IRSTI 28.23.01 https://doi.org/10.26577/jpcsit4220268

Aidana Karibayeva , Balzhan Abduali*

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: abduali.balzhan@kaznu.kz

AI-DRIVEN TEXT AND AUDIO SYNTHESIS FOR LOW-RESOURCE LANGUAGE DATASETS

Abstract. This paper presents an AI-driven methodology for constructing parallel text corpora and synthesizing parallel audio datasets for low-resource Turkic language pairs, namely Kazakh-Kyrgyz and Kazakh-Uzbek. The scarcity of high-quality linguistic and speech resources for these languages poses significant challenges for the development of neural machine translation and automatic speech recognition systems. The paper proposes and validates a methodology for parallel dataset construction driven by artificial intelligence, in which the selection of a translation system is guided by a set of criteria encompassing free accessibility, translation quality, and processing efficiency, assessed through experiments on a 2000-sentence test set. Based on this evaluation, large-scale parallel corpora for the Kazakh-Kyrgyz and Kazakh-Uzbek language pairs were generated using the selected AI system. A subsequent manual error analysis revealed that approximately 3.7% (Kazakh-Kyrgyz) and 4.3% (Kazakh-Uzbek) of the translations contained inaccuracies, indicating the need for post-editing and further model refinement. Audio synthesis experiments using MMS-TTS and TurkicTTS systems demonstrated that synthetic speech of near-natural quality can be generated for both languages, with NISQA scores reaching up to 4.44 for Uzbek. The findings confirm that the proposed methodology provides a practical and scalable foundation for expanding linguistic resources for low-resource Turkic languages and supporting further research in machine translation and speech processing.

Keywords: low-resource languages, methodology for creating datasets, parallel text corpora, AI systems, audio synthesis.

1. Introduction

There are more than 7,000 languages spoken worldwide, yet the majority of digital language technologies are developed for only a small subset of them [1]. While more than half of the world's population uses fewer than 20 major languages, a large number of languages remain underrepresented in the digital space. Many of these so-called low-resource languages are official languages of countries with relatively small populations, including several states in Central Asia such as Kazakhstan. The scarcity of digital and linguistic resources severely hampers the advancement of modern language technologies for these languages.

Low-resource languages often suffer from digital exclusion due to the lack of essential linguistic resources required for effective automatic language processing. The development of natural language processing (NLP) systems typically requires the availability of character encoding standards, lexical resources, and well-described linguistic rules, including grammar, syntax, and morphology. Although the problems associated with character encoding and the coverage of a core vocabulary have been partially solved by modern information technologies, the creation of comprehensive linguistic resources remains a difficult and labor-intensive task. Consequently, these languages suffer from an acute deficiency of extensive annotated datasets, lexicons, NLP toolkits, and digitized textual materials required for model development and evaluation. Compared to well-resourced languages like English, Chinese, Hindi, Spanish, or French, such insufficient data makes it considerably more difficult to construct reliable systems capable of handling machine translation, speech recognition, text classification tasks [2–4].

Kyrgyz and Uzbek fall into the category of low-resource languages, primarily due to the insufficient availability of high-quality linguistic data required to train contemporary natural language processing and machine translation systems. Among the most pressing challenges is the restricted access to parallel corpora for these languages. Although research interest in building and benchmarking machine translation systems for under-resourced Turkic languages has grown considerably in recent years [5, 6], the absence of substantial high-quality parallel datasets continues to be a fun-

damental bottleneck impeding further advances in this domain.

Several approaches have been proposed to address this problem. One solution involves the creation of synthetic corpora. For example, by constructing sentences based on a complete set of morphological rules and suffixes [7, 8]. One of the most widely used approaches today is to create synthetic parallel data by translating monolingual texts from one language to another using machine translation. However, this approach can also lead to significant difficulties in terms of translation quality. Machine translation systems for resource-constrained languages lack accuracy due to the limited amount of training data. Consequently, this may give rise to semantic distortions, grammatical inaccuracies, and incorrect interpretations of phrases, ultimately compromising the overall quality of the resulting parallel corpus.

A common approach to building synthetic parallel corpora for low-resource languages involves using an intermediate language, typically English [9, 10]. While this method helps compensate for the lack of direct translation data, it is not always effective for closely related language pairs such as Turkic languages. Routing translation through English often results in the loss of semantic and grammatical nuances specific to these languages, thereby reducing the quality of the generated data [11]. This highlights the need for direct corpus construction methods that better preserve the linguistic properties of such language pairs.

Publicly available parallel data for the Kyrgyz-Kazakh and Uzbek-Kazakh language pairs remains extremely scarce, making it difficult to develop and train effective neural machine translation (NMT) models. Research and openly accessible resources for these language pairs are also notably limited [12], underscoring the need for systematic approaches to constructing high-quality parallel corpora.

This article presents an AI-driven methodology for constructing parallel text corpora for the Kyrgyz-Kazakh and Uzbek-Kazakh language pairs, directly addressing the data scarcity challenges outlined above. The proposed approach leverages artificial intelligence techniques for both text and audio synthesis, aiming to enrich language resources for low-resource Turkic languages and to support further research in machine translation and related natural language processing tasks.

2. Materials and methods

This section outlines the methodological foundations of the proposed AI-driven approach to constructing parallel datasets for low-resource Turkic languages. It begins by reviewing the core challenges associated with data scarcity and examines existing resources and strategies for parallel corpus construction. Next, it discusses the role of modern AI-based systems in the automatic generation of both text and audio data, and motivates the selection of an AI-assisted pipeline for dataset creation. Finally, the section introduces the overall methodology adopted in this study for building and evaluating parallel corpora for the Kyrgyz-Kazakh and Uzbek-Kazakh language pairs.

2.1 Data Scarcity in Low-Resource Languages and Approaches to Parallel Corpus Creation

The scarcity of large, high-quality parallel text datasets represents one of the most significant obstacles for low-resource languages. Turkic languages including Kazakh, Kyrgyz, and Uzbek are particularly affected, given that openly accessible bilingual resources for these languages remain considerably limited [13, 14]. Efforts to bridge this gap have included developing bilingual dictionaries and compiling small parallel datasets for pairs such as Uyghur–Kazakh, Kazakh–Kyrgyz, and Kyrgyz–Uyghur, alongside pivot-based translation strategies that rely on English as a bridge language [14, 15]. Despite these efforts, the available resources fall short in both volume and linguistic coverage, rendering them insufficient for training robust neural machine translation systems.

Creating parallel data for languages with limited resources is a complex and multifaceted process that may involve various sources and methods. Existing parallel datasets represent the most straightforward option, and a number of large multilingual resources are available, including OPUS, TED Talks subtitles, and Europarl. OPUS stands out as one of the most extensively utilized open-access repositories of multilingual parallel corpora among the available resources [16–18]. Such datasets serve as a fundamental resource for training and evaluating machine translation systems, as well as for advancing research in multilingual natural language processing. Nevertheless, for a number of Turkic language pairs – notably Kazakh–Kyrgyz and Kazakh–Uzbek – the available resources remain inadequate in both

coverage and scale to effectively support the practical development of machine translation systems [19, 20].

Recent advances in neural machine translation and large multilingual models, such as attention-based NMT architectures, XLM-R, and the NLLB project, have substantially enhanced translation quality for low-resource languages through the utilization of large-scale multilingual pre-training and cross-lingual transfer learning techniques [21–24]. At the same time, these developments have made it possible to use machine translation systems not only as end-user tools but also as instruments for data generation, enabling the construction of synthetic parallel corpora from monolingual texts.

Several studies have discussed different strategies for corpus construction, including the use of open data sources, crowdsourcing, proprietary datasets, and human–machine collaboration [25-27]. In particular, AI-assisted and semi-automatic approaches have gained increasing attention due to their scalability and cost-effectiveness. However, the quality of automatically generated parallel data remains a critical issue, especially for low-resource languages, and typically requires careful selection of translation systems as well as subsequent quality assessment and filtering.

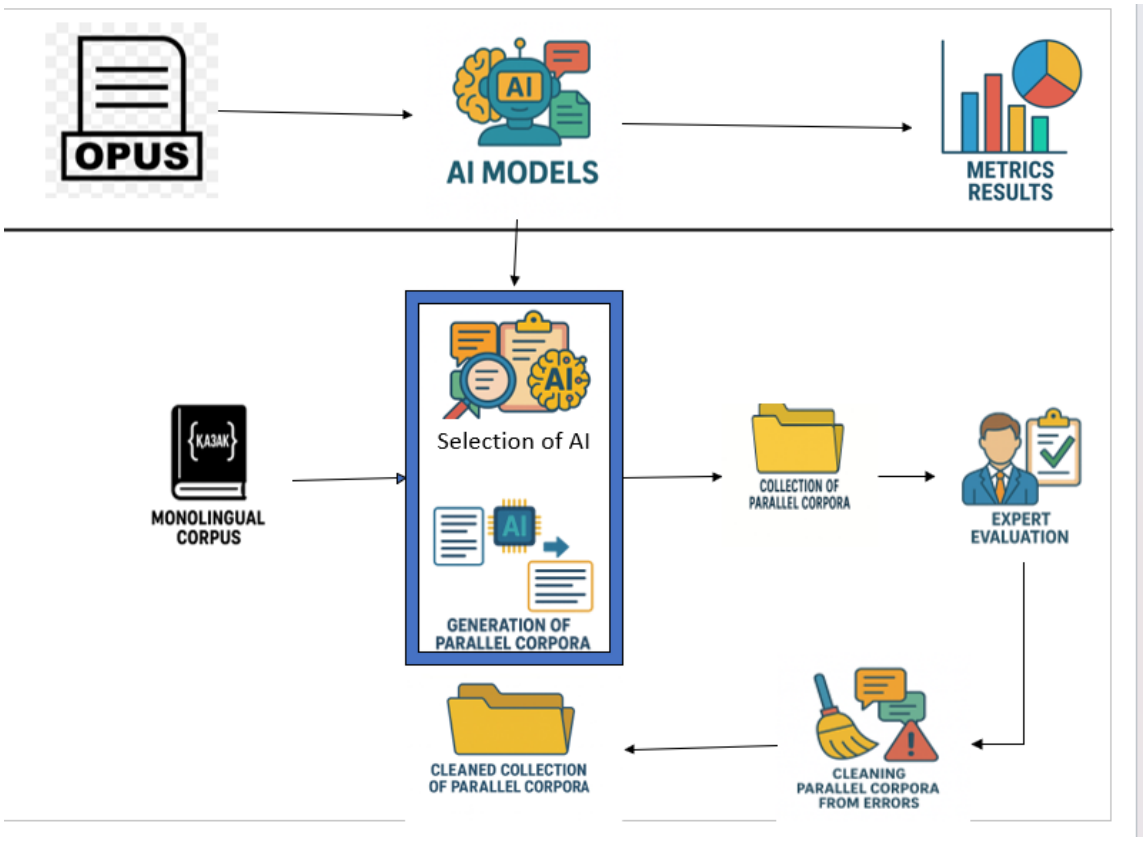
Building on these considerations, this study proposes a methodology that leverages artificial intelligence to construct parallel text corpora for the Kazakh–Kyrgyz and Kazakh–Uzbek language pairs. The approach seeks to integrate the scalability of contemporary machine translation systems with rigorous quality evaluation, with the ultimate goal of producing practical datasets suitable for training and benchmarking neural machine translation models.

2.2 Developed methodology

2.2.1 AI-Assisted Parallel Text Corpus Construction Methodology

As depicted in Figure 1, the methodology put forward in this study comprises three core phases: (1) preparation, (2) selection of an appropriate AI system, and (3) corpus construction.

Figure 1
Architecture of the proposed method for text



The preparation stage consists of several initial tasks. First, an extensive review of previous studies is carried out to investigate recent developments in machine translation for low-resource languages, with special attention given to Turkic language pairs such as Kazakh–Kyrgyz and Kazakh–Uzbek. Next, publicly available parallel corpora for these language pairs are collected from open resources, particularly the OPUS platform, followed by an analysis of their reliability and overall quality. At the same time, existing neural machine translation approaches and AI-driven translation systems applicable to the selected languages are examined. Based on the outcomes of the literature analysis, suitable translation evaluation measures are determined. Furthermore, a number of criteria for selecting AI systems are established, including translation performance, model design, computational requirements, and the accessibility of training resources. Finally, a balanced and domain-diverse benchmark dataset of parallel sentences is prepared to support an objective assessment of translation quality.

Several AI-based translation systems were evaluated against a set of predefined criteria, including accessibility, translation quality, and processing efficiency. The systems considered for comparison were Google Translate, NLLB, ChatGPT, DeepSeek, GitHub Copilot, and Gemma [32–37], which can be broadly divided into specialized translation systems and general-purpose large language models. Each system was used to translate a test set of Kazakh sentences into Kyrgyz and Uzbek, and the resulting outputs were assessed using the metrics described in Section 2.4. The system demonstrating the best overall performance in terms of translation quality, efficiency, and accessibility was subsequently selected for large-scale corpus generation.

The second stage is devoted to the comparative selection of AI systems. During this process, every candidate model translates the Kazakh portion of the evaluation dataset into Kyrgyz and Uzbek. The generated outputs are then assessed using the previously chosen evaluation metrics. Considering both the metric results and the predefined selection conditions, the candidate systems are systematically compared and filtered through several evaluation rounds until the most effective model is selected for further large-scale synthetic data generation.

The final stage focuses on corpus construction. At this stage, the chosen AI system is applied to translate the Kazakh component of an existing Kazakh–English parallel corpus into Kyrgyz and Uzbek. As a result, extensive synthetic parallel datasets for the Kazakh–Kyrgyz and Kazakh–Uzbek language pairs are created. These newly generated corpora can later serve as training and evaluation resources for neural machine translation systems.

2.2.2 Text-based audio synthesis

This study uses a cascade methodology for creating a text-based audio corpus. It aims to create a parallel text-audio corpus for Turkic languages. The source language is Kazakh, and the target languages are Uzbek and Kyrgyz.

The proposed approach differs from the traditional speech data collection and automatic recognition methodology. In the classical method, the quality of the corpus depends on the speakers, the writing environment, dialect features, and the accuracy of ASR/STT systems. In low-resource languages, this leads to transcription errors, phonetic distortions, and lower data quality.

Based on high-quality audio and text in the input language (Kazakh), first, supervised parallel text data is generated, and then converted into synthetic speech using TTS systems. This approach increases the scalability, repeatability, and controllability of the corpus creation process.

The stages of the proposed methodology consist of the following:

1 Kazakh data preparation. At this stage, the source texts in Kazakh are preprocessed: normalized, standardized, and unnecessary elements removed. Morphological and syntactic features are taken into account. This text is taken from the original high-quality audio dataset from NU corpora.

2 AI- based machine translation. The transcriptions are translated into Uzbek and Kyrgyz using the NLLB neural translation model. If necessary, additional grammatical and syntactic corrections are made, as in the previous text data processing section.

3. Speech synthesis – Kazakh, Uzbek, and Kyrgyz texts are converted into audio files using TTS systems. For each sentence, a corresponding "text – audio" pair is created.

In the study, only a limited number of TTS systems were identified that support the Kazakh language and our target languages (Uzbek and Kyrgyz). These systems are: MMS-TTS and TurkicTTS. The TTS systems were evaluated according to the following criteria: support for the target languages (Uzbek and Kyrgyz), availability under an open license, local deployment, scalability, naturalness, and suitability for generating a synthetic speech corpus. During the work, for 2 languages, the Kyrgyz

Based on the analysis, MMS-TTS and TurkicTTS were selected for the study. Although there are also systems for the Turkic languages, CoquiTTS and ElevenLabs, they were excluded because they did not fully meet the methodological requirements.

The CoquiTTS model does not support the Kyrgyz language. Therefore, additional training on a fixed Kyrgyz speech corpus would be required to use CoquiTTS in this study. Since such a corpus was not used in the study, the system was excluded from the analysis. As an additional limitation, the discontinuation of the Coqui company reduces the reliability of the system's long-term support. ElevenLabs, on the other hand, supports many languages, including Kyrgyz and Uzbek, and provides high-quality speech synthesis. However, the platform is commercial and closed. It does not offer open-source code, does not support local deployment, and limits the amount of generation available through tariff plans. In addition, dependence on external APIs reduces the reproducibility of experiments. For these reasons, ElevenLabs was excluded as a system that does not meet the requirements of openness, scalability, and independent scientific use.

MMS-TTS was selected as an open, multilingual system supporting more than 1,100 languages, including Uzbek and Kyrgyz. Its local deployment, lack of commercial restrictions, and ability to automatically generate audio data make it suitable for working with low-resource Turkic languages. In addition, the use of phonemic representations across languages increases their applicability to languages with limited speech resources.

TurkicTTS was selected as a special speech synthesis system for Turkic languages. Its advantage is its focus on the phonetic and morphological features of this language group, including agglutinativity, pronunciation, and word-form variability. Support for Uzbek and Kyrgyz languages, as well as the ability to further train on user data, make TurkicTTS suitable for the tasks of this study. Thus, MMS-TTS and TurkicTTS were selected as the most suitable systems for synthetic speech generation in Uzbek and Kyrgyz languages. One is a universal multilingual model, the other is a system adapted to Turkic languages. Using both together increases the study's objectivity.

4. Corpus formation – All texts and audio files are combined into a single structure:

{text KAZ} → {text UZB/KYR} → {audio KAZ (high-quality human voice)} → {audio UZB/KYR (synthetic)}

The advantages of the proposed methodology are as follows:

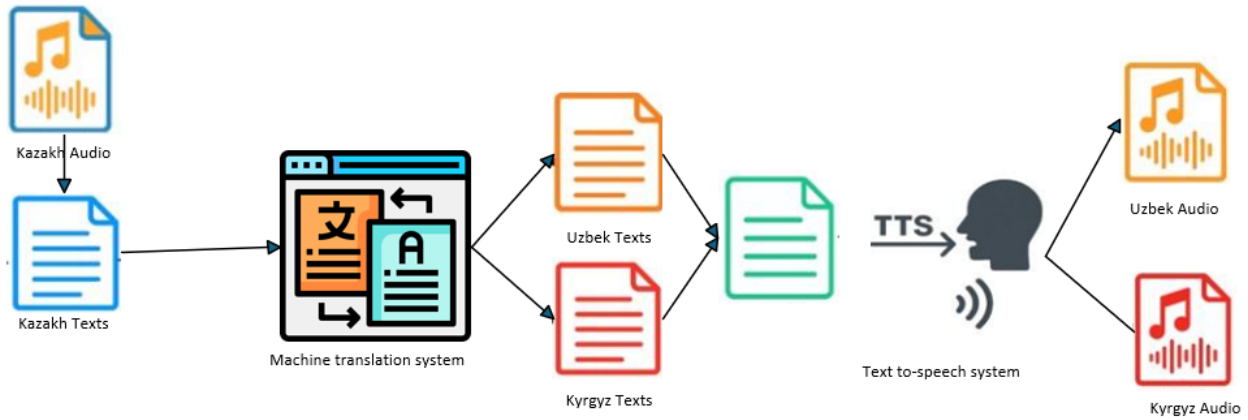
- High data quality control.
- Independent of narrators and recording conditions.
- Easy to scale and repeat.
- Reduces data collection costs and reduces transcription errors.

The proposed *text-based audio synthesis* methodology allows for the efficient and reliable formation of a parallel text-audio corpus in the Turkic language

An illustrative diagram of the proposed methodology is presented in Figure 2.

Figure 2

The scheme of Text-based audio synthesis Generation for Uzbek and Kyrgyz languages



2.3 Parallel Data Sources and Preprocessing

To conduct the experimental study, parallel text datasets were obtained from publicly available corpus resources. A review of existing multilingual corpora indicated that datasets suitable for Kazakh–Kyrgyz and Kazakh–Uzbek machine translation remain limited in both quantity and quality. The main available corpora from the OPUS collection are summarized in Table 1.

Table 1
Available OPUS Parallel Corpora for Kazakh–Kyrgyz and Kazakh–Uzbek

Name of Dataset	Parallel sentences	Dataset link
OPUS [15]	214 183 KZ-KG	https://opus.nlpl.eu/corpora-search/kk&ky
OPUS [15]	249 941 KZ-UZ	https://opus.nlpl.eu/corpora-search/kk&uz

According to these resources, 2000 Kazakh-Kyrgyz and Kazakh-Uzbek sentence pairs were taken to create a test set. Kyrgyz and Uzbek sentences were considered as gold standard references, and Kazakh sentences were used as source texts for translation and evaluation.

In addition to building large parallel corpora for low-resource Turkic languages, a monolingual Kazakh dataset was selected as the source material. The dataset was obtained from an open-access resource and originally contained ∽300 thousands parallel Kazakh–English sentence pairs [28]. In the context of this study, only the Kazakh side of the corpus was selected and utilized as the source input for automatic translation. Using the artificial intelligence system selected here, this corpus was translated into Kyrgyz and Uzbek, resulting in newly created Kazakh-Kyrgyz and Kazakh-Uzbek parallel corpora.

This approach allows for the efficient creation of large parallel datasets for low-resource language pairs by using monolingual or bilingual resources and modern artificial intelligence-based translation systems.

2.4 Quality metrics

2.4.1 Translation Quality Metrics for Text

Machine translation quality was evaluated using several complementary metrics, since a single evaluation method cannot fully reflect translation performance, especially for morphologically rich Turkic languages. The selected metrics measure translation quality from different perspectives, including lexical similarity, character-level correspondence, editing distance, and semantic adequacy.

Among the applied metrics, BLEU evaluates the overlap between generated and reference translations using n-gram matching and remains one of the most widely used metrics in machine translation research [29]. In contrast, WER focuses on the number of editing operations required to transform the generated sentence into the reference text, where lower values indicate better performance. To better capture morphological variations common in Kazakh, Kyrgyz, and Uzbek, the ChrF metric was additionally employed. Unlike word-based metrics, ChrF operates at the character level, making it more sensitive to agglutinative language structures [30]. Semantic and linguistic aspects of translation quality were assessed using METEOR and COMET. METEOR considers factors such as stemming, synonym matching, and word order, while COMET applies neural-based evaluation techniques to estimate translation fluency and semantic consistency [31].

2.4.2 Metrics for synthesized speech

To obtain a more comprehensive evaluation of synthesized speech quality, several objective assessment metrics were employed.

MOSNet was used for automatic prediction of perceptual speech quality scores comparable to the Mean Opinion Score (MOS).

DNSMOS evaluates the overall audio quality, naturalness of speech, and the level of distortion in the generated speech signal. Additionally, a neural network-based quality assessment was conducted using NISQA. Furthermore, SSL-MOS, based on self-learning representations, was used to assess the naturalness and intelligibility of synthesized speech.

3. Results

3.1 Text Corpus Generation Results

Table 2 presents the evaluation results of six AI-based translation systems on a test set of 2,000 Kazakh sentences from the OPUS datasets, comparing their performance across multiple metrics for both Kazakh–Kyrgyz and Kazakh–Uzbek language pairs.

Table 2
Translation model performance assessed on a test set of 2,000 Kazakh sentences drawn from the OPUS datasets

Metrics / System	BLEU		WER		METEOR		COMET		ChrF	
	kg	uz	kg	uz	kg	uz	kg	uz	kg	uz
Google Translator	12.00	13.00	0.93	0.91	0.08	0.06	0.70	0.71	23.02	24.41
Chat GPT	**37.60**	**37.30**	0.87	0.86	**0.16**	**0.16**	**0.91**	**0.89**	31.57	**32.78**
Nllb-200-3.3	30.40	31.40	0.88	**0.85**	0.13	0.13	0.78	0.72	27.12	25.35
DeepSeek	34.00	33.30	**0.88**	0.87	0.15	0.15	0.83	0.80	**31.58**	32.51
Copilot	27.20	25.00	0.89	0.88	0.15	0.13	0.82	0.76	31.40	30.70
Gemma-2-27b	32.00	32.00	0.87	0.87	0.14	0.14	0.81	0.81	30.27	31.36

The Figure 1 visually illustrates the performance differences between the evaluated translation systems, enabling a clear comparison of their effectiveness and facilitating the selection of the most suitable model.

Table 3 illustrates the translation time required by each AI system to process 300 thousand Kazakh sentences, revealing a substantial difference in efficiency between Gemma and NLLB, which completed the same task in approximately 70 days and 3 days, respectively.

Figure 1
Comparison of translation systems

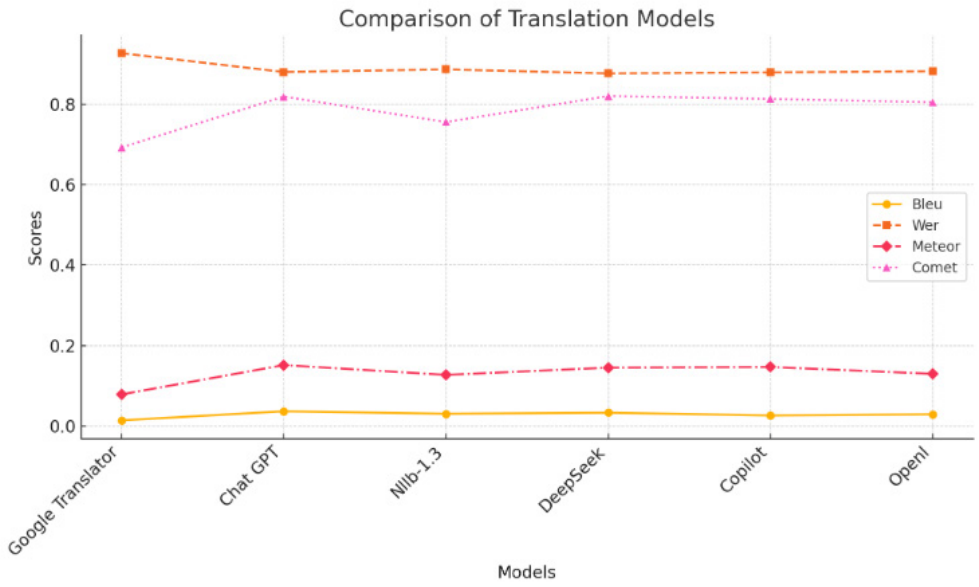


Table 3
Processing Time Required by Each AI System for Large-Scale Text Translation

Models	Estimated processing time required to translate a corpus of 300 thousand sentences
Gemma	~70 days
NLLB	~3 days

Table 4
Summary of Translation Error Analysis for Kazakh–Kyrgyz and Kazakh–Uzbek Language Pairs

Indicator	Kazakh–Kyrgyz	Kazakh–Uzbek
Total errors identified	11175	12930
Error rate	3.7%	4.3%
Main error types	Abbreviation, Repetition	Abbreviation, Repetition

The table 4 presents information about errors and quantity of errors. Manual inspection of the translated corpus revealed two predominant error categories: semantic errors, characterized by incorrect conveyance of meaning, and lexical errors, involving missing or inaccurately translated words. The total number of identified errors amounted to 11175 for the Kyrgyz and 12930 for the Uzbek translations. Addressing these errors led to a notable enhancement in the overall quality of the parallel corpus. Based on these figures, the estimated error rate stands at approximately 3.7% for the Kazakh–Kyrgyz and 4.3% for the Kazakh–Uzbek language pairs, encompassing inaccuracies in abbreviation translation and repetitions that were subsequently removed using a specially developed automated program.

3.2 Speech synthesis results

Based on the *proposed methodology for text-based audio synthesis*, a parallel corpus was obtained, containing the original Kazakh-language data and their Uzbek and Kyrgyz versions. The compiled data were uploaded to the Hugging Face platform. The test set of 1000 audio files is available at the following link: https://huggingface.co/datasets/KazNU-Lab/audio-parallel-1k-mms.

For the experimental section, 500 audio fragments of each male and female voice were selected from the NU Corpus. The dataset for the overall testing consisted of 1,000 audio recordings. This data was used for reference-free metrics that do not require a ground-truth recording to evaluate synthetic speech quality. They allow for the analysis of speech naturalness, intelligibility, and overall perceived quality.

Table 5
Results of synthetic audio generation for Uzbek

Language	DNSMOS	MOSNet	NISQA	SSL-MOS
MMS	3.33	4.04	4.26	3.24
TurkicTTS	3.31	3.99	4.44	2.39

Table 6
Results of synthetic audio generation for Kyrgyz

Language	DNSMOS	MOSNet	NISQA	SSL-MOS
MMS	3.10	3.75	4.00	2.60
TurkicTTS	3.15	3.80	4.20	1.90

The MMS system for Uzbek showed high results on the SSL-MOS metric (2.39 for TurkicTTS). This difference indicates that the MMS system's speech is prosodically (intonation, stress, speech tempo, pauses between words or phrases, rhythm, pitch / F0, speech intensity, emotional coloration) stylistically close to natural speech. Such a property is crucial for training ASR systems.

The TurkicTTS system for Uzbek has the advantage according to the NISQA metric (TurkicTTS 4.44, MMS 4.26). This indicates a high-quality speech signal: a low number of artifacts and consistent loudness and intonation are observed. The system was likely trained on clean studio data.

According to the DNSMOS and MOSNet results, both systems are practically on par. The difference is only 0.05 points, which is within the statistical margin of error.

The results indicate that both systems provide an acceptable level of quality for the Kyrgyz language. However, TurkicTTS achieves slightly higher scores than the MMS system across all three metrics.

The difference between the DNSMOS metrics is only 0.05 points; both metrics remain within an acceptable quality range, indicating that the synthetic speech is of sufficient quality.

The MOSNet metric evaluates speech naturalness and prosodic characteristics (TurkicTTS 3.80, MMS 3.75). Although the difference is small, TurkicTTS shows a consistent advantage. These values are lower than those obtained for Uzbek (3.99–

4.04), which is explained by the small amount of available training data for Kyrgyz and its phonetic characteristics.

The most significant advantage was observed in the NISQA metric, with 4.20 for TurkicTTS and 4.00 for MMS. The difference is 0. 20 points. This result is consistent with the trend observed for Uzbek (TurkicTTS – 4.44, MMS – 4.26), indicating that the TurkicTTS architecture consistently generates acoustically cleaner signals.

The fully-processed audio corpora are available at the following repositories: https://huggingface.co/datasets/KazNU-Lab/

Currently, the repositories are private pending copyright registration for the processed audio corpora. This is due to the lack of ready parallel audio corpora that simultaneously cover several Turkic languages in open access. Therefore, the created resource has both scientific and practical value for machine translation, speech synthesis, and multimodal processing.

4. Discussion

The evaluation results show that ChatGPT and DeepSeek achieve the highest performance for Kazakh-Kyrgyz and Kazakh-Uzbek translations according to several automatic indicators (Table 2). Here, ChatGPT shows a good performance in COMET and ChrF results, which indicates a high level of semantic sufficiency and fluency. DeepSeek has similar performance and can be considered a competitive solution. In contrast, NLLB-200-3.3 and Gemma-2-27b achieve slightly lower scores, indicating that although the models can perform the translation task, they do not reach the same level of accuracy and fluency as the leading systems. Despite their good performance, both ChatGPT and DeepSeek are not suitable for large-scale corpus generation in this study due to limited free access and practical limitations in processing large amounts of data. Among the models that can be used via an available API or local deployment, Gemma and NLLB offer realistic options. Gemma showed slightly better translation quality; however, its processing speed was found to be a significant limitation. As shown in Table 3, in order to improve processing speed, the NLLB-200-3.3 model was selected as the preferred option, as it provides a more optimal balance between translation quality and computational efficiency.

Using the NLLB-200-3.3 model, a large-scale machine translation project was conducted from Kazakh into Kyrgyz and Kazakh into Uzbek with the aim of creating a parallel corpus. The overall quality of the translation was generally satisfactory. However, during a systematic manual review of the translated text, a number of recurring errors were identified that require attention.

As a result of analyzing and cleaning the corpus from 300 000 sentences, shown in table 4, 8322–9307 repeated lines were identified and removed (depending on the language pair and corpus version). Removing these repetitions improved the structural and lexical consistency of the training material. Another important source of errors was the incorrect representation of abbreviations, toponyms, and official names. In the context of closely related Turkic languages, the models often replaced country names and official abbreviations with equivalents from another language. For example, cases of replacing the word “Kazakhstan” with “Kyrgyzstan” or “O’zbekiston” were recorded, as well as incorrect interpretation of abbreviations such as “КР” and “ЖШС”.

Comparative analysis showed that in the Kyrgyz part of the corpus, about 11000 out of 300 000 sentences were incorrect, which is about 3.7% of the total data. In the Uzbek part of the corpus, the number of corrections was slightly higher: 12930 examples (9300 lexical repetitions and 3630 incorrect abbreviations), which corresponds to 4.3% of the total number of sentences. These figures indicate the systematic nature of errors that occur during the automatic generation of synthetic data for closely related Turkic languages. Despite the relatively moderate proportion of errors (3.7–4.3%), such errors significantly affect the training process of neural machine translation models, contributing to overfitting, lexical-semantic distortions, and reduced generalization capabilities. Pre-cleaning improved the structural consistency of the corpus.

Despite these shortcomings, NLLB-200-3.3 demonstrates considerable promise for generating parallel data for low-resource Turkic language pairs. However, the findings confirm that automatic translation alone cannot guarantee sufficient corpus quality, and post-editing remains an essential step – particularly for handling abbreviations, named entities, domain-specific terminology, and complex syntactic constructions. Furthermore, fine-tuning the model on Kazakh–Kyrgyz and Kazakh–Uzbek specific data is expected to enhance its accuracy and reliability in future applications.

Following the completion of the initial manual review phase, the most obvious errors—such as in-

correct abbreviations, word repetitions, and clear syntactic and semantic inconsistencies–were corrected. This led to a significant improvement in the overall quality of the corpus. However, this phase should be considered preliminary.

The results of objective evaluations of synthetic speech quality in Kyrgyz and Uzbek demonstrate that the proposed Text-based audio synthesis methodology applies to both languages. Moreover, the results indicate that resource disparities between the languages directly affect synthesis quality. Overall, a similar trend was observed across both languages: while the TurkicTTS system consistently performed better on the NISQA metric, the results of the two systems were close for the DNSMOS and MOSNet scores. This indicates that the TurkicTTS system is inclined to generate a speech signal with high acoustic quality and few artifacts, and that, in terms of overall naturalness and perceptual quality, the MMS-TTS and TurkicTTS systems are comparable.

The results obtained for Uzbek demonstrate that synthetic speech was generated at a sufficiently high quality. In particular, the NISQA metrics scores ranging from 4.26 to 4.44 indicate that the synthesized speech is close to natural pronunciation. This demonstrates that the TTS systems used for Uzbek are suitable for creating a parallel audio corpus. Moreover, the systems' advantages are evident across various aspects: While TurkicTTS excels in acoustic clarity and signal quality, MMS-TTS achieved a higher SSL-MOS score, which evaluates prosodic naturalness. For example, the SSL-MOS score was 3.24 for MMS-TTS and 2.39 for TurkicTTS. This suggests that the MMS-TTS system may better capture the naturalness of intonation, rhythm, and speech flow.

Although the quality metrics obtained for Kyrgyz were lower than those for Uzbek, they still demonstrated quality close to natural speech. The low MOSNet, DNSMOS, and NISQA scores are explained by the limited availability of Kyrgyz speech resources, insufficient phonetic coverage in multilingual models, and the absence of dedicated Kyrgyz TTS systems.

Overall, TTS results demonstrate that the quality of synthetic speech is directly dependent on the level of a language's resource availability. Therefore, expanding the textual and speech corpora in Kyrgyz, improving phonetic normalization, and adapting TTS models remain important directions for future research. The proposed Text-based audio synthesis methodology provides a practical basis for creating parallel audio corpora for low-resource Turkic languages.

5. Conclusions

This study presented an AI-driven methodology for constructing parallel text corpora and synthesizing audio datasets for the Kyrgyz–Kazakh and Uzbek–Kazakh language pairs, addressing the critical shortage of high-quality linguistic and speech resources for these low-resource Turkic languages.

The experimental results demonstrated that while ChatGPT and DeepSeek achieved the highest translation quality, their practical limitations in processing large volumes of data made them unsuitable for large-scale corpus generation. Consequently, NLLB-200-3.3 was selected as the most viable option, offering a balanced trade-off between translation quality and computational efficiency. The large-scale translation of 300,000 Kazakh sentences into Kyrgyz and Uzbek revealed systematic errors affecting approximately 3.7% and 4.3% of the data respectively, primarily involving incorrect abbreviations, country name substitutions, and lexical repetitions. Although these error rates are relatively moderate, their impact on neural model training is significant, contributing to overfitting and lexical-semantic distortions.

With regard to audio synthesis, the evaluation results confirmed that the proposed text-based audio synthesis methodology is applicable to both Kyrgyz and Uzbek. The TurkicTTS system consistently demonstrated superior acoustic quality, while MMS-TTS showed advantages in prosodic naturalness, particularly for Uzbek. The lower synthesis quality observed for Kyrgyz is directly attributable to the limited availability of speech resources and the absence of dedicated Kyrgyz TTS systems, highlighting the need for expanded phonetic corpora and language-specific model adaptation as priorities for future research.

Given the rapid development of AI models and tools, there is room for further improvement of the obtained results and improvement of the proposed methodology. Future research will be aimed at fully correcting the identified errors and increasing the quality of the data.

The constructed Kyrgyz-Kazakh and Kazakh-Uzbek parallel corpora will serve as a basis for training "lightweight" LLM models that can operate on low computational resources. In addition,

the proposed methodology can be used to build parallel corpora for other Turkic languages and develop high-quality neural machine translation systems for resource-constrained Turkic languages.

Funding

This work was supported by the Ministry of Science and Higher Education of the Republic of Kazakhstan under the grant project «Study of automatic generation of parallel speech corpora of Turkic languages and their use for neural models» (grant number IRN AP23488624).

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, B.A.; methodology, A.K.; software, A.K. and B.A.; validation, B.A., and A.K.; formal analysis, B.A.; investigation, A.K.; resources, B.A. and A.K; data curation, A.K.; writing—original draft preparation, A.K. and B.A.; writing–review and editing, B.A. and A.K.; visualization, B.A.; supervision, A.K.; project administration, A.K. All authors have read and agreed to the published version of the manuscript.

References

1. Ethnologue. How Many Languages Are There in the World? Available online: https://www.ethnologue.com/insights/how-many-languages/ (accessed on 30 July 2025).
2. Cieri, C.; Maxwell, M.; Strassel, S.; Tracey, J. Selection Criteria for Low-Resource Language Programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portorož, Slovenia, 2016; European Language Resources Association (ELRA), pp. 4543–4549.
3. Ivasubramanian, R.; Umamaheswari, T.; Babu, S.B.G.; Inakoti, R.; Salome, J.; Y.M., Dr; Sivasubramanian, R. Natural Language Processing in Low-Resource Language Contexts. *Front. Health Inform*. 2024, 13, 1578–1584. doi:10.52783/fhi.vi.1776.
4. Pakray, P.; Gelbukh, A.; Bandyopadhyay, S. Natural language processing applications for low-resource languages. *Nat. Lang. Process*. 2025, 31, 183–197. doi:10.1017/nlp.2024.33.
5. Bekarystankyzy, A.; Mamyrbayev, O.; Mendes, M.; Fazylzhanova, A.; Assam, M. Multilingual end-to-end ASR for low-resource Turkic languages with common alphabets. *Sci. Rep*. 2024, 14, 10.1038/s41598-024-64848-1.
6. Tukeyev, U.; Amirova, D.; Karibayeva, A.; Sundetova, A.; Abduali, B. Combined Technology of Lexical Selection in Rule-Based Machine Translation. In *Recent Advances in Systems, Control and Information Technology*; Springer: Cham, Switzerland, 2017; pp. 491–500. doi:10.1007/978-3-319-67077-5_47.
7. Tukeyev, U.; Karibayeva, A.; Abduali, B. Neural Machine Translation System for the Kazakh Language Based on Synthetic Corpora. *MATEC Web Conf*. 2019, 252, 03006. doi:10.1051/matecconf/201925203006.
8. Ahmadnia, B.; Serrano, J.; Haffari, G. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP 2017, Varna, Bulgaria, 2017; INCOMA Ltd, pp. 24–30.
9. Elmadani, K.N.; Buys, J. Neural Machine Translation between Low-Resource Languages with Synthetic Pivoting. In *Proceedings of LREC-COLING 2024*, Torino, Italia, 2024; ELRA and ICCL, pp. 12144–12158.
10. Pontes, J.J.A. Bilingual Sentence Alignment of a Parallel Corpus by Using English as a Pivot Language. In *Proceedings of JISIC*, Quito, Ecuador, 2014; Association for Computational Linguistics, pp. 13–20.
11. Alekseev, A.; Turatali, T. KyrgyzNLP: Challenges, Progress, and Future. arXiv 2024, arXiv:2411.05503v1.
12. Riemland, M. Theorizing Sustainable, Low-Resource MT in Development Settings: Pivot-Based MT between Guatemala's Indigenous Mayan Languages. *Transl. Spaces* 2023, 12, doi:10.1075/ts.22018.rie.
13. Lin, D.; Murakami, Y.; Ishida, T. Towards Language Service Creation and Customization for Low-Resource Languages. *Information* 2020, 11, 67. doi:10.3390/info11020067.
14. Trieu, H.-L.; Tran, V.; Ittoo, A.; Nguyen, L. Leveraging Additional Resources for Improving Statistical Machine Translation on Asian Low-Resource Languages. ACM Trans. Asian Low-Resour. Lang. *Inf. Process*. 2019, 18, 1–22. doi:10.1145/3314936.
15. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012, pp. 2214–2218.
16. Karakanta, A.; Orrego-Carmona, D. Subtitling in Transition: The Case of TED Talks. 2023, doi:10.1075/ata.xx.07kar.
17. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand, 2005; pp. 79–86.
18. Tiedemann, J.; Thottingal, S. OPUS-MT—Building Open Translation Services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 2020.
19. Balahur, A.; Turchi, M. Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. *Comput. Speech Lang*. 2014, 28, 56–75. doi:10.1016/j.csl.2013.03.004.
20. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv 2015, arXiv:1409.0473.

21. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, 2020; pp. 28–39.

22. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th ACL*, 2020; pp. 8440–8451.

23. Costa-jussà, M.R.; Cross, J.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv 2022, arXiv:2207.04672.

24. Hou, X. Research on Translation Corpus Building with the Assistance of AI. In *Proceedings of the 2021 International Conference on Smart Technologies and Systems for IoT*, 2022; pp. 136–140.

25. Reixa, I.D.; et al. Corpus PaGeS: A Multifunctional Resource for Language Learning, Translation and Cross-Linguistic Research. In *Parallel Corpus for Contrastive and Translation Studies*; John Benjamins: Amsterdam, Netherlands, 2019; pp. 103–121.

26. Post, M. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, 2018; pp. 186–191.

27. Abduali, B.; Tukeyev, U.; Zhumanov, Z.; Israilova, N. Study of Kyrgyz-Kazakh Neural Machine Translation. In *Recent Challenges in Intelligent Information and Database Systems*; Nguyen, N.T., et al., Eds.; Springer: Singapore, 2025; CCIS, Vol. 2493, pp. . doi:10.1007/978-981-96-5881-7_21.

28. Zhumanov, Z.; et al. Integrated Technology for Creating Quality Parallel Corpus. In *Advances in Computational Collective Intelligence*; pp. 511–524.

29. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, 2002.

30. Popović, M. chrF: Character n-gram F-score for Automatic MT Evaluation. In *Proceedings of WMT 2015*, 2015.

31. Rei, R.; Sánchez-Cartagena, V.; Popović, M. COMET: A Neural Framework for MT Evaluation. In *Proceedings of EMNLP 2020*, 2020.

32. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv 2016, arXiv:1609.08144.

33. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. arXiv 2020, arXiv:2005.14165.

34. Costa-jussà, M.R.; Cross, J.; Fan, A.; Ghazvininejad, M.; Gu, J.; Gunasekara, C.; He, Y.; Kalbassi, E.; Liptchinsky, V.; Liu, Z.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv 2022, arXiv:2207.04672.

35. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., … & Zhang, Z. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://arxiv.org/abs/2501.12948

36. Ziegler, A.; Kalliamvakou, E.; Li, X.A.; Rice, A.; Rifkin, D.; Simister, S.; Sittampalam, G.; Aftandilian, E. Measuring GitHub Copilot's Impact on Productivity. *Commun*. ACM 2024, 67, 54–63. doi:10.1145/3633453.

37. Gemma Team; Mesnard, T.; et al. Gemma: Open Models Based on Gemini Research and Technology. arXiv 2024, arXiv:2403.08295.

Information about the authors:

Aidana Karibayeva – PhD, is an Acting Associate Professor at the Department of Information Systems of al-Farabi Kazakh National University. Dr. Karibayeva has more than 13 years of experience in artificial intelligence and machine learning. Her research interests include neural networks, deep learning applications, data mining, and speech recognition (Almaty, Kazakhstan, e-mail: karibayeva.aidana@kaznu.edu.kz. ORCID iD: 0000-0002-2023-1573).

Balzhan Abduali – Master of Engineering Science, Senior Lecturer at the Department of Artificial Intelligence and Big Data of al-Farabi Kazakh National University. She has more than 12 years of experience in machine learning. Her research interests include artificial intelligence, machine learning, and data mining (Almaty, Kazakhstan, e-mail: abduali.balzhan@kaznu.kz. ORCID iD: 0000-0003-0140-4181).

Submission received: 8 May, 2026
Revised: 25 May, 2026
Accepted: 25 May, 2026