволяет получить упрощенную фонетическую транскрипцию [23], до использования полифункциональных систем извлечения речевых характеристик речи, таких как openSMILE [25].

Совмещение технологий новейших систем распознавания и обработки речи с корпусными методами представляется более чем перспективным. Результаты такого технологического объединения позволят не только вывести традиционные корпусные лингвистические исследования на уровень «больших данных» в том числе и для устной речи, но и смогут служить в качестве обучающих данных для больших языковых моделей и создания сильного искусственного интеллекта [1], [2], [3]. В данном случае «большие данные» относятся не к текущим масштабам создаваемого ресурса, а к особенностям его структуры и потенциала, предназначенного для восполнения нишевого дефицита русскоязычных разговорных ресурсов, необходимых для развития прикладных речевых технологий.

Корпус КУРС/ESC представляет собой многоуровневый массив аутентичной повседневной русской речи с богатыми аннотациями (социо-, психолингвистическими, лексико-грамматическими), который может использоваться: 1) для дообучения и тестирования моделей автоматического распознавания речи (ASR) на реальных, зашумленных и спонтанных данных; 2) как источник редуцированной, разговорной и сленговой лексики для адаптации языковых моделей и синтеза речи; 3) в качестве «золотого стандарта» для оценки качества NLP-систем, работающих с устной речью. По количеству записанного речевого материала этот ресурс уже превосходит известный CSJ корпус спонтанной японской речи, который содержит 650 часов звукозаписей [27].

Запись повседневной русской речи методом суточной фиксации проводится в Санкт-Петербурге спустя 15 лет после первых записей корпуса ОРД и основана на схожей методологии. Это позволяет выстраивать серии записей этих двух корпусов в единую лонгитюдную цепочку, предоставляя уникальную возможность отслеживать изменения и развитие повседневного русского дискурса. Представляется, что подобные серии звукозаписей могут быть продолжены в будущем, что обеспечит возможность долгосрочного мониторинга и анализа динамики языковых и коммуникативных явлений.

Благодарности

Статья подготовлена в рамках проекта «Текст как Big Data: методы и модели работы с большими текстовыми данными» (ФИ-2025-32) для выполнения фундаментальных проектов тематического плана научных исследований, предусмотренных Государственным заданием Национального исследовательского университета «Высшая школа экономики» на 2025 год.

References[6]

[1] Muehlhauser Luke. What Is AGI? // Machine Intelligence Research Institute. — August 11, 2013. — Access mode: https://intelligence.org/2013/08/11/what-is-agi/.

[6] References, Scopus version

[2] Antebi L. General Artificial Intelligence // Artificial Intelligence and National Security in Israel. — Institute for National Security Studies, 2021. — Pp. 59–60.
[3] Chim J., Ive J., Liakata M. Evaluating Synthetic Data Generation from User Generated Text // Computational Linguistics. — 2024. — Pp. 1–44.
[4] Shibamoto Janet S. Japanese Sociolinguistics // Annual Review of Anthropology. — 1987. — Vol. 16. — Pp. 261–278.
[5] Shibata Takeshi. Study of Language Behavior over 24 Hours [Issledovanie yazykovogo sushchestvovaniya v techenie 24 chasov]. Translated by V.M. Alpatov // Alpatov V.M. (ed.). Linguistics in Japan [Yazykoznanie v Yaponii]. — Moscow: Raduga, 1983. — Pp. 134–141.
[6] Campbell N. Speech & Expression; the Value of a Longitudinal Corpus // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). — Eds. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva. — Lisbon, 2004. — Pp. 183–186.
[7] Reference Guide for the British National Corpus. — Access mode: http://www.natcorp.ox.ac.uk/docs/URG.xml. — Last accessed: 10/01/2025.
[8] Everyday Speech Corpus (ORD) Homepage. — Access mode: https://ord.spbu.ru/.
[9] Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // TSD 2009. — Eds. Matoušek, V., Mautner, P. — Lecture Notes in Artificial Intelligence (LNAI), vol. 5729. — Springer, Berlin-Heidelberg, 2009. — Pp. 250–257.
[10] Sherstinova T., Kolobov R., Mikhaylovskiy N. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level // SPECOM 2023. — LNCS, vol. 14338/14339. — Springer, 2023. — Pp. 43–56.
[11] Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9811. — Springer, Switzerland, 2016. — Pp. 659–666.
[12] Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus // SPECOM 2015. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9319. — Springer, Switzerland, 2015. — Pp. 268–276.
[13] Whisper Model Homepage. — Access mode: https://openai.com/index/whisper/.
[14] NTR Homepage. — Access mode: https://ntr.ai. — Last accessed: 2024/07/25.
[15] Kolpashchikova E.O. Writer Robin Dranattagor: Testing the Whisper Model on Russian-Language Spoken Speech [Pisatel' Robin Dranattagor: aprobatsiya modeli Whisper na russkoyazychnoi zvuchashchei rechi] // Socio- and Psycholinguistic Studies [Socio- i psikholingvisticheskie issledovaniya]. — 2023. — Issue 11. — Pp. 23–27.
[16] Khromov A.B. Five-Factor Personality Questionnaire: A Training Manual [Pyatifaktornyi oprosnik lichnosti. Uchebno-metodicheskoe posobie]. — Kurgan: KSU, 2000.
[17] Golovin G.V. Measuring Passive Vocabulary in Russian [Izmerenie passivnogo slovarnogo zapasa russkogo yazyka] // Socio- and Psycholinguistic Studies [Socio-i psikholingvisticheskie issledovaniya]. — No. 3. — 2015. — Pp. 148–159.
[18] ELAN Homepage. — Access mode: http://tla.mpi.nl/tools/tla-tools/elan. — Last accessed: 2024/07/25.
[19] MyStem Homepage. — Access mode: https://yandex.ru/dev/mystem/.
[20] American National Corpus (ANC) Homepage. — Access mode: https://anc.org/contribute/texts/.
[21] Prozhito Project Homepage [Proekt "Prozhito"]. — Access mode: https://prozhito.org/.
[22] Pishu Tebe Project Homepage [Proekt "Pishu tebe"]. — Access mode: https://pishutebe.ru/.
[23] Sherstinova T., Mikhaylovskiy N., Kolpashchikova E., Kruglikova V. Bridging Gaps in Russian Language Processing: AI and Everyday Conversations // 35th Conference of Open Innovations Association (FRUCT). — Tampere, Finland, 2024, April 24–26. — Pp. 253–258.
[24] Шерстинова Т. Ю. Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации) [The Most Frequent Words in Everyday Spoken Russian (in the gender dimension and depending on communication settings)] // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1-4 июня 2016 г.). Вып. 15 (22). М.: Изд-во РГГУ, 2016. С. 616-631.
[25] Eyben F., Weninger F., Gross F., Schuller B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor // Proceedings of the 21st ACM International Conference on Multimedia (MM '13). — 2013. — Pp. 835–838.
[26] Kotelnikov E., Razova E., Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? // Proceedings of the Conference on Artificial Intelligence and Natural Language — 2018. — Pp. 131–142.
[27] The corpus of Spontaneous Japanese: https://clrd.ninjal.ac.jp/csj/misc/preliminary/index_e.html — Last accessed: 28/03/2025.

References[7]

[1] Muehlhauser Luke. What Is AGI? // Machine Intelligence Research Institute. — August 11, 2013. — Access mode: https://intelligence.org/2013/08/11/what-is-agi/.
[2] Antebi L. General Artificial Intelligence // Artificial Intelligence and National Security in Israel. — Institute for National Security Studies, 2021. — Pp. 59–60.
[3] Chim J., Ive J., Liakata M. Evaluating Synthetic Data Generation from User Generated Text // Computational Linguistics. — 2024. — Pp. 1–44.
[4] Shibamoto Janet S. Japanese Sociolinguistics // Annual Review of Anthropology. — 1987. — Vol. 16. — Pp. 261–278.
[5] Сибата Такэси. Исследование языкового существования в течение 24 часов. Перевод В. М. Алпатова // Алпатов В. М. (сост.) Языкознание в Японии. Общая редакция И. Ф. Вардуля. — М.: Радуга, 1983. — С. 134–141.
[6] Campbell N. Speech & Expression; the Value of a Longitudinal Corpus // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). — Eds. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva. — Lisbon, 2004. — Pp. 183–186.
[7] Reference Guide for the British National Corpus. — Access mode: http://www.natcorp.ox.ac.uk/docs/URG.xml. — Last accessed: 10/01/2025.
[8] Everyday Speech Corpus (ORD) Homepage. — Access mode: https://ord.spbu.ru/.
[9] Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // TSD 2009. — Eds. Matoušek, V., Mautner, P. — Lecture Notes in Artificial Intelligence (LNAI), vol. 5729. — Springer, Berlin-Heidelberg, 2009. — Pp. 250–257.
[10] Sherstinova T., Kolobov R., Mikhaylovskiy N. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level // SPECOM 2023. — LNCS, vol. 14338/14339. — Springer, 2023. — Pp. 43–56.
[11] Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9811. — Springer, Switzerland, 2016. — Pp. 659–666.
[12] Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus // SPECOM 2015. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9319. — Springer, Switzerland, 2015. — Pp. 268–276.
[13] Whisper Model Homepage. — Access mode: https://openai.com/index/whisper/.
[14] NTR Homepage. — Access mode: https://ntr.ai. — Last accessed: 2024/07/25.
[15] Kolpashchikova E.O. Pisatel' Robin Dranattagor: aprobatsiya modeli Whisper na russkoyazychnoi zvuchashchei rechi // Socio- i psikholingvisticheskie issledovaniya (Socio- and Psycholinguistic Studies). — 2023. — Issue 11. — Pp. 23–27.
[16] Хромов А. Б. Пятифакторный опросник личности. Учебно-методическое пособие. — Курган: КГУ, 2000.
[17] Головин Г. В. Измерение пассивного словарного запаса русского языка // Социо-и психолингвистические исследования. — № 3. — 2015. — С. 148–159.
[18] ELAN Homepage. — Access mode: http://tla.mpi.nl/tools/tla-tools/elan. — Last accessed: 2024/07/25.
[19] MyStem Homepage. — Access mode: https://yandex.ru/dev/mystem/.
[20] American National Corpus (ANC) Homepage. — Access mode: https://anc.org/contribute/texts/.
[21] Проект «Прожито» Homepage. — Access mode: https://prozhito.org/.
[22] Проект «Пишу тебе» Homepage. — Access mode: https://pishutebe.ru/.
[23] Sherstinova T., Mikhaylovskiy N., Kolpashchikova E., Kruglikova V. Bridging Gaps in Russian Language Processing: AI and Everyday Conversations // 35th Conference of Open Innovations Association (FRUCT). — Tampere, Finland, 2024, April 24–26. — Pp. 253–258.
[24] Sherstinova T. The Most Frequent Words in Everyday Spoken Russian (in the gender dimension and depending on communication settings) // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 616 - 631.
[25] Eyben F., Weninger F., Gross F., Schuller B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor // Proceedings of the 21st ACM International Conference on Multimedia (MM '13). — 2013. — Pp. 835–838.
[26] Kotelnikov E., Razova E., Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? // Proceedings of the Conference on Artificial Intelligence and Natural Language — 2018. — Pp. 131–142.
[27] The corpus of Spontaneous Japanese: https://clrd.ninjal.ac.jp/csj/misc/preliminary/index_e.html — Last accessed: 28/03/2025.

[7] References, РИНЦ version