

УДК 004.934

## АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ: ФАКТОРЫ СНИЖЕНИЯ ТОЧНОСТИ ПРИ РАБОТЕ С РЕГИОНАЛЬНЫМ ВАРИАНТОМ РУССКОГО ЯЗЫКА (НА МАТЕРИАЛЕ ДАГЕСТАНА)

А. И. Кацнельсон<sup>1</sup> [0009-0009-2808-3243], О. Н. Ляшевская<sup>2</sup> [0000-0001-8374-423X]

<sup>1, 2</sup>Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия

<sup>1</sup>akatsnelson@hse.ru, <sup>2</sup>olesar@yandex.ru

### Аннотация

Работа посвящена сравнительному анализу качества автоматического распознавания стандартного русского языка и его дагестанского региолекта (дагестанского русского) – контактного идиома, формирующегося в условиях многоязычия Дагестана. Десять конфигураций ASR-систем (без дообучения), представляющих три архитектурных подхода (Conformer+RNNT, Conformer+CTC, Encoder-Decoder Transformer), протестированы на корпусах стандартного русского (GOLOS, RuDevices, примерно 20 и 90 тыс. высказываний) и дагестанского русского (DaGRuS, примерно 40 тыс. высказываний). Все модели демонстрируют падение метрики WER на 33–50% при переходе к дагестанскому русскому, что указывает на проблему сильного доменного сдвига. Анализ структуры ошибок выявил три основных механизма, вызывающих снижение WER: это русификация – систематическая замена дагестанских форм фонетически близкими стандартными русскими словами; удаление дискурсивных маркеров и OOV-лексики (этнонимов, топонимов); галлюцинации генеративных моделей на непонятных фрагментах. Установлено, что сквозные архитектуры-архитектуры деградируют слабее каскадных (33–35% против 40–42%), однако ни одна из моделей не достигает приемлемого качества (минимум WER около 43%). Сформулированы рекомендации для проектирования ASR-систем, устойчивых к региональной вариативности русского языка.

**Ключевые слова:** автоматическое распознавание речи, ASR, дагестанский региолект русского языка, региолект, региональные варианты речи, WER,

языковая вариативность, доменный сдвиг, *Conformer*, *Whisper*, *GigaAM*, галлюцинации.

## ВВЕДЕНИЕ

Современные системы автоматического распознавания речи (Automatic Speech Recognition, ASR) достигают значения WER (Word Error Rate, коэффициент ошибок на уровне слов) ниже 5% на эталонных корпусах стандартного русского языка. Русский язык существует не как единый стандарт, а как спектр региональных вариантов, формирующихся под влиянием контактных языков. По данным переписи 2020-2021 годов, более 20 млн жителей России говорят на языках коренных народов наряду с русским, что структурно влияет на их русскую речь.

Дагестан представляет собой крайний случай языкового разнообразия: в республике функционирует более 30 автохтонных языков. Русский выступает языком межнационального общения и подвержен влиянию на всех уровнях [1]. Дагестанский региолект русского языка (далее тире дагестанский, или региональный, русский) характеризуется специфическими консонантными фонетическими чертами, оригинальной лексикой (топонимы, этнонимы) и особыми дискурсивными структурами. В результате голосовые помощники и системы транскрипции работают для таких говорящих существенно хуже.

Целью настоящей работы было количественно оценить и лингвистически охарактеризовать разрыв в качестве ASR между стандартным и региональным русскими. Десять конфигураций ASR протестированы на корпусе DaGRuS; в качестве контрольных выбраны GOLOS и RuDevices. В качестве задач для поставленной цели рассмотрены следующие: выявление конкретных лингвистических контекстов деградации, анализ типов ошибок и соотнесение их с архитектурными свойствами моделей.

## СМЕЖНЫЕ РАБОТЫ

Устойчивость к региональной вариативности привлекает все большее внимание в литературе по ASR. В работе [2] продемонстрированы значительные различия WER для шотландского, индийского и нигерийского английских в коммерческих системах. В [3] показано, что дообучение на региональных данных снижает WER вплоть до 35% относительно исходного уровня. Для языков, отличных

от английского, многоязычные self-supervised-модели, протестированные на типологически разнообразных малоресурсных языках, сохраняют устойчивый разрыв в точности [4].

Для русского языка региональная вариативность в ASR остается практически неисследованной. В существующих бенчмарках, таких как GOLOS [5], преобладает стандартная речь. Корпус DaGRuS [6] представляет собой уникальную тестовую площадку для оценки устойчивости ASR к местным вариантам. В обзоре [7] отмечено, что архитектура декодера существенно определяет типы ошибок при сдвиге распределения, что согласуется с гипотезами, проверяемыми в настоящей работе.

Для смягчения таких доменных сдвигов в промышленных приложениях все чаще применяют федеративное обучение [8]. Однако анализ ошибок на диалектах показывает, что даже при использовании продвинутых методов по-прежнему трудно разработать единые модули нормализации, стабильно работающие на всех региональных вариантах.

## **ДАННЫЕ И МЕТОДИКИ**

### **Корпуса и лингвистический профиль**

Корпус DaGRuS [6] – это собрание спонтанных бытовых нарративов носителей дагестанских языков на русском, записанных в полевых условиях (39851 высказывание, 243 тыс. токенов<sup>1</sup>). Информанты – носители аварского, даргинского, кумыкского, лакского, лезгинского и других языков. Высказывания DaGRuS снабжены богатой ручной разметкой, включая технические события, неразборчивые фрагменты и нестандартные произносительные варианты с указанием канонических эквивалентов.

В качестве контрольных корпусов стандартного русского языка выступили GOLOS (части crowd и farfield) [5] и RuDevices (93241 высказывание). Для надежной оценки перед расчетом метрик применялись корпусно-специфичные конвейеры нормализации (разрешение разметки, стандартизация числительных и т. д.). Подробные описания нормализаторов представлены в Приложении А.

---

<sup>1</sup> Корпус DaGRuS доступен для исследовательских целей по запросу к авторам. Снимок корпуса, скрипты нормализации текста и конфигурации гиперпараметров инференса, использованные в настоящем исследовании, будут опубликованы в открытом репозитории.

---

В табл. 1 приведены основные лингвистические характеристики корпусов. DaGRuS демонстрирует «тяжелый хвост редкой лексики» (наклон Ципфа –1.175), представленный главным образом местными этнонимами и топонимами, отсутствующими в обучающих выборках ASR. Кроме того, в нем существенно выше доля частиц (13.0%) и междометий, что отражает спонтанный разговорный характер корпуса.

Табл. 1. Лингвистические характеристики корпусов.

| Характеристика       | GOLOS-crowd | GOLOS-far | RuDevices | DaGRuS  |
|----------------------|-------------|-----------|-----------|---------|
| Высказывания         | 9.896       | 1.915     | 93.241    | 39.851  |
| Токены               | 50.394      | 8.382     | 732.589   | 243.840 |
| Средняя длина (слов) | 5.09        | 4.38      | 7.86      | 6.12    |
| MSTTR                | 0.876       | 0.689     | 0.813     | 0.659   |
| Наклон Ципфа         | –0.857      | –0.747    | –1.047    | –1.175  |
| Частицы (%)          | 1.2         | 1.3       | 10.8      | 13.0    |
| Междометия (%)       | 0.2         | 0.5       | 2.0       | 2.1     |
| Местоимения (%)      | 3.7         | 2.3       | 10.4      | 8.5     |
| Гапаксы/словарь (%)  | 66.8        | 68.4      | 55.8      | 51.2    |
| 80% покрытие (леммы) | 1.189       | 525       | 970       | 356     |

### Предобработка и нормализация текста

Надежный расчет WER требует, чтобы эталонный текст и гипотеза модели были представлены в сопоставимом формате. Поскольку корпуса различаются конвенциями разметки, а модели ASR – форматом вывода, перед расчетом метрик применяют корпусно-специфичные конвейеры нормализации. Все нормализаторы имеют общую базу: приведение к нижнему регистру, унификация «ё → е» (большинство моделей ASR не различает эти символы), удаление пунктуации и схлопывание пробелов. Подробные описания корпусно-специфичных нормализаторов даны в Приложении А.

## Метрики оценки

Основная метрика – Word Error Rate (WER), определяемая как отношение суммарного числа операций редактирования на уровне слов (замены  $S$ , удаления  $D$ , вставки  $I$ ) к числу слов в эталоне:  $WER = (S + D + I)/N$ . Дополнительно используется Character Error Rate (CER), вычисляемая аналогично на уровне символов и менее чувствительная к выбору токенизации. WER разложена на три компонента: долю удалений, долю вставок и долю замен – для характеристики качественной структуры ошибок.

Статистическая значимость различия WER между корпусами оценивается с помощью U-критерия Манна – Уитни [9]. Доверительные интервалы для среднего WER получены бутстрэп-ресамплингом [10] (1000 итераций, 95%-ные процентильные интервалы).

## Модели ASR и гипотезы

Протестированы десять конфигураций на основе семи моделей (табл. 2). Выбор покрывает архитектурное разнообразие: Conformer+RNN-Transducer (авторегрессионный), Conformer+CTC (неавторегрессионный) и Encoder-Decoder Transformer (генеративный seq2seq). Представлены разные масштабы (от ~100 млн до 809 млн параметров) и режимы обучения (каскадные и сквозные (end-to-end)).

Табл. 2. Характеристики оцениваемых моделей ASR.

| Модель                 | Разработчик | Архитектура                  | Параметры, млн |
|------------------------|-------------|------------------------------|----------------|
| GigaAM-v2/v3 RNNT/CTC  | Sber        | Conformer + RNNT/CTC         | ~600           |
| GigaAM-v3-e2e RNNT/CTC | Sber        | E2E Conformer, joint CTC/Att | ~600           |
| Whisper-large-v3       | OpenAI      | Enc-Dec Transformer          | ~809           |
| T-One                  | T-Bank      | Conformer + Transducer       | ~100           |

Все модели оценивались «из коробки» в режиме zero-shot, без какого-либо целевого дообучения. Для Whisper-large-v3 (openai/whisper-large-v3) принудительно задавался языковой токен русского языка и использовалось «жадное декодирование» по умолчанию (beam size 1, temperature 0.0) для обеспечения детерминированности вывода.

Проверялись три гипотезы.

**H1 (влияние декодера):** RNN-T склонен к удалениям на входах с низкой уверенностью, CTC – к заменам, а seq2seq «галлюцинирует».

**H2 (совместный E2E и каскадные AM+LM):** каскадными мы называем модели с отдельным декодированием акустической и языковой моделей (например, стандартный GigaAM RNN-T), в отличие от чисто совместных сквозных архитектур. Мы предполагаем, что совместные E2E-модели формируют более гибкие представления, слабее привязанные к стандартным фонемам, и тем самым смягчают деградацию.

**H3 (доменная генерализация):** узкодоменные модели (например, T-One) деградируют сильнее, чем модели широкого домена.

## **РЕЗУЛЬТАТЫ: СТАНДАРТНЫЙ И ДАГЕСТАНСКИЙ РУССКИЙ**

### **Общая картина: градиент деградации**

На рис. 1 представлены результаты по четырем корпусам. Наборы данных можно рассматривать как последовательность: стандартная речь (GOLOS-crowd), стандартная речь в шумных условиях (GOLOS-farfield), спонтанная стандартная речь (RuDevices) и спонтанная речь местных вариантов (DaGRuS). Хотя WER растет на каждом шаге, переход к дагестанскому русскому дает резкий скачок, непропорциональный предшествующим приростам.

Все восемь моделей, протестированных и на GOLOS-crowd, и на DaGRuS, демонстрируют статистически значимую деградацию (критерий Манна – Уитни,  $p < 0.001$  для всех пар; размер эффекта  $r = 0.52 - 0.76$ ). Доверительные интервалы (95%-ный бутстрэп) не пересекаются ни для одной пары.

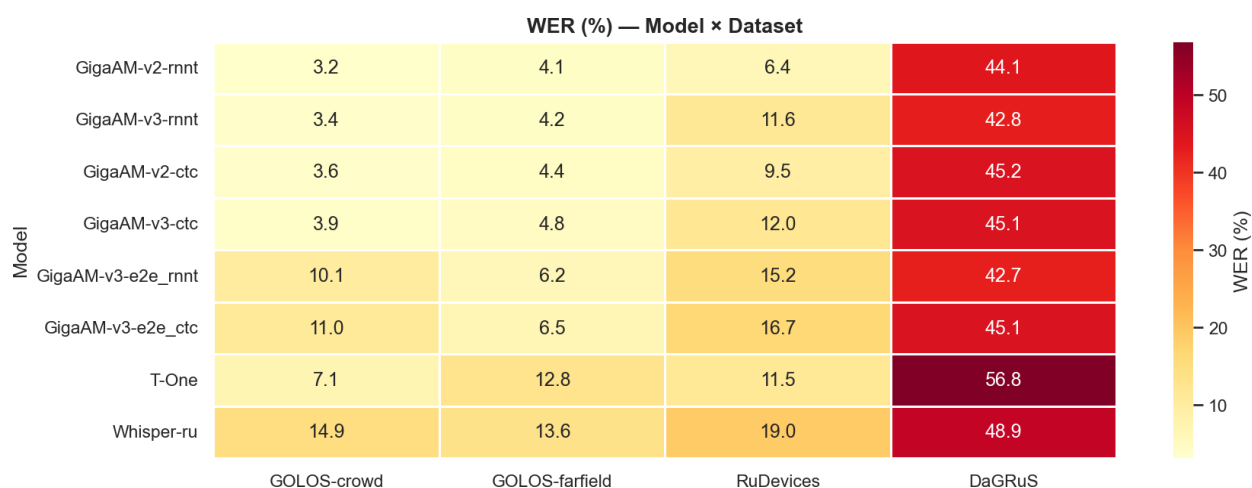


Рис. 1. Матрица WER (%) по корпусам и моделям. Прочерки обозначают отсутствие данных.

Переход от GOLOS-crowd к RuDevices (оба корпуса представляют стандартный русский, но последний – спонтанная речь, записанная на различных устройствах) дает умеренный рост WER: +6–7% для каскадных моделей GigaAM, +4–6% для e2e-моделей и +4% для Whisper. Однако переход от RuDevices к DaGRuS вызывает дополнительный скачок еще на 28–36%. Таким образом, даже с учетом эффектов спонтанности и шума основным фактором падения качества – является региональный фонетический компонент дагестанского русского.

Принципиально важно, что ранжирование моделей заметно изменяется в зависимости от набора данных. На GOLOS-crowd лучшие результаты показали каскадные GigaAM v2/v3 (WER 3–4%), на RuDevices лидирует GigaAM-v2-ctc (9.5%), а на DaGRuS – GigaAM-v3-e2e RNN-T (42.7%). Коэффициент ранговой корреляции Спирмена между WER на GOLOS и DaGRuS для восьми моделей равен 0.12 ( $p > 0.05$ ): качество на стандартном русском не является предиктором качества на дагестанском русском.

Более высокий WER моделей GigaAM-v3 E2E на GOLOS-crowd (10–11% против 3–4% у каскадных версий) объясняется встроенной текстовой нормализацией и пунктуацией: эти модели выводят числительные словами и вставляют знаки препинания (запятые и точки), что создает рассогласование с форматом эталона. Однако на DaGRuS, где речь спонтанна, а числительные естественно произносятся словами, этот эффект компенсируется, и E2E-модели демонстрируют наименьший абсолютный WER (42.7–45.1%).

### Эволюция структуры ошибок по корпусам

Разложение ошибок на удаления (Del), вставки (Ins) и замены (Sub) выявляет качественный сдвиг в характере ошибок при переходе к дагестанскому русскому (табл. 3).

Табл. 3. Профиль ошибок (Del/Ins/Sub) в процентах от числа слов эталона по корпусам.

| Модель              | GOLOS-crowd |      |      | RuDevices |      |      | DaGRuS |      |       |
|---------------------|-------------|------|------|-----------|------|------|--------|------|-------|
|                     | Del         | Ins  | Sub  | Del       | Ins  | Sub  | Del    | Ins  | Sub   |
| GigaAM-v3-e2e RNN-T | 1.49        | 0.69 | 5.42 | 1.98      | 2.36 | 6.60 | 8.11   | 3.60 | 10.17 |
| GigaAM-v3 RNN-T     | 0.28        | 0.19 | 1.79 | 2.23      | 1.36 | 5.37 | 8.58   | 3.24 | 9.87  |
| GigaAM-v3 CTC       | 0.27        | 0.23 | 2.10 | 2.00      | 1.49 | 5.78 | 7.70   | 3.48 | 12.81 |
| GigaAM-v2 CTC       | 0.25        | 0.21 | 1.88 | 1.54      | 1.04 | 4.92 | 7.90   | 2.51 | 11.59 |
| T-One               | 0.45        | 0.63 | 3.51 | –         | –    | –    | 11.96  | 2.78 | 15.29 |
| Whisper-ru          | 1.22        | 1.85 | 7.22 | 2.86      | 2.21 | 8.23 | 6.95   | 4.18 | 14.70 |

На GOLOS-crowd удаления минимальны (0.2–1.5%), доминируют замены. На RuDevices удаления возрастают до 1.5–2.9%: модели начинают пропускать слова в спонтанной речи. На DaGRuS удаления скачкообразно растут до 7–12%: модели систематически «молчат» на незнакомых фрагментах. Замены также увеличиваются в два-три раза (с 2–7% до 10–15%). Это означает, что на дагестанском русском модели не просто ошибаются чаще, они ошибаются качественно иначе.

Профиль ошибок на RuDevices сохраняет пропорции, наблюдаемые на GOLOS (доминирование замен, умеренные удаления). На DaGRuS эти пропорции нарушаются: доля удалений в общем числе ошибок возрастает с 10–15% до 30–40% для большинства моделей. Это говорит о том, что региональная фонетика создает принципиально иной тип трудности по сравнению с обычной спонтанной речью.

## АНАЛИЗ ОШИБОК

### Части речи

Анализ ошибок по частям речи (ЧР), размеченным с помощью rymorphy2 [11], позволяет локализовать конкретные зоны падения качества. В табл. 4 приведены доли неверно распознанных слов для двух моделей: GigaAM-v3-e2e RNN-T и Whisper-ru – на корпусах GOLOS-crowd и DaGRuS.

Табл. 4. Доля ошибок (%) по частям речи: GOLOS-crowd → DaGRuS.

| ЧР   | Описание        | GigaAM GOL | GigaAM DaG | Whis. GOL | Whis. DaG |
|------|-----------------|------------|------------|-----------|-----------|
| VERB | Глагол          | 4.2        | 54.3       | 11.6      | 59.4      |
| PREP | Предлог         | 6.5        | 47.3       | 12.1      | 48.6      |
| INFN | Инфинитив       | 3.3        | 49.3       | 13.3      | 55.6      |
| NOUN | Существительное | 30.3       | 63.0       | 31.5      | 64.7      |
| NUMR | Числительное    | 37.4       | 56.6       | 24.6      | 81.7      |
| INTJ | Междометие      | 33.3       | 62.3       | 24.1      | 65.4      |
| NPRO | Местоимение     | 2.2        | 38.4       | 10.2      | 43.0      |
| UNKN | Неизв.          | 84.9       | 97.6       | 81.2      | 95.4      |

Наиболее выраженная деградация наблюдается у глаголов (+50% у GigaAM, +48% у Whisper), инфинитивов и предлогов. Глаголы в дагестанском русском часто демонстрируют нестандартное управление и морфологию из-за интерференции родного языка (L1) говорящих [12]. Существительные, уже имеющие высокую базовую долю ошибок на GOLOS (~30%), достигают 63–65% на DaGRuS; эту категорию преимущественно наполняют внесловарные (Out-Of-Vocabulary, OOV) этнонимы и топонимы.

Особенно показателен случай числительных: у GigaAM-e2e ошибки растут лишь на +19%, тогда как у Whisper – на +57%. Причина состоит в архитектурной особенности: при декодировании Whisper предпочитает цифровую запись (например, «1961» вместо «тысяча девятьсот шестьдесят первого»). Напротив,

модели GigaAM E2E благодаря встроенной нормализации выводят числительные словами, что соответствует формату эталона DaGRuS.

Категория UNKN – слова, не распознанные морфологическим анализатором, – дает 95–98% ошибок на DaGRuS у всех моделей. Это прежде всего дагестанизмы, не имеющие русских эквивалентов: топонимы, этнонимы и бытовая лексика кавказского происхождения.

### **«Русификация» как основной паттерн замен**

Анализ конкретных замен выявляет систематический паттерн, который мы называем «русификацией»: модели заменяют дагестанские формы фонетически ближайшими стандартными русскими словами. Это предсказуемая стратегия декодера, обученного на стандартном русском: столкнувшись с незнакомой фонетической последовательностью, он ищет ближайшее соответствие в пространстве известной лексики. Иллюстративные примеры приведены в табл. 5.

Табл. 5. Примеры «русификации»: эталон и гипотезы двух моделей. Символ ✓ обозначает верную транскрипцию.

| Эталон                  | GigaAM-v3-e2e RNN-T    | Whisper-ru        |
|-------------------------|------------------------|-------------------|
| ...к грузинам           | к грузинам ✓           | к кукурузинам     |
| ...относились к кумуху  | к муху                 | к кумуху ✓        |
| наше село               | наше село ✓            | наши силу         |
| ...через село и вниз    | через село и вниз ✓    | через силу и вниз |
| ...наших овец ведь мало | наших овец ведь мало ✓ | наших везде мало  |

Русификация проявляется по-разному в зависимости от архитектуры. Как генеративная модель, Whisper активно «достраивает» последовательность до правдоподобного текста, подбирая фонетически близкие русские слова (*село* → *силу*, *овец* → *везде*, *грузинам* → *кукурузинам*). GigaAM-e2e, использующий символичный/ВРЕ-словарь, чаще выдает приблизительные незнакомые формы (например, *кумуху* → *муху*), что снижает семантическое искажение, но также

прибегает к «русификации» (*коршинцы* → *корчинцы*). Обе стратегии ведут к потере смысла, однако генеративный подход порождает особенно опасные «правдоподобные» ошибки (галлюцинации).

### **Дискурсивные маркеры как источник ошибок**

Маркеры обратной связи (*угу, ага, мм, ээ*) составляют около 15% токенов DaGRuS, но практически отсутствуют в обучающих данных стандартных корпусов. В результате модели справляются с ними плохо: например, слово *угу* (2668 вхождений) имеет долю ошибок свыше 50% у Whisper. Ключевой проблемой является сочетание этих маркеров с региональной фонетикой: стандартное русское *угу* артикулируется иначе, чем его дагестанский аналог. Кроме того, оценка этих ошибок осложняется неоднозначностью акустической реализации и непоследовательностью корпусной разметки. Подробный разбор наиболее частых удалений маркеров и проблем разметки приведен в Приложении Б.

### **Сегменты с нулевой точностью: где не справляется ни одна модель**

«Слепые зоны» – это высказывания, на которых все десять моделей дают WER выше 50%, – составили 23105 примеров (58% DaGRuS). Характерная лексика этих сегментов включает такие единицы, как рикванинский, лезгины, тюркский, ногойцы, Янгикент, табасаранский, полулак, Джамал, Тажутдин. Это дагестанские этнонимы, топонимы и антропонимы, для которых в обучающих данных моделей нет ни акустических, ни языковых соответствий.

На DaGRuS наблюдается отрицательная корреляция между длиной высказывания и WER ( $\rho$  Спирмена  $\approx -0.25$ ,  $p < 0.001$  для всех моделей), тогда как на GOLOS-crowd корреляция близка к нулю, а на RuDevices – слабо положительная (см. рис. 2). Этот «парадокс длины» объясняется природой коротких реплик в DaGRuS: высказывания из 1–3 слов часто представляют собой изолированные дагестанизмы, топонимы или маркеры обратной связи, где одна ошибка дает WER 100%. В более длинных фразах выше доля стандартной русской лексики, которая «разбавляет» общую долю ошибок.

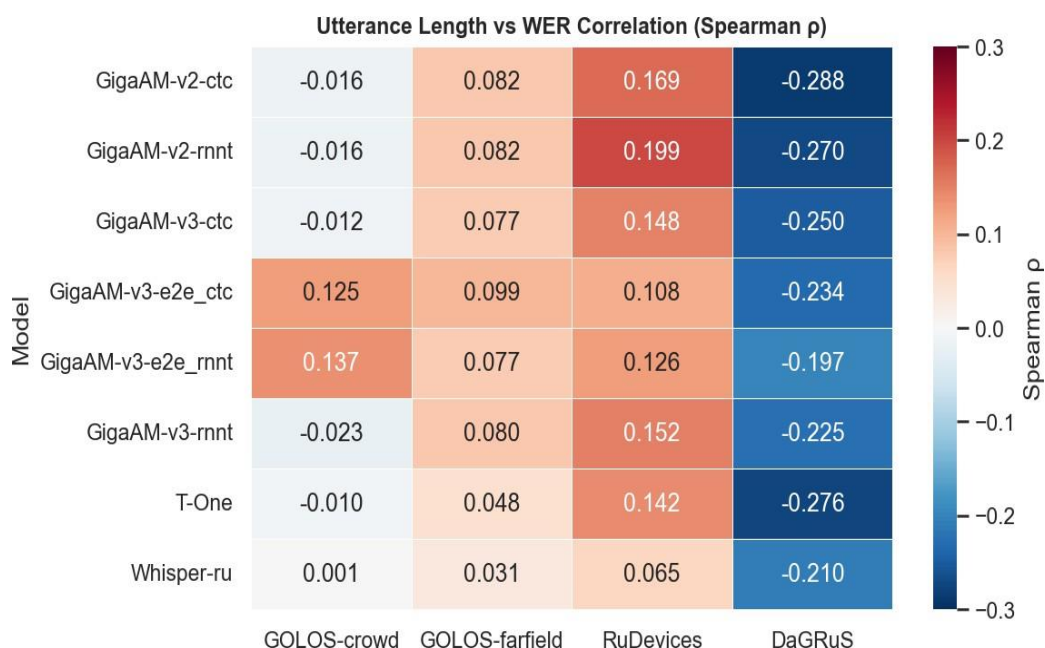


Рис. 2. Корреляция Спирмена между длиной высказывания и WER по моделям и корпусам.

### Стратегические контрасты: расхождение архитектур

Сравнение гипотез, порожденных разными моделями на одних и тех же примерах, выявляет контрастные стратегии обработки неразборчивого акустического входа. Расхождения наиболее выражены между семейством Conformer (GigaAM) и генеративным Whisper. GigaAM точнее работает с числительными (записывая их словами), топонимами и дискурсивными маркерами. Whisper, напротив, лучше справляется с контекстной реконструкцией в длинных высказываниях, состоящих из стандартной лексики.

Для систематического анализа этих ошибок необходимо различать фонетические замены и истинные галлюцинации. Следуя концептуальной рамке, предложенной для генеративных моделей [13], заменой мы считаем ошибочный вывод, фонетически мотивированный акустическим входом (например, замена *грузинам* на *кукурузинам*). Галлюцинацией ASR мы считаем порождение правдоподобно звучащего текста или повторяющихся циклов, которым во входном сигнале вообще не соответствуют акустические события.

Примечательной особенностью Whisper является окно обработки по умолчанию в 30 с, обеспечивающее широкий контекст. Это способствует контекстной реконструкции в длинных высказываниях со стандартной лексикой, однако на

неразборчивых региональных фрагментах генеративный декодер склонен к галлюцинациям. Особенно показательны следующие расхождения:

- *наши спускались в алазанскую долину к грузинам* – GigaAM: верно; Whisper: *к кукурузинам*;
- *вскоре после прихода Советской власти мы относились к кумуху* – Whisper: верно; GigaAM: *к муху*;
- *около трехсот пятисот овец* – GigaAM: словесная запись верна; Whisper: 300500;
- *Отгоняли* – GigaAM: верно; Whisper: *«отгорели»* – фонетическая, но семантически отклоняющаяся замена.

Паттерны вставок также существенно различаются. На DaGRuS модель GigaAM-e2e часто вставляет дискурсивные биграммы вроде *да да* (483 вхождения) или *э э* (393 вхождения), по сути удваивая уже имеющиеся маркеры. Whisper, напротив, порождает повторяющиеся числовые последовательности (например, 1988 1988 встретилось почти 17000 раз) и шаблонные фразы вроде *продолжение следует*, характерные для зацикливания seq2seq-декодера. Отсутствие этих паттернов в GOLOS и RuDevices подтверждает, что они запускаются именно неразборчивым акустическим входом.

Эти паттерны вставок не случайны: они, по-видимому, восходят к артефактам обучающих данных моделей. В обучающей выборке GigaAM, вероятно, много разговорных данных с *да да*, тогда как Whisper мог обучаться на транскриптах с повторяющимися датами или техническими заглушками.

## ОБСУЖДЕНИЕ

Разрыв в качестве между стандартным и дагестанским русскими (33–50%) существенно превышает разброс между контрастными архитектурами, оцененными на одном и том же корпусе. Даже наиболее устойчивая конфигурация (GigaAM v3-e2e) останавливается на WER 42.7% на дагестанском русском: одних лишь архитектурных оптимизаций недостаточно, чтобы решить задачу региональной языковой вариативности.

Полученные результаты указывают на ряд необходимых шагов по повышению устойчивости ASR. Во-первых, обучающие выборки необходимо расши-

рить тремя категориями, сейчас отсутствующими в основных корпусах: региональной фонетикой (акцентное произношение стандартных русских слов); региональной лексикой (этнонимы, топонимы и бытовые заимствования из кавказских и тюркских языков); дискурсивными маркерами спонтанной речи (сигналы обратной связи, хезитации и заполненные паузы). Во-вторых, предпочтительны сквозные (E2E) архитектуры: они дают снижение деградации на 7–8% по сравнению с каскадными системами на том же энкодере. Символьные или BPE-словари превосходят словарные (на уровне слов), поскольку позволяют модели частично транскрибировать незнакомые формы, а не удалять их целиком. В-третьих, генеративные (seq2seq) модели при обработке нестандартной речи требуют специальных мер против галлюцинаций – контроля температуры или штрафов за повторы. В-четвертых, узкодоменные модели (телефония, голосовые команды) критически уязвимы к речи местных вариантов; перед развертыванием в региональных контекстах, например в телефонной аналитике на Северном Кавказе, им необходима устойчивая доменная адаптация.

Отметим, что «приемлемое качество» зависит от приложения. Однако для большинства практических задач – голосовых помощников или телефонной аналитики – WER выше 15–20% делает систему практически непригодной [2]. Наблюдаемый здесь минимум WER в 43% явно выходит за границы практической полезности.

У работы есть ряд ограничений. Во-первых, DaGRuS представляет лишь один региональный вариант русского; полученные выводы не обязательно переносятся на другие контактно обусловленные разновидности (например, среднеазиатский, недагестанский кавказский или сибирский русский). Кроме того, GOLOS состоит преимущественно из прочитанной речи, тогда как DaGRuS содержит спонтанные диалоги. Часть наблюдаемой деградации может быть связана со стилистическими различиями, а не только с акцентом. RuDevices (стандартная спонтанная речь) дает промежуточные значения WER (9–19%), что помогает оценить стилистический вклад, однако строгое разделение этих факторов требует дополнительного контролируемого эксперимента. Во-вторых, наш конвейер нормализации намеренно удаляет нерусские сегменты (например, аварские или даргинские слова). Анализ того, как системы ASR опознают переключение кодов

или галлюцинируют на таких фрагментах, является важной задачей, однако мы сосредоточились на изолировании фонетического сдвига от стандартного к региональному русскому. В-третьих, проверка точного эффекта декодера путем сравнения моделей с одинаковыми энкодерами, но с разными декодерами (например, Whisper v3 и Whisper Turbo) выходила за рамки настоящей работы и представляет перспективное направление дальнейших исследований. Наконец, в отношении дискурсивных маркеров (Приложение Б) предварительный анализ показал, что флаг `remove_fillers` снижает базовый WER на всех корпусах, однако пропорциональный разрыв между стандартным и региональным русским остается тем же.

## **ЗАКЛЮЧЕНИЕ**

Все десять протестированных конфигураций ASR демонстрировали значительное падение качества на дагестанском русском: WER возрастает на 33–50% по сравнению со стандартным русским. Ни одна модель не достигла значения WER ниже 43%. Различия статистически значимы при большом размере эффекта ( $r = 0.52 - 0.76$ ).

Ошибки моделей являются неслучайными; выявлены три систематических механизма сбоев: «русификация» (замена дагестанских форм фонетически близкими стандартными русскими словами), удаление дискурсивных маркеров и внесловарной лексики, а также генеративные галлюцинации на неразборчивых фрагментах. Примечательно, что 58% высказываний DaGRuS попадают в «слепую зону», где все модели дают WER выше 50%.

Архитектура декодера определяет тип ошибки (RNN-T склонен к удалению, CTC – к заменам, seq2seq – к галлюцинациям), но не снижает общий масштаб проблемы. Сквозное обучение уменьшает деградацию на 7–8% по сравнению с каскадным подходом, а узкодоменные модели оказываются наиболее уязвимыми (например, +50% у T-One). Однако главным фактором выступает отсутствие региональных данных в обучающих выборках: разрыв между корпусами в несколько раз превышает разрыв между архитектурами.

Чтобы обеспечить инклюзивность цифровых сервисов, обучающие данные ASR необходимо целенаправленно расширять региональными вариантами рус-

ского языка, маркерами спонтанной речи и методами доменной адаптации к региональной фонетике. Корпус DaGRuS может служить бенчмарком для оценки устойчивости ASR-систем к региональной языковой вариативности.

### СПИСОК ЛИТЕРАТУРЫ

1. Добрушина Н.Р. Многоязычие в Дагестане, или зачем человеку три языка // Социологический журнал. 2007. № 1. С. 103–127.
2. Markl N. Language variation and algorithmic bias: Understanding algorithmic biases in British English automatic speech recognition // Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. P. 521–534.
3. Feng S., Kudina O., Halpern B.M., Scharenborg O. Quantifying bias in automatic speech recognition // arXiv:2103.15122. <https://doi.org/10.48550/arXiv.2103.15122>
4. Conneau A., Baevski A., Collobert R., Mohamed A., Auli M. Unsupervised cross-lingual representation learning for speech recognition // Proceedings of Interspeech 2021. 2021. P. 2426–2430.
5. Karpov N., Denisenko A., Minkin F. Golos: Russian dataset for speech research // Proceedings of Interspeech 2021. 2021. P. 1419–1423.
6. Dobrushina N., Daniel M., von Waldenfels R., Maisak T., Panova A. Corpus of Russian spoken in Daghestan. 2018. URL: <http://www.parasolcorpus.org/dagrus/> (дата обращения: 24.02.2026).
7. Hinsvark A., Delworth N., Del Rio M. et al. Accented speech recognition: A survey. 2021 // arXiv:2104.10747. 2021. <https://doi.org/10.48550/arXiv.2104.10747>
8. Ramesh G.V. et al. Federated representation learning for automatic speech recognition // Proc. 3rd Symposium on Security and Privacy in Speech Communication. 2023. P. 29–33. <https://doi.org/10.21437/SPSC.2023-5>
9. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other // The Annals of Mathematical Statistics. 1947. Vol. 18, No. 1. P. 50–60.
10. Efron B. Bootstrap methods: Another look at the jackknife // The Annals of Statistics. 1979. Vol. 7, No. 1. P. 1–26.

11. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages // *Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science*. Vol. 542. Cham: Springer, 2015. P. 320–332.

12. *Naccarato C., Panova A., Stoyanova N.* Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan // *Language Variation and Change*. 2021. Vol. 33, No. 3. P. 387–411.

13. *Guerreiro N.M., Voita E., Martins A.F.T.* Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation // *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023.

---

#### **ПРИЛОЖЕНИЕ А. КОНВЕЙЕРЫ НОРМАЛИЗАЦИИ ТЕКСТА**

DagrusNormalizer. Эталонный текст DaGRuS содержит полевую разметку, которую необходимо разрешить перед оценкой. Нормализатор работает следующим образом. (1) Маркеры технических событий, заключенные в знаки решетки (`#laughs#`, `#speaks another language#`), удаляются. (2) Аннотированные пары «слово{каноническая\*}» разрешаются к каноническому варианту (например, «`щас{сейчас*}`» → «сейчас»). Аннотации с маркером «%» (нестандартные формы без строгого канонического эквивалента) снимаются, сохраняется только поверхностная форма. (3) Отдельные маркеры неразборчивости «нрзб» удаляются. (4) Непомеченные технические фразы, случайно оставленные аннотаторами без знаков решетки (например, «говорит на другом языке»), удаляются точным сопоставлением со закрытым списком.

Для гипотез ASR, не содержащих разметки, нормализатор применяет отображение разговорных форм на канонические: словарь из 30 однозначных пар («щас → сейчас», «че → что», «тыща → тысяча») обеспечивает, что и произнесенная, и каноническая формы считаются эквивалентными при выравнивании. Например, если в эталоне стоит сейчас (разрешенное из разметки «`щас{сейчас*}`»), а модель выдает «щас», отображение преобразует гипотезу в сейчас, и сопоставление считается верным.

GolosNormalizer. В эталонных текстах GOLOS все числительные записаны словами (например, двадцать пять), тогда как некоторые модели ASR, в частности Whisper, выводят арабские цифры (25). Без нормализации каждое такое числительное засчитывается как ошибка. GolosNormalizer переводит цифры в гипотезе в русские слова с помощью библиотеки num2words, после чего применяется стандартная очистка.

SimpleNormalizer. Для RuDevices и других корпусов без особых конвенций разметки базовый нормализатор применяет только общий конвейер: нижний регистр, ё → е, удаление пунктуации и схлопывание пробелов.

Важное проектное решение касается дискурсивных заполнителей: «угу», «ага», «мм», «э». По умолчанию мы их не удаляем: если модель удаляет заполнитель, присутствующий в эталоне, это регистрируется как ошибка удаления, что лингвистически информативно. Для абляционных экспериментов предусмотрен необязательный флаг `remove_fillers`.

## **ПРИЛОЖЕНИЕ Б. ПРОФИЛЬ ОШИБОК НА ДИСКУРСИВНЫХ МАРКЕРАХ И ПРОБЛЕМЫ РАЗМЕТКИ**

Показатели WER для дискурсивных маркеров в DaGRuS несут неустранимую измерительную неопределенность. Разметка сигналов обратной связи в корпусе непоследовательна: одни аннотаторы систематически транскрибировали бэк-канальные сигналы (угу, мм), другие опускали их или включали в соседние реплики. Если маркер отсутствует в эталоне, но есть в аудио и порожден моделью, метрика фиксирует ложную вставку; напротив, если маркер есть в эталоне, но едва ли был отдельным речевым актом, его удаление завышает счетчик удалений. Граница между заполненной паузой и сигналом обратной связи часто неоднозначна даже для слушателя-человека.

Несмотря на эти проблемы разметки, для ряда биграмм наблюдается явная тенденция к систематическим удалениям. Наиболее часто удаляемые биграммы показывают склонность моделей отбрасывать повторенные дискурсивные заполнители:

- *да да* – 366 удалений у Whisper, 297 у GigaAM;
- *м м* – 266/234;
- *угу угу* – 141/95.

Таким образом, приводимые доли ошибок на дискурсивных маркерах представляют верхнюю оценку сбоев моделей и нижнюю оценку истинной трудности явления.

---

## AUTOMATIC SPEECH RECOGNITION: FACTORS REDUCING ACCURACY WHEN WORKING WITH A REGIONAL VARIANT OF THE RUSSIAN LANGUAGE (USING MATERIAL FROM DAGESTAN)

A. I. Katsnelson<sup>1</sup> [0009-0009-2808-3243], O. N. Lyashevskaya<sup>2</sup> [0000-0001-8374-423X]

<sup>1, 2</sup>*HSE University, Moscow, Russia*

<sup>1</sup>akatsnelson@hse.ru, <sup>2</sup>olesar@yandex.ru

### **Abstract**

We evaluate the quality of automatic speech recognition (ASR) for Standard Russian and its Dagestani regional variety – a contact idiom formed in the multilingual environment of Dagestan. We tested ten off-the-shelf ASR configurations in a zero-shot setting (without fine-tuning) representing three architectural approaches (Conformer+RNN-T, Conformer+CTC, and Encoder-Decoder Transformer) on datasets of Standard Russian (GOLOS, RuDevices; approx. 20K and 90K utterances respectively) and Dagestani Russian (DaGRuS; approx. 40K utterances). All models demonstrate a word error rate (WER) degradation of 33%–50% when evaluated on Dagestani Russian, highlighting a severe domain shift problem. Error structure analysis reveals three primary degradation mechanisms: “Russification” (systematic replacement of Dagestani forms with phonetically similar Standard Russian words); deletion of discourse markers and out-of-vocabulary (OOV) terms such as ethnonyms and toponyms; and generative model hallucinations on incomprehensible audio segments. We found that end-to-end architectures suffer less degradation than cascaded ones (33%–35% vs. 40%–42%), yet no model achieves acceptable performance (minimum WER  $\approx$  43%). Finally, we formulate recommendations for designing ASR systems robust to the regional variability of the Russian language.

**Keywords:** *automatic speech recognition, ASR, Dagestani Russian, regiolect, speech of local varieties, WER, language variation, domain shift, Conformer, Whisper, GigaAM, hallucinations.*

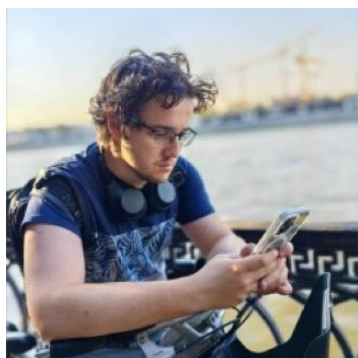
## REFERENCES

1. *Dobrushina N.R.* Mnogoyazychie v Dagestane, ili zachem cheloveku tri yazyka // *Sotsiologicheskii zhurnal*. 2007. No. 1. P. 103–127.
2. *Markl N.* Language variation and algorithmic bias: Understanding algorithmic biases in British English automatic speech recognition // *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022. P. 521–534.
3. *Feng S., Kudina O., Halpern B.M., Scharenborg O.* Quantifying bias in automatic speech recognition // *arXiv:2103.15122*. <https://doi.org/10.48550/arXiv.2103>
4. *Conneau A., Baevski A., Collobert R., Mohamed A., Auli M.* Unsupervised cross-lingual representation learning for speech recognition // *Proceedings of Interspeech 2021*. 2021. P. 2426–2430.
5. *Karpov N., Denisenko A., Minkin F.* Golos: Russian dataset for speech research // *Proceedings of Interspeech 2021*. 2021. P. 1419–1423.
6. *Dobrushina N., Daniel M., von Waldenfels R., Maisak T., Panova A.* Corpus of Russian spoken in Daghestan. 2018. URL: <http://www.parasolcorpus.org/dagrus/> (accessed: 24.02.2026).
7. *Hinsvark A. et al.* Accented speech recognition: A survey // *arXiv:2104.10747*. 2021. <https://doi.org/10.48550/arXiv.2104.10747>
8. *Ramesh G.V. et al.* Federated representation learning for automatic speech recognition // *Proc. 3rd Symposium on Security and Privacy in Speech Communication*. 2023. P. 29–33. <https://doi.org/10.21437/SPSC.2023-5>
9. *Mann H.B., Whitney D.R.* On a test of whether one of two random variables is stochastically larger than the other // *The Annals of Mathematical Statistics*. 1947. Vol. 18, No. 1. P. 50–60.
10. *Efron B.* Bootstrap methods: Another look at the jackknife // *The Annals of Statistics*. 1979. Vol. 7, No. 1. P. 1–26.
11. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages // *Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science*. Vol. 542. Cham: Springer, 2015. P. 320–332.
12. *Naccarato C., Panova A., Stoyanova N.* Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan // *Language Variation and Change*. 2021. Vol. 33, No. 3. P. 387–411.

13. *Guerreiro N.M., Voita E., Martins A.F.T.* Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.

---

## СВЕДЕНИЯ ОБ АВТОРАХ



**КАЦНЕЛЬСОН Артем Игоревич** – ML-разработчик, преподаватель Департамента компьютерной инженерии Национального исследовательского университета «Высшая школа экономики» (МИЭМ им. А. Н. Тихонова). Область научных интересов: прикладные системы искусственного интеллекта, мультиагентные системы, RAG-архитектуры, речевая аналитика, машинное зрение, прогнозирование временных рядов.

**Artem Igorevich KATSNELSON** is a ML developer, lecturer at the Department of Computer Engineering at the National Research University Higher School of Economics (Tikhonov MIEM). He has more than 5 years of experience in the development and implementation of commercial machine learning and AI systems. Research interests: applied artificial intelligence systems, multi-agent systems, RAG architectures, speech analytics, machine vision, time series forecasting. Author of more than 10 scientific publications. Two-time winner of the HSE Electronics Olympiad.

email: [akatsnelson@hse.ru](mailto:akatsnelson@hse.ru)

ORCID: <https://orcid.org/0009-0009-2808-3243>



**ЛЯШЕВСКАЯ Ольга Николаевна** – профессор Школы лингвистики факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики», ведущий научный сотрудник Института русского языка им. В. В. Виноградова РАН, кандидат филологических наук. Область научных интересов: компьютерная и корпусная лингвистика, лексическая и грамматическая семантика, грамматика конструкций, компьютерная лексикография. Активно участвует в развитии Национального корпуса русского языка.

**Olga Nikolayevna LYASHEVSKAYA** is Professor at the School of Linguistics, Faculty of Humanities, National Research University Higher School of Economics, leading research fellow at Vinogradov Russian Language Institute of the Russian Academy of Sciences, Candidate of Philological Sciences.

Her research interests include computational and corpus linguistics, lexical and grammatical semantics, construction grammar, and computational lexicography. She actively participates in the development of the National Corpus of the Russian Language.

email: olesar@yandex.ru

ORCID: <https://orcid.org/0000-0001-8374-423X>

*Материал поступил в редакцию 19 августа 2026 года*