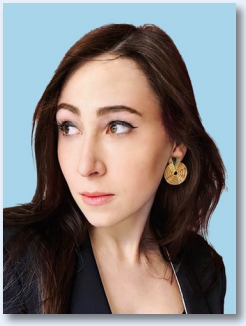
ПАРАЛЛЕЛИ МЕЖДУ ЕСТЕСТВЕННЫМ И ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

Индекс УДК 81-133, 004.82
Код ГРНТИ 16.31.21, 16.31.41
DOI: 10.22204/2587-8956-2025-123-04-122-135




Т.О. ШАВРИНА,
А.А. КОРНИЛОВ*

Мультиязычность в языковом моделировании: задачи, данные и возможности для типологических ресурсов

Последние достижения в области машинного обучения значительно улучшили возможности больших языковых моделей (Large Language Models, LLM), в том числе способности машинного перевода и машинного чтения. Тем не менее большинство языков мира остаются не покрытыми основными ресурсами, необходимыми для построения качественных речевых технологий и языковых моделей: корпусами текстов, аннотированными датасетами, достаточным количеством записей звучащей речи. Такие языки — языки с ограниченными письменными ресурсами — называют малоресурсными.

В настоящей статье мы представляем обзор современного состояния мультиязычности и поддержки малоресурсных языков в языковых моделях, а также проводим оценку способностей текущих моделей извлекать и классифицировать информацию из зачастую единственно доступного источника знаний для малоресурсных языков — дескриптивных грамматик. Мы предлагаем подход на основе метода дополненной генерации (Retrieval-Augmented Generation, RAG), позволяющий использовать такие описания для последующих задач, таких как машинный перевод. Наши тесты охватывают грамматические описания 248 языков из 142 языковых семей, фокусируясь на типологических характеристиках баз данных WALS [1] и Grambank [2].

Предлагаемый в работе подход обеспечивает первую комплексную оценку способности языковых моделей точно интерпретировать и извлекать лингвистические признаки в контексте, создавая критически важный ресурс для масштабирования технологий на малоресурсные языки. Код и данные доступны публично: https://github.com/al-the-eigenvalue/RAG-on-grammars.

Ключевые слова: многоязычность, языковые модели, бенчмарки, малоресурсные языки

* **Шаврина Татьяна Олеговна** — кандидат филологических наук, старший научный сотрудник Института языкознания РАН.
E-mail: rybolos@gmail.com
Корнилов Альберт Андреевич — бакалавр Высшей школы экономики.
E-mail: albert.kornilov801@gmail.com

1. Введение

1.1. Языковое разнообразие и его представленность в языковых моделях

Современные языковые модели трансформировали ландшафт обработки естественного языка, продвинувшись в машинном переводе, анализе текста, генерации речи и ряде других задач. Однако масштабный успех этих технологий сконцентрирован вокруг нескольких десятков, максимум сотен языков: как правило, самых широко распространённых. По различным оценкам, в мире насчитывается около 7000 языков[1] и около 70 тысяч диалектов, контактных языков, пиджинов и других языковых единиц[2], тогда как интернет-покрытие и параллельные корпуса для обучения моделей доступны лишь примерно для 1500. Таким образом, даже крупнейшие ресурсные и технические инициативы покрывают не более 20% от нижней оценки существующего языкового разнообразия.

Чтобы преодолеть разрыв между языками с высоким и низким объёмом ресурсов и сделать машинный перевод доступным для большего числа языков, были созданы новые бенчмарки перевода, ориентированные специально на малоресурсные языки. Конференция по машинному переводу (WMT[3]) теперь регулярно проводит открытые соревнования по машинному переводу для малоресурсных языков, например, для индийских[4] и африканских языков[5]; воркшопы, такие как AmericasNLP[6], поддерживают коренные языки [3]. А крупные коллаборации, такие как Masakhane[7] [4] и Aya[8], создали ресурсы для машинного перевода африканских [5], индонезийских языков [6] — в целом для более чем для 100 языков. Недавно корпус параллельного перевода FLORES-200 расширил покрытие данных для перевода до 200 языков. Помимо этих усилий, благодаря масштабной работе по очистке данных, фильтрации и идентификации языков исследователи смогли собрать данные и обучить модели машинного перевода для более чем 1000 языков [7].

На сегодняшний день самая многоязычная работа в области языкового моделирования — это модель для обработки звучащей речи Xeus [8]. Xeus — это открытая базовая модель для векторного представления звучащей речи, вспомогательная для многих применений и обученная почти на 1,1 млн часов неразмеченных аудиоданных на 4057 языках. Распределение данных внутри этой коллекции крайне неравномерное. Более чем у половины представленных языков объём записанной речи не превышает двух часов, что ставит их в категорию малоресурсных с точки зрения задач распознавания речи (по определению [9]). Таким образом, даже в самых масштабных многоязычных инициативах сохраняются существенные ограничения в объёмах данных для большинства языков.

Понятие «малоресурсный язык», или «low-resource language», в целом является относительным и зависит от конкретной задачи. В разных областях обработки языка минимальные пороговые значения различаются:

- для задач распознавания и обработки речи — менее двух часов аудиозаписей считается малоресурсным уровнем [9];

[1] 7159 — по базе данных Ethnologue (по состоянию на май 2025 г.). Электронный ресурс. URL: https://www.ethnologue.com/insights/how-many-languages/.

[2] 70 900 единиц в реестре LinguaSphere observatory (по данным переписи 2011 года). Электронный ресурс. URL: https://web.archive.org/web/20120614005015/http://www. linguasphere.info/.

[3] Электронный ресурс. URL: https://www2.statmt.org/wmt23/.

[4] Электронный ресурс. URL: https://www2.statmt.org/wmt23/indic-mt-task.html.

[5] Электронный ресурс. URL: https://www2.statmt.org/wmt23/african-mt-task.html.

[6] Электронный ресурс. URL: https://turing.iimas.unam.mx/americasnlp/2023_st.html.

[7] Электронный ресурс. URL: https://www.masakhane.io/.

[8] Электронный ресурс. URL: https://cohere.com/research/aya.

- для прикладных задач, таких как создание вопросно-ответных систем, тематической классификации докуметнов, извлечения именованных сущностей — менее 10 000 размеченных примеров [9];
- для обучения языковых моделей — менее 350 000 токенов текста.

Таким образом, малоресурсность — это не фиксированная характеристика языка, а динамическое понятие, которое определяется в зависимости от требований конкретных задач. Для задач обработки речи необходимы аудиозаписи, для обучения языковых моделей — большие текстовые корпуса, а для решения прикладных задач (например, машинного перевода или классификации) — размеченные данные. Один и тот же язык может считаться малоресурсным для одной задачи и вполне обеспеченным ресурсами — для другой. Поэтому при разработке мультиязычных моделей важно учитывать не только наличие данных, но и их вид, объём и качество в зависимости от цели использования.

Работа с привнесением многих языков в более высокоресурсную среду осложняется рядом фундаментальных проблем.

1. Дефицит данных: для успешной работы LLM требуются большие массивы данных. Однако даже минимальные ориентиры — два часа речи (для задач распознавания речи), 10 тысяч аннотированных примеров (для задач вроде классификации), 350 тысяч токенов текста (для языкового моделирования) — практически недостижимы для подавляющего большинства языков мира.
2. Отсутствие параллельных и монолингвальных корпусов: для малоресурсных языков крайне редко встречаются параллельные корпуса, которые лежат в основе современных моделей машинного перевода. Также отсутствуют большие монолингвальные коллекции, что делает обучение практически невозможным.
3. Проблемы валидации данных: даже если данные удаётся собрать, их качество и точность остаются под вопросом без участия носителей языка и профессиональных лингвистов.

В случаях, когда ресурсов практически нет (отсутствуют аудиозаписи, текстовые корпуса и размеченные примеры), единственным источником информации о языке остаются лингвистические ресурсы: дескриптивные грамматики и типологические базы данных. Такие материалы содержат систематизированное описание структуры языка (фонологии, морфологии, синтаксиса) и могут быть использованы как основа для создания моделей в условиях крайней малоресурсности.

Современные подходы, например Machine Translation from One Book (MTOB, [10]), показывают, что даже такие «ручные» источники можно интегрировать в работу с помощью методов, не требующих обучения языковой модели: дополненной генерации и длинного контекста.

Однако стоит отметить, что применение описательных грамматик для машинного перевода сталкивается с рядом проблем, таких как вариативность терминологии, нестандартные структуры и рассеянность релевантной информации. Кроме того, массивный тест для оценки текстовых вложений (MTEB, [11]) предоставляет детальную оценку текстовых вложений по различным задачам и языкам. Основная проблема остаётся в эффективном применении этих моделей к описательным грамматикам для малоресурсных языков, обычно поддерживаемых лишь лингвистическими материалами, такими как грамматики и словари.

Данная работа направлена на решение этих проблем, предлагая систематический подход к извлечению информации из описательных грамматик и создание масштабируемого метода для систематизации грамматических описаний. Ключевым аспектом предлагаемого подхода является метод дополненной генерации (RAG), позволяющий извлекать релевантную информацию из грамматик на основе конкретной типологической характеристики (например, порядок подлежащего, дополнения и сказуемого). На основе извлечён-

ных параграфов крупная языковая модель определяет значение этой характеристики (например, порядок слов «подлежащее—сказуемое—дополнение»).

В настоящей статье представлены следующие результаты:

1. Обзор состояния языковых моделей для применения к языковому разнообразию.
2. Первая масштабная лингвистическая оценка способностей машинного чтения крупных языковых моделей на материале описательных грамматик.
3. Методология на основе дополненной генерации (RAG), извлекающей релевантные параграфы из грамматик на основе заданной типологической характеристики (например, WALS 81A: порядок подлежащего, дополнения и сказуемого) и предоставляющей их в виде подсказок для крупной языковой модели с целью определения значений этих характеристик.

Предлагаемая архитектура, а также описанные тестовые наборы направлены на содействие дальнейшему развитию систем машинного перевода и повышение их качества и эффективности, а также на помощь лингвистам в типологических исследованиях путём частичной автоматизации извлечения данных из описательных грамматик.

Весь код работы открыт, распространяется под лицензией MIT: https://github.com/al-the-eigenvalue/RAG-on-grammars.

2. Обзор литературы

2.1. Метод дополненной генерации (RAG)

Метод дополненной генерации (retrieval-augmented generation, RAG) стал эффективным инструментом для повышения качества генерации в больших языковых моделях, дополняя их внутренние знания за счёт извлечения внешней информации. Вместо того чтобы полагаться исключительно на знания, встроенные в модель, RAG позволяет извлекать релевантные фрагменты из внешних источников в процессе генерации [12]. Этот подход доказал свою результативность в задачах открытого вопросно-ответного взаимодействия и суммирования документов, повышая достоверность информации и расширяя контекст.

Особенно перспективен метод RAG при работе с малоресурсными языками и сложными форматами, такими как лингвистические грамматики. Недавние исследования в области извлечения информации на основе запросов свидетельствуют о его применимости в условиях фрагментированных и неполных языковых данных. Настоящая работа применяет принципы RAG для повышения качества перевода и языкового моделирования с помощью описательных грамматик для языков с ограниченными ресурсами.

2.2. Обучение без примеров (Zero-shot Learning)

Zero-shot learning стало ключевым направлением в области обработки естественного языка, особенно с развитием фундаментальных моделей. Способность моделей к обобщению на новые задачи и языки без специфических обучающих данных критически важна для расширения применения NLP в отношении малоресурсных языков. Модели, такие как GPT-3 и GPT-4, демонстрируют впечатляющие возможности zero-shot для решения задач классификации, машинного перевода и др., что делает их незаменимыми инструментами в условиях ограниченного количества размеченных данных.

Тем не менее существующие zero-shot модели всё ещё испытывают трудности с языками, практически не представленными ни в монолингвальных, ни в билингвальных корпусах. Недавние исследования [10, 13] показали, что интеграция лингвистических описаний — таких как грамматики — может значительно улучшить качество zero-shot решений для таких языков. В этой работе мы продолжаем это направление, оценивая способность моделей использовать описательные грамматики в условиях отсутствия обучающих данных.

2.3. Использование грамматик в NLP

Несмотря на широкий спектр источников данных, используемых в NLP, таких как параллельные корпуса и размеченные датасеты, описательные грамматики остаются недостаточно используемым ресурсом. Между тем они содержат богатую структурированную информацию, которая особенно полезна для малоресурсных языков. Проекты Grambank и WALS показали, что типологические признаки можно систематизировать для лингвистического анализа и сопоставления языков.

Использование грамматик как источника знаний для языковых моделей всё чаще рассматривается как потенциально перспективное направление. В работах [10 и 14] подчёркивается, что лингвистическая информация из грамматик помогает строить обобщаемые представления о языках и их структурах, особенно в контексте задач машинного перевода и морфосинтаксического анализа. Тем не менее формализация, цифровизация и стандартизация таких ресурсов остаются открытыми вызовами. В данной статье мы частично преодолеваем этот барьер, интегрируя описательные грамматики в архитектуру RAG и предоставляя соответствующие тесты и инструменты.

2.4. Извлечение типологических признаков из грамматик

Существующие исследования по извлечению типологических признаков из грамматик предшествуют появлению крупных языковых моделей и основываются на правилах, классических методах машинного обучения и ранних версиях нейросетевых подходов. Серия работ [15] использует методы, требующие трудоёмкой аннотации семантических рамок. Работа [16] предлагает подход, применимый исключительно к бинарным типологическим признакам, а система, описанная в [17], ограничивается извлечением информации по ключевым словам.

В настоящей работе мы впервые применяем современные большие языковые модели для задачи извлечения типологических признаков и демонстрируем их потенциал в работе с лингвистическими описаниями.

2.5. Машинный перевод из одной грамматики

Ряд достижений в языковом моделировании открывает возможность использования лингвистических описаний: методы дополненной генерации для извлечения и генерации информации, способность моделей работать дескриптивными текстами в качестве затравок (prompt), наличие таких текстов в машиночитаемом формате.

Работа [10] представляет пример прикладного использования описательной грамматики малоресурсного языка для улучшения качества перевода на языки с чрезвычайно ограниченными ресурсами, показывая потенциал масштабных языковых моделей в преодолении разрыва между теоретической лингвистикой и практическими приложениями NLP.

Представленный в работе метод – Machine Translation from One Book — это новый подход к оценке способности языковых моделей выполнять перевод в условиях, когда язык целиком отсутствует в данных предварительного обучения. В отличие от традиционных бенчмарков машинного перевода, где модели обучаются на больших параллельных корпусах, MTOB предлагает моделям опереться только на единственный источник — дескриптивную грамматику и словарь языка Каламанг[1], представленные в оцифрованном виде. Такой сценарий моделирует реальные условия работы с экстремально малоресурсными и вымирающими языками, где отсутствуют как параллельные, так и монолингвальные данные.

Особенность подхода состоит в том, что модели получают в качестве контекста грамматическое описание языка и несколько примеров перевода, и на их основе

[1] Электронный ресурс https://en.wikipedia.org/wiki/Kalamang_language.

Input	**Kor kancing wa me an tur teba ma patin.**
Reference	My ankle bone, I fell and wounded it.
text-davinci-003	Even if I fall, my ankle bone is wounded.
gpt-3.5-turbo	I dream of a watch falling soon and injuring.
gpt-4	This ankle of mine is falling and progressively getting wounded.
Claude 2	This ankle of mine fell down while I was walking, and it got wounded.
Human	This ankle, I injured by falling.

Ил. 1. Качественный пример перевода с каламанга (kgv) на английский (Tanzer et al., 2024). Представлены ответы моделей в конфигурации с максимальным контекстом для каждой модели. Модель **text-davinci-003** правильно переводит лексику, но добавляет галлюцинированную конструкцию «even if». **GPT-3.5-turbo** отвлекается на нерелевантную лексику. **GPT-4** демонстрирует интересную ошибку: она интерпретирует значение частицы *teba* (маркер прогрессивного аспекта) как «progressively». **Claude 2** формулирует перевод странно, но почти правильно; однако в исходной фразе нет упоминания о хождении пешком, которое появляется на выходе модели. Человеческий перевод является точным, хотя и использует немного неестественную конструкцию для передачи топикализации и прогрессивного аспекта, присутствующих в оригинале

Input	**I'm getting pandanus, I want to make a mat.**
Reference	An padamualat repte kalifan paruotkin.
text-davinci-003	Kawat sie padamual, suka kangjie temun irar.
gpt-3.5-turbo	An kaloum bunga rampi, an suka rarie el.
gpt-4	An padamual gousat, suka an irar minggi.
Claude 2	An padamual rep=kin minggi kalifan paruo=kin
Human	An padamualat rep teba, elat paruotkin.

Ил. 2. Качественный пример перевода с английского на каламанг (eng→kgv) (Tanzer et al., 2024). Представлены ответы моделей в конфигурации с максимальным контекстом для каждой модели. **text-davinci-003** включает как релевантные, так и нерелевантные извлечённые слова с бессмысленной грамматикой. **GPT-3.5-turbo** и **GPT-4** начинают использовать местоимение «я» (*I*) и подбирают более последовательно релевантную лексику, однако перевод остаётся неграмматичным. **Claude 2** оперирует на уровне глосс, используя форму *=kin,* но упускает фонологические чередования, такие как *paruo=kin* → *paruotkin*; тем не менее содержание перевода в целом передано правильно: буквально «Я хочу добыть панданус и сделать из него циновку». Человеческий перевод использует более буквально переданную грамматику по сравнению с эталоном и применяет термин *el* (грубая циновка), а не *kalifan* (тонкая циновка); нам неизвестно, используется ли панданус действительно для обоих видов циновок

переводят новые предложения. Качество оценивается строго: язык должен отсутствовать в предобучении, а перевод основан только на грамматике и примерах. Для Каламанг используются латинские буквы, что облегчает задачу моделям. Задания — предложения для перевода в направлении «Каламанг — английский» или «английский — каламанг». Для оценки используются метрики машинного перевода и экспертная валидация.

3. Масштабирование на большее число языков

Несмотря на впечатляющие результаты [10] в задачах оценки пригодности одних лишь грамматик и словаря для работы с малоресурсным языком, остаётся открытым вопрос о том, насколько данный подход масштабируем и применим к другим языкам помимо каламанга, языкам с разными типологическими характеристиками и качеством грамматических описаний.

Далее рассмотрим эксперименты по расширению методологии MTOB для более широкой выборки языков [18]: мы предлагаем систематический подход к выбору типологически репрезентативной выборки языков, сбору и структурированию их грамматик, а также разработке автоматизированного метода оценки качества понимания таких грамматик языковыми моделями.

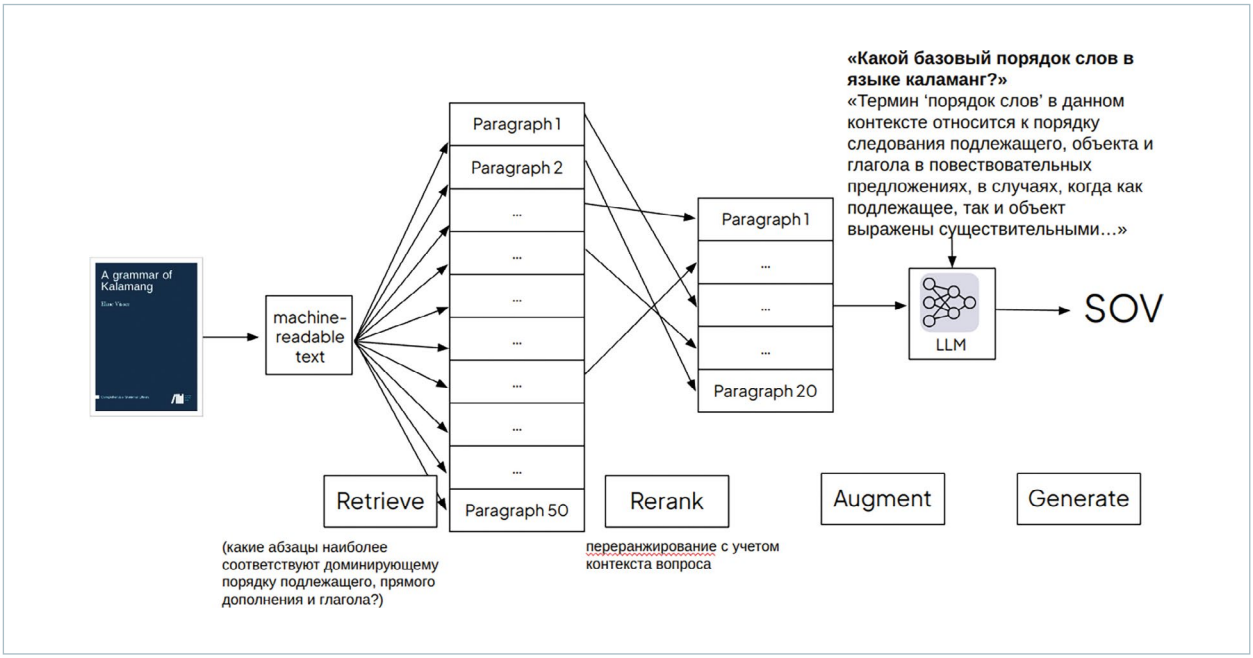
Как оценить способность языковых моделей масштабироваться на широкий спектр языков, особенно в тех случаях, когда имеются доступные грамматические описания, но отсутствует возможность привлечения носителей языка? Одним из возможных решений является использование не задач машинного перевода, а формата вопросно-ответных систем, где ответы модели можно валидировать на основе уже известных типологических характеристик языка. Такой подход позволяет не только оценивать корректность извлечения лингвистической информации, но и дополнительно проверять способность языковых моделей интерпретировать научный метаязык, используемый в дескриптивных грамматиках.

3.1. Базовая методология

Базовая методология строится как расширение подхода Retrieval-Augmented Generation (RAG), в котором модель сначала извлекает релевантные параграфы из грамматического описания языка с помощью классических методов информационного поиска (BM25), а затем использует языковую модель для генерации ответа на лингвистический вопрос.

Базовый метод дополненной генерации (Naive RAG) включает в себя базу документов (в нашем случае грамматики), компонент извлечения релевантных абзацев, процесс извлечения релевантных абзацев из базы документов на основе вопроса от пользователя, а также языковую модель, генерирующую ответ на основе этого вопроса и извлечённой информации.

Второй компонент — метод извлечения информации. Мы оцениваем BM25[1],



Ил. 3. Схема работы языковой модели с RAG на основе дескриптивной грамматики для вопросно-ответной системы

[1] Электронный ресурс: https://ru.wikipedia.org/wiki/Okapi_BM25.

Таблица 1

Распределение языков по макрорегионам

Макроареал	Всего языков
Африка	29
Австралия	9
Евразия	20
Северная Америка	25
Папуа Новая Гвинея	39
Южная Америка	26
Сумма	**148**

не зависящий от языка метод на основе частотности терминов, а также современные методы на основе эмбеддингов, представленные в рейтинге Massive Text Embedding Benchmark (MTEB) [11]. Такие методы устойчивы к лингвистическому разнообразию, поскольку описательные грамматики содержат примеры на исследуемых языках, включая диакритику и сегменты, редкие или отсутствующие в английском.

Третий компонент — формат затравки (prompt). Базовый шаблон включает абзац, вопрос о типологической характеристике, пояснение терминов и фиксированный набор ответов (см.: Приложение). Например, для признака WALS 81A «Преобладающий порядок: подлежащее, дополнение, сказуемое» возможные ответы: «SVO», «SOV», «VOS», «VSO», «OSV», «OVS», «Нет доминирующего порядка», а также «Недостаточно информации», если порядок не удаётся определить. Мы также реализуем стратегию затравок с расширенным описанием признака (из WALS или Grambank) и примерами — это вариант цепочки рассуждений (chain-of-thought prompting) [21].

Последний компонент — сама языковая модель. Мы используем GPT-4o, флагманскую модель OpenAI на май 2024 г. с улучшенными характеристиками по сравнению с GPT-4. Её задача — определить значение признака, например «4 падежа» для WALS 49A «Число падежей», на основе подсказки и параграфов из грамматики.

3.2. Данные

Бенчмарк для оценки метода RAG включает 148 описательных грамматик[1], поскольку бенчмарки с менее чем 100 примерами считаются ненадёжными: одна ошибка снижает точность более чем на один процент.

Случайный выбор грамматик мог бы исказить репрезентативность, создавая перекос в сторону языков из одних и тех же семей или географических регионов. Поэтому мы использовали метод Genus-Macroarea, описанный в [19] и реализованный в [20]. Как и в [20], мы берём распределение языков по макроареалам из списка родов WALS, автоматически выбираем грамматики из базы Glottolog References [22] и ограничиваем выбор одной грамматикой на род (genus). В отличие от Cheveleva, мы ограничили выбор грамматиками, написанными на английском языке.

Для оценки качества извлечения информации были выбраны четыре признака:

WALS 81A. Порядок подлежащего, дополнения и сказуемого, поскольку он зачастую явно указывается в грамматиках;

GB 107. Может ли стандартная негативная форма быть выражена аффиксом, клитикой или изменением глагола? Это бинарный признак, но трудный для наивного извлечения из-за вариативности терминологии. Составной признак, связанный с общевопросными предложениями (полярными вопросами);

[1] https://anonymous.4open.science/r/from-MTEB-to-MTOB/ground_truth_rag.csv.

Таблица 2

F1-оценки для всех конфигураций архитектуры

	F1 мера	Baseline	BM25	BM25+CoT
Все признаки	micro	0,5551 ± 0,0359	0,6892	**0,7027**
Все признаки	macro	0,2812 ± 0,0276	**0,7179**	0,6097
WALS 81A	weighted	0,5328 ± 0,0337	0,6694	**0,6890**
GB 107	weighted	0,5724 ± 0,0361	0,5957	**0,5959**
WALS 49A	weighted	0,3453 ± 0,0237	0,5314	**0,5542**
interrog.intonation only	weighted	0,4068 ± 0,0484	0,8480	**0,9007**
interrog.word order	weighted	0,9611 ± 0,0064	**0,9936**	0,9878
clause-initial particle	weighted	0,7371 ± 0,0337	0,9054	**0,9205**
clause-final particle	weighted	0,4661 ± 0,0428	0,7314	**0,7644**
clause-medial particle	weighted	0,6644 ± 0,0355	0,8439	**0,8888**
interrog. verb morphology	weighted	0,7102 ± 0,0160	0,8192	**0,8451**
tone	weighted	0,9104 ± 0,0121	0,9390	**0,9637**

Примечание: Все пять конфигураций используют промпты из Приложения с включением Wikipedia-резюме для соответствующих признаков. Колонка **Baseline** соответствует запросам к GPT-4o без использования материалов из грамматик (т.е. без применения RAG). **BM25** означает использование 50 параграфов, извлечённых методом BM25. **CoT** обозначает добавление цепочки рассуждений (Chain-of-Thought), т.е. инструкций и примеров из WALS или Grambank в промпт

WALS 116A. Он объединяет семь бинарных признаков из Grambank и оценивает способность моделей к рассуждению на основе нескольких реализаций одного феномена;

WALS 49A. Число падежей — количественный признак, требующий широкого охвата грамматики (разделы морфологии и синтаксиса).

Все признаки обладают следующими характеристиками:

- Интерпретируемость — признаки должны быть формулируемы в виде понятного лингвистического вопроса.
- Наличие аннотации в WALS или Grambank — для обеспечения возможности автоматической валидации.
- Независимость от конкретных языков — признаки должны быть универсальными и применимыми к разным языкам.
- Наличие явного описания в грамматике — чтобы их можно было извлечь с помощью RAG-архитектуры.

3.3. Результаты: оценка метода

Для оценки любой NLP-задачи по бенчмарку важно определить, обладает ли языковая модель знаниями о значениях признаков, включённых в процедуру оценки.

Чтобы установить базовый уровень (baseline) для GPT-4o, т.е. оценить его работу без привлечения грамматики, мы провели тест, исключив модуль извлечения из архитектуры RAG. Мы предложили GPT-4o определить значения всех признаков без использования извлечённых параграфов — модель получала только подсказку и краткое содержание статьи из Википедии по соответствующему признаку. Для каждого признака тест запускался десять раз, чтобы отразить разброс результатов и точнее различить случаи наличия и отсутствия знаний.

Результаты всех конфигураций метода RAG представлены в табл. 2. Все конфигурации RAG превосходят базовое решение (генерацию ответа без грамматики). Более высокие макроусреднённые значения F1-меры[1], по сравнению с микроусреднён-

[1] https://en.wikipedia.org/wiki/F-score.

ными, указывают на то, что метод RAG лучше справляется с частотными классами и испытывает сложности при наличии дисбаланса классов в типологических профилях языков мира.

Следует отметить, что далеко не все типологические признаки были успешно извлечены с высокой точностью. В частности, такие широко распространённые характеристики, как базовый порядок слов и выражение глагольного отрицания, несмотря на их частое упоминание в грамматических описаниях, демонстрируют низкое качество извлечения языковой моделью. Проведённый тест на контаминацию (Baseline) показал, что в ряде случаев модели уже обладают встроенными ассоциациями между языками и значениями признаков, что позволяет им правильно отвечать даже без обращения к тексту грамматики. Этот эффект может искусственно завышать показатели качества извлечения и затруднять объективную оценку работы модели по менее представленным и редким признакам.

Использование цепочек рассуждений (Chain-of-Thought prompts) и предоставление инструкций, основанных на примерах из WALS и Grambank, существенно улучшают интерпретацию модели, особенно для сложных грамматических конструкций. В то же время обработка нескольких признаков одновременно показала непредсказуемые результаты, подчёркивая сложность автоматического извлечения типологических данных из грамматик.

Расширение подхода MTOB может существенно выиграть от приведения описательных грамматик различных языков к унифицированному формату, используя базы данных, такие как Grambank или WALS.

Тем не менее даже в строго контролируемой среде, продемонстрированной в данной работе, дескриптивные лингвистические тексты остаются значительным вызовом. Несмотря на то, что машинное чтение в целом может рассматриваться как «решённая задача», результаты по извлечению лингвистических признаков показывают, что описательные грамматики остаются нетривиальным источником, выявляющим слабые стороны языковых моделей.

Заключение

Расширение подхода MTOB может существенно выиграть от стандартизации описательных грамматик различных языков в единый формат с опорой на базы данных, такие как Grambank или WALS. Тем не менее даже в неконтаминированной среде, рассмотренной в данной работе, описательные лингвистические тексты продолжают представлять значительную сложность.

В данной статье мы представили метод оценки решений, объединяющих дополненную генерацию (RAG) и большие языковые модели, с целью извлечения и классификации типологических признаков из описательных грамматик. Также мы представили метод для извлечения лингвистической информации, обладающий значительным потенциалом для улучшения NLP-систем на малоресурсных языках.

Кроме того, несмотря на распространённое мнение о решённости задачи машинного чтения, лингвистические работы остаются областью повышенной сложности. Хотя языковые модели значительно продвинулись в обработке различных типов текстов, полученные нами результаты указывают на то, что пока ещё рано говорить об их полной эффективности в домене описательной лингвистики.

Наши результаты закладывают основу для расширения возможностей языковых моделей в работе со сложными лингвистическими данными, такими как грамматические описания. Эта работа представляет собой важный шаг в направлении поддержки малоресурсных языков в NLP.

В дальнейшем возможны улучшение компонентов извлечения и классификации, расширение бенчмарка за счёт включения большего числа языков, а также исследование практических применений извлечения лингвистической информации, например, кросс-языковая типологическая аналитика или машинный перевод для крайне малоресурсных языков.

ЛИТЕРАТУРА

1. Dryer M.S., Haspelmath M. (eds.). WALS Online (v2020.4) [Data set]. Zenodo, 2013. DOI: 10.5281/zenodo.13950591.
2. Skirgård H., Haynie H., Passmore S. et al. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss // Science Advances. 2023. Vol. 9, № 16. Article eadg6175. DOI: 10.1126/sciadv.adg6175.
3. Ebrahimi A. et al. Findings of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages // Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP). Toronto, Canada: Association for Computational Linguistics, 2023. Pp. 206–219.
4. Lovenia H. et al. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, USA: Association for Computational Linguistics, 2024. Pp. 5155–5203.
5. Nekoto W. et al. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020. Pp. 2144–2160.
6. Winata G.I. et al. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Pp. 815–834.
7. Bapna A. et al. Building machine translation systems for the next thousand languages. arXiv preprint arXiv:2205.03983, 2022.
8. Chen W. et al. Towards Robust Speech Representation Learning for Thousands of Languages // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, USA: Association for Computational Linguistics, 2024. Pp. 10205–10224.
9. Garrette D., Mielens J., Baldridge J. Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). Vol. 1: Long Papers. Sofia, Bulgaria: Association for Computational Linguistics, 2013. Pp. 583–592.
10. Tanzer G., Suzgun M., Visser E., Jurafsky D., Melas-Kyriazi L. A benchmark for learning to translate a new language from one grammar book. arXiv preprint arXiv:2309.16575, 2023.
11. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: Massive Text Embedding Benchmark // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Pp. 2014–2037.
12. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33. Pp. 9459–9474.
13. Zhang K., Choi Y., Song Z., He T., Wang W.Y., Li L. Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions // Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024. Pp. 15654–15669.
14. Ponti E.M., Glavaš G., Majewska O., Liu Q., Vulić I., Korhonen A. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020. Pp. 2362–2376.
15. Virk S.M., Foster D., Sheikh M.A., Saleem R. A Deep Learning System for Automatic Extraction of Typological Linguistic Information from Descriptive Grammars // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Online: INCOMA Ltd., 2021. Pp. 1480–1489.

16. Hammarström H., Her O.-S., Allassonnière-Tang M. Term spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions // Selected Contributions from the Eighth Swedish Language Technology Conference (SLTC-2020). 2020. Pp. 27–34.
17. Kornilov A. Multilingual Automatic Extraction of Linguistic Data from Grammars // Proceedings of the Second Workshop on NLP Applications to Field Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Pp. 86–94.
18. Kornilov A., Shavrina T. From MTEB to MTOB: Retrieval-Augmented Classification for Descriptive Grammars. arXiv preprint arXiv:2411.15577, 2024.
19. Miestamo M., Bakker D., Arppe A. Sampling for variety // Linguistic Typology. 2016. Vol. 20, № 2. Pp. 233–296.
20. Cheveleva A. Neutralization of gender values in the plural. Bachelor's thesis. Moscow: HSE University, 2023.
21. Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q.V., Zhou D. et al. Chain-of-thought prompting elicits reasoning in large language models // *Advances in Neural Information Processing Systems (NeurIPS)*. 2022. Vol. 35. Pp. 24824–24837.
22. Hammarström H., Forkel R., Haspelmath M., Bank S. (eds.). *Glottolog 5.0* [Data set]. Zenodo, 2024. DOI: 10.5281/zenodo.8635585.

Приложение

Затравка для определения числа падежей (на основе WALS 49A);

Please determine the number of cases in the language <...>. The term "cases" in the context of this feature refers to productive case paradigms of nouns. Reply with one of the 9 following options: No morphological case-marking, 2 cases, 3 cases, 4 cases, 5 cases, 6-7 cases, 8-9 cases, 10 or more cases, Exclusively borderline case-marking. The feature value "Exclusively borderline case-marking" refers to languages which have overt marking only for concrete (or "peripheral", or "semantic") case relations, such as locatives or instrumentals. Categories with pragmatic (non-syntactic) functions, such as vocatives or topic markers, are not counted as case even if they are morphologically integrated into case paradigms. Genitives are counted as long as they do not encode categories of the possessum like number or gender as well, if they do not show explicit adjective-like properties. Genitives that may take additional case affixes agreeing with the head noun case ("double case") are not regarded as adjectival. 1. Provide the reasoning for the chosen option. 2. After the reasoning, output the word "Conclusion:" and the chosen option at the end of your response.

Multilinguality in Language Modeling: Tasks, Data, and Opportunities for Typological Resources

Tatiana Olegovna Shavrina — Ph.D. in Philology, Senior Researcher, Institute of Linguistics, Russian Academy of Sciences.
E-mail: rybolos@gmail.com

Albert Andreevich Kornilov — Bachelor's Degree, Higher School of Economics.
E-mail: albert.kornilov801@gmail.com

This paper addresses the significant challenge of building language technologies for the majority of the world's under-resourced languages, which lack the large text corpora and annotated datasets necessary for modern machine learning. While advances in Large Language Models

(LLMs) have revolutionized machine translation and reading comprehension, these models often underperform or fail entirely for languages with limited written resources.

We present an overview of current multilingual support in LLMs and evaluate their ability to understand the primary available knowledge source for such languages: descriptive grammars. To effectively utilize this structured but complex information, we propose a Retrieval-Augmented Generation (RAG) framework. This approach enables models to accurately extract and interpret linguistic features from grammatical texts, facilitating downstream tasks like machine translation. Our evaluation provides the first comprehensive assessment of model performance on this critical task, covering grammatical descriptions of 248 languages from 142 language families. The analysis focuses on the typological characteristics of the WALS [1] and Grambank [2] databases.

The proposed approach demonstrates the first comprehensive assessment of the ability of language models to accurately interpret and extract linguistic features in context, creating a critical resource for scaling technologies to under-resourced languages. Code and data from this study are made publicly available: https://github.com/al-the-eigenvalue/RAG-on-grammars.

Keywords: multilingualism, language models, benchmarks, under-resourced

REFERENCES

1. Dryer M.S., Haspelmath M. (eds.). WALS Online (v2020.4) [Data set]. Zenodo, 2013. DOI: 10.5281/zenodo.13950591.
2. Skirgård H., Haynie H., Passmore S. et al. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss // Science Advances. 2023. Vol. 9, № 16. Article eadg6175. DOI: 10.1126/sciadv.adg6175.
3. Ebrahimi A. et al. Findings of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages // Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP). Toronto, Canada: Association for Computational Linguistics, 2023. Pp. 206–219.
4. Lovenia H. et al. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, USA: Association for Computational Linguistics, 2024. Pp. 5155–5203.
5. Nekoto W. et al. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020. Pp. 2144–2160.
6. Winata G.I. et al. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Pp. 815–834.
7. Bapna A. et al. Building machine translation systems for the next thousand languages. arXiv preprint arXiv:2205.03983, 2022.
8. Chen W. et al. Towards Robust Speech Representation Learning for Thousands of Languages // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, USA: Association for Computational Linguistics, 2024. Pp. 10205–10224.
9. Garrette D., Mielens J., Baldridge J. Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). Vol. 1: Long Papers. Sofia, Bulgaria: Association for Computational Linguistics, 2013. Pp. 583–592.

10. Tanzer G., Suzgun M., Visser E., Jurafsky D., Melas-Kyriazi L. A benchmark for learning to translate a new language from one grammar book. arXiv preprint arXiv:2309.16575, 2023.
11. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: Massive Text Embedding Benchmark // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Pp. 2014–2037.
12. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33. Pp. 9459–9474.
13. Zhang K., Choi Y., Song Z., He T., Wang W.Y., Li L. Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions // Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024. Pp. 15654–15669.
14. Ponti E.M., Glavaš G., Majewska O., Liu Q., Vulić I., Korhonen A. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020. Pp. 2362–2376.
15. Virk S.M., Foster D., Sheikh M.A., Saleem R. A Deep Learning System for Automatic Extraction of Typological Linguistic Information from Descriptive Grammars // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Online: INCOMA Ltd., 2021. Pp. 1480–1489.
16. Hammarström H., Her O.-S., Allassonnière-Tang M. Term spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions // Selected Contributions from the Eighth Swedish Language Technology Conference (SLTC-2020). 2020. Pp. 27–34.
17. Kornilov A. Multilingual Automatic Extraction of Linguistic Data from Grammars // Proceedings of the Second Workshop on NLP Applications to Field Linguistics. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. Pp. 86–94.
18. Kornilov A., Shavrina T. From MTEB to MTOB: Retrieval-Augmented Classification for Descriptive Grammars. arXiv preprint arXiv:2411.15577, 2024.
19. Miestamo M., Bakker D., Arppe A. Sampling for variety // Linguistic Typology. 2016. Vol. 20. № 2. Pp. 233–296.
20. Cheveleva A. Neutralization of gender values in the plural. Bachelor's thesis. Moscow: HSE University, 2023.
21. Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q.V., Zhou D. et al. Chain-of-thought prompting elicits reasoning in large language models // Advances in Neural Information Processing Systems (NeurIPS). 2022. Vol. 35. Pp. 24824–24837.
22. Hammarström H., Forkel R., Haspelmath M., Bank S. (eds.). Glottolog 5.0 [Data set]. Zenodo, 2024. DOI: 10.5281/zenodo.8635585.