ISSN:2828-6235
Impact factor:13.77

International Journal of Literature and Language Studies(IJLLS)

ARTIFICIAL INTELLIGENCE-BASED DIGITAL MAPPING OF TURKIC TOPONYMS IN THE BUKHARA REGION: A LINGUISTIC AND GEOSPATIAL APPROACH

Yarashova Shahrizoda Alisher qizi
Bukhara State Universiti PhD Student
E-mail: sh.yarasheva@gmail.com
ORCID: https://orcid.org/0009-0009-8217-0380
Doi: https://doi.org/10.5281/zenodo.24403435

Abstract: Geographical names preserve linguistic, historical, and cultural information, yet toponymic data are often dispersed across maps, archives, scholarly publications, and local records. This study proposes an interdisciplinary framework for documenting, classifying, and digitally mapping Turkic toponyms in the Bukhara region of Uzbekistan. The framework integrates linguistic analysis, historical source documentation, semantic classification, geospatial referencing, and AI-assisted processing. Toponyms are treated as multidimensional entities whose linguistic structure, meaning, historical forms, documentary evidence, and location should be recorded together. AI is considered a support tool for variant identification, preliminary classification, pattern detection, and information retrieval, rather than an authority for determining etymology. The study outlines a database structure and workflow that can facilitate the creation of searchable, evidence-based toponymic resources and support the preservation of regional linguistic heritage. The framework remains conceptual and requires empirical testing with a verified corpus of geographical names.

Keywords: artificial intelligence; Turkic toponyms; digital toponymy; Bukhara region; geospatial mapping; GIS; linguistic heritage.

1. Introduction

Toponyms are linguistic units closely connected with geographical space, historical memory, and cultural identity. They may preserve information about settlement, natural features, social organization, ethnic relations, and historical linguistic contact. Their study therefore requires methods from linguistics, geography, history, and ethnography. Traditional toponymic research relies on fieldwork, archival materials, maps, linguistic comparison, and documentary evidence. These methods remain essential, but the growing volume and dispersion of toponymic information create opportunities for digital organization and comparative analysis.

Uzbekistan's geographical names reflect contact among Turkic, Persian, Arabic, and other linguistic traditions. The Bukhara region is especially relevant for studying such layers, as its toponyms may preserve information about settlement history, natural features, economic activity, and local cultural processes. Research on Uzbek toponyms and Bukhara's ethnonym-based names has highlighted structural, semantic, and linguacultural aspects of this heritage. However, linguistic descriptions, historical records, and geographical data are often maintained separately, limiting systematic comparison and spatial analysis.

GIS can support the spatial representation of place names, while artificial intelligence may facilitate text processing, variant matching, preliminary classification, and retrieval. Yet automated etymological interpretation requires caution: linguistic similarity alone cannot establish a name's origin. This study proposes an integrated framework for AI-assisted digital mapping of Turkic toponyms in Bukhara, treating each place name as a structured record linking linguistic description, historical evidence, and geographical location.

Research aim and questions

The aim is to develop a framework for digital documentation, linguistic classification, and geospatial mapping of Turkic toponyms in Bukhara through the integration of AI and GIS. The study asks: (1) which linguistic, historical, semantic, and geographical attributes should be recorded; (2) how AI can support identification and preliminary classification; (3) how linguistic and spatial data can be integrated; and (4) what limitations must be addressed when using AI for toponymic interpretation?

2. Literature Review

Toponymic research examines geographical names through phonetic, morphological, lexical, semantic, etymological, and historical dimensions. Place names may contain elements associated with natural features, settlements, ethnic groups, personal names, occupations, or cultural practices. Because names may undergo sound change, borrowing, reinterpretation, and administrative transformation, their meanings cannot always be inferred from contemporary forms alone.

Studies of Uzbek toponyms have considered structural and lexical features, including linguistic components of different origins. Research on Bukhara ethnoponyms has examined names formed from ethnonyms and Uzbek formants, illustrating the importance of morphological analysis. Geographical approaches have also emphasized classification, standardization, registers, transliteration, and thematic mapping. These findings suggest that a digital resource should preserve more than a name and coordinate: it should record linguistic structure, variants, semantic interpretation, sources, and uncertainty.

AI and digital humanities offer methods for identifying names in texts, detecting spelling variants, grouping similar forms, extracting information, and organizing large datasets. The methodological gap addressed here is the integration of linguistic, historical, and geospatial attributes within one structured resource, with AI-assisted processing subject to expert verification.

3. Materials and Methods

The proposed research design combines source collection, linguistic analysis, semantic classification, geospatial referencing, AI-assisted processing, and validation. The corpus may be assembled from maps, administrative registers, historical documents, toponymic dictionaries, scholarly studies, field records, interviews, and reliable digital resources. Each record should identify its source so that documented facts can be distinguished from interpretations. Historical and dialectal variants may be included when their evidence and status are recorded.

Linguistic and semantic analysis

Each name may be examined for orthographic variants, lexical components, morphological structure, language affiliation, and etymological evidence. Morphological descriptions may distinguish simple, derived, compound, and complex forms, where applicable. Etymological interpretations should be recorded as verified, probable, disputed, or unknown rather than treated as definitive without evidence.

A flexible semantic classification may include hydronymic, oronymic, oikonymic, ethnotoponymic, anthroponymic, occupational, descriptive, and historical-cultural categories. Multiple labels may be assigned to one name where justified.

Geospatial and AI processing

Geospatial processing links each name to a geographical object, coordinates, and administrative unit, then visualizes the records in GIS. Spatial uncertainty should be documented, especially for

historical or disappeared objects. AI may support named-entity recognition, spelling-variant matching, morphological pattern detection, semantic suggestions, similarity analysis, information extraction, and search. These outputs remain preliminary and require linguistic, source, and geographical validation.

4. Proposed Database and Results Framework

The proposed database represents each toponym as a multidimensional record connecting linguistic, geospatial, historical, semantic, documentary, and computational information. This design enables researchers to compare names across categories, retrieve related records, and map patterns without separating the name from its evidence.

Dimension	**Information recorded**
Linguistic	Lexical elements, morphology, language affiliation, variants
Geospatial	Coordinates, object type, administrative location
Historical	Attested forms, dates, historical name
Semantic	Meaning, naming motivation, category
Documentary	Source, evidence, reference, reliability
Computational	AI suggestions, similarity, processing and review status

Suggested database fields include a unique identifier, official and historical names, variants, object type, administrative unit, coordinates, linguistic affiliation, lexical components, morphological structure, semantic labels, etymological hypothesis and status, historical attestations, source, AI-generated suggestions, expert review status, and evidence level.

The framework may support queries about names containing specific Turkic elements, shared formants, semantic categories by district, historical variants, disputed etymologies, or records requiring further field verification. GIS visualization may produce linguistic, semantic, historical, administrative, and interactive maps. These are proposed analytical outputs, not completed results.

5. Discussion

The framework extends traditional toponymic documentation by enabling systematic comparison and spatial analysis across a structured corpus. While traditional linguistic description can explain an individual name, a digital resource can connect it with numerous other records, reveal recurrent patterns, and support retrieval of related evidence. The aim is not to replace fieldwork or archival analysis but to make their findings easier to organize, compare, and interpret.

AI's role must remain carefully defined. Automated systems may detect formal similarities or suggest categories, but these patterns do not necessarily establish shared etymology. Similar forms may result from coincidence, borrowing, sound change, or distinct historical processes. Separating machine-generated suggestions from expert-approved interpretations protects the reliability and transparency of the database.

Several limitations require attention: incomplete documentation, uncertain etymologies, AI-generated errors, variant-name complexity, uncertain historical coordinates, and inconsistent spelling or transliteration across sources. Future work should compile a verified corpus, develop

annotation standards, test entity recognition and variant matching, evaluate semantic classification against expert annotation, integrate GIS, and assess the accuracy and usefulness of the resulting digital map.

6. Conclusion

This study proposes an interdisciplinary framework for AI-assisted digital mapping of Turkic toponyms in the Bukhara region. It combines data collection, linguistic analysis, semantic classification, geospatial referencing, AI-assisted processing, and expert validation. Its main methodological contribution is the representation of place names as structured linguistic-geospatial entities rather than isolated map labels. Such a model may facilitate comparative analysis, pattern identification, information retrieval, and thematic visualization while supporting the preservation of regional linguistic heritage.

AI can assist with data processing and generate hypotheses, but etymological and historical interpretation must remain evidence-based and subject to scholarly verification. The proposed framework provides a basis for developing a digital toponymic database; its practical effectiveness must be demonstrated through corpus construction, computational testing, and empirical evaluation.

References

1. Abbasi, S., Asloniy, M. R., & Almasi, K. (2024). A look at the place names of Uzbekistan from a linguistic point of view. Oriental Journal of Philology, 4(03), 317–324. https://doi.org/10.37547/supsci-ojp-04-03-41
2. Kobilova, A. B., & Rakhmatova, A. K. (2025). Linguacultural features of toponymic units in the example of Bukhara region. The Lingua Spectrum, 3.
3. Mirakmalov, M. T., Ibragimova, R. A., Avezov, M. M., & Okhunjonova, D. K. (2023). Physical and geographical features of the toponyms of Uzbekistan. IOP Conference Series: Earth and Environmental Science, 1284, 012004. https://doi.org/10.1088/1755-1315/1284/1/012004
4. Safarova, M. (2025). Ethnotoponyms with Uzbek formants. International Journal of Artificial Intelligence, 5(05), 2640–2642. https://doi.org/10.71337/inlibrary.uz.ijai.108178