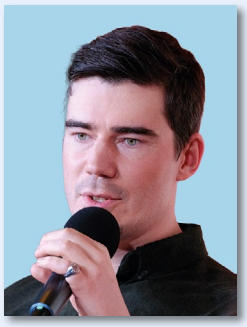
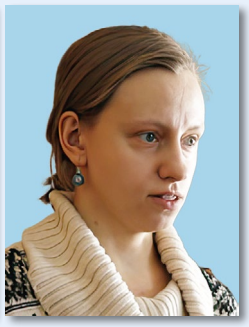
Индекс УДК 81-139
Код ГРНТИ 16.31.21
DOI: 10.22204/2587-8956-2025-123-04-46-55

Д.И. КОЛОМАЦКИЙ,
Е.В. КОРОВИНА*

Этический аспект использования технологий искусственного интеллекта в области сохранения языков коренных народов

В работе рассматриваются этические аспекты использования технологий искусственного интеллекта для языков коренных народов. Основное внимание уделяется анализу ряда заметных проектов. Часть из них не ставит своей непосредственной задачей ревитализацию языка или сохранение культурного наследия и потому может вызывать критику со стороны языковых сообществ. Анализируются и примеры успешных проектов, созданных и/или активно поддерживаемых представителями коренных народов. Предлагаемые к рассмотрению инициативы охватывают самые разные регионы мира: Африку, Северную и Южную Америку, Юго-Восточную Азию и Океанию. В статье также представлены рекомендации по дальнейшему развитию и поддержке таких инициатив, а также подчёркивается необходимость соблюдения этических принципов и особенно прав и интересов коренных народов. Настоящая работа призвана привлечь внимание к важности сохранения языков коренных народов с помощью современных технологий, что особенно актуально для Российской Федерации с её уникальным языковым разнообразием.

Ключевые слова: большие языковые модели, языки коренных народов, документирование языков, ревитализация языков, машинный перевод

Работа с так называемыми низкоресурсными языками (low resource languages) является одной из актуальных тем в области применения технологий искусственного интеллекта в лингвистике. Объём данных для обучения больших языковых моделей (Large Language Models, LLM, БЯМ) для таких языков, как английский, китайский, испанский, русский, достаточно велик, чтобы модели могли генерировать тексты очень высокого качества. Миноритарные же языки оказываются недостаточно охвачены такими технологиями, что лишь усиливает доминирование мажоритарных языков в мире. В России говорят более

* **Коломацкий Дмитрий Игоревич** — кандидат филологических наук, научный сотрудник отдела прикладной лингвистики Института языкознания РАН (ИЯз РАН).
E-mail: dk@iling-ran.ru
Коровина Евгения Владимировна — младший научный сотрудник отдела урало-алтайских языков ИЯз РАН.
E-mail: evkorovina@iling-ran.ru

чем на 150 языках (по данным ИЯз РАН), поэтому для нашей страны эта проблема также релевантна.

В последние годы компьютерные лингвисты обращают всё больше внимания на миноритарные языки. Одно из исследований такого рода представлено и в настоящем журнале (автор Т.О. Шаврина).

В связи с этим уместно вспомнить, что труд специалистов по ИИ и компьютерных лингвистов, занимающихся миноритарными языками, далеко не всегда направлен именно на их поддержку. Этический аспект работы с миноритарными языками вообще (включая просто сбор данных для словарей) завладел умами лингвистов не так давно, но к настоящему моменту уже появилось множество общих трудов на эту тему [1–4], а также case studies по конкретным примерам лингвистической работы в том или ином регионе [5].

В настоящей работе мы сосредоточимся на более узкой теме — этической стороне использования технологий ИИ применительно к миноритарным языкам. Если Т.О. Шаврина фокусируется на технологическом аспекте работы с низкоресурсными языками, мы обратим внимание на гуманитарный аспект на примерах конкретных инициатив.

Существующих проектов слишком много, чтобы дать их полный обзор в небольшой статье, поэтому представляется разумным предложить к рассмотрению несколько примечательных случаев, охватив этими примерами самые разные регионы мира.

В первую очередь мы приведём пример отношения исследователей к языку коренного народа как к средству, а не цели, а также обозначим сложность проблемы и риски, связанные с вовлечением крупных корпораций. Напротив, дальнейшие сюжеты показывают, как могут работать проекты в области ИИ, если они инициированы или хотя бы активно поддержаны языковыми сообществами.

Язык как средство: проекты, не инициированные изнутри коренных сообществ

Язык каламанг; «Machine Translation from One Book»

На языке каламанг говорят менее 200 человек, живущих на островах Карас у берегов полуострова Бомберай на западе острова Новая Гвинея (Индонезия). Его полное грамматическое описание было подготовлено в 2020 г., а в 2022 г. вышло в виде монографии [6]. Автор монографии стала соавтором работы о применении этого грамматического описания для порождения текстов на каламанг с помощью ИИ [7]. Эта работа получила широкий резонанс и стала как отправной точкой для других исследований [8], так и объектом критики ровно за то, что роль грамматического описания, возможно, переоценена [9].

Для настоящей работы интереснее тот факт, что сам язык в силу того, что он относится к «экстремально низкоресурсным языкам» (Extremely low-resource languages, XLR), хорошо подходит для эксперимента с БЯМ. В каком-то смысле такой статус языка стал для авторов достоинством. Основной целью исследования был «чистый» эксперимент, а не снабжение языкового сообщества новыми технологиями[1]. Результат эксперимента мог бы в лучшем случае стать позитивным побочным про-

[1] “Do LLMs really learn to perform new tasks by adaptation, or does adaptation simply draw out capabilities that the model had already learned? And does scaling pretraining data just improve performance because it implicitly scales up in-domain training data for every task? The best way to answer these questions is to evaluate on tasks that were unseen in the training data, but with these models being trained on increasingly opaque web-scale datasets, this can seem impossible <...>. **To address this challenge**, we turn to a field that is explicitly motivated and bottlenecked by a scarcity of web data: low-resource languages. <...> While translation tasks in general are well-represented in LLM training data, the Kalamang language in particular **has been held out from the web for sociohistorical reasons**, with the exception of the documentation in Visser (2022). **This means that Kalamang tasks are unseen to LLMs** but still feature the complexities that come with substantial real-world tasks.” [7] (полужирный шрифт мой. — *Д.К.*)

дуктом. В то же время стоит подчеркнуть, что носители языка были проинформированы о проводимом исследовании и согласились с использованием данных своего языка в качестве материала для БЯМ. При этом каких-либо данных об использовании разработанной модели среди сообщества говорящих на языке каламанг найти не удаётся.

Уастекский науатль; переводчик Google как благо или вред

Как мы уже могли видеть, при разработке больших языковых моделей языковые сообщества нередко воспринимаются сугубо как источник данных для обучения, что исключает их активное участие в процессе. Такое положение дел может вызывать неприятие со стороны коренных народов, а потому провоцировать конфликты между создателями моделей и носителями языков. Ярким примером подобной динамики стала разработка машинного переводчика уастекского варианта языка науатль (Мексика, около 1 млн носителей) для сервиса Google Translate.

Основным разработчиком проекта выступила Габриэла Салас Кабрера[1] — молодой специалист в области IT и социальных проектов родом из небольшого города в мексиканском штате Идальго, который входит в ареал распространения этого языка[2]. Для сбора языкового материала она обратилась к науа-язычным сообществам в соцсетях и мессенджерах, предложив сотрудничество на волонтёрской основе, что было обусловлено отсутствием финансирования (Салас Кабрера проходила в этот период неоплачиваемую стажировку в Google). Участники подписывали информированное соглашение, где оговаривались эти условия. Благодаря собранным данным в июле 2024 г. сервис стал доступен на платформе Google. Салас Кабрера получила признание: в 2024 г. британская телевещательная корпорация включила её в список «100 женщин, изменивших мир»[3], а мексиканские СМИ активно освещали её работу[4].

Однако 25 марта 2025 г. в соцсетях развернулась дискуссия[5] о возможной эксплуатации языкового сообщества для извлечения прибыли без адекватной компенсации. Критики указывали на неэтичность использования бесплатного труда носителей языка, бесполезность автоматического переводчика для сообщества и отсутствие консенсуса относительно его ценности. Часть пользователей отказалась от участия именно по этим причинам. В то же время другие участники дискуссии подчёркивали право каждого самостоятельно решать вопрос о сотрудничестве, а также отмечали, что резкость критики может быть связана с происхождением Салас Кабрера (молодая женщина из бедной семьи индейского происхождения). Даже сторонники проекта признавали его организационные недостатки: так, в рабочую группу добавляли всех желающих без предварительного отбора по мотивации. Также отмечалось и то, что первоначальное описание проекта не содержало чётких указаний о волонтёрском характере работы и отсутствии финансирования, что было исправлено лишь позднее.

Описанная ситуация демонстрирует, к чему могут привести наличие интереса крупной корпорации при отсутствии

[1] https://stemwomen.global/profile/gabriela-salas-cabrera (дата обращения: 14.05.2025).

[2] Нужно отметить, что лингвистическое образование у Салас Кабреры отсутствует.

[3] Quién es Gabriela Salas Cabrera, mexicana que está dentro de las 100 mujeres más influyentes… Telediario, 04.12.2024 https://www.telediario.mx/comunidad/gabriela-salas-cabrera-las-100-mujeres-mas-influyentes-de-la-bbc (дата обращения: 14.05.2025).

[4] Conoce a Gabriela Salas, la mexicana que logró incluir el Náhuatl al traductor de Google. Informador, 18.07.2024 https://www.informador.mx/mexico/Conoce-a-Gabriela-Salas-la-mexicana-que-logro-incluir-el-Nahuatl-al-traductor-de-Google-20240718-0105.html (дата обращения: 14.05.2025).

[5] Исходный пост позднее был удалён, но в распоряжении авторов есть ссылки на последующую дискуссию.

Ил. 1. Дж. Брикси.
Фото Aaron Balana

чёткой и ясной коммуникации с носителями языка.

Проекты, инициируемые или активно поддерживаемые представителями коренных народов

Чоктавский язык; мультимодальный корпус и чат-бот

Чоктавский язык — один из аборигенных языков Северной Америки, относится к группе мускогских языков. По данным Endangered Languages Project, на нём говорят от 9 до 11 тыс. человек, ему приписан статус языка, находящегося в уязвимом положении[1]; по данным исследовательницы языка Джэклин Брикси (ил. 1), число носителей не превышает 7 тыс., при этом сообщество Choctaw Nation of Oklahoma[2] насчитывает 220 тыс. «граждан» (citizens) [10]. Дж. Брикси создала мультимодальный корпус чоктавского языка ChoCo [11,12], чтобы затем на его основе разработать чат-бот под названием Masheli [13].

При разработке чат-бота изучались два варианта его функционирования: двуязычный (пользователь сам переключает язык) или с переключением кодов (что создаёт более полную имитацию живого разговора, так как реальные носители постоянно переключаются с английского на чоктавский и обратно) [14]. В оценке работы чат-бота обязательно принимали участие представители коренного народа, и протокол взаимодействия с ними подробно прописан в [14]. Вариант бота с переключением кода оказался в среднем предпочтительнее для участников эксперимента, которые называли бот «более дружелюбным».

Примечательно, что при создании Masheli было решено отказаться от использования БЯМ в части генерации ответа: БЯМ не обеспечили бы максимально корректного переключения кодов и не гарантировали бы точность вывода на чоктавском языке. В качестве технической основы был выбран NPCEditor, специально предназначенный для создания диалоговых систем в узких предметных областях [15]. Ответ не генерируется «на лету», а выбирается из коллекции, хотя запрос пользователя

[1] https://endangeredlanguages.com/lang/1692 (дата обращения: 10.05.2025).

[2] https://www.choctawnation.com/ (дата обращения: 10.05.2025).

Ил. 2. Коллектив Lelapa AI в 2025 г.

Фото с сайта agenticppa.com

анализируется статистическими методами (задача классификации текста). В этом смысле Masheli оказывается промежуточным продуктом между диалоговой системой на основе БЯМ и экспертной системой.

Усилия Дж. Брикси находят признание среди коренного населения штата Оклахома. Так, ей посвятила заметку издаваемая сообществом Choctaw Nation газета Biskinik[1]. В 2025 г. планируется выпустить словарь варианта чоктавского языка, на котором говорят в штате Миссисипи. Словарь будет сопровождаться разработанным Дж. Брикси приложением для распознавания речи, которое поможет в пользовании словарём тем носителям языка, которые не владеют им в письменной форме [10].

На наш взгляд, деятельность этой исследовательницы можно считать прекрасным примером применения технологий ИИ на благо коренных народов.

Пять языков ЮАР, далее вся Африка; Masakhane и Lelapa AI

Masakhane[2] — африканская инициатива по снабжению языков Африки средствами машинного перевода [16]. Проблемы и первые решения для пяти языков ЮАР описаны в [17]. Уже год спустя, в 2020 г., число поддерживаемых языков значительно выросло [18]. К сожалению, официальный репозиторий инициативы не обновляется и не принимает изменения в исходный код уже около трёх лет[3]. Однако двое из сооснователей Masakhane вошли в число создателей лаборатории Lelapa AI[4], которая продолжает разработки на тех же этических принципах (ил. 2).

Флагманским продуктом Lelapa AI стала коммерческая платформа Vulavula[5], сочетающая в себе возможности транскрипции, распознавания сущностей, диалога с пользователем и т.д. Машинный перевод и синтез речи по состоянию на май 2025 г.

[1] https://biskinik.com/wp-content/uploads/2025/02/mar2022-biskinik.pdf#page=12 (дата обращения: 10.05.2025).

[2] https://www.masakhane.io/ (дата обращения: 10.05.2025).

[3] https://github.com/masakhane-io/masakhane-mt (дата обращения: 10.05.2025).

[4] https://lelapa.ai (дата обращения: 10.05.2025).

[5] https://lelapa.ai/products/vulavula/ (дата обращения: 10.05.2025).

Ил. 3. Представители Te Hiku Media и групп маори.

Фото Te Hiku Media

находятся в статусе «coming soon». Для защиты прав коренных народов на их лингвистические ресурсы был создан фреймворк Esethu и составлена особая лицензия Esethu для распространения датасетов [19].

Пример Lelapa AI даёт основания полагать, что даже коммерческое использование языковых моделей может сочетаться с позитивным отношением к проекту в коренных языковых сообществах.

Язык маори; машинный перевод и обучающая платформа

Язык маори — полинезийский язык, относимый к австронезийской языковой семье, один из государственных языков Новой Зеландии. Endangered Language Project относит язык к категории угрожаемых.

Новозеландская медиакомпания Te Hiku Media, управляемая представителями различных групп (*iwi*) народа маори, уделяет особое внимание разработке технологических решений для документации и ревитализации языка. С появлением БЯМ компания обратилась к использованию этих средств, в частности к разработкам решений для автоматического распознавания речи, позволяющим транскрибировать речь маори с точностью 92% и речь с переключением кодов с точностью 82%[1]. Разработкам Te Hiku Media (ил. 3) посвящён ряд научных работ (например, [20, 21]). Авторы [20] отмечают, что разработанные для маори технологии планируется в будущем использовать для самоанского и гавайского языков. Описываемый в работах проект Papa Reo официально запущен и предлагает разработчикам целый пакет инструментов для распознавания и синтеза речи, а также отслеживания соблюдения этических принципов работы с данными («kaitiakitanga», по-английски «guardianship, trusteeship») на языке маори[2]. Te Hiku Media подчёркивает, что считает себя именно «хранителем, попечителем» (kaitiaki), а не владель-

[1] Māori Speech AI Model Helps Preserve and Promote New Zealand Indigenous Language. Nvidia blog. https://blogs.nvidia.com/blog/te-hiku-media-maori-speech-ai/ (дата обращения: 11.05.2025).

[2] https://papareo.io/ (дата обращения: 11.05.2025).

цем данных, получаемых от носителей языка[1].

Опыт Te Hiku Media иллюстрирует возможности, которыми обладают и которые могут реализовывать местные средства массовой информации, заинтересованные в сохранении языка коренного народа.

Рекомендации

Главный вывод из изученных нами примеров разработки ИИ-решений для языков коренных народов (как описанных в статье, так и оставшихся за её рамками) состоит в том, что в подобных проектах совершенно необходимо активное участие представителей коренных народов[2]. Если специалисты по ИИ активно контактируют с соответствующими языковыми сообществами и понимают их потребности, они верно сформируют и набор технических решений. Файн-тюнинг БЯМ (дообучение предобученной БЯМ для решения специфических задач) повышает точность работы модели по сравнению с обычной предобученной БЯМ. Если же ставится цель сделать модели доступными для непосредственного использования представителями сообществ, она может диктовать выбор решений, предъявляющих низкие требования к оборудованию. Например, может быть применена так называемая дистилляция знаний [22] или выбрана более энергоэффективная модель — скажем, из семейства Mistral [23][3]. В работе [24] даже вводится термин «Indigenous Language Models (ILMs)», обозначающий модели, оптимизированные для языков коренных народов.

В юридических документах, определяющих работу инициатив в области ИИ применительно к малочисленным языкам, должны быть прописаны права представителей языковых сообществ на предоставляемый ими материал, особенно если речь идёт о культурном и/или религиозном наследии коренного народа.

Заключение

В настоящей работе на нескольких ярких примерах предпринята попытка показать, что технологии ИИ могут служить сохранению языкового разнообразия, документации языкового наследия и даже ревитализации малых языков, если они применяются при активном участии представителей коренных народов и со стремлением принести пользу именно им. Мы надеемся, что эти соображения позитивно повлияют на развитие аналогичных проектов, создаваемых для малочисленных языков Российской Федерации.

ЛИТЕРАТУРА

1. Good J. Ethics in Language Documentation and Revitalization // The Oxford Handbook of Endangered Languages / Ed. K.L. Rehg, L. Campbell. Oxford University Press, 2018. Pp. 418–440. DOI: 10.1093/oxfordhb/9780190610029.013.21.
2. Holton G., Leonard W.Y., Pulsifer P.L. Indigenous Peoples, Ethics, and Linguistic Data // The Open Handbook of Linguistic Data Management / Ed. A.L. Berez-Kroeker et al. The MIT Press, 2022. Pp. 49–60. DOI: 10.7551/mitpress/12200.003.0008.
3. Marley T.L. Indigenous Data Sovereignty and the role of universities // Indigenous Data Sovereignty and Policy. 1st ed. London: Routledge, 2020. Pp. 157–168. DOI: 10.4324/9780429273957-11.

[1] A new vision of artificial intelligence for the people. MIT Technology Review https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/ (дата обращения: 11.05.2025).

[2] При этом, как говорилось в обсуждении проблемы языка науатль в соцсетях, не нужно воспринимать коренные народы как настолько уязвимые и нуждающиеся в помощи извне, что они не способны самостоятельно принимать решения о том, в каких проектах участвовать.

[3] Обсуждение использования Mistral для работы с языками аборигенов Австралии найдено на специализированных форумах, но академических статей на эту тему авторы пока не обнаружили.

4. Ruckstuhl K. Trust in Scholarly Communications and Infrastructure: Indigenous Data Sovereignty // Front. Res. Metr. Anal. 2022. Vol. 6. DOI: 10.3389/frma.2021.752336.
5. Ortenzi K.M. et al. Good data relations key to Indigenous research sovereignty: A case study from Nunatsiavut // Ambio. 2025. Vol. 54, № 2. Pp. 256–269. DOI: 10.1007/s13280-024-02077-6.
6. Visser E.A grammar of Kalamang. Berlin: Language Science Press, 2022. https://zenodo.org/record/6499927 (access date: 10.05.2025).
7. Tanzer G. et al. A Benchmark for Learning to Translate a New Language from One Grammar Book: arXiv:2309.16575. arXiv, 2024. DOI: 10.48550/arXiv.2309.16575.
8. Kornilov A., Shavrina T. From MTEB to MTOB: Retrieval-Augmented Classification for Descriptive Grammars: arXiv:2411.15577. arXiv, 2024. DOI: 10.48550/arXiv.2411.15577.
9. Aycock S. et al. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book? arXiv:2409.19151. arXiv, 2025. DOI: 10.48550/arXiv.2409.19151.
10. Brixey J. Using Artificial Intelligence to Preserve Indigenous Languages. USC Institute for Creative Technologies, 2025. https://ict.usc.edu/news/essays/using-artificial-intelligence-to-preserve-indigenous-languages/ (access date: 10.05.2025).
11. Brixey J., Pincus E., Artstein R. Chahta Anumpa: A Multimodal Corpus of the Choctaw Language // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018. Pp. 3371–3376. https://aclanthology.org/L18-1532.pdf. (access date: 10.05.2025).
12. Brixey J., Artstein R. ChoCo: a multimodal corpus of the Choctaw language // Lang Resources & Evaluation. 2021. Vol. 55, № 1. Pp. 241–257. DOI: 10.1007/s10579-020-09494-5.
13. Brixey J., Traum D. Masheli: A Choctaw-English Bilingual Chatbot // Conversational Dialogue Systems for the Next Decade / Ed. L.F. D'Haro, Z. Callejas, S. Nakamura Singapore: Springer Singapore, 2021. Vol. 704. Pp. 41–50. DOI: 10.1007/978-981-15-8395-7_4.
14. Brixey J., Traum D. Does a code-switching dialogue system help users learn conversational fluency in Choctaw? // Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). Albuquerque, New Mexico: Association for Computational Linguistics, 2025. Pp. 8–17. https://aclanthology.org/2025.americasnlp-1.2/ (access date: 10.05.2025).
15. Leuski A., Traum D. NPCEditor: Creating Virtual Human Dialogue Using Information Retrieval Techniques // AI Magazine. 2011. Vol. 32, № 2. Pp. 42–56. DOI: 10.1609/aimag.v32i2.2347.
16. Orife I. et al. Masakhane – Machine Translation For Africa: arXiv:2003.11529. arXiv, 2020. DOI: 10.48550/arXiv.2003.11529.
17. Martinus L., Abbott J.Z. A Focus on Neural Machine Translation for African Languages: arXiv:1906.05685. arXiv, 2019. DOI: 10.48550/arXiv.1906.05685.
18. Nekoto W. et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages: arXiv:2010.02353. arXiv, 2020. DOI: 10.48550/arXiv.2010.02353.
19. Rajab J. et al. The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages: arXiv:2502.15916. arXiv, 2025. DOI: 10.48550/arXiv.2502.15916.
20. Jones P.-L. et al. Kia tangata whenua: Artificial intelligence that grows from the land and people // Ethical Space: International Journal of Communication Ethics. 2023. Vol. 2023, № 2/3. DOI: 10.21428/0af3f4c0.9092b177.
21. Leoni G. et al. Solving Failure Modes in the Creation of Trustworthy Language Technologies // Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024. Torino, Italia, 2024. https://aclanthology.org/2024.sigul-1.39/ (access date: 11.05.2025).
22. Xu X. et al. A Survey on Knowledge Distillation of Large Language Models: arXiv:2402.13116. arXiv, 2024. DOI: 10.48550/arXiv.2402.13116.
23. Jiang A.Q. et al. Mistral 7B: arXiv:2310.06825. arXiv, 2023. DOI: 10.48550/arXiv.2310.06825.
24. Pinhanez C. et al. Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences: arXiv:2407.12620. arXiv, 2024. DOI: 10.48550/arXiv.2407.12620.

The Ethical Aspect of Using Artificial Intelligence Technologies in the Field of Indigenous Languages Preservation

Dmitry Igorevich Kolomatsky — Candidate of Philological Sciences, Research Fellow at the Department of Applied Linguistics, Institute of Linguistics, Russian Academy of Sciences (IL RAS).
E-mail: dk@iling-ran.ru

Evgeniya Vladimirovna Korovina — Junior Researcher at the Department of Ural-Altaic Languages at the IL RAS.
E-mail: evkorovina@iling-ran.ru

The paper discusses the ethical aspects of applying artificial intelligence technologies to indigenous languages. The analysis focuses on several prominent projects in this field. Some initiatives are not primarily aimed at language revitalization or cultural preservation, which may attract criticism from the language communities themselves. The paper also explores examples of successful projects created and/or actively supported by indigenous individuals. The proposed initiatives cover a wide range of regions across the globe, including Africa, North and South America, Southeast Asia, and Oceania. The article also provides recommendations for further development and support of these initiatives, emphasizing the importance of ethical considerations and respecting the rights and interests of indigenous peoples. This work aims to raise awareness about the significance of preserving indigenous languages through modern technology, which is particularly relevant for the Russian Federation with its unique linguistic diversity.

Keywords: large language models, indigenous languages, language documentation, language revitalization, machine translation

REFERENCES

1. Good J. Ethics in Language Documentation and Revitalization // The Oxford Handbook of Endangered Languages / Ed. K.L. Rehg, L. Campbell. Oxford University Press, 2018. Pp. 418–440. DOI: 10.1093/oxfordhb/9780190610029.013.21.
2. Holton G., Leonard W.Y., Pulsifer P.L. Indigenous Peoples, Ethics, and Linguistic Data // The Open Handbook of Linguistic Data Management / Ed. A.L. Berez-Kroeker et al. The MIT Press, 2022. Pp. 49–60. DOI: 10.7551/mitpress/12200.003.0008.
3. Marley T.L. Indigenous Data Sovereignty and the role of universities // Indigenous Data Sovereignty and Policy. 1st ed. London: Routledge, 2020. Pp. 157–168. DOI: 10.4324/9780429273957-11.
4. Ruckstuhl K. Trust in Scholarly Communications and Infrastructure: Indigenous Data Sovereignty // Front. Res. Metr. Anal. 2022. Vol. 6. DOI: 10.3389/frma.2021.752336.
5. Ortenzi K.M. et al. Good data relations key to Indigenous research sovereignty: A case study from Nunatsiavut // Ambio. 2025. Vol. 54, № 2. Pp. 256–269. DOI: 10.1007/s13280-024-02077-6.
6. Visser E.A grammar of Kalamang. Berlin: Language Science Press, 2022. https://zenodo.org/record/6499927 (access date: 10.05.2025).
7. Tanzer G. et al. A Benchmark for Learning to Translate a New Language from One Grammar Book: arXiv:2309.16575. arXiv, 2024. DOI: 10.48550/arXiv.2309.16575.
8. Kornilov A., Shavrina T. From MTEB to MTOB: Retrieval-Augmented Classification for Descriptive Grammars: arXiv:2411.15577. arXiv, 2024. DOI: 10.48550/arXiv.2411.15577.
9. Aycock S. et al. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book? arXiv:2409.19151. arXiv, 2025. DOI: 10.48550/arXiv.2409.19151.

10. Brixey J. Using Artificial Intelligence to Preserve Indigenous Languages. USC Institute for Creative Technologies, 2025. https://ict.usc.edu/news/essays/using-artificial-intelligence-to-preserve-indigenous-languages/ (access date: 10.05.2025).
11. Brixey J., Pincus E., Artstein R. Chahta Anumpa: A Multimodal Corpus of the Choctaw Language // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018. Pp. 3371–3376. https://aclanthology.org/L18-1532.pdf. (access date: 10.05.2025).
12. Brixey J., Artstein R. ChoCo: a multimodal corpus of the Choctaw language // Lang Resources & Evaluation. 2021. Vol. 55. № 1. Pp. 241–257. DOI: 10.1007/s10579-020-09494-5.
13. Brixey J., Traum D. Masheli: A Choctaw-English Bilingual Chatbot // Conversational Dialogue Systems for the Next Decade / Ed. L.F. D'Haro, Z. Callejas, S. Nakamura Singapore: Springer Singapore, 2021. Vol. 704. Pp. 41–50. DOI: 10.1007/978-981-15-8395-7_4.
14. Brixey J., Traum D. Does a code-switching dialogue system help users learn conversational fluency in Choctaw? // Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). Albuquerque, New Mexico: Association for Computational Linguistics, 2025. Pp. 8–17. https://aclanthology.org/2025.americasnlp-1.2/ (access date: 10.05.2025).
15. Leuski A., Traum D. NPCEditor: Creating Virtual Human Dialogue Using Information Retrieval Techniques // AI Magazine. 2011. Vol. 32, № 2. Pp. 42–56. DOI: 10.1609/aimag.v32i2.2347.
16. Orife I. et al. Masakhane – Machine Translation For Africa: arXiv:2003.11529. arXiv, 2020. DOI: 10.48550/arXiv.2003.11529.
17. Martinus L., Abbott J.Z. A Focus on Neural Machine Translation for African Languages: arXiv: 1906.05685.arXiv, 2019. DOI: 10.48550/arXiv.1906.05685.
18. Nekoto W. et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages: arXiv:2010.02353. arXiv, 2020. DOI: 10.48550/arXiv.2010.02353.
19. Rajab J. et al. The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages: arXiv:2502.15916. arXiv, 2025. DOI: 10.48550/arXiv.2502.15916.
20. Jones P.-L. et al. Kia tangata whenua: Artificial intelligence that grows from the land and people // Ethical Space: International Journal of Communication Ethics. 2023. Vol. 2023, № 2/3. DOI: 10.21428/0af3f4c0.9092b177.
21. Leoni G. et al. Solving Failure Modes in the Creation of Trustworthy Language Technologies // Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024. Torino, Italia, 2024. https://aclanthology.org/2024.sigul-1.39/ (access date: 11.05.2025).
22. Xu X. et al. A Survey on Knowledge Distillation of Large Language Models: arXiv:2402.13116. arXiv, 2024. DOI: 10.48550/arXiv.2402.13116.
23. Jiang A.Q. et al. Mistral 7B: arXiv:2310.06825. arXiv, 2023. DOI: 10.48550/arXiv.2310.06825.
24. Pinhanez C. et al. Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences: arXiv:2407.12620. arXiv, 2024. DOI: 10.48550/arXiv.2407.12620.