

<https://doi.org/10.22363/2687-0088-50181>

EDN: NKOHEU

Review article / Обзорная статья

Artificial intelligence in Russian sign language recognition: Bridging sign phonology and translation systems

Ilya N. PLAVUNOV¹ and Yulia N. EBZEEVA²  

¹*PAO Sberbank*

²*RUDN University, Moscow, Russian Federation*

 ebzeeva-jn@rudn.ru

Abstract

By 2026, automated sign language recognition has moved from research laboratories into applied domains, including mobile translators, educational software and inclusive services. For Russian Sign Language (RSL), however, this field remains fragmented and is rarely addressed from a linguistic perspective. The aim of the article is threefold: to systematise the state of research on AI-based recognition of RSL with an explicitly linguistic frame; to identify the principal gap, namely the absence of a continuous parallel RSL–spoken Russian corpus; and to propose a concrete component-based architecture for an RSL-to-Russian translation system grounded in published evidence. The review focuses on seven blocks: the phonology, morphology and prosody of RSL; the gap between Russian and international corpora (BSL Corpus, DGS-Korpus, How2Sign, RWTH-PHOENIX-Weather versus the NSTU corpus and TheRuSLan); the evolution of AI architectures from HMM-based continuous recognition to transformer-based translation and large language models; the Russian landscape of datasets and systems built around Slovo, Bukva and Logos; the linguistic validity criteria for synthetic sign data; the mapping between Stokoe-style sign phonology and the outputs of real-time pose estimators; and the choice of a language model for the spoken-Russian decoder. The paper argues that a successful RSL-to-Russian (Sign-to-Text) system must couple linguistic knowledge — Stokoe phonology, predicate typology, mouthings and information structure — with engineering components in a principled way. In particular, it proposes a pipeline that combines RTMPose Whole-body as a real-time front-end, a Logos-pretrained vision encoder, a Vector Quantization sign tokenizer in the spirit of SignLLM and a language decoder from the GigaChat family. The contribution of the study is a literature-grounded synthesis of the state of the art with explicit identification of research lacunae and a justified component architecture for a target system.

Keywords: *Russian sign language, sign language recognition, sign language translation, synthetic data, large language models, GigaChat*

For citation:

Plavunov, Ilya N. & Yulia N. Ebzeeva. 2026. Artificial intelligence in Russian sign language recognition: Bridging sign phonology and translation systems. *Russian Journal of Linguistics* 30 (3). 531–556. <https://doi.org/10.22363/2687-0088-50181> EDN: NKOEUE

Искусственный интеллект в распознавании русского жестового языка: от фонологии жеста к архитектуре системы перевода

Илья Николаевич ПЛАВУНОВ¹, Юлия Николаевна ЭБЗЕЕВА²  

¹ПАО «Сбербанк»

²Российский университет дружбы народов, Москва, Российская Федерация

 ebzeeva-jn@rudn.ru

Аннотация

К 2026 г. автоматическое распознавание жестовых языков стало прикладным с появлением мобильных переводчиков, образовательных приложений и инклюзивных сервисов. Автоматическое распознавание русского жестового языка (далее РЖЯ) развивается, однако эта область остается фрагментарной и нечасто попадает в собственно лингвистическое поле зрения. Цель статьи тройная: систематизировать состояние исследований по распознаванию РЖЯ методами искусственного интеллекта в лингвистической перспективе; обозначить лауну — отсутствие непрерывного параллельного корпуса РЖЯ↔звучащий русский; предложить конкретную компонентную архитектуру системы перевода РЖЯ→русский, обоснованную опубликованными исследованиями. Обзор сосредоточен на семи блоках: фонология, морфология и просодика РЖЯ; корпусный разрыв между российскими и зарубежными ресурсами (BSL Corpus, DGS-Corpus, How2Sign, RWTH-PHOENIX-Weather в сопоставлении с корпусом НГТУ и датасетом TheRuSLan); эволюция архитектур искусственного интеллекта от скрытых марковских моделей до трансформеров и больших языковых моделей; развитие российских баз данных и систем распознавания (Slovo, Bukva и Logos); лингвистические критерии валидности синтетических жестовых данных; соответствие между фонологией жеста, предложенной У. Стоуки (Stokoe 2005), и выходами детекторов позы в режиме реального времени; выбор языкового модуля для декодирования в звучащий русский язык. В статье утверждается, что успешная система перевода РЖЯ→русский требует методически грамотного сопряжения лингвистических знаний — фонологии Стоуки, типологии предикатов, маусингов, информационной структуры — с инженерными компонентами. В частности, предлагается сквозная архитектура, объединяющая RTMPose Whole-body как модуль предобработки данных в реальном времени, зрительный энкодер на основе корпуса Logos, токенизатор жестов с векторным квантованием в стиле архитектуры SignLLM и языковой декодер из семейства GigaChat. Научный вклад исследования заключается в систематизации текущего состояния предметной области, эксплицитном выделении исследовательских лаун и представлении теоретически обоснованной компонентной архитектуры целевой системы.

Ключевые слова: *русский жестовый язык, распознавание жестового языка, перевод жестовой речи, синтетические данные, большие языковые модели, GigaChat*

Для цитирования:

Плавунов И.Н., Эбзеева Ю.Н. Искусственный интеллект в распознавании русского жестового языка: от фонологии жеста к архитектуре системы перевода. *Russian Journal of Linguistics*. 2026. Vol. 30. № 3. Р. 531–556. <https://doi.org/10.22363/2687-0088-50181>
EDN: NKOHEU

1. Введение

К 2026 г. автоматическое распознавание жестовых языков (*sign language recognition*) перешло из лабораторий в прикладную плоскость: появляются мобильные переводчики, образовательные приложения, инклюзивные сервисы. Для русского жестового языка (далее РЖЯ) эта область также развивается, однако складывается она в условиях характерной диспропорции. Лингвистическое описание жестов и жестовых языков, в том числе РЖЯ, к настоящему моменту представлено серией исследований по фонологии, синтаксису, морфологии, прагматике, информационной структуре, использованию жестов при осуществлении синхронного перевода (Boutet et al. 2016, Cienki 2024, Dargue et al. 2019, Iriskhanova 2023, Kimmelman 2019, Kok et al. 2015, Leonteva et al. 2023). Параллельно развиваются российские базы данных (датасеты) и системы распознавания. Эти две линии работают, по сути, изолированно: лингвистические результаты редко учитываются при проектировании систем, а архитектурные решения, в свою очередь, редко анализируются в категориях, привычных лингвисту.

Существующая обзорная литература эту изолированность не снимает. Сводные обзоры (Rastgoo, Kiani & Escalera 2021, Núñez-Marcos et al. 2023, Tan et al. 2024) систематизируют прогресс архитектур ИИ для жестовых языков и в более новой работе (Tan et al. 2024) включают раздел о решениях на основе больших языковых моделей, однако удерживаются в рамке компьютерного зрения и ориентированы на наиболее ресурсные жестовые языки: американский жестовый язык ASL (American Sign Language), немецкий жестовый язык DGS (Deutsche Gebärdensprache), британский жестовый язык BSL (British Sign Language). Российские данные и лингвистическая специфика РЖЯ в них не представлены. С другой стороны, лингвистические работы по РЖЯ обращаются к проблематике автоматической обработки лишь точно. В результате к настоящему времени нет исследования, в котором были бы системно соотнесены лингвистические свойства РЖЯ с архитектурными возможностями современных систем распознавания и перевода и одновременно учтены российские корпусная и инженерная ситуации.

Настоящая статья ставит задачу заполнить эту лакуну. Цель работы — систематизировать состояние исследований по распознаванию РЖЯ методами искусственного интеллекта (далее ИИ) с лингвистической точки зрения; обозначить главную эмпирическую проблему области — отсутствие непрерывного параллельного корпуса РЖЯ↔звучащий русский, сопоставимого по объему с зарубежными ресурсами уровня RWTH-PHOENIX-Weather или

How2Sign; и на основе этой систематизации предложить компонентную архитектуру системы перевода РЖЯ→русский, в которой каждое архитектурное решение опирается на опубликованную литературу по лингвистике РЖЯ и по соответствующему классу инженерных решений. Вклад статьи, таким образом, тройной: обзорно-аналитический, проблемно-диагностический и архитектурно-конструктивный.

2. Лингвистическая специфика РЖЯ как контур требований к системе

Чтобы корректно поставить задачу автоматической обработки РЖЯ, недостаточно сказать, что речь идёт о классификации жестов: набор лингвистических оппозиций, которые должна различать целевая система, существенно богаче, чем уровень изолированных лексем. Настоящий раздел очерчивает контур требований к системе, выводя его из публикаций по фонологии, морфологии, синтаксису и просодике РЖЯ.

2.1. Фонология жеста: пять параметров

Современное фонологическое описание жеста восходит к работе У. Стоуки (Stokoe 2005), в которой американский жестовый язык (American Sign Language, ASL) был впервые разложен на конечный набор параметров, аналогичных фонемам звучащих языков: *tabula* (место артикуляции), *designator* (форма руки) и *signation* (движение) (Stokoe 2005). К этой триаде в последующей литературе была добавлена ориентация ладони, а в современной сводке (Sandler & Lillo-Martin 2006) и в русскоязычном учебнике (Буркова, Киммельман 2019) фонология жеста описывается через пять параметров: форма руки (*handshape*), ориентация ладони (*orientation*), движение (*movement*), место артикуляции (*location*) и немануальный компонент (*non-manual signals*). Этот терминологический аппарат на русском языке принципиален для последующего изложения: все дальнейшее описание признакового пространства, извлекаемого автоматическими системами, будет опираться именно на сетку из пяти параметров. Программная рамка лингвистического исследования РЖЯ как самостоятельного языка была сформулирована ранее в работе Е.В. Прозоровой (2007), указавшей на ценность изучения жестовых языков для лингвистической теории и типологии. Данная методическая рамка применяется в настоящей статье и переносится в плоскость взаимодействия с автоматическими системами. Концептуальный мост между фонологией, предложенной У. Стоуки (Stokoe 2005), и выходами детекторов позы (данные, которые описывают положение частей тела) строится в разделе 7; здесь он лишь отмечается.

2.2. Морфология и грамматика: грамматические оппозиции для распознавания

Синтаксис РЖЯ типологически нетривиален. На корпусном материале и в эксперименте по описанию картинок (Kimmelman 2012) показано, что

базовый порядок слов в простых предикациях с обычными глаголами (*plain verbs*) — SVO (subject–verb–object), тогда как в конструкциях с классификаторными предикатами (*classifier predicates*) реализуется порядок SOV (subject — object — verb), а в локативных описаниях фиксируются комбинации OSV / SOV / OVS, объяснимые модально обусловленным фактором («сначала фон, потом фигура, потом локативное отношение»). На том же материале документируются *V-sandwich*-конструкции с повтором глагола в конце клаузы и дополнительной немануальной наречной. Из этого следует методологический вывод, существенный для архитектуры систем: распознавание непрерывной жестовой речи (*continuous sign language recognition*, CSLR) обязано сигнализировать тип глагола, поскольку от него зависит ожидаемый порядок аргументов и, следовательно, корректное выравнивание в декодере перевода.

Морфологический уровень добавляет еще одно требование. Burkova et al. (2019) описывают числовую инкорпорацию в обозначениях времени РЖЯ — формы типа 3-МЕСЯЦ (с инкорпорированным числительным три) противопоставлены аналитической комбинации МЕСЯЦ + 3, и такие оппозиции принципиальны для семантики высказывания (Burkova et al. 2019). Этот уровень детализации систематически не покрывается классификаторами изолированных глоссов, в которых все варианты лексемы МЕСЯЦ относятся к одному классу.

Еще одна линия исследований — пространственная модуляция. С. Лидделл (Liddell 2003) ввел понятия реального пространства (*real space*), токен-пространства (*token space*), суррогатного пространства (*surrogate space*), а также феномен удерживающих жестов, или «буев» (*buys*), заключающийся в удержании референта нерабочей рукой. Он показал, что направленные жесты — указательные знаки и так называемые согласовательные глаголы (*indicating verbs*) — выполняют не лексическую, а грамматическую функцию. Так, одна и та же лексема ВИДЕТЬ может быть направлена на разные точки жестового пространства, кодируя морфосинтаксические различия в субъекте и объекте (Liddell 2003). Из этого следует, что классификация изолированных глоссов в существующих российских базах данных (Kapitanov et al. 2023, Ovodov et al. 2025) систематически не различает грамматические оппозиции, передаваемые направленностью жеста.

2.3. Маусинги и немануальные компоненты

РЖЯ в типологическом отношении выделяется интенсивным использованием беззвучных ротовых жестов. Bauer & Kyuseva (2022) на материале корпуса НГТУ (Burkova 2012–2015) сформировали типологию ротовых действий (*mouth actions*) с 27 377 аннотациями и показали, что объем ротовых действий в РЖЯ превышает соответствующие показатели австралийского жестового языка (Auslan) и нидерландского жестового языка (Nederlandse Gebarentaal, NGT). Авторы различают маусинги (*mouthings*) — артикуляции, воспроизводящие звучащее русское слово или его часть, — и ротовые жесты

(*mouth gestures*), не связанные с фонетикой звучащего русского (Bauer & Kyuseva 2022). Из данной типологической особенности вытекает ключевое техническое условие: автоматическая система, ограниченная мануальным каналом, теряет снимающую неоднозначность (дисамбигуирующую) информацию ротового компонента.

Близкий по методологической установке прецедент в смежной лингвистической области — датасет K-RSL (Imashev et al. 2020), содержащий 28 250 видеозаписей казахско-российского жестового языка с явной аннотацией мануальных и немануальных компонентов. Авторы эмпирически показали, что включение немануальных признаков повышает точность распознавания. Работа примечательна тем, что одна из первых явно задает интердисциплинарную рамку «лингвистическое понимание + распознавание», к которой настоящий обзор присоединяется со стороны РЖЯ.

Информационная структура РЖЯ, систематически описанная в монографии В.И. Киммельмана (Kimmelman 2019), задает второй просодический пласт требований. Топик и фокус кодируются движением бровей, наклоном и поворотом головы, удлинением знака и открытием рта. Типологически контраст между РЖЯ и НГТ обнаруживается прежде всего в зоне контрастивного фокуса, а не в зоне топикизации. Для автоматической системы это означает, что позовые экстракторы (*pose-extractors*) должны покрывать лицо и губы — иначе целые грамматические категории остаются невыделенными.

2.4. Лексическая вариативность как методологическая проблема разметки

Вариативность РЖЯ создает самостоятельную методологическую проблему для автоматического распознавания. В.И. Киммельман и его соавторы (Kimmelman et al. 2022) на материале лексической базы вариативности РЖЯ (термины родства и названия цветов) показали, что традиционный принцип разграничения фонологических вариантов и отдельных лексем — «если два жеста различаются ровно по одному фонологическому параметру, это варианты, иначе — разные лексемы» — приводит к нетранзитивным цепочкам и потому неудовлетворителен. Авторы предлагают графовое представление вариативности и сетевой анализ (Kimmelman et al. 2022). Региональная вариативность — московский и сибирский варианты РЖЯ, представленные в корпусе НГТУ (Burkova 2012–2015), — добавляет к этому социолингвистическое измерение. Для систем распознавания, обучаемых на изолированных глоссах, это означает, что разработчики обязаны принимать конвенциональные решения о том, какие варианты считать одной меткой, а какие — разными. Один из явных методологических ответов на эту проблему — введенная в Logos аннотация визуально близких групп знаков (*visually similar sign groups*, VSSign), которая фиксирует пары и тройки лексем, неразличимых по визуальным признакам (Ovodov et al. 2025).

Совокупный контур требований, выводимый из этого раздела, может быть сформулирован следующим образом. Система автоматического распознавания РЖЯ должна одновременно различать пять фонологических параметров жеста, учитывать тип предиката для синтаксической интерпретации, обрабатывать многоканальный сигнал (мануал, немануал, ротовой канал) и принципиально решать проблему лексической и региональной вариативности. К этому контуру будут возвращаться все следующие разделы, проверяя на соответствие каждой компонент архитектуры.

3. Корпусная база жестовых языков и положение РЖЯ

Эмпирическая база, на которой обучаются современные модели распознавания и перевода жестовых языков, складывается из аннотированных корпусов. Сопоставление российских и зарубежных корпусов дает количественный показатель отставания, которое в дальнейшем определяет архитектурные ограничения.

3.1. Международная корпусная база

Корпус британского жестового языка (BSL Corpus), созданный Британской Ассоциацией Глухих (2008–2011), включает данные 249 информантов в четырех жанрах — личные нарративы, свободные диалоги, интервью о принадлежности к сообществу глухих и лексическую эликситацию по 101 концепту, — с разметкой в ELAN и системой идентификационных глоссов (Schembri et al. 2013). Корпус немецкого жестового языка (DGS-Korpus), разрабатываемый Институтом немецкого жестового языка и коммуникации глухих при Гамбургском университете в рамках программы Академии наук, охватывает 330 информантов из 13 регионов Германии, при этом общий объем отснятого материала — около 560 часов разговорно-нарративных данных, из которых более 50 часов с переводом и глоссами, доступны в публичном релизе MY DGS — annotated (Konrad et al. 2020).

Стандартным эталонным набором данных для непрерывного распознавания и перевода жестовой речи остается корпус RWTH-PHOENIX-Weather: на материале телепрогнозов погоды собрано около 11 часов параллельного жестово-немецкого материала, словарь ≈ 1066 знаков, 8257 видеосегментов (Forster et al. 2014). Именно этот корпус с 2014 г. лежит в основании всех ключевых сопоставительных работ по архитектурам — от классических скрытых марковских моделей (Koller et al. 2015) до трансформерных подходов (Camgöz et al. 2020). Для ASL аналогичную роль играет How2Sign — параллельный корпус более 80 часов жестового и звучащего материала с 35 000+ предложений, причем трехчасовой подкорпус снят в студии Panoptic с 500+ камерами и пригоден для трехмерной разметки SMPL-X (Duarte et al. 2021).

3.2. Российские корпусные ресурсы

Базовый эмпирический ресурс по РЖЯ — онлайн-корпус, созданный в Новосибирском государственном техническом университете (далее НГТУ) в 2012–2015 гг. под руководством С.И. Бурковой (Burkova 2012–2015). Корпус включает московский и сибирский варианты РЖЯ, ELAN-разметку с глоссами по правой и левой руке, маусингами и отдельными немануальными подмодулями (брови, глаза, голова) и служит источником данных для последующих корпусных работ (Bauer & Kyuseva 2022, Burkova et al. 2019, Kimmelman 2012, 2019, Kimmelman et al. 2022). Параллельная линия — мультимедийная база данных, разработанная в лаборатории А.А. Карпова Санкт-Петербургского федерального исследовательского центра РАН (далее СПбФИЦ): Kinect 2.0, синхронные RGB-, depth- и инфракрасные каналы, 164 жеста, 13 информантов; материал доменно ограничен тематикой «супермаркет» (Kagirov et al. 2020). Та же база представлена в работе И.А. Кагирова и его коллег (Кагиров и др. 2020) в развернутом лингвистически ориентированном описании с фонологической аннотацией.

3.3. Количественное сопоставление данных, представленных в корпусах жестовых языков

Количественные пропорции наглядны: BSL Corpus — 249 информантов, DGS-Korpus — 330 информантов и более 50 часов в публичном доступе, корпус RWTH-PHOENIX-Weather — около 11 часов параллельного материала, корпус How2Sign — более 80 часов. Российские корпуса в этом измерении пока несопоставимы: корпус НГТУ существенно меньше по абсолютным масштабам, TheRuSLan содержит 13 информантов и 164 жеста. Ключевой структурный пункт состоит в том, что ни один из существующих российских ресурсов не является непрерывным корпусом со звучащим переводом. Изолированные знаки описаны в Slovo, Bukva и Logos, однако переход от изолированного распознавания знаков (*isolated sign language recognition*, ISLR) к непрерывному распознаванию (CSLR) и переводу жестовой речи (*sign language translation*, SLT) требует именно непрерывного параллельного материала уровня PHOENIX. Единственный непрерывный набор данных РЖЯ, упоминаемый в сравнительной таблице Slovo (Kapitanov et al. 2023), не имеет той корпусно-лингвистической верификации, которой обладает PHOENIX. Тем самым корпусный разрыв формируется не как «отставание», а как структурный барьер: без непрерывного параллельного корпуса РЖЯ↔русский невозможно обучить модель, сопоставимую по качеству с трансформерными решениями уровня BLEU-4 $\approx 21,80$ (Camgöz et al. 2020). Этот вывод задает повестку для разделов 6 (синтетические данные) и 8 (целевая архитектура с переносом).

4. Эволюция архитектур искусственного интеллекта для распознавания и перевода жестовых языков

Для соотнесения предложения о целевой архитектуре с историей ИИ, в этом разделе кратко прослеживается эволюция системных решений и фиксируется понятийная триада: изолированное распознавание жестового языка (ISLR), непрерывное распознавание жестового языка (CSLR) и перевод жестового языка (SLT).

4.1. Понятийная триада: ISLR, CSLR, SLT

Под изолированным распознаванием знаков (ISLR) понимается классификация одного жеста, представленного отдельным видеосегментом; под непрерывным распознаванием (*continuous sign language recognition*, CSLR) — распознавание последовательности глоссов в неразорванной жестовой речи; под переводом жестовой речи (*sign language translation*, SLT) — порождение текста на звучащем языке с учетом грамматических различий между жестовым и звучащим языками. Это различие проведено в работе (Camgöz et al. 2018), в которой авторы обосновывают переход от распознавания глоссов к собственно переводу как самостоятельной задаче, не сводимой к классификации последовательностей.

4.2. Архитектурная эволюция

В период 2010-х основной парадигмой CSLR оставались скрытые марковские модели в комбинации с временной классификацией без выравнивания (*Connectionist Temporal Classification*, CTC) и признаками, извлекаемыми классическими методами (Active Appearance Models, отдельные ручные трекеры). Ведущей разработкой того времени стала работа (Koller et al. 2015): авторы зафиксировали базу данных RWTH-PHOENIX-Weather как стандартный эталон и впервые открыто обозначили проблему зависимости системы от конкретного исполнителя жестов (*signer dependency*). Переход к нейросетевой парадигме был оформлен в публикации (Camgöz et al. 2018), авторы которой объединили сверточные сети (*convolutional neural networks*, CNN) с рекуррентными сетями (*recurrent neural networks*, RNN) и механизмом внимания (*attention*), а параллельный выпуск корпуса RWTH-PHOENIX-Weather 2014T открыл доступ к первому публичному корпусу с переводом на немецкий язык (Camgöz et al. 2018). На том же контрольном наборе данных метрика качества машинного перевода BLEU-4 (*Bilingual Evaluation Understudy*) для предложенной модели составила 18,13 при доступе к gloss-уровню и 9,58 без него.

Современный стандарт CSLR/SLT был оформлен в работе (Camgöz et al. 2018), где архитектура трансформера с функцией потерь CTC (*Connectionist Temporal Classification*) была применена для совместного решения задач распознавания (CSLR) и перевода жестовой речи (SLT) в рамках единой

архитектуры. Метрика BLEU-4 на корпусе PHOENIX14T составила 21,80, что более чем удвоило ранее достигнутые значения и зафиксировало планку для последующих исследований. Систематические обзоры (Rastgoo et al. 2021, Núñez-Marcos et al. 2023) описывают семейство архитектур этого периода — сверточные и трехмерные сверточные сети, рекуррентные сети с долгой краткосрочной памятью (*long short-term memory*, LSTM), графовые скелетные сети, мультимодальные модели — и фиксируют общие проблемы: отсутствие крупных параллельных корпусов, зависимость от исполнителя (*signer dependency*), а также невозможность прямого переноса моделей между разными жестовыми языками.

С 2024 г. в данной области наметился переход к использованию больших языковых моделей (*large language model*, LLM; далее — БЯМ) в качестве декодера. Архитектурный паттерн SignLLM (Gong et al. 2024) состоит в том, что между сырым видео и БЯМ ставится модуль векторного квантования (*Vector Quantization*, VQ): каждый отрезок видео превращается в дискретный токен из обучаемой книги кодов (кодбука). Последовательность таких токенов подается на вход БЯМ как «текст на жестовом языке». На тестовых наборах данных PHOENIX14T и CSL-Daily модель SignLLM демонстрирует наилучшие на сегодняшний день результаты (*state-of-the-art*) без использования *gloss*-аннотации, что методически принципиально для языков с дефицитной разметкой, т.е. для малоресурсных языков. Терминологическое замечание: под названием «SignLLM» в литературе известна и другая модель — SignLLM, ориентированная на мультязычную генерацию жестовой речи (*sign language production*) из текста (Fang et al. 2024). В настоящей статье под SignLLM везде понимается исключительно архитектура Gong et al., предназначенная для решения обратной задачи — перевода жестовой речи в звучащий язык (Gong et al. 2024).

4.3. Следствия для исследований РЖЯ

Современный уровень качества SLT (трансформер + CTC, BLEU-4 $\approx$ 21,80) предполагает наличие порядка 11 часов параллельного материала; однако для РЖЯ такого ресурса не существует. Соответственно, обоснованными альтернативными путями оказываются: *gloss-free* обучение без использования глосс (*gloss-free*) с БЯМ-декодером в стиле SignLLM; перенос обучения через универсальный жестовый энкодер; а также синтетическое расширение (аугментация) корпуса. Все три направления обсуждаются в следующих разделах и интегрируются в архитектурное предложение, представленное в разделе 8.

5. Распознавание РЖЯ: российские датасеты и системы

Представление о том, что уже сделано в России в области автоматической обработки РЖЯ, складывается из двух исследовательских линий —

петербургской лаборатории А.А. Карпова в СПбФИЦ и московской лаборатории искусственного интеллекта Сбера (далее СберАИ). Их разнесенность по задачам, аппаратным платформам и методологии формирует достаточно полную картину покрытия задач — и тем явственнее очерчивает оставшиеся лакуны.

5.1. Группа А. А. Карпова (СПбФИЦ)

Линия команды А.А. Карпова в СПбФИЦ имеет более длинную предысторию, существенную для постановки настоящего обзора. Еще в 2011 г. А.А. Карпов описал характеристики РЖЯ, дал аналитический обзор существующих в мире систем жестовой нотации и систем синтеза «текст–в–жесты» и предложил концепцию универсальной мультимодальной системы аудиовизуального синтеза звучащей и жестовой речи по тексту на русском языке (Карпов 2011). Данное исследование послужило одной из первых работ на стыке лингвистики РЖЯ и компьютерной обработки данных, фактически заложив основу для цикла изысканий, развитием которых стали проект TheRuSLan и последующий обзор. Помимо самого датасета TheRuSLan (раздел 3), исследовательский коллектив под руководством А.А. Карпова представил его лингвистически ориентированное описание с фонологической аннотацией и отдельной интерпретацией трехмерного материала, что важно для последующего изучения параметров жеста по depth-картам (Кагиров и др. 2020). В сводном виде состояние области на 2021 г. зафиксировано в обзоре Д.А. Рюмина, И.А. Кагирова, А.А. Аксенова и А.А. Карпова (Рюмин и др. 2021). Это первый русскоязычный академический обзор моделей и методов автоматического распознавания жестов и жестовых языков, в котором систематизированы проблемы детектирования артикуляторов, распознавания их конфигурации и сегментации жестов в потоке речи, а также сделан вывод о перспективности двухпоточных сверточных и рекуррентных архитектур. Данная работа остается в инженерной плоскости и предшествует эпохе Slovo, Bukva, Logos, SignLLM и больших языковых моделей. Настоящая статья продолжает заданный вектор исследований, дополняя его лингвистической рамкой (раздел 2), синтезом российских датасетов поколения 2023–2025 гг. (раздел 5.2) и компонентной архитектурной интеграцией с семейством GigaChat (раздел 8).

Еще одним значимым направлением исследований данного коллектива является мультимодальное распознавание речи и жестов на сенсорах мобильных устройств. Так, в работе (Ryumin et al. 2023) предложены две глубокие нейросетевые архитектуры, реализующие аудиовизуальное распознавание речи с интеграцией жестового канала, а также три стратегии слияния каналов: на уровне предсказания, признака и модели. С точки зрения настоящего обзора эта линия дает два аргумента: во-первых, российский подход к созданию систем реального времени (realtime-систем) на одной RGB-камере смартфона технически реализуем; во-вторых, для РЖЯ-системы целесообразно

изначально планировать мультимодальное слияние (fusion) звукового и видеопотока, а не только обработку видеопотока.

5.2. Группа СберAI

Slovo — крупный краудсорсинговый ISLR-датасет РЖЯ: 20 000 видеозаписей FullHD, 1000 классов изолированных жестов, 194 информанта (Kapitanov et al. 2023). Сбор организован на двух краудсорсинговых платформах — Yandex Toloka для записи и ABC Elementary для валидации и временной разметки; перед каждым этапом краудворкеры проходят экзамен по РЖЯ с порогом не ниже 80 %. Каждое видео длится 1–4 секунды; для участков, на которых сигнализирующий еще готовится к жесту или уже его завершил, введен отдельный класс «no event». Авторы подчеркивают, что в выборку вошли наиболее частотные жесты без дактилологии и сложных составных конструкций. Эту нишу — дактиль РЖЯ — закрывает Bukva: 3757 видеозаписей, 33 класса, соответствующих буквам алфавита (включая динамические — *й, ж, х*), 155 глухих и слабослышащих экспертов; архитектура распознавания опирается на Temporal Shift Module поверх MobileNet с инференсом в реальном времени на центральном процессоре (Kvanchiani et al. 2024). Тем самым целевая система РЖЯ→русский получает дактильный модуль, без которого передача имен собственных и терминов невозможна.

Logos занимает в этой линии ключевую методологическую позицию. По заявленным авторами характеристикам, это самый крупный публичный датасет ISLR в мире по числу сигнализирующих — 381 информант (Ovodov et al. 2025). Принципиальная инновация — введение явной аннотации визуально близких групп знаков (*visually similar sign groups*, VSSign): для жестов, неразличимых по чисто визуальным параметрам, но имеющих разные значения в зависимости от контекста, вводится отдельная разметка, что снимает часть проблемы лексической вариативности (Kimmelman et al. 2022). Предложенная архитектура опирается на базовую модель (бэкбон) MViTv2-S с инициализацией Kinetics-400, классификатор на полносвязном слое, кросс-энтропию со сглаживанием меток (label smoothing) и вспомогательную задачу регрессии временных границ знака. Главный для настоящей статьи результат таков: модель, предобученная на корпусе Logos, превосходит наилучшие достижения (state-of-the-art) на датасете WLASL-2000 (американский жестовый язык) и демонстрирует конкурентоспособные результаты (competitive-показатели) на AUTSL (турецкий жестовый язык). Авторы прямо утверждают, что модель, предобученная на Logos, может использоваться как универсальный энкодер для других жестовых языков, в том числе в режиме обучения по нескольким примерам (few-shot learning). Данный факт эмпирически меняет традиционное представление о РЖЯ как о малоресурсном (low-resource) языке, заимствующем из DGS / ASL: предобучение на материале РЖЯ оказывается *донором* для других жестовых языков.

В индустриальной плоскости картину СберАИ дополняет технический обзор подходов к распознаванию и переводу жестовой речи, опубликованный командой SberDevices в формате блог-публикации на платформе Habr (Петрова и др. 2024). Обзор не является академической публикацией и не имеет лингвистической рамки; в нем детально описаны пять архитектурных решений — TwoStream-SLR/SLT, Multi-Modality Transfer Learning Baseline, Self-Mutual Distillation Learning, SignBERT+ и CorrNet — оцениваемых на бенчмарках PHOENIX-2014T и CSL-Daily. Первоисточники этих архитектур указаны в самом обзоре и здесь не воспроизводятся. Настоящая статья соотносится с этим индустриальным взглядом не как с альтернативой, а как с дополнением, причем расхождение между ними носит не риторический, а методический характер и объясняется тремя факторами.

Во-первых, **парадигмальный сдвиг**: все пять архитектур, рассматриваемых в работе Е. Петровой, У. Быковой, К. Кванчани (Петрова и др. 2024), остаются в gloss-based линии — каждая из них решает задачу в формате Sign2Gloss с функцией потерь CTC, после чего глосс-цепочка переводится в текст модулем mBART-large-cc25. Настоящая статья выбирает gloss-free пайплайн с VQ-токенизатором и БЯМ-декодером (раздел 8), что обусловлено отсутствием для РЖЯ непрерывного gloss-аннотированного корпуса, сопоставимого с PHOENIX-2014T или CSL-Daily, и невозможностью в этих условиях обучить gloss-based архитектуру на нужном масштабе данных.

Во-вторых, **выбор декодера**: индустриальный обзор удерживает mBART-large-cc25 как стандартный многоязычный модуль перевода глосс в текст, тогда как настоящая статья мотивированно переходит к семейству GigaChat (GigaChat team 2025) по трем лингвистически осмысленным критериям — преобладание русского в обучающих данных, длина контекстного окна и гибкость развертывания (раздел 8.2).

В-третьих, **хронология компонентов**: ключевые элементы предлагаемого пайплайна опубликованы уже после Habr-обзора 13 февраля 2024 г. — паттерн SignLLM (Gong et al. 2024) представлен в апреле 2024 г. на CVPR; датасет Logos и универсальный энкодер на его основе (Ovodov et al. 2025) опубликованы в мае 2025 г. на arXiv и в ноябре 2025 г. на EMNLP; GigaChat как семейство с описанной архитектурой и открытыми весами (GigaChat team 2025) представлен в июне 2025 г.; алфавитный модуль Bukva (Kvanchiani et al. 2024) — в октябре 2024 г.

Таким образом, упомянутый индустриальный обзор и настоящая статья не противоречат друг другу, а продолжают одну исследовательскую линию, но в принципиально иных методологических координатах: первая работа представляет собой обзор архитектур мирового уровня на основе глосс (survey gloss-based) по состоянию на начало 2024 г., вторая — лингвистически обоснованное архитектурное предложение для РЖЯ→русский, использующее работы 2024–2025 гг.

5.3. Решенные задачи и оставшиеся лакуны

В совокупности российские исследования включают: изолированное распознавание жестового языка (ISLR) на крупном словаре (корпус Logos), распознавание дактильного алфавита (Bukva), мобильную аудиовизуальную архитектуру (Ryumin et al. 2023), а также разметку карт глубины (depth-разметку) (датасет TheRuSLan) (Кагиров и др. 2020). В то же время оставшимися лакунами являются: непрерывный параллельный корпус РЖЯ↔русский, синтаксическая разметка с типом предиката (Kimmelman 2012), просодическая разметка немануальных маркеров (Kimmelman 2019), морфологическая разметка инкорпорированных числительных (Burkova et al. 2019). Корпус Logos в этой картине — точка опоры для тезиса о компонентной архитектуре: универсальность энкодера обеспечивает перенос обучения, частично компенсирующий корпусный дефицит непрерывного материала.

6. Синтетические данные как ответ на корпусный дефицит

Корпусный разрыв, описанный в разделе 3, ставит вопрос о том, можно ли частично компенсировать его генеративными методами. Ответ оказывается положительным, но обусловленным лингвистическими критериями валидности.

6.1. Генерация жестовой речи: обратная задача распознавания

Первой архитектурой современного типа для генерации жестовой речи (*sign language production*, SLP) стала архитектура Progressive Transformers (Saunders et al. 2020) на основе трансформера с механизмом встречного декодирования (*counter-decoding*) для порождения непрерывных трехмерных последовательностей поз по тексту, с инициализацией скелета OpenPose-выходами PHOENIX14T и поднятием в 3D через инверсную кинематику. Существенным шагом вперед стала работа тех же авторов (Saunders et al. 2021), в которой базовая архитектура была расширена до многоканальной генерации (одновременного моделирования мануального и немануального каналов) с использованием смесей плотностей (*Mixture Density Networks*) для моделирования стохастичности и просодической вариативности. Авторы прямо характеризуют одноканальную генерацию как порождающую «робота» и потому непригодную в качестве обучающего сигнала для распознавания.

6.2. Диффузионные модели и 4D-аватары

В 2024 г. область сделала шаг в сторону диффузионных моделей: Neural Sign Actors (Baltatzis et al. 2024) — диффузионная модель, обученная на курированном крупном датасете 4D-аватаров с английскими транскриптами (т.е. с текстовыми описаниями поз, жестов или артикуляции); на выходе — параметры SMPL-X для каждого кадра, обусловленные на текст. Полученные

жестовые последовательности визуально реалистичнее решений на основе скелетных моделей (skeleton-based), представленных в (Saunders et al. 2020, 2021). Однако для РЖЯ это направление упирается в отсутствие аналогичного 4D-датасета на русском материале: 3D-подкорпус How2Sign из 500-камерной студии Panoptic (Duarte et al. 2021), на материале которого обучены сравнимые модели для ASL, не имеет российского аналога.

6.3. Лингвистические критерии валидности синтетических данных

Из требований, выводимых в разделе 2, и из сводных результатов по маусингам (Bauer & Kyuseva 2022) и информационной структуре (Kimmelman 2019) следует список условий, без которых синтетический материал не пригоден для обучения CSLR/SLT-моделей. Многоканальность: одновременное порождение мануальной составляющей, немануальных компонентов и ротового канала. Просодическая непрерывность: вариативность движений бровей и наклонов головы должна быть градуальной, а не сводимой к дискретным маркерам. Региональная репрезентация: представлены должны быть, как минимум, московский и сибирский варианты РЖЯ. Грамматическая разметка: синтетический материал должен явно различать обычные и классификаторные (входящие в классификаторные конструкции) глаголы (Kimmelman 2012). При нарушении этих условий синтетические данные могут улучшить ISLR-классификатор, но не обеспечат требуемого качества для CSLR/SLT.

6.4. Соотнесение с целевой архитектурой

В целевой системе синтетические данные следует использовать на этапе предобучения зрительного энкодера для расширения покрытия лексики и просодической вариативности; они *не заменяют* реальный корпус РЖЯ для верификации, и финальная оценка качества системы должна проводиться на естественной речи носителей.

7. Распознавание поз и рук в реальном времени: соотнесение с фонологией жеста

Настоящий раздел является центральным концептуальным мостом статьи: именно здесь компьютерное зрение становится лингвистически осмысленным. Аргумент состоит в том, что выходы современных детекторов позы и рук в реальном времени (*real-time pose and hand estimation*) прямо соответствуют пяти фонологическим параметрам жеста, и именно благодаря этому они могут служить признаковым пространством, понятным как лингвисту, так и инженеру.

7.1. Эволюция инструментов

Первой широкодоступной системой оценки позы человека в реальном времени стал инструмент OpenPose (Cao et al. 2021). Данная открытая

архитектура предназначена для многопоточного двумерного оценивания позы человека с использованием полей сродства частей тела (Part Affinity Fields) и позволяет за один проход извлекать ключевые точки тела, кистей рук и лица (body-, hand- и face-ландмарки). Над ее выходами обучается семейство классических моделей распознавания действий — прежде всего пространственно-временные графовые сверточные сети (Spatial-Temporal Graph Convolutional Networks, ST-GCN) (Yan et al. 2018), ставшие канонической архитектурой для скелетных методов. Параллельно Google Research предложил фреймворк MediaPipe Hands. Данное решение для локального выполнения на устройствах (on-device) базируется на двух моделях: детекторе ладоней (palm detector) и модели определения ключевых точек кисти (hand landmark model). Архитектура позволяет в режиме реального времени (real-time inference) извлекать по 21 точке для каждой кисти руки в формате 2.5D на мобильных графических процессорах (Zhang et al. 2020).

Современный SOTA для распознавания в режиме реального времени позы всего тела — RTMPose (Jiang et al. 2023). Ключевые числовые характеристики: 75,8 % AP (Average Precision) на COCO при более чем 90 кадрах в секунду на центральном процессоре Intel; 67,0 % AP на COCO-WholeBody при более чем 130 кадрах в секунду на графическом процессоре; 72,2 % AP на COCO при более чем 70 кадрах в секунду на смартфонном чипе Snapdragon 865. Whole-body-вариант выдает 21 точку на каждую кисть — то же количество, что и MediaPipe Hands.

7.2. Соответствие пяти фонологических параметров и выходов детекторов

Фонологическая сетка из пяти параметров (Буркова, Киммельман 2019, Stokoe 2005) непосредственно отображается на выходы перечисленных детекторов. Форма руки (*handshape*) реконструируется из относительных координат 21 точки кисти; ориентация ладони — из нормали к плоскости, вычисляемой по тем же ключевым точкам; место артикуляции — из позиции запястья относительно скелета корпуса; движение — из временной производной координат; немануальный компонент — из ландмарок лица и позиции головы. Таким образом, выбор RTMPose Whole-body или функционально близкого инструмента в качестве модуля предобработки оказывается не «инженерным предпочтением», а отражением пятипараметрового описания жеста.

7.3. Ограничения для лингвистической разметки

При этом современные инструменты, работающие в режиме реального времени (real-time), не пригодны в качестве непосредственного источника тонкой просодической разметки. Так, А. Кузнецова и В.И. Киммельман (Kuznetsova & Kimmelman 2024) показали, что фреймворк MediaPipe Holistic не обеспечивает достаточной точности при отслеживании движений бровей и наклонов головы на материале казахско-русского жестового языка.

Исследование проводилось на базе корпуса K-RSL (Kazakh-Russian Sign Language), разработанного в рамках междисциплинарного датасета K-RSL (Imashev et al. 2020) с эксплицитной разметкой мануальных и немануальных компонентов. Авторы пришли к выводу, что без коррекции и дообучения данный алгоритм непригоден для решения лингвистических задач, требующих максимальной точности. Соответственно, если для классификации лексики (Slovo, Bukva, Logos) точность MediaPipe является приемлемой, то для лингвистической разметки немануальной просодики целесообразен переход на более точные детекторы позы (например, RTMPose Whole-body) и/или разработка специальной корректирующей модели поверх извлеченных ключевых точек.

8. Целевая архитектура системы РЖЯ→русский: компонентный подход и выбор языкового модуля

Настоящий раздел сводит предыдущие тезисы в конкретное архитектурное предложение. Сначала описывается методический шаблон, на который опирается синтез — гибридная схема с БЯМ-декодером; затем обосновывается выбор семейства языковых моделей; затем формулируется компонентная архитектура целевой системы; наконец, обсуждаются открытые компоненты и условие масштабирования.

8.1. SignLLM как методический шаблон

Архитектурный паттерн, на который опирается предлагаемая система, последовательно описан под названием SignLLM (Gong et al. 2024). Между сырым видео жестовой речи и БЯМ-декодером ставится модуль векторного квантования (*Vector-Quantized Visual Sign module*): каждый отрезок видео отображается в дискретный токен из обучаемого кодбука; последовательность таких токенов представляет сжатое визуальное содержание жестового высказывания. Дополнительный модуль реконструкции и выравнивания книги кодов (Codebook Reconstruction and Alignment) преобразует токены символического уровня (character-level) в представления словесного уровня (word-level), после чего вход подается в БЯМ — в исходной работе LLaMA или T5 — и тот порождает перевод на звучащий язык. На тестовых наборах данных (бенчмарках) PHOENIX14T и CSL-Daily SignLLM демонстрирует наилучшие на сегодняшний день результаты (state-of-the-art) без привлечения поуровневой аннотации (gloss-аннотации). Принципиальная роль БЯМ-декодера состоит в том, что он компенсирует отсутствие эксплицитной разметки глосс за счет собственных языковых способностей, сформированных предобучением на текстах, и тонкой настройки на парах (VQ-токены, перевод). Для РЖЯ это особенно важно: разметка глосс существует только для изолированных жестов (корпус Logos), но отсутствует для непрерывных жестовых дискурсов.

8.2. Выбор языкового модуля: семейство GigaChat

В качестве языкового декодера для системы РЖЯ→русский предлагается семейство моделей GigaChat (GigaChat team 2025) — российская семья БЯМ с публично описанной архитектурой, открытой базовой версией и старшими проприетарными моделями, доступными через программный интерфейс. Обоснование выбора опирается на три лингвистически осмысленных критерия и сознательно остается в этой плоскости.

Прежде всего, **ориентация на русский язык**. Многоязычные модели обучаются преимущественно на английских и западноевропейских текстах, и русский в их обучающих данных занимает периферийную позицию. Семейство GigaChat, напротив, разработано как русскоориентированное: модель спроектирована с нуля для работы с русским языком, доля русского в обучающем корпусе существенно выше, чем у многоязычных аналогов сопоставимого размера, и итоговое качество оценивается прежде всего на материале русского языка (GigaChat team 2025). Для задачи перевода жестовой речи в звучащий русский это означает более высокую языковую беглость в порождаемом тексте — в частности, в морфологически и синтаксически сложных конструкциях. Во-вторых, **длинный контекст**. Контекстное окно старшей открытой модели GigaChat-A3B составляет 131 072 токена; при типичной плотности РЖЯ 2–4 знака в секунду перевод даже минутных высказываний требует контекстного окна, существенно превышающего границы одного предложения; такая длина контекста обеспечивает устойчивую работу и в минутных и более длинных жестовых дискурсах. В-третьих, **гибкость развертывания**. GigaChat доступен в двух режимах: открытые веса базовой модели — для локального дообучения под жестовые VQ-токены, что критично при ограничениях, связанных с приватностью данных носителей жестовой речи, — и доступ через программный интерфейс к старшим проприетарным версиям семейства для прототипирования и сравнительной оценки качества. Эта двойственность отличает GigaChat от ряда англоязычных альтернатив, доступных только через программный интерфейс. Технические детали реализации — внутренняя структура смеси экспертов (Mixture of Experts, MoE), алгоритмы тонкой настройки (fine-tuning), оценки углеродного следа (выбросов CO₂) при обучении лежат за рамками лингвистической интерпретации и описаны в первоисточнике (GigaChat team 2025); в настоящем обзоре они не воспроизводятся.

8.3. Целевая компонентная архитектура

Совокупный пайплайн системы РЖЯ→русский может быть описан пятью последовательными компонентами. (1) **Модуль предобработки** — RTMPose Whole-body (Jiang et al. 2023) для извлечения скелета корпуса, ландмарок 21 точки на каждую кисть и ландмарок лица в режиме реального вре-

мени; этот выбор согласуется с пятипараметровым фонологическим описанием жеста (раздел 7) и обеспечивает покрытие немануального и ротового каналов, требуемое разделом 2. (2) **Зрительный энкодер** — архитектура MViTv2-S, как в Logos (Ovodov et al. 2025), с инициализацией Kinetics-400 и дообучением на Logos для функционирования в качестве универсального жестового энкодера; дополнительная адаптация — переход от ISLR-режима к непрерывному потоку с помощью скользящего окна и вспомогательной задачи регрессии границ знака, реализованной в Logos. (3) **Синтетическая аугментация** — на этапе предобучения энкодера используются данные, порожденные многоканальным SLP (Saunders et al. 2021) и/или диффузионной моделью (Baltatzis et al. 2024); условия лингвистической валидности заданы в разделе 6. (4) **Токенизатор жестов** — модуль векторного квантования по образцу архитектуры SignLLM (Gong et al. 2024) и обучаемый на материале корпуса Logos и на ограниченном непрерывном корпусе РЖЯ; на выходе формирует последовательность дискретных «жестовых токенов». (5) **Языковой декодер** — модель из семейства GigaChat (GigaChat team 2025), функционирует либо на основе открытой базовой версии для локального дообучения на парах ⟨VQ-последовательность, русский перевод⟩ при ограничениях по приватности данных, либо на базе старших проприетарных моделей семейства, подключаемых через программный интерфейс при доступности облачной инфраструктуры. На выходе генерируется текст на звучащем русском языке.

8.4. Открытые компоненты и условие масштабирования

Архитектура реализуется преимущественно на открытых компонентах: RTMPose доступен под лицензией Apache, Logos — открытый ресурс с публичным предобучением Kinetics-400, MViTv2 — открытая модель, паттерн SignLLM описан в открытой публикации. Языковой модуль — GigaChat — также допускает развертывание в двух режимах. Принципиальное бутылочное горлышко системы (bottleneck) — параллельный корпус ⟨непрерывное РЖЯ-видео, русский перевод⟩. Без него тонкая настройка языкового декодера невозможна на нужном масштабе; режим программного интерфейса в этом случае может использоваться лишь в качестве базового решения (baseline) без доменной адаптации. Соответственно, формирование непрерывного параллельного корпуса РЖЯ↔русский следует считать исследовательской задачей номер один для российского научно-инженерного сообщества в области распознавания жестовых языков.

9. Открытые задачи

Задачи, формирующие повестку российских исследований в области распознавания РЖЯ и научно обоснованные в существующей литературе, могут быть резюмированы в следующих пяти пунктах. Прежде всего, **непрерывный параллельный корпус РЖЯ↔звучащий русский** уровня DGS-Korpus

(Konrad et al. 2020) или How2Sign (Duarte et al. 2021) — от 50 часов с ELAN-разметкой, переводом и метаданными региональной принадлежности сигнализирующих. Без такого ресурса ни SLT-системы уровня RWTH-PHOENIX-Weather (Forster et al. 2014, Camgöz et al. 2020), ни тонкая настройка языкового декодера в архитектуре раздела 8 невозможны на нужном масштабе. Во-вторых, **4D-разметка РЖЯ** — российский аналог Panoptic-подкорпуса How2Sign и датасета Neural Sign Actors (Baltatzis et al. 2024) для обучения многоканальных SLP-моделей и диффузионных аватаров. В-третьих, **просодическая автоматическая разметка через специализированные детекторы позы**, развивающая результаты А. Кузнецовой и В.И. Киммельмана (Kuznetsova & Kimmelman 2024): обучение корректирующих моделей, функционирующих поверх архитектуры RTMPose Whole-body для извлечения тонких просодических признаков информационной структуры. В-четвертых, **региональная репрезентативность**: существующие российские датасеты собирались преимущественно методами краудсорсинга (crowdsourcing), без эксплицитной стратификации (выборочного контроля) по регионам РЖЯ, и московский вариант, по-видимому, в них преобладает; целенаправленный сбор по сибирскому варианту и крупным региональным центрам остается открытой задачей. В-пятых, **грамматическая разметка как вспомогательная задача**: включение в архитектуру Logos и его наследников вспомогательных классификационных голов — тип предиката (обычный глагол vs классификаторный), наличие инкорпорированного числительного, наличие маусинга на основе слов русского языка — для повышения точности синтаксически сложных конструкций и интерпретируемости выходов системы.

10. Заключение

Проведенный обзор позволяет свести разнородный материал — лингвистическое описание РЖЯ, корпусную картину, эволюцию архитектур ИИ, российские датасеты и системы, синтетические данные и детекторы позы человека — в связную картину, на основании которой формулируется итоговая архитектура целевой системы.

На лингвистическом уровне обзор показывает, что РЖЯ — самостоятельная языковая система, не сводимая к ASL и DGS ни фонологически, ни синтаксически, ни в части немануального и ротового каналов; типологические данные о порядке слов, числовой инкорпорации, маусингах и информационной структуре, накопленные за последние полтора десятилетия, формируют контур требований, которому должна отвечать любая система распознавания и перевода РЖЯ. На корпусном уровне видно, что российская ресурсная база беднее зарубежной и, главное, не содержит непрерывного параллельного корпуса со звучащим переводом — и именно это, а не недостаток инженерных решений, ограничивает сегодня переход к непрерывному распознаванию жестового языка (CSLR) и переводу жестового языка (SLT). На архитектурном

уровне сопоставление эпох (НММ — трансформеры — большие языковые модели) показывает прогрессирующее усиление языкового модуля, что для РЖЯ имеет особое значение, поскольку gloss-разметка существует только для изолированных знаков.

Из этого синтеза вырастают итоговые положения обзора. Фонология пяти параметров жеста, предложенная У. Стоуки, оказывается лингвистически осмысленным признаковым пространством, прямо отображаемым на выходы современных whole-body-детекторов позы — и именно это соответствие делает RTMPose Whole-body не «инженерным предпочтением», а архитектурным выбором, поддержанным теорией. Российский ресурс Logos эмпирически продемонстрировал способность функционировать как универсальный жестовый энкодер с переносом на американский и турецкий жестовые языки, что переворачивает традиционную картину «РЖЯ как заимствующий язык» и делает Logos обоснованной точкой опоры зрительного компонента. Дефицит непрерывного корпуса частично, но не полностью, может быть компенсирован синтетическими данными — при условии многоканальности и просодической осмысленности генерации, отсутствие которых превращает синтетический материал в обучающий шум. Семейство языковых моделей GigaChat удовлетворяет лингвистически осмысленным критериям выбора декодера для системы РЖЯ→русский: специализация на русском языке в обучающих данных, длинный контекст, сопоставимый с минутными жестовыми дискурсами, и двойной режим развертывания — открытые веса для локального дообучения и доступ к старшим версиям через программный интерфейс.

Сводя эти положения в архитектурное предложение, целевая система перевода РЖЯ→русский может быть описана как пайплайн RTMPose Whole-body → Logos-предобученный зрительный энкодер → токенизатор с векторным квантованием по образцу архитектуры SignLLM → декодер из семейства GigaChat, с многоканальной синтетической аугментацией на этапе предобучения энкодера. Каждый компонент в этой схеме обоснован опубликованной литературой и направлен на решение задач, обусловленных лингвистическими требованиями к системе.

Из проведенной систематизации следует и повестка ближайших исследований. Приоритетной задачей российских исследователей распознавания жестовых языков остается создание непрерывного параллельного корпуса РЖЯ↔русский уровня DGS-Korpus или How2Sign. Без него тонкая настройка языкового декодера в предлагаемой архитектуре невозможна на нужном масштабе данных. За этой задачей следуют 4D-разметка для обучения многоканальных генеративных моделей, региональная репрезентативность датасетов, просодическая авторазметка через специализированные детекторы позы человека и грамматическая разметка как вспомогательная задача распознавания.

Вклад данного исследования видится в систематизации области автоматической обработки РЖЯ с лингвистической точки зрения, в эксплицитном

выделении корпусных лакун как структурного барьера, а также в разработке теоретически обоснованной компонентной архитектуры целевой системы, в которой лингвистические свойства РЖЯ и инженерные компоненты сопряжены не риторически, а методически.

Вклад авторов

Плавун И.Н. — формулирование идеи, гипотезы и цели исследования; разработка методов и процедур, сбор данных; написание первоначального текста статьи.

Эбзеева Ю.Н. — редактирование, доработка текста; теоретический анализ данных.

Список литературы / References

- Буркова С.И., Киммельман В.И. (отв. ред.). Введение в лингвистику жестовых языков. Русский жестовый язык. Новосибирск: Изд-во НГТУ, 2019. [Burkova, Svetlana I. & Vadim I. Kimmelman (eds.). 2019. *Vvedenie v lingvistiku zhestovykh yazykov. Russkii zhestovyi yazyk: uchebnik* (Introduction to the linguistics of sign languages. Russian Sign Language). Novosibirsk: Izd-vo NGTU. (In Russ.)].
- Кагиров И.А., Рюмин Д.А., Аксенов А.А., Карпов А.А. Мультимедийная база данных жестов русского жестового языка в трехмерном формате // *Вопросы языкознания*. 2020. № 1. С. 104–123. <https://doi.org/10.31857/S0373658X0008302-1> [Kagirov, Ildar A., Dmitry A. Ryumin, Alexander A. Axyonov & Alexey A. Karpov. 2020. A multimedia database of Russian Sign Language items in 3D. *Voprosy Jazykoznanija* (1). 104–123. (In Russ.)].
- Карпов А.А. Компьютерный анализ и синтез русского жестового языка // *Вопросы языкознания*. 2011. № 6. С. 41–53. <http://vja.ruslang.ru/ru/archive/2011-6/41-53> [Karpov, Alexey A. 2011. Computer analysis and synthesis of Russian Sign Language. *Voprosy Jazykoznanija* (6). 41–53. (In Russ.)].
- Петрова Е., Быкова У., Кванчиани К. Распознавание и перевод жестовых языков: обзор подходов, 2024. Корпоративный блог SberDevices на платформе Habr, 13 февраля 2024 г. <https://habr.com/ru/companies/sberdevices/articles/792660/> (дата обращения: 30 апреля 2026 г.). [Petrova, Elizaveta, Ulyana Bykova & Karina Kvanchiani. 2024. *Sign language recognition and translation: A review of approaches*. SberDevices corporate blog at Habr, 13 February 2024. (In Russ.)].
- Прозорова Е.В. 2007. Российский жестовый язык как предмет лингвистического исследования // *Вопросы языкознания*. 2007. № 1. С. 44–61. <http://vja.ruslang.ru/ru/archive/2007-1/44-61> [Prozorova, Elena V. 2007. Russian Sign Language as a subject of linguistic study. *Voprosy Jazykoznanija* (1). 44–61. (In Russ.)].
- Рюмин Д.А., Кагиров И.А., Аксенов А.А., Карпов А.А. Аналитический обзор моделей и методов автоматического распознавания жестов и жестовых языков // *Информационно-управляющие системы*. 2021. № 6. С. 10–20. <https://doi.org/10.31799/1684-8853-2021-6-10-20> [Ryumin, Dmitry A., Ildar A. Kagirov, Alexander A. Axyonov & Alexey A. Karpov. 2021. An analytical review of models and methods for automatic recognition of gestures and sign languages. *Informatsionno-upravlyayushchie sistemy* (6). 10–20. (In Russ.)].
- Baltatzis, Vasileios, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng & Stefanos Zafeiriou. 2024. Neural Sign Actors: A diffusion model for 3D sign language production from text. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 1985–1995. Piscataway, NJ: IEEE.

- Bauer, Anastasia & Maria Kyuseva. 2022. New insights into mouthings: Evidence from a corpus-based study of Russian Sign Language. *Frontiers in Psychology* 12. 779–958. <https://doi.org/10.3389/fpsyg.2021.779958>
- Burkova, Svetlana I. (project leader). 2012–2015. *Russkii zhestovyi yazyk: korpus / Russian Sign Language Corpus*. Novosibirsk: Novosibirsk State Technical University. <http://rsl.nstu.ru/> (accessed 30 April 2026).
- Boutet, Dominique, Aliyah Morgenstern & Alan Cienki. 2016. Grammatical aspect and gesture in French: A kinesiological approach. *Russian Journal of Linguistics* 20 (3). 132–151.
- Burkova, Svetlana, Elena Filimonova, Vadim Kimmelman, Valeria Kopylova & Nina Semushina. 2019. Lexical expressions of time in Russian Sign Language. *Sign Language Studies* 19 (2). 175–203. <https://doi.org/10.1353/sls.2018.0031>
- Camgöz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney & Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 7784–7793. Piscataway, NJ: IEEE. <https://doi.org/10.1109/CVPR.2018.00812>
- Camgöz, Necati Cihan, Oscar Koller, Simon Hadfield & Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, 10023–10033. Piscataway, NJ: IEEE. <https://doi.org/10.1109/CVPR42600.2020.01004>
- Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei & Yaser Sheikh. 2021. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (1). 172–186. <https://doi.org/10.1109/TPAMI.2019.2929257>
- Cienki, Alan. 2024. Self-focused versus dialogic features of gesturing during simultaneous interpreting. *Russian Journal of Linguistics* 28 (2). 227–242. <https://doi.org/10.22363/2687-0088-34572>
- Duarte, Amanda, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres & Xavier Giró-i-Nieto. 2021. How2Sign: A large-scale multimodal dataset for continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, 2735–2744. Piscataway, NJ: IEEE. <https://doi.org/10.1109/CVPR46437.2021.00276>
- Fang, Sen, Lei Wang, Ce Zheng, Yapeng Tian & Chen Chen. 2024. *SignLLM: Sign Language Production Large Language Models*. arXiv:2405.10718. <https://arxiv.org/abs/2405.10718>
- Forster, Jens, Christoph Schmidt, Oscar Koller, Martin Bellgardt & Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 1911–1916. Paris: ELRA.
- GigaChat team. 2025. GigaChat Family: Efficient Russian language modeling through Mixture of Experts architecture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg, PA: ACL. (Also arXiv:2506.09440. <https://arxiv.org/abs/2506.09440>)
- Gong, Jia, Lin Geng Foo, Yixuan He, Hossein Rahmani & Jun Liu. 2024. LLMs are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 18362–18372. Piscataway, NJ: IEEE. <https://doi.org/10.1109/CVPR52733.2024.01738>
- Imashev, Alfarabi, Medet Mukushev, Vadim Kimmelman & Anara Sandygulova. 2020. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The K-RSL. In Raquel Fernández & Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*,

- 631–640. Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.conll-1.51>
- Irishanova, Olga K., Alan Cienki, Maria V. Tomskaya & Alexandra I. Nikolayeva. 2023. Silent, but salient: Gestures in simultaneous interpreting. Research result. *Theoretical and Applied Linguistics* 9 (1). 99–114. <https://doi.org/10.18413/2313-8912-2023-9-1-0-7>
- Jiang, Tao, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li & Kai Chen. 2023. *RTMPose: Real-time multi-person pose estimation based on MMPose*. arXiv:2303.07399. <https://arxiv.org/abs/2303.07399>
- Kagirov, Ildar, Denis Ivanko, Dmitry Ryumin, Alexander Axyonov & Alexey Karpov. 2020. TheRuSLan: Database of Russian Sign Language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 6079–6085. Paris: ELRA.
- Kapitanov, Alexander, Karina Kvanchiani, Alexander Nagaev & Elizaveta Petrova. 2023. Slovo: Russian Sign Language dataset. In Henrik I. Christensen, Peter Corke, Renaud Detry, Jean-Baptiste Weibel & Markus Vincze (eds.), *Computer Vision Systems. ICVS 2023* (Lecture Notes in Computer Science 14253), 63–73. Cham: Springer. https://doi.org/10.1007/978-3-031-44137-0_6
- Kimmelman, Vadim. 2012. Word order in Russian Sign Language. *Sign Language Studies* 12 (3). 414–445. <https://doi.org/10.1353/sls.2012.0001>
- Kimmelman, Vadim. 2019. *Information Structure in Sign Languages: Evidence from Russian Sign Language and Sign Language of the Netherlands* (Sign Languages and Deaf Communities 10). Berlin & Boston: De Gruyter Mouton. <https://doi.org/10.1515/9781501510045>
- Kimmelman, Vadim, Anna Komarova, Lyudmila Luchkova, Valeria Vinogradova & Oksana Alekseeva. 2022. Exploring networks of lexical variation in Russian Sign Language. *Frontiers in Psychology* 12. Article 740734. <https://doi.org/10.3389/fpsyg.2021.740734>
- Kok, Kasper, Kirsten Bergmann, Alan Cienki & Stefan Kopp. 2015. Mapping out the multifunctionality of speakers' gestures. *Gesture* 15. 37–59. <https://doi.org/10.1075/gest.15.1.02kok>
- Koller, Oscar, Jens Forster & Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141. 108–125. <https://doi.org/10.1016/j.cviu.2015.09.013>
- Konrad, Reiner, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn & Marc Schulder. 2020. *MEINE DGS — annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS — annotated. Public corpus of German Sign Language [Dataset]*. Universität Hamburg. <https://doi.org/10.25592/dgs.corpus-3.0>
- Kuznetsova, Anna & Vadim Kimmelman. 2024. *Testing MediaPipe Holistic for linguistic analysis of nonmanual markers in sign languages*. arXiv:2403.10367. <https://arxiv.org/abs/2403.10367>
- Kvanchiani, Karina, Petr Surovtsev, Alexander Nagaev, Elizaveta Petrova & Alexander Kapitanov. 2024. *Bukva: Russian Sign Language alphabet*. arXiv:2410.08675. <https://arxiv.org/abs/2410.08675>
- Leonteva, Anna V., Alan Cienki & Olga V. Agafonova. 2023. Metaphoric gestures in simultaneous interpreting. *Russian Journal of Linguistics* 27 (4). 820–842. <https://doi.org/10.22363/2687-0088-36189>
- Liddell, Scott K. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.

- Núñez-Marcos, Adrián, Olatz Perez-de-Viñaspre & Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications* 213 (B). Article 118993. <https://doi.org/10.1016/j.eswa.2022.118993>
- Ovodov, Ilya, Petr Surovtsev, Karina Kvanchiani, Alexander Kapitanov & Alexander Nagaev. 2025. Logos as a well-tempered pre-train for sign language recognition. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose & Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 24340–24353. Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1238>
- Rastgoo, Razieh, Kourosh Kiani & Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications* 164. Article 113794. <https://doi.org/10.1016/j.eswa.2020.113794>
- Ryumin, Dmitry, Denis Ivanko & Elena Ryumina. 2023. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors* 23 (4). Article 2284. <https://doi.org/10.3390/s23042284>
- Sandler, Wendy & Diane Lillo-Martin. 2006. *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.
- Saunders, Ben, Necati Cihan Camgöz & Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In Andrea Vedaldi, Horst Bischof, Thomas Brox & Jan-Michael Frahm (eds.), *Computer Vision — ECCV 2020 (Lecture Notes in Computer Science 12356)*, 687–705. Cham: Springer. https://doi.org/10.1007/978-3-030-58621-8_40
- Saunders, Ben, Necati Cihan Camgöz & Richard Bowden. 2021. Continuous 3D multi-channel sign language production via progressive transformers and mixture density networks. *International Journal of Computer Vision* 129 (7). 2113–2135. <https://doi.org/10.1007/s11263-021-01457-9>
- Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Sally Reynolds & Kearsy Cormier. 2013. Building the British Sign Language Corpus. *Language Documentation & Conservation* 7. 136–154.
- Stokoe, William C. 2005. Sign language structure: An outline of the visual communication systems of the American Deaf [Reprint of Stokoe 1960]. *Journal of Deaf Studies and Deaf Education* 10 (1). 3–37. <https://doi.org/10.1093/deafed/eni001>
- Tan, Sihan, Nabeela Khan, Zhaoyi An, Yoshitaka Ando, Rei Kawakami & Kazuhiro Nakadai. 2024. A review of deep learning-based approaches to sign language processing. *Advanced Robotics* 38 (23). 1649–1667. <https://doi.org/10.1080/01691864.2024.2442721>
- Yan, Sijie, Yuanjun Xiong & Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 7444–7452. Palo Alto, CA: AAAI Press. <https://doi.org/10.1609/aaai.v32i1.12328>
- Zhang, Fan, Valentin Bazarevsky, Andrei Vakunov, Andrey Tkachenka, George Sung, Chuo-Ling Chang & Matthias Grundmann. 2020. *MediaPipe Hands: On-device real-time hand tracking*. arXiv:2006.10214. <https://arxiv.org/abs/2006.10214>

Article history:

Received: 15 May 2026

Accepted: 15 July 2026

Bionotes:

Ilya N. PLAVUNOV is Director of GenAI Transformation and Technological Development at Sberbank (Moscow Bank), advisor to government institutions, expert in AI and digital transformation with over 20 years of experience in managing high-tech projects: from founding IT companies to leading large-scale digital initiatives at Sberbank and Rosneft. Research interests: multi-agent AI systems, innovative digital products, and data-driven management.

Yulia N. EBZEEVA is a Doctor of Sociology, First Vice-Rector (Vice-Rector for Academic Affairs), Head of the Department of Foreign Languages at the Faculty of Philology, RUDN University. She is a member of the QS International Advisory Committee. She has presented at the Council of Europe, and has repeatedly served as an expert on linguistic and migration issues. Her research interests include French lexicology and stylistics, translation theory, intercultural communication, sociolinguistics, migration studies and educational policy.

e-mail: ebzeeva-jn@rudn.ru

<https://orcid.org/0000-0002-0043-7590>

Сведения об авторах:

Илья Николаевич ПЛАВУНОВ — директор по ГенИИ-трансформации и технологическому развитию ПАО «Сбербанк» (Московский банк), консультант государственных учреждений, эксперт в области ИИ и цифровой трансформации. Имеет более 20 лет опыта в управлении высокотехнологичными проектами: от основания IT-компаний до руководства крупными цифровыми инициативами в Сбербанке и Роснефти. Сфера научных интересов: мультиагентные ИИ-системы, инновационные цифровые продукты и управление на основе данных.

Юлия Николаевна ЭБЗЕЕВА — доктор социологических наук, Первый проректор (проректор по образовательной деятельности), заведующая кафедрой иностранных языков филологического факультета РУДН. Является членом международного научного комитета QS, выступала в Совете Европы, многократно являлась экспертом по лингвистическим и миграционным проблемам. Сферу ее научных интересов составляют лексикология и стилистика французского языка, теория перевода, межкультурная коммуникация, социолингвистика, миграциология и образовательная политика.

e-mail: ebzeeva-jn@rudn.ru

<https://orcid.org/0000-0002-0043-7590>