

ПРОБЛЕМА НЕСКОЛЬКИХ ПИСЬМЕННОСТЕЙ И НОРМАЛИЗАЦИИ ТЕКСТОВ В МАШИННОМ ПЕРЕВОДЕ УЗБЕКСКОГО И КАРАКАЛПАКСКОГО ЯЗЫКОВ

М.А.Узакова¹, У.Ташкент,

1. Международный университет Азии

Mas'ul muallif: Узакова Мамура Абдураимовна, Международный университет Азии.

Received: 15 Oktabr 2026

Accepted: 20 Oktabr 2026

Published: 25 Oktabr 2026

Copyright: © 2026 by the author(s). Licensee Buxoro xalqaro universiteti ilmiy axborotnomasi.

Iqtibos uchun: Узакова, М. А. ПРОБЛЕМА НЕСКОЛЬКИХ ПИСЬМЕННОСТЕЙ И НОРМАЛИЗАЦИИ ТЕКСТОВ В МАШИННОМ ПЕРЕВОДЕ УЗБЕКСКОГО И КАРАКАЛПАКСКОГО ЯЗЫКОВ / М. А. Узакова, У. Ташкент, // Buxoro xalqaro universiteti ilmiy axborotnomasi. — 2026. — Т. 10, № 2. — В. 276–278. — DOI: 10.67780/buxoro-xalqaro-universiteti-ilmiy-axborotnomasi.2026.febccd8c.

Аннотация. В статье рассматривается влияние графической неоднородности письменных текстов на качество машинного перевода для узбекского и каракалпакского языков, которые в разные периоды использовали латиницу и кириллицу и до сих пор существуют в обеих графиках с несколькими вариантами орфографии. Показано, что смешение алфавитов, неоднозначность соответствий между графическими системами и вариативность апострофов и диакритики приводят к фрагментации словаря, росту числа неизвестных токенов и снижению качества нейросетевого перевода.

Ключевые слова: машинный перевод, транслитерация, нормализация текста, узбекский язык, каракалпакский язык, малоресурсные языки, тюркские языки, латиница, кириллица.

O'ZBEK VA QORAQALPOQ TILLARIDA MASHINA TARJIMASIDA BIR NECHTA YOZUV TIZIMI VA MATNLARNI ME'YORLASHTIRISH MUAMMOSI

Annotatsiya. Maqolada turli davrlarda lotin va kirill yozuvlaridan foydalanib kelgan, hozirgi vaqtda ham bir nechta grafik tizim va imlo variantlarida qo'llanayotgan o'zbek va qoraqalpoq tillarida yozma matnlarning grafik jihatdan bir xilda emasligi mashina tarjimasi sifatiga qanday ta'sir ko'rsatishi ko'rib chiqiladi. Alifbolarning aralash qo'llanilishi, grafik tizimlar o'rtasidagi muvofiqlikning bir xil emasligi, shuningdek, apostrof va diakritik belgilarning turlicha ishlatilishi lug'at birliklarining tarqoqlashuviga, noma'lum tokenlar sonining ortishiga va neyron tarmoqlarga asoslangan tarjima sifatining pasayishiga olib kelishi ko'rsatib beriladi.

Kalit so'zlar: mashina tarjimasi, transliteratsiya, matnni me'yorlashtirish, o'zbek tili, qoraqalpoq tili, resurslari kam tillar, turkiy tillar, lotin yozuvi, kirill yozuvi.

THE PROBLEM OF MULTIPLE WRITING SYSTEMS AND TEXT NORMALIZATION IN MACHINE TRANSLATION OF UZBEK AND KARAKALPAK

Abstract. This article examines the impact of graphic inconsistency in written texts on the quality of machine translation for Uzbek and Karakalpak, languages that have used both Latin and Cyrillic scripts at different historical periods and continue to exist in both scripts with multiple spelling variants. It demonstrates that the mixing of alphabets, ambiguous correspondences between writing systems, and variations in the use of apostrophes and diacritical marks lead to lexical fragmentation, an increase in the number of unknown tokens, and a decline in the quality of neural machine translation.

Keywords: machine translation, transliteration, text normalization, Uzbek language, Karakalpak language, low-resource languages, Turkic languages, Latin script, Cyrillic script.

Введение. В условиях стремительного развития цифровых технологий и систем искусственного интеллекта машинный перевод становится одним из важнейших инструментов межъязыковой коммуникации, обработки текстовой информации и расширения доступа к научным и образовательным ресурсам. Особую актуальность данное направление приобретает для языков с ограниченными цифровыми ресурсами, к которым относятся узбекский и каракалпакский языки. Несмотря на постепенное расширение электронных корпусов, развитие национальных языковых технологий и внедрение нейросетевых моделей перевода, качество автоматической обработки текстов на этих языках по-прежнему зависит от ряда лингвистических и технологических факторов. Одним из таких факторов выступает неоднородность письменных систем и отсутствие единообразия в представлении языковых единиц в цифровой среде.

Узбекский и каракалпакский языки характеризуются историческим использованием различных графических систем, прежде всего кириллицы и латиницы. Переходные процессы в письменности, сохранение ранее созданных текстовых материалов и одновременное функционирование разных графических вариантов обусловили формирование неоднородного массива цифровых данных. В результате одна и та же лексическая единица может встречаться в нескольких

<https://doi.org/10.67780/buxoro-xalqaro-universiteti-ilmiy-axborotnomasi.2026.febccd8c>

графических формах, отличающихся использованием букв, апострофов, диакритических знаков и орфографических вариантов. Для человека такие различия зачастую не препятствуют пониманию текста, однако для автоматизированных систем они могут создавать существенные трудности при распознавании слов, определении их значений и установлении межъязыковых соответствий.

Особенно важное значение в данном контексте приобретает нормализация текста, представляющая собой совокупность процедур приведения графических и орфографических вариантов к согласованной форме. Нормализация позволяет уменьшить количество формально различных, но лексически тождественных единиц, повысить согласованность языковых данных и создать более устойчивую основу для обучения моделей машинного перевода. Развитие нейросетевого машинного перевода (НМП) существенно улучшило качество перевода для языков с большими параллельными корпусами, однако для малоресурсных языков результаты остаются нестабильными [6, 7]. К числу таких языков относятся узбекский и каракалпакский, принадлежащие к тюркской семье: узбекский входит в карлукскую (юго-восточную) ветвь, каракалпакский относится к кыпчакской (северо-западной) [11]. Кроме нехватки параллельных данных, для них характерна особая проблема, которая в работах по НМП обычно остаётся на втором плане: нестабильность письменной формы текста.

Узбекский язык в XX веке сменил несколько графических систем: арабскую письменность, латиницу (1920-е гг.), кириллицу (с 1940 г.) и вновь латиницу после принятия закона о введении алфавита на латинской основе в 1993 г., с последующими изменениями алфавита в 1995 г. и позднее [11]. Каракалпакский язык прошёл сходный путь, а переход на латиницу был закреплён в 2016 г. и затем уточнялся. В результате в открытых источниках одновременно присутствуют тексты в кириллице, в латинице разных редакций и «смешанные» тексты, где фрагменты записаны в разных системах или с произвольной заменой специфических букв.

Для НМП это означает, что одно и то же слово может быть представлено десятками графических вариантов. Токенизаторы, обученные на данных такой природы, фрагментируют словарь, а модель обучается воспринимать варианты как разные слова. Проблема усиливается агглютинативной морфологией: ошибка в одном символе суффикса приводит к потере целой словоформы.

Подсловная сегментация методами BPE и SentencePiece стала стандартом [4,5], но её эффективность зависит от согласованности обучающих данных. Исследование Rust и соавторов показывает, что качество токенизатора для конкретного языка напрямую связано с качеством модели на задачах этого языка [8]. Для агглютинативных языков и данных с графической неоднородностью это означает повышенную фрагментацию слов.

Отдельных работ, посвящённых совместному влиянию графической неоднородности и нормализации на перевод именно между узбекским и каракалпакским языками, автору не известно; этот пробел и определяет актуальность настоящего исследования.

Материалы и методы

На основании анализа алфавитов обоих языков выделены следующие группы проблем.

Часть соответствий кириллицы и латиницы зависит от контекста. Например, в узбекском языке буква «е» передаётся как *ue* в начале слова и после гласных и как *e* в остальных позициях, а буквы «ц», «ъ», «ь» встречаются преимущественно в заимствованных из русского слов и требуют особых правил. Обратная транслитерация (из латиницы в кириллицу) также неоднозначна для сочетаний *sh*, *ch*, *ng*, которые могут читаться как отдельные звуки или как последовательность [11].

Для каракалпакского языка в кириллице используются буквы *Ə, F, K, H, Ø, Y, Ÿ, X*, а в латинице - *Á, Ġ, Í, Ñ, Ó, Ú*. На практике, особенно в интернете, эти символы заменяют ближайшими ASCII-символами или кириллическими буквами внешне схожей формы [11].

В узбекской латинице буквы *o'* и *g'* записываются с помощью апострофа, который в реальных текстах представлен символами *U+02BB*, *U+2018*, *U+2019*, *U+0027* и другими. Для модели это разные символы, а для читателя - один и тот же [10].

Встречаются тексты, где часть слов записана кириллицей, часть латиницей, а также слова с гомоглифами, где латинская буква подставлена вместо кириллической и наоборот.

Тексты разных периодов отражают разные редакции алфавитов, а государственные нормативные акты, регулирующие их, неоднократно уточнялись.

Предлагаемый конвейер предобработки состоит из трёх последовательных этапов:

Определение графики для каждого предложения или его фрагмента определяется доля символов латиницы, кириллицы и специфических знаков. Для коротких или смешанных фрагментов применяется классификатор на основе символьных *n*-грамм, различающий узбекскую и каракалпакскую графику и их редакции.

Нормализация выполняется:

- приведение к форме Unicode NFC;
- унификация апострофов и модификаторов к единому символу;
- замена гомоглифов на основании окружающего контекста;
- восстановление диакритических знаков по словарю и контексту;
- приведение орфографических вариантов к выбранной редакции алфавита.

Транслитерация реализуются две версии: правилочная (детерминированные таблицы соответствий с

<https://doi.org/10.67780/buxoro-xalqaro-universiteti-ilmiy-axborotnomasi.2026.febccd8c>

контекстными правилами и словарём исключений для заимствований) и нейросетевая (модель «символ - символ» на основе архитектуры transformer, обученная на выровненных парах слов в двух графиках). Правильная версия служит базовой линией и источником синтетических данных для обучения нейросетевой.

Для оценки транслитерации используется набор выровненных текстов в двух графиках объёмом [N] предложений. Для оценки перевода используются тестовые наборы пары узбекский - каракалпакский, подготовленные по образцу FLORES [7] и проверенные носителями языка, объёмом [N] предложений. Для обучения применяются параллельные данные из открытых источников, количество и происхождение которых указывается явно.

Конфигурации сравнения:

B0: базовая модель без предобработки;

B1: только унификация Unicode и апострофов;

B2: нормализация и правильная транслитерация в единую графику;

B3: нормализация и нейросетевая транслитерация;

B4: дообучение модели с аугментацией, при которой в обучающие данные искусственно вносятся замены графики и орфографический шум.

Закключение. Графическая неоднородность является самостоятельным фактором, влияющим на качество машинного перевода для узбекского и каракалпакского языков. Предложенный конвейер, включающий определение графики, нормализацию символов и орфографии и контекстно-зависимую транслитерацию, позволяет свести разнородные тексты к единому представлению и тем самым уменьшить фрагментацию словаря.

Перспективными направлениями являются: использование безтокенных и символьных моделей, совместное обучение транслитерации и перевода, расширение подхода на киргизский и другие тюркские языки, а также создание открытых размеченных ресурсов по графическим вариантам, которые могут быть использованы научным сообществом.

Список использованной литературы

- [1] Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting of the ACL. 2002. P. 311–318.
- [2] Popović M. chrF: character n-gram F-score for automatic MT evaluation // Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015. P. 392–395.
- [3] Rei R., Stewart C., Farinha A. C., Lavie A. COMET: A neural framework for MT evaluation // Proceedings of EMNLP. 2020. P. 2685–2702.
- [4] Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units // Proceedings of the 54th Annual Meeting of the ACL. 2016. P. 1715–1725.
- [5] Kudo T., Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing // Proceedings of EMNLP: System Demonstrations. 2018. P. 66–71.
- [6] Э. Назилова, Ш.Абидова, М.Узакова "Туркий тиллар учун икки тилли электрон таржиманинг модел ва алгоритмлари" Мухаммад ал-хоразмий номидаги тошкент ахборот технологиялари университети. Монография. Тошкент-2022 й. Монография. Тошкент-2022. «Бизнес-полиграф».– 102 б. ББК:22.18:81.2
- [7] Nazirova Elmira, Abidova Shakhnoza, Uzakova Mamura, Usmanova Kamola CREATING OPEN-CORRES PARALLEL CORPUS BASES FOR MACHINE TRANSLATION J.: "Bulletin of TUIT: Management and Communication Technologies". (05.00.00; OAK rayosat qarori 30.07.2020. № 01-10/1103) 2025. –7p. <https://doi.org/10.61663/251tuitmct5>.
- [8] E. Nazirova, Sh.Abidova, M. Uzakova "Mathematical Models of Complex Word Formation in Uzbek, Karakalpak and Kyrgyz Languages belonging to the Turkic Language Family" International conference "Modern Problems of Applied Science and Engineering" – 2024, 2-3 may
- [9] E.Sh. Nazirova, M.A. Uzakova LINGUO MATHEMATICAL MODELS OF SENTENCES BY TYPES OF UZBEK LANGUAGE FOR ELECTRONIC TRANSLATION SCIENTIFIC REVIEW OF THE PROBLEMS AND PROSPECTS OF MODERN SCIENCE AND EDUCATION. GERMANY Vol. 1 No. 1 (2025)
- [10] <https://zenodo.org/badge/DOI/10.5281/zenodo.15049689.svg>
- [11] Nazirova Elmira Shodmonovna, Abidova Shakhnoza Baxadirovna, Uzakova Mamura Abduraimovna, Kholmuminov Yusuf Kholbuta o'g'li "LINGUISTIC MATHEMATICAL MODELS AND ALGORITHMS OF THE PROCESS PHONETIC ANALYSIS IN THE UZBEK LANGUAGE" SCIENCE AND EDUCATION: DEVELOPMENT VECTORS ASTANA, KAZAKHSTAN, JANUARY 2025 UDK 378 (063) BKB 74.58 C 940 28
- [12] Закон Республики Узбекистан «О введении в действие алфавита узбекского языка, основанного на латинской графике» от 2 сентября 1993 г. (с последующими изменениями).