

e-ISSN 3060-5059

DOI: 10.67227



IJTIMOIIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNAL

Scientific and Innovative Research in
Social and Humanitarian Studies

Vol. 3 • Issue 10

2026



Founded in 2024



publishscience.uz



publish_science_journal@mail.ru



PEER REVIEWED • OPEN ACCESS

RAQAMLI KOMMUNIKATSIYA SHAROITIDA MASHINA TARJIMASI VA O'ZBEK TILINING LINGVISTIK IDENTITETI

Baxromova Madina Yaxshibay qizi

Toshkent amaliy fanlar universiteti, magistr.

e-mail: madinabaxromova2599@gmail.com ORCID: <https://orcid.org/0009-0001-4283-5589>

Annotatsiya.

Mashina tarjimasi raqamli kommunikatsiyada o'zbek tilidan foydalanish imkoniyatlarini kengaytirishi mumkin, biroq avtomatlashtirilgan tarjima natijalarining mavjudligi o'z-o'zidan lingvistik identiklikning saqlanishini anglatmaydi. Ushbu maqolada tarjima texnologiyalari grammatik, leksik va madaniy ifoda imkoniyatlarini cheklamagan holda tilning funksional qo'llanilishini qanday sharoitlarda qo'llab-quvvatlashi mumkinligi o'rganiladi. Tadqiqotda o'zbek va turkiy tillarning lingvistik resurslari, mintaqaviy tarjima infratuzilmasi hamda nisbatan kichik til jamoalari uchun ishlab chiqilgan til texnologiyalariga oid ilmiy ishlarning maqsadli tanlangan narrativ sharhidan foydalaniladi. Konseptual sintez asosida raqamli muhitda mavjudlik, lingvistik reprezentatsiya va til egalari qabul qilish vakolati o'rtasidagi farqlar ajratib ko'rsatiladi. Maqolada ushbu uch jihatni o'zaro bog'liq holda ko'rib chiqish zarurligi asoslanadi: tarjima qilingan kontent hajmining ortishi, agar tarjima natijalari ishonchsiz bo'lsa yoki til egalari o'z tillarini ifodalash uchun foydalaniladigan resurslarga ta'sir ko'rsata olmasa, cheklangan ahamiyat kasb etadi. Maqolada korpus xilma-xilligi, ma'lumotlarning kelib chiqishi, terminologik rivojlanish va til jamoasi tomonidan ekspert baholash jarayonlarini o'zaro bog'lovchi lingvistik resurslarni boshqarish modeli taklif etiladi. Tadqiqotning ilmiy hissasi til identikligining saqlanishiga oid miqdoriy jihatdan o'lgan natijalarni da'vo qilishdan ko'ra, raqamli inklyuziya til identikligini qo'llab-quvvatlashi mumkin bo'lgan mexanizmlarni aniqlashtirishdan iborat.

Kalit so'zlar: mashina tarjimasi; o'zbek tili; til identikligi; raqamli inklyuziya; korpuslarni boshqarish; lingvistik xilma-xillik; resurslari cheklangan tillar.

МАШИННЫЙ ПЕРЕВОД И ИДЕНТИЧНОСТЬ УЗБЕКСКОГО ЯЗЫКА В ЦИФРОВОЙ КОММУНИКАЦИИ

Бахромова Мадина Яхшибай кизи

Ташкентский университет прикладных наук, магистрант.

Аннотация.

Машинный перевод способен расширить использование узбекского языка в цифровой коммуникации, однако само по себе наличие результатов автоматизированного перевода не свидетельствует о сохранении языковой идентичности. В данной статье рассматриваются условия, при которых переводческие технологии могут обеспечивать полноценное функциональное использование языка без ограничения грамматических, лексических и культурных возможностей выражения. В исследовании применяется целенаправленный нарративный обзор научных работ, посвящённых языковым ресурсам узбекского и тюркских языков, региональной переводческой инфраструктуре, а также языковым технологиям, разработанным для сравнительно небольших языковых сообществ. На основе концептуального синтеза разграничиваются цифровая доступность, языковая репрезентация и полномочия носителей языка в принятии решений. Обосновывается необходимость комплексного рассмотрения этих трёх аспектов: увеличение объёма переведённого контента имеет ограниченную ценность, если результаты перевода ненадёжны или носители языка не могут влиять на ресурсы, используемые для представления их языка. В статье предлагается модель управления языковыми ресурсами, объединяющая разнообразие корпусов, происхождение данных, развитие терминологии и экспертную оценку со стороны языкового сообщества. Научный вклад исследования заключается в определении механизмов, посредством которых цифровая инклюзия может способствовать поддержанию языковой идентичности, а не в утверждении о наличии эмпирически измеренного эффекта её сохранения.

Ключевые слова: машинный перевод; узбекский язык; языковая идентичность; цифровая инклюзия; управление корпусными ресурсами; языковое разнообразие; языки с ограниченными ресурсами.

MACHINE TRANSLATION AND UZBEK LANGUAGE IDENTITY IN DIGITAL

Baxromova Madina Yaxshibay qizi

Tashkent University of Applied Sciences, Master's Student.

Abstract.

Machine translation can expand the use of Uzbek in digital communication, but the availability of automated output alone does not demonstrate the preservation of linguistic identity. This article examines the conditions under which translation technology can support functional participation without narrowing grammatical, lexical, and cultural expression. It uses a purposive narrative review of research on Uzbek and Turkic language resources, regional translation infrastructure, and language technologies developed for smaller linguistic communities. The conceptual synthesis distinguishes digital availability, linguistic representation, and speaker authority. It argues that these dimensions must be addressed together: more translated content offers limited value when output is unreliable or when speakers cannot influence the resources used to represent their language. The article proposes a resource-governance framework connecting corpus diversity, provenance, terminological development, and community review. Its contribution is to specify mechanisms through which digital inclusion may support language identity, rather than to claim a measured preservation effect.

Keywords: machine translation; Uzbek; language identity; digital inclusion; corpus governance; linguistic diversity; low-resource languages.

Introduction. The possibility of using a language online depends partly on the tools available to produce, retrieve, and translate its content. Machine translation can reduce the work required to communicate across linguistic boundaries and can make information accessible in a language that users prefer. For Uzbek, this opportunity raises a question that exceeds model accuracy: how can automated translation enlarge the language's digital functions while maintaining the range of meanings and forms through which its speakers express themselves?

The issue should not be framed as a choice between technology and tradition. A language can acquire new terminology and communication practices without losing continuity. Conversely, producing more text in that language does not automatically guarantee a stronger position. Repetitive, poorly edited output may increase volume while offering users a restricted or misleading representation of how the language works.

This article therefore distinguishes participation from representation. Participation concerns the ability to use Uzbek in digital activities. Representation concerns the quality and diversity of the Uzbek that systems produce and learn from. The research question is which resource and governance conditions allow machine translation to strengthen both. The argument is that corpus quality, speaker involvement, and opportunities for correction are necessary links between technological availability and meaningful linguistic inclusion.

LITERATURE REVIEW

Research relevant to this question follows several different routes. Li et al. (2016) describe manually aligned Uzbek–English and Turkish–English resources at the morpheme level. Their work provides direct evidence that linguistic annotation can address structural distinctions obscured by word-level alignment. Tukeyev et al. [2026] examine the adaptation of open-source models for several Turkic languages, including Uzbek. These studies locate technological development within specific linguistic resources rather than treating language coverage as sufficient in itself.

A regional contribution is provided by Mamasaidov and Shopulatov [2024], who develop translation resources and open models for Karakalpak, including an Uzbek–Karakalpak language pair. The relevance to Uzbek is infrastructural: regional communication need not always be organised exclusively around a globally dominant language. The study concerns Karakalpak and should not be recast as a direct evaluation of Uzbek linguistic identity.

Other research explicitly connects translation tools with language maintenance. Prieto et al. [2024] consolidate parallel resources for four Colombian Indigenous languages and build baseline systems. Rikters et al. [2022] examine translation for Livonian in a very small speaker context. These cases support considering translation resources as components of a wider language ecosystem. They do not establish that the same interventions have equivalent effects in Uzbek.

The unresolved conceptual issue is how to move from the existence of a model or corpus to a claim about identity. Technical studies can show resource construction or translation capability without measuring changes in speaker practice, cultural continuity, or language-domain retention. The present article addresses that distinction through a framework of conditions and mechanisms. It does not present resource development as empirical proof of preservation.

METHODOLOGY

This study is a narrative literature review and conceptual synthesis. Its source set was selected purposively from the research identified in the underlying topic: direct Uzbek resource development, contemporary Turkic machine translation, regional multilingual infrastructure, and translation projects explicitly connected with language maintenance. Publication metadata and central research descriptions were checked against publisher or scholarly repository records. The five studies cited in the review constitute a focused set rather than an exhaustive account of the field.

The synthesis compares each contribution according to the resource it develops, the language context it addresses, and the kind of conclusion its evidence supports. Findings about datasets and models are distinguished from the present article's interpretation of their relevance to language identity. No pooled effect size, benchmark ranking, systematic-

review screening count, or original interview result is produced.

The analysis uses three categories: digital availability, linguistic representation, and speaker authority. These categories form the article's proposed framework. They are not variables already measured across the cited studies. The approach is suitable for clarifying research questions and institutional priorities, while its purposive selection limits claims about the completeness of the literature.

LANGUAGE IDENTITY AS FUNCTIONAL AND EXPRESSIVE CAPACITY

Language identity is defined here as the connection between a community's use of a language and its ability to express socially meaningful distinctions through that language. This working definition includes grammatical resources, terminology, registers, and culturally situated meanings. It does not imply that a national community possesses a single voice or that all change is undesirable.

Functional capacity concerns where Uzbek can be used: for learning, administration, professional discussion, cultural production, and everyday communication. Expressive capacity concerns what speakers can say in those settings and how precisely they can say it. Translation technology may improve the first while falling short on the second. A website can offer an Uzbek version whose unclear terminology makes it difficult to use.

Speaker authority concerns who can judge those limitations and influence their correction. If a tool is available but its training resources, terminology choices, and correction procedures are inaccessible, speakers have little control over recurring distortions. This is a governance issue as well as a linguistic one. The framework therefore evaluates language support through usability, representation, and opportunities to contest the system's decisions.

DIGITAL AVAILABILITY AND ITS LIMITS

Machine translation can support access in two directions. Translation into Uzbek may help readers encounter material that would otherwise require proficiency in another language. Translation from Uzbek may help writers reach international audiences while composing in their preferred language. These possibilities create a practical reason to invest in translation tools without requiring a claim that the tools alone preserve identity.

Whether availability results in sustained use depends on quality and context. A translation that readers distrust may be bypassed. A tool that cannot handle a relevant domain may leave its users dependent on another language for that domain. A technically strong resource may also remain unused if it is difficult to access or poorly integrated into everyday workflows.

Consequently, preservation-oriented evaluation needs evidence beyond the number of supported language pairs. Relevant future indicators include repeated use, the range of activities completed in Uzbek, editorial correction burdens, and users' judgements of comprehension. These are proposed indicators, not outcomes observed in this article. They would help distinguish a model's formal language coverage from its contribution to actual communicative practice.

CORPUS DIVERSITY AND LINGUISTIC REPRESENTATION

A model learns from a selected body of language rather than from the language in its entirety. For Uzbek translation development, this makes corpus composition a consequential decision. Official prose, journalism, instructional material, conversation, and literary writing present different registers and communicative expectations. A resource dominated by one category may perform acceptably in that category while providing weak support elsewhere.

Corpus expansion should therefore be accompanied by documentation of genre, origin, date, script, and translation status where those fields are available. Human-authored originals, professional translations, and synthetic translations should be identifiable. Keeping these distinctions visible enables developers to investigate whether particular sources are associated with recurrent errors rather than treating all text as interchangeable.

The implications of Li et al. [2016] also extend to annotation priorities. A resource that represents morphological relationships can make language-specific structure visible to research and evaluation. The conceptual inference is that structural fidelity is one component of reliable representation. It does not follow that morpheme annotation is the optimal training method for every modern model; that question requires controlled comparisons.

Representation should also allow variation. A preservation framework that accepts only one narrow register risks excluding legitimate uses of Uzbek. Standardisation is useful where consistent meaning is required, but it should not erase differences among conversational, literary, professional, and regional practices. Corpus documentation can make these differences examinable rather than invisible.

TERMINOLOGY AND CULTURAL MEANING

Terminology links translation quality to the capacity to discuss new realities in Uzbek. A preservation-oriented approach should ask whether a term is intelligible, established in the relevant domain, and consistent with the intended audience. The existence of a foreign element does not, by itself, prove a loss of identity. Borrowing, explanation, and the development of new expressions can all be appropriate responses to new concepts. The practical problem is arbitrary inconsistency. If the same institutional or technical concept is translated differently across connected documents, readers may struggle to recognise it. A machine-readable terminology resource can record the preferred equivalent, the relevant field, alternative forms, and circumstances in which explanation is needed. Such a resource should remain revisable as actual usage changes. Culture-specific meanings require additional care. A term such as mahalla can denote a neighbourhood, a community, or an institutionally specific arrangement, depending on context. A translation decision should preserve the relevant distinction rather than assume one universal equivalent. This is an illustrative linguistic example, not a recorded error from a commercial system. The direction of translation matters. An Uzbek reader may need an idiomatic explanation of a foreign institution, while an international reader may need an Uzbek term retained with a brief gloss. These choices connect preservation with audience understanding. Keeping every word unchanged is no more reliable than replacing every unfamiliar concept with a broad approximation.

A FRAMEWORK FOR RESOURCE GOVERNANCE

The proposed framework begins with documented resource collection. Universities, publishers, translators, and language specialists can identify texts suitable for reuse and record their provenance and permitted uses. Collection should prioritize missing genres and domains rather than volume alone. A corpus release should make its coverage and exclusions clear.

The second component is linguistic curation. Review should check alignment, meaning, orthography, and domain terminology, with sampling procedures matched to the intended application. Corrections should be linked to the affected source records. This makes curation cumulative: future releases can preserve verified improvements rather than repeating the same editorial work.

The third component is speaker participation with a defined route to action. Consultation has limited value if reviewers' observations never change the dataset, terminology resource, or output. Participants need a way to report problems, receive a decision, and see the resulting revision. Disagreements should be documented, especially where several expressions are acceptable.

The fourth component is longitudinal assessment. Institutions should distinguish short-term translation improvements from longer-term effects on use. A revised model may reduce a class of errors, while evidence about identity requires observation of communication practices and users' experiences over time. The framework thus connects technical maintenance with a research agenda rather than asserting an already demonstrated cultural outcome.

DISCUSSION AND LIMITATIONS. The synthesis yields three conceptual propositions. First, availability creates an opportunity for inclusion but does not establish successful use. Second, reliable representation depends on the selection and curation of language resources. Third, the ability of speakers and specialists to influence those resources helps make support accountable to the people whose language is represented. These propositions explain why technical and social objectives should be evaluated separately and then connected. International maintenance projects offer useful mechanisms, but their contexts differ from Uzbek. A severely endangered language, an Indigenous-language community, and a nationally established language with uneven computational resources do not share identical needs. The transferable lesson concerns deliberate resource development, not an assumption of equal vulnerability or equal outcomes. This article is limited by a small purposive literature set and the absence of original user data. It does not show that machine translation has caused linguistic simplification in Uzbek or that a governance intervention has prevented it. These possibilities remain research questions. Future studies should combine translation audits with evidence about uptake, editing practice, domain use, and speakers' assessments of meaningful linguistic choice.

CONCLUSION. Machine translation can support Uzbek language identity when increased digital access is accompanied by reliable representation and opportunities for speakers to shape the technology. The proposed framework connects these aims through corpus diversity, documented provenance, terminology development, linguistic review, and correction procedures. The contribution is a distinction between technological availability and demonstrated preservation. A translation model may provide useful infrastructure, while claims about identity require further evidence concerning use and expressive capacity. Treating preservation as a continuing process of informed development allows Uzbek to participate in new digital activities without reducing its linguistic diversity to a single machine-generated norm.

REFERENCES

1. Li X., Tracey J., Grimes S., Strassel S. Uzbek-English and Turkish-English Morpheme Alignment Corpora // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). – European Language Resources Association, 2016. – P. 2925–2930. – URL: <https://aclanthology.org/L16-1467/>.
2. Mamasaidov M., Shopulatov A. Open Language Data Initiative: Advancing Low-Resource Machine Translation for Karakalpak // Proceedings of the Ninth Conference on Machine Translation. – Association for Computational Linguistics, 2024. – P. 606–613. – DOI: 10.18653/v1/2024.wmt-1.48.
3. Prieto J., Martinez C., Robles M., Moreno A., Palacios S., Manrique R. Translation Systems for Low-Resource Colombian Indigenous Languages, a First Step towards Cultural Preservation // Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024). – Association for Computational Linguistics, 2024. – P. 7–14. – DOI: 10.18653/v1/2024.americasnlp-1.2.
4. Rikters M., Tomingas M., Tuisk T., Ernštreits V., Fishel M. Machine Translation for Livonian: Catering to 20 Speakers // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. – 2022. – Vol. 2: Short Papers. – P. 508–514. – Association for Computational Linguistics. – DOI: 10.18653/v1/2022.acl-short.55.
5. Tukeyev U., Shormakova A., Karibayeva A., Rakhimova D., Abduali B., Amirova D., Rakhmanberdi N., Aliyev R. An Integrated Approach to Adapting Open-Source AI Models for Machine Translation of Low-Resource Turkic Languages // Computers. – 2026. – Vol. 15, № 2. – Article 73. – DOI: 10.3390/computers15020073.