

КОММУНИКАТИВНАЯ ПРЕДВЗЯТОСТЬ ASR ПРИ РАСПОЗНАВАНИИ АТИПИЧНОЙ РЕЧИ НА РУССКОМ ЯЗЫКЕ

А. В. Колмогорова¹ [0000-0002-6425-2050], Е. В. Явшиц² [0000-0001-6344-1565],
М. С. Майорова³ [0009-0009-2106-7689], С. В. Дмитриева⁴ [0009-0001-1949-5360]

¹⁻⁴Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия

¹akolmogorova@hse.ru, ²kate_yavshitz@mail.ru, ³svyatoslavna.m@gmail.com,
⁴sofiadmitrieva82@gmail.com

Аннотация

Статья посвящена проблеме коммуникативной предвзятости систем автоматического распознавания речи при работе с атипичной речью на русском языке. Актуальность исследования обусловлена активным внедрением технологий искусственного интеллекта в социальную сферу и государственное управление в России, что предполагает обеспечение их доступности для различных категорий граждан, включая людей с речевыми нарушениями. Введено понятие коммуникативной предвзятости ASR-моделей как свойства системы в условиях неопределенности (атипичная речь, шум, паузы) принимать решения о заполнении смысловых и акустических лакун на основе собственной обучающей выборки. В эмпирической базе исследования выбран корпус RuAphasiaBank, включающий записи речи пациентов с различными типами афазии (моторной, сенсорной, семантической, сенсомоторной) разной степени тяжести, а также нейротипичных информантов. Для сравнительного анализа были отобраны шесть ASR-моделей, работающих с русским языком: две открытые (GigaAM, T-One) и четыре коммерческие (Yandex SpeechKit, Any2Text, Charla, Speech2Text). Качество распознавания оценено с использованием метрик WER и CER, а также с помощью анализа соотношения вставок, удалений и замен на уровне слов и символов. Результаты показали, что модели реализуют различные стратегии при столкновении с атипичной речью: «паническое укорачивание» (GigaAM, Yandex SpeechKit), «лексическое сохранение» в ущерб точности (T-One) и «стабильная» стратегия сбалансированных ошибок (Any2Text, Speech2Text, Charla). Лингвистический анализ

выявил, что каждая модель порождает специфический искаженный коммуникативный стиль: от разговорной диалогической речи до стиля, ассоциируемого с когнитивными нарушениями или употреблением молодежного сленга. На основе полученных данных моделируются потенциальные риски для социального самочувствия и правовой защищенности пациентов с афазией при использовании голосовых помощников в государственных учреждениях.

Ключевые слова: речевые технологии, речевая база данных, автоматическое распознавание речи, атипичная речь, афазия, коммуникативная предвзятость, голосовые ассистенты, ассистивные технологии, инклюзивность.

ВВЕДЕНИЕ

Активное развитие технологий искусственного интеллекта в России и его быстрое проникновение в социальную сферу ставят перед разработчиками и исследователями ряд вопросов, связанных с качеством работы, доступностью современных технологий различным категориям пользователей и их объективно существующими ограничениями. Эти вопросы становятся особенно актуальными в связи со стремлением государства внедрить искусственный интеллект в социальную сферу: «В рамках проекта «Искусственный интеллект» планируется внедрить технологии искусственного интеллекта в экономику, социальную сферу и государственное управление»¹. При этом согласно Указу Президента «О развитии искусственного интеллекта в Российской Федерации» к основным принципам развития и использования технологий искусственного интеллекта в России относятся, в первую очередь, «обеспечение защиты прав и свобод человека, гарантированных законодательством Российской Федерации, в том числе права на труд, и предоставление гражданам возможности получать знания и

¹ Нацпроект «Экономика данных и цифровая трансформация государства» [Электронный ресурс]. Офиц. сайт национальных проектов РФ.
URL: <https://xn--80akpjgfh4a0d.xn--p1ai/naczproekt-ekonomika-dannyh-i-czifrovaya-transformacziya-gosudarstva/> (дата обращения: 17.08.2025)

приобретать навыки для успешной адаптации к условиям цифровой экономики»², а также безопасность, понимаемая, в том числе, как «предупреждение и минимизация рисков возникновения негативных последствий использования технологий искусственного интеллекта». К таким негативным последствиям следует относить не только риски в сфере информационной и экономической безопасности, но и репутационные потери, снижение уровня социального (субъективного) настроения [1], основанного, в частности, на социальном самочувствии. Последнее есть «определенное состояние переживания индивида по поводу комфортности или дискомфорта своего бытия в социуме» [2]. Иными словами, проблематизирующий контекст настоящего исследования составляет предположение о том, что возможны случаи, когда технологии могут вносить дисбаланс в психологическое равновесие пользующегося ими индивида, вызывать чувство фрустрации, ухудшая тем самым его социальное самочувствие. Эта гипотеза особенно актуальна тогда, когда мы говорим о людях с заболеваниями, накладывающими определенные ограничения на социальное взаимодействие, и при этом предлагаем технологии, предполагающие смягчение этих ограничений. Подобные технологии включены в категорию ассистивных, понимаемых как «любое внешнее изделие (включая оборудование, инструменты, приборы или программное обеспечение), основная цель которого – поддерживать или улучшать функционирование и независимость человека и тем самым способствовать его благополучию» [3].

Работа посвящена описанию результатов предварительной оценки качества работы шести моделей автоматического распознавания речи (Automatic Speech Recognition – ASR) с корпусом речи пациентов с системным речевым расстройством – афазией. Ее целью – системно представить наблюдения, свидетельствующие о том, что каждая из моделей имеет собственные стратегии получения результата в ситуации работы с атипичной речью, но в большинстве случаев это приводит к комплексному, в случае каждой из моделей – к специфическому, искажению коммуникативного стиля говорящих, что способно повлечь за

² Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» [Электронный ресурс]. Официальный интернет-портал правовой информации. URL:

https://www.consultant.ru/document/cons_doc_LAW_335184 (дата обращения: 17.08.2025)

собой риски в сфере социального самочувствия пользователей с нарушениями речи.

ПОНЯТИЕ КОММУНИКАТИВНОЙ ПРЕДВЗЯТОСТИ: КАК ЭТО РАБОТАЕТ В ОТНОШЕНИИ МОДЕЛЕЙ ASR

В коммуникативной лингвистике сложилась и существует определенная традиция изучения способов передачи чужой речи – ксенопоказателей. К их числу относят модальные частицы *мол, типа, якобы* [4], специфическую просодику [5] и мимические движения, не свойственные пересказывающему чужую речь в его обычном коммуникативном поведении. Между таким сугубо лингвистическим подходом и проблематикой моделей ASR достаточно мало общего. Тем не менее вопрос о стратегиях передачи чужой речи важен как в лингвистическом, так и в технологическом контекстах. Так, например, в [5] отмечено, что при передаче содержания чужих мнений или высказываний возможны две стратегии: либо говорящий ориентируется на свои представления, не заботясь о том, согласится ли с такой передачей субъект передаваемого мнения или высказывания, либо говорящий ориентируется на представления субъекта передаваемого мнения или высказывания, даже если сам их не разделяет.

В техническом контексте же обсуждается несколько иное, но близкое по сути [6] понятие «предсказательной предвзятости» моделей: когда качество предсказания моделей ASR зависит от того, насколько подаваемые на вход данные отличаются от тех, на которых модель обучалась. Ставится вопрос о том, что подавляющее большинство моделей обучается и тестируется на данных, репрезентирующих речь некоторой приоритезируемой компаниями-разработчиками социальной группы. Качество моделей оценивается также на ограниченном количестве датасетов, активно циркулирующих внутри индустриальных и исследовательских сообществ. Ошибки и эрративные паттерны, которые допускают модели, работая с нетипичной или атипичной речью не принимаются во внимание, не собираются и не анализируются. Таким образом, по нашему мнению, разработчики ASR моделей фактически становятся создателями определенной языковой политики, оказывая предпочтение одним идиомам и маргинализируя другие. Иными словами, речь идет о том, что модели также могут «выбирать» ту или иную стратегию для передачи распознанного чужого слова.

Таким образом, объединяя лингвистическую и технологическую концепции, мы будем понимать под коммуникативной предвзятостью ASR-моделей такое свойство системы, при котором в условиях неопределенности (атипичная речь, шум, паузы) модель принимает решение о заполнении лакун, опираясь на собственную обучающую выборку, а не на реальное коммуникативное намерение говорящего, что приводит к систематическому искажению речевого портрета пользователя.

ДАННЫЕ, МЕТОДЫ И МОДЕЛИ

Записи атипичной речи производились с использованием микрофонов iPhone 13–16 во время клинических логопедических занятий с пациентами одного из нейрореабилитационных центров. Для обеспечения конфиденциальности доступ в Интернет был отключен. Занятия проходили в тихом помещении, пациент и терапевт сидели за столом, телефон размещался на расстоянии 30–40 см от говорящего, что обеспечивало минимальный уровень фонового шума. Записи речи нейротипичных участников собирались в сопоставимых акустических условиях. Все участники – как нейротипичные лица, так и пациенты с афазией (ПА) – дали информированное согласие на участие в записи. Возрастное распределение ПА и нейротипичных информантов коррелировали, гендерное распределение – нет. В датасете представлены различные типы речи: монолог, диалог, чтение и пересказ. Основные параметры данных представлены в табл. 1, более подробное описание датасета можно найти в [9].

В сравнительный анализ были включены шесть моделей ASR, в документации которых указано, что при их обучении использовались данные на русском языке, а в продвигающем тексте отмечалось, что модель предназначена в том числе для анализа данных на русском языке. Две модели являются открытыми (GigaAM и T-One), еще 4 – коммерческими (Yandex SpeechKit, Any2text, Charla, Speech2Text).

Поскольку для каждой записи уже имелись проверенные экспертами эталонные транскрипты, сделанные вручную, основной методологией стала оценка качества распознавания при помощи стандартных метрик WER (Word Error Rate) [10] и CER (Character Error Rate) [11]. Однако мы решили не просто измерить мет-

рики для каждой из моделей на выборках: а) различных типов афазий, б) различных степеней выраженности нарушений, в) различных типов речи, но и посмотреть в каждом из 18 случаев соотношение вставок, замен и удалений, чтобы проанализировать стратегии, которые выбирают модели при столкновении с атипичными записями.

Табл. 1. Основные параметры датасета.

Параметр	Категория	Объем
Записи	Общая длительность	14 ч
	ПА	9.7 ч
	Нейротипичные	4.3 ч
Информанты	ПА	70 чел.
	Логопеды	8 чел.
	Нейротипичные	20 чел.
Гендер	ПА m/f	54/16 чел.
	Логопеды m/f	0/8 чел.
	Нейротипичные m/f	18/6 чел.
Средний возраст (в годах)	ПА m/f	53.4/61.3
	Логопеды m/f	0/41.2
	Нейротипичные m/f	56.1/70.2

РЕЗУЛЬТАТЫ

Сравнение значений WER при распознавании записей речи пациентов с разными типами афазий демонстрирует (см. рис. 1), что наиболее успешные по этой метрике модели изменяются в зависимости от типа афазии, а аутсайдер неизменен – это Yandex SpeechKit, хотя и его поведение нестабильно: в случае семантической афазии лидируют замены, а при сенсорной – удаления. GigaAM показывает наилучшие значения WER при комплексной моторной, семантической и сенсомоторной афазии, но при этом так же, как и аутсайдер, предпочитает удалять непонятое.

T-One выделяется высоким уровнем замен (до 25%) при относительно низком CER (рис. 2) (15–24%), что говорит о стратегии «лексического сохранения» – модель жертвует точностью на уровне символов, но пытается сохранить слова;

его лучший результат достигается на семантической афазии (WER 29%, CER 19%). Speech2Text, Charla и Any2Text занимают промежуточное положение (WER 27–41%, CER 18–26%) (рис. 1 и 2), демонстрируя сбалансированное соотношение удалений и замен; Speech2Text выделяется аномально высокой долей вставок на семантической афазии (10.1%), а Any2Text показывает лучший результат на эфферентной моторной афазии (WER 32%, CER 18%).

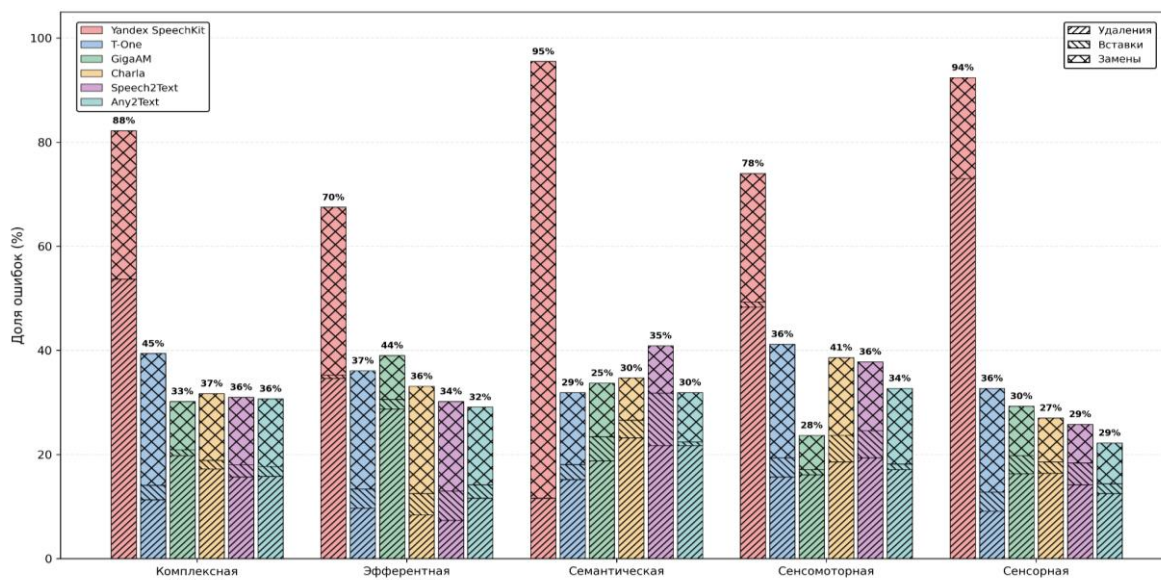


Рис. 1. WER и соотношение вставок/удалений/замен по разным типам афазии.

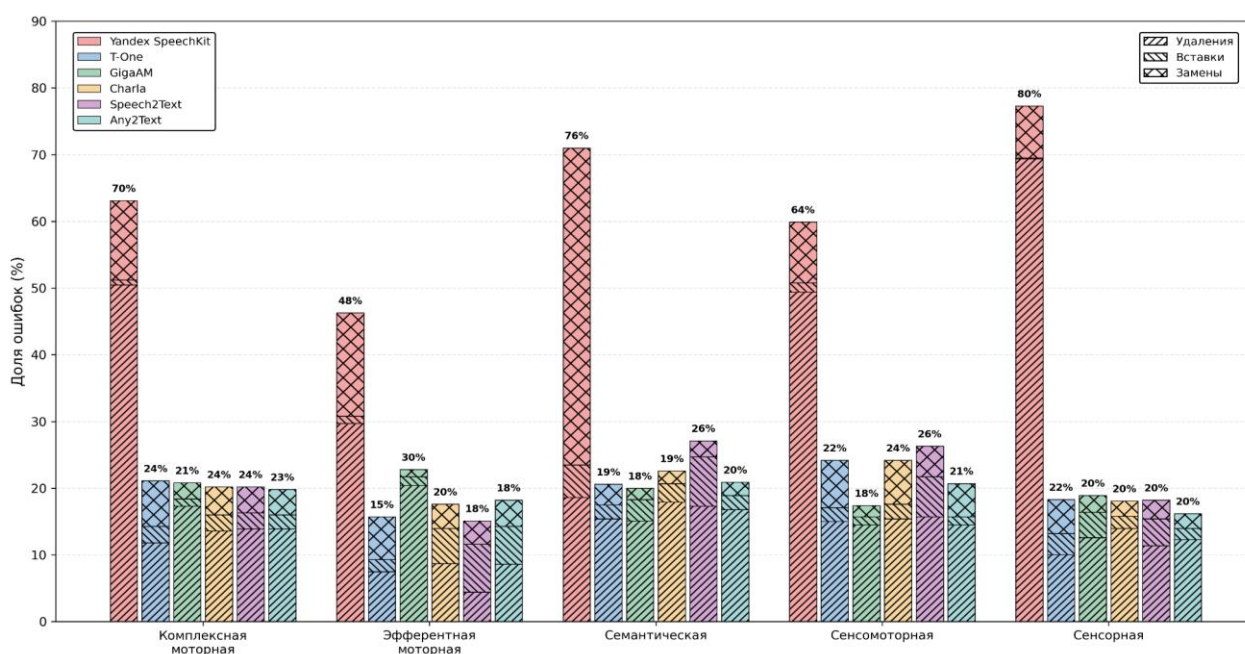


Рис. 2. CER и соотношение вставок/удалений/замен по разным типам афазии.

Поскольку в датасете представлены записи и нейротипичных информантов (монолог и диалог), то при анализе результатов обработки записей с различной степенью выраженности речевого нарушения (рис. 3 и 4) заметен четкий градиент: все модели показывают рост WER от нейротипичных к тяжелой афазии. Исключение составляет Yandex SpeechKit: у этой модели WER и так высокий на нейротипичных (89%), и при афазии он почти не растет, потому что «потолок» достигнут. GigaAM демонстрирует самые низкие значения WER, которые умеренно повышаются вместе с ростом тяжести нарушения: 17% (нейротипичные) → 24% (легкая) → 32% (средняя) → 35% (тяжелая). T-One и Charla показывают наиболее крутой подъем от легкой к тяжелой (40% → 47% и 30% → 41% соответственно).

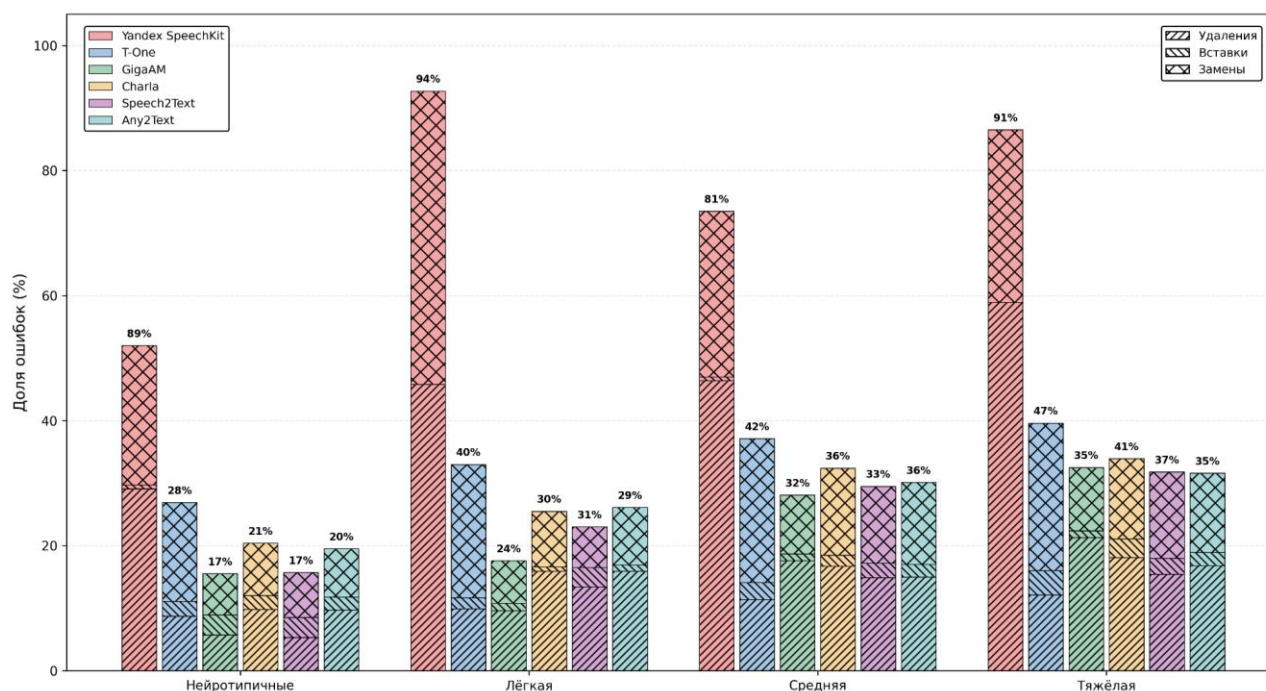


Рис. 3. WER и соотношение вставок/удалений/замен по степени тяжести афазий.

При распознавании речи, имеющей разную степень речевой девиации, на уровне символов (CER) все модели демонстрируют значительно более низкий процент ошибки, чем на уровне слов (WER). Это говорит о том, что большинство ошибок связано с неправильным определением границ слов, а не с искажением отдельных букв. Остальные тенденции коррелируют с WER. Однако, например,

T-One показывает CER 15–25% с преобладанием замен (4.3–7.1%) над удалениями (7.1–13.1%) – модель сохраняет слова, но искажает отдельные символы внутри них. Speech2Text, Charla и Any2Text занимают промежуточное положение (CER 18–26%) со сбалансированным ростом всех типов изменений по сравнению с эталонными транскриптами, при этом у Speech2Text на легкой степени нарушения CER (18%) выше, чем у GigaAM (15%), но ниже, чем у Any2Text (20%).

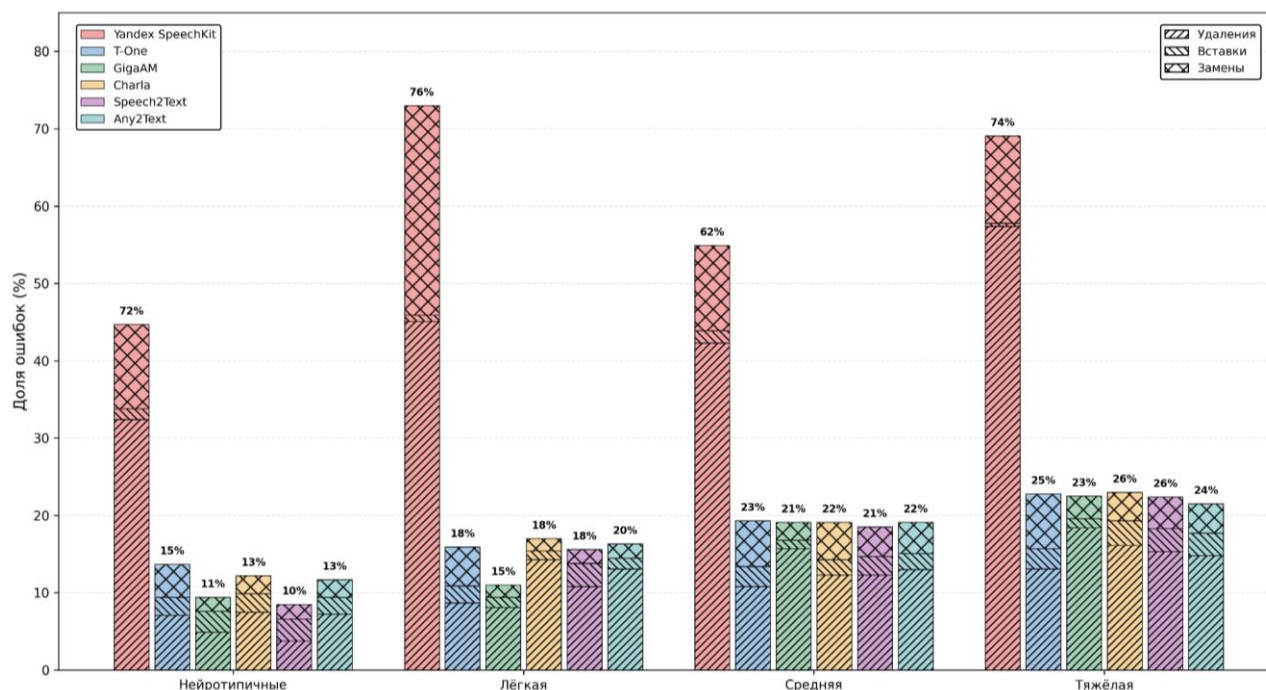


Рис. 4 CER и соотношение вставок/удалений/замен по степени тяжести афазий.

Сравним теперь модели по типам речи (рис. 5 и 6). Медианные значения WER и CER для всех моделей, за исключением Yandex SpeechKit, на уровне слов демонстрируют ожидаемую тенденцию: самые низкие показатели ошибок наблюдаются при распознавании нейротипичных записей, тогда как самые высокие – при обработке пересказов у ПА. Yandex SpeechKit имеет самые высокие показатели ошибок как для нейротипичных, так и для афазических записей диалогов; распознавание же монологов этой моделью значимо лучше для нейротипичных, чем для ПА. GigaAM имеет наиболее сбалансированную и устойчивую производительность на всех типах речи, превосходя другие модели по WER в самом сложном случае – пересказе, хотя при этом использует чаще всего удаления.

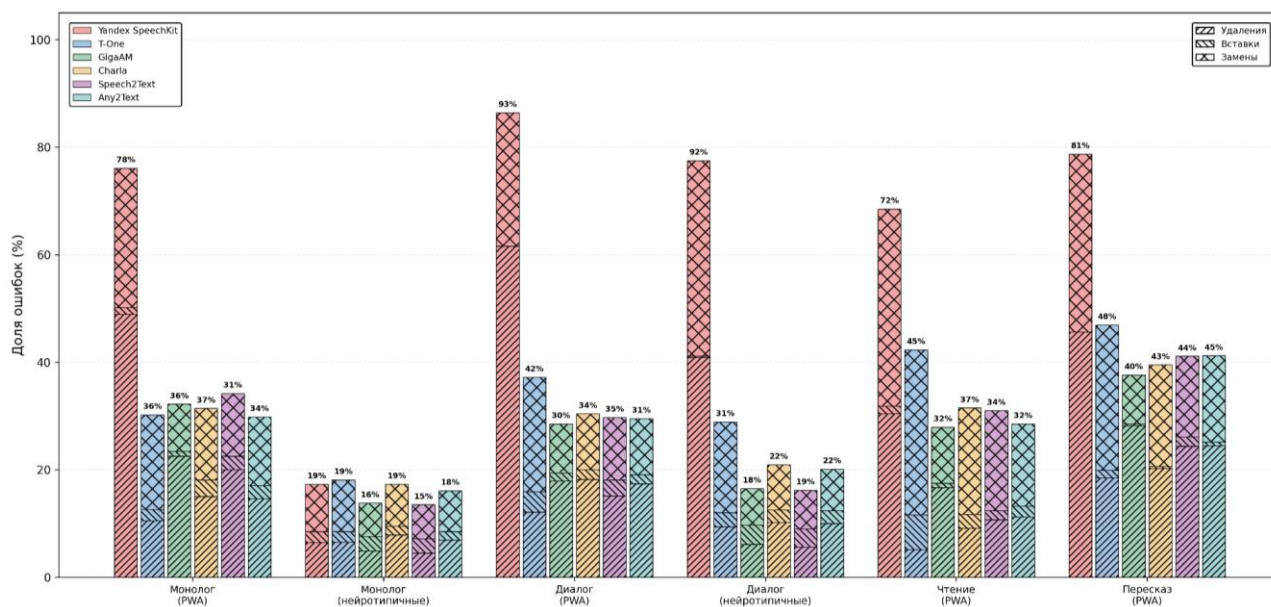


Рис. 5. WER и соотношение вставок/удалений/замен по типу речи у пациентов с афазией.

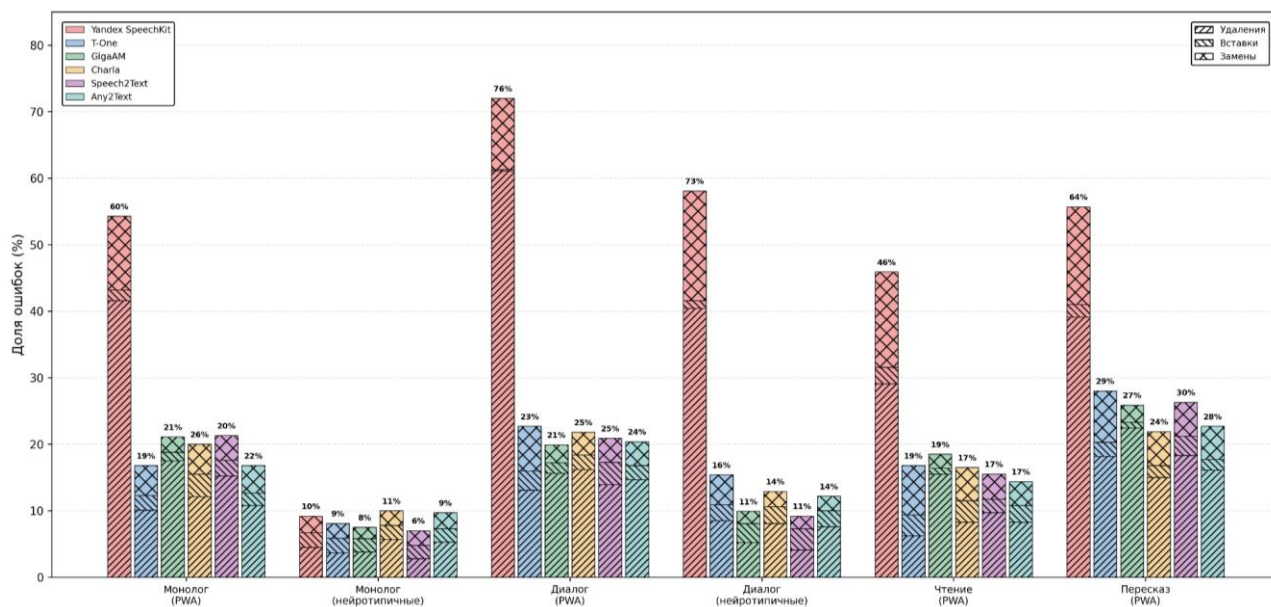


Рис. 6. CER и соотношение вставок/удалений/замен по типу речи.

Speech2Text показывает сравнительно хорошие результаты на монологах (WER – 31%, CER – 20%), но теряет свое преимущество в интерактивных или нарративных контекстах афазической речи. Any2Text и Charla имеют стабильные средние результаты, в то время как T-One отстает, приоритезируя замены.

Таким образом, в результате проведенного анализа установлены три различные стратегии моделей при распознавании речи в условиях неопределенности, в нашем случае – при обработке атипичной речи. GigaAM и Yandex SpeechKit используют «паническое укорачивание» – агрессивное сжатие распознанного текста по сравнению с оригиналом. T-One старается сохранить лексическое наполнение в ущерб точности на уровне символов. Any2Text, Speech2Text и Charla демонстрируют «стабильную» стратегию, при которой длина выходного текста практически не зависит от снижения качества речевых данных (например, при усугублении грубости речевого нарушения у говорящего).

Табл. 2. Результаты сравнения замеров метрик качества распознавания русскоязычных движков автоматического распознавания речи.

ПА						
ASR Model	Yandex SpeechKit	T-One	GigaAM	Charla	Speech2Text	Any2Text
WER	88%	43%	32%	37%	35%	34%
CER	69%	23%	21%	23%	23%	23%
Нейротипичные информанты						
ASR Model	Yandex SpeechKit	T-One	GigaAM	Charla	Speech2Text	Any2Text
WER	89%	28%	17%	21%	17%	20%
CER	72%	15%	11%	13%	10%	13%

Yandex SpeechKit показывает наихудшее значение метрики WER (88%) (табл.2) на фоне чрезвычайно высокого количества удалений и замен. Тем не менее GigaAM имеет самый низкий показатель WER (32%) за счет большого количества удалений на фоне минимального количества вставок и замен. T-One (WER = 43%) выделяется самой высокой долей замен на уровне слов. Большой разрыв между CER и WER (WER = 43% при CER = 23%) для T-One позволяет пред-

положить, что модель правильно распознает большинство символов, но допускает одно-два незначительных искажения на слово – этого как раз достаточно, чтобы все слово было отвергнуто метрикой WER. Такое расхождение указывает на фонетическую близость ошибок: модель испытывает трудности с распознаванием редких или искаженных фонем, типичных для афазической речи, но лексический инвентарь текста в значительной степени сохраняется. Остальные модели – Charla (WER = 37%), Speech2Text (WER = 35%) и Any2Text (WER = 34%) – демонстрируют примерно схожие значения метрик (табл.2). Однако само по себе низкое значение WER или CER не может быть абсолютным показателем качества распознавания. То же самое можно сказать и о соотношении удалений/вставок/замен.

Проанализируем качественно лингвистический характер удалений/вставок/замен и попробуем спрогнозировать, какие черты коммуникативного стиля фиксируются моделями при распознавании и как это может повлиять на характер социального взаимодействия в случае, если клиент с нарушением речи решит воспользоваться услугами голосового помощника в банке или ином учреждении.

Например, при распознавании семантической афазии GigaAM стремится структурировать высказывание, приводя его к некоторым типовым синтаксическим структурам, характерным для разговорной речи. Следует отметить, что семантическая афазия сложна тем, что пациент достаточно случайным образом употребляет служебные слова, ему не удается передача пространственных и причинно-следственных отношений, неправильно употребляются слова и конструкции с непрямым значением. Например, пациент на вопрос логопеда о том, кем он работал, пытается сказать, что он ходил на судне в море, был моряком, но ему удается передать только простую пространственную схему – некоторый сверток смысла, передающий ритм его профессиональной жизни «туда-сюда (ходил в рейсы)»:

(1) [эталонная расшифровка] *мне а я это как его вот туда-сюда туда-сюда;*

(2) [аутпут модели]: *Мне это как его? Вот – Туда-сюда, туда-сюда.*

Речь пациента изобилует повторами (парафазиями) двух первых слогов слова *конечно: коне коне коне*. Модель удаляет часть повторов, а то, что было

неразборчиво, заменяет на другие асемантические повторы – *отрадио, радио*, но при этом отчетливо делит транскрипт на короткие ясные диалогические реплики:

(4) [аутпут модели] *Дети живут отдельно отрадио, радио. – Отдельно. – Да, конечно.*

По мере ухудшения качества речи (при движении от слабой степени выраженности речевого нарушения к сильной) GigaAM удаляет хезитации, числительные (например, *тринадцать*), имена-отчества (*Иван Сергеевич*) и многосложные слова (*вернувшись*), заменяя их на более короткие формы: хезитации, предлоги (чаще всего – *в, на*), междометия, обсценизмы, местоимения *я* и *ты*, производятся замены литературных форм на более разговорные (*поскользнулся* → *подскользнулся*). Таким образом, речь говорящего в расшифровках GigaAM предстает как более разговорная, диалогическая, используются короткие, но емкие реплики.

Модель от Yandex лучше всего справляется с речью пациентов с эфферентной моторной афазией. Для такой речи характерны многократные повторы слогов, замены одних звуков на другие (похожие):

(5) [эталонная расшифровка] *доча есть двадцать два года и иже оза занимается оу не знаю она живет ну две нрзб учиться не знаю пор пор званию медик полощ зубы протезирует.*

Yandex SpeechKit правильно расшифровывает ключевое числительное *двадцать два* и глагол *занимается*, но остальные слова заменяет рандомными формами: междометиями (*ага, ну*) и формами, частотными для речи молодежи (*ВК, Про образование* (название подкаста)) и игрового сленга (*Зерай* и *Карвез* – имена персонажей *World of Warcraft*):

(6) [аутпут модели] *это Надо двадцать два года и Ага занимается Она жалеет Ну два Катя Вк Про образование Карвез зерай.*

Yandex SpeechKit испытывает наибольшие трудности при распознавании речи пациентов с уже упоминавшейся выше семантической афазией. Например, тот же фрагмент, что и в примере 1, данная модель распознает как:

(7) *Мы живем техникой Что выжили не теряя ни один Благодарю Челябинка джэйли диана Нет То О это ж деньги И какими же ладонями была ему музейными какая то феноменальная ночь.*

Рандомно подобранные лексемы отражают, скорее, коммуникативный стиль героев манг, нежели повседневную речь рядового говорящего. По мере огрубления речевого расстройства в записях (от легкой до тяжелой степени) модель Yandex SpeechKit заменяет слова случайным образом на другие существующие, часто выступающие как контекстуальные синонимы (*чиновников* → *хирургов*, т. е. название профессий, видов деятельности) или синтагматические соседи (*много* → *суп*, частотная коллокация *много супа*).

Таким образом, Yandex SpeechKit искажает коммуникативный стиль «диктора» (пациента с речевым расстройством) до неузнаваемости, делая смысловую сторону почти полностью бессмысленной, а форму высказывания – ассоциирующейся с молодым человеком, активно пользующимся социальными сетями и сленгом геймеров.

T-One как будто пытается достигнуть наибольшей интерпретабельности говоримого. Например, фраза в примере 3 распознается очень четко благодаря тому, что все асемантические слова просто удаляются:

(8) *дети живут отдельно отдельно.*

По мере ухудшения качества речи модель обнаруживает постепенное усиление следующих тенденций: удалению подвергаются не только короткие слова (*ой, там*), несуществующие формы (*пэцхоус, чизля, пахля*), многосложные слова (например, *захихикал, белоствольная, революции*), но и термины родства (*жена, внук, дед, дедок*); вставляются обцензизмы, производятся замены «с толка» (*а – нагорская, ну – звезда*), а также усечение первого слога оригинального многосложного слова, в результате которого появляется другое, тоже существующее, слово с иным значением (*благодарит – дарит, оторвалось – рвало, убежал – жал*). Модель стремится быть морфологически изобретательной и легко манипулирует близкими по фонемному составу словами. Особенно это проявляется при распознавании чтения текста пациентом. Сравните:

(9) *ориг – я косточку бросил за окошко;*

t_one – я кошечку брошу за окошко.

Таким образом, порождаемое высказывание при поверхностном знакомстве выглядит достаточно правильным и в большинстве случаев отвечает речевому стандарту, однако передаваемое содержание сильно искажено по сравнению с оригиналом, что порождает у слушающего когнитивные диссонансы

между соседними контекстами. Такая речь, в свою очередь, может ассоциироваться с человеком, имеющим когнитивные нарушения.

Speech2text интересна тем, что на месте неразборчивого фрагмента довольно часто вставляет обсценизмы («нрзб» → «п...з...ц»), а также многократно перебирает варианты по одному услышанному слогу («петь», «пять», «почти», «пас»). Склонность к морфологическим экспериментам приводит к тому, что модель часто производит морфологические замены глаголов по лицу, виду, времени, словообразовательной модели, типа предикации (полная/свернутая): *дочитывала* → *зачитывала*, *накинув* → *накинула*; *убежали* → *сбежали*, *сбросился* → *бросился*. Еще один любопытный паттерн – замена на «угу» («лезняком» → «угу»), т. е. модель просто «сдается» и ставит маркер согласия вместо непонятого слова. Лучше всего модель справляется с эфферентной моторной афазией, достаточно неплохо подбирая контекстуальные варианты, и опираясь на отдельные услышанные слоги. Так, реплику о дочери, работающей зубным техником (пр.5), модель распознает как:

(10) *дочь есть, двадцать два года ее или она занимается, не знаю, она живет, ну две группы в городе учится не знаю по образованию медик волос зубы паразитизирует, где просторечное доча заменяется на литературное дочь, оза правильно достраивается до она, ну две нрзб превращаются в две группы, а пор пор званию медик – в по образованию медик*. Преимуществом модели является ее способность, поняв общий контекст (например, образование и род занятий), достраивать непонятое самыми частотными коллокациями. Однако специальное слово из медицинского контекста *протезирует*, было заменено на оценочное общеупотребительное *паразитирует*.

Контекст же из примера с семантической афазией (пример 3) модель же передала идеально с точки зрения узуса, удалив все неразборчивые и эрративные формы:

(11) *Дети живут отдельно от вас? Отдельно, да.*

Иными словами, модель порождает такой коммуникативный стиль, который ассоциируется с говорящим, репрезентирующим кооперативный тип коммуникативной личности, склонным соглашаться с собеседником [12], владеющим русской литературной нормой и имеющим достаточно большой словарный запас, но в качестве междометий нередко использующим обсценизмы.

Any2text выделяет вставки, делающие речь максимально разговорной, даже близкой к просторечной – местоимения я и ты, обценизмы (*бл...ть, них...я*), имена (например, *Владимир*), термины родства (*жена, сын, внучка*), неправильные формы. При этом заменяет редкие, неправильные и даже частотные правильные формы выдуманными (*ивняк* → *явнях*, *защищаль* → *сосисали*, *удоками* → *ударками*, *взгляда* → *гляда*, *принадлежностями* → *пренаслежностями*, *насили* → *насылили*). В результате фраза о дочери-зубном технике (пример 5) выглядит так:

(12) *Дочь есть. двадцать два года ее. Или... Она занимается... не знаю она живет ну две группы короткий хочется Не знаю. Борзование медика. Уболость зубы паразитизирует.*

Короткие сегменты, многоточия, маркирующие паузы и хезитации, повторы «не знаю», междометия *ну* делают речь очень разговорной, а *борзование* и *уболость* делают ее странной.

Отметим, что модель Charla демонстрирует аналогичные изменения оригинальной расшифровки. Иными словами, порождаемый моделями коммуникативный стиль может быть ассоциирован с таким говорящим, который владеет исключительно разговорной или просторечной речью, возможно, имеет психические или когнитивные нарушения.

ОБСУЖДЕНИЕ

Проведенные эксперименты показали, что при оценке качества распознавания метрики CER и WER дают только общую оценку, которую необходимо дополнять качественно-количественным анализом вставок, удалений и замен. Так, например, среди трех моделей Charla, Any2text и Speech2text при сопоставимых значениях метрик и соотношений вставок/удалений/замен Speech2text генерирует расшифровки, которые могут быть ассоциированы слушающим с коммуникативным стилем носителя языка, владеющего узуальной повседневной разговорной речью, и склонного соглашаться с собеседником и использовать короткие обценизмы в качестве междометий. Остальные две модели выдают такие распознанные высказывания, которые реализуют коммуникативный стиль, скорее, свойственный малограмотному человеку, владеющему преимущественно

просторечием и имеющему некоторые психические девиации, например, находящемуся в состоянии алкогольного опьянения или под воздействием психотропных веществ. T-One порождает формально правильные высказывания, которые противоречат изначальному смыслу речевого фрагмента, что может создать у слушающего впечатление о том, что собеседник имеет некоторые когнитивные нарушения.

Лучшая по метрикам модель GigaAM продемонстрировала стратегию панического удаления, но при этом ей удастся трансформировать усеченные распознанные фразы в короткие диалогические реплики, соответствующие представлению о живой повседневной диалогической речи. Наихудшая же модель от Yandex, демонстрируя ту же стратегию панического удаления, все же делает вставки и замены, хотя бы в малой степени пытаясь компенсировать удаленное, но использует при этом молодежный сленг и сленг геймеров.

Таким образом, можно заключить что наша гипотеза о коммуникативной предвзятости моделей подтвердилась, но в каждом случае в разной степени. Так, в наибольшей степени это верно для Yandex SpeechKit: по-видимому, модель обучалась на записях спонтанной диалогической речи представителей молодежных сообществ, пользователей социальных сетей, которых компания одновременно рассматривает как своих основных потребителей. Скорее всего, GigaAM обучалась на спонтанных повседневных диалогах взрослых (возможно, 30+) жителей мегаполисов (Москва, Санкт-Петербург), имеющих высшее образование, однако не исключено и обучение на некотором сэмпле атипичной речи, поскольку модель хорошо работает с ошибочными и неполными формами. Speech2text опирается на обучающие выборки записей разговорной речи информантов, возможно, имеющих более низкий образовательный ценз, чем информанты, речь которых «слушала» GigaAM. Можно предположить, что остальные три модели не обучались на повседневной разговорной речи, хотя оказались знакомыми с нормированной речью актеров, дикторов, иных, не спонтанных, говорящих.

В дальнейших исследованиях планируется проверить сформулированные на основе лингвистического анализа признаки коммуникативной предвзятости в экспериментальной работе с информантами-носителями русского языка. В про-

водимой в настоящее время работе мы ставим перед информантами задачу проанжировать шестерки транскриптов по степени субъективной понятности смысла сообщаемого, на следующем этапе экспериментальное задание будет предложено сотрудникам госучреждений и сервисов.

Перспективу исследования составляет также дополнение датасета новыми записями речи пациентов и нейротипичных информантов в тех же условиях и их расшифровками. Несбалансированность записей по типам афазии и степени тяжести речевого нарушения составляет серьезное ограничение данных, которое исследовательская команда в полной мере осознает.

ЗАКЛЮЧЕНИЕ

Проведенный анализ шести систем ASR демонстрирует, что предложенное нами определение коммуникативной предвзятости следует считать рабочим и достаточно продуктивным, поскольку нам удалось установить и описать признаки коммуникативной предвзятости для большинства моделей, участвовавших в экспериментах на датасете RuAphasiaBank. То, в какой степени коммуникативная предвзятость может навредить социальному самочувствию пациентов с афазией при пользовании голосовыми сервисами и ассистентами, только предстоит оценить на практике – в настоящей работе мы лишь смоделировали одну из ситуаций гипотетически. Тем не менее уже сама постановка вопроса обращает нас к широкой проблеме технологической инклюзии, которая требует своего обсуждения.

Благодарности

Статья и данные были подготовлены в рамках Программы фундаментальных исследований НИУ ВШЭ (HSEBR-2025-032). Исследование одобрено Комиссией по внутривузовским опросам и этической оценке эмпирических исследовательских проектов Высшей школы экономики. Авторы являются зарегистрированными правообладателями описанной базы данных RuAphasiaBank: база данных фонограмм речи пациентов с афазией зарегистрирована в Федеральной службе по интеллектуальной собственности Российской Федерации (№ 6.0021-2025).

СПИСОК ЛИТЕРАТУРЫ

1. Харченко С.В. Феномен и природа социального настроения // Мир науки, культуры, образования. 2011. № 1. С. 168–170.
2. Михайлова Л.И. Социальное самочувствие и восприятие будущего россиянами // Социологические исследования. 2010. № 3. С. 45–49.
3. *World Health Organization*. Priority assistive products list: improving access to assistive technology for everyone, everywhere. Geneva, 2016.
URL: <https://www.who.int/publications/i/item/priority-assistive-products-list> (дата обращения: 20.08.2026).
4. Левонтина И.Б. Об арсенале ксенопоказателей в русском языке // Вопросы языкознания. 2020. № 3. С. 52–77.
5. Шклярчук Е.Я. Маркеры-ксенопоказатели в русской устной речи: дискуссионный статус единиц // Вестник ЮУрГУ. Серия: Лингвистика. 2025. № 2. С. 70–78.
6. Markl N., McNulty S.J. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR // Proc. of the 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA), 2022. P. 6328–6339.
7. Вассерман Л.И., Дорофеева С.А., Меерсон Я.А. Методы нейропсихологической диагностики: практическое руководство. СПб.: Стройлеспечать, 1997. 123 с.
8. Лурия А.Р. Высшие корковые функции человека. 3-е изд. СПб.: Питер, 2008. 128 с.
9. Колмогорова А.В., Явшиц Е.В., Никольская О.Н., Сугян А.Х. Сервисы автоматического распознавания речи на русском языке: исследование качества и инклюзивности на примере датасета RuAphasiaBank // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Т. 24, № 1. С. 113–127.
10. Morris A.C., Maier V., Green P. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition // Proc. Interspeech. 2004. P. 2765–2768.
11. Thennal K.D., James Jesin, Gopinath D.P., Ashraf K.M. Advocating Character Error Rate for Multilingual ASR Evaluation // arXiv.org. 2024.
<https://doi.org/10.48550/arXiv.2410.07400>

12. Седов К.Ф. Речевое поведение и типы языковой личности // Культурно-речевая ситуация в современной России. Екатеринбург, 2000. С. 298–312.

COMMUNICATIVE BIAS IN ASR RECOGNITION OF ATYPICAL RUSSIAN SPEECH

A .V. Kolmogorova¹ [0000-0002-6425-2050], E. V. Yavshits² [0000-0001-6344-1565],
M. S. Mayorova³ [0009-0009-2106-7689], S. V. Dmitrieva⁴ [0009-0001-1949-5360]

^{1–4}*National Research University Higher School of Economics, Moscow, Russia*

¹akolmogorova@hse.ru, ²kate_yavshitz@mail.ru, ³svyatoslavna.m@gmail.com,

⁴sofiadmitrieva82@gmail.com

Abstract

The article addresses the problem of communicative bias in automatic speech recognition systems when processing atypical speech in Russian. The relevance of the study is driven by the active integration of artificial intelligence technologies into the social sphere and public administration in Russia, which presupposes ensuring their accessibility for various categories of citizens, including people with speech disorders. The authors introduce the concept of communicative bias of ASR models as a property of a system that, under conditions of uncertainty (atypical speech, noise, pauses), makes decisions about filling semantic and acoustic gaps based on its own training data rather than the speaker's actual communicative intention. The empirical basis of the study is the RuAphasiaBank corpus, which includes speech recordings from patients with various types of aphasia of varying severity, as well as neurotypical informants. For comparative analysis, six ASR models operating with Russian were selected: two open-source (GigaAM, T-One) and four commercial (Yandex SpeechKit, Any2Text, Charla, Speech2Text). Recognition quality was assessed using WER and CER metrics, as well as a qualitative analysis of the ratio of insertions, deletions, and substitutions at the word and character levels. The results show that the models implement different strategies when encountering atypical speech: “panic shortening” (GigaAM, Yandex SpeechKit), “lexical preservation” at the expense of accuracy (T-One), and a “stable” strategy of balanced errors (Any2Text, Speech2Text, Charla). Linguistic analysis

reveals that each model produces a specific distorted communicative style, ranging from colloquial dialogic speech to a style associated with cognitive impairments or the use of youth slang. Based on the obtained data, potential risks to the social well-being and legal security of aphasia patients when using voice assistants in public institutions are modeled.

Keywords: *speech technology, speech database, automatic speech recognition, atypical speech, aphasia, communicative bias, voice assistants, assistive technologies, inclusivity.*

REFERENCES

1. *Kharchenko S.V.* Fenomen i priroda sotsialnogo nastroyeniya // *Mir nauki, kultury, obrazovaniya*. 2011. № 1. P. 168–170.
2. *Mikhaylova L.I.* Sotsialnoe samochuvstvie i vospriyatie budushchego rossiyanami // *Sotsiologicheskie issledovaniya*. 2010. № 3. P. 45–49.
3. *World Health Organization*. Priority assistive products list: improving access to assistive technology for everyone, everywhere. Geneva, 2016.
URL: <https://www.who.int/publications/i/item/priority-assistive-products-list> (accessed: 20.08.2026).
4. *Levontina I.B.* Ob arsenale ksenopokazateley v russkom yazyke // *Voprosy yazykoznaniiya*. 2020. № 3. P. 52–77.
5. *Shklyaruk E.Ya.* Markery-ksenopokazateli v russkoy ustnoy rechi: diskussionnyy status edinit // *Vestnik YuUrGU. Seriya: Lingvistika*. 2025. № 2. P. 70–78.
6. *Markl N., McNulty S.J.* Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR // *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. European Language Resources Association (ELRA), 2022. P. 6328–6339.
7. *Vasserman L.I., Dorofeeva S.A., Meerson Ya.A.* Metody neyropsikhologicheskoy diagnostiki: prakticheskoe rukovodstvo. SPb.: Stroylespechat, 1997. 123 p.
8. *Luriya A.R.* Vysshie korkovye funktsii cheloveka. 3-e izd. SPb.: Piter, 2008. 128 p.
9. *Kolmogorova A.V., Yavshits E.V., Nikolskaya O.N., Sugyan A.Kh.* Servisy avtomaticheskogo raspoznavaniya rechi na russkom yazyke: issledovanie kachestva i

inklyuzivnosti na primere dataseta RuAphasiaBank // Vestnik NGU. Seriya: Lingvistika i mezhkulturnaya kommunikatsiya. 2026. Vol. 24, № 1. P. 113–127.

10. *Morris A.C., Maier V., Green P.* From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition // Proc. Interspeech. 2004. P. 2765–2768.

11. *Thennal K.D., James Jesin, Gopinath D.P., Ashraf K.M.* Advocating Character Error Rate for Multilingual ASR Evaluation // arXiv.org. 2024.
<https://doi.org/10.48550/arXiv.2410.07400>

12. *Sedov K.F.* Rechevoe povedenie i tipy yazykovoy lichnosti // Kulturno-rechevaya situatsiya v sovremennoy Rossii. Ekaterinburg, 2000. P. 298–312.

СВЕДЕНИЯ ОБ АВТОРАХ



КОЛМОГОРОВА Анастасия Владимировна – Профессор, доктор филологических наук, заведующая лабораторией языковой конвергенции и профессор департамента филологии НИУ ВШЭ – Санкт-Петербург. Является методологическим руководителем проекта. Под ее руководством сформировалась научная школа, объединяющая более 20 исследователей, 14 из которых – кандидаты наук. Имеет опыт реализации прикладных проектов, включая разработку систем автоматической обработки текстовых и речевых данных по заказам промышленных и коммерческих организаций. Владеет теоретическими основами, так и практическими инструментами компьютерной лингвистики.

Anastasia Vladimirovna KOLMOGOROVA is a Professor, Doctor of Philological Sciences, head of the Laboratory for Language Convergence, and professor at the Department of Philology at HSE University – St. Petersburg. She serves as the project's methodology lead. She heads a scientific school that currently brings together over 20 researchers, 14 of whom have successfully defended their PhD theses under her supervision. Prof. Kolmogorova has extensive experience in managing applied projects, including those involving automated processing of textual and speech data commissioned by industrial and commercial enterprises. She is proficient in both the theoretical foundations and practical methods of computational linguistics.

email: akolmogorova@hse.ru

ORCID: 0000-0002-6425-2050



ЯВШИЦ Екатерина Валерьевна – специалист в области речевых технологий и компьютерной лингвистики, чья профессиональная деятельность сосредоточена на междисциплинарном пересечении автоматического распознавания речи и обработки естественного языка (NLP). Научно-исследовательские интересы охватывают широкий спектр направлений, включая цифровую обработку речевых сигналов, речевую реабилитацию, устойчивое распознавание диктора и речи в условиях акустических шумов, а также речевую аналитику и ее применение для улучшения клиентского опыта (CX).

Ekaterina Valerievna YAVSHITS is a specialist in speech technologies and computational linguistics, whose professional work lies at the interdisciplinary intersection of automatic speech recognition (ASR) and natural language processing (NLP). The author's research interests cover a broad spectrum of areas, including digital speech signal processing, speech rehabilitation, robust speaker and speech recognition in noisy acoustic environments, as well as speech analytics and its application to enhancing customer experience (CX).

email: kate_yavshitz@mail.ru

ORCID: 0000-0001-6344-1565



МАЙОРОВА Милана Святославна – специалист в области компьютерной лингвистики, работающий на стыке анализа данных, машинного обучения и обработки естественного языка. В сферу научных интересов входят прикладная, когнитивная и корпусная лингвистика, а также изучение нарушенной речи.

Milana Svyatoslavna MAYOROVA is a specialist in computational linguistics, working at the intersection of data analysis, machine learning, and natural language processing. Research interests include applied, cognitive, and corpus linguistics, as well as the study of disordered speech.

email: svyatoslavna.m@gmail.com

ORCID: 0009-0009-2106-7689



ДМИТРИЕВА Софья Владимировна – магистрант программы «Языковые технологии в бизнесе и образовании» НИУ ВШЭ – Санкт-Петербург, специалист по разметке данных ООО «КВИНТ». Занимается разметкой текстовых и аудиоданных для голосовых роботов и контролем качества их работы.

Sofia Vladimirovna DMITRIEVA is a master's student in the «Language Technology in Business and Education» program at HSE University – St. Petersburg, and a data labeling specialist at ООО «KVINT». She labels text and audio data for voice bots and monitors the quality of their performance. She has experience preparing, testing, and optimizing dialogue scenarios for voice assistants.

email: sofiadmitrieva82@gmail.com

ORCID: 0009-0001-1949-5360

Материал поступил в редакцию 21 августа 2026 года