International Journal of Informatics and Data Science Research
ISSN 2997-3961 (Online)
Vol. 2, No. 8, Aug 2025,
Available online at: https://scientificbulletin.com/index.php/IJIDSR

Availability of Labeled Data Sources in the Uzbek Domain: a Study of Publicly-Available Speech, Text, Image and Multimodal Resources

Fattoev Azizbek
Bachelor student of Computer Science and Software Engineering, Faculty of Information and Communication Technologies, Inha University in Tashkent
fattoevazizbek@gmail.com

Article information:
Manuscript received: 4 Jun 2025; **Accepted:** 10 Jul 2025; **Published:** 28 Aug 2025

Abstract: Modern machine learning and natural language processing systems require large, labeled datasets. Languages and countries with small digital footprints—so-called low-resource contexts—face structural barriers to participation in AI research. This paper presents a systematic, contemporary survey of publicly available labeled datasets for the Uzbek domain across four modalities: speech, text, images, and multimodal (paired) data. We map existing resources, identify critical gaps (e.g., script variation, code-switching, institutional lock-up), and analyze cross-cutting challenges that inhibit dataset creation and reuse. Finally, we propose concrete, actionable strategies—crowdsourcing campaigns, synthetic data generation, cross-lingual transfer, and a national open dataset hub—to accelerate Uzbek AI capacity for academic researchers and AI companies.

Keywords: Uzbek language; low-resource languages; datasets; speech corpora; text corpora; multimodal; crowdsourcing; dataset hub.

1. Introduction

1.1 Motivation

Large, labeled, and standardized datasets are foundational to modern AI. Supervised learning, transfer learning, and evaluation benchmarks all rely on public corpora that are large enough and diverse enough to support robust models. Uzbek—the primary language of Uzbekistan—has, to date, a far smaller publicly available dataset ecosystem than higher-resource languages. This deficit constrains both local research and the ability of Uzbek-language products to reach production quality.

1.2 Research gap and objectives

While there exist systematic surveys and community efforts focused on other low-resource language families (e.g., Indic, African), Uzbek has not been comprehensively mapped in a contemporary, modality-wise public dataset inventory for AI. This paper aims to:

1. Document the publicly available contemporary datasets for Uzbek across four domains: speech, text, images, and multimodal data;
2. Identify and analyze critical gaps and structural barriers to dataset creation and access;

and

3. Propose practical strategies for researchers, industry, and policymakers to accelerate dataset availability in Uzbekistan.

Target audience: academic researchers and AI companies working on Uzbek or comparable low-resource languages.

2. Methodology

2.1 Scope

This study focuses exclusively on publicly available contemporary datasets (open-access or openly hosted academic releases). Proprietary or closed institutional collections were not included in the inventory but are discussed when relevant as a barrier. Modalities examined: (a) Speech (ASR, TTS), (b) Text (monolingual, parallel, annotated corpora), (c) Images (photographs, annotated visual corpora), and (d) Multimodal (paired text+image, audio+video+text, OCR datasets).

2.2 Data collection approach

The inventory and analysis were constructed through:

- A literature and repository review (academic publications, dataset repositories such as Hugging Face Datasets, Mozilla Common Voice, GitHub, Kaggle, Zenodo);
- Examination of national and institutional portals (e.g., government open data portals and local research group releases);
- Collation of community-reported datasets and documented corpus releases from universities and research labs.

Note: Where interviews or direct data sharing with local organizations were feasible, findings were used to corroborate public reporting. The work intentionally excludes proprietary corpora held by private firms or government agencies that are not publicly releasable.

3. Domain-by-Domain Analysis

3.1 Speech

Existing public resources include community voice campaigns and small academic ASR corpora (often read speech). Missing resources include large conversational and spontaneous speech corpora, domain-specific corpora (medical, legal), and dialectally-balanced collections covering rural/urban varieties. Key challenges are script variation (Latin vs. Cyrillic), code-switching with Russian, and limited speaker metadata.

3.2 Text

Existing public resources include Uzbek Wikipedia, limited news and blog articles, and small annotated efforts such as POS/UD treebanks and narrow-domain NER datasets. Missing are large monolingual corpora comparable to Common Crawl scales, high-quality large-scale annotated corpora for tasks like NER, QA, sentiment, and larger parallel corpora for MT. Key challenges are alphabet fragmentation between Latin and Cyrillic, digitization gaps, and lack of standardized annotation tooling and guidelines.

3.3 Images

Existing public resources are few, typically small cultural or heritage photo sets. Missing resources include medical imaging datasets, traffic/road datasets, and product recognition datasets tied to local retail contexts. Challenges include privacy and legal constraints on

medical and surveillance data and a lack of organized acquisition pipelines.

3.4 Multimodal (paired data)

Very limited public multimodal resources exist. There are sparse Uzbek instances in multilingual image-caption datasets and near-zero Uzbek-focused paired corpora such as lecture videos with transcripts, sign language corpora, or visual question answering datasets. Challenges include the high cost of collection, script OCR hurdles, and alignment difficulties.

4. Cross-cutting Issues

Legal and ethical barriers: Much useful data remain in banks, hospitals, universities, and private archives, often inaccessible due to privacy or contractual restrictions.

Standardization and benchmarks: There is no widely accepted Uzbek benchmark suite for NLP or vision tasks, reducing reproducibility and comparability.

Digitization and infrastructure gaps: Many printed archives remain undigitized or in scanned, non-machine-readable formats.

Collaboration shortfalls: Limited cooperation between academia, industry, and government reduces sustained open initiatives.

5. Proposed Solutions and Roadmap

5.1 Crowdsourcing & community engagement

- Launch an Uzbek public voice collection campaign (e.g., modeled on Mozilla Common Voice) including code-switched utterances and speaker metadata collection.
- Encourage Wikipedia expansion and curated web-scraping with license checks.
- Organize annotation hackathons and student annotation sprints.

5.2 Synthetic data generation

- Use back-translation and paraphrase generation to expand small labeled sets with human validation.
- Generate synthetic TTS audio for ASR augmentation paired with validated transcripts.
- Apply generative models for domain-specific image synthesis where ethical.

5.3 Cross-lingual & transfer learning

- Use multilingual pretraining and transfer from related Turkic languages (Turkish, Kazakh) and high-resource contact languages (Russian).
- Leverage mBERT/mT5 style models and fine-tune on limited Uzbek datasets.

5.4 Standardized toolchains and best practices

- Publish annotation guidelines in Uzbek (Latin and Cyrillic) and provide open-source labeling tools.
- Promote script normalization standards and conversion utilities.

5.5 National Open Dataset Hub

- Establish a centralized, curated portal for Uzbek datasets with clear licensing and dataset metadata standards (DOIs, validation metrics).
- Encourage public-private partnerships to release sanitized/anonymized datasets under conditional access when necessary.

6. Prioritization & Implementation Timeline (Suggested)

Priority focus areas:

1. Speech (ASR) & Text (LM/NLP): Immediate—crowdsource speech, expand monolingual corpora, and produce seed annotated datasets.

2. OCR & Digitization: Short term—digitize historical and institutional documents.

3. Multimodal collections: Medium term—curate educational videos and align transcripts.

4. Image datasets: Medium-long term—build medical and traffic datasets subject to legal clearances.

7. Conclusion

This survey documents a modest but meaningful set of contemporary public Uzbek datasets across speech and text, with very limited coverage in images and virtually no Uzbek-focused multimodal corpora. Principal obstacles include alphabet variation (Latin/Cyrillic), code-switching with Russian, lack of standardized annotation practices, digitization deficits, and institutional data lock-in. A combined strategy of community crowdsourcing, synthetic augmentation, cross-lingual transfer, standardized tooling, and a national open dataset hub is recommended to accelerate Uzbek AI capacity.

Appendix

Table 1 – Inventory Summary (high-level)

Modality	Representative Public Resources (examples)	Major Gaps	Main Challenges
Speech	Community voice campaigns; small academic ASR corpora	Conversational speech; domain-specific corpora; dialect coverage	Script variation; code-switching; metadata scarcity
Text	Uzbek Wikipedia; small annotated NER/POS datasets; small parallel sets	Large monolingual web corpora; large annotated corpora for QA, sentiment	Cyrillic vs. Latin fragmentation; digitization gaps
Images	Cultural/heritage photos; ad-hoc community uploads	Medical images; traffic/road datasets; product catalogs	Privacy/legal constraints; lack of organized acquisition
Multimodal	Sparse Uzbek instances in multilingual image-caption corpora	Paired video+transcript, lecture corpora, VQA	Collection cost; script OCR; lack of alignment

Recommendations for Stakeholders

- Researchers & universities: Start annotation courses and student projects that feed into public datasets; publish guidelines and baseline models.
- Industry & startups: Where ethically possible, explore anonymized dataset releases or research partnerships; contribute annotation effort and compute to public initiatives.
- Policymakers: Sponsor a National AI Dataset Hub with seed funding and clear policy for privacy-preserving release.

- International partners: Provide technical assistance in annotation tools, benchmarks, and capacity building.

Selected References and Resources

1. Every Language Matters: Building a More Inclusive Digital Future | United Nations Development Programme https://www.undp.org/digital/blog/every-language-matters-building-more-inclusive-digital-future
2. ISSAI - Institute of Smart Systems and Artificial Intelligence https://issai.nu.edu.kz/uzbek-asr/
3. issai/Uzbek_Speech_Corpus · Datasets at Hugging Face https://huggingface.co/datasets/issai/Uzbek_Speech_Corpus
4. FERUZASPEECH: A 60 HOUR UZBEK READ SPEECH CORPUS WITH PUNCTUATION, CASING, AND CONTEXT https://arxiv.org/html/2410.00035v1
5. Multilingual Speech Recognition for Turkic Languages https://www.mdpi.com/2078-2489/14/2/74?type=check_update&version=1
6. UD_Uzbek-UT https://universaldependencies.org/treebanks/uz_ut/index.html
7. A comprehensive dataset and neural network approach for named entity recognition in the Uzbek language – PubMed https://pubmed.ncbi.nlm.nih.gov/39811531/
8. Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation – PubMed https://pubmed.ncbi.nlm.nih.gov/38708296/
9. (PDF) Text classification dataset and analysis for Uzbek language https://www.researchgate.net/publication/368877295_Text_classification_dataset_and_analysis_for_Uzbek_language
10. Parallel texts dataset for Uzbek-Kazakh machine translation - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC10904177/
11. Open data portal of the Republic of Uzbekistan https://data.egov.uz/eng/ai-data
12. GitHub - google-research-datasets/wit: WIT (Wikipedia-based Image Text) Dataset is a large multimodal multilingual dataset comprising 37M+ image-text sets with 11M+ unique images across 100+ languages. https://github.com/google-research-datasets/wit
13. When Languages Lack Data: Making Open-Source LLMs Work for Banking in Uzbekistan | by Aiphoria Team | Aiphoria | Medium https://medium.com/aiphoria/when-languages-lack-data-making-open-source-llms-work-for-banking-in-uzbekistan-663e6b3891b8