IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Methodology for creating datasets of parallel sentences in low-resource languages by using AI

Balzhan Abduali[1*], Marek Milosz[2], Ualsher Tukeyev[3], Aidana Karibayeva[4]

[1,3,4]Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan.
[2]Faculty of Electrical Engineering and Computer Science, Lublin University of Technology, Lublin, Poland.

Corresponding author: Balzhan Abduali (*Email: balzhanabdualy@gmail.com*)

Abstract

This study addresses the crucial problem of data scarcity for low-resource languages, with a particular focus on a methodology for creating parallel corpora in two low-resource languages. The lack of large-scale, high-quality bilingual datasets significantly hinders the development of neural machine translation systems for such languages. This study proposes and validates a methodology for creating such datasets. The methodology involves selecting an AI system to generate a parallel corpus based on criteria of accessibility (free access), translation quality, and efficiency, based on a test dataset of 1000 sentences. Subsequently, a substantial parallel corpus of Kyrgyz-Kazakh was created using the selected AI system. However, manual error analysis revealed that approximately 0.5% of the translations contained inaccuracies, indicating the need for further post-editing and model refinement. This study contributes to the development of resources for low-resource language pairs and provides practical guidance on the effective creation of parallel corpora using modern AI systems.

Keywords: AI systems, Kazakh-Kyrgyz language pair, Low-resources languages, Methodology for creating datasets, Parallel sentences.

DOI: 10.53894/ijirss.v8i9.10605
Funding: This research was funded and financed by the grant project "Study of neural models for the formation of speech transcripts and minutes of meetings in Turkic languages" IRN AP23487816 of the Ministry of Science and Higher Education of the Republic of Kazakhstan.
History: Received: 12 August 2025 / **Revised:** 15 September 2025 / **Accepted:** 17 September 2025 / **Published:** 10 October 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

There are over 7,000 different languages in the world [1]. A significant portion (44%) are threatened with extinction, primarily because fewer than 1,000 people speak them. On the other hand, more than half of humanity speaks the 20 most populous languages. Low-resource languages make up the middle. Therefore, their number can be estimated at approximately 4,600. Many of them are official languages in countries with relatively small populations, such as Kazakhstan.

Low-resource languages can be digitally excluded. Effective language automatic processing requires: character encoding, vocabulary, and linguistic rules (grammar, syntax, morphology, pronunciation, etc.). While the problems of character encoding and vocabulary have been better or worse solved by Information Technology (IT), linguistic rules require much more research effort. Even more problems with low-resource languages arise during the development of machine translation systems.

Low-resource languages lack sufficient linguistic resources, such as the large-scale annotated corpora, dictionaries, computational tools, and digitised texts required for developing effective Natural Language Processing (NLP) systems. Unlike high-resource languages such as English, Mandarin Chinese, Hindi, Spanish, or French, low-resource languages face significant challenges in data availability, making it difficult to train and evaluate models for tasks like machine translation, speech recognition, and text classification [2-4].

Turkic languages (such as Kazakh, Kyrgyz, Uzbek, Tatar, Azerbaijani, Turkish, etc.) are also considered low-resource, as most of them suffer from limited availability of high-quality linguistic data and the parallel corpus necessary for training NLP models. In recent years, the development and evaluation of machine translation systems for low-resource Turkic languages have gained increasing attention in the field of natural language processing [5, 6].

For many low-resource languages, the lack of parallel sentence data poses a serious challenge, significantly limiting the development of effective machine translation systems and other NLP applications. Without a high-quality parallel corpus for specific languages, it is impossible to train accurate models for translation and other tasks such as text analysis and generation. One solution for the lack of parallel data is to create a synthetic corpus by generating sentences based on a complete set of suffixes [7, 8]. Another common approach is creating a synthetic corpus using machine translation, translating the source text from one language into another. Monolingual texts are used to create parallel data through machine translation. However, the method faces challenges, primarily poor translation quality. Machine translation for low-resource languages may be insufficiently accurate, especially when the system is trained on limited data. Translation errors may include incorrect meaning transfer, grammatical mistakes, and misinterpretation of phrases, all of which negatively affect the quality of the resulting parallel corpus.

Many studies on low-resource languages have explored the use of a pivot language—most commonly English—for generating a synthetic parallel corpus [9, 10]. This approach helps overcome the lack of direct translation data between two under-resourced languages by leveraging the rich linguistic resources available for English [11]. However, this method is not always effective, especially for closely related languages such as Kyrgyz and Kazakh. Using English as an intermediate step can lead to the loss of semantic and grammatical nuances specific to Turkic languages, ultimately reducing the quality of the resulting data. Therefore, more direct and linguistically informed approaches to corpus creation are necessary for such language pairs.

For the Kazakh-Kyrgyz language pair, publicly available parallel data are extremely limited, which presents a significant challenge for developing and training effective Neural Machine Translation (NMT) models. There is a noticeable lack of research and publications on NMT for the Kazakh-Kyrgyz language pair [12].

An effective methodology and process of creating a corpus of parallel sentences in the Kazakh and Kyrgyz languages are presented in this article.

2. Low-Resource Languages and Creation of Parallel Sentences Datasets

One of the main challenges in working with low-resource languages is the lack of sufficient parallel sentence datasets, which are essential for training and evaluating machine translation models. This scarcity highlights the urgent need to develop and expand a parallel corpus for such languages [13, 14]. For low-resource languages, namely for Turkic languages and Indonesian languages, bilingual dictionaries were obtained for the language pairs Uyghur-Kazakh, Kazakh-Kyrgyz, Kyrgyz-Uyghur [14]. For Asian language pairs – Japanese, Indonesian, Malay paired with Vietnamese, an innovative approach is proposed to build a bilingual corpus from comparable data and phrase pivot translation on an existing bilingual corpus of the languages paired with English [15].

The creation of parallel data is a multifaceted process that requires the use of various methods, especially for low-resource languages. This is where the idea of using AI comes in. Initially, existing parallel datasets can be utilised, which is the simplest and fastest approach. Many languages already have available parallel data, such as:

- OPUS — an extensive repository of parallel datasets in various languages, including data for many language pairs [16]
- TED Talks — subtitles for TED talks are often available in multiple languages, allowing the creation of a parallel dataset [17]
- Europarl — parallel dataset from European Parliament proceedings in multiple languages [18].

OPUS is an open-access collection of multilingual parallel datasets compiled from various sources that are widely used for machine translation development. The OPUS database is maintained and continuously updated by Uppsala University (Sweden) [16]. The OPUS datasets include several key resources:

- Europarl – Official documents of the European Parliament.
- GNOME, KDE, Ubuntu – Software interfaces and technical documentation.
- Tanzil – Multilingual translations of the Quran.
- OpenSubtitles – Movie subtitles in multiple languages.

- WikiMatrix – Multilingual parallel texts extracted from Wikipedia.

OPUS datasets are widely used to evaluate machine translation quality, train multilingual models, and gather data for low-resource languages. Since OPUS texts cover various styles and topics, they are highly suitable for training neural translation systems. OPUS datasets can be accessed via Hugging Face Datasets, OPUS API, or processed using tools such as Moses and FastText Tiedemann and Thottingal [19] and in Balahur and Turchi [20] the quality of translation using real data source from various platforms can be examined, including news websites and internet resources. This allows for an assessment of how machine translation systems perform under actual usage conditions. In recent years, the field of Machine Translation (MT) and NLP has undergone significant changes due to the introduction of deep learning methods and neural networks. One of the first breakthroughs in this area was the transition from statistical translation methods to neural networks, where they proposed a NMT model with an Attention Mechanism, which significantly improved translation results compared to previous methods Bahdanau, et al. [21]. Koehn and Knowles [22] outline and discuss six key challenges in NMT, including data sparsity, handling rare words, and difficulties in translating long sentences. The work emphasised the need for larger training datasets and architectural modifications to overcome these issues. In Conneau, et al. [23] XLM-R was introduced, a cross-lingual language model that demonstrated improvements in low-resource translation by leveraging multilingual pre-training. Meta AI’s "No Language Left Behind" (NLLB) project further advanced translation quality for underrepresented languages Costa-Jussà, et al. [24].

Hou [25] identifies four primary sources of data for corpus construction: open internet resources, corpus data, user-generated content, and machine-generated data. The paper also outlines four approaches to creating an AI-assisted corpus: using third-party open sources, crowdsourcing, training models on proprietary data, and joint Corpus creation by humans and machines. The crowdsourcing approach is particularly noteworthy, as it allows for the expansion of the corpus through contributions from both professional and non-professional translators. In addition, the author emphasises the importance of data multimodality (video, audio, text) and the shift from traditional dictionaries to the concept of multilingual terminology management. The paper also highlights the challenges of integrating AI algorithms into real-world applications, particularly in education and real-time online translation. Reixa, et al. [26] present recent research in the field of a parallel corpus, covering both the development of new resources and the improvement of methods for their utilisation. The volume discusses the role of parallel corpora using the German-Spanish language pair as an example in translation studies and contrastive linguistics, as well as technical aspects of alignment, annotation, and search. The paper also introduces current projects on the creation of parallel and multimodal corpora in Europe, including real-world use cases. Furthermore, it highlights the significance of such corpora for developing bilingual resources, language teaching, and machine translation tasks.

These studies were reviewed to identify the effective strategies for creating and utilising parallel corpora for low-resource languages. They provide valuable insights into the advantages and limitations of current datasets, translation models, and data acquisition methods. Building upon this foundation, the present research focuses on generating a high-quality corpus for the Kazakh-Kyrgyz language pair.

3. Methods and Materials

3.1. Developed Methodology

Figure 1 shows the three-phase methodology for creating datasets of parallel sentences in low-resource languages using AI.

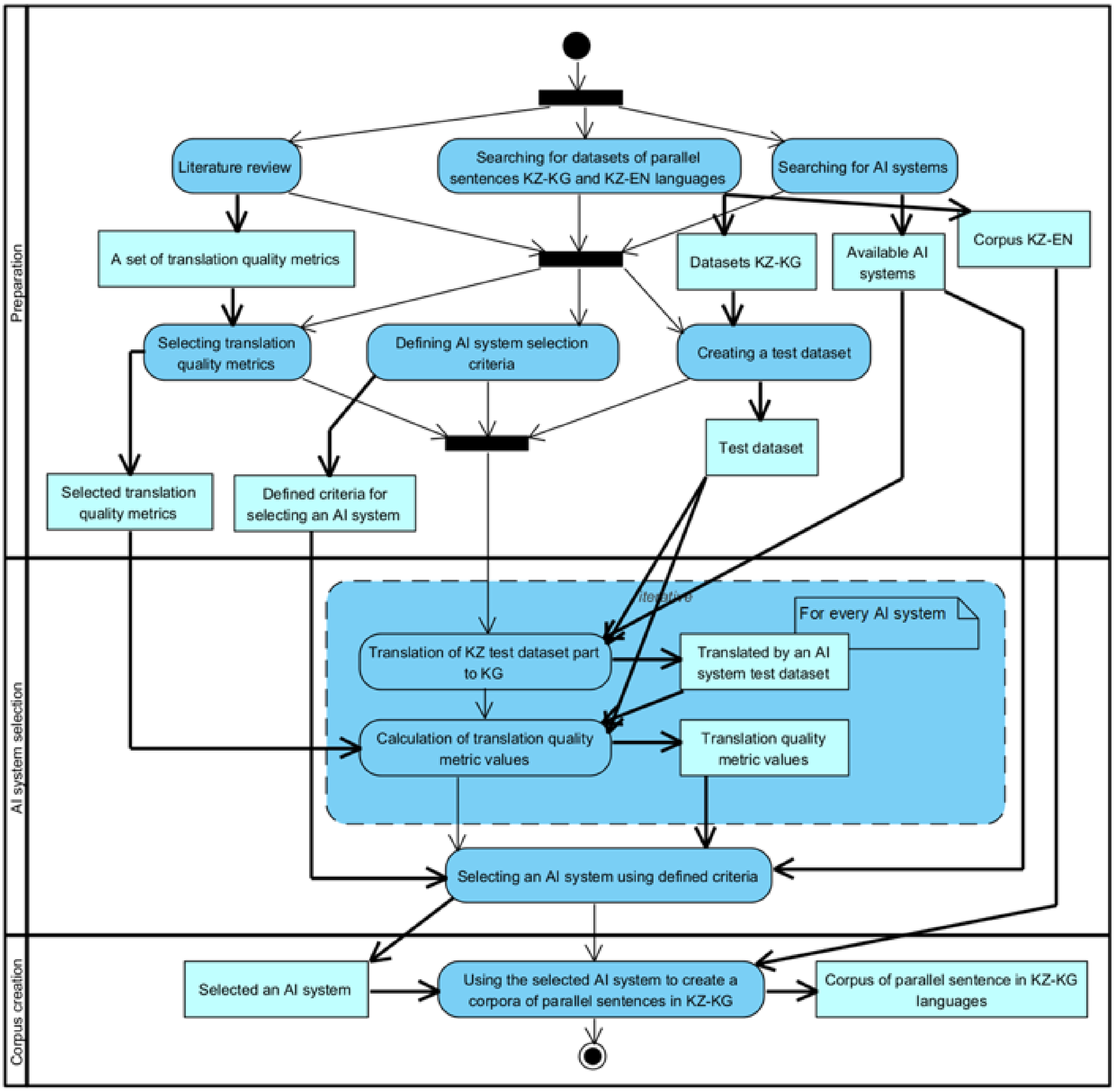
Preparation
AI system selection
Corpus creation
Literature review
Searching for datasets of parallel sentences KZ-KG and KZ-EN languages
Searching for AI systems
A set of translation quality metrics
Datasets KZ-KG
Available AI systems
Corpus KZ-EN
Selecting translation quality metrics
Defining AI system selection criteria
Creating a test dataset
Test dataset
Selected translation quality metrics
Defined criteria for selecting an AI system
iterative
For every AI system
Translation of KZ test dataset part to KG
Translated by an AI system test dataset
Calculation of translation quality metric values
Translation quality metric values
Selecting an AI system using defined criteria
Selected an AI system
Using the selected AI system to create a corpora of parallel sentences in KZ-KG
Corpus of parallel sentence in KZ-KG languages

Figure 1.
Detailed Workflow of Developed Methodology (abbreviations of lan-guage names: KZ – Kazakh, KG – Kyrgyz, EN – English).

The methodology (Figure 1) consists of three phases:
1. Preparation.
2. AI system selection.
3.Corpus creation.
During the first phase of research, the following activities will be carried out:

- Literature review – conducting an extensive review of the current state of research in machine translation for low-resource languages, focusing on the Kazakh-Kyrgyz language pair and related technologies.
- Searching for datasets of parallel sentences KZ-KG and KZ-EN languages – identifying available parallel corpus for the Kazakh-Kyrgyz and Kazakh-English language pairs from open data sources, such as OPUS, and evaluating their quality and coverage.
- Searching for AI systems – investigating available AI systems and NMT models that can be used for the Kazakh-Kyrgyz language pair, focusing on pre-trained models and open-source solutions.
- Selecting translation quality metrics – choosing relevant evaluation metrics based on a literature review to assess the quality and performance of the translation systems.
- Defining AI system selection criteria – establishing clear criteria for selecting the most suitable AI model for translation tasks, considering factors like model architecture, efficiency, training data availability, and translation accuracy.

- Creating a test dataset – curating a test dataset of parallel sentences from selected sources, ensuring it is balanced and representative of different domains for comprehensive evaluation of translation quality.

The second phase of the methodology consists of iteratively performed translation quality tests by individual AI systems (activities: Translation of KZ test dataset part to KG and Calculation of translation quality metric values), final AI system selection (activity: Selecting an AI system using defined criteria), and a method of step-by-step elimination of possibilities.

The third and final phase, Corpus creation, will consist of the translation of the KZ part of the corpus KZ-EN into the KG language. Details of this process are shown in Figure 2.

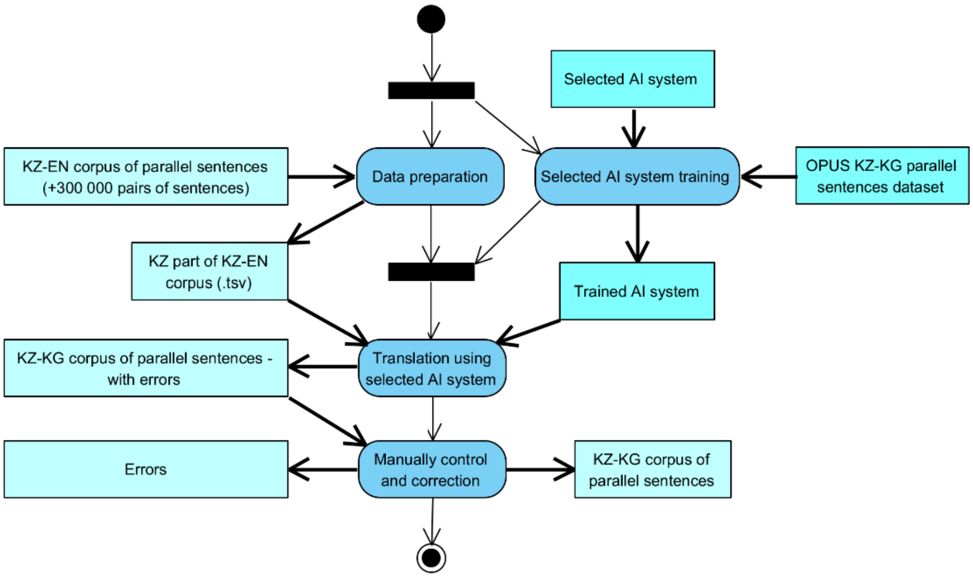
Selected AI system
KZ-EN corpus of parallel sentences (+300 000 pairs of sentences)
Data preparation
Selected AI system training
OPUS KZ-KG parallel sentences dataset
KZ part of KZ-EN corpus (.tsv)
Trained AI system
KZ-KG corpus of parallel sentences - with errors
Translation using selected AI system
Errors
Manually control and correction
KZ-KG corpus of parallel sentences

Figure 2.
Corpus Preparation Process Using an AI System.

During the computational work (translation), the hardware and software presented in Table 1 were used.

Table 1.
Server specifications for corpus translation.

Specification	**Value**
Graphics Card	NVIDIA RTX 4090 24 GB
Graphics Memory Type	GDDR6X
Graphics Memory Size	24 GB
CUDA Cores	16,384
Core Clock Speed	2.23 GHz
RAM	128 GB DDR4/DDR5
RAM Type	DDR4 or DDR5
Network Interfaces	10Gb Ethernet (or higher)
Power Supply	850 W or higher

3.2. Translation Quality Metrics

Translation quality metrics are essential for evaluating the effectiveness of machine translation, and commonly used metrics for assessing translation quality include BLEU, TER, and WER. However, these metrics often provide limited evaluation, which is why additional metrics for syntactic and semantic accuracy, such as COMET and chrF, are also used to offer a more comprehensive assessment of translation quality. The SacreBLEU baselines in the corpus use the following metrics from SacreBLEU [27]. To assess the quality of the translated text, the following metrics are utilised:

- BLEU (Bilingual Evaluation Understudy) measures the overlap between machine-generated translations and reference translations. Sentence-level BLEU scores were calculated using the `sentence_bleu` function from SacreBLEU in Python [28].
- WER (Word Error Rate) measures the number of errors (insertions, de-letions, and substitutions) in the translated text. A lower WER indicates better translation accuracy.
- ChrF (Character n-gram F-score) evaluates translations at the character level, making it particularly useful for assessing morphologically rich languages. The “sacrebleu.sentence_ChrF” function was utilised to compute segmental-level ChrF scores, which were then averaged [29].
- METEOR (Metric for Evaluation of Translation with Explicit ORdering) evaluates translations based on synonymy, stemming, and word order.
- COMET is a neural-based evaluation metric that considers semantic adequacy and fluency [30].

Table 2 was compiled based on an analysis of the reviewed scientific publications. It incorporates indicators drawn from these sources, and the final metric was approximately derived from the aggregated results [27-30]. Table 2 summarises the key evaluation metrics used to compare the performance of the selected translation models and explains how these results inform model selection.

Table 2.
Interpretation of Evaluation Metrics and Model Selection Criteria

Metrics	Value range	Interpretation
BLEU	0-100	>50 – Excellent 30–50 – Good 10–30 – Fair <10 – Poor
WER	0–1 Example: WER of 0.8 means that there is an 80% error rate for compared sentences.	<0.2 – Excellent 0.2 – 0.4 Acceptable >0.4– Poor
ChrF	0-100	>60 – Excellent 40–60 – Good <40 – Weak
METEOR	0-1	>0.5 – Excellent 0.3-0.5 – Moderate <0.3 – Low
COMET	0-1	>0.5 – High quality 0.3–0.5 – Acceptable <0.3 – Weak

Source: Reixa, et al. [26]**;** Post [27]**;** Papineni, et al. [28]**;** Popović [29] **and** Rei, et al. [30].

3.3. Formatting of Mathematical Components

For the evaluation of the selected translation AI models, a parallel dataset was extracted from the OPUS repository, as presented in Table 3. From this dataset, 1, 000 Kazakh–Kyrgyz sentence pairs were selected. The Kyrgyz part of the corpus was treated as the gold reference, while the Kazakh sentences were used as the source for translation. An inspection of available resources [16] revealed that there are several dozen datasets that include both Kazakh and Kyrgyz [16]. As discussed in Abduali, et al. [31] these data were utilised for the training of NMT models, providing a foundation for evaluating model performance. However, only three of them are suitable for use in machine translation tasks shown in Table 3.

Table 3.
Parallel datasets for machine translation in KZ-KG languages.

Dataset name	Dataset link	Number of parallel sentences
OPUS Tiedemann [16]	https://opus.nlpl.eu/results/kk&ky/Corpus-result-table	102 345
NTREX Federmann, et al. [32]	https://huggingface.co/datasets/davidstap/NTREX	2 000
Flores 101 Goyal, et al. [33]	https://huggingface.co/datasets/severo/flores_101	2 000

To create a high-quality parallel corpus for a low-resource Kazakh-Kyrgyz pair, the first step involved selecting an existing monolingual Kazakh dataset consisting of approximately 300,000 sentences. The Kazakh corpus used in this study was obtained from an open-access dataset presented in Zhumanov and Tukeyev [34]. The original dataset consists of 302530 parallel sentence pairs in Kazakh and English. For the purposes of this research, only the Kazakh portion of the corpus was extracted and used as the source data for generating a Kazakh–Kyrgyz parallel corpus via machine translation.

3.4. AI Systems and Selecting Criteria

Several translation systems were considered for this task and evaluated according to a set of predefined criteria.
The selection criteria included (in order of importance):

- Accessibility – the availability of the system for large-scale use, with preference for free or open access.
- Translation quality – the linguistic accuracy and contextual relevance of the output.
- Translation efficiency – the capability of the system to handle and translate large amounts of text.

Based on Internet searching results, the following AI-model translation systems were selected for comparison and evaluation: Google Translate, NLLB, ChatGPT, DeepSeek, GitHub Copilot, and Gemma. There are quite a lot of AI-model translation systems on the Internet, such as DeepL, Claude, etc. However, the purpose of selecting AI-model translation systems is to generate significant volumes of parallel corpora, which for many systems requires significant computational time and finance, and thus does not meet the accessibility criterion.

The AI-models under consideration can be divided into two groups: specialised (Google Translate, NLLB, GitHub Copilot) and general-purpose (ChatGPT, DeepSeek, and Gemma).

One of the most well-known and popular translation tools is Google Translate, which uses neural networks to train translation models based on large volumes of bilingual data. Since its launch in 2006, the system has greatly improved thanks to the use of Google Neural Machine Translation (GNMT), which introduced sequence models with deep neural network-based learning. Today, Google Translate supports over 100 languages and is one of the most widely used translation tools worldwide [35].

NLLB (No Language Left Behind), developed by Meta, is a machine translation specialized AI-model designed to support a large number of languages, including rare and under-resourced ones. This is a significant achievement in the field of machine translation, as the model demonstrates excellent results in translating languages with limited training data. NLLB also employs deep learning and transformer-based approaches to improve translation quality [24].

GitHub Copilot, developed by GitHub and powered by OpenAI, is an artificial intelligence tool that helps developers write code. GitHub Copilot uses the OpenAI Codex model, which allows generating code in various programming languages based on text prompts. This can also be useful for translation tasks, where automating code generation can speed up the process of integrating translation into software systems. GitHub Copilot significantly boosts productivity by providing solutions to programming tasks in real-time [36].

GPT (Generative Pretrained Transformer), developed by OpenAI, is one of the most successful examples of applying transformers in NLP. GPT is a general-purpose AI-model used not only for translation but also for a variety of other tasks, such as text generation, summarization, and dialogue systems. While GPT has shown good results in the context of natural language processing, its application for translation is limited and requires additional fine-tuning on specialised translation datasets to improve quality [37].

DeepSeek is a general-purpose AI-model aimed at improving translation quality by applying more complex neural network architectures. An important aspect of DeepSeek's operation is the use of multitask learning to process different types of texts and increase the model's flexibility [38].

Gemma is a family of lightweight, open-source large language models (LLMs) developed by Google. Introduced in March 2024, Gemma is based on the research and technology behind Google's Gemini models. Designed to be efficient and accessible, Gemma models are available in two sizes—2 billion and 7 billion parameters—and come with both pretrained and fine-tuned checkpoints [39].

4. Results of Using the Methodology

4.1. Translation Quality Assessment Results by Various AI Systems

Table 4 presents the results obtained based on the evaluation metrics, providing an overview of the translation quality achieved by the selected AI systems.

Table 4.
Evaluation results of translation models on 1,000 Kazakh sentences from the OPUS datasets, assessed using BLEU, ChrF, Meteor, Comet, and WER metrics.

Metrics System	BLEU	WER	METEOR	COMET	ChrF
Google Translator	14.0	0.92	0.078	0.692	23.01
Chat GPT	**36.6**	0.87	**0.151**	0.818	31.56
Nllb-200-3.3	30.4	0.88	0.126	0.755	27.12
DeepSeek	33.0	**0.87**	0.145	**0.819**	**31.56**
Copilot	26.2	0.87	0.146	0.812	31.38
Gemma-2-27b	31.0	0.87	0.136	0.802	30.25

4.2. Selection of an AI System

Table 5.
Time Consumption for Translating Kazakh Texts Using Different AI systems.

Indicator	Gemma-2-27	NLLB-200-3.3
Translation speed	3 sentences per minute	300 sentences per minute
Time for full translation of 302 530 sentences	~2.5 months	~2 days

As demonstrated in the comparison table, the Gemma model requires a sig-nificantly longer time to process the text body, whereas the NLLB model com-pletes the same task in a substantially shorter period. Consequently, due to its higher efficiency and faster performance, the NLLB model was selected for further use.

4.3. Parameters of the Corpus Created

As shown in Table 6, the resulting volume of parallel sentence pairs was stored in a single TSV file with a total size of 139.5 MB, containing approximately 10,000,000 words.

Table 6.
Kazakh-Kyrgyz Parallel Corpus Created via Automatic Translation of 302530 Kazakh Sentences Using the NLLB Model.

Corpus name	Quantity of sentences	Quantity of words	Size of file
KZ-KG	302 530	~ 10 000 000	139.5 MB

4.4. Translation Errors and Their Correction

Table 7 presents examples of translation errors along with their descriptions.

Table 7.
Examples of sentences with errors and missing elements identified from the translated KZ-KG corpus.

Original Kazakh text (presented in Latin script)	Translated text into Kyrgyz using NLLB-200-3.3 (presented in Latin script)	Explanation of the identified errors
Memleket basshysy Nursultan Nazarbaev Resey Federaciyasynyn Prezidenti Vladimir Putinge Donetsk mańynda bolǵan TU-154 jolaýshylar uşhagynyn apatynan adamdardyń qaza bolýyna baılanysty kóńil aıtty.	Prezident Vladimir Putinge Donečk shaarynyn janyndaǵy TU-154 uçagynyn kyıraşynan kaza bolgondorgo bajlanyshtuu kóńil aıtty.	The Kazakh phrase "Memleket basshysy Nursultan Nazarbaev Resey Federaciyasynyn Prezidenti (Head of State Nursultan Nazarbayev President of the Russian Federation)" translated to Kyrgyz only as "Prezident (President)". The translation must be as "Mamleket bashchysy Nursultan Nazarbaev Rossija Federacijasynyn Prezidenti" The Kazakh word "jolaýshylar" (passengers) was not translated. The correct Kyrgyz equivalent should be "jürgünçülör." The Kazakh word "adamdardyń (peoples)" was not translated. The correct Kyrgyz equivalent should be "adamdar".
QR prem'er-ministri Asqar Mamin Arys qalasyndaǵy zardap shekken úılerdi, áleýmettik nısandar men injenerlik jeliilerdi qalpyna keltirý jumystarynyń barysymen tanysý maqsatynda jumys saparymen Túrkistan oblysyna bardy.	Kyrgyzstandyn prem'er-ministri Askar Mamin shaardaghy kyıragan üıldördü, socialdyk obyektterdi jana injenerdik tarmaktardy kalybyna keltirüü ishterinin jürüşü menen taanyshuu maksatynda Türkstan oblastyna ish sapary menen bardy.	The Kazakh abbreviation "QR (RK)" was incorrectly translated as "Kyrgyzstandyn (Kyrgyzstan)"; it should be translated as "KR (RK)" or "Kazakstannyn (Kazakhstan)." The Kazakh phrase "Arys qalasyndaǵy (in the city of Arys)" was translated only as "shaardaghy (in the city)," omitting the name of the city.
QR prem'er-ministri Asqar Maminnıń tóraghalygymen ótken úkimet otyrysynda "Eńbek" nátizheli jumyspen qamtýdy jáne jappai kásipkerlikti damytýdyń 2017–2021 jyldarǵa arnalǵan memlekettik baǵdarlamasyn iske asyrý barysy qaraldy.	Ökmöttün jıyynynda 2017–2021-zyldarga "Emgekti" natyıjaluu paıdalanuu jana massalyk ishkerdikti önüktürüü boıuncha mamlekettik programmany ishke ashyruunun jürüşü qaraldy.	The Kazakh phrase "QR prem'er-ministri Asqar Maminnıń tóraghalygymen ótken (chaired by the Prime Minister of the Republic of Kazakhstan, Askar Mamin)" was not translated.

During the manual verification process, several types of errors were identified, including semantic errors (incorrect translations of meanings) and lexical errors (missing or incorrectly translated words). The total number of identified errors was 25,584, of which 17680 were corrected. These corrections contributed to a significant improvement in the overall quality of the parallel corpus. As shown in Table 6, the total number of words in the parallel corpus is approximately 10,000,000. Considering the linguistic similarity between Kazakh and Kyrgyz, it can be assumed that the Kyrgyz portion contains around 5,000,000 words. Based on the identified 25,584 translation errors, the estimated error rate is approximately 0.5% of the translations that contained inaccuracies in the translation of abbreviations and repetitions, which were removed by a specially developed program.

5. Discussion

The evaluation results clearly indicate that Chat GPT and DeepSeek are the most suitable models for the translation of Kazakh to Kyrgyz in Table 4, based on their high performance across multiple metrics. Chat GPT stood out with superior results in COMET (0.818) and ChrF (31.56), indicating its ability to produce fluent and accurate translations. DeepSeek, while slightly behind Chat GPT, also showed competitive performance with a COMET score of 0.819 and a ChrF score of 31.56, making it another strong contender. In contrast, NLLB-200-3.3 and Gemma-2-27b performed adequately but fell short in comparison to the leading models. Their results suggest that while they can handle translation tasks, they may not provide the same accuracy and fluency required for high-quality corpus creation. Given these findings, Chat GPT and DeepSeek are identified as the most promising candidates for future research and the creation of a high-quality Kazakh-Kyrgyz parallel corpus. However, these systems do not provide free access suitable for processing large-scale datasets. Following them in terms of performance are the Gemma and NLLB models, both of which can be used via API. Among these, Gemma showed slightly more accurate translation results and was initially selected for further use. However, as demonstrated in Table 5, Gemma required a relatively long time to process each sentence. Gemma requires, on average, 20 seconds to translate a single sentence. This means that translating a corpus of 302,530 sentences would take approximately 2.5 months, which is a significantly long processing time. This estimate holds even when the model is run on a high-performance computer equipped with a powerful GPU and ample memory. Therefore, the NLLB model was chosen for translating the corpus, as it offered an optimal compromise between processing efficiency and translation quality.

The study focused on the automatic translation of sentences from Kazakh to Kyrgyz using the NLLB-200-3.3 model. The translation process was carried out on a high-performance computing system, which enabled the efficient processing of large volumes of data and the creation of a Kazakh-Kyrgyz parallel corpus. The results indicated a generally high quality of translation, particularly in standard syntactic constructions and commonly used expressions. However, several errors and inconsistencies were observed upon manual inspection of the translated output, revealing some of the limitations of the model when applied to closely related Turkic languages.

Table 7 presents additional examples of incorrect or inaccurate translations. One of the prominent issues was the incorrect handling of abbreviations. For example, the abbreviation “QR” (short for Qazaqstan Respublikasy, Republic of Kazakhstan) was occasionally mistranslated as “Kyrgyzstan” (Kyrgyzstan), “Kyrgyz” (Kyrgyz), “Kyrgyz Respublikasy” (Republic of Kyrgyz), suggesting that the model may have incorrectly inferred meaning based on contextual frequency rather than semantic accuracy. This points to challenges in the model’s ability to distinguish between similar geopolitical terms within closely related languages. The NLLB model also struggles with the translation of abbreviated terms such as JSHS (Jauapkershilik shekteýli seriktestik, “Limited Liability Partnership”), UBT (Ulttyq Biryńǵaı test, “Unified National Test”), and AQ (Akcionerlik Qoǵam, “Joint-Stock Company”). In some cases, these abbreviations are omitted entirely, while in others, they are replaced with unrelated or inaccurate words, leading to a loss of meaning. In addition, there were instances where multiple Kazakh words were compressed into a single Kyrgyz word, leading to a loss of semantic content. For example, the Kazakh sentence “QR prem'er-ministri Asqar Maminnıń tóraghalygymen ótken úkimet otyrysynda ‘Eńbek’ nátizheli jumyspen qamtýdy jáne jappai kásipkerlikti damytýdyń 2017–2021 jyldarǵa arnalǵan memlekettik baǵdarlamasyn iske asyrý barysy qaraldy” (English: “At the government meeting chaired by the Prime Minister of the Republic of Kazakhstan, Askar Mamin, the progress of the implementation of the state program ‘Enbek’ for productive employment and the development of mass entrepreneurship for 2017–2021 was reviewed”) was translated into Kyrgyz as “Ökmöttün jıyynynda 2017–2021-zyldarga ‘Emgekti’ natyıjaluu paıdalanuu jana massalyk ishkerdikti önüktürüü boıuncha mamlekettik programmany ishke ashyruunun jürüşü qaraldy” (English: “At the government meeting, the progress of the implementation of the state program ‘Emgek’ for effective employment and the development of mass entrepreneurship for 2017–2021 was reviewed”). Notably, the NLLB model omits the entire clause “QR prem'er-ministri Asqar Maminnıń tóraghalygymen ótken” (“chaired by the Prime Minister of the Republic of Kazakhstan, Askar Mamin”) and fails to translate it. Instead, it omits or provides inaccurate translations for such phrases, which leads to incomplete or incorrect translations. Such reductions compromise the quality of sentence alignment and the overall equivalence of meaning in the parallel corpus, which are critical for downstream tasks such as machine translation training and evaluation. Despite these challenges, the NLLB-200-3.3 model demonstrated potential for generating parallel data for low-resource language pairs within the Turkic family. However, to ensure a high-quality corpus, post-editing remains essential, particularly for domain-specific terms, abbreviations, and named entities. Moreover, it is recommended that the model be fine-tuned on dedicated Kazakh-Kyrgyz datasets to improve its accuracy and contextual understanding in future applications.

After manual verification and identification of translation errors, corrections were made wherever possible. As a result, the parallel sentences corpus has been significantly improved, with an overall quality increase based on the reduction of errors, making it more reliable for further research and practical applications. It is important to note that these errors were identified during the first round of verification, and additional rounds of checks will be conducted to identify and correct any remaining issues. As shown in Table 6, the number of tokens decreased after translation, which is one aspect. On the other hand, lexical errors were identified, as seen in Table 7, where certain words were either not translated at all or were translated using abbreviations.

The preliminary stage of corpus cleaning removed only the most obvious errors, including abbreviations, word duplication, and syntactic and semantic inconsistencies. Future efforts will focus on expanding the cleaning process to include identifying and correcting other semantic and lexical inconsistencies.

6. Conclusions and Future Work

In this study, various AI systems for generating parallel corpus were checked on the test dataset, and the most suitable system was selected based on predefined criteria such as accessibility, translation quality, and efficiency. Using the proposed methodology on the selected large language model NLLB 200 3.3B, a parallel corpus of 302,530 sentence pairs was successfully created for the Kyrgyz-Kazakh language pair.

However, the generated corpus was not flawless; several manual translation errors were identified. The ratio of errors of the translations contained inaccuracies in the translation of abbreviations and repetitions, which were removed by a specially developed program, was small: 0.5% of all words in the developed corpus.

Due to the rapid development and refinement of AI models and tools, it will be possible to improve the parameters achieved in this study. It will also be possible to refine the proposed methodology.

Future work will focus on correcting the remaining errors and improving the data. The generated parallel Kyrgyz-Kazakh corpus will be used to train “lighter” LLM models to produce high-quality translations that can be implemented on lower-end computing hardware. Additionally, the current proposed methodology will be applied to generating of parallel corpora for other Turkic languages and create high quality NMT systems for low-resource Turkic languages.

References

[1] Ethnologue, "How many languages are there in the world?," 2025. https://www.ethnologue.com/insights/how-many-languages/

[2] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey, "Selection criteria for low resource language programs," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4543-4549.

[3] R. Ivasubramanian, T. Umamaheswari, S. B. G. Babu, R. Inakoti, and J. Y. M. Salome, Dr., "Natural language processing in low-resource language contexts," *Frontiers in Health Informatics*, vol. 13, no. 8, pp. 1578–1584, 2024.

[4] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," *Natural Language Processing*, vol. 31, no. 2, pp. 183-197, 2025. https://doi.org/10.1017/nlp.2024.33

[5] A. Bekarystankyzy, O. Mamyrbayev, M. Mendes, A. Fazylzhanova, and M. Assam, "Multilingual end-to-end ASR for low-resource Turkic languages with common alphabets," *Scientific Reports*, vol. 14, no. 1, p. 13835, 2024. https://doi.org/10.1038/s41598-024-64848-1

[6] U. Tukeyev, D. Amirova, A. Karibayeva, A. Sundetova, and B. Abduali, "Combined technology of lexical selection in rule-based machine translation," presented at the International Conference on Computational Collective Intelligence, 2017.

[7] U. Tukeyev, A. Karibayeva, and B. Abduali, "Neural machine translation system for the kazakh language based on synthetic corpora," *MATEC Web of Conferences*, vol. 252, p. 03006, 2019.

[8] A. Karibayeva, B. Abduali, and D. Amirova, "Formation of the synthetic corpus for kazakh on the base of endings complete system," in *Turklang-2018 Proceedings of International Conference*, 2018.

[9] B. Ahmadnia, J. Serrano, and G. Haffari, "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, pp. 24-30.

[10] K. N. Elmadani and J. Buys, "Neural machine translation between low-resource languages with synthetic pivoting," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 12144-12158.

[11] J. d. J. A. Pontes, "Bilingual sentence alignment of a parallel corpus by using English as a pivot language," in *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, 2014, pp. 13-20.

[12] A. Alekseev and T. Turatali, "KyrgyzNLP: challenges, progress, and future," presented at the International Conference on Analysis of Images, Social Networks and Texts, 2024.

[13] M. Riemland, "Theorizing sustainable, low-resource MT in development settings: Pivot-based MT between Guatemala’s indigenous Mayan languages," *Translation Spaces*, vol. 12, no. 2, pp. 231-254, 2023. https://doi.org/10.1075/ts.22018.rie

[14] D. Lin, Y. Murakami, and T. Ishida, "Towards language service creation and customization for low-resource languages," *Information*, vol. 11, no. 2, p. 67, 2020. https://doi.org/10.3390/info11020067

[15] H.-L. Trieu, D.-V. Tran, A. Ittoo, and L.-M. Nguyen, "Leveraging additional resources for improving statistical machine translation on asian low-resource languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 18, no. 3, pp. 1-22, 2019. https://doi.org/10.1145/3314936

[16] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," *Lrec*, vol. 2012, pp. 2214-2218, 2012.

[17] A. Karakanta and D. Orrego-Carmona, *Subtitling in transition: The case of TED Talks*. Netherlands: John Benjamins Publishing Company, 2023, pp. 130-156.

[18] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," presented at the The Tenth Machine Translation Summit Proceedings of Conference, 2005.

[19] J. Tiedemann and S. Thottingal, "OPUS-MT--building open translation services for the world," presented at the Annual Conference of the European Association for Machine Translation, 2020.

[20] A. Balahur and M. Turchi, "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis," *Computer Speech & Language*, vol. 28, no. 1, pp. 56-75, 2014. https://doi.org/10.1016/j.csl.2013.03.004

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017.

[23] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[24] M. R. Costa-Jussà *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[25] X. Hou, "Research on translation corpus building with the assistance of ai," presented at the 2021 International conference on Smart Technologies and Systems for Internet of Things (STS-IOT 2021), 2022.

[26] I. D. Reixa, S. F. Lanza, T. J. Juliá, E. L. Lamas, and B. Lübke, *Corpus PaGeS: A multifunctional resource for language learning, translation and cross-linguistic research*. Amsterdam, Netherlands: John Benjamins Publishing Company, 2019, pp. 103-121.

[27] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation*, 2018.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.

[29] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015, pp. 392-395.

[30] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," *Proceedings of EMNLP* 2020.

[31] B. Abduali, U. Tukeyev, Z. Zhumanov, and N. Israilova, "Study of kyrgyz-kazakh neural machine translation," presented at the Asian Conference on Intelligent Information and Database Systems, 2025.

[32] C. Federmann, T. Kocmi, and Y. Xin, "NTREX-128–news test references for MT evaluation of 128 languages," in *Proceedings of the First Workshop on Scaling up Multilingual Evaluation*, 2022, pp. 21-24.

[33] N. Goyal *et al.*, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics,* vol. 10, pp. 522-538, 2022.

[34] Z. Zhumanov and U. Tukeyev, "Integrated technology for creating quality parallel corpora," presented at the International Conference on Computational Collective Intelligence, 2021.

[35] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144,* 2016.

[36] A. Ziegler *et al.*, "Measuring github copilot's impact on productivity," *Communications of the ACM,* vol. 67, no. 3, pp. 54-63, 2024. https://doi.org/10.1145/3633453

[37] T. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems,* vol. 33, pp. 1877-1901, 2020.

[38] D. Guo *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948,* 2025.

[39] G. Team *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295,* 2024.