Факультет гуманитарных наук | Фундаментальная и прикладная лингвистика | Нижний Новгород

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

**Докладчик:** Жилина Полина Павловна 21ФиПЛ-2

**Научный руководитель:** к. филол. н., старший научный сотрудник Центра языка и мозга, Хоменко Анна Юрьевна

**Соруководитель:** приглашенный преподаватель, Бадасян Александра Арсеновна

# Актуальность

- Устная речь – источник когнитивной и психоэмоциональной информации.
- Точная транскрипция – основа лингвистического и диагностического анализа.
- Автоматическое распознавание речи (Automatic Speech Recognition, ASR) позволяет масштабировать обработку звучащей речи.
- Аннотированные корпуса – ключевой ресурс для анализа устной речи в том числе в диагностических и исследовательских целях.

# Новизна

В отличие от большинства русскоязычных работ, ограничивающихся использованием различных версий модели Whisper для автоматического распознавания речи (Мамаев, 2023; Sherstinova и др., 2024), **в настоящем исследовании сравнивается эффективность трёх современных ASR-моделей — *whisper-large-v3-turbo, nemo-fastconformer-ru-rnnt, gigaam-v2-rnnt***, две из которых редко применяются в лингвистических исследованиях, но при этом демонстрируют высокую точность распознавания русской устной речи на различных наборах данных (Иступаков, 2025).

## Объект:

устная речь, содержащая достоверную и недостоверную информацию.

## Предмет:

- особенности устной речи, которые либо не поддаются фиксации средствами автоматического распознавания речи (ASR), либо воспроизводятся неверно, приводя к ошибкам в транскриптах.
- особенности дискурсивной разметки автоматически созданных транскриптов.

## Цель работы:

создание корпуса устной речи для диагностических целей, содержащего как правдивые так и ложные высказывания, с применением современных алгоритмов автоматического распознавания речи

# Задачи:

- проанализировать научную литературу по теме исследования;
- **получить и сравнить автоматические транскрипты с эталонными ручными расшифровками**, подготовленными для анализа ошибок распознавания;
- **классифицировать ошибки** с точки зрения их лингвистической и диагностической значимости;
- **выявить наиболее эффективный алгоритм распознавания речи** для создания последующих транскриптов в корпусе;
- **разработать протокол разметки транскриптов**, учитывая специфику диагностического корпуса;
- **сформировать структуру корпуса** и подготовить его к размещению в виде открытого репозитория.

# Методы:

- **Автоматического создания транскриптов** с использованием ASR-моделей *whisper-large-v3-turbo*, *nemo-fastconformer-ru-rnnt* и *gigaam-v2-rnnt*.
- **Ручного создания транскриптов** для создания эталонных текстов.
- **Сравнительного анализа** для сопоставление автоматически созданных и ручных транскриптов, классификация и локализация ошибок.
- **Дискурсивной разметки** для разработки протокола аннотирования речевых данных в корпусе.

# Материал исследования:

- **60 аудиофайлов (20 человек);**
- **Возрастная группа: от 17 до 67 лет;**
- **Общая продолжительность материала: 231 минута 37 секунд;**
- **Устные рассказы респондентов**, собранные в три этапа:
  - **Этап 1.** Ложный рассказ:
    - Респондент по заданному плану рассказывал о вымышленном событии, заведомо зная, что должен лгать; после монолога интервьюер задавал уточняющие вопросы.
  - **Этап 2.** Правдивый рассказ:
    - Респондент описывал реальное событие, также с опорой на план. Вопросы задавались после рассказа.
  - **Этап 3.** Свободный выбор:
    - Респондент сам решал, будет ли его рассказ правдивым или вымышленным. Интервьюер также задавал уточняющие вопросы.

# Автоматические метрики оценки качества распознавания речи

**WER (Word Error Rate)** - показатель, оценивающий частоту ошибок на уровне слов.

**CER (Character Error Rate)** - показатель, оценивающий частоту ошибок на уровне символов.

**WIP (Word Information Preserved)** – показывает, насколько эффективно система сохраняет информацию из оригинального высказывания.

**WIL (Word Information Lost)** – измеряет долю утраченной информации в результате ошибок распознавания.

*Чем ниже значения метрик **WER**, **CER** и **WIL**, тем выше точность и качество распознавания.

# Результаты

Оценка производилась на аудиоданных общей длительностью **71 минута 41 секунд** (30.95% от общего кол-ва записанного материала)

| метрика (%) / ASR - модель | *nemo-fastconformer-ru-rnnt* | *whisper-large-v3-turbo* | *gigaam-v2-rnnt* |
|---|---|---|---|
| **WER** | 11.09 | 16.46 | **6.18** |
| **CER** | 4.4 | 9.94 | **2.13** |
| **WIL** | 17.42 | 20.84 | **9.54** |
| **WIP** | 82.58 | 79.16 | **90.46** |

Таблица 1. Результаты оценки качества на незашумленном аудиоматериале

| метрика (%) / ASR - модель | *nemo-fastconformer-ru-rnnt* | *whisper-large-v3-turbo* | *gigaam-v2-rnnt* |
|---|---|---|---|
| **WER** | 22.48 | 24.77 | **13.18** |
| **CER** | 9.11 | 16.63 | **6.28** |
| **WIL** | 34.47 | 32.11 | **19.53** |
| **WIP** | 65.53 | 67.89 | **80.47** |

Таблица 2. Результаты метрик оценки качества на зашумленном аудиоматериале

# Результаты (классификация ошибок)

- **Ошибки, не искажающие смысл**
  - Затрагивают лишь формальные грамматические характеристики слова (например, род, число, падеж у существительных и прилагательных):

| | |
|---|---|
| эталонный транскрипт<br>гипотеза | отвезти другой моей **подруге**<br>отвезти другой моей **подруги** |

- **Ошибки, искажающие смысл.** В результате появляется:
  - полностью другая лексема(ы), не связанная с контекстом, но схожая по звуковому облику:

| | |
|---|---|
| эталонный транскрипт<br>гипотеза | **мне** дают бланк я заполняю этот **бланк**<br>**не** дают бланк я заполняю этот **банк** |

  - несуществующий звуковой аналог:

| | |
|---|---|
| эталонный транскрипт<br>гипотеза | не было **толкучки** никто на пятки не наступал<br>не было **толпучки** никто на пятки не наступал |

# Результаты

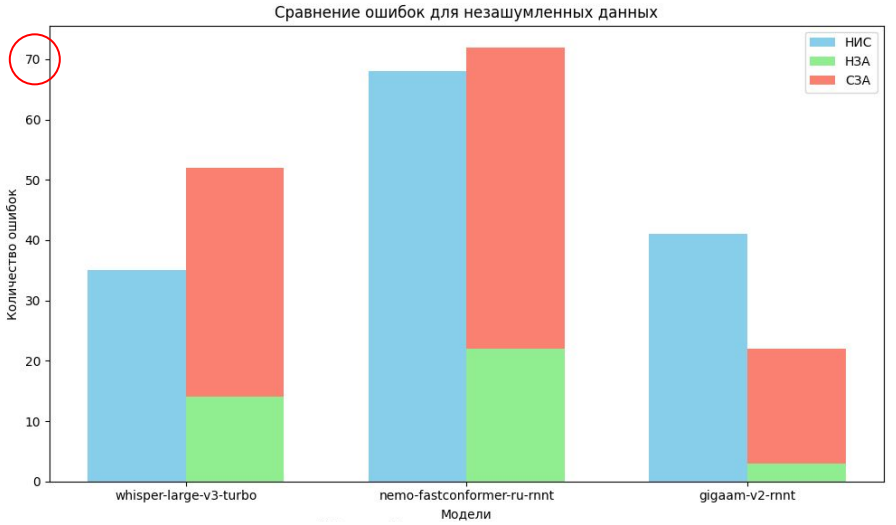
**Для каждой модели было выявлнено общее количество ошибок :**

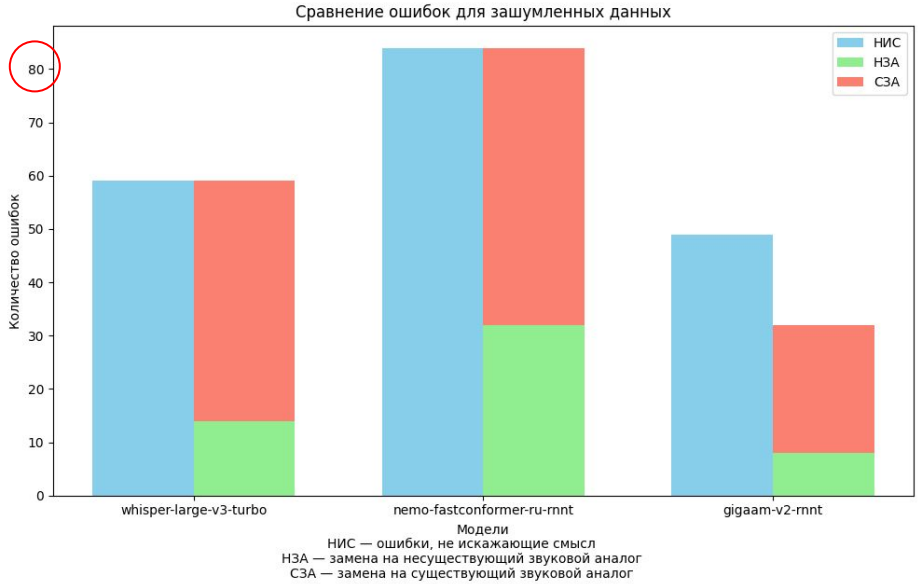
- **на незашумленных данных:**
- ***whisper-large-v3-turbo* – 87 ошибок**
- ***nemo-fastconformer-ru-rnnt* – 140 ошибок**
- ***gigaam-v2-rnnt* – 62 ошибки**

**Общее количество токенов: 4812**

- **на зашумленных данных:**
- ***whisper-large-v3-turbo* – 123 ошибки**
- ***nemo-fastconformer-ru-rnnt* – 189 ошибок**
- ***gigaam-v2-rnnt* – 81 ошибка**

**Общее количество токенов: 4167**



# Результаты (ошибки пропусков)

| Модель распознавания речи/вид ошибки | Служебные слова (предлоги, союзы, частицы) | Знаменательные слова | Целые предложения/отдельные клаузы | Хезитационные паузы | Самоперебивания |
|---|---|---|---|---|---|
| ***whisper-large-v3-turbo*** | ✓ | ✓ | ✓ | ✓ | ✓ |
| ***nemo-fastconformer-ru-rnnt*** | ✓ | ✓ (только личные местоимения) | × | ✓ | ✓ |
| ***gigaam-v2-rnnt*** | ✓ | ✓ (только личные местоимения) | × | ✓ | ✓ |

Таблица 3. Распределение видов пропусков по моделям автоматического распознавания речи

# Протокол разметки

**Цель:** разработка унифицированной системы аннотаций, обеспечивающей релевантность корпуса для задач диагностики правдивости/ложности устной речи.

**Основные принципы:**

- Разделение речевого потока на элементарные дискурсивные единицы (ЭДЕ).
- Фиксация речевых сбоев (паузы, повторы, фальстарты, самокоррекции) как индикаторов когнитивной нагрузки и коммуникативной неуверенности.

**Теоретическая база:**

- Исследования, выявляющие взаимосвязь речевых сбоев и восприятия лжи (Chen et al., 2020; Loy, Rohde, Corley, 2017; Solà-Sales и др., 2023; Loy, Rohde, Corley, 2018).
- Работа «Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи» (Подлесская В. И., Кибрик А. А., 2007) акцентирующая внимание на значимости системной фиксации речевых сбоев для изучения когнитивных и прагматических аспектов устного дискурса, подчеркивая их роль как ключевого объекта аннотирования в корпусах устной речи.

**Практическая реализация:**

- Протокол основан на системе разметки проекта «Рассказы о сновидениях...» (Кибрик А. и др., 2009), адаптированной для автоматического анализа (регулярные выражения, подсчет частот).

**Значимость:** создание инструмента, сочетающего традиции корпусной лингвистики и требования к машинной обработке, для поддержки последующих исследований в области диагностики речевой лжи.

# Протокол разметки

- **Произведена замена 7 существующих обозначений** (незаполненные и заполненные паузы, пограничные паузы, обозначение интонации многоточия, отображение удлинения в произнесение слов, специальная фиксация пересекающихся реплик, обозначение неразборчивых звуковых отрезков)**;**
- **Введено 3 новых обозначения** (|E| – знак границы неполноценной ЭДЕ, |S| – знак границы незавершенного предложения, |ES| – знак границы ЭДЕ совмещенный с границей предложения)**;**
- **Произвелена замена существующих обозначений в связи с введением новых символов** (фиксации сильного фальстарта, обознаение сплита (разрыва внутри ЭДЕ), бозначение односторонней и обычной парентезы).

| **Пример отображения сплита:** | **Пример отображения парентезы:** |
|---|---|
| *Там сидели мои друзья \|E\|*<br>*дво-о-е, ...(0.3)*<br>*или тро-о-е,*<br>*и брат мой старший был.* | *Я вчера прихожу домой \|E\|*<br>*(А было темно-о уже.)*<br>*слышу,*<br>*дверь хлопает, ...(0.3)*<br>*и кто-то мне из темноты:*<br>*«Вы не подскажете?,*<br>*где тут магазин?»* |

С подробной версией протокола вы можете ознакомиться, перейдя по QR-коду:

README

Code Blame 1.92 MB Raw

View raw

# Truth-Or-Lie-Speech-Corpora

| Аудиофайл | Транскрипт | Разметка |
|---|---|---|
| Ж18 (1) 3.mp3 | ссылка | Ж18 (1) 3_2.txt |
| Ж18 (1) врёт.mp3 | ссылка | Ж18 (1) врёт.mp3 |
| Ж18 (1) правда.mp3 | ссылка | Ж18 (1) правда.txt |
| Ж18 (2) 3.mp3 | ссылка | Ж18 (2) 3.txt |
| Ж18 (2) врёт.mp3 | ссылка | Ж18(2) врет.txt |
| М19 (1) говорит правду.mp3 | ссылка | М19(1)врет.txt |
| М19 (2) говорит правду.mp3 | ссылка | М19_3.txt |
| М19 (2) 3.mp3 | ссылка | М19(2)3.txt |
| М18(1) врёт.mp3 | ссылка | М18(1)_врёт.txt |
| М18(2) правда.mp3 | ссылка | М18(2)_правда.txt |

Code Blame 32 lines (31 loc) · 2.73 KB Raw

| | Время начала – чч:мм:сс.мс | Время окончания – чч:мм:сс.мс | Длительность – чч:мм:сс.мс | Рассказчик |
|---|---|---|---|---|
| 2 | 00:00:00.080 | 00:00:01.690 | 00:00:01.610 | Меня зовут Алексеева Анна, |
| 3 | 00:00:02.320 | 00:00:04.500 | 00:00:02.180 | я ходила на концерт со своими друзьями \|E\| |
| 4 | 00:00:04.510 | 00:00:06.340 | 00:00:01.830 | на концерт Рамиля. ...(0.4) |
| 5 | 00:00:06.340 | 00:00:09.410 | 00:00:03.070 | Инициатором концерта была я. ...(0.5) |
| 6 | 00:00:09.412 | 00:00:15.451 | 00:00:06.039 | А(0.3) так…(0.5) мне очень понравились эмоции от этого концерта, |
| 7 | 00:00:15.451 | 00:00:18.686 | 00:00:03.235 | а(0.2) времяпровождение, |
| 8 | 00:00:18.706 | 00:00:19.608 | 00:00:00.902 | а(0.1) очень круто. \|S\| |
| 9 | 00:00:24.784 | 00:00:27.529 | 00:00:02.745 | Я бы порекомендовала своим друзьям сходить на такой концерт, |
| 10 | 00:00:27.529 | 00:00:32.588 | 00:00:05.059 | а(0.2) чтобы отвлечься ...(0.3) от своих проблем, |
| 11 | 00:00:32.608 | 00:00:35.824 | 00:00:03.216 | а(0.2) чтобы провести круто время, |
| 12 | 00:00:35.843 | 00:00:36.882 | 00:00:01.039 | ну и всё. |

Code Blame 4 lines (2 loc) · 2.92 KB Raw

| | |
|---|---|
| 1 | Я Алексеева Анна. Аа, Я н= совсем… ну вот [наверное] на этой неделе, аа, лазила в инстаграме и наткнулась на шоу-рум. Мне там очень сильн |
| 2 | |
| 3 | |
| 4 | Угу. Аа, какого цвета было платье? – Черное. – Черное. Ам, так, второй вопрос. Эм, в каком районе Челябинска, да, в Челябинске же это про |

# Итоги

- Модель *gigaam-v2-rnnt* выбрана для финальной работы с корпусом.
- Современные ASR-системы не фиксируют паралингвистические элементы (паузы, фальстарты, самоперебивы) и требуют дообучения на специализированных данных.
- Все модели допускают как ошибки, искажающие смысл высказываний, так и ошибки, затрагивающие лишь грамматические категории, однако их частотность и выраженность различаются.
- Разработан протокол аннотирования, основанный на существующих принципах разметки устных корпусов, фиксирующий речевые сбои как потенциальные маркеры лжи и адаптированный для автоматического анализа.
- Корпус создан как открытый ресурс для дальнейшего расширения и исследований в области диагностики правдивости/ложности устной речи.

# Ограничения

# Список источников


1. Мамаев И. Д. АВТОМАТИЧЕСКАЯ Р АСШИФРОВКА ЗАПИСЕЙ УСТНОЙ РЕЧИ: ТЕСТИРОВАНИЕ ПРОГР АММЫ WHISPER1. – 2023.
2. 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.
3. 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.
4. 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.
5. Loy J. E., Rohde H., Corley M. Effects of disfluency in online interpretation of deception //Cognitive Science. – 2017. – Т. 41. – С. 1434-1456.
6. 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.
7. Кибрик А. и др. (ред.). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. – Litres, 2022.
8. Solà-Sales S. et al. Analysing deception in witness memory through linguistic styles in spontaneous language //Brain sciences. – 2023. – Т. 13. – №. 2. – С. 317.
9. Wieczorkowska A. Methodology for Obtaining High-Quality Speech Corpora //Applied Sciences. – 2025. – Т. 15. – №. 4. – С. 1848.

# Результаты

| Модель распознавания речи/вид ошибки | Замены, не влияющие на смысл (%) | Замены, искажающие смысл (%) | Замена на несуществующий звуковой аналог (%) | Замена на существующую лексему, не связанную с контекстом, но схожую по звуковому облику (%) |
|---|---|---|---|---|
| ***whisper-large-v3-turbo*** | 52% | 48% | 24% | 76% |
| ***nemo-fastconformer-ru-rnnt*** | 56% | 44% | 38% | 62% |
| ***gigaam-v2-rnnt*** | 40% | 60% | 25% | 75% |

Таблица 4. Распределение ошибок замены в транскриптах моделей автоматического распознавания речи (в условиях фоновых шумов)

# Результаты

**Для каждой модели было выявлнено общее количество ошибок (на незашумленных данных):**

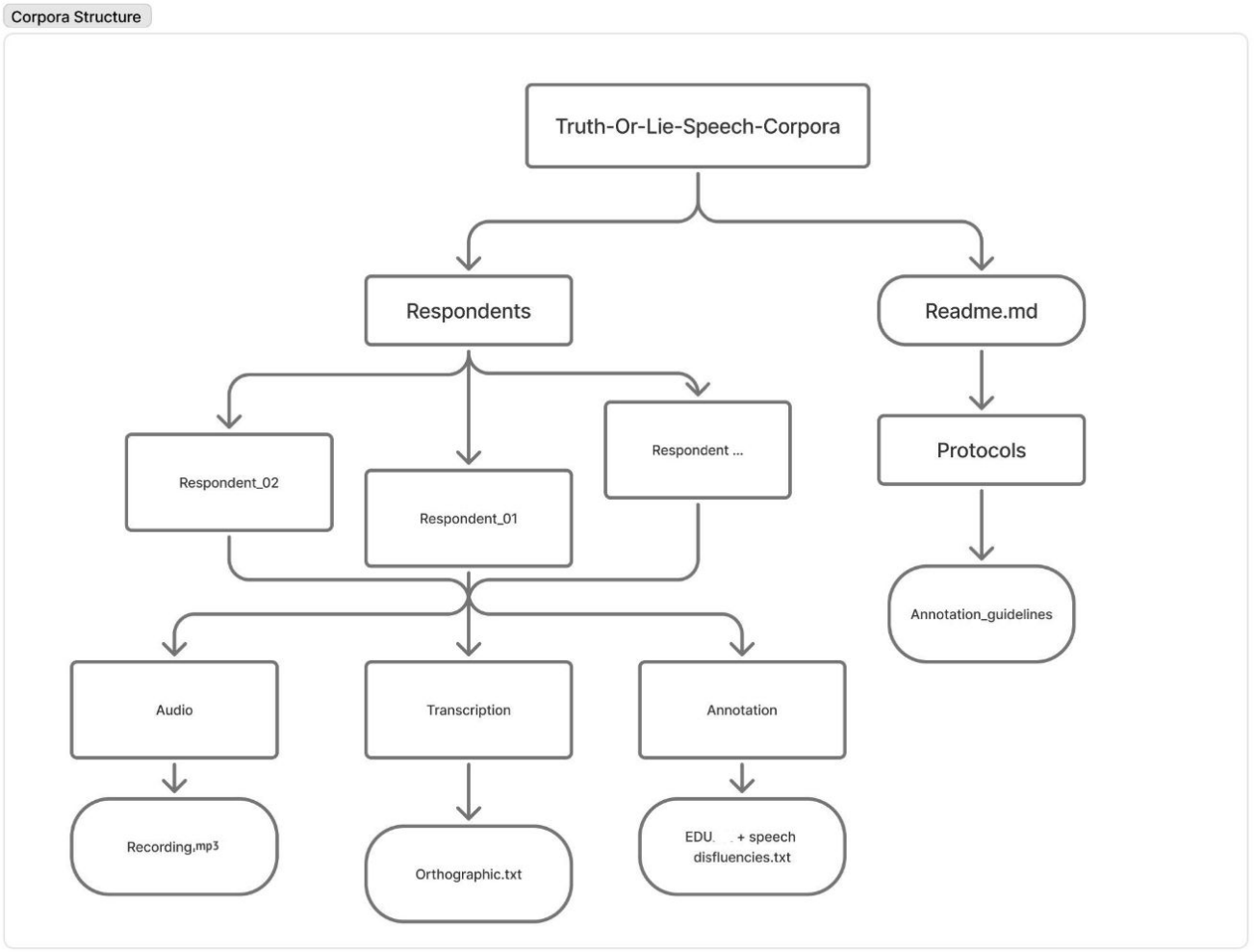
- ***whisper-large-v3-turbo – 87 ошибок***
- ***nemo-fastconformer-ru-rnnt – 140 ошибок***
- ***gigaam-v2-rnnt – 62 ошибки***

***Для зашумленных данных:***

**Подсчитано распределение ошибок по выделенным классам:**

| Модель распознавания речи/вид ошибки | Замены, не влияющие на смысл лексемы (%) | Замены, искажающие смысл лексемы (%) | Замена на несуществующий звуковой аналог (%) | Замена на существующую лексему, не связанную с контекстом, но схожую по звуковому облику (%) |
|---|---|---|---|---|
| ***whisper-large-v3-turbo*** | 39% | 61% | 28% | 72% |
| ***nemo-fastconformer-ru-rnnt*** | 49% | 51% | 31% | 69% |
| ***gigaam-v2-rnnt*** | 67% | 33% | 13% | 87% |

# Создание архитектуры корпуса



# Результаты

Результаты автоматической оценки качества транскриптов, созданных тремя ASR-алгоритмами в нормальных 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 | **2.13** |
| WIL (Word Information Lost) | 17.42 | 20.84 | **9.54** |
| WIP (Word Information Preserved) | 82.58 | 79.16 | **90.46** |

Таблица 1. Сравнение метрик качества автоматического распознавания речи для трёх моделей ASR в условиях без зашумлений

Оценка производилась на аудиозаписях общей длительностью 31 минуту 21 секунду (15% от общего кол-ва незашумленного материала)

# Результаты

Результаты автоматической оценки качества транскриптов, созданных тремя ASR-алгоритмами (зашумлённые условия):

| 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** |
| WIL (Word Information Lost) | 34.47 | 32.11 | **19.53** |
| WIP (Word Information Preserved) | 65.53 | 67.89 | **80.47** |

Таблица 2. Сравнение метрик качества автоматического распознавания речи для трёх моделей ASR в условиях с зашумлениями

Оценка проводилась на аудиозаписях из 6 аудиофайлов, в которых общая длительность монологической речи составила 24 минуты 16 секунд. Эта длительность соответствует всей монологической зашумленной речи, представленной в данном наборе данных (10.48% от общего материала).

# Промежуточные итоги

- Модель ***gigaam-v2-rnnt*** **продемонстрировала лучшие результаты** среди протестированных систем.
- **Автоматическая транскрипция** рассматривается как **предварительный этап** подготовки итоговых транскриптов корпуса.
- Наиболее обоснованным на текущем этапе развития технологий остаётся **гибридный подход** (assisted transcription) (Bazillon, Esteve, Luzzati, 2004), сочетающий скорость автоматического распознавания с точностью и глубиной ручной корректировки.

# Методы:

- **ASR:** автоматическая транскрипция с использованием моделей *whisper-large-v3-turbo*, *nemo-fastconformer-ru-rnnt* и *gigaam-v2-rnnt*.
- **Ручное создание транскриптов:** создание эталонных текстов для последующего сравнения с автоматическими транскриптами.
- **Сравнительный анализ:** сопоставление транскриптов, классификация и локализация ошибок.
- **Дискурсивная разметка:** разработка протокола аннотирования и структуры корпуса.