ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2026. Том 19. Выпуск 8 | 2026. Volume 19. Issue 8
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Оценка автоматического упрощения немецкоязычных художественных текстов с использованием больших языковых моделей

Хохлова М. В., Шер А. П.

Аннотация. Целью исследования является апробация применения больших языковых моделей для упрощения немецкоязычных художественных текстов. В работе дан краткий обзор существующих подходов к решению данной задачи, а также статистических метрик, которые используются для оценки сложности текстов и их удобочитаемости. Для значений ряда количественных показателей, полученных на материале оригинальных и упрощенных текстов, были проведены два вида статистического анализа (дисперсионный и корреляционный), которые были использованы для проверки гипотезы о существовании различий между разными группами текстов, а также взаимосвязи параметров. Научная новизна исследования заключается в том, что впервые была проведена оценка полученных текстов с привлечением тринадцати квантитативных показателей, а также было установлено, что при упрощении немецкоязычных художественных текстов большими языковыми моделями в изменении значений показателя лексической плотности нет статистически значимых различий. Результаты анализа показали, что тексты, упрощенные с помощью моделей Llama 3 и MarianMT, характеризуются лексикой более низких уровней по шкале CEFR; кроме того, модель Llama 3 существенно сокращает длину текстов при упрощении.

EN

Evaluation of automatic simplification of German-language literary texts using large language models

M. V. Khokhlova, A. P. Sher

Abstract. The aim of this study is to investigate the applicability of large language models for the simplification of German-language literary texts. The paper provides a brief overview of existing approaches to this task, as well as statistical metrics used to assess text complexity and readability. Two types of statistical analysis (analysis of variance and correlation analysis) were applied to the values of a range of quantitative indicators obtained from original and simplified texts. These methods were used to test the hypothesis concerning the existence of differences between different groups of texts, as well as the relationships between the analyzed parameters. The scientific novelty of the study lies in the first comprehensive evaluation of the generated simplified texts based on thirteen quantitative indicators. The study also demonstrates that, in the process of simplifying German-language literary texts with large language models, changes in lexical density do not exhibit statistically significant differences. The results of the analysis show that texts simplified using the Llama 3 and MarianMT models are characterized by vocabulary corresponding to lower levels of the CEFR scale. Furthermore, the Llama 3 model significantly reduces text length during the simplification process.

Введение

Актуальность настоящего исследования продиктована тем, что в современной лингвистике существует необходимость в изучении как механизмов, лежащих в основе языковой сложности, так и способов ее снижения при помощи автоматических подходов. Особое значение данная проблема приобрела в связи с активным развитием больших языковых моделей (БЯМ), поскольку качество получаемых при помощи них результатов требует лингвистической оценки. Дополнительно необходимо отметить следующее соображение, также определяющее актуальность исследования: БЯМ, которые используются для решения разнообразных задач

Научная статья (original research article) | https://doi.org/10.30853/phil20260378

и показывают высокие результаты при упрощении или перефразировании текстов, были обучены на больших объемах данных, в которых художественная литература занимает только часть, в то время как наибольшая доля приходится на публицистику и Интернет-коммуникацию. Следовательно, необходимо оценить результаты автоматических процедур применительно к художественному стилю.

Актуальность работы также обусловлена тем, что задача упрощения текстов имеет в настоящее время большое значение, а полученные результаты могут найти широкое применение в разных областях знаний. В ряде случаев специализированные тексты, например, юридические или медицинские, требуется перефразировать для того, чтобы они стали понятны носителям языка, не обладающим соответствующей профессиональной подготовкой. Термины, а также сложные синтаксические конструкции могут заменяться более простыми аналогами, благодаря чему неспециалисты имеют возможность получить из документов нужную информацию.

Для достижения цели исследования необходимо было решить следующие задачи:

1. Дать обзор существующих методов и моделей для упрощения текстов.
2. Сформировать пилотный корпус оригинальных немецкоязычных художественных текстов.
3. Выполнить эксперимент по упрощению текстов с учетом контекста с применением БЯМ.
4. Оценить полученные адаптации с помощью метрик удобочитаемости и лексической сложности.
5. Провести статистический анализ и оценить взаимосвязь исследуемых характеристик текстов.

Методы исследования обусловлены целью и совокупностью поставленных задач. Для автоматического упрощения текстов был использован экспериментальный метод, который предусматривает применение БЯМ к текстам с использованием единой стратегии генерации. Оценка качества упрощенных текстов осуществлялась при помощи методов квантитативной лингвистики с применением метрик удобочитаемости, лексической сложности и лексического разнообразия. Дисперсионный и корреляционный статистический анализ был использован для проверки гипотезы о существовании различий между разными группами текстов, а также выявления взаимосвязей между их количественными характеристиками.

Материалом исследования послужили оригинальные и упрощенные при помощи выбранных БЯМ (mT5, MarianMT и Llama 3) тексты на немецком языке. В качестве оригинальных художественных текстов были отобраны рассказы Иоганна Петера Хебеля, Роберта Вальзера и Франца Кафки: объем каждого из произведений не превышал 500 слов, чтобы обеспечить их более быструю обработку при помощи БЯМ. Всего было рассмотрено 188 текстов общим объемом 55058 слов.

Теоретическую базу исследования составляют работы, посвященные упрощению текстов (англ. text simplification), которое может выполняться экспертами или автоматически, а также исследования по автоматической оценке их сложности применительно к преподаванию/изучению иностранных языков. Для решения задачи автоматического упрощения текстов первоначально применялись методы, основанные на правилах, которые подразумевали не только перестраивание синтаксических структур, но и привлечение словарей синонимов (Chandrasekar, Doran, Srinivas, 1996; Carroll, Minnen, Canning et al., 1998). В работе (Specia, 2010) было предложено рассматривать упрощение текстов как внутриязыковой перевод со сложного языка на простой, что позволило применять методы статистического машинного перевода к параллельным корпусам текстов – оригинальным текстам соответствовали их версии, упрощенные экспертами.

С развитием нейронных сетей рассматриваемая задача стала включать в себя генерацию целого текста на основе заданного с помощью таких архитектур, как RNN (англ. Recurrent Neural Network) и LSTM (англ. Long Short-Term Memory), к которым позднее добавился трансформер (англ. Transformer) (Vaswani, Shazeer, Parmar et al., 2017). В результате процесс упрощения стал включать несколько этапов: обученные на параллельных корпусах нейросетевые модели позволили одновременно производить лексическую замену, перефразирование и удаление или добавление информации. В работе (Alva-Manchego, Martin, Bordes et al., 2020) были предложены варианты организации корпусов, позволяющие более точно оценить результаты проведенных моделями процедур.

Оценка эффективности используемых алгоритмов традиционно производится с помощью таких метрик, как BLEU (Papineni, Roukos, Ward et al., 2002) или SARI (Xu, Napoles, Pavlick et al., 2016), которые требуют сравнения автоматически упрощенных текстов с текстами, которые были отредактированы экспертами. Однако существующие метрики подвергаются критике из-за неспособности описывать комплексное упрощение (Alva-Manchego, Scarton, Specia, 2021; Sulem, Abend, Rappoport, 2018).

Для немецкого языка на данный момент существует несколько параллельных корпусов, используемых в работах по упрощению текстов: например, APA (нем. Austria Presse Agentur), Wikipedia и другие (Ebling, Battisti, Kostrzewa et al., 2022). На материале этих корпусов проводятся исследования, включающие в себя, например, упрощение текстов до заданного уровня шкалы CEFR или добавление дополнительной информации в упрощенный текст (Spring, Rios, Ebling, 2021; Hewett, Asghari, Stede, 2024). Оценке сложности учебных текстов для преподавания русского языка как иностранного уделяется внимание в работах (Солнышкина, Андреева, 2026; Солнышкина, Андреева, Гайнутдинова, 2026).

Практическая значимость работы заключается в возможности использования исследуемых БЯМ в прикладных задачах: при упрощении или комментировании текстов, ориентированных на обучающихся, которые владеют немецким языком на начальном или среднем уровнях. Полученные результаты могут найти применение при разработке рекомендаций по работе с БЯМ, в преподавании иностранных языков, в частности, при создании учебных материалов (например, упражнений) и учебных пособий.

Обсуждение и результаты

Сложность текста определяется рядом факторов, проявляющихся на разных уровнях (например, на лексическом, синтаксическом или дискурсивном), и может получить оценку при помощи экспертов-лингвистов (Seiffe, Kallel, Möller et al., 2022) или в результате психолингвистических экспериментов (например, с привлечением айтрекинга (Vaněček, Kursch, Şen et al., 2026)). Помимо этого, с помощью специально разработанных формул можно измерить удобочитаемость текста – то, насколько легко тест читается, вне зависимости от предполагаемой сложности. Данные метрики основываются на статистических показателях, и некоторые из них носят название индексов удобочитаемости. Удобочитаемость текста характеризует сложность восприятия текста читателем и вычисляется на основе различных параметров с учетом особенностей конкретного языка. Например, могут учитываться средняя длина слова, средняя длина предложения, количество редких слов и др.

Индекс Флеша (англ. Flesch Reading Ease, FRE), традиционно используемый в многочисленных исследованиях, был введен в работе (Flesch, 1948) для английского языка:

$$FRE_{English} = 206{,}835 - 1{,}015 \times ASL - 84{,}6 \times ASW,$$

где ASL (англ. Average Sentence Length) – средняя длина предложения в словах;

ASW (англ. Average Syllables per Word) – средняя длина слова в слогах.

Полученные значения позволяют охарактеризовать тексты следующим образом: 0 – сложный текст, 100 – простой текст.

Для немецкого языка формула была адаптирована с учетом большей средней длины слов по сравнению с английским языком (Amstad, 1978) и имеет следующий вид:

$$FRE_{German} = 180 - ASL - (58{,}5 \times ASW).$$

Еще один индекс удобочитаемости под названием LIX (швед. Läsbarhetsindex) был разработан Карлом-Хуго Бьернссоном (Björnsson, 1968), при этом для его вычисления длина сложных слов измеряется не в слогах, а в символах:

$$LIX = ASL + L \div W \times 100,$$

где L – количество слов длиннее 6 символов;
W – количество всех слов текста.

Вместе с тем для более комплексной оценки необходимо использовать и другие метрики, которые учитывают не только длину слов, предложений или текстов (см. более полный обзор в работе (Захарова, Савина, 2020)): к числу метрик лексической сложности, например, относятся также метрики лексического разнообразия, описывающие долю уникальных слов в тексте, и лексической плотности, описывающие долю знаменательных слов в тексте.

Так, широко используются показатели TTR (англ. Type-Token Ratio), RTTR (англ. Root Type-Token Ratio) и MTLD (англ. Measure of Textual Lexical Diversity), которые принимают во внимание долю уникальных слов в тексте и вычисляются по следующим формулам:

$$TTR = T \div W,$$

где T – количество уникальных слов;

$$RTTR = T \div \sqrt{W},$$
$$MTLD = W \div F,$$

где F – количество факторов текста (последовательных отрывков текста, на которых значение TTR не падает ниже заданного порогового значения).

Лексическая плотность, входящая в число метрик лексической сложности и описывающая долю знаменательных слов в тексте, рассчитывается следующим образом:

$$D = C \div W,$$

где C – количество знаменательных слов текста.

Значения, которые принимают индексы удобочитаемости и показатели лексической сложности, должны учитываться при оценивании общей сложности текстов (Оборнева, 2005). Лексическое упрощение является важной частью общего упрощения текстов, и в рамках этой подзадачи разрабатываются инструменты для лексической замены сложных слов в тексте для разных языков (Qiang, Li, Zhu et al., 2020; Truică, Stan, Apostol, 2023; Большакова, 2024).

Помимо этого, лексика соответствует определенным уровням владения языком по шкале CEFR. Как правило, более сложные тексты характеризуются лексикой более высокого уровня. Этот фактор, как и вышеперечисленные, тоже оказывает влияние на общую сложность текста.

В работе были использованы лексические минимумы для уровней A1, A2 и B1 по шкале CEFR (примеры списков представлены в Таблице 1). Для выделения лексики, относящейся к выбранным уровням, оригинальные тексты были лемматизированы с помощью библиотеки spaCy [spaCy]. Была выполнена их предобработка, в результате которой каждая лемма входила в одну из четырех групп: в группу одного из трех перечисленных выше уровней (A1, A2 или B1) или четвертую группу, куда попали все остальные леммы. Лексика из последней группы была обозначена как «сложная», то есть требующая в дальнейшем замены более простыми синонимами.

Таблица 1. *Примеры лексики уровней A1, A2, B1 для немецкого языка*

A1	A2	B1	«сложная» лексика
Unterricht (урок, занятие)	furchtbar (ужасный, страшный)	beschädigen (повреждать, портить)	Verwunderung (удивление, изумление)
kaputt (сломанный, испорченный)	Taschengeld (карманные деньги)	Kneipe (пивная, трактир)	Handwerksbursche (подмастерье)
kaufen (покупать)	Verkehr (движение (транспорта), транспорт)	Küste (побережье)	gnädig (милостивый, благосклонный)
teuer (дорогой)	vorbereiten (готовить, подготавливать)	wertlos (ничего не стоящий, бесполезный)	bekanntlich (как известно, общеизвестно)
Frühstück (завтрак)	wenigstens (по крайней мере, хотя бы)	behaupten (утверждать, заявлять)	betrachten (рассматривать, смотреть на)

Снижение доли слов, выходящих за пределы лексических минимумов для начальных уровней владения языком, характерно для упрощенных текстов (Дмитриева, Лапошина, Лебедева, 2021). Взаимосвязь количественных характеристик текста с приписанным ему уровнем CEFR исследуется в рамках решения задачи об упрощении текстов для изучающих второй язык (Zhang, Lu, 2025).

Для подбора корректной синонимической замены были привлечены материалы из словаря синонимов Duden (2016), при этом все приведенные синонимы слова рассматривались как равноправные. Например, для лексемы “gnädig” был получен следующий синонимический ряд: 1) *entgegenkommend* (предупредительный, отзывчивый), *liebenswürdig* (любезный, обаятельный), *wohlwollend* (доброжелательный, благожелательный), *zuvorkommend* (внимательный, предупредительный); 2) *gütig* (добрый, добросердечный), *milde* (мягкий, снисходительный), *nachsichtig* (снисходительный, терпимый); *(geh.) barmherzig* (милосердный, сострадательный).

В рамках исследования был проведен эксперимент по автоматическому упрощению немецкоязычных художественных текстов с помощью БЯМ архитектуры трансформер: модели mT5 и MarianMT типа энкодер-декодер, а также модель Llama 3 типа декодер. Для моделей mT5 и Llama 3 были сформулированы промпты на разных языках в соответствии с тем, какой язык обеспечивал более точное выполнение приведенной инструкции: для mT5 – на английском языке, для Llama 3 – на немецком. Модель MarianMT не требовала инструкции, поскольку предназначена исключительно для машинного перевода.

Следующие промпты привели к наиболее точному выполнению задачи:

- mT5: “Rewrite this German text using simpler German words, but keep the original meaning.” / Перепиши этот текст на немецком при помощи более простых немецких слов, но сохрани содержание;
- Llama 3: “Ersetze schwierige Wörter durch einfache Synonyme, ohne den Inhalt zu verändern.” / Замени сложные слова простыми синонимами, не изменяя содержания.

На вход всем моделям подавались короткие немецкие рассказы в оригинальном виде. Дополнительно модели mT5 и Llama 3 получали инструкцию упростить прикрепленный текст. Таким образом, модели mT5 и Llama 3 должны были перефразировать оригинальный текст с использованием более простой лексики, в то время как при помощи модели MarianMT необходимо было осуществить перевод текста в два этапа (первоначально с немецкого языка на английский, а затем обратно, чтобы в процессе перевода сложные слова автоматически были заменены простыми синонимами – теми, которые относились бы к более низким уровням по шкале CEFR). В итоге для каждого из 188 оригинальных текстов было создано три упрощенных варианта – по одному при помощи каждой модели. На основе полученных данных был собран датасет из 752 текстов, разбитых на группы по источнику (оригинальный текст или упрощенный одним из трех способов). В Таблице 2 приведен пример упрощения для короткого юмористического рассказа Иоганна Петера Хебеля “ Des Dieben Antwort” («Ответ вора»): 1. Einem Dieb, der sich mit Reden mausig machen wollte, sagte jemand: “Was wollt Ihr? Ihr dürft ja gar nicht mehr in Eure Heimat zurückkehren und müsst froh sein, wenn man Euch hier duldet.” / Одному вору, который хотел пустить людям пыль в глаза своими речами, кто-то сказал: – Чего Вы хотите? Ведь вам нельзя даже вернуться на родину, и Вы должны быть рады уже тому, что Вас здесь терпят (здесь и далее перевод выполнен авторами статьи. – *М. Х., А. Ш.*). 2. “Meint Ihr?” sagte der Dieb; “meine Herren daheim haben mich so lieb, ich weiss gewiss, wenn ich heimkäme, sie liessen mich nimmer fort.” / – Вы так думаете? – ответил вор. – Мои земляки так меня любят, что я совершенно уверен: вернись я домой, они бы меня уже больше никогда не отпустили.

Оценка полученных адаптированных текстов осуществлялась следующим образом. Во-первых, для всех текстов было рассчитано процентное соотношение лексики по уровням шкалы CEFR: доля слов, относящихся к уровням A1, A2, B1, а также доля «сложных» слов, которые не вошли ни в одну из вышеперечисленных категорий. Во-вторых, на этапе анализа и для оригинальных, и для упрощенных текстов были посчитаны метрики удобочитаемости, лексической плотности и лексического разнообразия. Кроме того, для упрощенных текстов были оценены доля добавленных при упрощении сложных слов среди всех новых слов и доля правильно подобранных

синонимов среди всех сложных слов оригинального текста. В-третьих, был выполнен однофакторный дисперсионный анализ: при помощи критерия Краскела-Уоллиса была проверена гипотеза о равенстве медиан всех метрик для четырех групп текстов (оригинальные и упрощенные тремя разными моделями). Дополнительно был проведен корреляционный анализ для оценки взаимосвязи параметров для каждой из четырех групп текстов.

Таблица 2. *Пример упрощения предложений оригинального текста*

Оригинальный текст	Llama 3	mT5	MarianMT
Einem Dieb, der sich mit Reden mausig machen wollte, sagte jemand: “Was wollt Ihr? Ihr dürft ja gar nicht mehr in Eure Heimat zurückkehren und müsst froh sein, wenn man Euch hier duldet”	Ein Mann, der versucht war, sich als Betrüger auszugeben, wurde von jemandem angesprochen: “Was möchtest du? Du darfst gar nicht mehr nach Hause gehen und bist froh, wenn man dich hier toleriert”	Ein Kind, der sich mit Worten mögen wollte, sagte zu einem: “Was willt ihr? Ihr dürft ja gar nicht mehr in Ihrem Heimat zurückkehren und müsste froh, wenn man euch hier duldet”	Jemand sagte zu einem Dieb, der sich mit Reden mausen lassen wollte: “Was willst du? Ihr sollt nicht mehr in eure Heimat zurückkehren und froh sein, wenn ihr hier geduldet werdet”
“Meint Ihr?” sagte der Dieb; “meine Herren daheim haben mich so lieb, ich weiss gewiss, wenn ich heimkäme, sie liessen mich nimmer fort”	“Meinst du das?” fragte der Mann; “meine Familie hat mich sehr gern, ich bin sicher, dass sie mich nie wieder loslassen würde, wenn ich nach Hause käme”	“Meint ihr?” “Meine Herren in der Nähe haben mich so lieb, dass ich weiß, wenn ich heimkäme, sie liessen mich noch lange”	“Glaubst du?” sagte der Dieb: “Meine Herren zu Hause lieben mich so sehr, dass ich weiß, dass sie mich nie gehen lassen würden, wenn ich nach Hause kam”

Все тексты – и оригинальные, и упрощенные – были оценены с помощью метрик удобочитаемости и лексической сложности. Стандартные метрики для оценки моделей упрощения (SARI, SAMSA, ROUGE, BLEU и другие) не были использованы, так как они предполагают сравнение автоматически упрощенных текстов с эталонными упрощенными текстами, что на данный момент осталось за рамками настоящего исследования.

В качестве метрик удобочитаемости были использованы описанные выше индексы FRE и LIX, в то время как для оценки лексической плотности была найдена доля знаменательных слов по отношению ко всем словам в рассматриваемом тексте. Помимо этого, для оценки лексического разнообразия использовались уже упомянутые показатели TTR, RTTR и MTLD, вычисленные с помощью библиотеки LexicalRichness (lexicalrichness: A small Python…).

Результаты дисперсионного анализа продемонстрировали, что существуют статистически значимые различия для каждой из тринадцати метрик в четырех группах текстов ($p << 0,05$), поэтому дополнительно был проведен апостериорный тест Данна для попарных сравнений. Значения медиан для вычисленных коэффициентов разных групп показаны в Таблице 3 (дополнительно выделены те из них, которые отличаются от остальных).

Таблица 3. *Значения медиан метрик по группам*

	Llama	MarianMT	mT5	Original
flesch	**76,86**	73,26	73,12	72,04
lix	**31,54**	34,51	35,62	37,14
lexical_density	0,48	0,47	0,48	**0,49**
ttr	**0,56**	0,52	0,53	0,53
rttr	**7,75**	8,74	8,94	9,02
mtld	58,29	56,16	59,73	62,76
length	**196,00**	303,00	298,50	308,50
a1	**12,00**	**10,00**	7,50	8,00
a2	**12,00**	**10,00**	8,00	8,00
b1	**14,00**	**12,00**	11,00	11,00
other	**66,00**	**71,00**	77,00	76,00
new_other	32,00	32,00	20,50	0,00
duden_syn	**17,00**	**10,50**	**8,00**	0,00

Показатель *other* характеризует долю «сложных» слов, присутствующих в оригинальном и в упрощенном текстах, а показатель *new_other* – долю тех «сложных» слов, которые появились только в упрощенном тексте.

Для текстов, упрощенных с помощью моделей Llama 3 и MarianMT, в значениях были обнаружены отличия от всех остальных текстов по ряду метрик. По всем метрикам, кроме лексической плотности, MTLD и доли новых сложных слов, тексты, упрощенные моделью Llama 3, существенно отличались от остальных: они характеризовались большей удобочитаемостью и лексическим разнообразием, вместе с тем они были значительно короче и включали в себя большую долю простой лексики, а также большее количество корректно выбранных синонимов. Кроме того, более простой лексикой по сравнению с оригинальными текстами и текстами, упрощенными с помощью модели mT5, характеризовались тексты, упрощенные моделью MarianMT, в то время как лексическая плотность оригинальных источников была выше. В Таблице 4 представлен короткий рассказ “Der Rekrut” («Рекрут») Иоганна Петера Хебеля и сопоставленные ему упрощенные фрагменты.

В ряде случаев модели выполняют пересказ, а не упрощение. Это особенно характерно для модели mT5: она использует грамматические конструкции, отличные от конструкций оригинального текста, но оставляет оригинальную лексику. Кроме того, именно в текстах, упрощенных с помощью mT5, содержится больше всего ошибок, которые относятся к содержанию текста. Также необходимо отметить, что полученные тексты содержат грамматические и лексические ошибки. Так, например, модель Llama 3 заменяет оригинальное “ein Seemann” (моряк) на “ein Seeleute” (моряки, экипаж), а модель MarianMT использует глагол “atmen” (дышать) в значении “поймать”. Модель mT5 иногда оставляет оригинальный текст практически без изменений, несмотря на большое количество сложной лексики.

Таблица 4. *Пример оригинального и упрощенных текстов*

Оригинальный текст	Zum schwäbischen Kreiskontingent kam im Jahr 1795 ein Rekrut, so ein schöner, wohlgewachsener Mann war. Der Offizier fragte ihn, wie alt er sei. Der Rekrut antwortete: “Einundzwanzig Jahr. Ich bin ein ganzes Jahr lang krank gewesen, sonst wär' ich zweiundzwanzig”
Перевод	В 1795 году в швабский окружной воинский контингент прибыл новобранец – красивый статный человек. Офицер спросил его, сколько ему лет. Рекрут ответил: «Двадцать один год. Целый год я болел, а не то мне было бы двадцать два»
Llama 3	Ein junger Mann aus Schwaben trat im Jahr 1795 in das Kreiskontingent ein. Er war ein hübscher und kräftiger Mann. Ein Offizier fragte ihn nach seinem Alter. Der junge Mann antwortete: “Ich bin zwanzig Jahre alt. Ich war ein Jahr lang krank, sonst wäre ich einundzwanzig geworden”
mT5	Im Jahr 1795 wurde ein Rekrut auf dem schwäbischen Kreiskontingent, so ein schönes, wohlgewachsener Mann. Der Offizier fragte ihn, wie alt er sei. Die Rekrute antwortete: “Einhalbzig Jahre. Ich bin ein ganzer Jahr krank, sonst wäre ich zweiundzwanzig”
MarianMT	Im Jahre 1795 kam ein Rekrut in das Schwäbische Bezirkskontingent, so daß ein schöner, gut gewachsener Mann war. Der Offizier fragte ihn, wie alt er sei. Der Rekrut antwortete: “Einundzwanzig Jahre. Ich bin ein ganzes Jahr krank, oder ich wäre zweiundzwanzig”

Выполненный корреляционный анализ для значений рассмотренных выше метрик (в число анализируемых метрик для группы оригинальных текстов не входили показатели, которые описывали новые слова, – доля новых сложных слов и доля корректно выбранных синонимов) показал следующее, его результаты для каждой из четырех групп текстов представлены на Рисунках 1-4. Подтверждается наличие корреляции для всех четырех групп текстов между следующими парами: FRE – LIX, длина текста – TTR, длина текста – RTTR. Подтверждается отсутствие зависимости значений MTLD от длины текстов.

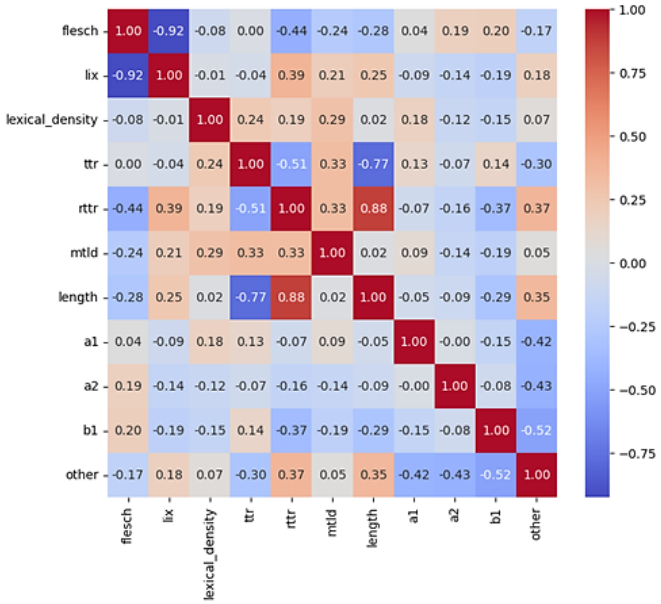

Рисунок 1. *Результаты корреляционного анализа для оригинальных текстов*

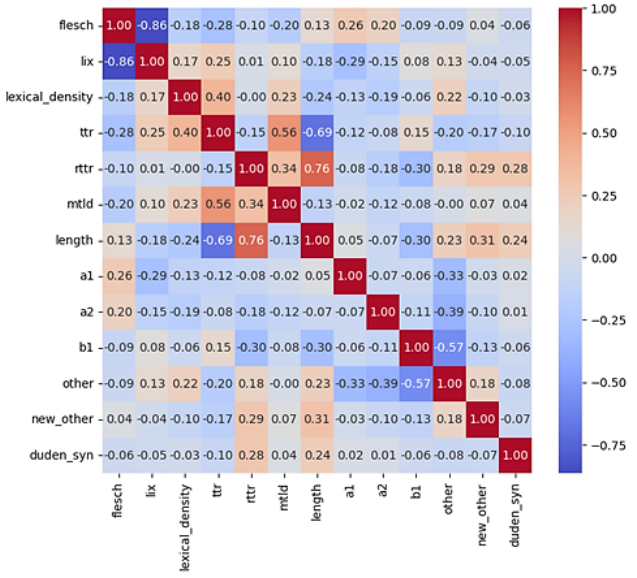

Рисунок 2. *Результаты корреляционного анализа для упрощенных текстов (модель Llama 3)*

Отметим, что для текстов, упрощенных моделью Llama 3, дополнительно отмечена умеренная положительная корреляция (0,56) между показателями MTLD и TTR (Рисунок 2), что может также косвенно подтверждать их большее лексическое разнообразие.

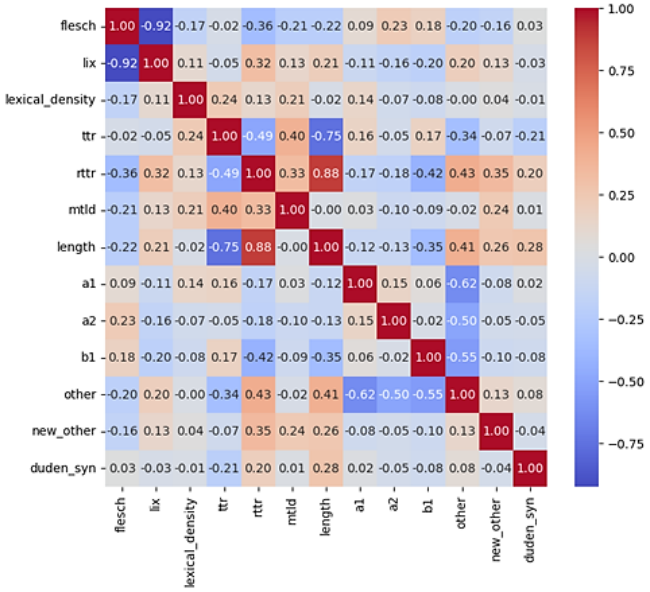

Рисунок 3. *Результаты корреляционного анализа для упрощенных текстов (модель mT5)*

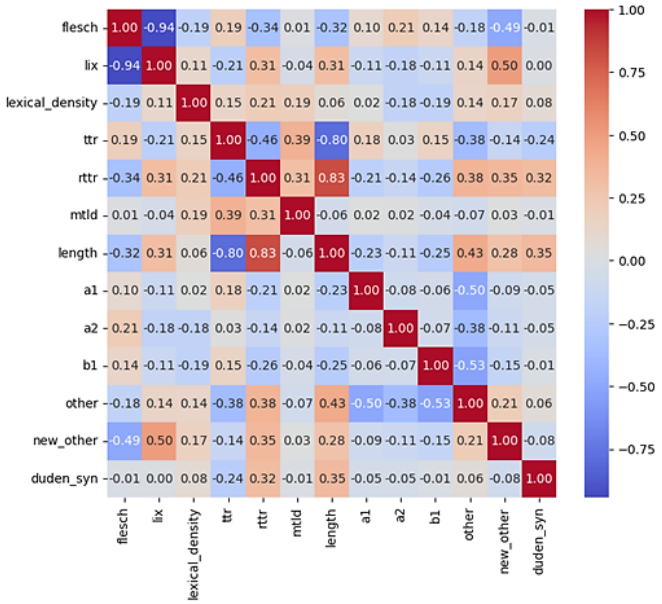

Рисунок 4. *Результаты корреляционного анализа для упрощенных текстов (модель MarianMT)*

Во всех группах наблюдалась сильная положительная корреляция (>0,75) метрики RTTR и длины текста. Отметим, что, несмотря на предварительно ограниченную длину текстов (<500 слов), метрики TTR и RTTR для всех групп коррелируют отрицательно, за исключением текстов, упрощенных моделью Llama 3 (в этом случае значение коэффициента корреляции статистически значимо не отличается от нуля). Метрика RTTR демонстрирует слабую положительную корреляцию с долей сложных слов, долей новых сложных слов и долей корректно подобранных синонимов и слабую отрицательную корреляцию с долей простых лексических единиц для уровней A1, A2, B1. Таким образом, ожидаемо, что большим лексическим разнообразием характеризуются тексты с большим количеством различных синонимов и сложных слов.

Заключение

В работе было апробировано применение больших языковых моделей для автоматического упрощения немецкоязычных художественных текстов. Полученные результаты позволили сформулировать следующие выводы. Во-первых, несмотря на активное использование БЯМ в задаче упрощения текстов, вопрос об оценке

качества получаемых результатов остается открытым, что подтвердило актуальность использования количественных метрик. Во-вторых, для проведения эксперимента был сформирован пилотный параллельный корпус немецкоязычных художественных текстов, что дало возможность сопоставления результатов и позволило провести сравнение между оригинальными и адаптированными фрагментами на лексическом и грамматическом уровнях. В-третьих, эксперимент по использованию БЯМ показал различия между моделями. Было обнаружено, что наиболее качественные результаты продемонстрировала модель Llama 3, которая по сравнению с моделями mT5 и MarianMT обеспечивает более последовательное упрощение художественных текстов согласно ряду показателей: она позволяет получить более короткие тексты, которые отличаются большей удобочитаемостью, менее сложной лексикой и большим числом корректных синонимических замен. Полученные данные позволяют предположить большую гибкость декодерной БЯМ в отличие от специализированных моделей, которые нацелены на решение отдельных задач обработки естественного языка. В-четвертых, оценка качества упрощения материала, выполненная при помощи тринадцати квантитативных показателей, позволила всесторонне охарактеризовать изменения данной группы текстов. В-пятых, корреляционный анализ метрик показал наличие взаимосвязей между исследуемыми характеристиками текстов. Было подтверждено, что метрики лексического разнообразия TTR и RTTR сильно зависят от длины текста, в то время как метрика MTLD демонстрирует устойчивость к изменению данного показателя. Также было установлено, что на общую сложность текста влияют его длина, удобочитаемость, доля уникальных слов, доля знаменательных слов, распределение простой и сложной лексики по уровням шкалы CEFR.

Перспективы дальнейшего исследования связаны, прежде всего, с расширением корпуса художественных произведений, которое подразумевает включение иных авторов, а также с привлечением других современных БЯМ. Представляется важным изучение влияния различных стратегий промптинга и параметров генерации на степень упрощения текста и на передачу его содержания.

Материалы исследования | Research materials

1. Duden. Das Wörterbuch der Synonyme: 100.000 Synonyme für Alltag und Beruf / Dudenredaktion. 3. Aufl. Berlin: Dudenverlag, 2016. https://www.duden.de/woerterbuch/synonyme
2. lexicalrichness: A small Python module to compute textual lexical richness // PyPI: сайт. https://pypi.org/project/lexicalrichness/
3. spaCy: Industrial-strength Natural Language Processing in Python: официальный сайт. https://spacy.io/

Источники | References

1. Большакова С. А. Система автоматической адаптации русскоязычных текстов и ее практическая значимость // Проблемы искусственного интеллекта. 2024. № 3. https://doi.org/10.24412/2413-7383-2024-3-45-54
2. Захарова Е. Ю., Савина О. Ю. Лексическое разнообразие текста и способы его измерения // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2020. Т. 6. № 1 (21). https://doi.org/10.21684/2411-197X-2020-6-1-20-34
3. Оборнева И. В. Автоматизация оценки качества восприятия текста // Вестник Московского городского педагогического университета. Серия: Информатика и информатизация образования. 2005. № 5.
4. Солнышкина М. И., Андреева М. И. Оценка сложности учебного текста по русскому как иностранному: перспективы и вызовы использования искусственного интеллекта // Русистика. 2026. Т. 24. № 1. https://doi.org/10.22363/2618-8163-2026-24-1-120-137
5. Солнышкина М. И., Андреева М. И., Гайнутдинова Д. З. Оценка сложности текстов по русскому как иностранному: инструментарий и алгоритм на основе больших языковых моделей // Актуальные проблемы филологии и педагогической лингвистики. 2026. № 2. https://doi.org/10.29025/2079-6021-2026-2-53-66
6. Дмитриева А. А., Лапошина А. Н., Лебедева М. Ю. Количественное исследование стратегий упрощения в адаптированных текстах для изучающих русский язык на уровне L2 // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (2021), Москва, 16-19 июня 2021 года. М.: Российский государственный гуманитарный университет, 2021. Вып. 20.
7. Alva-Manchego F., Martin L., Bordes A., Scarton C. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. http://doi.org/10.18653/v1/2020.acl-main.424
8. Alva-Manchego F., Scarton C., Specia L. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification // Computational Linguistics. 2021. Vol. 47. № 4. https://doi.org/10.1162/coli_a_00418
9. Amstad T. Wie verständlich sind unsere Zeitungen?: Diss. Zürich: Universität Zürich, 1978.
10. Björnsson C. H. Läsbarhet. Stockholm: Liber, 1968.
11. Carroll J., Minnen G., Canning Y., Devlin S., Tait J. Practical simplification of English newspaper text to assist aphasic readers // Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. Madison, 1998.
12. Chandrasekar R., Doran C., Srinivas B. Motivations and methods for text simplification // COLING 1996: Proceedings of the 16th International Conference on Computational Linguistics. 1996. Vol. 2. https://doi.org/10.3115/993268.993361

13. Ebling S., Battisti A., Kostrzewa M., Pfütze D., Rios A., Säuberli A., Spring N. Automatic Text Simplification for German // Frontiers in Communication. 2022. Vol. 7. Art. 706718. https://doi.org/10.3389/fcomm.2022.706718

14. Flesch R. A new readability yardstick // Journal of Applied Psychology. 1948. Vol. 32. № 3. https://doi.org/10.1037/h0057532

15. Hewett F., Asghari H., Stede M. Elaborative Simplification for German-Language Texts // Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Kyoto: Association for Computational Linguistics, 2024.

16. Papineni K., Roukos S., Ward T., Zhu W. Bleu: A method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), Philadelphia, Pennsylvania. Philadelphia: ACL, 2002. https://doi.org/10.3115/1073083.1073135

17. Qiang J., Li Y., Zhu Y., Yuan Y., Wu X. LSBert: A Simple Framework for Lexical Simplification. 2020. arXiv:2006.14939 [cs.CL]. https://doi.org/10.48550/arXiv.2006.14939

18. Seiffe L., Kallel F., Möller S., Naderi B., Roller R. Subjective Text Complexity Assessment for German // Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France. Marseille: European Language Resources Association, 2022.

19. Specia L. Translating from Complex to Simplified Sentences // Computational Processing of the Portuguese Language. PROPOR 2010. Lecture Notes in Computer Science. Berlin – Heidelberg: Springer, 2010. Vol. 6001. https://doi.org/10.1007/978-3-642-12320-7_5

20. Spring N., Rios A., Ebling S. Exploring German Multi-Level Text Simplification // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021. http://doi.org/10.26615/978-954-452-072-4_150

21. Sulem E., Abend O., Rappoport A. BLEU is Not Suitable for the Evaluation of Text Simplification. 2018. arXiv:1810.05995 [cs.CL]. https://doi.org/10.48550/arXiv.1810.05995

22. Truică C., Stan A., Apostol E. SimpLex: A lexical text simplification architecture // Neural Computing and Applications. 2023. Vol. 35. https://doi.org/10.1007/s00521-022-07905-y

23. Vaněček D., Kursch M., Şen E., Rehman I. U., Yorulmaz Y. I. Exploring how textual complexity affects cognitive load during reading: an eye-tracking study // Scientific Reports. 2026. Vol. 16. Art. 22446. https://doi.org/10.1038/s41598-026-51704-7

24. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. 2017. arXiv:1706.03762 [cs.RL]. https://arxiv.org/abs/1706.03762

25. Xu W., Napoles C., Pavlick E., Chen Q., Callison-Burch C. Optimizing statistical machine translation for text simplification // Transactions of the Association for Computational Linguistics. 2016. Vol. 4. https://doi.org/10.1162/tacl_a_00107

26. Zhang X., Lu X. Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels // Studies in Second Language Acquisition. 2025. Vol. 47. № 5. https://doi.org/10.1017/S0272263125101125

Информация об авторах | Author information

RU

Хохлова Мария Владимировна[1], к. филол. н.
Шер Анна Петровна[2]
[1, 2] Санкт-Петербургский государственный университет

EN

Maria Vladimirovna Khokhlova[1], PhD
Anna Petrovna Sher[2]
[1, 2] St. Petersburg State University

[1] *m.khokhlova@spbu.ru,* [2] *ap_sher@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 08.07.2026; опубликовано online (published online): 07.08.2026.

Ключевые слова (keywords): большие языковые модели; автоматическое упрощение текстов; сложность текстов; удобочитаемость текстов; немецкий язык; large language models; automatic text simplification; text complexity; readability; German language.