ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет гуманитарных наук**

**Программа подготовки бакалавров по направлению**

**45.03.03 «Фундаментальная и прикладная лингвистика»**

Жилина Полина Павловна

# ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Создание корпуса устной речи для диагностических целей с использованием автоматических алгоритмов распознавания речи

Рецензент
приг. преп.,
А.Д. Микулинский

Научный руководитель
к. ф. н., доцент, с.н.с.
ЦЯиМ, А. Ю. Хоменко;
Соруководитель
приг. преп., А.А. Бадасян

Нижний Новгород, 2025

**Содержание**

Введение....3
**Глава 1. Изучение устной речи в контексте корпусной лингвистики....9**
1.1. Устная речь как первичная форма языка и объект лингвистического анализа....9
1.2. Корпусы речи: классификация, принципы построения и специфика специализированных устных корпусов....11
1.2.1. Корпусы речи и их разновидности....11
1.2.2. Устный корпус....15
1.2.3. Конструирование речевого (устного) корпуса....15
1.2.4. Специализированный корпус....17
1.3. Диагностические корпуса....20
1.4. Виды разметки речевых корпусов....21
1.5. Основные способы создания транскрипций устной речи....24
1.6. Оценка качества ASR моделей....26
1.6.1. Автоматически рассчитываемые метрики оценки качества моделей...26
1.6.2. Лингвистически ориентированные подходы к классификации ошибок систем автоматического распознавания речи....29
**Глава 2. Трансформация аудиоданных в диагностически значимый корпус: автоматическая транскрипция, анализ ошибок, аннотирование данных и структурирование репозитория....33**
2.1. Описание материала....33
2.2. Создание транскриптов для корпуса....34
2.2.1. Предобработка данных....35
2.2.2. Отбор модели для автоматической транскрибации речи....36
2.2.2.1. Подсчет автоматических метрик....36
2.2.2.2. Лингвистический анализ ошибок ASR-моделей....39
2.2.2.2.1. Ошибки замен....40
2.2.2.2.1. Ошибки удалений....45
2.3. Создание разметки для диагностического корпуса....48
2.3.1. Разработка протокола аннотирования....49
2.4. Создание архитектуры корпуса....55
**Заключение....57**
**Список литературы....59**
**Приложение....65**

## Введение

Фиксация и анализ устной речи занимают центральное место во множестве прикладных задач – от педагогической и судебной диагностики до социолингвистических исследований и оценки когнитивного состояния говорящего. В условиях стремительного развития технологий и растущего интереса к использованию речевых данных в междисциплинарных исследованиях **актуальным** становится вопрос об эффективной, воспроизводимой и масштабируемой трансформации устной речи в формат, пригодный для систематического анализа. Для проведения диагностики важно не только наличие аудиозаписей речевого материала, но и его точный транскрипт, обеспечивающий возможность повторного анализа и сопоставление различных речевых признаков. Так, например, в педагогической практике наличие транскриптов позволяет экспертам повторно оценить речь ребёнка (Васютина, 2007), а в судебной лингвистической экспертизе аудиозаписи и их текстовые транскрипты рассматриваются в комплексе для извлечения как физиологических характеристик говорящего, так и для анализа содержания речи (Рычкалова, 2002). В социологических исследованиях транскрипты устных интервью являются основой анализа, поскольку позволяют точно зафиксировать как вербальные, так и паравербальные особенности высказываний, что критически важно для последующей интерпретации (Готлиб и др., 2019).

Современные технологии автоматического распознавания речи (Automatic Speech Recognition, ASR) открывают новые возможности для масштабного и более оперативного создания транскриптов, особенно в условиях дефицита как человеческих, так и временных ресурсов для расшифровки. Однако эффективность ASR-систем может существенно варьироваться в зависимости от качества и условий записи, индивидуальных речевых особенностей говорящего (качество артикуляции, темпа речи), наложения реплик наличии

нескольких участников на аудиозаписи, что делает необходимым их оценку в контексте задач лингвистического анализа и диагностики.

На сегодняшний день модель Whisper от OpenAI, а также её модификации, адаптированные под конкретные задачи, например, распознавание акцентной речи или обработку малоресурсных языков, заняли устойчивое место среди инструментов автоматического распознавания речи, широко применяемых в корпусной и прикладной лингвистике (Graham, Roll, 2024; Liu, Yang, Qu, 2024). Популярность этой системы во многом объясняется её открытой архитектурой, мультилингвальной направленностью и устойчивостью к шуму, что делает её удобным решением для разнообразных исследовательских задач. В ряде недавних работ продемонстрировано использование Whisper и для русскоязычных задач, в частности при создании специализированных корпусов и автоматизации транскрибирования (Мамаев, 2023; Sherstinova и др., 2024).

**Новизна** настоящего исследования состоит в том, что в дополнение к системе Whisper в нём рассматриваются и более современные модели автоматического распознавания речи, в частности архитектура FastConformer и основанная на ней модель GigaAM-RNNT, специально разрабатываемая для русскоязычных данных. Архитектура FastConformer представляет собой один из наиболее современных подходов в области ASR и активно применяется в различных исследовательских и прикладных проектах благодаря высокой точности и способности эффективно обрабатывать длинные аудиофрагменты (Noroozi и др., 2024; Rekesh и др., 2023). GigaAM, в свою очередь, обучалась на масштабном корпусе русской речи, что делает её особенно релевантной для задач транскрибирования на русском языке.

Несмотря на высокие технические характеристики, данные модели сравнительно редко используются в лингвистических исследованиях, что отчасти связано с их более сложной интеграцией по сравнению с Whisper. Выбор этих моделей в данном исследовании обусловлен стремлением как опереться на широко используемую и доступную систему, так и оценить

потенциал современных архитектур, демонстрирующих передовые результаты в ASR. Их сравнительный анализ в контексте задачи создания корпуса устной речи позволяет выявить сильные и слабые стороны каждого подхода, а также определить направления для последующего совершенствования автоматической транскрипции в лингвистических и диагностических целях.

**Объектом** данного исследования выступает устная речь, продуцируемая в условиях правдивого и ложного высказывания. **Предметом** исследования являются особенности автоматической транскрипции устной речи с использованием современных моделей распознавания речи (ASR), включая типологию и лингвистическую значимость возникающих ошибок, а также принципы создания корпуса, ориентированного на распознавание и диагностику правдивых и ложных высказываний.

**Целью** настоящей работы является создание корпуса устной речи для диагностических целей, включающего как монологическую, так и диалогическую речь, с применением современных алгоритмов автоматического распознавания речи и последующим лингвистическим анализом полученных транскриптов.

Для достижения поставленной цели необходимо решить ряд **задач**:

- проанализировать научную литературу, посвящённую созданию речевых корпусов, включая исследования, ориентированные на диагностические цели и применение технологий автоматического распознавания речи (ASR) в лингвистике;
- получить автоматические транскрипции с использованием ASR-моделей, основанные на различных архитектурах – Transformer (Whisper), Conformer (FastConformer) и RNN-T (в частности, реализованной в модели GigaAM-RNNT);
- подготовить вручную транскрипты, предназначенные для использования в качестве эталона;
- сопоставить ручные и автоматические транскрипции для выявления ошибок распознавания;

- классифицировать ошибки с точки зрения их лингвистической и диагностической значимости;
- выявить наиболее эффективный алгоритм распознавания речи для создания последующих транскриптов;
- разработать протокол разметки транскриптов, учитывая дискурсивные и просодические характеристики речи;
- сформировать структуру корпуса и подготовить его к размещению в виде открытого репозитория.

В процессе данного исследования для решения поставленных задач были использованы следующие **методы**:

- метод автоматического распознавания речи (ASR) – технология преобразования звучащей речи в текст с использованием нейросетевых моделей. В рамках настоящего исследования метод применялся для автоматической транскрипции устной речи с помощью трёх моделей, основанных на различных архитектурах, включая Transformer (Whisper), FastConformer и RNN-T (GigaAM).
- метод ручного создания транскриптов – для создания эталонных текстов и последующего сравнения с автоматическими результатами.
- сравнительный анализ – для выявления различий между автоматическими и ручными транскриптами, локализации и классификации ошибок.
- кластеризация – для группировки ошибок по лингвистическим параметрам и выявления закономерностей в их распределении.
- лингвистический анализ – для оценки значимости и влияния различных типов ошибок на интерпретацию устной речи.
- методы корпусной лингвистики – при разработке структуры и принципов аннотирования речевого корпуса.
- описание протокола разметки – как прикладной метод стандартизации аннотирования данных для дальнейшего анализа.

**Теоретико-методологической базой** данного исследования послужили работы отечественных и зарубежных лингвистов – В.П. Захарова, Д.С.

Богданова, О.Ф. Кривновой, Строкин Г. С., А.Я. Подрабинович, а также А. Вецорковской, которые позволили подробно рассмотреть структуру речевого корпуса, его основные компоненты и этапы создания, учитывающие не только технические и лингвистические аспекты, но и пользовательский опыт. Значительный вклад в развитие методов оценки качества транскрипций и анализа особенностей устной речи внесли исследования Т.Ю. Шерстиновой, Р.А. Колобова, Н.В. Михайловского, Сэма О'Коннора Рассела, Дж. Гессингера, А. Красон и Н. Харт. Данные работы фокусируются на применении алгоритмов автоматического распознавания речи (ASR) в корпусных лингвистических исследованиях, выявляют основные трудности ASR-моделей при транскрибировании спонтанной устной речи и анализируют влияние различных факторов на качество получаемых транскриптов. Методологической основой разработки протокола разметки и параметров аннотирования послужили работы В.И. Подлесской и А.А. Кибрика: «Рассказы о сновидениях. Корпусное исследование устного русского дискурса», «Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи», «Проблема сегментации устного дискурса и когнитивная система говорящего». Из этих исследований были заимствованы критерии выделения элементарных дискурсивных единиц (ЭДЕ) в спонтанной речи, а также сформировано системное представление об их типологических особенностях. Данные работы оказались особенно ценными при формировании параметров разметки, адаптированных к целям настоящего корпуса.

**Материалом** исследования послужили 60 аудиозаписей, собранных у 20 респондентов в возрасте от 18 до 67 лет. В ходе эксперимента участники воспроизводили как правдивые, так и ложные высказывания в разных речевых форматах: в виде монологов и в диалогах с интервьюером. Этот материал использовался как основа для построения корпуса, а также для сравнительного анализа эффективности ASR-моделей.

**Теоретическая значимость исследования** заключается в выявлении лингвистически релевантных ошибок автоматического распознавания речи и

анализе того, какие особенности устной речи остаются недоступными для автоматической транскрипции. Работа вносит вклад в развитие теоретических подходов к созданию диагностически ориентированных корпусов речи, а также в методологию лингвистического аннотирования звучащей речи. **Практическая значимость** состоит в создании структурированного корпуса устной русской речи, который может быть использован в лингвистических, диагностических и технологических задачах, включая обучение моделей ASR, разработку речевых интерфейсов и проведение повторных исследований. Также в работе проводится сравнительный анализ ASR-моделей, что дает исследователям представление о возможностях современных архитектур в условиях нерегламентированной устной речи.

Таким образом, данная работа объединяет задачи лингвистического анализа и оценки эффективности технологий автоматического распознавания речи, обеспечивая как научную, так и прикладную ценность полученных результатов.

# Глава 1. Изучение устной речи в контексте корпусной лингвистики

## 1.1. Устная речь как первичная форма языка и объект лингвистического анализа

Вопрос о соотношении устной и письменной речи, а также о приоритете одной формы по отношению к другой остается предметом активных дискуссий в современной лингвистике. Представители структурной традиции, такие как Ф. де Соссюр и Л. Блумфилд, подчёркивали первичность устной формы, рассматривая письмо как вторичную, производную систему репрезентации языка. Сходной точки зрения придерживаются и исследователи, работающие в рамках биологических, когнитивных и культурно-исторических подходов. Так, например, советский психолог Д. Б. Эльконин отмечал, что уже в младенчестве ребенок демонстрирует элементы лепета, воспроизводя интонационные контуры речи окружающих, постепенно переходя к фразовым единицам и полноценным высказываниям. В то время как навыки чтения и письма формируются значительно позже и требуют предварительного освоения звуковой стороны языка, то есть способности «адекватно воссоздавать звуковую форму слова по его буквенному изображению». Историко-культурный аргумент в пользу первичности устной речи представил Уолтер Онг (Ong, 1982), указав, что письменность существует лишь около 6000 лет, в то время как устная речь не менее 50 000 лет. Более того, лишь ограниченное число языков развили полноценные письменные традиции, и во многих сообществах письмо до сих пор либо отсутствует, либо используется крайне ограниченно. Это подчёркивает сравнительно недавний характер письменности как культурной технологии и первичность устной речи.

Устная речь в ряде научных работ трактуется прежде всего как акустический, звуковой сигнал, возникающий в результате сложной артикуляционной и физиологической деятельности человека. В психолингвистических исследованиях восприятия речи она описывается как непрерывный речевой поток, последовательно обрабатываемый слушающим субъектом на уровне звуков, слов, грамматических структур и семантики, что

подчёркивает необходимость точной акустико-фонетической разметки звучащих текстов для последующего анализа (Венцов, 2015). В учебных пособиях по фонетике речь определяется как физическое явление, производимое за счет скоординированной работы речевого аппарата и центральной нервной системы, включая образование воздушной струи, фонации и артикуляции, что подтверждает её звуковую, физиологическую природу (Князев, Пожарицкая, 2011). Нередко подчёркивается, что устная речь передается по звуковому каналу, выступая материальной формой речевого общения, в отличие от визуально фиксированной письменной речи (Васильева, Коньков, 2015). Совокупность этих подходов к рассмотрению и определению речи позволяет утверждать, что устная речь, в своей сущности, является акустическим сигналом, что предопределяет как специальные методы её исследования, так и технические требования к ее фиксации и последующей обработке, а также обуславливает комплексность подходов к её изучению.

До появления средств звукозаписи, особенно до 1940-х годов, исследование устной речи было серьезно затруднено из-за отсутствия технических возможностей для ее фиксации. Ученым приходилось, полагаясь на собственные восприятие и память, записывать услышанное вручную, что превращало устную речь в письменную форму и иногда искажало или фиксировало не все её особенности. Как отмечают Чейф и Таннен (1987), именно из-за невозможности собирать и анализировать аудиоданные устная речь долгое время оставалась вне основного поля лингвистических исследований. В современных же работах авторы снова подчёркивают, что для полноценного анализа устной речи необходимо её обязательное переложение в письменную форму. Несмотря на то, что технологии звукозаписи сегодня позволяют точно фиксировать речевой сигнал, сам по себе аудиофайл остаётся малопригодным для детального лингвистического анализа, ведь устная речь есть непрерывный, протекающий во времени сигнал, который невозможно одновременно воспринимать и интерпретировать без его визуального представления. Как подчёркивает Кибрик (2003), графическая фиксация

речевого сигнала, т.е. его транскрипция, является необходимым условием для анализа, поскольку позволяет «удержать» многомерность звучащей речи и сделать её доступной для изучения. Аналогичную позицию разделяют и авторы проекта Звукового корпуса русского языка, которые настаивают на том, что только совмещение аудиозаписей с транскриптами позволяет изучать спонтанную речь в её естественной форме и применять результаты анализа в прикладных и образовательных целях (Богданова-Бегларян, 2015). Кроме того, в статье Т. В. Бердниковой (2024) подчёркивается, что без объективной фиксации и графического представления звуковой формы речи невозможно достоверно анализировать особенности устного дискурса, в особенности спонтанной речи. Таким образом, даже при наличии технической возможности звукозаписи, перевод устной речи в письменный формат остаётся ключевым этапом ее научного осмысления и изучения.

## 1.2. Корпусы речи: классификация, принципы построения и специфика специализированных устных корпусов

### 1.2.1. Корпусы речи и их разновидности

Существует большое разнообразие лингвистических корпусов, и в научном сообществе принято разделять их на некоторые типы. Одну из таких типологий предлагает Н. В. Козлова, которая выделяет три основные группы корпусов: письменные, устные и смешанные. Письменные корпуса в понимание исследователей – систематизированные собрания текстов, отражающие письменную форму языка в разных жанрах и стилях, предназначенные для лингвистического анализа (Френсис, Кучер, 1979). Они могут включать материалы из художественной и научной литературы, публицистики, блогов, официальных документов и других источников, обеспечивая тем самым жанровую и стилистическую репрезентативность языковых данных. Устные корпуса в свою очередь представляют собой структурированные коллекции речевых фрагментов, снабженные программными средствами доступа к аудио- и текстовым данным (Кривнова, 2006). Такие корпуса включают орфографические транскрипции аудиозаписей или аудиозаписи,

сопровождаемые фонетическими и/или просодическими аннотациями. Как правило, они создаются в строго контролируемых условиях, имеют ограниченный объем и часто используются в социолингвистических и дискурсивных исследованиях, а не исключительно для фонетического анализа. И, наконец, смешанные корпуса объединяют как устные, так и письменные тексты, что позволяет более полно отразить функционирование языка в разных формах. Ярким примером такого ресурса является Национальный корпус русского языка (НКРЯ), включающий как письменные тексты различных жанров, так и аудиозаписи устной речи (публичных выступлений, радиопередачи и др.).

Н.В. Козлова отмечает, что типология корпусов также может основываться на количестве представленных в них языков. Одноязычные корпуса могут охватывать язык в целом либо ограничиваться его специализированными разновидностями, например, профессиональной лексикой медицины или права. В свою очередь, двуязычные и многоязычные корпуса делятся по критерию наличия единых принципов отбора языкового материала на сопоставимые – состоящие из текстов на разных языках, подобранных по сходным критериям, и параллельные – включающие оригинальные тексты и их переводы на другие языки. Еще одним важным критерием является временной охват: синхронные корпуса отражают состояние языка в определённый исторический момент, тогда как диахронные позволяют отслеживать его развитие на протяжении времени, фиксируя языковые изменения в разные эпохи.

Особое значение для исследований имеет и степень аннотированности корпуса. Размеченные (аннотированные) корпуса по мнению В.П. Захарова должны содержать различные уровни разметки: метатекстовую (сведения об источнике, авторе, дате и т.п.), лингвистическую (морфологическую, синтаксическую, семантическую) и экстралингвистическую (просодию, жесты и др.). Неразмеченные корпуса содержат лишь «сырые» тексты без дополнительной аннотации.

Наконец, учитывается и доступность ресурса: корпуса могут быть открытыми (доступными для всех), частично открытыми (с ограничениями по доступу) и закрытыми (коммерческими).

Следует отметить, что представленная Н. В. Козловой типология во многом соотносится с классификациями, разработанными зарубежными исследователями, в частности Дж. Синклером, Дж. Торруэллой и Дж. Листери. Вместе с тем В. П. Захаров в своём пособии по корпусной лингвистике расширяет предложенную типологию, опираясь на дополнительные классификационные признаки. Так, он выделяет особый критерий специфичности, позволяющий различать диалектные, разговорные, терминологические и смешанные корпуса. Кроме того, Захаров предлагает учитывать жанровую принадлежность, динамичность корпуса (фиксированный или обновляемый), а также его функциональное назначение, различая исследовательские и иллюстративные корпуса. Эти уточнения можно рассматривать как более детализированную, углубленную версию типологии, тогда как в большинстве исследований преобладают более обобщенные подходы к классификации корпусов.

Что представляет особый интерес для исследователя, так это возможность конструирования собственного корпуса, адаптированного под конкретные исследовательские задачи. Однако прежде чем перейти к рассмотрению этапов этого процесса, необходимо обозначить ключевые принципы, без соблюдения которых лингвистическое собрание речевых и письменных материалов не может считаться полноценно сформированным корпусом. Речь идет о ряде обязательных характеристик, в совокупности определяющих статус и функциональность корпуса, которые были представлены в работе К.П. Чилингарян. Автору удалось вывести данные критерии благодаря подробному изучению понимания термина «корпус» разными исследователями. Мы же в работе постарались распределить данные критерии по функционально-содержательному признаку на три группы:

1. **Технические и формальные параметры.** В современном мире, где компьютерные технологии распространены довольно широко, корпус должен быть представлен в цифровом формате, то есть размещён на электронном носителе или доступен через сеть Интернет. Кроме того, объём корпуса должен быть достаточным для обеспечения репрезентативности и надежности результатов анализа. При этом важно учитывать специфику исследовательских целей: для широких лингвистических обобщений предпочтительны более масштабные корпуса, тогда как при изучении узких тем или специализированных разновидностей языка вполне допустимы корпуса меньшего объема, что может быть дополнительно обусловлено сложностью сбора данных. Важной характеристикой также выступает открытый характер корпуса, что предполагает его постоянное обновление и актуализацию.
2. **Принципы отбора и состава материала.** Включаемые в корпус материалы должны быть аутентичными, то есть отражать реальные коммуникативные практики языка. Его отбор не может быть случайным: он осуществляется в соответствии с заранее определенными лингвистическими или экстралингвистическими задачами, что и отличает корпус от коллекций текстов/аудио записей. При этом корпус должен быть репрезентативным, то есть текстовая выборка должна адекватно представлять ту или иную разновидность языка – будь то временной, жанровый, социолингвистический или иной аспект.
3. **Аналитическая пригодность.** Важным требованием является наличие лингвистической разметки: текстам (в том числе речевым транскриптам) и их компонентам должны быть прописаны специальные метки, описывающие грамматические, лексические, структурные или экстралингвистические особенности. Кроме того, корпус должен обеспечивать возможности классификации

материалов по определенному признаку, определенному исследователем: жанру, тематике, степени специализированности и другим релевантным параметрам.

### 1.2.2. Устный корпус

Особенности устного корпуса определяются спецификой самой речевой деятельности, отражающей спонтанность, вариативность и многоплановость устной коммуникации. Как подчеркивается в исследовании А. Вецорковской (2025), устная речь в корпусе может представлять собой как заранее подготовленный текст (чтение), так и спонтанные монологи, диалоги или полилоги с типичными для устного общения элементами: паузами, повторами, самокоррекциями и неразборчивыми фрагментами. Особенно важно, что такие корпуса фиксируют просодические и паралингвистические особенности речи, включая темп, интонацию, акцентуацию, а также неартикулированные звуки, сопровождающие речь, например, смех, вздохи или иные звуковые проявления невербального взаимодействия (Кибрик, Майсак, 2021). В отличие от письменных корпусов, устные корпуса требуют обязательного наличия аудиокомпонента и, как правило, включают многоуровневую аннотацию, где звуковой сигнал соотносится с текстовой транскрипцией и метаинформацией о дикторе и ситуации общения (Богданов, Кривнова, Подрабинович, 2024). Кроме того, важной характеристикой устных корпусов является неоднородность условий записи – от студийных до мобильных, что требует специальной фильтрации и контроля качества звука.

### 1.2.3. Конструирование речевого (устного) корпуса

Современные методологии построения речевых корпусов демонстрируют как общую стратегическую направленность, так и вариативность в конкретных решениях, что обусловлено исследовательскими задачами, техническими возможностями и профилем предполагаемых пользователей этих корпусов. Анализируя подходы, предложенные, с одной стороны, Д. С. Богдановым, О. Ф. Кривновой и А. Я. Подрабинович (2024), а с другой – А. Вецорковской (2025), можно выявить как структурное единство

этапов, так и акценты, отражающие разные исследовательские приоритеты. Оба подхода едины в понимании необходимости чёткого целеполагания на начальном этапе: корпус не может быть нейтральным ресурсом – его архитектура выстраивается строго в соответствии с задачами, будь то автоматическое распознавание речи или лингвистический анализ.

На уровне формирования речевой базы оба подхода исходят из принципов репрезентативности и параметризации дикторской базы: пол, возраст, родной язык, социолингвистические признаки дикторов являются обязательными категориями. В то же время, если в российской методике основной акцент делается на профиль диктора и тематическую направленность материала, то в западной концепции подчёркивается техническая и пользовательская сторона процесса – выбор аудиоформатов, организация записи вне студийных условий, контроль качества аудиозаписи и удобство интерфейса для конечного пользователя.

Сходство проявляется также в понимании роли аннотации: обе модели предполагают многоуровневую разметку, начиная от орфографической транскрипции и заканчивая специализированными лингвистическими метками. Под последними подразумеваются, например, просодическая аннотация (фиксация ударений, интонационных контуров, пауз), прагматическая разметка (обозначение речевых актов, таких как просьба, согласие, уточнение), а также стилистическая маркировка, отражающая эмоциональную окраску или регистр речи. Такие метки существенно расширяют аналитический потенциал корпуса, особенно в задачах, связанных с распознаванием интонации, изучением спонтанной речи или автоматической классификацией дискурсивных стратегий. Вецорковска также подчёркивает важность документирования и открытости метаданных, а также предлагает проводить UX-тестирование на этапе разработки инструментов, что редко акцентируется в традиционных моделях.

Подытоживая анализ подходов, можно сказать, что их интеграция позволяет выстроить целостную и гибкую методологию, сочетающую концептуальную строгость академического проектирования с ориентацией на

технологическую реализуемость и прикладную эффективность. В таком синтезе нам видится перспективная траектория развития современных речевых корпусов. Для наглядности представим этапы создания корпуса, учитывающие оба подхода на рисунке 1.

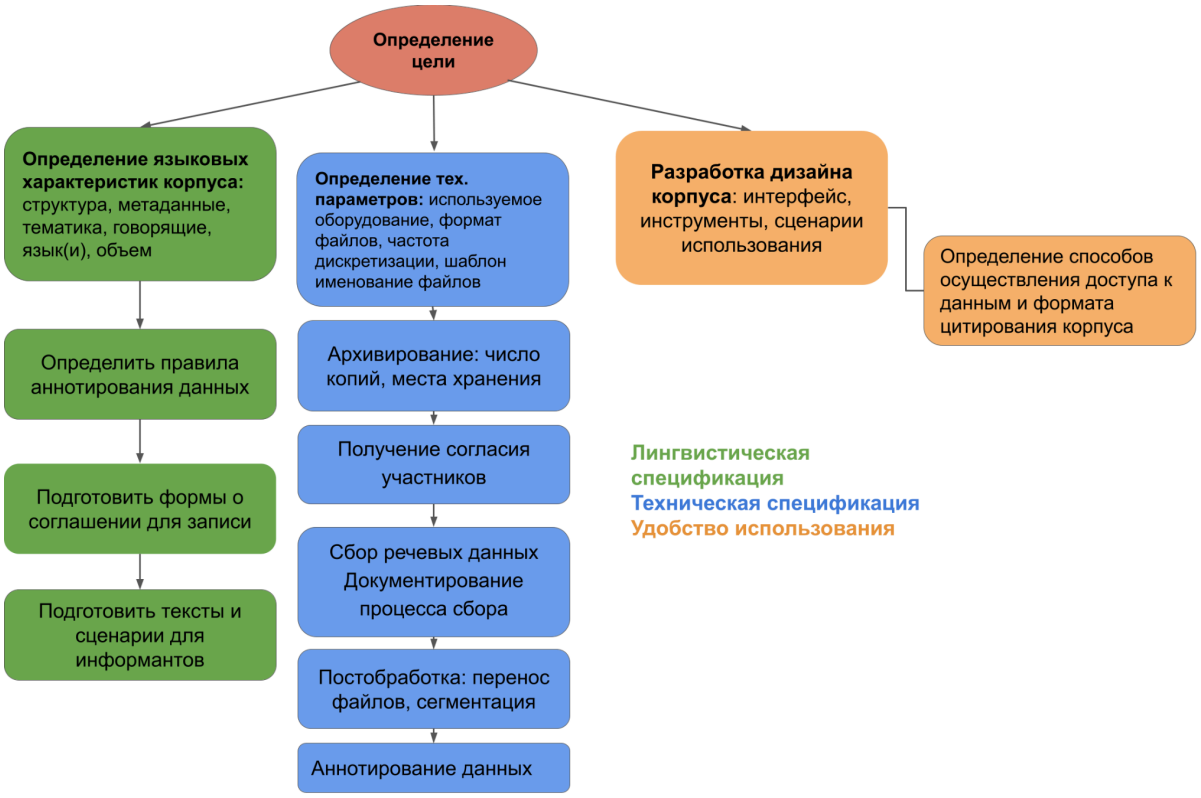


Рисунок 1. Этапы создания речевого корпуса.

### 1.2.4. Специализированный корпус

Специализированным корпусом называют лингвистический корпус, сформированный с целью изучения или анализа определённой области языка, жанра, стиля, тематики или функциональной сферы речи. Такой корпус отличается узкой направленностью и содержит тексты или речевые данные, относящиеся к конкретной предметной области, профессиональной сфере, жанру или типу дискурса (McEnery & Hardie, 2012; Wieczorkowska, 2025). В научной литературе специализированный корпус часто противопоставляется общему (или национальному) корпусу, который стремится охватить язык в целом и все его функциональные стили.

Рассмотрим подробнее, какие задачи могут решать специализированные корпуса на конкретных примерах.

Корпуса, ориентированные на обучение иностранным языкам, представляют собой особый тип специализированных корпусов. Они включают устную и/или письменную речь не-носителей языка и часто аннотированы с указанием языковых ошибок, уровня владения, родного языка и других параметров. Авторы работы *The Need for a Speech Corpus* подчёркивают важность создания специализированных корпусов спонтанной устной речи для обучения иностранцев, осваивающих английский язык. Корпус, представленный в этом исследовании, ориентирован на конкретную учебную задачу – развитие навыков восприятия и понимания живой разговорной речи. Материалы представляют собой записи реальных коммуникативных ситуаций между носителями языка, что обеспечивает обучающимся аутентичный контекст употребления языковых единиц. Это позволяет фиксировать не только формальные особенности устной речи, такие как редукции, ассимиляции, просодические контуры и частотные интонационные паттерны, но и социолингвистические вариации, связанные с региональными акцентами (например, ирландским, американским, австралийским вариантами английского). Также известными корпусами для изучающих английский язык как иностранный являются ICLE (International Corpus of Learner English), LINDSEI (Louvain International Database of Spoken English Interlanguage) и другие.

Корпуса устной речи играют ключевую роль в обучении и тестировании алгоритмов автоматического распознавания речи (ASR) и синтеза речи (TTS), обеспечивая аудиоданные с точными транскрипциями и аннотациями. Один из наиболее популярных англоязычных ресурсов – LibriSpeech (Panayotov и др, 2015), основанный на аудиокнигах и широко применяемый в задачах ASR благодаря разнообразию голосов и высокому качеству записи. Корпус VCTK (Yamagishi и др., 2019) содержит речь от носителей британского английского с различными акцентами и активно используется в задачах TTS и клонирования

голоса. Switchboard (Godfrey и др., 1992) представляет собой богатый ресурс спонтанных диалогов, полезный для моделирования живого общения. Мультиязычные корпуса, такие как Common Voice от Mozilla и FLEURS от Google (Conneau и др., 2023), предоставляют открытые данные на десятках языков, включая русский, с подробной аннотацией, что делает их особенно ценными для обучения моделей в условиях ограниченных ресурсов и для кросслингвистических исследований. На русском языке также доступны крупные специализированные корпуса: Open STT, разработанный SberDevices, включает более 16 000 часов аудиоданных и применяется для обучения ASR-моделей; а также отечественные аналоги зарубежных ресурсов – RuLibriSpeech и Common Voice 12.0. Использование таких специализированных ресурсов позволяет учитывать особенности произношения и диалектных вариаций, что повышает точность ASR-систем и естественность синтезированной речи.

Отдельную категорию составляют специализированные корпуса устной речи, направленные на сохранение языков народов России, изучение контактных разновидностей русского языка, а также на диалектологический и сравнительно-исторический анализ. К числу таких проектов относятся устные корпуса башкирского языка деревни Рахметово и села Баимово, устный корпус абазинского языка, созданные Международной лабораторией языковой конвергенции НИУ ВШЭ. Для исследования контактной разновидности русского языка, формирующейся в специфической социолингвистической среде, создаются такие корпуса, как *Corpus of Russian spoken in Daghestan* (Dobrushina N. et al., 2018) и *Corpus of Russian spoken in Chuvashia* (K.A. Bayda et al., 2018). Значимыми также являются проекты, ориентированные на диалектологическое описание и языковую архивацию в рамках задач сохранения языкового наследия, например, корпус говора села Малинино и корпус опочецких говоров. Аннотированная структура этих ресурсов делает их ценным инструментом для исследований в области морфосинтаксической типологии русских диалектов.

Список таких корпусов далеко не исчерпывающий, но приведённые выше примеры демонстрируют разнообразие задач, которые могут решаться с помощью специализированных корпусов, и подчёркивают их прикладную значимость в различных областях лингвистики и прикладных технологий.

### 1.3. Диагностические корпуса

Корпусы, используемые для диагностических задач, характеризуются тем, что они позволяют выявлять и анализировать особенности речи, которые могут служить индикаторами речевых нарушений, патологий или других лингвистических феноменов, важных для диагностики (Богданова-Бегларян, 2015). Как правило, такие корпуса включают аудиозаписи спонтанной речи, диалогов, заданий на называние или пересказ текста, а также их транскрипции и аннотации, что делает возможным фиксацию как лингвистических, так и паралингвистических признаков и структурных характеристик, имеющих диагностическую ценность.

Большинство подобных корпусов ориентированы на изучение нарушений речи у детей или взрослых с неврологическими и психиатрическими расстройствами. Так, в *Cummings Corpus* представлены записи детей с фонетическими нарушениями, *Torrington Eaton* фокусируется на анализе спонтанной речи и называния предметов, а *CMU Kids* содержит материалы чтения детьми как с нарушениями, так и без. Однако, как подчёркивают М. Шанин, Ю. Зафар, Б. Ахмед (Shahin, Zafar, Ahmed, 2019), высокая вариативность акустических характеристик речи, обусловленная возрастом, диалектом, типом и степенью выраженности нарушений, требует разработки специализированных корпусов с тщательно продуманной схемой аннотирования.

В обзоре речевых корпусов, используемых в клинической лингвистике, авторы подчёркивают, что выбор схемы аннотирования определяется как типом речевых нарушений, так и задачами конкретного исследования (Khudyakova и др, 2022). Так, в корпусе *Toronto Bilingual Aphasia* применяется многоуровневая аннотация, включающая морфосинтаксические и прагматические параметры, с

акцентом на выявление отклонений в структуре высказываний у билингвальных носителей с афазией.

В рамках проекта *Discourse Diversity Database (3D)* была разработана оригинальная схема аннотирования, ориентированная на комплексное описание отклонений речевой деятельности у разных групп пациентов, включая лиц с афазией, деменцией и другими когнитивными нарушениями. Эта схема охватывает несколько уровней анализа: дискурсивные стратегии (например, повторы, переформулирование), структурные особенности (фиксация начала, завершения или нарушения целостности высказываний), а также прагматические и интерактивные элементы (например, отклонение от темы).

Выбор именно такой многоуровневой схемы аннотирования обусловлен необходимостью учитывать разнообразные проявления речевых нарушений, выходящих за рамки фонетических или грамматических ошибок. Так, при диагностике афазии важно фиксировать затруднения в организации дискурса, частоту речевых замещений, персеверации и аграмматизмы, а не ограничиваться только отклонениями в произношении. Таким образом, аннотирование выступает не просто способом структурирования данных, но и полноценным диагностическим инструментом, позволяющим уточнять характер и степень нарушения речевой функции.

### 1.4. Виды разметки речевых корпусов

Как уже отмечалось на протяжении нескольких разделов в корпусной лингвистике аннотация речевых корпусов представляет собой ключевой этап подготовки языкового материала, обеспечивающий его структурирование и многоуровневую интерпретацию. В статье С. Грис, А. Берез (2017) «Linguistic annotation in/for corpus linguistics» довольно подробно описаны все типы разметки для разных речевых корпусов (устного и письменного). Однако автор отмечает, что разметка, обычно ассоциирующаяся с письменным корпусом может быть при желании исследователя переложена и на транскрипты устного материала. Рассмотрим представленные автором типы аннотации ниже.

Первичным шагом для письменных корпусов в этом процессе, как правило, выступает лемматизация – приведение словоформ к их базовой, словарной форме, что позволяет учитывать лексемы независимо от морфологических вариаций и упрощает подсчёт частотности и анализ контекстов. На этом основании строится морфосинтаксическая разметка, в рамках которой каждому токену присваиваются грамматические признаки: часть речи, число, род, время, вид и другие категории. Эти сведения важны не только для уточнения синтаксической функции слова в контексте, но и как база для более сложных уровней аннотации. Следующим этапом служит синтаксическая разметка, направленная на описание структуры предложения с помощью синтаксических деревьев. Она может опираться на модели фразовой (иерархической) грамматики или зависимостной структуры, фиксируя и визуализируя отношения между словами и синтаксические связи. На наиболее сложном уровне осуществляется семантическая аннотация, ориентированная на интерпретацию значений слов и выражений в контексте. Этот тип разметки может включать как автоматическое различение значений многозначных слов (разрешение лексической омонимии), так и маркировку таких феноменов, как метафора, метонимия, семантические роли и др. кто-то там отмечает, что последовательность указанных уровней аннотации позволяет обеспечить комплексную формализацию языковых данных, что, в свою очередь, открывает широкие возможности для проведения как количественных, так и качественных лингвистических исследований, а также интеграции корпусов в прикладные задачи от машинного перевода до лексикографической обработки.

Разметка устных корпусов значительно комплекснее по сравнению с письменными, поскольку требует учёта не только вербализованного содержания, но и множества паралингвистических и невербальных компонентов, специфичных для спонтанной устной коммуникации. Базовой и обязательной формой аннотации является орфографическая транскрипция, обеспечивающая первичную фиксацию речевого материала, однако для более точного анализа часто применяется фонемная или фонетическая аннотация.

Фонемная аннотация позволяет идентифицировать звуки на уровне фонем, тогда как фонетическая учитывает вариативность произношения, аллофонию и другие артикуляционные особенности; последняя, как правило, выполняется вручную и требует значительных затрат времени (Oostdijk и Boves, 2008).

Важным направлением в разметке устных корпусов является просодическая аннотация, охватывающая элементы интонации, акцентуации и ритмической организации высказываний. В этом контексте в международной практике как правило применяются как дискурсивно ориентированные системы (например, Discourse Transcription), фиксирующие интонационные единицы, так и более точные фонологические модели, такие как ToBI, которые позволяют анализировать структуру интонационного контурирования вплоть до уровня слоговой структуры (Jun, 2005).

Существенным компонентом также является аннотация невербального поведения, включая жесты и мимику, особенно в мультимодальных корпусах или при работе с жестовыми языками. В таких проектах используется специализированное программное обеспечение (например, ELAN, iLex) и сложные схемы разметки, включая HamNoSys и дескриптивные системы для фаз и типов жестов (Hanke, 2004; Kipp и др., 2007). Кроме того, существующий тип интерактивная аннотация играет ключевую роль в анализе естественной устной коммуникации, так как позволяет отразить структуру взаимодействия участников: порядок и пересечение реплик, паузы, реплики-сигналы обратной связи, темп речи и такие особенности устной речи, как паузы, запинки, повторы, самокоррекции и другие отклонения от плавного речевого потока. Здесь наибольшее распространение получила система Conversation Analysis (Jefferson, 2004), ориентированная на фиксацию «очередей говорящего» (turn-at-talk) и последовательной организации дискурса. Все эти типы аннотации направлены на то, чтобы максимально полно зафиксировать сложную природу устной речи, отражая её многоуровневую структуру и динамику, что делает их незаменимыми в лингвистическом исследовании спонтанной и диалогической речи.

Возвращаясь к просодической аннотации устной речи, стоит упомянуть обзорное исследование Ю. Д. Абаева (2021), в котором последовательно рассматриваются принципы транскрибирования звучащего текста в корпусной лингвистике – от докорпусных транскрипционных традиций до современных подходов, реализованных в звуковых корпусах в отечественной практике. Особое внимание автор уделяет основным параметрам просодической аннотации, которые являются обязательными при разметке корпусов устной речи. К числу таких параметров автор относит, прежде всего, членение высказывания на синтагмы – структурно-семантические единицы речи (или ЭДЕ), представляющие собой минимальные интонационно завершённые отрезки. Вторым неотъемлемым элементом, по мнению исследователя, выступает маркировка направления движения тона в интонационном центре (ядре), поскольку именно движение тона позволяет судить об иллокутивном типе высказывания, а также о его дискурсивной завершённости или незавершённости. Тем не менее, как подчёркивает Ю. Д. Абаева (2021), в корпусе «Один речевой день» данная аннотация в минимальной версии транскрипции отсутствует, однако при этом указывается иллокутивный тип реплики, что также является значимым для прагматической интерпретации данных. Кроме того, в корпусах, содержащих образцы живой спонтанной речи, обязательным элементом аннотации становятся паузы: как структурные маркеры границ между синтагмами, так и паузы-хезитации. Наряду с этим подлежат разметке и речевые сбои – такие как фальстарты, самокоррекции, повторы, — как важные характеристики спонтанной речевой деятельности. Эти параметры, обозначенные в исследовании, формируют основу для разработки комплексной просодической разметки, что актуально и для задач, решаемых в рамках данного проекта.

### 1.5. Основные способы создания транскрипций устной речи

К первому способу транскрибации устной речи относится ручное создание транскриптов. Этот способ фиксации речи отличается высокой точностью, особенно когда речь идет о передаче сложных акцентов, наличии

нескольких говорящих, специфической терминологии или плохом качестве записи. В дополнение к этому, только ручная транскрипция предоставляет возможность фиксировать нюансы, важные для диагностики и анализа речи: хезитационные паузы, самоперебивы, обрывы фраз и пр. Однако, стоит отметить, что несмотря на свою точность, ручная транскрипция имеет свои недостатки в виде больших затрат временных и человеческих ресурсов (требует участия квалифицированного специалиста или проведения валидации качества в несколько этапов).

Вторым и наиболее современным способом создания транскриптов устной речи стало использование автоматических систем распознавания речи в текст, так называемых Speech-to-text (STT) или Automatic Speech Recognition (ASR) алгоритмов. Их применение для достижения цели создания транскриптов для корпусов устной речи началось совсем недавно, когда качество распознавания вышло на довольно хороший уровень. Этот подход стал популярен из-за своей высокой скорости и масштабируемости, но ему пока что не удалось достичь того же уровня точности, как у ручной транскрипции, так как алгоритмы, в большинстве своем ориентированы не под конкретные исследовательские задачи, им важно зафиксировать как можно больший объем информации, но с потерями важных для конкретной области особенностей звучащей речи. В современной лингвистике и смежных дисциплинах, изучающих устную речь, данный подход становится всё более популярным (Russell и др., 2024; Latif и др., 2023). Как отмечается в исследованиях, современные ASR-алгоритмы демонстрируют высокую эффективность в задачах транскрибации, особенно при работе с четкой дикцией и стандартным произношением, где уровень ошибок (WER) может составлять всего 2–3% (Russell и др., 2024). Однако в случае спонтанной речи, характеризующейся наличием нелексических элементов (например, вокалическим заполнением пауз *ээ, мм*), перекрывающимися репликами собеседников в диалогах и вариативностью произношения, точность транскрибации снижается, а WER достигает от 30 % до 99 % (Sherstinova и др., 2024). Это подчёркивает

необходимость тщательного выбора моделей посредством предварительного тестирования их перформанса перед их использованием для создания лингвистических ресурсов.

Оптимальным на данном этапе вариантом становится гибридная или полуавтоматическая транскрипция, комбинирующая автоматическое распознавание с последующей ручной проверкой и корректировкой текста траснкрипта. Как итог, мы получаем баланс между скоростью и точностью, что позволяет быстро обрабатывать большие объемы данных, сохраняя высокое качество (Bazillon, Esteve, Luzzati, 2008).

### 1.6. Оценка качества ASR моделей

#### 1.6.1. Автоматически рассчитываемые метрики оценки качества моделей

При выборе модели преобразования речи в текст (speech-to-text, STT) важно опираться на объективные метрики, отражающие точность и надежность распознавания. Наиболее широко используемой метрикой является Word Error Rate (WER) – показатель, оценивающий частоту ошибок на уровне слов. WER рассчитывается как отношение суммы замен (S), удалений (D) и вставок (I) к общему числу слов в эталонной транскрипции (N):

$$WER = \frac{I + D + S}{N(\text{Длина эталона})} \times 100\%.$$

Однако данная метрика имеет ограничения, особенно при оценке языков с богатой морфологией, каким и является русский язык. При подсчете WER за ошибку будет засчитано целое слово, даже если оно верно распознано лексически, но на морфологическом уровне произошла замена, например, той же флексии (например, эталонная транскрипция: *плох**ое** утро было сегодня*, ASR-транскрипция: *плох**ой*** (S) *утро было сегодня*). В этом случае слово *плохой* засчитывается как полностью неверное, несмотря на практически полное совпадение основы слова и части флексии. Таким образом, из-за довольно жесткой оценки итоговые показатели метрики для всего траснкрипта могут искажать реальную картину качества распознавания. В таких случаях более информативной оказывается метрика Character Error Rate (CER), оценивающая

ошибки на уровне символов. CER рассчитывается аналогично WER, но с учетом количества символов:

$$CER = \frac{I + D + S}{N(\text{Длина эталона в символах})} \times 100\%.$$

В дополнению к этому в исследовании Д. Джеймс и др. (2005) было выявлено, что оценка качества CER обладает более высокой корреляцией с субъективными оценками качества распознавания текста носителями языка, что подтверждает адекватность использования данной метрики, в задачах, где CER демонстрирует большую устойчивость к морфологическим вариациям и орфографическим особенностям языка, что делает её предпочтительной при оценке качества ASR для любого флективного языка. Кроме того, данная метрика устраняет влияние орфографических вариаций, обеспечивая объективность оценки даже в случаях допустимых вариантов написания. Это становится особенно полезно, когда мы говорим о фиксации произнесения слов в их нелитературной норме при помощи условной орфографической транслитерации (например, *что→чё*, *говорит→грит*). В таблице 1 наглядно представлено сопоставление расчета данных метрик и их итоговые значения.

Таблица 1. Сопоставление расчета метрик WER и CER.

| Транскрипторы | Транскрипция | Ошибки WER | Ошибки CER | WER (%) | CER (%) |
|---|---|---|---|---|---|
| эксперт | учитель не разрешила переписать мне работу | | | | |
| модель | учитель разрешила приписать мне работу | **пропуск** частицы "не" (D=1), **замена** "переписать"→"приписать" (I=1) | **пропуск:** "н", "е", пробел (D=3) **замена:** "е"→"и", "р"→"п" в приставке слова (S=2) | 33 | 11.9 |
| эксперт | я лазила в инстаграме и наткнулась на шоурум | | | | |
| модель | я а лазила в инстаграме наткнулась на шаурму | **вставка** буквы "а" (I=1), **пропуск** союза "и" (D=1), **замена** слова "шоурум"→"шаурму" (S=1) | **вставка:** "а", пробел (I=2) **пропуск:** союза "и", пробел (D=2) **замена:** "о"→"а", "у"→"а" в слове "шоурум"(S=2) | 37.5 | 15.91 |

Помимо таких базовых метрик, как WER (Word Error Rate) и CER (Character Error Rate), для более точной оценки качества работы ASR-систем используются и альтернативные показатели. Одной из таких метрик является

Match Error Rate (MER). В отличие от WER, которая рассчитывает количество ошибок (вставок, удалений и замен) по отношению к числу слов в эталонной транскрипции, MER нормирует ошибки по совокупному количеству слов, участвующих в сопоставлении, то есть учитывает как слова в эталоне, так и слова, распознанные системой. Это делает MER более сбалансированной метрикой в условиях, когда длины эталонной и гипотетической транскрипций сильно различаются (Morris, Maier, Green, 2004.). MER рассчитывается по следующей формуле:

$MER = \frac{S+D+I}{S+D+I+H} = 1 - \frac{H}{S+D+I+H}$, где $H$ – количество правильно распознанных слов, $S$ – количество замен, $D$ – количество удалений, $I$ – количество вставок.

Следующим шагом к более точному представлению производительности ASR-систем являются метрики, позволяющие оценивать не только количественные ошибки, но и то, насколько успешно система сохраняет содержательную информацию высказывания. В этом контексте особое значение приобретают показатели Word Information Preserved (WIP) и Word Information Lost (WIL), предложенные Э. Моррисом и его коллегами (Morris, Maier, Green, 2004). В отличие от MER, которая фокусируется на подсчёте всех несовпадений на уровне слов, WIP измеряет долю лексической информации, точно воспроизведённой системой. То есть она показывает, какая часть исходных слов была корректно распознана, и тем самым может служить индикатором сохранения смысла высказывания. Метрика WIL, в свою очередь, отражает объём информации, утерянной в результате распознавания, и вычисляется как дополнение к WIP (WIL = 1 – WIP). С точки зрения лингвистического анализа, эти показатели особенно полезны при оценке качества транскрипции в задачах, где критически важна точная передача значимых слов, например, при тематическом анализе, дискурсивной сегментации или автоматическом извлечении информации. Кроме того, в отличие от базовых метрик, WIP и WIL чувствительны именно к сохранению/утрате лексических единиц, а не только к

числу технических ошибок, что делает их особенно релевантными при анализе текстов, насыщенных содержательной лексикой. Таким образом, применение этих метрик позволяет оценить качество ASR не только формально, но и с точки зрения её способности сохранять семантическую структуру высказывания.

### 1.6.2. Лингвистически ориентированные подходы к классификации ошибок систем автоматического распознавания речи

Помимо существующих метрик качества, с помощью которых мы можем определить, насколько хорошо модель справляется с распознаванием аудиоматериала, исследователи нередко прибегают к изобретению собственных классификаторов, в которых они стараются рассмотреть ошибки, допущенные ASR-моделями более подробно. Так, например, в работе Зафар и др. (2004) была предложена система, объединяющая девять категорий ошибок, возникающих при транскрибировании медицинских записей как человеком, так и машинным алгоритмом. В данной работе предлагается рассмотреть актуальные для нас 6 категорий ошибок, отмеченных конкретно у ASR-модели, использованной в упомянутом исследовании.

Наиболее распространёнными оказались ошибки произношения (ориг. annunciation errors), связанные с нечёткой артикуляцией, фоновым шумом или особенностями речи диктора. В условиях быстрой или прерывистой речи возникали добавление лишних слов (ориг. added words) или их пропуски (ориг. deleted words), чаще всего такими артефактами становились служебные слова: союзы, предлоги и артикли. Примечательно, что именно данные виды ошибок легли в основу будущей метрики WER. Значительную долю заняли «словарные ошибки» (ориг. dictionary errors), возникающие в том случае, когда термин отсутствовал в словаре языковой модели и заменялся на фонетически близкую последовательность слов, зачастую не имеющую смысловой связи с исходным термином (например, *benazepril* → *nasal pearl* в медицинском дискурсе). Грамматические искажения, такие как некорректное определение временных форм, неверное распознавание части речи слова, посредством замены одного суффикса на другой или опущением флексий классифицировались как ошибки

распознавания морфемного состава слова (ориг. suffix errors) и часто были связаны с неполным проговариванием его окончания. Наконец, омонимическими заменами (ориг. homonym errors), например, *male → mail*, считались фонетические совпадения слов, которые присутствовали в словаре языковой модели, что отличает данный тип ошибок от ошибок словаря, но по каким-то причинам не были распознаны верно.

Впоследствии данная типология была расширена рядом эмпирических исследований, каждое из которых привнесло важные акценты в понимание природы и факторов, влияющих на распознавание речи. Так, Гладфелтер и Сорджел (Gladfelter, Soergel, 2005), анализируя корпус устных исторических интервью, подтвердили наличие трёх основных типов ошибок (вставок, замен и пропусков) и дополнили их отдельной категорией, связанной с неверным определением границ слов. Они обратили внимание на влияние таких просодических и паралингвистических характеристик, как акцент, темп и эмоциональная окраска высказывания, подчёркивая, что именно эти параметры значительно усложняют задачу автоматического распознавания спонтанной речи. Эту линию продолжили Дж. Чое и М. Чан (Choe и др., 2022; Chan и др., 2022), сфокусировавшись на диалектной вариативности внутри одного языка. Их исследования показали, что отклонения от нормативного произношения, связанные с региональными особенностями говорящих, могут снижать точность ASR на 8–11%. При этом особое внимание уделялось фонетической природе ошибок: у носителей китайского языка, для которых английский являлся вторым языком, были характерны замены звонких согласных на глухие и наоборот, связанные с отсутствием в их родном языке противопоставления по звонкости/глухости, например, *bag → back*. Для носители вьетнамского и тайского языков модель демонстрировала сложности распознавания передних гласных (например, *beat [i] → bit [ɪ]*), что объяснялось отсутствием в родном языке носителей оппозиции гласных по напряжённости. Таким образом, исследования показали, что качество распознавания напрямую зависит от типологических особенностей родного языка говорящего и адаптированности

модели распознавания речи. В то же время работа Адда-Декер и Ламэль (Adda-Decker, Lamel, 2005) внесла в исследование ASR-ошибок ещё один важный аспект – гендерный. Авторы выявили, что женская речь, как правило, распознаётся точнее, чем мужская. Как выяснили исследователи, причины кроются в различиях на трёх уровнях: фонетическом, лексическом и дискурсивном. На фонетическом уровне мужская речь чаще характеризуется редуцированной артикуляцией, в частности – сокращённой длительностью согласных (/t/, /d/) и гласных (/ɪ/, /ʊ/, /ʌ/) в английском языке, что повышает вероятность пропусков реально произнесённых звуков. На лексическом уровне мужчины чаще использовали разговорные формы, не всегда представленные в словаре модели (например, «yeah» вместо «yes»), тогда как женщины предпочитали более нормативные употребления. Наконец, на дискурсивном уровне мужская речь содержала до 50% больше заполняющих пауз и повторов, что увеличивало количество ошибок вставки. Этот трёхуровневый подход позволил связать особенности ASR-ошибок с социальными и языковыми переменными, ранее оставшимися вне поля зрения большинства классификаций.

В работе Leveraging Lexical and Grammatical Errors было предложено типологическое деление ошибок на три категории: лексические, грамматические и частеречные. Лексические ошибки затрагивали в основном знаменательные части речи (существительные, глаголы, прилагательные) и оказывали значительное влияние на смысловую интерпретацию текста. Грамматические ошибки чаще касались функциональных слов (артиклей, местоимений, предлогов), а частеречные были связаны с подменой одной части речи другой, зачастую обусловленной неполным анализом морфологических признаков. Такая классификация была разработана с учётом многоязычной природы современных ASR-систем, для которых типологические различия между языками (например, агглютинативный или флективный строй) напрямую влияют на характер и распределение ошибок.

Особый интерес для настоящего исследования представляют работы, посвящённые оценке качества транскрипций русскоязычного аудиоматериала. Так, в исследовании Д.И. Мамаева и Е.И. Риехакайнен (2023) была предложена классификация ошибок автоматического распознавания речи, включающая два типа: критичные и некритичные. Критичными признавались ошибки, способные искажать семантику высказывания и, как следствие, нарушать его восприятие в транскрибированном виде. К данной категории относились, в частности, ошибки в распознавании имён собственных, числительных, а также полная замена фонетической и морфологической формы глагола. Кроме того, критичными признавались случаи опущения придаточных или целых простых предложений, существенно влияющих на интерпретацию высказывания. Результаты данного анализа коррелируют с наблюдениями Т. Шерстиновой (2023), которая в собственном исследовании отметила, что автоматические алгоритмы испытывают затруднения при распознавании редких топонимов и имён собственных. При этом распространённые антропонимы, такие как *Иванов* или *Сидоров*, как правило, распознаются корректно. Также в её работе указывается, что модели часто допускают ошибки в определении падежных форм и глагольных окончаний, что особенно важно в морфологически богатом русском языке. В противоположность этому, некритичные (или коммуникативно незначимые) ошибки, по мнению Мамаева и Риехакайнен, не оказывают существенного влияния на восприятие и интерпретацию транскрипта. Среди них – вариативность форм частиц и союзов (например, *чтобы* → *чтоб*), чередование кратких и полных форм прилагательных, замены между повелительным и изъявительным наклонением, изменения рода и числа существительных. К этой же категории были отнесены случаи опущения заполнителей пауз (*э, а*), повторов, а также нарушения синтаксического согласования, компенсируемые в контексте. Эти наблюдения находят подтверждение и в работе Т. Шерстиновой (2023), где подчёркивается, что подобные расхождения, как правило, не препятствуют пониманию и интерпретации текста.

Таким образом, анализ ошибок, совершаемых системами автоматического распознавания речи, не может ограничиваться лишь количественными метриками вроде WER, CER, MER и др.. Он требует многоуровневого лингвистического подхода, включающего рассмотрение фонетических, морфологических, лексических и дискурсивных характеристик. Изучение ошибок с этой точки зрения позволяет не только глубже понять природу ошибок в работе ASR-моделей, но и предложить пути для их доработки – от пополнения лексических ресурсов до настройки языковых моделей под конкретные исследовательские задачи. В рамках настоящей работы также предпринимается попытка проанализировать ошибки, допускаемые выбранными ASR-системами, с опорой на классификационные подходы, разработанные в упомянутых исследованиях, с целью выявления типичных паттернов ошибок и определения их критичности для дальнейшего использования той или иной модели для транскрибации устной речи.

# Глава 2. Трансформация аудиоданных в диагностически значимый корпус: автоматическая транскрипция, анализ ошибок, аннотирование данных и структурирование репозитория

## 2.1. Описание материала

В корпус устной речи вошли 60 аудиозаписей, фиксирующие речь людей возрасте от 18 до 67 лет. Всего к записи для исследования было привлечено 20 человек, каждый из которых был записан в момент рассказа историй. Набор данных включает записи респондентов, предоставляющих как правдивую, так и ложную информацию в различных форматах, включая монологическую речь

(когда участники рассказывали свои собственные истории) и диалогическую речь (когда участники отвечали на вопросы интервьюера). Эксперимент проводился в три отдельных этапа.

На первом этапе респонденту было необходимо составить рассказ о недавно произошедшем с ним событии (походе на концерт или спектакль) по предоставленному плану монолога. Респондент предварительно был проинформирован, что ему придётся говорить неправду. В ходе рассказа интервьюер делал для себя пометки и затем в конце задавал уточняющие вопросы, чтобы выяснить детали рассказанной истории. Конкретных временных ограничений не выставлялось. На втором этапе респонденты должны были вспомнить и рассказать о наиболее ярком событии, которое произошло с ними на самом деле, с опорой на план монолога. На этот раз рассказ должен был содержать только правдивые высказывания. Как и на первом этапе, интервьюер в конце рассказа респондента задавал уточняющие вопросы. На третьем этапе респонденту предлагалось самому выбрать, рассказывать о выдуманном или реально произошедшим с ним событии. За его рассказом также следовали вопросы. Итак, общая продолжительность собранного материала составила 2 часа 17 минут 51 секунду.

### 2.2. Создание транскриптов для корпуса

Первый этап данного исследования заключался в оценке производительности трех ASR-моделей на конкретном наборе речевых данных, собранных в определенных акустических условиях с использованием конкретного оборудования (мобильного устройства с петличным микрофоном). Такой подход обусловлен тем, что, как отмечают Geirhos и др. (2020), высокая оценка качества модели на материале одного корпуса не гарантирует сопоставимых результатов при работе с иными наборами данных. Это подчёркивает необходимость локального тестирования систем автоматического распознавания речи, особенно в контексте задач, связанных с транскрибированием специализированных речевых материалов.

### 2.2.1. Предобработка данных

Все аудиозаписи содержали в себе монологическую и диалогическую речь. В исследованиях неоднократно подтверждалось (Sherstinova, 2024), что наличие нескольких спикеров (которое в свою очередь нередко приводит к наложению реплик говорящих друг на друга) довольно сильно влияет на качество распознавания речи, поэтому во избежание аномально низких показателей метрик качества для тестируемых ASR-моделей для всего текста был произведен их подсчет отдельно для первой части записи, где респондент рассказывает историю без вопросов и комментариев интервьюера, и для второй части, где для уточнения деталей рассказанной респондентом истории, проводящий эксперимент интервьюер задает дополнительные вопросы. Также в отдельную категорию были выделены аудиофайлы, с речью, записанной в зашумленной обстановке. В итоге получилось выделить три выборки: монологическая речь (12 аудиофайлов), диалогическая речь (12 аудиофайлов), записи в зашумленной обстановке (6 аудиофайлов) из общего набора данных. Общая продолжительность тестового набора аудифоайлов составила 1 час 13 минуты 29 секунд.

Перед проведением подсчета метрик тексты, полученные из моделей, и сами референсы были обработаны в соответствии с общепринятыми правилами нормализации текстов (Microsoft Corporation, 2025):

1. Приведены к нижнему регистру;
2. Удалены все знаки препинания;
3. Числа приведены к единому формату (в нашем случае: прописан словами);
4. Из оригинальных транскриптов были удалены вокалические заполнители пауз (например, *аа, ээ, эм, ам* и др.), поскольку современные алгоритмы автоматического распознавания речи, находящиеся в открытом доступе, пока не обеспечивают их точную идентификацию. Это связано, по-видимому, с тем, что данные алгоритмы преимущественно ориентированы на распознавание лексически значимых единиц и не

предназначены для обработки подобных просодических феноменов. Данный вывод подтверждается результатами исследования, в котором установлено, что наиболее часто игнорируемыми элементами при автоматической транскрипции являются паузы, заполненные вокалическим содержанием (Russell и др., 2024).

### 2.2.2. Отбор модели для автоматической транскрибации речи

В настоящем исследовании для получения автоматических транскрипций использовались три модели ASR, размещённые в открытом доступе. Для проведения первичного тестирования была выбрана библиотека **onnx-asr**, специально разработанная для быстрого запуска и сравнения моделей без необходимости установки дополнительных зависимостей и устранения конфликтов между версиями библиотек в среде Python. С её помощью были загружены и протестированы следующие модели: *nemo-fastconformer-ru-rnnt*, *gigaam-v2-rnnt* и *whisper-large-v3-turbo*.

Выбор этих моделей был обусловлен результатами предварительного тестирования, проведённого авторами библиотеки **onnx-asr** на тестовой части датасета Russian LibriSpeech, где модели показали высокие показатели точности распознавания среди своих конкурентов. В частности, по метрикам CER (Character Error Rate) и WER (Word Error Rate) были получены следующие значения:

- *whisper-large-v3-turbo*: CER – 2.63%, WER – 10.08%
- *nemo-fastconformer-ru-rnnt*: CER – 2.63%, WER – 11.62%
- *gigaam-v2-rnnt*: CER – 1.10%, WER – 5.22%

#### 2.2.2.1. Подсчет автоматических метрик

В целях оценки качества каждой модели, 30% аудиофайлов корпуса (охватывающих 53,6% общей длительности материала) упомянутые ранее были вручную транскрибированы для создания эталонной версии, которая использовалась в качестве основы для сравнения с результатами, сгенерированными каждой из моделей автоматического распознавания речи

(ASR). Результаты по итогам тестирования для каждой модели представлены в таблице 1.

Таблица 2. Сравнение метрик качества автоматического распознавания речи (WER, CER, MER, WIL, WIP) для трёх моделей ASR в условиях без зашумлений.

| метрика (%) / ASR - модель | *nemo-fastconformer-ru-rnnt* | *openai/whisper-large-v3-turbo* | *gigaam-v2-rnnt* |
|---|---|---|---|
| WER (Word Error Rate) | 11.09 | 16.46 | **6.18** |
| CER (Character Error Rat) | 4.4 | 9.94 | **2.62** |
| MER (Match Error Rate) | 10.94 | 16.76 | **6.10** |
| WIL (Word Information Lost) | 17.42 | 20.84 | **9.54** |
| WIP (Word Information Preserved) | 82.58 | 79.16 | **90.46** |

Проведённый сравнительный анализ трёх ASR-моделей (*nemo-fastconformer-ru-rnnt*, *gigaam-v2-rnnt* и *whisper-large-v3-turbo*) выявил значимые различия в их способности обеспечивать точную дословную транскрипцию устной речи. Наивысшую эффективность продемонстрировала модель *gigaam-v2-rnnt*, показавшая минимальные значения WER (6.18%) и CER (2.62%), что указывает на высокую точность распознавания как на уровне слов, так и символов. Кроме того, низкие значения WIL (9.54%) и высокие значения WIP (90.46%) свидетельствуют о способности данной модели минимизировать потери смысловой информации, что особенно важно в контексте задач дословной фиксации речи. Модель *nemo-fastconformer-ru-rnnt* продемонстрировала средние показатели, уступая *gigaam-v2-rnnt* по всем ключевым метрикам: WER составил 11.09%, CER – 6.48%, WIL – 14.64%, а WIP – 85.36%. Тем не менее, её результаты остаются стабильными и приемлемыми для задач транскрибирования звучащей речи в связи с показателями автоматической оценки качества. В то же время, модель *whisper-large-v3-turbo* показала наихудшие результаты по всем метрикам качества. Это, вероятно, связано с её мультиязычной архитектурой, менее приспособленной к морфологической и фонетической специфике русской речи.

Теперь рассмотрим результаты, которые получились после транскрибирования диалогической части в таблице 3.

Таблица 3. Сравнение метрик качества автоматического распознавания речи (WER, CER, MER, WIL, WIP) для трёх моделей ASR в условиях без зашумлений, диалогическая речь.

| метрика (%) / ASR - модель | *nemo-fastconformer-ru-rnnt* | *openai/whisper-large-v3-turbo* | *gigaam-v2-rnnt* |
|---|---|---|---|
| WER (Word Error Rate) | 42.97 | 29.49 | **25.70** |
| CER (Character Error Rat) | 24.73 | **21.16** | 23.54 |
| MER (Match Error Rate) | 42.47 | 28.36 | **25.43** |
| WIL (Word Information Lost) | 48.42 | 34.99 | **29.11** |
| WIP (Word Information Preserved) | 51.58 | 65.01 | **70.89** |

Результаты оценки качества работы ASR-моделей в условиях диалогического общения (ответы на вопросы интервьюера) подтверждают, что их эффективность напрямую зависит от акустических особенностей аудиоданных. Наличие двух спикеров, наложение реплик и сниженная разборчивость речи интервьюера (обусловленная отсутствием второго петличного микрофона) приводят к выраженному ухудшению ключевых метрик. Например, модель *gigaam-v2-rnnt*, всё ещё демонстрирующая наилучшие показатели в распознавании диалогической речи, однако теряет устойчивость: её WER и CER возрастают до 25.70% и 23.54% соответственно, что указывает на скачок метрик в среднем на 20% по сравнению с «идеальными» условиями в виде записи монологической речи в незашумленной обстановке. В отличие от *gigaam-v2-rnnt*, модели *whisper-large-v3-turbo* удается сохранить относительную стабильность – вариативность её WER и CER не превышает 13%, вероятно, благодаря обучению на большем количестве разнородных данных. Наибольшая деградация наблюдается у *nemo-fastconformer-ru-rnnt*, для которой наблюдается рост WER до 42.97% и WIL до 48.42%, что может свидетельствовать о недостаточной адаптивности алгоритма к диалогической и тихой речи, представленной конкретно в нашем наборе данных.

Помимо этого в исследовании системы были протестированы и для отдельной категории данных: зашумленных, которые были записаны в торговом центре, где естественно присутствует посторонний шум. Метрики качества алгоритмов в данном случае отражены в таблице 4.

Таблица 4. Сравнение метрик качества автоматического распознавания речи (WER, CER, MER, WIL, WIP) для трёх моделей ASR в условиях с зашумлениями.

| метрика (%) / ASR - модель | *nemo-fastconformer-ru-rnnt* | *openai/whisper-large-v3-turbo* | *gigaam-v2-rnnt* |
|---|---|---|---|
| WER (Word Error Rate) | 22.48 | 24.77 | **13.18** |
| CER (Character Error Rat) | 9.11 | 16.63 | **6.28** |
| MER (Match Error Rate) | 22.25 | 24.48 | **12.94** |
| WIL (Word Information Lost) | 34.47 | 32.11 | **19.53** |
| WIP (Word Information Preserved) | 65.53 | 67.89 | **80.47** |

Здесь мы можем наблюдать то, что значение метрик по сравнению с идеальными условиями не так сильно поменялось, как в случае фиксации диалогической речи. По прежнему наиболее качественной в своих показателях остается *gigaam-v2-rnnt,* в то время как модели *whisper-large-v3-turbo* и *nemo-fastconformer-ru-rnnt* показывают результаты хуже, но близкие друг к другу, что в очередной раз выделяет *gigaam-v2-rnnt* как наиболее стабильно работающую систему.

**2.2.2.2. Лингвистический анализ ошибок ASR-моделей**

В рамках настоящего исследования была поставлена задача детального изучения ошибок, допускаемых алгоритмами автоматического распознавания речи (ASR). Автоматическая транскрипция звучащей речи сопряжена с рядом трудностей, связанных как с акустическими, так и с лингвистическими особенностями речевого материала – от фонетических редукций до синтаксической фрагментарности и нелинейности высказываний. Однако не все ошибки распознавания равнозначны: одни носят формальный характер и несущественно искажают структуру высказывания, тогда как другие могут влиять на содержание и приводить к потере значимых лексических или прагматических единиц (Мамаев, 2023). В этой связи лингвистический анализ

ошибок становится необходимым этапом, позволяющим выявить типологию и частотность искажений, а также определить, насколько конкретная модель способна обеспечивать адекватное представление речевого потока в транскрибированном виде. Это, в свою очередь, служит обоснованием выбора той или иной модели для дальнейшего использования в задачах, где критически важна точность и полнота фиксации устной речи.

**2.2.2.2.1. Ошибки замен**

Прежде всего рассмотрим категорию ошибок, связанных с заменой лексем, т.к. они составляют значимую часть нарушений в автоматической транскрипции. На основании их анализа, а также с опорой на теоретические положения, изученные на предыдущих этапах исследования, была разработана классификация, позволяющая разделить ошибки на две основные группы:

- **Ошибки, не влияющие на смысл**. В эту категорию входят формальные морфологические отклонения, при которых сохраняется лексическая идентичность: например, замена грамматической формы слова (изменение рода, числа или падежа), не приводящая к появлению новой лексемы.
- **Ошибки, искажающие смысл**. К данным ошибкам относятся случаи, когда модель подставляет вместо исходного слова:
    - полностью другую лексему, не связанную с контекстом, но схожую по звуковому облику;
    - несуществующий звуковой аналог.

Анализ ошибок, зафиксированных в транскрипциях модели *whisper-large-v3-turbo*, выявил преобладание смыслоизменяющих замен. Они составляют около 60% от общего числа. Эти ошибки, как уже было сказано, можно условно разделить на две категории в зависимости от их природы и степени воздействия на интерпретацию высказывания. Наиболее объемную группу составляют ошибки, вызванные фонетическим сходством слов, при котором модель подбирает лексемы с аналогичным звучанием, не принимая во внимание семантический контекст. Наглядные примеры приведены в таблице 5.

Таблица 5. Примеры ошибок модели *whisper-large-v3-turbo*, при которых вместо ожидаемой лексемы подставляется полностью иное слово, не соответствующее контексту, но совпадающее частично по фонетическому профилю.

| *whisper-large-v3-turbo* | |
|---|---|
| эталонный траснкрипт | я решила съездить и **отвезти** его приезжаю |
| гипотеза | я решила съездить и **отвести** его приезжа |
| эталонный траснкрипт | **мне** дают бланк я заполняю этот **бланк** |
| гипотеза | **не** дают бланк я заполняю этот **банк** |
| эталонный траснкрипт | все на **свете** липнет |
| гипотеза | все на **сети** липнет |
| эталонный траснкрипт | по моему **вот да** после двенадцати |
| гипотеза | по моему **вода** после двенадцати |
| эталонный траснкрипт | пять пар и **репетиция** |
| гипотеза | пять пар и **петиция** |
| эталонный траснкрипт | мы записываем как будто **на допросе** |
| гипотеза | мы записываем как будто **надо просить** |
| эталонный траснкрипт | вот она **отделала** липкой лентой |
| гипотеза | вот она **делала** липкой лентой |
| эталонный траснкрипт | что является **гарантом** их возвращения |
| гипотеза | что является **грантом** их возвращения |

В пределах этой категории различимы два уровня искажения:

- Минимальные фонетические изменения, сохраняющие исходную часть речи и корневую морфему слова (например, *делала → отделала*). Такие замены сопровождаются подстановкой или удалением приставки, которые в данной категории образуют совершенно новые лексемы.
- Существенные отклонения, сопровождающиеся перестановкой или выпадением звуков, что приводит к замене на лексему иной части речи или функции в высказывании (*вот да → вода*, *на допросе → надо просить*).

Подобные ошибки свидетельствуют о склонности модели опираться преимущественно на акустическую форму высказывания, при этом игнорируя

его семантическую целостность. Ввиду ограниченного объёма выборки на данном этапе представляется затруднительным определить достоверное количественное соотношение указанных подтипов искажений. Тем не менее их фиксация будет полезна для последующей классификации и валидации ошибок в рамках дальнейших исследований.

Случаи генерации моделью невалидных слов, отсутствующих в лексиконе русского языка, представлены в таблице 6.

Таблица 6. Примеры ошибок, при которых модель *whisper-large-v3-turbo* подставляет вместо исходного слова звуковую последовательность, фонетически похожую на оригинал, но не являющуюся существующим словом.

| *whisper-large-v3-turbo* | |
|---|---|
| эталонный траснкрипт | не было тол толкотни **толкучки** никто на пятки не наступал |
| гипотеза | не было толкотни **толпучки** никто на пятки не наступал |
| эталонный траснкрипт | обратная сторона **медали** этой работы |
| гипотеза | обратная сторона **бенда ли** этой работы |
| эталонный траснкрипт | концерт группы **кровь** и слёзы |
| гипотеза | концерт группы **крофе** и слезы |

Хотя количественно такие ошибки встречаются реже, чем замены на существующие слова (на 31%), они имеют более деструктивный характер, поскольку нарушают саму лексическую валидность результата.

Для модели *nemo-fastconformer-ru-rnnt* зафиксировано приблизительное равновесие между смыслоизменяющими и формальными ошибками: последние составляют 51% от общего числа, первые – 49% соответственно. Формальные ошибки для данной модели преимущественно затрагивают морфологические характеристики слов: род, число, падеж у существительных и прилагательных Результаты представлены в таблице 7.

Таблица 7. Ошибки модели *nemo-fastconformer-ru-rnnt,* не влияющие на семантику.

| *nemo-fastconformer-ru-rnnt* | |
|---|---|
| эталонный траснкрипт | оно как как **школьное** платье **было** |
| гипотеза | он как **школьная** платье **была** |
| эталонный траснкрипт | никто на **пятки** не наступал |
| гипотеза | никто на **пятке** не наступал |
| эталонный траснкрипт | дорогая но достаточно можно сказать **подходящая** |
| гипотеза | дорогая но достаточно можно сказать **подходящее** |

В отличие от предыдущих систем, *gigaam-v2-rnnt* демонстрирует наибольшую устойчивость к смыслоизменяющим ошибкам. Доля ошибок, не влияющих на смысл, составляет 67%, а смысловых – лишь 33%. Примеры отражены в таблице 8.

Таблица 8. Ошибки модели *gigaam-v2-rnnt,* не влияющие на семантику.

| *gigaam-v2-rnnt* | |
|---|---|
| эталонный транскрипт | отвезти другой моей **подруге** |
| гипотеза | отвезти другой моей подруги |
| эталонный транскрипт | покатушки на катке имени парка **терешковой** |
| гипотеза | покатушки на катке имени парка **терешкова** |
| эталонный транскрипт | было **кучу** народу я значит стою |
| гипотеза | было **куча** народу я значит стою |
| эталонный транскрипт | понравились эмоции от этого концерта **времяпровождение** |
| гипотеза | понравились эмоции от этого концерта **времяпровождения** |

Кроме того, количество слов, замененных на несуществующие лексемы сведено к минимуму, их только 13% от всех смыслоизменяющих замен, то есть это 4,3% от общего числа ошибок (например, *шоурум→шауру, медали→мендалия*). Такое распределение, вероятно, связано с тем, что языковая модель была обучена конкретно на материалах на русском языке, что позволило ей точнее распознавать грамматические структуры и отсеивать маловероятные или нерелевантные слова.

Анализ транскрипций аудиофайлов, записанных в зашумлённых условиях, показал, что распределение ошибок между категориями (влияющих

на смысл и не влияющих) сохраняет тенденции, зафиксированные при обработке аудио, записанных в акустически благоприятной среде (без фоновых шумов). В частности, для модели *whisper-large-v3-turbo* доли ошибок составили 52% (искажающих смысл) против 48% (не влияющих на смысл); для *nemo-fastconformer-ru-rnnt* – 56% и 44% соответственно; а для *gigaam-v2-rnnt* наблюдается обратное соотношение – 40% к 60%. Эти данные свидетельствуют о стабильности поведения моделей даже в условиях акустического шума и подтверждают их устойчивость распределения ошибок к варьированию внешних характеристик записи. Интересно то, что в данном наборе появились имена собственные, с которыми часто не справляются модели распознавания речи, особенно, если эти лексемы не являются распространенными в языке (Sherstinova, 2024). Результаты представлены сразу для трех моделей в таблице 9. Можно предположить, что для отдельных моделей в каждом случае имя собственное является ошибкой словаря, как упоминалось в исследовании А. Зафар (Zafar A. и др., 2004). Однако мы снова наблюдаем даже на таком ограниченном наборе, как модель *gigaam-v2-rnnt* справляется с двумя из трёх случаев распознавания имен собственных.

Таблица 9. Сопоставление результатов распознавания имён собсвенных тремя моделями (*whisper-large-v3-turbo, nemo-fastconformer-ru-rnnt, gigaam-v2-rnnt*).

| | |
|---|---|
| эталонный тарснкрипт | вот видео про ворону которая **карлушенька** |
| гипотеза (nemo-fastconformer-ru-rnnt) | знаешь тут видео про ворону которые **карлуженка** |
| гипотеза (openai/whisper-large-v3-turbo) | знаешь вот видео про ворону который **карла ушенко** |
| гипотеза (gigaam-v2-rnnt) | знаешь вот видео про ворону которая **карлушенька** |
| эталонный тарснкрипт | вообще без выходного пособия пойду в свободно в цирк **дю солей** |
| гипотеза (*nemo-fastconformer-ru-rnnt*) | вообще просто без выходного пособия пойду с тобой в цирк **досулей** |
| гипотеза (*openai/whisper-large-v3-turbo*) | вообще просто без выходного пособия пойду свободно в цирк **для сулей** |
| гипотеза (*gigaam-v2-rnnt*) | вообще просто без выходного пособия пойду свободный в цирк **для сулей** |
| эталонный тарснкрипт | не поверишь к кому к юрию **куклачеву** |
| гипотеза (*nemo-fastconformer-ru-rnnt*) | не поверишь кому юрия **пухлачева** |
| гипотеза (*openai/whisper-large-v3-turbo*) | не поверишь кому юрию **куклачеву** |
| гипотеза (*gigaam-v2-rnnt*) | не поверишь кому юрию **куклачеву** |

Таким образом, по результатам анализа замен можно заключить, что модель *gigaam-v2-rnnt* демонстрирует наименьшую степень искажения смысла

высказываний и в наибольшей степени сохраняет их семантическую полноту по сравнению с остальными моделями.

**2.2.2.2.1. Ошибки удалений**

На уровне удалений модель *whisper-large-v3-turbo* демонстрирует тенденцию к пропуску не только отдельных служебных слов, что мы можем наблюдать и у двух других ASR-моделей, но и целых синтаксических конструкций. Подобная особенность была зафиксирована в исследовании И. Д. Мамаева (2023), что позволяет говорить о системной природе этого явления. Так, в одной из аудиозаписей модель не распознала обширный начальный фрагмент высказывания, в котором содержалось, по меньшей мере, шесть клауз: *«в общем это произошло на моих новогодних каникулах в общем я как приехал с поезда я сразу написал всем своим друзьям нужно обязательно всем всем собраться отметить как бы предстоящие новогодние праздники ну я приехал там двадцать девятого тридцатого числа по-моему ну где-то вот».* В таких случаях итоговый транскрипт теряет значительный объём информации, включая инициирующие реплики, вступления и указания на последовательность событий, что отрицательно сказывается на связности и содержательной полноте фиксирующегося высказывания. При подсчете было отмечено, что практически 20% удалений у модели *whisper-large-v3-turbo* содержат как минимум одну клаузу.

Следует отметить, что модели *gigaam-v2-rnnt* и *nemo-fastconformer-ru-rnnt* демонстрируют более высокую устойчивость к удалению знаменательных частей речи по сравнению с *whisper-large-v3-turbo*. Вместе с тем, они также совершают пропуски предлогов, союзов и дискурсивных частиц (например, *вот, ну*). Причём последние играют важную роль в структурной и прагматической организации речи.

Необходимо подчеркнуть, что при подготовке транскриптов неполные или усечённые слова, возникшие в результате самоисправлений, намеренно не исключались, поскольку предполагалось, что некоторые из них могут быть распознаны ASR-моделями. Тем не менее, ни одна из исследуемых моделей не

продемонстрировала способности к адекватному распознаванию таких единиц. Все случаи самоисправлений (12 штук), отмеченные при ручной транскрипции, были полностью удалены в их автоматических версиях. Конкретные примеры и контексты приведены в таблице 10.

Таблица 10. Фиксация удалённых самоперебиваний в автоматических транскриптах.

| | |
|---|---|
| эталонный транскрипт | я попросила моего **мл=** молодого человека приехать |
| гипотеза (*gigaam-v2-rnnt*) | я попросила моего ** молодого человека приехать |
| гипотеза (*nemo-fastconformer-ru-rnnt*) | я попросила моего ** молодого человека |
| гипотеза (*openai/whisper-large-v3-turbo*) | я попросила ***** ** молодого человека |
| эталонный транскрипт | вот куранты **прозв=** пробили двенадцать вот |
| гипотеза (*gigaam-v2-rnnt*) | вот куранты ***** пробили двенадцать вот |
| гипотеза (*nemo-fastconformer-ru-rnnt*) | вот куранта ***** пробили двенадцать |
| гипотеза (*openai/whisper-large-v3-turbo*) | вот куранта ***** пробили двенадцать |
| эталонный транскрипт | не было **тол=** толкотни толкучки |
| гипотеза (*gigaam-v2-rnnt*) | не было *** толкотни толпучки |
| гипотеза (*nemo-fastconformer-ru-rnnt*) | не было *** толкотни толпучки |
| гипотеза (*openai/whisper-large-v3-turbo*) | не было *** толкотни толпучки |
| эталонный транскрипт | через **пят=** пятнадцать минут никто не пришел |
| гипотеза (*gigaam-v2-rnnt*) | через *** пятнадцать минут никто не пришел |
| гипотеза (*nemo-fastconformer-ru-rnnt*) | через *** пятнадцать минут никто не пришел |
| гипотеза (*openai/whisper-large-v3-turbo*) | через *** пятнадцать минут никто не пришел |

Подводя итог анализа ошибок удаления, можно выделить уязвимые стороны модели *whisper-large-v3-turbo*, в частности её склонность к пропуску протяжённых фрагментов речи – от отдельных фраз до полноценных синтаксических конструкций, что ведёт к существенным потерям информации. В то же время все исследуемые модели демонстрируют систематическую тенденцию к удалению незнаменательных единиц (предлогов, союзов, дискурсивных частиц), а также к неспособности фиксировать самоисправления. Эти элементы часто оказываются нераспознанными из-за своей фонетической редуцированности, слабой артикуляции и высокой степени слияния с

последующими знаменательными словами, что, возможно, затрудняет их извлечение алгоритмами.

Подводя общий итог лингвистического анализа, важного для оценки пригодности исследуемых моделей к использованию в качестве транскрипторов устной звучащей речи для диагностических целей, следует отметить, что полученные результаты коррелируют с автоматическими метриками качества. Наилучшие показатели продемонстрировала модель *gigaam-v2-rnnt*, что, вероятно, связано с её узкой адаптацией под русский язык. В текущем исследовании именно эта модель обеспечивает наиболее полные и интерпретируемые транскрипты. Тем не менее, даже при высоких показателях точности, автоматические транскрипты остаются лишь предварительным этапом: для применения в диагностических и научных целях необходима ручная проверка и доразметка. Особенно это важно при работе со спонтанной речью, где значимыми оказываются не только лексические единицы, но и паралингвистические характеристики – самоперебивы и заполнители пауз. Современные ASR-системы, не обученные на специализированных лингвистических корпусах, как правило, игнорируют эти феномены, тем самым теряя важную диагностическую и дискурсивную информацию. В связи с этим наиболее рациональной стратегией на текущем этапе развития технологий является использование гибридного подхода (assisted transcription), предложенного Т. Базиллон, И. Эстеве., Д. Лаззати (Bazillon, Esteve, Luzzati, 2004) сочетающего скорость и масштабируемость ASR с точностью и глубиной ручной доразметки. Как итог, модель *gigaam-v2-rnnt* была выбрана в качестве базового транскрибирующего алгоритма для обработки оставшейся части данных корпуса, с последующей ручной валидацией для обеспечения необходимого уровня качества и полноты транскрипции. Итоговые транскрипты, прошедшие ручную валидацию, а также автоматические транскрипции, полученные на этапе тестирования трех моделей, представлены в Приложении 1. Код, реализующий расчёт автоматических метрик, а также

вывод конкретных ошибок находится в репозитории, располагающимся по ссылке в Приложении 2.

### 2.3. Создание разметки для диагностического корпуса

После завершения процедуры транскрибирования всех 60 аудиозаписей осуществляется этап их аннотирования, направленный на обеспечение релевантности корпуса для решения диагностических задач. Однако, чтобы процесс аннотирования был воспроизводимым и применимым к новым речевым данным, поступающим в корпус, необходимо разработать унифицированный набор обозначений разметки, отражающий специфику задач, решаемых в рамках исследования.

Одним из центральных компонентов разметки стало разграничение речевого потока на элементарные дискурсивные единицы (ЭДЕ) – минимальные синтаксико-семантические и просодически завершённые отрезки речи, широко используемые в корпусных исследованиях устного дискурса. Вместе с тем, в целях повышения диагностического потенциала корпуса, в настоящем исследовании было принято решение дополнительно использовать разметку, с помощью которой возможно фиксировать речевые сбои.

В ряде автоматических и эмпирических исследований показано, что признаки лжи и правды могут быть зафиксированы как на вербальном, так и на невербальном уровне (Loy, Rohde, Corley, 2017; Solà-Sales и др., 2023; Loy, Rohde, Corley, 2018). Однако в рамках настоящего проекта акцент сделан на речевые сбои как потенциальные индикаторы когнитивной нагрузки и коммуникативной неуверенности, сопровождающих акт лжи. Такие феномены, как незаполненные и заполненные паузы, повторы, фальстарты и самокоррекции, получают всё большее внимание в лингвистических и психолингвистических исследованиях (Кибрик, Подлесская, 2007: 1). Особенно ценно включение таких феноменов в корпус в свете результатов исследования *Acoustic-Prosodic and Lexical Cues to Deception and Trust* (Chen и др., 2020), где показано, что некоторые типы речевых сбоев – в частности, хезитационные паузы – могут устойчиво ассоциироваться с ложью в восприятии слушателей.

Однако и менее надёжные с точки зрения статистики признаки (такие как задержки перед ответом на вопрос, повторы, самокоррекции) оказываются значимыми для анализа, так как отражают расхождение между когнитивными механизмами планирования речи и социальным восприятием достоверности.

Таким образом, включение обозначений, регистрирующих речевые сбои, представляется обоснованным и методологически значимым шагом, направленным на фиксацию потенциально релевантных признаков ложности, расположенных на стыке вербального, просодического и прагматического уровней речи.

### 2.3.1. Разработка протокола аннотирования

В качестве основы для аннотирования создаваемого корпуса была взята система разметки, применённая в проекте *«Рассказы о сновидениях и другие корпуса звучащей речи»*, в которой уже были реализованы обозначения ключевых типов речевых сбоев. В частности, в данной системе использовались специальные маркеры для фиксации обрывов слов (=), а также слабых и сильных фальстартов (|| и == соответственно), что позволило задать начальный ориентир при разработке протокола разметки. Тем не менее, поскольку одной из перспектив данного проекта является последующий автоматический анализ речевых проявлений с использованием программных инструментов, исходная система была адаптирована с учётом требований к машинной обработке. В частности, предложенный формат разметки оптимизирован для автоматического подсчёта частот и распределений дискурсивных элементов с помощью, например, регулярных выражений, что делает его более универсальным и пригодным для масштабируемого анализа.

Разработка протокола велась совместно с научной учебной группой «Комплинг» и включала серию итеративных доработок, направленных на повышение точности, формализованности и воспроизводимости аннотации. В существующую схему были внесены следующие корректировки:

1. Замена существующих обозначений:

a. **Незаполненные и заполненные паузы.** В отличие от традиционной системы, где длительность незаполненных пограничных и внутренних пауз отражается количеством «висячих точек» или букв (Кибрик и др., 2009), в данной разметке все упомянутые выше паузы обозначаются унифицированным троеточием (…), а их длительность указывается численно в скобках. Это также касается пауз с вокалическим гласным или сонорным заполнением (пример оформления: *э(0.2), м(1.0)*). Пример реализации разметки из корпуса:

*А-а вот как раз* ***э(0.3)*** *когда мы пришли,*
*я во-первых был поражен тем,*
***м(0.2)*** *какая атмосфера царила вокруг..* ***…(0.4)***
*Это просто было незабываемо..*

b. **Пограничные паузы** фиксируются в конце текущей элементарной дискурсивной единицы (ЭДЕ), а не в начале последующей, как это принято, например, в модели транскрибирования, предложенной А. А. Кибриком (2009). Данное решение обусловлено, прежде всего, техническими и визуальными соображениями, связанными с упрощением автоматизированной обработки данных. Пример реализации разметки из корпуса:

*попила водички,*
*села на следующий автобус,*
*доехала,*
*и пошла на пары.* ***...(0.8)***

c. В связи с унификацией обозначения незаполненных пауз с использованием троеточия (…), было принято решение переопределить графическое **обозначение интонации многоточия**, или иллокутивно-фазового значения неполноты

информации. Для её фиксации используется символ из двух точек (..), что позволяет избежать смешения с паузами и обеспечивает однозначность при автоматической обработке транскриптов.

*а я уже не помню,*

*в какие-то игры там начали играть**..***

*там просто вот у= сходить с ума**..***

*там на улицу выходили**..***

*кричали чё-то**..***

d. Теперь отображение **удлинения в произнесение слов** фиксируется той же литерой, которая пишется в орфографически корректном варианте слова. Пример реализации из корпуса:

*Да,*

*а(0.3) **с-с** к= хреном,*

*таки**е-е** зелёненькая пачка,*

*самые дешёвые,*

*на что ему ...(0.1) в принципе только хватает.*

e. Было предложено фиксировать **пересекающиеся реплики** следующим образом: *Это было незабываемое [событие! Р2: Да, ]согласна с тобой*. Такое оформление позволяет сохранить визуальную наглядность синхронной речи и обеспечить возможность автоматизированного подсчёта наложений: одна пара квадратных скобок считается за одно пересечение реплик.

f. Все **случаи цитаций** на данном этапе разметки обозначаются простыми кавычками («»). В дальнейшем предполагается обрабатывать каждый конкретный случай отдельно. Пример реализации из корпуса:

*Они сказали: ...(0.2)*

*«Мы пойдём на речку,*

*а то купаться охота.»*

g. Обозначение **неразборчивых звуковых отрезков** (<НРЗБ>) заменено на {НРЗБ}, поскольку использование угловых скобок может вступать в конфликт с возможной XML-разметкой, предусмотренной для дальнейшей машинной обработки корпуса. Кроме того, рекомендуется **придерживаться одного варианта транскрипции** для каждого речевого отрезка, избегая дублирующих трактовок внутри самого транскрипта. В случае сомнений альтернативные интерпретации предлагается выносить в аннотационные комментарии, сопровождающие соответствующий участок. Пример реализации из корпуса:

*Э(0.2) в юности, |S|*

*(**{НРЗБ}** ну уже как бы у меня возраст такой,*

*что часто вспоминаешь |E|*

*всё-таки что же с тобой было..*

h. В случае, если предложение состоит из нескольких элементарных дискурсивных единиц (ЭДЕ), каждая из которых выражает **различное иллокутивное значение**, разметка предполагает указание иллокуции для каждой ЭДЕ отдельно. Такой подход позволяет более точно отразить прагматическую структуру высказывания. Пример реализации в корпусе:

*Нет,*

*ну ты погляди на неё¡,*

*как она пляшет!,*

*кружится!*

2. Введение новых обозначений:

a. **|E|** – знак границы ЭДЕ, в случаях, когда мы не можем считать её полноценной.
b. **|S|** – знак границы предложения, когда оно не может считаться завершенным.
c. **|ES|** – знак границы ЭДЕ совмещенный с границей предложения (если эти границы попадают на одну ЭДЕ).

В связи с новыми обозначениями в разметку были внесены следующие замены:

a. Вместо использования символов (==), теперь для фиксации **сильного фальстарта** применяется сочетание отсутствия знаков пунктуации в конце элементарной дискурсивной единицы (ЭДЕ) с последующим использованием специальных маркеров |E| или |ES|. Пример реализации из корпуса:

   *Вчера такой день был, …(0.3)*
   *что просто* ***|E|***
   *э(0.2) солнышко светило,*
   *пары рано кончились..*

b. Вместо использования символов (—), применяемых для обозначения начала и конца **сплита** (разрыва ЭДЕ), теперь для фиксации разрыва используется маркер |E|, который ставится в конце ЭДЕ перед вставкой. Таким образом, |E| сигнализирует о прерывании синтаксически/просодически связной структуры вставным элементом. Конец вставки не маркируется специально, если продолжающаяся после неё часть оформляется как отдельная полноценная ЭДЕ.

   *Там сидели мои друзья |E|*
   *дво-о-е, …(0.3)*
   *или тро-о-е,*
   *и брат мой старший был.*

c. Обозначение **парентезы** (вставки одного предложения внутрь другого). Знак |E| используется только в тех случаях, когда вставке предшествует

неполноценная ЭДЕ (например, прерывание или отсутствие финального пунктуационного оформления). Если же вставке предшествует полноценная ЭДЕ (с финальной просодией и пунктуацией), никакого дополнительного маркера не требуется – остаются только круглые скобки, указывающие на границы парентезы.

*Я вчера прихожу домой* ***|E|***

***(****А было темно-о уже.****)***

*слышу,*

*дверь хлопает, …(0.3)*

*и кто-то мне из темноты:*

*«Вы не подскажете?,*

*где тут магазин?»*

d. Вместо использования символа (*, ранее применявшегося для фиксации **односторонней парентезы** (начатой, но не завершённой вставки), теперь используется маркер |S|. Маркер |S| ставится в конце фрагмента, где предполагалась вставка, но говорящий не завершил её, и вернулся (или переключился) к другой линии изложения. Таким образом, |S| обозначает обрыв парентезы и одновременно сигнализирует о структурной незавершённости высказывания. Пример реализации в корпусе:

*Я хотел сказать,* ***|S|***

*А ты видела кстати?,*

*ливень вчера был!,*

*да сильный какой.*

Все остальные явления оформляются в прежнем формате разметки корпуса «Рассказы о сновидениях».

На данный момент протокол оформлен в виде удобной табличной схемы, предназначенной для аннотаторов, участвующих в проекте НУГ. Таблица содержит основные обозначения размечаемых явлений, краткие примечания, уточняющие критерии их различения, а также примеры реализации с целью обеспечения единообразного и осмысленного подхода к аннотированию устной

речи. С протоколом можно ознакомиться по ссылке в разделе Приложения (Приложение 3).

Данная аннотаци уже легла в основу разметки 60 аудиофайлов корпуса.

### 2.4. Создание архитектуры корпуса

Итоговый корпус устной речи, сформированный в рамках данного исследования, предполагает структурированное и открытое хранение всех компонентов, обеспечивающее как воспроизводимость результатов, так и возможность его дальнейшего пополнения.

Структура корпуса организована по принципу слотов, каждый из которых содержит определённый тип данных, относящийся к одному и тому же респонденту. Для каждого из 60 респондентов предусмотрены следующие слоты:

1. **Аудиофайлы** – оригинальные записи высказываний, сохранённые в формате .mp3 с параметрами 44.1 кГц и 16 бит.
2. **Орфографическая транскрипция** – вручную проверенные и откорректированные текстовые расшифровки аудиофайлов, оформленные в формате .txt;
3. **Аннотированные транскрипции** – файлы, содержащие лингвистическую разметку по ключевым параметрам (включая самоисправления, дискурсивные единицы, иллокутивные акты, ЭДЕ и др.), в формате .txt.

На главной странице репозитория размещён файл README.md, содержащий:

- описание структуры корпуса и его назначения;
- протокол разметки, поясняющий правила аннотации речевых явлений;
- структурированную таблицу с гиперссылками на соответсвующие файлы для каждго респондента;
- возможность обратной связи с целью предложения дополнительных категорий разметки или уточнений к уже существующим.

Корпус задуман как открытый и расширяемый: предполагается его дальнейшее пополнение речевыми данными и углубление аннотационного слоя по мере роста исследовательского интереса. Его наглядная структура изображена на рисунке 2.

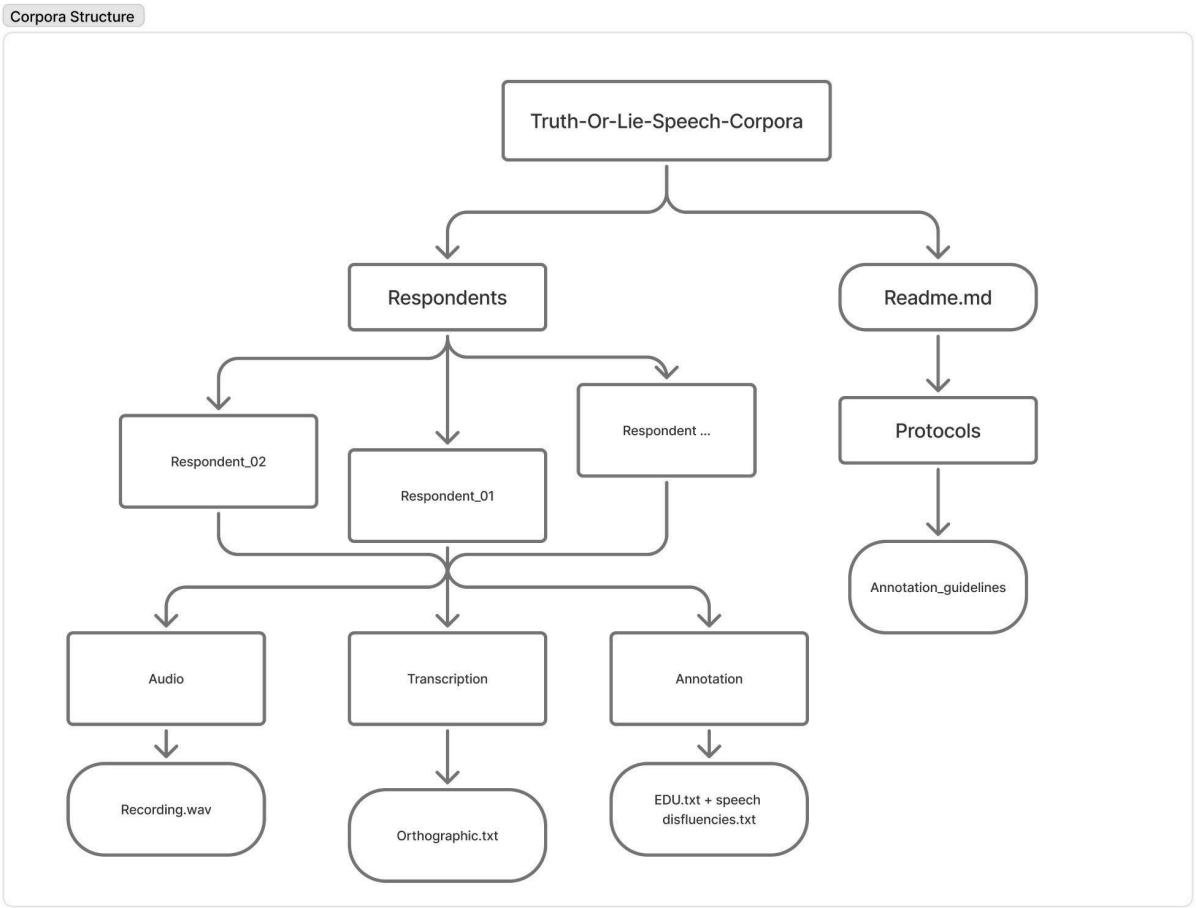


Рисунок 2. Архитектура корпуса устной речи, предназначенного для диагностических целей.

Ссылка на репозиторий с корпусом находится в Приложении 4.

## Заключение

В рамках данного исследования были протестированы и сравнены три современные модели автоматического распознавания речи (ASR) с точки зрения качества транскрипции и способности фиксировать просодические и дискурсивные особенности, в частности речевые сбои. Результаты анализа показали, что современные алгоритмы распознавания речи на текущем этапе развития недостаточно приспособлены к решению задач, связанных с фиксацией речевых сбоев (в частности самоперебивов, заполнений пауз хезитаций), и требуют дополнительного обучения на специализированных наборах данных, адаптированных под эти цели. Кроме того, проведенное сравнение позволило выявить характер ошибок, допускаемых различными моделями: так, модели *whisper-large-v3-turbo* и *nemo-fastconformer-ru-rnnt* чаще всего допускали искажения, изменяющие смысл высказывания, тогда как модель *gigaam-v2-rnnt* проявила большую устойчивость к подобным ошибкам, ограничиваясь преимущественно заменой флексий, что, в свою очередь, влияло лишь на грамматические категории, оставляя семантику лексем неизменной. Анализ работы моделей на различных типах данных – монологах, диалогах, записях с зашумленным фоном – продемонстрировал, что некоторые модели демонстрируют устойчивость даже в условиях акустических помех, в то время как другие теряют способность адекватно фиксировать реплики участников диалога. Совокупность результатов, полученных с использованием автоматических метрик и лингвистического анализа, позволила сформировать комплексное представление о текущих возможностях и ограничениях современных ASR-систем в исследуемой области. По итогам анализа для дальнейшей работы в качестве транскрибатора с последующей ручной валидацией была выбрана модель *gigaam-v2-rnnt*.

Особое внимание в ходе исследования было уделено разработке протокола разметки речевых данных, с акцентом на фиксацию речевых сбоев как потенциальных маркеров лжи. Этот процесс потребовал обращения к вопросам восприятия речевых сбоев как индикаторов правдивости или

ложности высказываний. Разработанная система аннотации основана на существующих стандартах разметки, используемых в фундаментальных исследованиях, и дополнена рядом усовершенствований, направленных на обеспечение автоматического подсчёта частот тех или иных типов речевых сбоев для каждой из категорий правдивых и ложных высказываний. Такой подход, как представляется, создаёт основу для будущих исследований роли речевых сбоев в диагностике лжи и правды.

В то же время создание корпуса, его открытость и доступность для дальнейшего пополнения, а также возможность его адаптации под новые уровни аннотаций, делает данный ресурс важным вкладом в развитие задач по диагностике правдивости и ложности информации в устной речи. В будущем предполагается расширение корпуса за счёт новых данных и углубление уровней разметки, что позволит более полно исследовать механизмы восприятия и производства речевых актов в контексте лжи и правды.

## Список литературы


1. Абаева Ю. Д. Принципы просодического транскрибирования звучащего текста в корпусных исследованиях //Филологические науки. Вопросы теории и практики. – 2021. – Т. 14. – №. 2. – С. 553-557.
2. Бердникова Т. В. Исследование спонтанности и подготовленности звучащей речи в судебной экспертизе: к постановке проблемы //Теория и практика судебной экспертизы. – 2024. – Т. 18. – №. 4. – С. 6-11.
3. Богданов Д. С., Кривнова О. Ф., Подрабинович А. Я. Современный инструментарий для разработки речевых технологий //Информационные технологии и вычислительные системы. – 2004. – №. 2. – С. 11-24.
4. Богданова-Бегларян Н. В. и др. Звуковой корпус русского языка: новая методология анализа устной речи //ЯЗЫК И МЕТОД. – 2015. – С. 357-372.
5. Божович Е. Д. Проблематика развития речи ребенка и обучения чтению в трудах ДБ Эльконина //Культурно-историческая психология. – 2014. – Т. 10. – №. 1. – С. 26-33.
6. Васильева В. В., Коньков В. И. Устная речь: практикум //Петерб. гос. ун-т, Ин-т «Высш. шк. журн. и мас. коммуникаций. – 2015. – Т. 100.
7. Васютина И. А. Диагностика обученности учащихся составлению связного устного высказывания на заданную тему //Сибирский педагогический журнал. – 2007. – №. 7. – С. 291-297.
8. Венцов А.В. Изучение восприятия устной речи: в поисках оптимального метода //Language and Method. – 2015. – Т. 2015. – №. 2. – С. 327-334.
9. Готлиб А. С. и др. Процедуры и методы социологического исследования. Практикум. – 2010.
10. Захаров В. П. Корпусная лингвистика: Учебно-методическое пособие //СПб.: СПбГУ. – 2005.
11. Кибрик А. А., Майсак Т. А. Правила дискурсивной транскрипции для описательных и документационных исследований //Rhema. Рема. – 2021. – №. 2. – С. 23-45.

12. Кибрик А. А., Подлесская В. И. К созданию корпусов устной русской речи: принципы транскрибирования //Научно-техническая информация. Серия. – 2003. – Т. 2. – №. 6. – С. 5-11.
13. Князев С. В., Пожарицкая С. К. Современный русский литературный язык: фонетика, орфоэпия, графика, орфография. – 2011.
14. Козлова Н. В. Лингвистические корпуса: определение основных понятий и типология //Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. – 2013. – Т. 11. – №. 1. – С. 79-89.
15. Кривнова О. Ф. Речевые корпуса на новом технологическом витке //Речевые. – 2008.
16. Мамаев И. Д. АВТОМАТИЧЕСКАЯ РАСШИФРОВКА ЗАПИСЕЙ УСТНОЙ РЕЧИ: ТЕСТИРОВАНИЕ ПРОГРАММЫ WHISPER1. – 2023.
17. Подлесская В. И., Кибрик А. А. Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи //Научно-техническая информация. Серия. – 2007. – Т. 2. – С. 2-23.
18. Рычкалова Л. А. Лингвистические методы анализа звучащей речи в криминалистике //Юрислингвистика. – 2002. – №. 3. – С. 105-113.
19. Чилингарян К. П. Корпусная лингвистика: теория vs методология //Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. – 2021. – Т. 12. – №. 1. – С. 196-218.
20. Эльконин Д. Б. К проблеме периодизации психического развития в детском возрасте. – 1989.
21. Adda-Decker, M., & Lamel, L. (2005). Do speech recognizers prefer female speakers? // Interspeech 2005: Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4–8, 2005. P. 2205–2208.
22. Alsaawi A. Spoken and written language as medium of communication: A self-reflection //International Journal of Applied Linguistics and English Literature. – 2019. – Т. 8. – №. 2. – С. 194-198.

23. Ardila R. et al. Common voice: A massively-multilingual speech corpus //arXiv preprint arXiv:1912.06670. – 2019
24. Bazillon T., Estève Y., Luzzati D. Transcription manuelle vs assistée de la parole préparé et spontanée //Revue TAL. – 2008.
25. Chafe, W. and Tannen, D. (1987) 'The relation between written and spoken language', Annual Review of Anthropology, 16, pp. 383-407.
26. Chan M. P. Y. et al. Training and typological bias in ASR performance for world Englishes //INTERSPEECH. – 2022. – С. 1273-1277.
27. Chen X. et al. Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies //Transactions of the Association for Computational Linguistics. – 2020. – Т. 8. – С. 199-214.
28. Choe J. et al. Language-specific effects on automatic speech recognition errors for world Englishes //Proceedings of the 29th international conference on computational linguistics. – 2022. – С. 7177-7186.
29. Conneau A. et al. Fleurs: Few-shot learning evaluation of universal representations of speech //2022 IEEE Spoken Language Technology Workshop (SLT). – IEEE, 2023. – С. 798-805.
30. Francis W. N., Kučera H. Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers. – Brown University, Department of Linguistics, 1979.
31. Godfrey J. J., Holliman E. C., McDaniel J. SWITCHBOARD: Telephone speech corpus for research and development //Acoustics, speech, and signal processing, ieee international conference on. – IEEE Computer Society, 1992. – Т. 1. – С. 517-520.
32. Graham C., Roll N. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits //JASA Express Letters. – 2024. – Т. 4. – №. 2.
33. Gries S. T., Berez A. L. Linguistic annotation in/for corpus linguistics //Handbook of linguistic annotation. – 2017. – С. 379-409.

34. Hanke, T.: HamNoSys - representing sign language data in language resources and language processing contexts. In: Streiter, O., Chiara, C. (eds).: Proceedings of the Workshop Representation and Processing of Sign Languages, LREC 2004, pp. 1–6. ELRA, Paris (2004)
35. James J. et al. Advocating character error rate for multilingual asr evaluation //arXiv preprint arXiv:2410.07400. – 2024.
36. Jefferson, G.: Sequential aspects of storytelling in conversation. In: Schenkein, J. (ed.) Studies in the Organization of Conversational Interaction, pp. 219–248. Academic Press, New York (1978)
37. Jun, S.A. (ed.): Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford (2005)
38. K.A. Bayda (Ivanova), M.A. Kholodilova, A.D. Yegorova (Kozhemjakina), E.A. Romanova, T.E. Remizova, A.A. Storozheva, N.K. Tarasova, A.A. Zorina, V.A. Morozova, A.B. Panova, N.R. Dobrushina. ChuvashRus Corpus. 2018. Moscow: Linguistic Convergence Laboratory, HSE University. (Available online at URL: https://lingconlab.ru/ChuvashRus, accessed on 10.05.2025.)
39. Khudyakova M. et al. Discourse diversity database (3D) for clinical linguistics research: Design, development, and analysis //Bakhtiniana: Revista de Estudos do Discurso. – 2022. – T. 18. – C. 32-57.
40. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gesture: how to economically capture timing and form. Lang. Resour. Eval. 41(3/4), 325–339 (2007)
41. Latif S. et al. Can large language models aid in annotating speech emotional data? uncovering new frontiers //arXiv preprint arXiv:2307.06090. – 2023.
42. Liu Y., Yang X., Qu D. Exploration of Whisper fine-tuning strategies for low-resource ASR //EURASIP Journal on Audio, Speech, and Music Processing. – 2024. – T. 2024. – №. 1. – C. 29.
43. Loy J. E., Rohde H., Corley M. Effects of disfluency in online interpretation of deception //Cognitive Science. – 2017. – T. 41. – C. 1434-1456.

44. Loy J. E., Rohde H., Corley M. Cues to lying may be deceptive: Speaker and listener behaviour in an interactive game of deception //Journal of Cognition. – 2018. – Т. 1. – №. 1. – С. 42.
45. McEnery T., Hardie A. Corpus linguistics: Method, theory and practice. – Cambridge University Press, 2011.
46. Morris, A. C., Maier, V., & Green, P. D. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In Proceedings of Interspeech 2004 (pp. 2765–2768). International Speech Communication Association.
47. Morris, A. C., Maier, V., & Green, P. D. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In Proceedings of Interspeech 2004 (pp. 2765–2768). International Speech Communication Association.
48. Nina Dobrushina, Michael Daniel, Ruprecht von Waldenfels, Timur Maisak, Anastasia Panova. 2018. Corpus of Russian spoken in Daghestan. Moscow: Linguistic Convergence Laboratory, HSE University. (Available online at http://www.parasolcorpus.org/dagrus/, accessed on 10.05.2025.)
49. Noroozi V. et al. Stateful conformer with cache-based inference for streaming automatic speech recognition //ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2024. – С. 12041-12045.
50. Ong W. J., Hartley J. Orality and literacy. – Routledge, 2013.
51. Oostdijk, N., Boves, L.: Preprocessing speech corpora. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics: An International Handbook, vol. 1, pp. 642–663. Walter de Gruyter, Berlin (2008)
52. Panayotov V. et al. Librispeech: an asr corpus based on public domain audio books //2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2015. – С. 5206-5210.

53. Rekesh D. et al. Fast conformer with linearly scalable attention for efficient speech recognition //2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). – IEEE, 2023. – С. 1-8.
54. Russell S. O. C. et al. What automatic speech recognition can and cannot do for conversational speech transcription //Research Methods in Applied Linguistics. – 2024. – Т. 3. – №. 3. – С. 100163.
55. Shahin M., Zafar U., Ahmed B. The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results //IEEE Journal of Selected Topics in Signal Processing. – 2019. – Т. 14. – №. 2. – С. 400-412.
56. Shahin M., Zafar U., Ahmed B. The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results //IEEE Journal of Selected Topics in Signal Processing. – 2019. – Т. 14. – №. 2. – С. 400-412.
57. Sherstinova T. et al. Bridging gaps in Russian language processing: AI and everyday conversations //2024 35th Conference of Open Innovations Association (FRUCT). – IEEE, 2024. – С. 665-674.
58. Sinclair J. EAGLES Preliminary recommendations on Corpus Typology, 1996. Режим доступа: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (дата обращения: 10.04.2025).
59. Solà-Sales S. et al. Analysing deception in witness memory through linguistic styles in spontaneous language //Brain sciences. – 2023. – Т. 13. – №. 2. – С. 317.
60. Wieczorkowska A. Methodology for Obtaining High-Quality Speech Corpora //Applied Sciences. – 2025. – Т. 15. – №. 4. – С. 1848.
61. Yamagishi J. et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92) //University of Edinburgh. The Centre for Speech Technology Research (CSTR). – 2019. – С. 271-350.
62. Zafar A. et al. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription //International Journal of Medical Informatics. – 2004. – Т. 73. – №. 9-10. – С. 719-730.