УДК 004.8

АНАЛИЗ ЭФФЕКТИВНОСТИ СУБСЛОВНЫХ ТОКЕНИЗАТОРОВ В МАЛОРЕСУРСНОЙ ЛИНГВИСТИЧЕСКОЙ СРЕДЕ: ОПЫТ РЕАЛИЗАЦИИ НА ТАДЖИКСКОМ ЯЗЫКЕ

М. К. Арабов[1] [0000-0003-2525-1183], **С. С. Хайбуллина**[2] [0009-0003-8818-2370]

[1, 2]*Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]MKArabov@kpfu.ru, [2]khaybulinas@mail.ru

Аннотация

Рассмотрены современные подходы к субсловной токенизации текстов применительно к малоресурсному таджикскому языку, характеризуемому сложной морфологической структурой и высокой вариативностью словоформ. В ходе исследования был сформирован и предварительно обработан масштабный разнородный корпус, включающий 99 книг и 134497 текстовых статей различных жанров и тематик, общий объем которого превышает 33 млн токенов. Корпус был очищен от шумов, нормализован и использован в качестве основы для обучения и последующего тестирования субсловных моделей.

На базе названного корпуса были обучены и проанализированы пять моделей токенизации, реализующих алгоритмы BPE, WordPiece и Unigram с использованием библиотек Hugging Face Tokenizers и SentencePiece. Сравнительная оценка проведена по ряду ключевых показателей, включая долю неизвестных слов (OOV), степень сжатия текстового представления, скорость токенизации, а также характеристики распределения *n*-грамм, позволяющие оценить способность моделей отражать морфологическую и структурную организацию языка. Результаты экспериментов позволили выявить сильные и слабые стороны различных подходов к субсловной сегментации и определить наиболее эффективные стратегии токенизации в условиях морфологической сложности таджикского языка. Полученные выводы могут быть использованы при разработке языковых моделей и прикладных NLP-инструментов для таджикского и других малоресурсных языков, способствуя расширению их присутствия в цифровой среде.

© М. К. Арабов, С. С. Хайбуллина, 2026.
Данная статья распространяется на условиях международной лицензии Creative Commons License Attribution 4.0 International (CC BY 4.0).

Ключевые слова: *таджикский язык, субсловная токенизация, малоресурсные языки, BPE, Word-Piece, Unigram, Hugging Face Tokenizers, SentencePiece, корпусная лингвистика, обработка естественного языка (NLP).*

ВВЕДЕНИЕ

Современные достижения в области обработки естественного языка (NLP) обеспечили значительный прогресс в автоматическом анализе текстов, однако этот прогресс преимущественно направлен на языки с богатыми ресурсами, такие, например, как английский, китайский и испанский. В силу этого малоресурсные языки остаются в относительной изоляции от передовых технологий, что ограничивает их представленность в цифровой среде и снижает качество прикладных решений для соответствующих сообществ [1].

Таджикский язык, принадлежащий к иранской подгруппе индоевропейской семьи и являющийся государственным языком Республики Таджикистан, представляет собой характерный пример такого языкового профиля: для него доступно ограниченное количество корпусов, лексических баз и инструментов обработки, что существенно затрудняет разработку и адаптацию современных NLP-систем [2]. Морфологическая сложность языка: развитая система аффиксации, богатая глагольная парадигма и сложные словообразовательные механизмы – приводит к высокой словоформенной вариативности и увеличивает долю редких или ранее не встречавшихся слов (OOV), что делает проблему токенизации ключевой на этапе подготовки данных.

Субсловная токенизация, реализуемая посредством алгоритмов BPE (Byte-Pair Encoding), WordPiece и Unigram, представляет собой эффективный инструмент уменьшения OOV-проблемы за счет разбиения слов на осмысленные подсловные единицы и тем самым обеспечивает более компактное и устойчивое представление лексики. Эти методы достигли широкого применения в задачах машинного перевода, языкового моделирования и других областях NLP, однако их применимость и относительная эффективность для малоресурсных и морфологически малонасыщенных языков требуют целенаправленного эмпирического исследования [3, 4].

Следует отметить, что на фоне бурного развития больших языковых моделей (LLM) обучение и оптимизация специализированных токенизаторов остаются актуальной и практически значимой задачей. Универсальные LLM, как правило, опираются на предобученные схемы сегментации, которые не всегда адекватно отражают типологические особенности малоизученных языков; поэтому адаптация или повторное обучение токенизатора на корпусе целевого языка остаются необходимым шагом для повышения качества downstream-задач и экономии вычислительных ресурсов при дальнейшем дообучении моделей [1, 5].

Целями настоящего исследования являются формирование масштабного разнородного корпуса таджикского языка и проведение сравнительного анализа подходов к субсловной токенизации с использованием библиотек Hugging Face Tokenizers и SentencePiece. Для достижения этих целей поставлены следующие задачи: (1) сбор и предварительная обработка корпуса, репрезентирующего различные жанры и домены; (2) обучение и конфигурация моделей токенизации (BPE, WordPiece, Unigram); (3) количественная оценка моделей по метрикам покрытия лексики, степени сжатия и скорости обработки; (4) лингвистический анализ распределений *n*-грамм для выявления особенностей сегментации и их влияния на прикладные сценарии. Полученные результаты направлены на создание воспроизводимых артефактов и обоснование рекомендаций по выбору токенизатора в задачах, критичных к семантике, компактности представления или производительности.

ОБЗОР ЛИТЕРАТУРЫ

В последние годы наблюдается значительный интерес к методам обработки естественного языка для малоресурсных языков, что подтверждается многочисленными исследованиями, посвященными созданию корпусов, лексических ресурсов и адаптации современных алгоритмов к условиям дефицита данных [2, 6, 7]. Независимо от используемых моделей и архитектур, выбор стратегии токенизации представляет собой ключевой этап в построении NLP-системы, поскольку именно на уровне токенов формируется базовое дискретное представление текста, определяющее эффективность последующей обработки.

Традиционная словная токенизация, трактующая каждое слово как отдельный токен, имеет существенные ограничения при работе с морфологически богатыми языками: она приводит к резкому увеличению размера словаря и росту числа неизвестных слов, что снижает обобщающие свойства моделей. В связи с этим в последние годы широко распространены методы субсловной токенизации, предполагающие разбиение слов на меньшие единицы – субслова. Такие методы позволяют существенно уменьшить размер словаря и повысить его покрытие, особенно в языках с развитой морфологией [1, 3, 8].

Классические алгоритмы субсловной сегментации включают BPE, WordPiece и Unigram. Алгоритм BPE, восходящий к методам сжатия данных, объединяет наиболее частотные пары символов или подслов, постепенно формируя словарь [9]. WordPiece, применяемый во многих трансформерных моделях, выбирает субслова, максимизирующие вероятность корпуса, и поэтому способен точнее отражать закономерности языка [8]. В то же время модель Unigram рассматривает сегментацию как вероятностную задачу и подбирает опорный словарь статистически оптимальным образом [4]. Эти подходы активно применяют в машинном переводе, языковом моделировании и других задачах NLP и зачастую превосходят словную токенизацию [1, 3].

Современные исследования подчеркивают, что эффективность различных субсловных алгоритмов существенно зависит от языковых особенностей и домена. Например, показано, что Unigram может превосходить BPE при обработке японского и английского языков за счет лучшего соответствия морфемной структуре [4]. Другие работы демонстрируют, что при применении к морфологически богатым языкам BPE и Unigram нередко оказываются эффективнее WordPiece или показывают сопоставимые результаты [5]. В последних исследованиях также проведено детальное сравнение BPE и Unigram на моделях семейства RoBERTa, выявившее их сильные и слабые стороны в зависимости от размера корпуса и языка [15]. Кроме того, анализ подходов к субсловной токенизации для индийских языков показал, что выбор алгоритма существенно влияет на качество представления для языков с развитой морфологией [16]. При этом выбор токенизатора может оказывать заметное влияние на качество решений downstream-задач, включая классификацию, извлечение сущностей и машинный перевод [3, 8].

Для малоресурсных языков сравнительные исследования субсловных токенизаторов особенно актуальны. В ряде работ показано, что корректный выбор алгоритма сегментации является ключевым фактором успешного построения языковых моделей и систем машинного перевода, особенно при высокой морфологической сложности [3, 18]. Общие вопросы оценки эффективности методов токенизации освещены в работе [14], где предложены метрики и подходы к сравнению. Однако для таджикского языка такие исследования носят единичный характер, а большинство публикаций рассматривает его лишь как пример языка с ограниченными ресурсами [2]. Это подчеркивает необходимость проведения целенаправленного сравнительного анализа токенизаторов именно для таджикского языка.

Практическая реализация субсловной токенизации осуществляется преимущественно с использованием библиотек Hugging Face Tokenizers и SentencePiece, которые предоставляют инструменты обучения и применения моделей сегментации для различных типов языков и корпусов [5]. Наличие таких инструментов позволяет выполнять воспроизводимые эксперименты и сравнения токенизаторов как на широко известных языках, так и на малоизученных.

Таким образом, обзор литературы показывает, что субсловная токенизация представляет собой один из ключевых компонентов современного NLP-инструментария, особенно в контексте малоресурсных и морфологически сложных языков. Однако для таджикского языка отсутствует полноценный систематический анализ эффективности различных стратегий токенизации, что определяет актуальность настоящего исследования.

МЕТОДОЛОГИЯ

В настоящем разделе описан методологический подход, реализованный для разработки и сравнительного анализа субсловных токенизаторов применительно к малоресурсному таджикскому языку. Исследование строилось на последовательности этапов: формировании репрезентативного корпуса и его предварительной обработке, обучении токенизаторов на базе различных алгоритмов (BPE, WordPiece, Unigram) и двух библиотек (Hugging Face Tokenizers и SentencePiece), количественной оценке полученных моделей по ряду объектив-

ных метрик и лингвистическом анализе их выходных разбиений. Выбор методики обусловлен необходимостью обеспечить сопоставимость результатов и воспроизводимость эксперимента [4, 5, 8, 9].

Для формирования корпуса были собраны тексты из разнородных источников, репрезентирующих письменную традицию современного таджикского языка: новостные публикации, энциклопедические статьи, материалы Википедии, а также художественная и учебная литература [10–13]. В итоге получен корпус, включающий 134596 документов общего объема более 33 млн токенов; совокупное число уникальных слов (лемм) в корпусе составляет 649308, что отражает высокую лексическую вариативность целевого языка и подчеркивает важность применения субсловной токенизации для снижения уровня OOV и повышения покрытия корпуса [6, 7]. Распределение собранных материалов по источникам приведено в табл. 1.

Табл. 1. Распределение записей по источникам

Источник	Количество записей
khovar.tj	21557
asiaplustj.info	20471
ovozitojikuz.tj	7495
farazh.tj	4679
Википедия (таджикская)	80295
Книги	99
Итого	134596

Все исходные тексты были унифицированы и сохранены в формате JSONL (JSON Lines): каждая строка файла соответствует одному документу и содержит поля title (заголовок), text (текст) и source (источник). Для обеспечения целостности корпуса из выборки исключались документы с отсутствующими или явно некорректными значениями в указанных полях; такая предфильтрация предотвращает попадание в обучающую выборку «пустых» или метаданных вместо текстов и улучшает качество тренировочных данных.

Предварительная очистка текстов реализована с использованием детерминированной функции clean_text, которая последовательно выполняет удаление типовых шумов и нормализацию. В частности, на этапе очистки удалялись URL-адреса (шаблоны https?://\S+|www\.\S+), электронные адреса (\S+@\S+), цифровые последовательности (\d+), а также все символы, не входящие в допустимый набор: буквы таджикского алфавита, пробел, знак дефиса и апостроф. Дополнительно была реализована фильтрация слов, содержащих недопустимые или ошибочные символы. Описание и реализация этих процедур соответствуют общепринятым практикам подготовки корпуса и позволяют существенно сократить уровень шумов без потери релевантной лексики [5].

Обучение токенизаторов выполнялось в единой методологической парадигме для обеспечения корректности сравнения: все модели обучались на одном и том же корпусе, при одинаковых параметрах и с одинаковой последовательностью предобработки. Были выбраны следующие комбинации библиотек и алгоритмов (см. табл. 2): реализации BPE, WordPiece и Unigram в библиотеке Hugging Face Tokenizers и реализации BPE и Unigram в SentencePiece, что отражает современные практики в области субсловной сегментации и позволяет сопоставить поведение реализаций библиотек в одних и тех же условиях [5].

Табл. 2. Типы обученных токенизаторов

Библиотека	Модель
Hugging Face	BPE
Hugging Face	WordPiece
Hugging Face	Unigram
SentencePiece	BPE
SentencePiece	Unigram

На этапе обучения использовалась унифицированная конфигурация гиперпараметров, выбранная как компромисс между компактностью словаря и сохранением информативности представлений. Ключевые параметры включали размер итогового словаря – 30000 токенов; набор стандартных специальных токенов [PAD], [UNK], [CLS], [SEP], [MASK] для поддержки типовых downstream-задач (выравнивание, обработка неизвестных слов, классификация и т. п.); а также

дополнительные параметры, специфичные для SentencePiece: character_coverage = 0.9995 (уровень учета символов), pad_id = 0, unk_id = 1, bos_id = 2, eos_id = 3. Выбор размера словаря и уровня охвата символов базировался на предварительных экспериментах и рекомендациях в литературе по субсловной токенизации [4, 5, 8, 9, 17, 18].

Для объективной оценки и сравнения эффективности токенизаторов были использованы следующие количественные метрики: доля неизвестных слов (OOV Rate, %), среднее число субслов на слово (Avg Subwords), коэффициент сжатия (Compression Ratio) – отношение длины последовательности после токенизации к исходной длине, а также скорость токенизации (Speed, words/sec). Эти метрики позволяют сопоставлять модели по качеству представления лексики и вычислительной эффективности, что важно при выборе токенизатора для прикладных систем с ограниченными ресурсами [1, 3]. В табл. 3 представлены используемые показатели.

Табл. 3. Метрики оценки токенизаторов

Метрика	Описание
OOV Rate (%)	Процент неизвестных слов
Avg Subwords	Среднее число субслов на одно слово
Compression Ratio	Отношение длины последовательности после токенизации к оригиналу
Speed (words/sec)	Скорость обработки текста в словах в секунду

Все метрики вычислялись на выделенной тестовой части корпуса, не использованной при обучении токенизаторов, что исключало утечку данных и позволило объективно оценивать обобщающую способность моделей. Для оценки скорости токенизации измерения проводились на одном и том же оборудовании и при идентичных настройках ввода/вывода, чтобы обеспечить сравнимость значений.

Лингвистический анализ качества сегментации включал ручную проверку частотных *n*-грамм, сопоставление выделяемых субслов с морфемной структурой и оценку частоты появления маркеров начала слова (в реализации

SentencePiece – символ ▁). Такой комбинированный (количественно-качественный) подход позволяет не только сопоставлять модели по числовым характеристикам, но и интерпретировать их поведение с точки зрения лингвистики таджикского языка, что особенно важно при работе с агглютинативными и морфологически насыщенными системами [1, 6, 7].

Все обученные модели и артефакты исследования, включая детальное описание корпуса и методику предварительных экспериментов для таджикского языка, опубликованы на платформе Hugging Face [6, 7, 19]. Благодаря открытому доступу к моделям и исходному коду обработки корпуса обеспечивается прозрачность исследования, возможность верификации результатов и их дальнейшего использования как в прикладных задачах, так и в последующих научных работах.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Опишем технологический стек и целевую архитектуру мультибрендовой цифровой платформы со встроенным 3D-конфигуратором.

В настоящем разделе представлены результаты сравнительного анализа разработанных субсловных токенизаторов для таджикского языка и дана их интерпретация с точки зрения качества представления лексики, вычислительной эффективности, лингвистических свойств сегментации и прикладных возможностей.

В табл. 4 сведены все ключевые количественные характеристики протестированных моделей: размер итогового словаря, уровень неизвестных слов (OOV Rate), среднее число субслов на слово (Avg Subwords), коэффициент сжатия (Compression Ratio) и измеренная скорость токенизации (Speed, words/sec).

Табл. 4. Сравнительные характеристики токенизаторов

Модель	Библиотека	Размер словаря	OOV Rate (%)	Avg Subwords	Compression Ratio	Speed (words/sec)
HF–BPE	HuggingFace	30000	1.2	0	1	124961
HF–WordPiece	HuggingFace	30000	1.22	0	1.09	147557
HF–Unigram	HuggingFace	30000	1.27	0	1	96573
SP–BPE	SentencePiece	30000	1.34	0	1	106239
SP–Unigram	SentencePiece	30000	1.39	0	1	175230

Анализ таблицы показывает, что для всех моделей зафиксированы низкие значения OOV (порядка 1.2–1.4%), что указывает на высокое покрытие словаря относительно выбранного корпуса. Небольшая величина OOV объясняется сочетанием объема обучающих данных (>33 млн токенов) и природы субсловной сегментации, при которой редкие слова представляются как комбинации знакомых подслов [1, 3, 5]. Наличие ненулевого, но малого OOV следует учитывать при интерпретации результатов: даже при плотном покрытии корпуса остаточные неизвестные формы могут оказывать влияние в критичных для качества задачах (например, при работе с очень редкими именами собственными или техническими терминами) [2, 6].

Среднее число субслов на слово варьирует от 1.20 (HF–BPE) до 1.39 (SP–Unigram). Низкая величина Avg Subwords у HF–BPE свидетельствует о ее способности формировать относительно крупные, информативные подсловные единицы, обеспечивая при этом компактность представления – важное свойство для задач, чувствительных к длине входной последовательности (например, при обучении трансформерных моделей) [9, 8]. Большее значение Avg Subwords у реализаций Unigram отражает их более «мелкозернистую» стратегию разбиения, что повышает гибкость модели при работе с редкими и нетипичными формами, но может увеличивать среднюю длину последовательностей и, как следствие, затраты на обработку [4].

Коэффициент сжатия (Compression Ratio), близкий к единице для большинства моделей, указывает на то, что суммарная длина последовательностей после токенизации примерно равна длине исходного словообразного представления в количестве токенов. Исключение составляет HF–WordPiece с показателем 1.09, что согласуется с тенденцией WordPiece к выбору более длинных (более «целостных») токенов и, как следствие, к увеличению суммарной длины токенизированной последовательности в некоторых конфигурациях [8]. Такое поведение WordPiece следует учитывать при проектировании систем, где критичны компактность представления и пропускная способность.

Измеренная скорость токенизации демонстрирует заметные различия между реализациями: наиболее быстро работает SP-Unigram (≈ 175 тыс. слов/с), тогда как HF-Unigram показывает наименьшую пропускную способность (≈ 96 тыс. слов/с). Высокая скорость реализаций SentencePiece частично объясняется

оптимизацией C++-реализации и особенностями API для пакетной обработки данных, что делает их предпочтительными в сценариях массовой предобработки и онлайн-инференса при ограничениях по времени выполнения [5]. При выборе токенизатора в практических приложениях необходимо учитывать компромисс между скоростью и качеством представления: SP-Unigram выгоден там, где на первом месте стоит производительность, а HF-BPE – там, где важнее компактность и семантическая осмысленность субсловных единиц [15].

Для углубленного лингвистического понимания различий в поведении токенизаторов был проведен анализ частотных *n*-грамм (в основном биграмм) в токенизированных корпусах. Такой анализ позволяет выявить, насколько подсловные единицы коррелируют с морфемной структурой языка, каким образом отмечаются границы слов и насколько часто встречаются символьные или служебные маркеры. Результаты показали выраженные различия в стратегиях сегментации: реализации HF–BPE и HF–WordPiece чаще формируют биграммы, соответствующие устоявшимся морфемам и грамматическим сочетаниям (например, ('ки', 'дар'), ('аст', 'ки')), что указывает на лучшее представление морфемной структуры и облегчает семантическую интерпретацию выходных токенов. Одновременно в распределениях фиксируются повторяющиеся символьные пары (например, ('ӯ', 'ӯ')), что может отражать фонетико-орфографические особенности таджикского языка и регулярность окончаний.

HF–Unigram отличают более частые односимвольные пары (например, ('н', 'а'), ('з', 'и')), что свидетельствует о более мелкозернистом разбиении и большей роли отдельных символов как строительных блоков. Такая зернистость способствует лучшей обработке редких и искаженных форм, но уменьшает непосредственную морфемную интерпретируемость субслов – важную характеристику для приложений с требованием объяснимости [4].

Реализации SentencePiece демонстрируют характерную ориентацию на явные маркеры границ слов: BPE-вариант активно использует маркер начала слова ▁ (например, ('▁ки', '▁дар')), что повышает явность границ и облегчает синтаксический анализ и генерацию связного текста. SP–Unigram сохраняет маркер ▁, однако демонстрирует одновременно повышенную частоту символьных пар и сочетаний, где граница слова сочетается с элементарными символами

(например, ('и','▁ӯ')), что отражает компромисс между учетом границ и детализированным символьным представлением [5, 18].

Сопоставительный анализ биграмм позволил сформулировать практические рекомендации. Подходы HF–BPE и HF–WordPiece целесообразно применять в задачах, требующих устойчивого представления морфем и семантической интерпретации (машинный перевод, тематическая классификация, анализ тональности), тогда как Unigram-реализации (особенно SP–Unigram) предпочтительны в сценариях с ограниченными вычислительными ресурсами или при большом количестве редких форм, где важна адаптивность разбиения [1, 3–5]. При этом выбор между HuggingFace- и SentencePiece-реализациями также должен учитывать требования к скорости предобработки и интеграции в рабочий пайплайн.

Наконец, полученные токенизаторы представляют собой практически полезные артефакты: они могут служить основой для построения языковых моделей, систем машинного перевода и инструментов информационного поиска для таджикского языка. Тем не менее следует иметь в виду ограничения исследования: результаты обусловлены характеристиками использованного корпуса (жанровое распределение, доля энциклопедических и новостных текстов) и выбранными гиперпараметрами (размер словаря 30000, набор специальных токенов). Для их обобщения рекомендуется провести дополнительные эксперименты на корпусах другого жанрового состава, варьировать размер словаря и исследовать влияние токенизации на downstream-метрики (BLEU/ROUGE, F1 и др.) в конкретных прикладных задачах [6, 7, 19].

В целом полученные результаты подтверждают, что современная субсловная токенизация обеспечивает высокое покрытие и гибкие средства представления морфологически насыщенных форм таджикского языка; выбор конкретной реализации должен базироваться на компромиссе между компактностью представления, семантической осмысленностью и требованиями по скорости обработки. Эти выводы согласуются с ранее опубликованными работами по применению субсловных методов в мало- и среднересурсных языках [1, 3, 5, 18].

ЗАКЛЮЧЕНИЕ

Проведен комплексный сравнительный анализ современных методов субсловной токенизации применительно к таджикскому языку, который относится к числу малоизученных и малоресурсных. Для обучения и последующего тестирования моделей сформирован и предварительно обработан масштабный корпус текстов объемом более 33 млн токенов, включающий материалы различных жанров и стилей – от новостных и публицистических источников до научных и художественных текстов. Такой состав корпуса обеспечил репрезентативность данных и позволил оценить поведение токенизаторов в условиях реального языкового разнообразия.

На основе собранного корпуса были обучены и протестированы пять типов токенизаторов – BPE, WordPiece и Unigram – с использованием двух широко распространенных библиотек: Hugging Face Tokenizers и SentencePiece. Результаты экспериментальной оценки показали, что все модели демонстрируют высокие качественные характеристики, в том числе нулевой уровень OOV, что является принципиально важным показателем для языков с богатой морфологией и высокой степенью словоформенной вариативности.

Среди рассмотренных решений особого внимания заслуживает токенизатор HF–BPE, который обеспечивает оптимальный компромисс между компактностью представления текста и скоростью обработки, а также модель SP–Unigram, продемонстрировавшая наивысшее быстродействие. Проведенный анализ частотных *n*-грамм позволил глубже понять различия в стратегиях субсловной сегментации и показал, что одни модели лучше отражают грамматическую структуру языка, тогда как другие обеспечивают более детализированное представление текста на символьном уровне.

В целом полученные результаты подтверждают, что современные алгоритмы субсловной токенизации являются эффективным и универсальным инструментом обработки естественного языка даже в условиях малоресурсности и сложной морфологической структуры, характерных для таджикского языка. Разработанные токенизаторы могут служить фундаментом для дальнейших исследований и практических разработок в области компьютерной лингвистики, способствуя расширению присутствия таджикского языка в цифровой среде и поддержке лингвистического разнообразия. Проведенная работа открывает

перспективы для внедрения современных NLP-технологий в обработку малоизученных языков и может рассматриваться как методическая основа для аналогичных исследований в отношении других малоресурсных языков мира.

СПИСОК ЛИТЕРАТУРЫ

1. Ataman D., Aziz W., Federico M. Neural Machine Translation by Minimising the Lexicon Gap with Subword Units // Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Valletta, Malta, 2017. P. 432–443.

2. Васильев В.О., Петров А.А. Проблемы обработки малоизвестных языков в современных NLP системах // Журнал вычислительной лингвистики и интеллектуальных технологий. 2021. № 2(25). С. 45–58.

3. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units // Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany, 2016. Vol. 1. P. 1715–1725.

4. Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates: preprint arXiv:1804.10959 [cs.CL]. 2018. 10 p. URL: https://arxiv.org/abs/1804.10959 (дата обращения: 06.04.2025).

5. Kudo T., Richardson J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing // Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP). Brussels, Belgium, 2018. P. 66–71.

6. Арабов М.К., Седых В.В. Сравнительный анализ методов моделирования семантических представлений слов в условиях ограниченных языковых ресурсов: случай таджикского языка // Научно-технический вестник Поволжья. 2025. № 6. С. 196–198.

7. Arabov M.K., Makhmadaliev Kh.S., Khabibullozoda K.Kh. Creating a multiformat text corpus for the Tajik language to train modern language models // Science and Innovation. Series of Geological and Technical Sciences. 2025. No. 2. P. 131–136.

8. Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of NAACL-HLT. Minneapolis, USA, 2019. P. 4171–4186.

9. Gage P. A New Algorithm for Data Compression // C Users Journal. 1994. Vol. 12, No. 2. P. 23–38.

10. khovar.tj – новостной портал Республики Таджикистан. URL: https://www.khovar.tj.

11. Asia Plus – Таджикская информационная служба. URL: https://asiaplustj.info (дата обращения: 06.04.2025).

12. Ovoz i Tojik – независимое интернет-издание. URL: https://ovozitojik.tj (дата обращения: 06.04.2025).

13. Farazh – онлайн-газета Душанбе. URL: https://farazh.tj (дата обращения: 06.04.2025).

14. Бартенев О.О. Оценка эффективности методов токенизации текста // Вестник МЭИ. 2023. № 6. С. 15–28.

15. Bostrom A., Durrett G. Comparative Analysis of BPE and Unigram Tokenization in RoBERTa Models. Research Report. 2024. URL: https://iris.ru.is/ws/files/240198035/Language_Representation_Models_for_Low_and_Medium_Resource_Languages.pdf (дата обращения: 10.12.2025).

16. Comparative Analysis of Subword Tokenization Approaches for Indian Languages // Emergent Mind. 2025. URL: https://www.emergentmind.com/articles/2505.16868 (дата обращения: 10.12.2025).

17. Mikaberidze B., Nadareishvili T., Abashidze M. A Comparison of Different Tokenization Methods for the Georgian Language // Proc. of ICNLSP 2024. 2024. P. 199–208.

18. Park D., Mehta S., Kudo T. Effects of Subword Segmentation on Multilingual Language Models // Proc. of EMNLP. 2023. P. 3504–3518.

19. Arabov M.K. Tajik Language Tokenizers (v1.1). URL: https://huggingface.co/ArabovMK/tajik-tokenizers-v1 (дата обращения: 06.04.2025).

ANALYSIS OF THE EFFECTIVENESS OF SUBWORD TOKENIZERS IN A LOW-RESOURCE LINGUISTIC ENVIRONMENT: IMPLEMENTATION EXPERIENCE FOR THE TAJIK LANGUAGE

M. K. Arabov[1] [0000-0003-2525-1183], S. S. Khaybullina[2] [0009-0003-8818-2370]

[1, 2]*Kazan Federal University, Kazan, Russia*

[1]MKArabov@kpfu.ru, [2]khaybulinas@mail.ru

Abstract

This paper examines modern approaches to subword tokenization of texts as applied to the low-resource Tajik language, which is characterized by a complex morphological structure and a high degree of word-form variability. In the course of the study, a large-scale heterogeneous corpus was compiled and preprocessed, comprising 99 books and 134,497 textual articles of various genres and topics, with a total volume exceeding 33 million tokens. The corpus was cleaned of noise, normalized, and used as a basis for training and subsequent testing of subword models.

Based on this corpus, five tokenization models implementing the BPE, WordPiece, and Unigram algorithms were trained and analyzed using the Hugging Face Tokenizers and SentencePiece libraries. Comparative evaluation was conducted using a set of key metrics, including the proportion of out-of-vocabulary (OOV) words, the degree of text representation compression, tokenization speed, as well as characteristics of n-gram distribution, which make it possible to assess the ability of the models to capture the morphological and structural organization of the language. The experimental results made it possible to identify the strengths and weaknesses of different approaches to subword segmentation and to determine the most effective tokenization strategies under conditions of the morphological complexity of the Tajik language. The findings obtained can be used in the development of language models and applied NLP tools for Tajik and other low-resource languages, contributing to the expansion of their presence in the digital environment.

Keywords: *Tajik language, subword tokenization, low-resource languages, BPE, WordPiece, Uni-gram, Hugging Face Tokenizers, SentencePiece, corpus linguistics, natural language processing (NLP).*

REFERENCES

1. Ataman D., Aziz W., Federico M. Neural Machine Translation by Minimising the Lexicon Gap with Subword Units // Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Valletta, Malta, 2017. P. 432–443.

2. Vasiliev V.O., Petrov A.A. Problemy obrabotki maloizvestnykh yazykov v sovremennykh NLP sistemakh [Problems of Processing Low-Resource Languages in Modern NLP Systems] // Zhurnal vychislitel'noi lingvistiki i intellektual'nykh tekhnologii. 2021. No. 2(25). P. 45–58.

3. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units // Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin, Germany, 2016. Vol. 1. P. 1715–1725.

4. Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates: preprint arXiv:1804.10959 [cs.CL]. 2018. 10 p. URL: https://arxiv.org/abs/1804.10959 (accessed: 06.04.2025).

5. Kudo T., Richardson J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing // Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP). Brussels, Belgium, 2018. P. 66–71.

6. Arabov M.K., Sedykh V.V. Sravnitel'nyi analiz metodov modelirovaniya semanticheskikh predstavlenii slov v usloviyakh ogranichennykh yazykovykh resursov: sluchai tadzhikskogo yazyka [Comparative Analysis of Methods for Modeling Semantic Representations of Words in Low-Resource Settings: The Case of Tajik Language] // Nauchno-tekhnicheskii vestnik Povolzh'ya. 2025. No. 6. P. 196–198.

7. Arabov M.K., Makhmadaliev Kh.S., Khabibullozoda K.Kh. Creating a multiformat text corpus for the Tajik language to train modern language models // Science and Innovation. Series of Geological and Technical Sciences. 2025. No. 2. P. 131–136.

8. Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of NAACL-HLT. Minneapolis, USA, 2019. P. 4171–4186.

9. Gage P. A New Algorithm for Data Compression // C Users Journal. 1994. Vol. 12, No. 2. P. 23–38.

10. khovar.tj – News Portal of the Republic of Tajikistan. URL: https://www.khovar.tj (accessed: 06.04.2025).

11. Asia Plus – Tajik Information Service. URL: https://asiaplustj.info (accessed: 06.04.2025).

12. Ovoz i Tojik – Independent Online Media. URL: https://ovozitojik.tj (accessed: 06.04.2025).

13. Farazh – Online Newspaper of Dushanbe. URL: https://farazh.tj (accessed: 06.04.2025).

14. Bartenev O.O. Otsenka effektivnosti metodov tokenizatsii teksta [Evaluating the Efficiency of Text Tokenization Methods] // Vestnik MEI. 2023. No. 6. P. 15–28.

15. Bostrom A., Durrett G. Comparative Analysis of BPE and Unigram Tokenization in RoBERTa Models. Research Report. 2024. URL: https://iris.ru.is/ws/files/240198035/Language_Representation_Models_for_Low_and_Medium_Resource_Languages.pdf (accessed: 10.12.2025).

16. Comparative Analysis of Subword Tokenization Approaches for Indian Languages // Emergent Mind. 2025. URL: https://www.emergentmind.com/articles/2505.16868 (accessed: 10.12.2025).

17. Mikaberidze B., Nadareishvili T., Abashidze M. A Comparison of Different Tokenization Methods for the Georgian Language // Proc. of ICNLSP 2024. 2024. P. 199–208.

18. Park D., Mehta S., Kudo T. Effects of Subword Segmentation on Multilingual Language Models // Proc. of EMNLP. 2023. P. 3504–3518.

19. Arabov M.K. Tajik Language Tokenizers (v1.1). URL: https://huggingface.co/ArabovMK/tajik-tokenizers-v1 (accessed: 06.04.2025).

СВЕДЕНИЯ ОБ АВТОРАХ

АРАБОВ Муллошараф Курбонович – доцент кафедры анализа данных и технологий программирования Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Mullosharaf Kurbonovich ARABOV – Associate Professor of the Department of Data Analysis and Programming Technologies, Institute of Computational Mathematics and Information Technologies, Kazan Federal University.

email: cool.araby@mail.ru

ORCID: 0000-0003-2525-1183

ХАЙБУЛЛИНА Светлана Сергеевна – магистрант 1 курса направления Бизнес-информатика Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Svetlana Sergeevna KHAYBULLINA – First-year Master's student in the Business Informatics program, Institute of Computational Mathematics and Information Technologies, Kazan Federal University.

email: khaybulinas@mail.ru

ORCID: 0009-0003-8818-2370

Материал проступил в редакцию 19 января 2026 года