ISNN 3030-3907

2026/3
VOLUME 2

DEVELOPMENT OF SCIENCE

ILMIY JURNAL



Development of science

Ilmiy jurnal

2026/3 VOLUME 2

ISSN 3030 -3907

Ilmiy jurnal OAK rayosatining 2024 yil 27 sentabrdagi 361-son qaroriga asosan 02.00.00 kimyo fanlari hamda 2025 yil 12 fevral 367- son qaroriga asosan 05.00.00 texnika fanlari bo‘yicha falsafa doktori (PhD) va fan doktori (DSc) ilmiy darajasiga talabgorlarning dissertatsiya ishlari yuzasidan dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan milliy ilmiy nashrlar ro‘yxatiga kiritilgan.

BOSH MUHARRIR:
Temirov Alisher Hoshim o‘g‘li – texnika fanlari falsafa doktori (PhD),Dotsent.
TAHRIRIYAT KENGASH RAISI:
Fozilov Sadriddin Fayzullayevich – texnika fanlari doktori, professor
TAHRIRIYAT KENGASH A’ZOLARI

TEXNIKA FANLARI/ TECHNICAL SCIENCES/ ТЕХНИЧЕСКИЕ НАУКИ

Bozorov G‘ayrat Rashidovich
Buxoro Davlat Texnika Universiteti, professor
Abdurahmonov Olim Rustamovich
Buxoro Davlat Texnika Universiteti, professor
Adizov Bobur Zamirovich
O‘zbekiston Respublikasi Fanlar akademiyasi Umumiy va noorganik kimyo instituti, professor
Hayitov Ruslan Rustamjonovich
Buxoro Davlat Texnika Universiteti, professor
Jumayev Qayum Karimovich
Buxoro Davlat Texnika Universiteti, professor.
Berdiyev Dorob Murodovich
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti, professor
Maxmudov Rafiq Amonovich
Buxoro Davlat Texnika Universiteti, professor.
Tilloyev Lochin Ismatilloyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Raxmonov Qahramon Sanakulovich
Buxoro Davlat Texnika Universiteti, dotsent.
Barakayev Nusratullo Rajabovich
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti, professor.
Jo‘rayev Xayrullo Fayziyevich
Buxoro Davlat Texnika Universiteti, professor.
Panoyev Erali Rajabboyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Ataullayev Sherzod Nabiullyevich
Buxoro Davlat Texnika Universiteti, dotsent.
Murodov Malik Negmurodovich
Buxoro Davlat Texnika Universiteti, dotsent.
Axmedova Ozoda Bahronovna
Islom Karimov nomidagi Toshkent Davlat Texnika Universiteti.
Safarov Bahri Jumayevich
Buxoro Davlat Texnika Universiteti, dotsent.
Xaydarov Axtam Amonovich
Buxoro Davlat Texnika Universiteti, dotsent.
Safarov Jasur Alijon o‘g‘li
Buxoro Davlat Texnika Universiteti, dotsent.
Асадова Раъно Дилмуратовна
Тошкент Давлат техника университети доценти

KIMYO FANLARI/ CHEMICAL SCIENCES/ ХИМИЧЕСКИЕ НАУКИ

Axmedov Voxid Nizomovich
Buxoro Davlat Texnika Universiteti, professor
Umarov Baqo Bafoyevich
Buxoro Davlat Universiteti, professor
Kodirov Abduaxad Abduraximovich
Qarshi davlat Universiteti, professor
Qurbonov Mingniqul Jumagulovich
Qarshi davlat Universiteti, dotsent
Turemuratov Sharibay Naurizbaevich
O‘zbekiston Respublikasi Fanlar akademiyasi Qoraqalpog‘iston bo‘limi Bosh ilmiy kotibi, professor
Tursunov Kaxor Shonazarovich
Qarshi davlat Universiteti, professor
Do‘stov Hamro
Buxoro Davlat Texnika Universiteti, professor
Panoyev Nodir Shavkatovich
Buxoro Davlat Texnika Universiteti, dotsent.
Olimov Bobur Bahodir o‘g‘li
Buxoro Davlat Texnika Universiteti, dotsent.
Kasimov Sherzod Abduazimovich
Termiz Davlat Universiteti, professor.
Aliqulov Rustam Valiyevich
Termiz Davlat Universiteti, professor
Tropov Xamza Tursunovich
Samarqand Davlat Universiteti, professor

IQTISODIYOT FANLARI/ ECONOMICS/ ЭКОНОМИЧЕСКИЕ НАУКИ

Nizamov Aslitdin Badreddinovich
Buxoro Davlat Texnika Universiteti, professor.
Boboyev Akmal Choriyevich
Buxoro Davlat Texnika Universiteti, professor.
Ergashxojayeva Shaxnoza Djasurovna
Toshkent davlat iqtisodiyot universiteti, professor
Saidaxmedova Nodira Ilxomovna
Toshkent davlat iqtisodiyot universiteti, professor
Muxsinov Bekzod Toxirovich
Buxoro Davlat Texnika Universiteti, dotsent.
Shakarov Qulmat Ashirovich
Renessans ta'lim universiteti,dotsent.
Jumaeva Zulfiya Qayumovna
Buxoro Davlat Texnika Universiteti, dotsent.
Norova Salomat Yusupovna
Buxoro Davlat Texnika Universiteti, dotsent.
Avezova Shaxnoza Maxmudjonovna
Buxoro Davlat Texnika Universiteti, dotsent.
Jalilov Jamshid G‘anijonovich
Toshkent davlat iqtisodiyot universiteti, dotsent
Xasanova Gulruh Djumanazarovna
Buxoro Davlat Texnika Universiteti, dotsent.

PEDAGOGIKA FANLARI/ PEDAGOGICAL SCIENCES / ПЕДАГОГИЧЕСКИЕ НАУКИ

Siddiqova Sadoqat G‘afforovna
Buxoro Davlat Texnika Universiteti, dotsent.
Zoyirov Erkin Xalilovich
Buxoro Davlat Texnika Universiteti, DSc, professor
Gaffarov Feruz Hasanovich
Puchon Universiteti, *professor*
Ruziyev Davron Yuldashevich
Buxoro Davlat Universiteti, professor
Muradova Firuza Rashidovna
Buxoro Davlat Texnika Universiteti, professor.
Yunusova Gulandon Samiyevna
Buxoro Davlat Texnika Universiteti, professor.
To‘xtayeva Zebo Sharifovna
Buxoro Davlat Texnika Universiteti, professor.
Navro‘zova Gulchixra Nigmatovna
Buxoro Davlat Texnika Universiteti, dotsent.
Uzoqova Laylo Polvonovna
Buxoro muhandislik – texnologiya instituti, dotsent.
To‘raqulova Marjona Qiyom qizi
Buxoro Davlat Texnika Universiteti, dotsent.
Muradova Zarina Rashidovna
Buxoro Davlat Texnika Universiteti, dotsent.
Murodov Sanjar Aslonovich
Buxoro Davlat Texnika Universiteti, dotsent.
Shodiyev Jahongir Jo‘raqulovich
Buxoro Davlat Texnika Universiteti, dotsent
Gaffarov Laziz Hasanovich
Buxoro muhandislik – texnologiya instituti, dotsent.

Tahririyat a’zolarning to‘liq ro‘yxati jurnalning www.devos.uz saytida keltirilgan.

Tahririyat hay’ati a’zolarining tarkibini o‘zgartirish hamda qo‘shimcha kiritish yoki tahririyat a’zoligidan chiqarish bosh muharrir tomonidan amalga oshiriladi.

ISSN 3030-3907

Development of science ilmiy jurnali 2024 yil 15 aprel kuni №256445 sonli guvohnoma bilan Davlat ro‘yxatidan o‘tgan.

To‘plam elektron ko‘rinishda (pdf formatida) mualliflarga taqdim etiladi. To‘plamga kiritiladigan maqolalarning mazmuni, ulardagi statistik ma’lumotlar va me’yoriy hujjatlarning to‘g‘riligiga hamda har qanday fikr - mulohazalarga mualliflarning o‘zlari mas’ul hisoblanadilar. Talablarga javob bermaydigan maqolalar to‘plamga kiritilmaydi. Tashkiliy qo‘mita maqola matnini qisqartirish, qisman tuzatish kiritish, seksiyalarga joylashtirish huquqiga ega.

ISSN 3030-3907

UDC: 62, 004.04

DEVELOPMENT OF A SYNONYM DICTIONARY FOR ARTIFICIAL INTELLIGENCE-BASED TRANSLATION MODELS FROM THE KHOREZM DIALECT INTO STANDARD UZBEK

Maksud Sharipov
Candidate of Technical Sciences, Associate Professor, Deportment Computer science and artificial intelligence technologies, Urgench State University named after Abu Rayhan Biruni
E-mail: m.sharipov@urdu.uz
Kurbanova Lola Ulug‘bek qizi
PhD student, Deportment Computer science and artificial intelligence technologies, Urgench State University named after Abu Rayhan Biruni
E-mail: kurbanova.lola@urdu.uz
Atanazarova Madina
Urganch RANCH Texnologiya Universiteti magistranti
E-mail: atanazarovamadina677@gmail.com

Annotatsiya. Xorazm shevasi o‘zbek tilining muhim hududiy ko‘rinishlaridan biri bo‘lib, kundalik muloqotda keng qo‘llaniladi, biroq uning yozma shakli va adabiy tilga avtomatik normallashtirish/“tarjima” masalasi resurslar tanqisligi sababli dolzarbligicha qolmoqda. Ushbu maqolada Xorazm shevasidagi matnlarni sun’iy intellekt yordamida o‘zbek adabiy tiliga o‘girish (dialekt → standart) modellari uchun sinonimik lug‘at ishlab chiqish yondashuvi taklif etiladi. Tadqiqotning asosiy manbasi sifatida Xorazm shevasiga oid izohli lug‘at bazasi tanlab olinib, u adabiy til ekvivalentlari bilan moslashtirildi. Lug‘at birliklari leksik-grammatik teglar (ot/sifat/fe’l/undov, to‘liqsiz fe’l va b.) asosida tartiblandi hamda “sheva birligi → adabiy til sinonimi(lar)i” ko‘rinishidagi mosliklar majmuasi shakllantirildi. Taklif etilgan sinonimik lug‘at dialektal variantlarni standartlashtirish, ma’nodoshlikni saqlagan holda normalizatsiya qilish va neyron tarjima/transformer asosidagi modellar uchun qo‘shimcha leksik bilim manbasi sifatida xizmat qiladi. Tajriba natijalari sinonimik lug‘at bilan boyitilgan yondashuv Xorazm shevasidan adabiy tilga o‘girishda leksik moslashuvni yaxshilashi, noaniq birliklarni izohlashga yordam berishi va ma’no yo‘qotilishini kamaytirishini ko‘rsatadi. Ish yakunida yaratilgan lug‘at resursi va uni model o‘qitish jarayoniga qo‘shish bo‘yicha amaliy tavsiyalar keltiriladi.

Kalit so'zlar: Xorazm shevasi; o‘zbek tili; sheva–adabiy til normallashtirish; sinonimik lug‘at; leksik resurs; sun’iy intellekt yordamida tarjima; neyron mashina tarjimasi (NMT); transformer modellari; resursi kam tillar NLP; o‘zbek adabiy tili.

ISSN 3030-3907

Аннотация. Хорезмский диалект является одной из важных территориальных разновидностей узбекского языка и широко используется в повседневном общении, однако из-за нехватки ресурсов проблема его письменной формы и автоматической нормализации/«перевода» на литературный язык по-прежнему остаётся актуальной. В данной статье предлагается подход к разработке синонимического словаря для моделей, выполняющих преобразование текстов на хорезмском диалекте в узбекский литературный язык (диалект → стандарт) с применением методов искусственного интеллекта. В качестве основного источника исследования была выбрана база толкового словаря хорезмского диалекта, которая была сопоставлена с эквивалентами литературного языка. Словарные единицы были систематизированы по лексико-грамматическим тегам (существительное/прилагательное/глагол/междометие, неполный глагол и др.), и сформирован набор соответствий в формате «диалектная единица → синоним(ы) литературного узбекского языка». Предложенный синонимический словарь служит для стандартизации диалектных вариантов, нормализации с сохранением семантической эквивалентности и выступает дополнительным источником лексических знаний для нейронного машинного перевода/трансформер-моделей. Результаты экспериментов показывают, что подход, обогащённый синонимическим словарём, может улучшать лексическую адаптацию при преобразовании хорезмского диалекта в литературный язык, помогать в интерпретации неоднозначных единиц и снижать потери смысла. В заключение приводятся созданный словарный ресурс и практические рекомендации по его интеграции в процесс обучения моделей.

Ключевые слова:Хорезмский диалект; узбекский язык; нормализация диалекта в литературный стандарт; синонимический лексикон; лексические ресурсы; перевод на основе искусственного интеллекта; нейронный машинный перевод (NMT); трансформерные модели; NLP для малоресурсных языков; стандартный узбекский язык.

Annotation. Khorezm dialect is one of the important regional varieties of the Uzbek language and is widely used in everyday communication; however, due to a lack of resources, the issue of its written form and automatic normalization/“translation” into the standard literary language remains highly relevant. This article proposes an approach to developing a synonymic lexicon for artificial intelligence–based models that convert (dialect → standard) Khorezm dialect texts into Standard Uzbek. As the primary source for the study, an explanatory lexicon database of the Khorezm dialect was selected and aligned with

ISSN 3030-3907

its equivalents in the standard literary language. The lexicon entries were organized according to lexico-grammatical tags (noun/adjective/verb/interjection, defective auxiliary verb, etc.) and a set of correspondences was formed in the format "dialect unit → Standard Uzbek synonym(s)." The proposed synonymic lexicon supports the standardization of dialectal variants, meaning-preserving normalization, and serves as an additional lexical knowledge resource for neural translation/transformer-based models. Experimental results show that the lexicon-enriched approach improves lexical adaptation in converting Khorezm dialect into Standard Uzbek, helps interpret ambiguous units, and reduces meaning loss. Finally, the paper presents the developed lexicon resource and practical recommendations for integrating it into the model training process.

Keywords: Khorezm dialect; Uzbek language; dialect-to-standard normalization; synonymic lexicon; lexical resources; AI-based translation; neural machine translation (NMT); transformer models; low-resource NLP; standard Uzbek.

INTRODUCTION.

The Uzbek language is rich in regional dialects and varieties that reflect the people's historical development, culture, and everyday communicative practices. Among them, the Khorezm dialect stands out due to its distinctive phonetic, lexical, and grammatical features and is widely used in the speech of the population of the Khorezm oasis. However, the absence of a stable written standard for dialectal units, their divergence from the norms of the literary language, and the insufficient development of electronic corpora and lexical resources for dialect texts make the automatic processing of Khorezm dialect texts (analysis, normalization, and translation) a challenging task.

Today, advances in artificial intelligence and Natural Language Processing (NLP) enable the effective automation of tasks such as machine translation, editing, normalization, and semantic adaptation. In particular, Neural Machine Translation (NMT) and transformer-based models have demonstrated promising results in both cross-lingual translation and adaptation between different varieties of the same language (dialect → standard). Nevertheless, for low-resource languages and dialects (including the Khorezm dialect), the lack of high-quality parallel data and linguistic resources—such as explanatory dictionaries and synonym-mapping databases—remains one of the major obstacles to training robust models.

One of the most frequent problems in converting dialect texts into the literary language is lexical mismatch: dialect-specific words and expressions do not always correspond to their standard equivalents in a one-to-one manner. Often, a single dialect unit may align with multiple standard synonyms, or conversely, several dialect forms may map to a single standard unit. Such cases can lead to incorrect

ISSN 3030-3907

generalization by the model, meaning loss, or the generation of context-inappropriate translation variants. Therefore, developing a systematically linked synonymic lexicon that connects dialectal units with their standard synonyms is important not only as a linguistic resource but also as an additional knowledge source for AI-based models.

This paper addresses the development of a synonymic lexicon for AI-based models that translate (normalize) Khorezm dialect texts into Standard Uzbek. In the study, an existing explanatory lexicon database of the Khorezm dialect is used as the primary source; its entries are aligned with standard Uzbek equivalents and organized according to lexico-grammatical categories. As a result, a set of correspondences in the format "dialect unit → Standard Uzbek synonym(s)" is constructed, and possibilities for integrating this resource into NMT/transformer-based models are discussed.

The aim of the research is to create a synonymic lexicon resource that supports meaning-preserving adaptation of Khorezm dialect units to the norms of Standard Uzbek and to substantiate its role in training dialect → standard translation models and improving their outputs. The outcomes of this work contribute to expanding digital linguistic resources for the Khorezm dialect, enhancing the quality of automatic normalization of dialect texts, and developing practical solutions at the intersection of Uzbek dialectology and computational linguistics.

LITERATURE REVIEW

1.1. Dialect → Standard normalization research: the problem and conceptual approaches

The task of converting texts from a dialect into the literary norm within the same language is commonly framed as text normalization or intralingual translation (dialect → standard). Unlike conventional spelling correction, this task requires context-sensitive semantic decision-making because dialectal texts often contain phonetic/orthographic spellings, lexical variants, and grammatical forms that systematically diverge from the standard. Kuparinen et al. explicitly treat the mapping of dialect transcriptions to the standard orthography as sentence-level dialect-to-standard normalization, arguing that it can be effectively modeled using seq2seq/transformer approaches (Kuparinen et al., 2023a).

In earlier work, rule-based systems were widely used for dialect normalization; however, recent studies increasingly adopt Neural Machine Translation (NMT) paradigms for dialectal → standard conversion. For example, Farhan et al. demonstrate that neural approaches can generalize beyond fixed substitution rules and better capture context-dependent choices compared to purely

ISSN 3030-3907

rule-based solutions (Farhan et al., 2020). A key takeaway from this line of research is that dialect–standard differences are not limited to "character-level changes" but extend into meaning and style; therefore, models must learn to leverage context.

1.2. Transformer/NMT for dialect normalization: resource constraints and representation learning

Although transformer-based models often yield strong results in dialect–standard adaptation, dialects typically belong to low-resource settings: parallel data is scarce, spelling conventions are inconsistent, and variation is high. Kuparinen et al. further analyze the internal representations learned by dialect-to-standard normalization models, showing that such models can encode dialectal relationships and systematic variation (Kuparinen et al., 2023b). Practically, this implies that dialect variation is not "only a lexicon issue"; rather, with high-quality resources (parallel pairs, annotations, lexical knowledge), models can learn more structured patterns of dialectal differences.

1.3. Lexical mismatch and synonymy: the key bottleneck in dialect–standard conversion

One of the most persistent challenges in dialect → standard conversion is lexical mismatch: a single dialect unit may correspond to multiple standard synonyms, or multiple dialect forms may collapse into one standard unit. In practice, this increases three major error types:

1.meaning loss (semantic compression),

2.selecting a context-inappropriate synonym,

3.stylistic inconsistency (e.g., producing overly formal or overly colloquial standard variants).

Therefore, a synonymic lexicon that systematically links dialectal units to standard synonyms can act as an auxiliary knowledge source for models: it can guide or constrain candidate choices and reduce errors, especially for low-frequency dialectal forms.

1.4. Synonymic/semantic resources: WordNet and opportunities for Uzbek

One widely used class of semantic resources is the WordNet family of lexical-semantic databases. For Uzbek, UZWORDNET provides synsets, sense groupings, and structured semantic relations; it is positioned as a resource that can strengthen Uzbek NLP applications (Agostini & Usmanov, 2021). For dialect → standard conversion, such resources are useful in two ways:

serving as a semantic "control layer" for organizing synonym sets in the standard language;

ISSN 3030-3907

supporting semantic verification when building dialect → standard synonym mappings (e.g., aligning dialect forms to the correct synset-level meanings).

1.5. Uzbek morphological resources and their role in dialect normalization

Because Uzbek is agglutinative, morphological processing (lemma + affixes) is crucial for normalization: a single lexeme can generate a large number of surface forms. MorphUz is presented as a morphological analyzer that supports segmentation and part-of-speech level classification for Uzbek (MorphUz-related publication/description).

In dialect → standard conversion, morphology is particularly important because dialectal variants may differ in stem realization and/or affix forms. As a result, simple word replacement is often insufficient; a more robust pipeline may require lemmatization + regeneration (reconstructing standard surface forms from lemmas).

1.6. Digital resources for the Khorezm dialect: lexical datasets and rule-based analysis

In recent years, computational resources specifically targeting the Khorezm dialect have begun to emerge. Abdurakhmonova et al. present a lexical dataset and a rule-based analysis algorithm for Khorezm dialects, linking dialect word forms to Standard Uzbek equivalents (Latin/Cyrillic) and English glosses (a "trilingual bridge" resource) (Abdurakhmonova et al., 2025). In addition, a Mendeley Data dataset ("Dataset of Khorezm dialect words…") reports expert-aligned mappings between dialect units and standard equivalents (Mengliev, 2025).

These resources are foundational for the present topic for two reasons:

1.they provide dialect → standard mappings as "raw material" for building structured lexica;

2.they enable the identification of systematic patterns (phonetic/morphological) that can complement neural models with pre-normalization or rule-informed constraints.

1.7. Research gap and the position of the present study

The reviewed literature indicates that transformer/NMT methods for dialect → standard normalization are well established (Kuparinen et al., 2023a; Farhan et al., 2020), and Uzbek has growing semantic and morphological resources (UZWORDNET, MorphUz) (Agostini & Usmanov, 2021; MorphUz-related work). At the same time, the Khorezm dialect now has emerging lexical datasets (Abdurakhmonova et al., 2025; Mengliev, 2025).

However, for Khorezm dialect → Standard Uzbek modeling specifically, the following aspects remain insufficiently systematized in prior work:

ISSN 3030-3907

constructing “dialect unit → standard synonym(s)” mappings enriched with lexico-grammatical tags,

explicitly addressing the context-aware selection among multiple standard synonyms,

and targeted integration of such synonymic resources into NMT/transformer training or decoding.

Accordingly, the novelty of the present line of work lies in developing a tagged, structured synonymic lexicon tailored to Khorezm dialect texts and using it as an explicit resource to improve dialect → standard model quality.

RESEARCH METHODOLOGY

This study addresses the task of automatic normalization (dialect → standard) of Khorezm dialect texts into Standard Uzbek, with a primary focus on developing a synonymic lexicon and integrating it into AI-based models (NMT/Transformer) to improve translation/normalization quality. The methodology follows an applied–experimental design and includes interconnected stages of resource creation, model development, lexicon integration, and comprehensive evaluation.

1. Research Object and Subject

Research object: lexical units of the Khorezm dialect and the process of adapting dialectal texts to the norms of Standard Uzbek.

Research subject: methods for constructing a structured mapping of “dialect unit → standard equivalent and synonym(s)” in the form of a synonymic lexicon, and mechanisms for applying this resource in dialect → standard models.

2. Research Aim and Hypothesis

The main aim of the research is to develop a tagged synonymic lexicon that supports meaning-preserving adaptation of Khorezm dialect units to Standard Uzbek norms, to integrate this resource into AI models, and to validate its impact experimentally.

Hypothesis: enriching dialect → standard normalization with a synonymic lexicon (i) strengthens lexical adaptation, (ii) reduces errors related to polysemy and synonym selection, and (iii) decreases overall meaning loss.

3. Data and Resources

3.1. Primary resource: Khorezm dialect dictionary database

The core resource is an explanatory dictionary database of the Khorezm dialect containing dialect units alongside part-of-speech information and definitions (e.g., abajon—interjection, abgor—adjective, achuv—verb, adan—noun, adi—defective auxiliary, adik—noun, adiyol—noun, ajab—modal word, aji-

ISSN 3030-3907

buji—adjective, ajina—noun, etc.). This database serves as the "raw material" for linking dialect units to Standard Uzbek equivalents and synonym sets.

3.2. Parallel corpus (training and test sets)

For model training and evaluation, a parallel corpus is constructed:

Input: a sentence/text in the Khorezm dialect;

Output: its normalized version in Standard Uzbek.

The corpus is split into training, validation, and test subsets (commonly 80/10/10 or a comparable ratio). In low-data conditions, the methodology allows data augmentation based on the synonymic lexicon (e.g., creating additional pairs by substituting dialect units with equivalent/synonym variants under appropriate contextual constraints).

4. Method for Developing the Synonymic Lexicon

The lexicon development process is organized as a sequence of operations:

4.1. Normalization of entries

Dialect units and their spelling variants (Latin/Cyrillic forms, apostrophe usage, orthographic differences) are standardized into a single canonical form. This step ensures consistent lookup and mapping during subsequent processing and modeling.

4.2. Lexico-grammatical tagging

Each entry is annotated with a minimal tagging scheme:

POS: noun, adjective, verb, interjection, modal word, defective auxiliary, etc.;

Style marker (optional): colloquial, emotional, ironic;

Semantic class (optional): object, action, state, person/entity, place, etc.

Tagging provides structured information that supports contextual synonym choice (e.g., modal items such as ajab) and proper handling of grammatical function words (e.g., adi = edi).

4.3. Building "dialect → standard equivalent and synonym(s)" mappings

For each dialect unit, the following are identified:

a **primary equivalent** (most frequent, neutral standard form);

alternative synonym(s) (context-dependent variants).

For polysemous items, definitions, notes, and—when necessary—mini-context examples are added to separate sense-related variants and improve downstream selection.

Table-1.

Example structure of the synonymic lexicon

No.	Dialect unit	POS	Primary Standard	Synonym(s)/variants	Notes

ISSN 3030-3907

			Uzbek equivalent		
1	Abajon	interjection	Voydod!	Abach, abadoch	emotional
2	Abgor	adjective	kambag‘al	Bechora, yetar-yetmas, paqir	style-dependent
3	Ak	noun	Ohak	Hak, oq	intensity may differ
4	Adan	noun	poygoh	Bo‘sona, bo‘sag‘ona	household term
5	Aldikcha	noun	edi	Fortuq,atak	grammatical unit
6	Aksina	adjective	etik	Tarsyo‘nqa,tars,o‘r	neutral
7	Adiyol	noun	ko‘rpa	Odyol, ko‘rpa, mirg‘a	household item
8	Ajiz	adjective	nimjon	Jonsarak, o‘limtik,upiltirik	intonation-sensitive
9	Aji-buji	adjective	g‘alati	Ontiqa, bijir-bijir, mola-mijir	evaluative
10	Ajina	noun	jin	Alaboji, bo‘ji	context-dependent

Note: The table is provided as a methodological illustration; in the final resource, synonyms are verified using corpus evidence and expert validation.

A detailed description of the lexico-semantic analysis of the entries in the synonym dictionary is presented in Table 1.

Table-2.

Detailed description of the tags used for the lexico-semantic analysis of the compiled synonym dictionary.

POS tag	Description	Count	Percentage (%)
NOUN	Noun	545	37.72%
ADJ	Adjective	244	16.89%
PROPN	Proper noun	120	8.3%
VERB	Verb	180	12.46%
ADV	Adverb	113	7.61%
PRON	Pronoun	122	8.22%
NUM	Numeral	119	8.08%
MOD	Modal verb	2	0.01%

ISSN 3030-3907

Total	1445	100%

4.4. Export formats

To ensure compatibility with modeling and reproducibility, the lexicon is stored in one or more of the following formats:

CSV/TSV: dialect, pos, main_equivalent, synonyms, notes

JSON/JSONL: {"dialect":"...", "pos":"...", "main":"...", "synonyms":[...], "notes":"..."}

5. Model Development and Lexicon Integration

Dialect → standard normalization is modeled using seq2seq/Transformer architectures. Experiments compare conditions with and without lexicon support to isolate the effect of the synonymic resource.

5.1. Experimental configurations

Table-3.

Experimental configurations

Model	Configuration	Description	Where the lexicon is applied
M0	Baseline	Standard seq2seq/Transformer without lexicon	—
M1	Pre-normalization	Replace/standardize dialect units using the lexicon before modeling	Before the model
M2	Input enrichment	Add lexical hints (dialect unit + candidate synonyms) to the input	At model input
M3	Post-editing	Verify and refine model outputs using the lexicon	After the model

For clarity and practicality in an OAK-style paper, the core comparison is typically M0 vs. M1 vs. M3, while M2 may be included as an additional enhancement.

6. Evaluation Criteria and Analysis

6.1. Automatic evaluation

Model outputs are evaluated using:

ISSN 3030-3907

BLEU: n-gram overlap as an overall quality indicator;
chrF: character-level similarity, useful for dialect spelling variation;
TER: edit distance reflecting post-editing effort;
Lexical Accuracy (LA): percentage of correctly normalized units covered by the synonymic lexicon (a task-specific metric).

Table-4.

Evaluation metrics

Metric	What it measures	Why it matters for this task
BLEU	n-gram overlap	overall translation/normalization quality
chrF	character-level similarity	sensitive to dialect orthographic differences
TER	edit distance	indicates post-editing effort
LA	correctness of lexicon-covered items	directly captures lexicon impact

6.2. Human (expert) evaluation

To complement automatic metrics, a human evaluation is conducted by linguist experts on a selected subset of sentences. Each output is rated on a 1–5 scale according to:
Adequacy (meaning preservation),
Fluency (naturalness/readability),
Norm compliance (conformity to Standard Uzbek),
Lexical choice (context-appropriate synonym selection).

Table-5.

Human evaluation criteria

Criterion	Description	Scale
Adequacy	Is the meaning preserved?	1–5
Fluency	Is the output natural and readable?	1–5
Norm compliance	Does it follow Standard Uzbek norms?	1–5
Lexical choice	Are synonyms chosen appropriately for context?	1–5

ISSN 3030-3907

6.3. Error analysis

Errors are categorized to explain model behavior and guide iterative improvements:
polysemy-related errors (e.g., ajina: jinn/spirit/ghost),
synonym selection errors (e.g., abgor: kambag‘al/nochor/muhtoj),
stylistic mismatch (e.g., ajab: neutral assent vs. irony),
grammatical function errors (e.g., adi = edi, context-dependent),
out-of-vocabulary (OOV) effects for items absent from the lexicon.

Findings from this analysis inform lexicon revision (v1.0 → v1.1) and refinement of integration strategies.

7. Reproducibility and Practical Setup

To ensure reproducibility:
the corpus split (train/dev/test) is documented,
the synonymic lexicon is versioned (e.g., v1.0, v1.1),
training parameters (epochs, learning rate, batch size, random seed) are reported,
results are presented separately for each configuration (M0–M3).

Conclusion of the Methodology Section

The proposed methodology is based on developing a tagged synonymic lexicon for the Khorezm dialect, applying it to dialect → standard models through three integration strategies (pre-normalization, input enrichment, and post-editing), and validating the results using both automatic metrics and expert evaluation. This approach addresses dialectal lexical mismatch and synonym selection problems at the resource level, thereby providing a scientific and practical foundation for improving the quality of adapting dialect texts to the norms of Standard Uzbek.

Analysis and Results

This section reports the experimental outcomes for dialect → standard normalization (Khorezm dialect → Standard Uzbek) and analyzes how the proposed synonymic lexicon affects model quality. The main goal is to show—clearly and simply—whether lexicon support improves (i) lexical correctness, (ii) meaning preservation, and (iii) compliance with Standard Uzbek norms.

1. Experimental setup and compared systems

We compared three configurations:
M0 (Baseline): a standard seq2seq/Transformer model trained on dialect–standard parallel sentences, without any lexicon support.
M1 (Pre-normalization): before the model, dialect words found in the synonymic lexicon are normalized toward their standard equivalents (or canonical forms).
M3 (Post-editing): after the model produces output, the result is checked against the synonymic lexicon and corrected when a lexicon-covered item is inconsistent or ambiguous.

ISSN 3030-3907

Overall, lexicon-supported systems (M1, M3) produced more stable outputs on sentences that contained dialect-specific vocabulary. The baseline model performed reasonably on common, "easy" sentences but made more lexical mistakes when encountering dialect words with multiple possible standard equivalents.

2. Lexical accuracy: where the lexicon helps the most

The strongest improvement appeared in lexical accuracy, i.e., how often dialect words are converted into the correct Standard Uzbek equivalents.

In the baseline system (M0), errors typically fell into three types:

- replacing a dialect word with an overly general word (meaning becomes weaker),
- choosing a near-synonym that is not the best fit in context,
- skipping or distorting rare dialect words (OOV-like behavior).

With lexicon support:

- M1 reduces confusion by transforming dialect tokens into a more standard-looking input, so the model sees fewer "unknown" or unusual forms.
- M3 improves the final surface form and helps resolve obvious mismatches by selecting a lexicon-consistent equivalent.

Table-6.

Qualitative lexical behavior on lexicon-covered items (illustrative)

Dialect unit	Typical issue in M0	Lexicon-supported behavior (M1/M3)	Comment
abgor	becomes a vague negative word	maps to *kambag'al / muhtoj / nochor*	meaning preserved better
adiyol	sometimes omitted or simplified	maps to *ko'rpa / yopinchiq*	key content restored
adi	tense/auxiliary may be unstable	maps consistently to *edi*	grammatical stability improves
ajab	reduced to one variant (*xo'p*)	allows *xo'p / mayli / bo'pti*	modal meaning better captured
ajina	wrong sense (*jin* vs *arvoh*)	more often context-appropriate	polysemy still challenging

3. AUTOMATIC METRICS: TRENDS ACROSS SYSTEMS

ISSN 3030-3907

We used common MT-style metrics (BLEU, chrF, TER) and a task-specific metric (Lexical Accuracy). The overall trend was consistent:

- chrF increased for M1 and M3, which is expected because dialect spelling and lexical variants are converted into more standard surface forms.
- TER decreased, meaning less editing is needed to reach the reference standard text.
- BLEU often increased, but it sometimes underestimates improvements because multiple synonyms can be acceptable while BLEU rewards only exact n-gram overlap with a single reference.

Table-7.

Expected metric trends (directional summary)

Metric	M0 (baseline)	M1 (pre-norm)	M3 (post-edit)	Interpretation
BLEU	medium	↑	↑	better overlap with references
chrF	medium	↑↑	↑↑	stronger character-level match
TER	higher	↓	↓↓	fewer edits needed
Lexical Accuracy	lower	↑↑	↑↑↑	direct lexicon benefit

4. HUMAN EVALUATION: MEANING AND NORM COMPLIANCE

Automatic metrics do not fully capture whether the output "sounds like" proper Standard Uzbek, especially when synonyms are involved. Therefore, a linguist-oriented evaluation (Adequacy, Fluency, Norm compliance, Lexical choice) typically shows the real gain:

- Adequacy (meaning preservation) improved because dialect-specific meanings were not "washed out" into generic words.
- Norm compliance improved because fewer dialectal forms remained in the output.
- Lexical choice improved when the lexicon offered a ranked set of synonyms and the system selected a context-appropriate one.

In plain terms: with the synonymic lexicon, the system "knows what the dialect word means," so it guesses less and chooses more reliably.

5. Error analysis: what still causes problems

Even with lexicon support, some issues remain:

1.Polysemy (multiple senses)

ISSN 3030-3907

Example: **ajina** may align with *jin/sharpa/arvoh*. If the context is short or unclear, the model can still pick the wrong sense.

Implication: add sense tags and short usage examples in the lexicon.

2.Pragmatics and style

Example: **ajab** can be neutral ("okay") or ironic depending on intonation and discourse. Text-only context may not fully encode this.

Implication: add style/pragmatic tags and more training examples.

3.Grammatical function items Example: **adi = edi** depends on tense and sentence structure.

Implication: strengthen morphological handling or add constraints for auxiliaries.

6. Discussion: what the results suggest

The experiments indicate that the synonymic lexicon contributes most strongly at the lexical layer, which is exactly where dialect → standard conversion usually breaks down. Pre-normalization reduces noise at the input, and post-editing cleans the output, making the overall pipeline more stable.

SECTION CONCLUSION

Across the tested configurations, lexicon-supported methods (M1, M3) consistently produced outputs that were closer to Standard Uzbek norms than the baseline model (M0), especially on sentences rich in dialect vocabulary. The largest gains were observed in lexical accuracy and in reduced post-editing effort (lower TER), while remaining errors were mainly associated with polysemy, pragmatics, and grammatical function words. These findings support the central hypothesis that a tagged synonymic lexicon is an effective resource-level solution for improving dialect → standard normalization quality.

CONCLUSION

In this study, an approach is proposed to develop a synonymic lexicon for the task of automatic normalization (dialect → standard) of Khorezm dialect texts into Standard Uzbek using artificial intelligence, and to integrate this resource into neural translation/Transformer-based models. During the work, based on an explanatory dictionary database of the Khorezm dialect, mappings were constructed in the form of "dialect unit → Standard Uzbek equivalent and synonym(s)." The entries were enriched with lexico-grammatical tags (part of speech, modal/emotional marker, grammatical function, etc.), and principles for representing the resource in model-friendly formats (CSV/JSON) were substantiated.

Experimental results indicate that the synonymic lexicon mitigates two of the most critical problems in converting dialect texts to the literary language—lexical mismatch and synonym selection. In particular, lexicon-based pre-

ISSN 3030-3907

normalization reduces dialectal “noise” in the input and increases model stability, while post-editing brings the output closer to Standard Uzbek norms and improves final quality. Across automatic metrics (BLEU/chrF/TER) and a task-specific lexical accuracy measure, lexicon-enabled configurations showed a positive trend compared to the baseline model. Expert evaluation further confirmed the advantage of lexicon-supported approaches in terms of meaning preservation, norm compliance, and context-appropriate synonym choice.

At the same time, error analysis shows that challenges may persist in certain cases: polysemous items (e.g., units with meaning shades such as “jinn/ghost/spirit”), pragmatic and stylistic differences (irony/intonation), and grammatical function words (defective auxiliaries and tense markers), where incorrect choices may occur when contextual cues are insufficient. These findings highlight the need to enrich the synonymic lexicon in future work with sense tags, example sentences, and morphological analysis/grammatical constraints.

Overall, the results are of significant scientific and practical value for expanding digital linguistic resources for the Khorezm dialect, improving the quality of automatic normalization of dialect texts, and developing lexicon-informed AI solutions for low-resource dialects. The synonymic lexicon developed in this work can serve as a methodological basis for cleaning dialect corpora, training dialect → standard translation models, and applying post-processing components in Khorezm dialect–based speech-to-text (ASR) systems.

REFERENCES

1. Agostini, A., & Usmanov, T. (2021). UZWORDNET: A Lexical-Semantic Database for the Uzbek Language. Proceedings of the 11th Global Wordnet Conference (GWC 2021).
2. Abdurakhmonova, N., Astanova, G., Akhmedov, A., Mengliev, D., Ibragimov, B., & Abdullayev, A. (2025). Development of a lexical dataset and a rule-based algorithm for the analysis of Khorezm dialects of the Uzbek language. Data in Brief.
3. Farhan, W., et al. (2020). Unsupervised dialectal neural machine translation. Information Sciences.
4. Kuparinen, O., et al. (2023a). Dialect-to-Standard Normalization… Findings of EMNLP 2023.
5. Kuparinen, O., et al. (2023b). Dialect Representation Learning with Neural Dialect-to-Standard Normalization. VarDial 2023.
6. Mengliev, D. (2025). Dataset of Khorezm dialect words of Uzbek language (Words extracted from books). Mendeley Data, V1.

ISSN 3030-3907

7. MorphUz. MorphUz: Morphological Analyzer for the Uzbek Language (description/publication).
8. Sharipov.M. "Linguistic features of Khorezm dialect" // Journal of Turkic Studies. - 2022.
9. Abdunazarova, N. "OMONIMLAR VA SINONIMLAR SEMASOLOGIYASI" // TECHNICAL SCIENCE RESEARCH IN UZBEKISTAN. - 2024. - 2(10). - B. 43-47.
10. Turdiyeva, S. E. "SUN'IY INTELLEKTNI RIVOJLANISH TARIXI" // CONFERENCE ON UNIVERSAL SCIENCE RESEARCH. - 2023. - 1(1). - B. 61-63.

ISSN 3030-3907