AMERICAN JOURNALS
Open Access, Peer Reviewed Journals

American Journal of Research in Humanities and Social Sciences
ISSN (E): 2832-8019 Volume 40, | September - 2025

EVALUATION SUITE FOR UZBEK SPEECH AND BILINGUAL TRANSLATION AT SCALE

Sukhrob Avezov Sobirovich
PhD, Lecturer in the Department of Russian Language and Literature
Bukhara State University
senigama1990@mail.ru

ABSTRACT

In this article we present a reproducible evaluation suite for Uzbek speech and bilingual translation at scale. It unifies ASR, TTS, MT, and speech-to-text evaluation across Latin/Cyrillic scripts, dialectal variation, and code-switching. We provide task-specific metrics, error taxonomies and human protocols, and we report baseline scores and reliability to support fair, longitudinal benchmarking.

KEYWORDS

Uzbek ASR, bilingual translation, evaluation metrics, code-switching, Latin/Cyrillic, SacreBLEU, chrF, COMET.

Introduction

Reliable evaluation is the precondition for real progress in low-resource speech and translation. Uzbek presents distinctive evaluation challenges: two active scripts (Latin and Cyrillic), agglutinative morphology with productive derivation, widespread Russian code-switching, and strong dialectal variation across regions and media. We propose a unified evaluation suite that treats these facts as first-class design constraints. The suite spans automatic speech recognition (ASR), text-to-speech (TTS), machine translation (MT), and speech-to-text translation (ST), and it blends reference-based metrics with targeted error analyses and human protocols. By standardizing tokenization, normalization, scoring, and reporting, and by curating balanced test sets with explicit difficulty tags (noise, dialect, script, domain), the suite enables fair comparison across systems and time, supports robust model selection for deployment, and lowers the barrier to future community contributions.

Methods and Literature Review

We ground our metrics in widely adopted work and enforce reportable, reproducible settings. For MT and ST text outputs, we require SacreBLEU [3] with fixed tokenization and versioned signatures, chrF [2] to capture character-level adequacy on morphologically rich Uzbek, and a learned quality estimator such as COMET [4] to approximate human adequacy/fluency when references are imperfect or style-diverse. For ASR we report WER and CER with transparent normalization: punctuation stripped, case/symbol mapping unified, and Uzbek diacritics (o‘, g‘) preserved rather than collapsed. For TTS we combine crowd-MOS with 95% CIs and an ASR-CER proxy to approximate intelligibility under channel variation. We adopt latency-aware measures (e.g., Average Lagging) for streaming ST/MT but

keep them optional to avoid penalizing offline systems. Critically, all metrics are accompanied by short qualitative digests with examples, so that users see where systems fail, not just how much.

Our design choices are shaped by the Uzbek language situation and by standardization lessons from the MT community [1]. [3] and reference-free evaluation [4]. Unlike generic leaderboards, our suite encodes linguistic priors:

1. script awareness — parallel scoring in Latn and Cyrl and a consistent transliteration layer for cross-script experiments;
2. code-switch robustness — explicit tags for RU-UZ switches with span-level error tallies;
3. morphology sensitivity — character-level metrics and targeted probes for affixal errors;
4. domain balance — news, conversational chat, broadcast, and public-service speech;
5. noise realism — SNR bands and real device/channel mixtures. For human studies, we define concise rubrics and short tasks that raters can complete reliably, and we report inter-rater agreement. All components are versioned and auditable: test set hashes, preprocessing pipelines, and per-item scores are stored with metadata for longitudinal tracking.

Results

We demonstrate the suite on three internally compiled, legally shareable evaluation bundles:

a. ASR-Clean/Noisy (15 h each) with balanced Latn/Cyrl and four dialect zones;
b. UZ↔RU MT (2×2,000 sentences) with 25% code-switch spans;
c. UZ→RU ST (6 h). Systems are baselines meant to anchor the suite: a conformer-based ASR model fine-tuned on public Uzbek speech, a mid-size Transformer MT pair trained on filtered OPUS-like web data, a speech-encoder–text-decoder ST model, and a Tacotron-style TTS adapted to Uzbek phonology. Scores are illustrative of the suite's reporting style rather than the last word on model performance.

Table 1. Example scores produced by the suite (SacreBLEU signatures recorded; 95% CIs for MOS)

Task	Subset	Primary metric(s)	Score
ASR	Clean (Latn/Cyrl)	WER / CER	9.8% / 4.1%
ASR	Noisy (SNR 0–15 dB)	WER / CER	16.7% / 7.8%
MT UZ→RU	Mixed domains	BLEU / chrF / COMET	29.3 / 57.8 / 0.56
MT RU→UZ	Mixed domains	BLEU / chrF / COMET	24.7 / 53.2 / 0.49
ST UZ→RU	Conversational	BLEU / AL (s)	21.5 / 3.2
TTS UZ	Read speech	MOS (5-pt) / ASR-CER	4.12 ± 0.08 / 6.1%

Beyond headline metrics, the suite outputs structured error taxonomies. For ASR, the top categories on Noisy were: (i) diacritic omissions (o‘→o; g‘→g), 27%; (ii) cross-script confusions within named entities, 19%; (iii) clitic boundary errors in interrogatives, 14%. For MT RU→UZ, most degradations involved derivational morphology (wrong affix stacking) and script preservation for quoted RU spans. ST inherited both ASR deletions and MT lexical mismatches, which the suite surfaces via alignment-based attribution so users can see whether an error is speech or translation-driven.

Human studies show consistent judgments under concise rubrics. For MT adequacy/fluency (3-point scales) on 400 UZ→RU items, inter-rater agreement reached $\kappa = 0.62$ (adequacy) and $\kappa = 0.58$

(fluency). For TTS MOS with 30 raters across three devices, between-device variance was small (Levene's $p > 0.1$), but channel-noisy playback lowered MOS by ~0.12 on average, which matched the ASR-CER increase of 0.9 pp. The suite ships with calculator notebooks that convert these raw judgments to confidence intervals and facilitate power analysis for study sizing.

Scalability is primarily an engineering constraint. Our batch scoring of 10k MT segments completes in minutes on a single CPU node; streaming ST evaluation at 1× real time is feasible with light instrumentation. Storage and governance are treated as first-class: every test item carries a stable ID, license, script tag, dialect label (if known), SNR bin (for audio), and domain tag, and per-segment metric dumps are retained so that new metrics can be retro-applied without re-decoding.

Discussion

Three findings stand out. First, character-aware metrics are not optional for Uzbek. BLEU alone over-rewards surface token overlap while under-penalizing affixal errors; chrF correlates better with rater adequacy on RU→UZ, and COMET provides a useful tie-breaker when multiple outputs cluster in BLEU. Second, script awareness prevents misleading gains. If a model «solves» UZ→RU by copying RU spans indiscriminately, BLEU can rise while human adequacy falls; our cross-script diagnostics expose this failure mode. Third, attribution matters in speech translation: conflating ASR and MT errors leads to the wrong fixes; the suite's alignment-based decomposition points modelers to the true bottleneck.

Conclusion

A language-aware, reproducible evaluation suite turns Uzbek from a «special case» into a tractable engineering target. By standardizing preprocessing, metrics, error taxonomies, and human protocols across ASR, TTS, MT, and ST, and by baking in script and code-switch awareness, the suite supports fair comparison, actionable diagnostics, and longitudinal tracking.

References

1. Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – С. 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. – С. 392-395.
3. Post M. A call for clarity in reporting BLEU scores //arXiv preprint arXiv:1804.08771. – 2018.
4. Rei R. et al. COMET: A neural framework for MT evaluation //arXiv preprint arXiv:2009.09025. – 2020.
5. Sobirovich S. A. A PRAGMATICALLY ORIENTED APPROACH TO GENERATIVE LINGUISTICS //CURRENT RESEARCH JOURNAL OF PHILOLOGICAL SCIENCES. – 2024. – Т. 5. – №. 04. – С. 69-75.
6. Авезов С. КОРПУСНАЯ ЛИНГВИСТИКА: НОВЫЕ ПОДХОДЫ К АНАЛИЗУ ЯЗЫКА И ИХ ПРИЛОЖЕНИЯ В ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ //International Bulletin of Applied Science and Technology. – 2023. – Т. 3. – №. 7. – С. 177-181.