RUDN Journal of Language Studies, Semiotics and Semantics
2313-2299 (Print), ISSN 2411-1236 (Online)
Вестник Российского университета дружбы народов.
Серия: ТЕОРИЯ ЯЗЫКА. СЕМИОТИКА. СЕМАНТИКА

2026 Vol. 17 No. 2 557–568
http://journals.rudn.ru/semiotics-semantics

DOI: 10.22363/2313-2299-2026-17-2-557-568
EDN: LSTNOS

Научная статья / Research Article

Токен как алгоритмическая «тень» морфемы

К.П. Жикулина

Российский университет дружбы народов, *Москва, Российская Федерация*
✉zhikulina-kp@rudn.ru

Аннотация. Актуальность исследования критического этапа обработки текста в эпоху доминирования больших языковых моделей обусловлена системным расхождением между статистически выделенными границами токенов и лингвистически обоснованными границами морфем в токенизаторах, что проявляется в работе с языками с богатой морфологией. Научная новизна исследования заключается в рассмотрении *токена* не как технической единицы, а как лингвистического артефакта — *алгоритмической «тени» морфемы*, утратившей семантическую мотивированность, но сохранившую внешнее сходство. Цель исследования — выявить лингвистические потери, возникающие при статистической сегментации слов и оценить степень расхождения между токенными и морфемными границами. Методы исследования: сопоставительный лингвистический анализ, корпусный анализ, критический дискурс-анализ и типологическое моделирование. В ходе исследований разработана и эмпирически обоснована типология морфологических искажений при BPE-токенизации, описывающая три типа искажений и расхождений между морфемными и токенными границами: фрагментация, контаминация и аберрация. На материале лексем подорожник и высокотехнологичный экспериментально доказана трехступенчатая деградация морфологической прозрачности при переходе от одноязычных к двуязычным и мультиязычным BPE-моделzv. Установлено, что мультиязычные токенизаторы полностью утрачивают морфологические свойства в отношении отдельного языка, множа статистические артефакты, которые не идентифицируются классическим набором единиц морфологии и носителями данного языка.

Ключевые слова: искусственный интеллект, компьютерная лингвистика, машинное обучение, BPE-токенизация, НЛП-системы, большая языковая модель, сублексемная сегментация, морфология, морфемизация, подслово

Финансирование: публикация подготовлена в рамках грантовой поддержки научных проектов РУДН 050227-0-000.

Конфликт интересов: автор заявил об отсутствии конфликтов интересов.

История статьи: дата поступления: 20.01.2026; дата приема в печать: 10.02.2026.

Для цитирования: *Жикулина К.П.* Токен как алгоритмическая «тень» морфемы // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2026. Т. 17. № 2. С. 557–568. https://doi.org/10.22363/2313-2299-2026-17-2-557-568 EDN LSTNOS

© Жикулина К.П., 2026
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License
https://creativecommons.org/licenses/by-nc/4.0/legalcode

Token as an Algorithmic ‘Shadow’ of a Morpheme

Kristina P. Zhikulina

RUDN University, *Moscow, Russian Federation*
✉zhikulina-kp@rudn.ru

Abstract. The relevance of the study of a critical stage of text processing in the era of the dominance of large language models is due to the systemic divergence between statistically identified token boundaries and linguistically justified morpheme boundaries in tokenizers, which is particularly evident when working with rich morphology languages. The scientific novelty of the study lies in considering the token not as technical part, but as a linguistic artifact – an algorithmic ‘shadow’ of a morpheme that has lost its semantic motivation but preserved its external resemblance. The aim of the research is to identify linguistic losses that occur during statistical segmentation of words and to assess the degree of divergence between token and morpheme boundaries. Research methods: comparative linguistic analysis, corpus analysis, critical discourse analysis and typological modeling. During the research, a typology of morphological distortions in BPE tokenization has been developed and empirically substantiated, describing three types of distortions and divergence between token and morpheme boundaries: fragmentation, contamination, and aberration. Using the lexemes ‘podorozhnik’ and ‘vysokotekhnologichny’ as material, the three-stage degradation of morphological transparency in the transition from monolingual to bilingual and multilingual BPE models has been experimentally proven. It has been established that multilingual tokenizers completely lose morphological properties in relation to a particular language multiplying statistical artifacts that are not identified by a classical set of morphological units and native speakers of this language.

Keywords: artificial intelligence, computational linguistics, machine learning, BPE-tokenization, NLP-systems, LLM, subword segmentation, morphology, morphemization

Founding: The publication has been supported by the RUDN University Scientific Projects Grant System, project No 050227-0-000.

Conflict of interests: The author stated that there were no conflicts of interest.

Article history: Received: 20.01.2026; Accepted: 10.02.2026.

For citation: Zhikulina, K.P. (2026). Token as an Algorithmic ‘Shadow’ of a Morpheme. *RUDN Journal of Language Studies, Semiotics and Semantics, 17*(2), 557–568. (In Russ.). https://doi.org/10.22363/2313-2299-2026-17-2-557-568 EDN LSTNOS

Введение

Стремительное развитие генеративных моделей искусственного интеллекта (далее — ИИ), способных имитировать мыслительные процессы человека, обучаться, «рассуждать», поставило перед лингвистикой принципиально новые вопросы. Если ранее язык рассматривался как система знаков, существующая независимо от способов ее фиксации и передачи, то в эпоху больших языковых моделей (LMM) язык все чаще предстает как объект алгоритмической обработки, подчиняющийся не только программным закономерностям, но и лингвистическим. Центральным звеном, опосредующим взаимодействие между человеческим текстом и LMM, становится *машинное обучение* (ML). Под *машинным обучением* понимается узкоспециализированная область, в основу которой положен принцип обучения машин точному

выполнению определенных задач с помощью выявления закономерностей и детального прогнозирования. Благодаря подходу, называемому *непрерывным обучением*, машина способна адаптироваться под разные задачи и обстоятельства, совершенствовать свои функции и выдавать более точные результаты.

Практически все современные нейросетевые системы — от ChatGPT до Llama — используют *сублексемную сегментацию,* при которой слова разбиваются на единицы меньше слова, но больше символа — *подслова*. Наиболее распространенным методом такой сегментации является Byte Pair Encoding (BPE), адаптированный для задач обработки естественного языка Р. Сенрихом с соавторами [1].

Токенизаторы, использующиеся в современных нейросетевых системах, все чаще демонстрируют системное расхождение между статистически выделенными границами токенов и лингвистически обоснованными границами морфем. Именно на этапе *токенизации* закладываются границы членения, которые впоследствии определяют, как LMM интерпретирует язык.

Это расхождение не является случайным и незначительным — оно носит системный характер и особенно проявляется в языках с богатой морфологией, таких как русский, немецкий, тюркские языки. Модель оперирует единицами, внешне напоминающими морфемы, но не наследуют их семантическую и структурную мотивированность. Возникает парадоксальная ситуация: инструмент, созданный для более эффективной обработки языка, порождает лингвистические артефакты, ставящие под сомнение саму возможность классического морфологического анализа в рамках архитектур нейронных сетей. Как справедливо отмечает О.И. Максименко, современный этап развития языкознания характеризуется сближением традиционного лингвистического описания с методами алгоритмической обработки текста, что требует пересмотра устоявшихся понятий в контексте их алгоритмической операционализации [2]. В настоящем исследовании такому пересмотру подвергается категория морфемы, которая в условиях *сублексемной сегментации* утрачивает статус минимально значимой единицы и замещается статистически выделенным токеном — алгоритмической «тенью».

Актуальность исследования обусловлена не только теоретической значимостью поставленной проблемы, но и ее практическими импликациями. Современные исследования в области NLP фокусируются на улучшении технических параметров токенизации — сжатии словарей, скорости обработки, фиксации и распознаванию редких слов (см., например, [3–6]), — игнорируя при этом *морфологическую прозрачность* как критерий качества. Ее утрата при алгоритмической обработке ведет не просто к техническим погрешностям, но и к лингвистической аберрации — искажению способа существования языка в цифровой среде. Проблемным местом оказываются современные *токенизаторы*, как правило, являющиеся мультиязычными: обучаются на разнородных текстах, но должны одинаково хорошо работать для сотен языков одновременно. Такой подход неизбежно приводит к *билингвально-*

ориентированному или *полилингвальному алгоритму членения*, при котором границы *токенов* определяются не морфемной структурой одного языка, а частотной статистикой, усредненной по всем языкам обучающего корпуса. Мнимая универсальность достигается ценой утраты лингвистической специфики: языки с богатой морфологией начинают сегментироваться по подобию языков с другой структурой — преимущественно английского языка, что порождает феномен *лексической интерференции*.

Таким образом, исследование ставит перед собой следующие задачи:

1) провести сравнительный анализ трех типов *токенизации* (однозначной, двуязычной и мультиязычной) на материале русских лексем *подорожник* и *высокотехнологичный* с прозрачной морфемной структурой;

2) разобрать классификацию типов морфологических искажений при BPE-сегментации, описав их лингвистическую природу и механизм возникновения;

3) предложить предварительную систему критериев для оценки морфологической прозрачности *токенизаторов*;

4) очертить границы применимости статистической сегментации для задач, требующих морфологического анализа.

Теоретическая значимость исследования заключается в пересмотре онтологического статуса *токена* как лингвистической категории — *алгоритмической «тени» морфемы*, внешне подобной морфеме, но лишенной семантической и структурной мотивированности. *Токен* впервые осмысляется как *псевдоморфемная единица* — алгоритмическая проекция морфемы, утратившая этимологическую, парадигматическую и словообразовательную мотивированность, но сохранившая подобие. Введенное понятие *алгоритмической «тени» морфемы* позволяет концептуализировать принципиальное различие между языковым наследием и цифровым компромиссом.

Практическая значимость исследования состоит в том, что его результаты позволяют разработчикам NLP-систем выявлять слабые места существующих *токенизаторов* и выбирать стратегию сегментации в зависимости от языка и задачи, предлагают критерии для проектирования лингвистически адекватных *токенизаторов* для низкоресурсных и морфологически богатых языков, а также служат дидактическим материалом для универсальных курсов по компьютерной лингвистике, искусственному интеллекту в филологии и алгоритмической лингвистике.

Обсуждение и результаты исследования

Обработка естественного языка (NLP) направлена на решение таких задач, как обучение компьютеров понимать, интерпретировать и генерировать живую речь. В компьютерной лингвистике NLP означает, прежде всего, обучение ИИ распознаванию закономерностей в человеческой речи, текстовых источниках и использованию этих же закономерностей для анализа пользовательского запроса, расшифровки и создания реакции, то есть ответа.

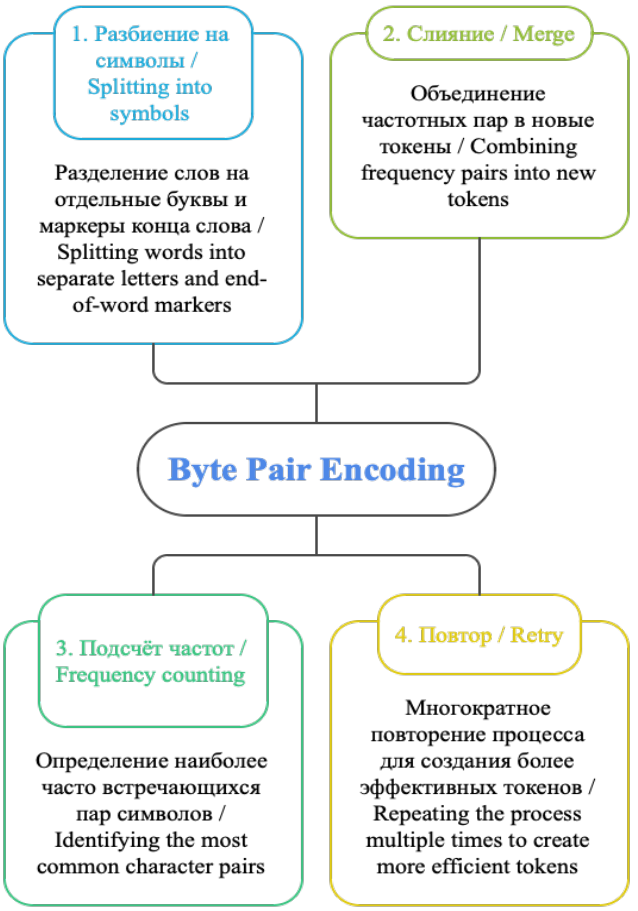
Одним из фундаментальных процессов в NLP является *токенизация*, при которой происходит преобразование исходного текстового материала в формат, воспринимаемый машиной для анализа и моделирования. Текст разбивается на фрагменты — маленькие единицы, называемые *токенами*. *Токенами* могут выступать как отдельные слова так и части слов или символы, в зависимости от требуемого уровня фрагментарности и детализации. Они помогают понимать контекст и разрабатывать модель для обработки естественного языка, называемую сегодня *большой языковой моделью (LLM)*.

В машинном обучении существует несколько типов *токенизации*. *Токены* могут различаться в зависимости от используемого метода *токенизации* и уровня лингвистической абстракции, под которой в исследуемой нами области подразумеваются обобщенные классы, свойства и неосязаемые сущности. Например, ограничения современных нейронных систем машинного перевода связаны с тем, что такие модели используют словарный запас в 30 000–500 000 *токенов*, постоянно обращаясь к внешним словарям — источникам данных, создавая искусственный барьер, не соответствующий природе естественного языка: для продуктивных словообразовательных процессов, таких как транслитерация имен собственных в русском языке, необходимо применять методы обработки данных, связанных с единицами меньше, чем целое слово, одновременно включающее префиксы, суффиксы, интерфиксы и окончания. Р. Сенрих, Б. Хэддоу и А. Бёрч экспериментальным путем выяснили, что *сублексемная сегментация* может решать проблемы, связанные с распознаванием и нахождением неизвестных слов без обращения к внешним словарям. Также они доказали, что подобный метод обработки данных позволяет улучшать качество перевода. По полученным данным, современные механизмы *сублексемной сегментации* повышают качество и эффективность перевода с английского на русский язык с 6,6 до 18,3 % [1. P. 1716–1718].

Для сжатия данных часто используется алгоритм Byte Pair Encoding (BPE), адаптированный под сегментацию слов на *сублексемы* в задачах по обработке естественного языка (рис.).

В результате *морфемизации* получается словарь, в котором существуют отдельные буквы, частотные биграммы, целые слова (в случае высокой частотности), морфемы, слоги и любые другие устойчивые последовательности. Необходимость для такого членения в нейронных сетях возникает из-за невозможности обработки бесконечных потоков данных. Если оставлять целые слова, словарь будет слишком большим, и на этапе обучения LLM не сможет оперировать данными в условиях ограниченной памяти. Словарь позволяет сохранять и использовать фиксированный и компактный набор *токенов* (от 50000 до 100000 в современных моделях), а также сборку слов из *подслов*, даже если их не было в обучении.

Проблема эффективной токенизации для больших языковых моделей остается актуальной и важной в эпоху ИИ. Исследователи и разработчики ищут баланс между компрессией словаря и способностью нейронной сети к морфологическому восприятию и *морфемно-морфологической обработке* (табл.).



Этапы алгоритмической морфемизации методом BPE
Источник: составлено К.П. Жикулиной с помощью Napkin.ai[1]
The stages of algorithmic morphemization by the BPE method
Source: compiled by Kristina P. Zhikulina with Napkin.ai

На основе сопоставительного анализа трех поколений BPE-*токенизаторов* (одноязычный, двуязычный и мультиязычный) были выделены следующие критерии:

- ***Фрагментация***, которая наблюдается, когда одна морфема последовательно разбивается на несколько статистически частотных *токенов* в корпусе, однако совпадение с морфемой частичное (например, *до* и *рож* вместо *дорож*);
- ***Контаминация***, при которой происходит соединение частей разных морфем (например, интерфикс -о- с началом или концом корней) и образуется *гибридный токен*;

[1] Нейронная сеть для генерации изображений Napkin.ai // napkin.ai. Режим доступа: https://www.napkin.ai (дата обращения: 20.01.2026).

• ***Аберрация*** происходит тогда, когда *токен* не имеет аналога в морфемной структуре используемого языка, при этом не является слогом и не несет никакого семантического или структурного смысла, превращаясь в статистический артефакт (например, *со* и *ко*).

Типология морфологических искажений при сублексемной сегментации (на примере лексем «подорожник» и «высокотехнологичный») Typology of morphological distortions in subword segmentation (using words 'podorozhnik' and 'vysokotekhnologichny' as examples)

Вид токенизации / Type of tokenization	**Границы токенов / Boundaries of tokens**	**Лингвистическое искажение / Linguistic distortion**
Правильное морфемное членение слова *подорожник*: по | дорож | ник / The correct morphemic division of the word *psyllium*: po | dorozh | nik		
Одноязычная BPE (только русский язык) / Monolingual BPE (only Russian language)	по | до | рож | ник / po | do | rozh | nik	• *Фрагментация и статическая аппроксимация* Корень *дороож* разбит на частотные встречаемости пар символов *до* и *рож*, а приставка *по-* и суффикс *-ник-* узнаваемы. Морфологически неточное членение, но узнаваемое. • Fragmentation and static approximation The root *dorozh* is divided into frequency of occurrence pairs of characters *do* and *rozh*, and the prefix *po-* and the suffix *-nik-* are recognizable. Morphologically inaccurate division, but recognizable.
Двуязычная (или совместная) BPE (английский и русский языки) / Bilingual BPE (or collaborative) (English and Russian languages)	по | дор | ож | ник / po | dor | ozh | nik	• *Контаминация и лексическая интерференция* Корень *дороож* членится под английские слова *door (дор)* и частотный токен окончания англо-французских слов *ож (garage)*. Размытие морфем происходит под влиянием паттернов языка-донора. • Contamination and lexical interference The root *dorozh* is divided into English word *door* and the frequency token ending of English-French words *ozh* (garage). Morpheme blurring occurs under influence of patterns of the donor language.
Мультиязычная BPE (более 100 языков) / Multilingual BPE (over 100 languages)	под | оро | жник / pod | oro | zhnik	• *Статистическая аберрация* Токены не соответствуют морфемам и слогам: *под* – английское *under*, *оро* – итальянское *золото*, *-жник* – артефакт, или бессмысленный набор символов.

Продолжение табл. / Continuation of the Table

Вид токенизации / Type of tokenization	**Границы токенов / Boundaries of tokens**	**Лингвистическое искажение / Linguistic distortion**
		• *Statistical aberration* Tokens do not correspond to morphemes and syllables: *pod* is English *under*, *opo* – is Italian *gold*, *zhnik* – is an artifact, or a meaningless set of characters.
Правильное морфемное членение слова *высокотехнологичный*: высок | о | техн | о | лог | ич | н | ый / The correct morphemic division of the word *high-tech*: vysok | o | tehn | o | log | ich | n | yj		
Одноязычная BPE (только русский язык) / Monolingual BPE (only Russian language)	высоко | тех | но | ло | гич | ный / vysoko | teh | no | lo | gich | nyj	• *Фрагментация с элементами аппроксимации* Корень *высок* соединён с интерфиксом *-о-*, корень *тех* выделен неверно, потеря второго интерфикса *-о-*, корень *лог* и суффикс *-ич-* разделены на два токена *ло* и *гич*, не выделен отдельно суффикс *-н-*, окончание воспроизведено как *-ный*. Морфологическая путаница присутствует, но слово остаётся опознаваемым. • *Fragmentation with approximation elements* The root *vysok* is connected to the prefix *-o-*, the root *teh* is highlighted incorrectly, the loss of the second prefix *-o-*, the root *log* and suffix *-ich-* are divided into two tokens *lo* and *gich*, the suffix *-n-* is not highlighted separately, the ending is reproduced as *-nyj*. Morphological confusion is present, but the word remains recognizable.
Двуязычная (или совместная) BPE (английский и русский языки) / Bilingual (or collaborative) BPE (English and Russian languages)	вы | со | ко | те | хно | ло | гич | ный / vy | so | ko| te | hno | lo | gich | nyj	• *Фрагментация с контаминацией* Токен *вы* отделён от корня, *со* и *ко* становятся статистическими артефактами, *хно* – гибридный токен, соединяющий часть корня *техн* и интерфикс *-о-*. Интерференция английского *technology* ведёт к разрыву корня. / • *Fragmentation with contamination* The token *vy* is separated from the root, *so* and *ko* become statistical artifacts, and *hno* is a hybrid token connecting a part of the root *tehno* and prefix *-o-*. The interference of English technology leads to a rupture of the root.

Окончание табл. 1 Ending of the Table

Вид токенизации / Type of tokenization	**Границы токенов / Boundaries of tokens**	**Лингвистическое искажение / Linguistic distortion**
Мультиязычная BPE (более 100 языков) / Multilingual BPE (over 100 languages)	высок | отех | ноло | гич | ный / vysok | oteh | nolo | gich | nyj	• *Контаминация с аберрацией* *Отех* – соединение корня с интерфиксом *-о-*, *ноло* – гибридный токен от английского и латинского *nolo*. Морфологическая структура полностью разрушена, границы токенов утрачены - нет соответствий ни морфемам, ни слогам. / • *Contamination with aberration* *Oteh* – it is a compound of the root with the prefic *-o-*, *nolo* – it is a hybrid token from English and Latin. The morphological structure is completely destroyed; the boundaries of the tokens are lost – there are not correspondences to either morphemes or syllables.

Источники: токенизация выполнена К.П. Жикулиной с использованием OpenAI Tokenizer (tiktoken) для GPT-4 [7], Hugging Face Transformers (AutoTokenizer.from_pretrained("meta-llama/Llama-3-8b")) [6], интерактивного инструмента токенизации GPT Tokenzier [4], демонстративный BPE-токенизатор FatimaALzahrani с пошаговой визуализацией [1].
Source: compiled by Kristina P. Zhikulina with using OpenAI Tokenizer (tiktoken) for GPT-4 [7], Hugging Face Transformers (AutoTokenizer.from_pretrained("meta-llama/Llama-3-8b")) [6], interactive tokenization tool GPT Tokenzier [4], demonstrative FatimaALzahrani BPE tokenizer with step-be-step visualization [1].

Границы между типами морфологических расхождений определяются степенью семантической мотивированности *токена*: в случае фрагментации морфема узнаваема, но бывает раздроблена; в случае контаминации морфема неузнаваема, а *токен* выступает гибридным образованием; в случае аббревиации морфема отсутствует, а *токен* — носитель *псевдоморфемного смысла*.

Описанные выше искажения являются следствием детерминированного характера классического BPE-*токенизатора*: каждое слово получает одно статистически вероятное частотное членение [1]. Однако Т. Кудо обращает внимание на то, что *сублексемная сегментация* неоднозначна: одно и то же слово может быть разделено множеством способов даже при условии использования фиксированного словаря [7]. Предложенный им метод доказывает, что, с лингвистической точки зрения, алгоритмическая «тень» морфемы может переставать быть фиксированной и становиться проекциями одно и того же языкового объекта. Вероятностный взгляд Т. Кудо на *токенизацию* открывает перспективу для снижения описанных выше морфологических искажений и потерь с помощью обучения модели на определенном множестве «теней», а не сфокусированной в одной точке.

Существуют и более продвинутые инструменты, как SentencePiece, который не требует предварительной сегментации текста на слова

и рассматривает пробел как отдельный символ, что делает токенизацию обратимой и независимой от языка [8. P. 67–70]. Но на практике оказывается, что алгоритм перестает получать подсказки о том, где проходят границы слов, и вынужден восстанавливать их по частотности: «тень» морфемы становится более автономной и менее привязанной к языковой реальности.

О.В. Ломакина предлагает разграничение ценностных констант и ценностных переменных, что имеет место и в алгоритмической лингвистике [9. С. 623–625]. В системе *сублексемной сегментации* также можно говорить о морфемных константах — корнях, приставках, суффиксах, которые сохраняют свою целостность даже при многократном прохождении через BPE-сжатие [10. С. 28–29], и морфемных переменных — единицах, членимость которых варьируется в зависимости от обучающего корпуса модели, степени мультиязычности и числа операций слияния. Так, на примере слова *подорожник* (см. Табл.), корень *дорож* в русском языке, при однозначной *токенизации*, демонстрирует константное поведение, а в мультиязычной модели превращается в переменную — утрачивается морфемная идентичность.

Заключение

Типология морфологических расхождений позволяет говорить о неравноценном лингвистическом членении, а также о невозможности следования классической морфологической схеме русского языка в современных условиях технических возможностей ИИ, а именно — ограничениям памяти и скорости обработки слишком больших массивов данных. Анализ показывает тесную связь теоретической морфологии и инженерных практик: интервенцию в морфологический строй языка. *Типология морфологических расхождений* — это инструмент лингвистической рефлексии над алгоритмическими практиками. Она позволяет определить, что именно происходит с морфемой при BPE-сегментации, классифицировать степень и характер искажения, оценивать универсальность каждого отдельного языка.

Классическая морфология оперирует устойчивыми и семантически расширенными морфемами, границы которых определяет этимология, парадигматика, словообразование и словоизменение. Однако можно отметить, что статистическая сегментация в нейронных сетях не дополняет базовые принципы, а конкурирует с ними, предлагая иную онтологию членения речи — по частоте. *Токен* выступает в качестве алгоритмической «тени» морфемы, так как является порождением алгоритма, морфема же дана языком. Также «тень» лишь повторяет очертания, но не сущность: *токен* внешне схож с морфемой, но не имеет пока что этимологических, парадигматических и словообразовательных гнезд. Морфема есть в языке, но *токен* – это то, что осталось от морфемы после прохождения через алгоритм сжатия, однако он может указывать на нее, но не повторять или замещать. Можно сказать, что морфема является языковым наследием, а токен — компромисс цифровой стратификации языка. Чем более универсальной становится LLM, тем более искаженной становится алгоритмическая «тень» морфемы.

Список литературы

1. *Sennrich R., Haddov B., Brich A.* Neural Machine Translation of Rare Words with Subword Units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. N.-Y. : Cornell University, 2016 P. 1715–1725. https://doi.org/10.48550/erXiv.1508.07909
2. *Максименко О.И.* Формализованная и цифровая лингвистика. М. : Государственный университет просвещения, 2024. EDN: NPXJZB
3. *Никитина Е.Н., Станкевич М.А., Ларионов Д.С.* Эмотивные глаголы с отрицанием: опыт интерпретации данных автоматического анализа текстов // Медиалингвистика. 2025. Т. 12. № 2. С. 174–198. https://doi.org/10.21638/spbu22.2025.201 EDN: TDQIWC
4. *Гудков В.В., Митренина О.В., Соколов Е.Г., Коваль А.А.* Языковой перенос нейросетевого обучения для частеречной разметки Санкт-Петербургского корпуса агиографических текстов (СКАТ) // Вестник Санкт-Петербургского университета. Язык и литература. 2023. Т. 20. № 2. С. 268–282. https://doi.org/10.21638/spbu09.2023.205 EDN: VQHOZM
5. *Польщикова О.Н.* Становление и формирование терминологии компьютерной лингвистики // Вопросы журналистики, педагогики, языкознания. 2022. Т. 41. № 3. С. 590–607. https://doi.org/10.52575/2712-7451-2022-41-3-590-607 EDN: HTWXDI
6. *Горожанов А.И., Степанова Д.В.* Интерпретация художественного произведения: корпусный подход // Филологические науки. Вопросы теории и практики. 2022. № 15 (1). С. 203–208. https://doi.org/10.30853/phil20220020 EDN: TCZLAF
7. *Kudo T.* Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates // arXiv.org. 2018. Режим доступа: https://arxiv.org/abs/1804.10959 (дата обращения: 18.12.2025).
8. *Kudo T., Richardson J.* SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels. P. 66–71. 2018 https://doi.org/10.48550/arXiv.1808.06226
9. *Ломакина О.В.* Метаязык современной лингвокультурологии: паремиологический уровень // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2025. Т. 16. № 3. С. 622–637. https://doi.org/10.22363/2313-2299-2025-16-3-622-637 EDN: DQGURR
10. *Зонова Д.Ю.* Принцип работы и проблемы «generative pre-trained transformer artificial intelligence» // Вестник науки и образования. 2023. № 8 (139). С. 28–31. EDN: AFNBRE

References

1. Sennrich, R., Haddov, B., & Brich, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1715–1725). N-Y.: Cornell University publ. https://doi.org/10.48550/erXiv.1508.07909
2. Maksimenko, O.I. (2024). *Formalized and Digital Linguistics*. Moscow: State University of Education publ. (In Russ.). EDN: NPXJZB
3. Nikitina, E.N., Stankevich, M.A., & Larionov, D.S. (2025). Psych Verbs Used with Negation: A Way of Interpretation of Automatic Text Analysis Data. *Media Linguistics*, *12*(2), 174–198. (In Russ.). https://doi.org/10.21638/spbu22.2025.201 EDN: TDQIWC
4. Gudkov, V.V., Mitrenina, O.V., Sokolov, E.G., & Koval, A.A. (2023). Language-Based Transfer Learning Approaches for Part-of-Speech Tagging on Saint Petersburg Corpus of Hagiographic Texts (Skat). *Vestnik of Saint Petersburg University. Language and Literature*, *20*(2), 268—282. (In Russ.). https://doi.org/10.21638/spbu09.2023.205 EDN: VQHOZM

5. Polshchykova, O.N. (2022). Formation of Computational Linguistics Terminology. *Issues in Journalism, Education, Linguistics*, *41*(3), 590–607. (In Russ.). https://doi.org/10.52575/2712-7451-2022-41-3-590-607 EDN: HTWXDI
6. Gorozhanov, A.I., & Stepanova, D.V. (2022). Work of Fiction Interpretation: Corpus Approach. *Philology. Theory & Practice*, *15*(1), 203–208. (In Russ.). https://doi.org/10.30853/phil20220020 EDN: TCZLAF
7. Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates // arXiv.org. URL: https://arxiv.org/abs/1804.10959 (accessed: 18.12.2025).
8. Kudo, T., & Richardson, J. (2018). SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66–71). Brussels. https://doi.org/10.48550/arXiv.1808.06226
9. Lomakina, O.V. (2025). Metalanguage of Modern Linguaxiology: Paremiological Level. *RUDN Journal of Language Studies. Semiotics and Semantics*, *16(*3), 622–637. (In Russ.). https://doi.org/10.22363/2313-2299-2025-16-3-622-637 EDN: DQGURR
10. Zonova, D.Yu. (2023). The Principle of Operation and Problems of "Generative Pre-trained Transformer Artificial Intelligence". *Bulletin of Science and Education*, 8 (139), 28–31. (In Russ.). EDN: AFNBRE

Сведения об авторе:
Жикулина Кристина Петровна, кандидат филологических наук, ассистент кафедры общего и русского языкознания филологического факультета, Российский университет дружбы народов (117198, Российская Федерация, г. Москва, ул. Миклухо-Маклая, д. 6); *сфера научных интересов*: искусственный интеллект, компьютерная лингвистика, нейролингвистическое программирование, когнитивистика, голосовые помощники, нейронные сети, современный русский язык; *e-mail*: zhikulina-kp@rudn.ru
ORCID: 0000-0003-2488-4616; SPIN-код: 5871-9336, Scopus Author ID: 59236754500.

Information about the author:
Kristina P. Zhikulina, PhD in Philology, Assistant of the Department of General and Russian Linguistics, Faculty of Philology, RUDN University (6 Miklukho-Maklaya str., Moscow, Russian Federation, 117198); *Research interests*: artificial intelligence, computational linguistics, neuro-linguistic programming, cognitive science, voice assistance, neural networks, modern Russian language; *e-mail*: zhikulina-kp@rudn.ru
ORCID: 0000-0003-2488-4616; SPIN-code: 5871-9336, Scopus Author ID: 59236754500.