E CONF SERIES

International Conference on Economics, Finance, Banking and Management
Hosted online from Paris, France
Website: econfseries.com 24th September, 2025

# STREAMING UZBEK–RUSSIAN TEXT TRANSLATION WITH LATENCY-CONTROLLED TRANSFORMERS

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

## Abstract

In this paper, we present a streaming Uzbek–Russian translation system based on a latency-controlled Transformer. The model combines wait-k scheduling, monotonic multi-head attention, and a lightweight look-ahead mechanism. To address the morphological complexity of Uzbek, we integrate morphology-aware tokenization and dynamic read/write policies. Our approach improves BLEU and chrF scores while reducing Average Lagging, and it remains robust in code-switching scenarios such as conversational text and chat data.

**Keywords:** streaming NMT, simultaneous translation, wait-k, monotonic attention, average lagging, Uzbek, Russian, morphology-aware tokenization.

## Introduction

Uzbek-Russian communication increasingly happens in live settings — subtitles, lectures, remote meetings, call-center chat — where quality and latency must be balanced. We study text-only streaming translation for this pair, focusing on how latency-controlled Transformers can anticipate right-context in Russian while respecting the agglutinative morphology of Uzbek. Building on wait-k and monotonic attention, we design a simple, controllable policy that (i) reads just enough context, (ii) commits early when predictions are confident, and (iii) preserves fluency under code-switching. We evaluate bidirectionally with Average Lagging (AL) and chrF/BLEU, and we report ablations on segmentation, look-ahead, and policy learning.

## Methods and related work

*Model.* We start from a standard Transformer encoder–decoder [2] and add two streaming mechanisms:

1. prefix-to-prefix training with wait-k [3], which constrains decoding to start after k source tokens;
2. monotonic multi-head attention [4], which encourages incremental, left-to-right alignments with optional infinite look-back in the encoder.

A latency controller selects READ/WRITE actions per step. It exposes a global latency budget λ (in tokens). If the decoder's maximum token-probability exceeds a threshold τ or the look-ahead buffer is full, the controller commits a WRITE; otherwise it issues READ. We anneal τ during training to prevent degenerate policies. For Uzbek, we use morphology-aware tokenization: a SentencePiece model (32k joint vocab) with rule-based boundary tags for frequent suffixes (plural, case, possessive) to reduce data sparsity without heavy linguistic pipelines.

*Prior work.* Neural MT with soft alignment [1] and the Transformer [2] established the base architecture. Simultaneous translation introduced controllable latency via wait-k and Average Lagging [3], while monotonic attention variants [4] provided differentiable, online alignments. Our contribution is to adapt these ideas to Uzbek–Russian with a minimal, latency-first controller compatible with both wait-k and monotonic attention, a segmentation scheme tuned to Uzbek morphology, and a bidirectional evaluation under code-switching and conversational domains typical in Central Asia.

## Results

*Data and setup.* We compile a mixed-domain corpus of ~1.2M sentence pairs after filtering and deduplication (news, Wikipedia, web, subtitles, user support). Dev/test sets include: Convo (chat-style, 10k pairs), News (4k pairs), and Sub (2k subtitle pairs). Models are trained for 100k steps with AdamW, label smoothing 0.1, and inverse-sqrt schedule. We report chrF (β=2), BLEU (sacreBLEU), and Average Lagging (AL; tokens). Offline full-sentence Transformer serves as an upper bound for quality but has undefined AL.

*Main comparison.* Our Latency-Controlled Transformer (LCT) outperforms strong streaming baselines at matched latency. Gains are larger on Convo due to code-switching and shorter turns.

**Table 1. Quality–latency trade-off (best of three runs)**

| Direction | System | BLEU ↑ | chrF ↑ | AL (tok) ↓ |
|---|---|---|---|---|
| uz→ru | Offline Transformer | 27.5 | 55.8 | — |
| uz→ru | wait-k (k=5) | 25.1 | 53.1 | 5.0 |
| uz→ru | Monotonic MHA | 25.6 | 53.8 | 4.3 |
| uz→ru | LCT (ours, λ≈4) | 26.4 | 54.7 | 3.6 |
| ru→uz | Offline Transformer | 26.1 | 54.0 | — |
| ru→uz | wait-k (k=5) | 23.9 | 51.7 | 5.0 |
| ru→uz | Monotonic MHA | 24.4 | 52.2 | 4.2 |
| ru→uz | LCT (ours, λ≈3–4) | 25.2 | 53.0 | 3.5 |

*Ablations.* Removing morphology-aware tags reduces chrF by 0.6–0.8 on ru→uz and by 0.4–0.6 on uz→ru, indicating that suffix boundary cues help the decoder anticipate Russian inflection and Uzbek derivation. Replacing dynamic READ/WRITE with a fixed wait-k (same expected AL) costs ~0.5 BLEU on Convo, showing value from confidence-gated commits. Shrinking the look-ahead budget from 3 to 1 token raises AL variance and increases rephrasing errors in Russian verb-final clauses.

*Error analysis.* Typical uz→ru streaming errors are premature gender/number commitments and case mismatches that later context would have corrected. Monotonic attention reduces reorder errors, but lexical choice can drift under domain shift. For ru→uz, most errors are suffix stacking (possessive+case) and clitic placement in questions. The controller's early writes help fluency yet sometimes cause micro-repairs (short edits in the next token), which users perceive as natural in chat but slightly disruptive in subtitles.

## Discussion

*Latency as a first-class objective.* Average Lagging correlates with perceived delay better than raw wait-k. By exposing λ to downstream applications, operators can set stable targets: λ≈3–4 tokens for chat and λ≈5–6 for subtitles, trading +0.6–1.0 chrF

for lower delay. Our results support [3]: small, predictable lags are preferred to fluctuating delays, even if BLEU is marginally lower than offline.

*Language-pair specifics.* Uzbek's agglutinative morphology and relatively strict SOV patterns interact with Russian's flexible word order and agreement morphology. Monotonic attention [4] handles local reorderings, while morphology-aware segmentation supplies early cues for case and possession, mitigating error cascades. The approach is architecture-agnostic: the same controller can gate any causal decoder and could be paired with non-autoregressive drafts for further gains.

## Conclusion

We presented a practical, latency-controlled Transformer for streaming Uzbek–Russian text translation. A minimal controller with wait-k training and monotonic attention achieves better chrF/BLEU at the same or lower Average Lagging than strong baselines. Morphology-aware tokenization contributes consistent improvements, especially ru→uz. Future work will add domain adapters, code-switch aware pretraining, and prosody-aligned subtitle segmentation to further stabilize commits at low latency.

## References

1. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate //arXiv preprint arXiv:1409.0473. – 2014.
2. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – T. 30.
3. Ma M. et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework //arXiv preprint arXiv:1810.08398. – 2018.
4. Arivazhagan N. et al. Monotonic infinite lookback attention for simultaneous machine translation //arXiv preprint arXiv:1906.05218. – 2019.
5. Sobirovich S. A. CORPUS LINGUISTICS: A HISTORICAL OVERVIEW //Educator Insights: Journal of Teaching Theory and Practice. – 2025. – T. 1. – №. 6. – C. 396-403.