

DOI: 10.15514/ISPRAS-2026-38(6)-16



Dependence of Text Intrinsic Dimension on Language

^{1,2} M.A. Kuchina, ORCID: 0009-0008-2865-3050 <kuchina.ma@phystech.edu>² A.S. Kildyakov, ORCID: 0009-0005-9591-0983 <kildyakov@ipu.ru>¹ I.D. Papay, ORCID: 0009-0004-8323-1660 <papai.id@phystech.edu>¹ Moscow Institute of Physics and Technology,

9, Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russia.

² V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences,
65, Profsoyuznaya street, Moscow 117997, Russia.

Зависимость внутренней размерности текста от языка

^{1,2} М.А. Кучина, ORCID: 0009-0008-2865-3050 <kuchina.ma@phystech.edu>² А.С. Кильдяков, ORCID: 0009-0005-9591-0983 <kildyakov@ipu.ru>¹ И.Д. Папай, ORCID: 0009-0004-8323-1660 <papai.id@phystech.edu>¹ НИУ Московский физико-технический институт,

Россия, 141700, Московская обл., г. Долгопрудный, Институтский пер., д. 9.

² Институт проблем управления им. В.А. Трапезникова Российской академии наук,
Россия, 117997, г. Москва, ул. Профсоюзная, д. 65.

Аннотация. В работе исследуется зависимость внутренней размерности текста от языка. Проверяется гипотеза о том, что внутренняя размерность инвариантна относительно перевода. Исследование проведено на мультязычном корпусе текстов, состоящем из текстов одинакового содержания, переведённых на несколько языков. Для каждого текста на основе скрытых представлений токенов, полученных с помощью языковой модели, вычислена внутренняя размерность, после чего проведён сравнительный анализ полученных значений. В результате анализа гипотеза об инвариантности внутренней размерности при переводе отвергается; подтверждается, что внутренняя размерность зависит от языка. Показано, что синтетические языки с богатой морфологией дают более высокие значения, чем аналитические.

Ключевые слова: внутренняя размерность; инвариантность перевода; скрытые представления текстов; обработка естественного языка.

Для цитирования: Кучина М.А., Кильдяков А.С., Папай И.Д. Зависимость внутренней размерности текста от языка. Труды ИСП РАН, 2026, том 38, вып. 6, часть 1, стр. 243–258. DOI: 10.15514/ISPRAS-2026-38(6)-16.

Благодарности: Авторы работы выражают благодарность Андрею Валериевичу Грабовому (Институт проблем управления им. В.А. Трапезникова Российской академии наук, НИУ Московский физико-технический институт), эксперту курса «Автоматизация научных исследований» МФТИ, за постоянное внимание к настоящей работе и высказанные замечания, безусловно повысившие качество статьи.

Abstract. This work investigates the dependence of the intrinsic dimension of a text on its language. The hypothesis that the intrinsic dimension is invariant under translation is tested. The study is conducted on a multilingual corpus consisting of texts with the same content, translated into several languages. For each text, the intrinsic dimension is computed based on token embeddings obtained from a language model, followed by a comparative analysis of the resulting values. The analysis rejects the hypothesis of invariance under translation, confirming that the intrinsic dimension depends on the language. It is shown that synthetic languages with rich morphology give higher values than analytic ones.

Keywords: intrinsic dimension; translation invariance; text embeddings; natural language processing.

For citation: Kuchina M.A., Kildyakov A.S., Papay I.D. Dependence of Text Intrinsic Dimension on Language. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 6, part 1, pp. 243-258 (in Russian). DOI: 10.15514/ISPRAS-2026-38(6)-16.

Acknowledgements. The authors gratefully acknowledge Andrey Valerievich Grabovoy (V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow Institute of Physics and Technology), an expert of the MIPT course «Automation of Scientific Research», for his continued attention to this work and for his valuable comments, which have undoubtedly improved the quality of this work.

1. Введение

Современные языковые модели преобразуют текст в многомерные векторные пространства, геометрические свойства которых активно изучаются [1, 2]. Одной из важных характеристик таких пространств является внутренняя размерность – мера, определяющая минимальное количество параметров, необходимых для описания набора точек в многомерном пространстве. Маленькое значение внутренней размерности говорит об однородности и связности данных, а высокая внутренняя размерность характерна для неоднородных и разрозненных данных [3].

Существуют различные подходы к оценке внутренней размерности: локальные (MLE, TwoNN), фрактальные (MADA) и глобальные (PCA) [4]. Для текстовых данных оказывается эффективным метод, основанный на устойчивой гомологии (persistent homology dimension, PHD) [5, 6]. В отличие от локальных методов эта оценка устойчива к шуму и более эффективна на выборках небольшого размера [7, 8].

Однако эти исследования сфокусированы на свойствах PHD в рамках одного языка. Известно, что мультязычные модели имеют различное качество решения задач для разных языков [9]. Также исследования показывают, что в мультязычных моделях скрытые векторные представления содержат как общие компоненты, так и специфичные для конкретного языка [10]. Уже на этапе токенизации тексты могут быть по-разному разбиты на токены в зависимости от морфологических свойств языка [11]. Всё это указывает на то, что языковые модели по-разному работают с различными языками, в связи с чем внутренняя размерность скрытых представлений текста тоже может изменяться.

В данной работе исследуется зависимость внутренней размерности текста от языка. Для оценки размерности используется метод PHD. Множество скрытых представлений токенов

текста рассматривается как облако точек, предположительно лежащих вблизи многообразия малой размерности. Размерность этого многообразия называется внутренней размерностью. Размерность PND оценивает сложность текста [7]. Ранее было показано, что размерность PND устойчива к перефразированию [8]. В нашей работе проверяется, сохраняется ли эта инвариантность при переводе текста на другой язык, и выявляются языковые факторы, влияющие на значение размерности PND. Подтверждение инвариантности относительно перевода открыло бы перспективы использования метода PND в прикладных кросс-языковых задачах. Например, мог бы быть построен поисковый индекс текста путем разбиения текста на фрагменты, вычисления внутренней размерности каждого фрагмента и формирования вектора из полученных значений. Построение такого индекса позволило бы сравнивать тексты, написанные на разных языках, без необходимости получения прямого перевода, что актуально для кросс-языковых задач, например, поиска плагиата [12], суммаризации [13] или поиска переводных заимствований [14].

Для проверки гипотезы об инвариантности PND относительно перевода использован набор данных Wiki-40B [15]. На выборке из 500 статей, переведённых на 18 языков, с помощью модели Qwen2.5-0.5B [16] получены скрытые представления токенов и вычислены размерности PND. Статистический анализ показывает, что для разных языков значения внутренней размерности коррелируют друг с другом, но статистически значимо отличаются. То есть внутренняя размерность не инвариантна относительно перевода. Обнаружена связь внутренней размерности с морфологическим типом языка: для синтетических языков (русский, немецкий, польский и др.) среднее значение размерности PND выше, чем для аналитических (английский, французский, испанский и др.). Анализ токенизации показывает, что в синтетических языках слова разбиваются BPE на большее число токенов, что расширяет пространство скрытых представлений и увеличивает оценку размерности. Полученные результаты воспроизводятся на дополнительных моделях XLM-RoBERTa-base [17] и Gemma-4-E2B [18], что свидетельствует о независимости полученных результатов от выбранной архитектуры.

2. Обзор литературы

Оценка внутренней размерности данных состоит в определении минимального числа параметров, необходимых для описания данных в пространстве признаков. В работе [4] предложена классификация методов оценки внутренней размерности. Среди них выделяют методы, основанные на оценке максимального правдоподобия (Maximum Likelihood Estimation, MLE), методы, использующие расстояния до ближайших соседей, а также фрактальные подходы. Недавние исследования [19, 20] систематизируют современные методы по типу используемой геометрической информации.

Одним из подходов являются методы, основанные на топологическом анализе данных. К таким методам относится размерность PND [5, 6]. В работе [7] предложен эффективный алгоритм вычисления PND через минимальное остоное дерево и установлена связь PND с обобщающей способностью нейронных сетей. PND-размерность уже активно применяется для анализа текстов. Так, например, работе [8] размерность PND применена для детекции ИИ-сгенерированных текстов. Авторами показано, что размерность PND различается для ИИ-сгенерированных и написанных человеком текстов, при этом ее значение инвариантно к перефразированию, смене модели и предметной области. В работе [21] исследовано влияние жанра текста на PND: показано, что тексты с научным содержанием имеют меньшую внутреннюю размерность, чем художественные тексты. Для коротких текстов предложена модификация Short-PND [22]. Однако вопрос о том, влияет ли язык, на котором написан текст, на PND, остаётся малоизученным. Инвариантность PND при переводе могла бы стать основой для развития новых методов решения различных прикладных кросс-языковых задач.

Чтобы понять, может ли размерность PND зависеть от языка, необходимо понимать, содержится ли информация о языке в скрытых представлениях больших языковых моделей. В работе [10] показано, что скрытые представления мультязычных моделей содержат как общие, так и специфичные для конкретного языка компоненты. В работе [23] показано, что в больших языковых моделях информация о языке кодируется в небольшом наборе «разреженных» размерностей, и предложен способ выявления этих размерностей без дообучения. В работе [24] также показано, что в больших языковых моделях языки различаются не по внешним признакам написания, а по скрытым представлениям. Эти результаты демонстрируют, что скрытые представления текстов в мультязычных моделях содержат информацию о языке, что может приводить к зависимости внутренней размерности от языка.

Скрытые представления текста и, следовательно, их внутренняя размерность зависят также от токенизации текста. Исследования показывают, что производительность мультязычных моделей зависит от используемого токенизатора [25]. Также показано, что токенизация влияет на качество работы языковых моделей по-разному в зависимости от морфологической сложности языка [26]. Исследования показывают, что для малоресурсных языков качество токенизации и машинного перевода может отличаться [27, 28]. В работе [11] рассмотрен алгоритм токенизации BPE (byte-pair encoding), и показано, что в этом случае токенизация по-разному работает для языков разных морфологических типов. В [29] показано, что токенизаторы, учитывающие морфологию, дают более устойчивое обобщение. Таким образом, можно ожидать, что внутренняя размерность текста, вычисленная по скрытым представлениям токенов, будет систематически отличаться для языков в зависимости от их морфологии.

3. Постановка задачи

Пусть задана коллекция текстов X на языках из конечного множества A на темы из конечного множества T . Пусть $x_{t,a}$ – текст на языке $a \in A$ на тему $t \in T$. Под темой текста будем подразумевать его содержание: тексты x_{t,a_1} и x_{t,a_2} являются переводами друг друга для любых $a_1, a_2 \in A$. Тогда вся коллекция текстов:

$$D = \{x_{t,a} | t \in T, a \in A\}.$$

Каждому тексту $x_{t,a}$ поставлено в соответствие число $\rho(x_{t,a}) \in \mathbb{R}$ – внутренняя размерность текста. Обозначим среднее значение внутренней размерности для языка a :

$$\overline{\rho_a} = \frac{1}{|T|} \sum_{t \in T} \rho(x_{t,a_i}).$$

Гипотеза H_0 : $\overline{\rho_{a_1}} = \overline{\rho_{a_2}} = \dots = \overline{\rho_{a_{|A|}}}$ – средние значения внутренней размерности одинаковы для всех языков.

Цель работы – проверить гипотезу H_0 . В случае её отвержения определить, для каких языков различия статистически значимы, а также проанализировать возможные причины этих различий.

4. Метод

Пусть $x_{t,a}$ – текст на языке a на тему t . После токенизации текст представим в виде последовательности токенов:

$$x_{t,a} = [w_{t,a}^1, w_{t,a}^2, \dots, w_{t,a}^{N_{t,a}}],$$

где $N_{t,a}$ – длина текста $x_{t,a}$ в токенах. С помощью языковой модели M для каждого токена $w_{t,a}^i$ получим его скрытое представление:

$$e_{t,a}^i = M(w_{t,a}^i) \in \mathbb{R}^m.$$

Таким образом, весь текст $x_{t,a}$ отображается в множество точек в пространстве $\mathbb{R}^m$:

$$E_{t,a}=\{e_{t,a}^1,\dots,e_{t,a}^{N_{t,a}}\}\subset\mathbb{R}^m.$$

Предполагается, что точки множества $E_{t,a}$ лежат вблизи некоторого гладкого многообразия малой размерности d , вложенного в $\mathbb{R}^m$. Значение размерности гладкого многообразия d может быть интерпретировано как число степеней свободы, определяющее лексическое разнообразие и смысловую сложность текста: чем более разнообразна лексика и чем больше смысловых конструкций используется, тем выше размерность многообразия.

Оценка внутренней размерности PHD основана на понятии устойчивой гомологии. Используется следующий подход к изучению формы облака точек. Рассматривается порог r – радиус шара, описанного вокруг каждой точки, который последовательно увеличивается. При увеличении r появляются и исчезают топологические признаки разной размерности: признаки размерности 0 – связные компоненты, размерности 1 – циклы, размерности 2 – туннели и т.д. Набор i -мерных признаков – устойчивая гомология PH_i . Для каждого такого признака $\gamma \in PH_i$ фиксируется момент его рождения t_{birth} и момент смерти t_{death} . Момент рождения – это наименьший порог, при котором признак появляется, а момент смерти – порог, при котором он исчезает. Разность $l(\gamma)=t_{death}-t_{birth}$ называется временем жизни признака. Оценка PHD основана на анализе полученных времён жизни для разных признаков. Важный частный случай возникает при рассмотрении только 0-мерных признаков – компонент связности. В этом случае оценка PHD может быть выражена через минимальное остовное дерево (англ. Minimum Spanning Tree, MST) – связный граф без циклов, соединяющий все точки с минимальной суммой длин рёбер. В этом случае длины рёбер MST будут соответствовать временам жизни компонент связности.

Рассмотрим конечное множество точек. Построим на этом множестве минимальное остовное дерево. Для параметра $\alpha>0$ определим взвешенную сумму длин его рёбер:

$$S_\alpha=\sum_{l\in MST}|l|^\alpha.$$

Известно, что если точки распределены случайно на d -мерном многообразии, то с ростом числа точек n выполняется

$$\log E_\alpha(X_n)\approx(1-\frac{\alpha}{d})\log n+\text{const},$$

Оценивая коэффициент наклона линейной регрессии $\log S_\alpha$ от $\log n$, получаем искомую размерность d . Эту оценку будем называть размерностью PHD. Для каждого текста вычислим внутреннюю размерность и обозначим полученное значение $\rho(x_{t,a})$. Полученные числовые характеристики используем в экспериментах для проведения статистического анализа и проверки гипотезы H_0 о равенстве средних значений $\bar{\rho}_a$ по языкам.

5. Эксперименты

5.1 Формирование мультязычной выборки

В качестве исходного корпуса текстов использован набор данных Wiki-40B, содержащий статьи из Википедии на 40 языках. Для каждого языка статьи представлены в виде текстов со служебной разметкой (специальные токены ``_START_ARTICLE``, ``_START_SECTION``, ``_START_PARAGRAPH``, ``_NEWLINE``). Перед вычислением PHD эта разметка была удалена из каждого текста.

Для проведения эксперимента выбрано 18 языков. Полный список языков представлен в табл.1.

Из набора данных Wiki-40B было отобрано 500 английских статей с различными идентификаторами `wikidata_id`. Для каждой из этих статей получены машинные переводы на остальные 17 языков. Таким образом, итоговая коллекция составляет $500\times 18=9\,000$ текстов, где все тексты являются переводами единого исходного текста на разные языки.

Табл. 1. Список языков, использованных в эксперименте.
Table 1. Languages used in the experiment.

Язык	Код	Язык	Код
Английский	en	Французский	fr
Немецкий	de	Испанский	es
Нидерландский	nl	Итальянский	it
Шведский	sv	Португальский	pt
Норвежский	no	Каталанский	ca
Датский	da	Румынский	ro
Русский	ru	Сербский	sr
Польский	pl	Чешский	cs
Украинский	uk	Болгарский	bg

Это позволяет минимизировать шум, связанный с возможными расхождениями в оригинальных статьях Википедии на разных языках. В качестве исходного языка для перевода был выбран английский, так как для него доступны наибольшие объёмы параллельных корпусов, что обеспечивает более высокое качество машинного перевода на другие языки по сравнению с переводом с любого другого языка-источника.

Для получения скрытых представлений токенов использована языковая модель Qwen2.5-0.5B. В качестве представления текста взяты выходы последнего скрытого слоя для каждого токена. Выбор Qwen в качестве основной модели обусловлен её поддержкой большого количества языков и открытой архитектурой. Для проверки воспроизводимости результатов на других архитектурах все ключевые эксперименты воспроизведены на моделях XLM-RoBERTa-base и Gemma-4-E2B.

5.2 Проверка инвариантности PHD относительно перефразирования

Ранее в работе [8] было показано, что оценка внутренней размерности инвариантна относительно перефразирования текста. В эксперименте воспроизведены эти результаты для нашего мультязычного корпуса и модели Qwen2.5-0.5B, чтобы подтвердить применимость метода в наших условиях и создать основу для дальнейшей проверки инвариантности PHD относительно перевода.

Для этого для каждой из 500 исходных английских статей был получен парафраз с помощью двух последовательных переводов: с английского на русский и затем обратно на английский. Для каждого из 500 исходных английских текстов и для каждого из 500 полученных парафразов вычислена оценка внутренней размерности (PHD): для каждого текста с помощью модели Qwen2.5-0.5B получены скрытые представления токенов (выход последнего скрытого слоя), после чего на множестве этих скрытых представлений вычислена внутренняя размерность PHD.

Для полученных наборов PHD для оригинальных текстов и парафразы были посчитаны коэффициенты корреляции Пирсона, Спирмена и Кендалла. Их значения составили 0,72 (Пирсон), 0,80 (Спирмен) и 0,62 (Кендалл). Во всех случаях $p\text{-value}\approx 0$. Визуализация сравнения представлена на рис. 1. Полученные результаты говорят о том, что значения PHD для оригинальных и перефразированных текстов сильно скоррелированы, и корреляция статистически значима.

Результаты эксперимента показали, что средняя абсолютная разность PHD между оригиналом и парафразом составила 1,3, что обусловлено наличием выбросов. Медианная

абсолютная разница составила 0,3. Так как медиана робастна к выбросам, она лучше отражает действительную величину расхождения между PHD оригинала и парафраз.

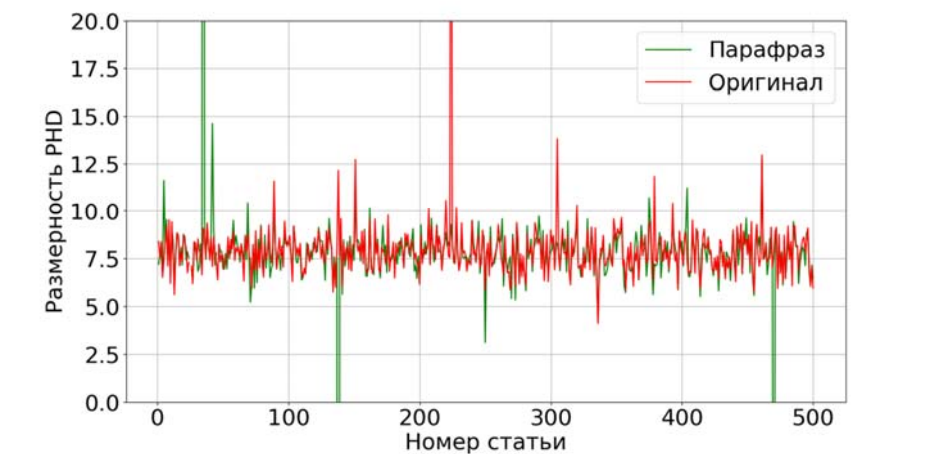


Рис. 1. Сравнение оценок PHD для оригинальных английских текстов и их парафразов. Графики для оригиналов и парафразов практически совпадают, что свидетельствует о близости значений PHD и высокой корреляции между ними.

Fig. 1. Comparison of PHD estimates for original English texts and their paraphrases. The plots for the originals and paraphrases almost coincide, indicating close PHD values and a high correlation between them.

5.3 Исследование зависимости PHD от языка

Для каждого из 500 текстов на 18 языках вычислено значение PHD. Основные статистики (среднее, медиана, стандартное отклонение, минимум, максимум, 25-й и 75-й процентиля, разброс значений) для каждого языка для модели Qwen2.5-0.5B приведены в табл. 2. Распределение PHD по языкам для всех трёх моделей визуализировано на рис. 2.

Согласно результатам эксперимента, значение PHD зависит от языка текста. Наблюдаемые различия визуально соответствуют делению языков на аналитические и синтетические в соответствии с их морфологическим типом. В аналитических языках грамматические значения (время, падеж, число и др.) выражаются преимущественно с помощью служебных слов и определённого порядка слов. Слова видоизменяются минимально. К аналитическим относятся: английский, французский, испанский, итальянский, португальский, каталанский, болгарский, нидерландский, шведский, датский, норвежский. В синтетических языках грамматические значения передаются с помощью приставок, суффиксов и окончаний. Это приводит к большому разнообразию форм одного и того же слова. К синтетическим относятся: немецкий, русский, украинский, польский, чешский, сербский, румынский. По результатам эксперимента для синтетических языков наблюдаются более высокие значения PHD, чем для аналитических. Статистически это предположение проверено в разделе 5.3.2.

5.3.1 Анализ корреляции PHD между языками

Для оценки степени сходства значений PHD для разных языков вычислены коэффициенты корреляции Пирсона, Спирмена и Кендалла для всех пар языков. Все корреляции статистически значимы (p-value < 0,05). Среднее значение коэффициента Пирсона по всем

парам составляет 0,59, Спирмена – 0,58, Кендалла – 0,41. Все попарные корреляции Пирсона представлены в табл. 3.

Табл. 2. Основные статистики распределения PHD для 18 языков. Видны существенные различия между языками, что говорит о влиянии языка на PHD.

Table 2. Summary statistics of PHD values distributions for 18 languages.

Язык	Среднее	Медиана	Ст. откл.	Мин.	Мак.	25%	75%	Диапазон
en	8,26	8,25	0,77	6,0	11,21	7,75	8,77	5,21
de	9,16	9,17	0,96	6,2	12,26	8,46	9,76	6,06
nl	9,02	8,99	0,9	6,24	12,12	8,4	9,58	5,88
sv	8,8	8,82	0,82	6,14	11,52	8,21	9,33	5,37
fr	8,28	8,23	0,82	5,44	10,67	7,76	8,89	5,23
es	8,37	8,38	0,9	5,87	11,26	7,73	8,94	5,4
it	8,54	8,56	0,88	5,55	10,93	7,94	9,1	5,38
pt	8,47	8,43	0,9	5,72	10,86	7,88	9,06	5,15
ru	9,09	9,07	1,04	5,97	12,34	8,36	9,81	6,38
pl	9,12	9,13	0,98	6,22	12,77	8,44	9,72	6,55
uk	8,35	8,35	0,77	5,91	10,67	7,8	8,81	4,75
sr	7,86	7,8	0,69	5,65	10,03	7,4	8,33	4,38
no	8,79	8,75	0,8	6,09	11,4	8,27	9,34	5,31
da	8,78	8,74	0,8	5,8	11,79	8,25	9,29	6,0
ca	8,46	8,47	0,82	6,08	11,77	7,86	9,05	5,69
ro	9,15	9,14	0,93	6,42	12,03	8,54	9,78	5,61
cs	8,95	8,84	0,91	5,98	11,71	8,31	9,58	5,72
bg	8,38	8,34	0,77	6,3	10,49	7,87	8,88	4,19

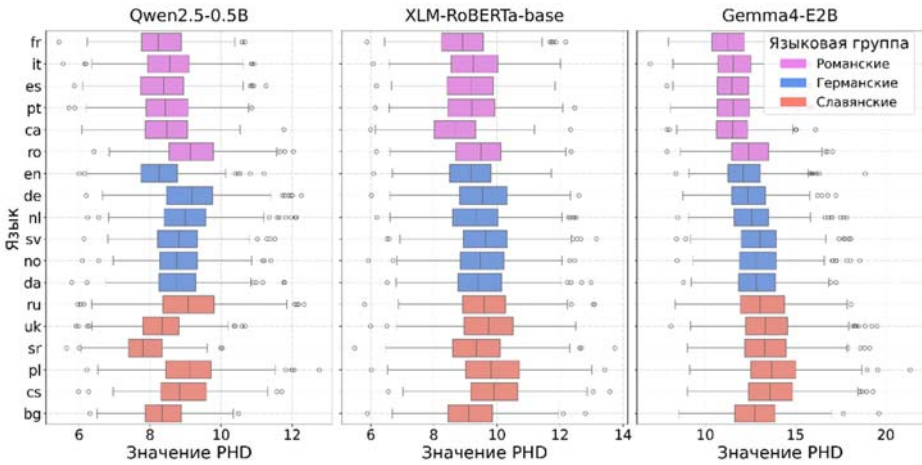


Рис. 2. Распределение значений PHD для 18 языков для трёх моделей: Qwen2.5-0.5B, XLM-RoBERTa-base и Gemma-4-E2B. Видны существенные различия: для одних языков (немецкого, русского, польского, ...) значения PHD в целом выше, чем для других (английского, французского, ...).

Fig. 2. Distribution of PHD values across 18 languages for three models: Qwen2.5-0.5B, XLM-RoBERTa-base, and Gemma-4-E2B. Substantial differences are observed: some languages (German, Russian, Polish, ...) exhibit generally higher PHD values than others (English, French, ...).

Табл. 3. Попарные коэффициенты корреляции Пирсона между языками.

Table 3. Pearson correlation coefficients between languages.

Язык	bg	ca	cs	da	de	en	es	fr	it	nl	no	pl	pt	ro	ru	sr	sv	uk
bg	1,00	0,55	0,59	0,48	0,55	0,34	0,55	0,54	0,54	0,50	0,51	0,60	0,55	0,56	0,64	0,66	0,51	0,64
ca	0,55	1,00	0,57	0,59	0,65	0,51	0,72	0,72	0,72	0,62	0,65	0,67	0,71	0,68	0,64	0,42	0,60	0,54
cs	0,59	0,57	1,00	0,53	0,60	0,42	0,58	0,55	0,60	0,60	0,58	0,67	0,56	0,55	0,58	0,47	0,54	0,57
da	0,48	0,59	0,53	1,00	0,64	0,51	0,62	0,61	0,60	0,64	0,68	0,59	0,60	0,62	0,61	0,44	0,66	0,47
de	0,55	0,65	0,60	0,64	1,00	0,56	0,72	0,67	0,71	0,67	0,66	0,68	0,69	0,64	0,65	0,41	0,65	0,51
en	0,34	0,51	0,42	0,51	0,56	1,00	0,59	0,61	0,55	0,54	0,50	0,51	0,60	0,51	0,48	0,28	0,50	0,34
es	0,55	0,72	0,58	0,62	0,72	0,59	1,00	0,79	0,78	0,64	0,66	0,69	0,81	0,62	0,68	0,42	0,61	0,54
fr	0,54	0,72	0,55	0,61	0,67	0,61	0,79	1,00	0,77	0,63	0,63	0,66	0,79	0,63	0,65	0,41	0,60	0,50
it	0,54	0,72	0,60	0,60	0,71	0,55	0,78	0,77	1,00	0,61	0,66	0,68	0,77	0,66	0,68	0,42	0,62	0,54
nl	0,50	0,62	0,60	0,64	0,67	0,54	0,64	0,63	0,61	1,00	0,62	0,64	0,64	0,61	0,58	0,39	0,62	0,46
no	0,51	0,65	0,58	0,68	0,66	0,50	0,66	0,63	0,66	0,62	1,00	0,62	0,65	0,66	0,62	0,44	0,71	0,50
pl	0,60	0,67	0,67	0,59	0,68	0,51	0,69	0,66	0,68	0,64	0,62	1,00	0,68	0,63	0,70	0,47	0,59	0,57
pt	0,55	0,71	0,56	0,60	0,69	0,60	0,81	0,79	0,77	0,64	0,65	0,68	1,00	0,65	0,70	0,42	0,60	0,52
ro	0,56	0,68	0,55	0,62	0,64	0,51	0,62	0,63	0,66	0,61	0,66	0,63	0,65	1,00	0,66	0,47	0,62	0,55
ru	0,64	0,64	0,58	0,61	0,65	0,48	0,68	0,65	0,68	0,58	0,62	0,70	0,70	0,66	1,00	0,55	0,59	0,63
sr	0,66	0,42	0,47	0,44	0,41	0,28	0,42	0,41	0,42	0,39	0,44	0,47	0,42	0,47	0,55	1,00	0,43	0,57
sv	0,51	0,60	0,54	0,66	0,65	0,50	0,61	0,60	0,62	0,62	0,71	0,59	0,60	0,62	0,59	0,43	1,00	0,49
uk	0,64	0,54	0,57	0,47	0,51	0,34	0,54	0,50	0,54	0,46	0,50	0,57	0,52	0,55	0,63	0,57	0,49	1,00

Из табл. 3 видно, что наибольшие значения коэффициента корреляции Пирсона наблюдаются для пар родственных языков романской группы: испанский-португальский (0,81), французский-португальский (0,79), французский-испанский (0,79), испанский-итальянский (0,78).

Наименьшие значения коэффициента корреляции Пирсона получены для пар, содержащих сербский язык: английский-сербский (0,28), нидерландский-сербский (0,39), немецкий-сербский (0,41), французский-сербский (0,41). Также низкая корреляция наблюдается для пар, состоящих из английского языка и славянских языков: английский-украинский (0,34) и английский-болгарский (0,34).

Для коэффициентов корреляции Спирмена и Кендалла результаты аналогичные: наиболее высокая корреляция наблюдается для языков романской группы, наиболее низкая – для пар английского и языков славянской группы, а также для пар, включающих сербский язык.

На дополнительных моделях XLM-RoBERTa-base и Gemma4-E2B аналогично вычислены попарные коэффициенты корреляции Пирсона, Спирмена и Кендалла между значениями PHD для всех пар языков. В табл. 4 приведены средние значения этих коэффициентов для каждой модели.

Табл. 4. Средние значения попарных коэффициентов корреляции PHD между языками для трёх языковых моделей.

Table 4. The average values of the pairwise correlation coefficients of PHD between languages for the three language models.

Модель	Пирсон	Спирмен	Кендалл
Qwen2.5-0.5B	0,59	0,58	0,41
XLM-RoBERTa-base	0,78	0,77	0,59
Gemma4-E2B	0,64	0,63	0,47

Для дополнительных моделей все попарные корреляции статистически значимы (p-value < 0,05) и положительны, что говорит о том, что вне зависимости от модели значения PHD для одного текста на разных языках связаны, но смещены под влиянием языкового фактора. Для всех трёх моделей повторяются закономерности: наиболее высокая корреляция наблюдается в парах родственных языков, наиболее низкая – в парах английского языка с языками славянской группы.

5.3.2 Статистический анализ различий PHD между языками

Для проверки наличия различий в значениях PHD между языками проведён дисперсионный анализ ANOVA для связанных выборок. Результаты анализа (F-Value = 237,06, p-value ≈ 0) показали, что гипотеза о равенстве среднего значения PHD для всех языков отвергается.

Для проверки наличия статистически значимых различий между парами языков проведён пост-хок критерий Неменьи с поправкой на множественные сравнения. По результатам теста статистически значимые различия наблюдаются для 117 из 153 пар языков, что соответствует 76,5%. Незначимые различия (p-value > 0,05) характерны для родственных языков: испанский-португальский, немецкий-нидерландский, русский-украинский, шведский-норвежский, а также для некоторых пар, включающих английский и романские языки (английский-французский, английский-испанский).

Для проверки наличия статистически значимой разницы между синтетическими и аналитическими языками для каждой статьи вычислено среднее PHD по аналитическим языкам и среднее PHD по синтетическим языкам. В силу отсутствия предположения о нормальности выборок использован парный критерий Уилкоксона. Результаты показали, что среднее значение PHD для синтетических языков статистически значимо выше, чем для

аналитических: $W = 16914$, $p\text{-value} \approx 0$. Среднее значение PHD для синтетических языков составляет 8,81, для аналитических – 8,56.

Для дополнительных моделей получены аналогичные результаты., Согласно ANOVA для связанных выборок, различия в PHD между языками статистически значимы: для XLM-RoBERTa-base $F\text{-Value} = 184,72$, $p\text{-value} \approx 0$, для Gemma4-E2B $F\text{-Value} = 348,19$, $p\text{-value} \approx 0$. Пост-хок критерий Неменки с поправкой на множественные сравнения показывает, что для XLM-RoBERTa-base число пар языков, для которых PHD статистически значимо различается, составляет 120 из 153 (78,4%), для Gemma4-E2B – 124 из 153 (81,1%). Эти значения близки к результатам, полученным для Qwen2.5-0.5B. Незначимые различия ($p\text{-value} > 0,05$), аналогично эксперименту с Qwen2.5-0.5B, наблюдаются в основном для родственных языков: испанский–португальский, немецкий–нидерландский, русский–украинский, шведский–норвежский, а также для некоторых пар, включающих английский и романские языки. На дополнительных моделях проверено наличие статистически значимой разницы между аналитическими и синтетическими языками. Для XLM-RoBERTa-base среднее значение PHD в группе аналитических языков составляет 9,22, в группе синтетических – 9,64 (критерий Уилкоксона: $W = 4161$, $p\text{-value} \approx 0$). Для Gemma-4-E2B значения составляют 12,19 и 13,21 соответственно ($W = 600$, $p\text{-value} \approx 0$). Во всех трёх моделях различия между синтетическими и аналитическими языками статистически значимы, что подтверждает устойчивость выявленной закономерности к выбору архитектуры.

5.4 Анализ влияния токенизации на значение PHD

Для объяснения наблюдаемых различий PHD между языками проведён анализ характеристик токенизации используемой модели Qwen2.5-0.5B. Эта модель использует алгоритм токенизации byte-BPE. Согласно алгоритму, при токенизации слова сначала разбиваются на отдельные байты, а затем в словарь итеративно добавляются наиболее часто встречающиеся пары токенов, уже находящихся в словаре.

Для каждого из 18 языков на основе тех же 500 переведённых текстов были вычислены следующие метрики:

- среднее число токенов на слово;
- среднее число символов в одном токене;
- доля коротких токенов (длиной до 3 символов);
- доля токенов средней длины (длиной 3-10 символов);
- доля длинных токенов (длиной более 10 символов).

Полученные значения представлены в табл. 5.

По данным из табл. 5 видно, что синтетические языки (русский, польский, украинский, сербский, чешский) имеют существенно более высокое среднее число токенов на слово (2,6-3,7) по сравнению с аналитическими (1,4–2,2). Напротив, среднее число символов в токене для синтетических языков ниже (2,0–2,8), чем для аналитических (3,0–4,5), что отражает более мелкое дробление. Среднее значение числа токенов на слово для аналитических языков составляет 2,00, для синтетических – 2,75; средняя длина токена – 3,30 и 2,61 соответственно. Согласно критерию Уилкоксона, различия статистически значимы: $W = 0$, $p\text{-value} \approx 0$ в обоих случаях.

Для дополнительных моделей XLM-RoBERTa-base и Gemma4-E2B также проведен анализ токенизации. В XLM-RoBERTa-base используется алгоритм SentencePiece с Unigram, в Gemma4-E2B – SentencePiece с BPE. Для этих моделей вычислены основные метрики: среднее число токенов на слово и среднее число символов в токене. В табл. 6 приведены средние значения метрик по аналитическим и синтетическим языкам для каждой модели.

Табл. 5. Метрики токенизации модели Qwen2.5-0.5B для 18 языков. Видна зависимость метрик от морфологии языка: языки с более богатым словоизменением разбиваются на большее число токенов меньшей длины.

Table 5. Tokenization metrics of the Qwen2.5-0.5B model for 18 languages. A dependence on language morphology is visible: languages with richer inflection are split into more tokens of shorter length.

Язык	Среднее число токенов на слово	Среднее число символов в токене	Доля коротких токенов	Доля токенов средней длины	Доля длинных токенов
en	1,393	4,48	25,8	70,6	3,6
es	1,707	3,61	29,3	70,0	0,7
fr	1,763	3,54	28,9	70,0	1,1
pt	1,808	3,45	33,6	65,9	0,5
it	1,92	3,39	31,5	68,2	0,3
ca	1,945	3,11	38,5	61,2	0,3
nl	2,023	3,29	32,2	67,6	0,2
de	2,079	3,47	30,0	69,5	0,5
no	2,136	3,07	36,0	63,8	0,2
ro	2,173	2,91	42,6	57,3	0,2
da	2,186	3,04	38,6	61,1	0,3
sv	2,208	3,06	37,4	62,3	0,3
pl	2,544	2,84	45,0	54,7	0,3
ru	2,601	2,82	22,5	68,6	8,9
bg	2,903	2,26	29,9	67,7	2,5
cs	2,936	2,3	59,4	40,6	0,1
sr	3,284	1,99	39,2	59,5	1,3
uk	3,66	1,96	41,2	57,5	1,3

Для всех моделей различия метрик в группах синтетических и аналитических языков статистически значимы, согласно критерию Уилкоксона: $W = 0$, $p\text{-value} \approx 0$ для всех моделей и обеих метрик. Таким образом, закономерность, полученная для Qwen2.5-0.5B, повторяется для других моделей: синтетические языки разбиваются на большее число более коротких токенов, чем аналитические.

Дополнительно вычислена корреляция между средним числом токенов на слово и средним значением PHD для всех 18 языков. Значение коэффициента корреляции Спирмена для XLM-RoBERTa-base составляет 0,77 ($p\text{-value} < 0,05$), для Gemma-4-E2B – 0,88 ($p\text{-value} < 0,05$). Полученные результаты подтверждают связь значения PHD с токенизацией: для языков, в которых слова разбиваются на большее число токенов, значение PHD выше.

Табл. 6. Сравнение средних значений метрик токенизации по группам синтетических и аналитических языков для трёх языковых моделей. Для каждой модели в группе синтетических языков среднее число токенов на слово выше, а средняя длина токена ниже, чем в группе аналитических языков.

Table 6. Comparison of mean tokenization metrics for three language models across analytic and synthetic language groups. For each model, the synthetic group exhibits a higher mean number of tokens per word and a lower mean token length compared to the analytic group.

Модель	Алгоритм токенизации	Среднее число токенов на слово (аналит.)	Среднее число токенов на слово (синт.)	Среднее число символов в токене (аналит.)	Среднее число символов в токене (синт.)
Qwen2.5-0.5B	byte-BPE	2,00	2,75	3,30	2,61
XLM-RoBERTa-base	SentencePiece (Unigram)	1,61	1,86	4,00	3,74
Gemma4-E2B	SentencePiece (BPE)	1,72	2,11	3,79	3,33

6. Заключение

В работе проведено исследование зависимости оценки внутренней размерности (PHD) текста от языка, на котором он написан. Эксперименты выполнены на выборке из 500 статей, переведённых на 18 языков, использована основная модель Qwen2.5-0.5B и дополнительные модели XLM-RoBERTa-base и Gemma-4-E2B. В результате экспериментов статистически подтверждена зависимость внутренней размерности не только от содержания текста, но и от языка, на котором он написан. Показано, что значения PHD для текстов одного и того же содержания, переведённых на разные языки, коррелируют между собой. Однако в отличие от перефразирования, внутренняя размерность не инвариантна относительно перевода. Полученные результаты демонстрируют связь PHD с морфологическим типом языка – аналитическим или синтетическим. Все результаты, полученные на основной модели, воспроизводятся на дополнительных, что говорит о независимости полученных результатов от выбора архитектуры языковой модели.

В синтетических языках грамматические значения (время, падеж, число и др.) выражаются преимущественно с помощью аффиксов (приставок, суффиксов, окончаний). В аналитических языках эти значения передаются служебными словами и порядком слов (например, в английском – вспомогательные глаголы was, have, going to и др.). Следствием этого является меньшее разнообразие форм одного и того же слова в аналитических языках. В процессе обучения токенизатора на аналитических языках слова быстрее «запоминаются» как целые единицы, поскольку они изменяются минимально. Таким образом, при токенизации одно слово в аналитических языках разбивается на меньшее число токенов.

В синтетических языках одно слово может иметь большое число различных форм, употребляющихся в зависимости от падежа, времени, числа. Это приводит к тому, что токенизатор не «запоминает» слово целиком, а выделяет его неизменяемые морфемы. В результате этого при токенизации слово разбивается на большее число токенов.

Увеличение числа токенов в синтетических языках расширяет пространство скрытых представлений текста. Вследствие этого увеличивается внутренняя размерность текста. Таким образом, для синтетических языков значение внутренней размерности получается больше, чем для аналитических. Данная закономерность повторяется на всех трёх использованных моделях независимо от алгоритма токенизации, что свидетельствует об устойчивости выявленной закономерности.

Список литературы / References

[1]. Lee A., Weber M., Viégas F. B., Wattenberg M. Shared Global and Local Geometry of Language Model Embeddings. arXiv preprint, 2025. DOI: 10.48550/arXiv.2503.21073.

[2]. Viswanathan K., Gardinazzi Y., Panerai G., Cazzaniga A., Biagetti M. The Geometry of Tokens in Internal Representations of Large Language Models. arXiv preprint, 2025. DOI: 10.48550/arXiv.2501.10573.

[3]. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, Cambridge, Massachusetts, 2016. 775 p.

[4]. Campadelli P., Casiraghi E., Ceruti C., Rozza A. Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. Mathematical Problems in Engineering, 2015, vol. 2015, Article ID 759367, 21 p. DOI: 10.1155/2015/759567.

[5]. Adams H., Aminian M., Farnell E., Kirby M., Mirth J., Neville R., Peterson C., Shonkwiler C. A Fractal Dimension for Measures via Persistent Homology. Topological Data Analysis: The Abel Symposium 2018. pp. 1-31. DOI: 10.1007/978-3-030-43408-3_1.

[6]. Schweinhart B. Persistent Homology and the Upper Box Dimension. Discrete & Computational Geometry, 2021, vol. 65, no. 2, pp. 331-364. DOI: 10.1007/s00454-019-00145-3.

[7]. Birdal T., Lou A., Guibas L. J., Simsekli U. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. Advances in Neural Information Processing Systems 34, 2021, pp. 6776-6789. DOI: 10.48550/arXiv.2111.13171.

[8]. Tulchinskii E., Kuznetsov K., Kushnareva L., Cherniavskii D., Nikolenko S., Burnaev E., Barannikov S., Piontkovskaya I. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. Advances in Neural Information Processing Systems 36, 2023, pp. 39257-39276. DOI: 10.48550/arXiv.2306.04723.

[9]. Shani C., Reif Y., Roll N., Jurafsky D., Shutova E. The Roots of Performance Disparity in Multilingual Language Models: Intrinsic Modeling Difficulty or Design Choices? arXiv preprint arXiv:2601.07220, 2026. DOI: 10.48550/arXiv.2601.07220.

[10]. Chang T., Tu Z., Bergen B. The Geometry of Multilingual Language Model Representations. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. DOI: 10.18653/v1/2022.emnlp-main.9.

[11]. Parra I. Morphological Typology in BPE Subword Productivity and Language Modeling. arXiv preprint, 2024. DOI: 10.48550/arXiv.2410.23656.

[12]. Bjelobaba S., Foltýnek T., Glendinning I., Krásničan V., Dlabolová D. H. Academic Integrity: Broadening Practices, Technologies, and the Role of Students. Ethics and Integrity in Educational Contexts, vol. 4. DOI: 10.1007/978-3-031-16976-2.

[13]. Grashchenkov K., Grabovoy A., Khabutdinov I. A Method of Multilingual Summarization For Scientific Documents. Ivannikov ISPRAS Open Conference (ISPRAS), 2022, pp. 24-30.

[14]. Кузнецова П. В., Бахтеев О. Ю., Чехович Ю. В. Методы обнаружения переводных заимствований в больших текстовых коллекциях. Информатика и ее применения, 2021. Т. 15, № 1. С. 30-41. DOI 10.14357/19922264210105. EDN BQFZAZ. / Kuznetsova, R. V., Bakhteev, O. Yu., Chekhovich, Yu. V. Methods for Detecting Translated Borrowings in Large Text Collections. Informatics and Applications, 2021, Vol. 15, No. 1, pp. 30–41. DOI: 10.14357/19922264210105. EDN: BQFZAZ. (in Russian).

[15]. Guo M., Dai Z., Vrandečić D., Al-Rfou R. Wiki-40B: Multilingual Language Model Dataset. In Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 2440-2452.

[16]. Yang A. et al. Qwen2.5 Technical Report. arXiv preprint, 2025. DOI: 10.48550/arXiv.2412.15115.

[17]. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440-8451.

[18]. El Abd S., Aggarwal V., Algayres R., et al. Gemma 4 Technical Report. arXiv preprint, 2026. DOI: 10.48550/arXiv.2607.02770.

[19]. Uchendu A. Le T. Unveiling Topological Structures in Text: A Comprehensive Survey of Topological Data Analysis Applications in NLP. arXiv preprint, 2024. DOI: 10.48550/arXiv.2411.10298v2.

[20]. Hopp M. D. S., Labatut V., Amalvy A., Dufour R., Stone H., Jach H., Murayama K. Persistent Homology of Topic Networks for the Prediction of Reader Curiosity. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 28121-28132. DOI: 10.18653/v1/2025.acl-long.1364.

[21]. Pedashenko V., Kushnareva L., Khassan Nibal Y., Tulchinskii E., Kuznetsov K., Zharchinskii V., Maximov Y., Piontkovskaya I. Unveiling Intrinsic Dimension of Texts: from Academic Abstract to Creative

- Story. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 7916-7944. DOI: 10.48550/arXiv.2511.15210.
- [22]. Wei D., Mao M., Li Z., Zhang Y. Short-PHD: Detecting Short LLM-generated Text with Topological Data Analysis After Off-topic Content Insertion. arXiv preprint, 2025. DOI: 10.48550/arXiv.2504.02873.
- [23]. Zhong C., Cheng F., Liu Q., Murawaki Y., Chu C., Kurohashi S. Language Lives in Sparse Dimensions: Toward Interpretable and Efficient Multilingual Control for Large Language Models. arXiv preprint, 2025. DOI: 10.48550/arXiv.2510.07213.
- [24]. Kim J., Lee S. How Language Directions Align with Token Geometry in Multilingual LLMs. arXiv preprint, 2025. DOI: 10.48550/arXiv.2511.16693.
- [25]. Rust P., Pfeiffer J., Vulić I., Ruder S., Gurevych I. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 3118-3135. DOI: 10.18653/v1/2021.acl-long.243.
- [26]. Toraman C., Yılmaz E. H., Şahinç O., Özcelik O. Impact of Tokenization on Language Models: An Analysis for Turkish. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, pp. 1-21. DOI: 10.1145/3578707.
- [27]. Asvarov A., Grabovoy A. Neural Machine Translation System for Lezgian, Russian and Azerbaijani Languages. 2024 Ivannikov Ispras Open Conference (ISPRAS), 2024, pp. 1-7. DOI: 10.1109/IS-PRAS64596.2024.10899143.
- [28]. Asvarov A., Grabovoy A. The Impact of Multilinguality and Tokenization on Statistical Machine Translation. 2024 35th Conference of Open Innovations Association (FRUCT), 2024, pp. 149-157. DOI: 10.23919/FRUCT61870.2024.10516416.
- [29]. Fusco A., Bianchessi M. L. P., Sgrizzi T., Zanollo A., Chesi C. Linguistic Units as Tokens: Intrinsic and Extrinsic Evaluation with BabyLM. In: Proceedings of the First BabyLM Workshop, Suzhou, China, November 2025, pp. 496-507. DOI: 10.18653/v1/2025.babylm-main.35.

Информация об авторах / Information about authors

Марина Александровна КУЧИНА – студентка 4 курса бакалавриата Физтех-школы прикладной математики и информатики Московского физико-технического института, кафедра Анализа данных, техник лаборатории «Интеллектуального анализа данных» Института проблем управления им. В.А. Трапезникова РАН с 2026 года. Сфера научных интересов: машинное обучение, искусственный интеллект, обработка естественных языков, статистический анализ скрытых представлений языковых моделей.

Marina Aleksandrovna KUCHINA – a 4th-year bachelor’s student at the Phystech School of Applied Mathematics and Computer Science, Moscow Institute of Physics and Technology, Department of Data Analysis, technician of the Laboratory of Data Mining at the V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences since 2026. Research Interest: machine learning, artificial intelligence, natural language processing, statistical analysis of hidden representations in language models.

Александр Сергеевич КИЛЬДЯКОВ – научный сотрудник лаборатории «Интеллектуального анализа данных» Института проблем управления им. В.А. Трапезникова РАН с 2025 года. Сфера научных интересов: машинное обучение, искусственный интеллект, обработка естественных языков, распознавание образов, высоконагруженные прикладные системы.

Alexandr Sergeevich KILDYAKOV – Scientist of the Laboratory of Data Mining at the V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences since 2025. Research Interest: machine learning, artificial intelligence, natural language processing, pattern recognition, high-load applications.

Иван Дмитриевич ПАПАЙ – студент 2 курса магистратуры Физтех-школы прикладной математики и информатики Московского физико-технического института, кафедра Интеллектуальных систем. Сфера научных интересов: вероятностные текстовые модели, причинный вывод, оптимальный транспорт.

Ivan Dmitrievich PAPAY – a 2nd-year master’s student at the Phystech School of Applied Mathematics and Computer Science, Moscow Institute of Physics and Technology, Department of Intelligent Systems. Research Interest: probabilistic text models, causal learning, optimal transport.