KazParC: Kazakh Parallel Corpus for Machine Translation

Rustem Yeshpanov, Alina Polonskaya, Huseyin Atakan Varol
Institute of Smart Systems and Artificial Intelligence
Nazarbayev University, Astana, Kazakhstan
{rustem.yeshpanov, alina.polonskaya, ahvarol}@nu.edu.kz

Abstract

We introduce KazParC, a parallel corpus designed for machine translation across Kazakh, English, Russian, and Turkish. The first and largest publicly available corpus of its kind, KazParC contains a collection of 371,902 parallel sentences covering different domains and developed with the assistance of human translators. Our research efforts also extend to the development of a neural machine translation model nicknamed Tilmash. Remarkably, the performance of Tilmash is on par with, and in certain instances, surpasses that of industry giants, such as Google Translate and Yandex Translate, as measured by standard evaluation metrics, such as BLEU and chrF. Both KazParC and Tilmash are openly available for download under the Creative Commons Attribution 4.0 International License (CC BY 4.0) through our GitHub repository.

Keywords: English, Kazakh, KazParC, machine translation, parallel corpus, Russian, Tilmash, Turkish

1. Introduction

Machine translation (MT) refers to the use of computer systems tasked to automatically translate between languages with or without human intervention (Hutchins, 1995). Beyond its fundamental role in linguistic translation, MT demonstrates great versatility extending to practical applications in various domains. These applications include accessing and gaining information in another language (Wiesmann, 2019), language learning and teaching (Lee, 2020), facilitating professional translation tasks (Craciunescu et al., 2004), and providing multilingual customer service (Barrera et al., 2016; Lewis et al., 2012).

MT approaches include rule-based, statistical, and neural methods. Statistical machine translation (SMT) gained ground over rule-based machine translation (RBMT) in the late 1990s thanks to its ability to learn from large bilingual corpora, making it more adaptable to different language pairs and contexts. However, the dominance of SMT was challenged by the emergence of neural machine translation (NMT) in the mid-2010s, when NMT models with the sequence-to-sequence network (Sutskever et al., 2014) displayed unprecedented translation quality and fluency, as well as the ability to handle a wide range of linguistic phenomena (Stahlberg, 2020), leading to their widespread adoption.

Modern MT models are typically trained on large-scale parallel corpora containing pairs of source and target language texts, also known as bitexts (Jurafsky and Martin, 2009). Similar to many other domains of natural language processing (NLP), MT faces a resource imbalance. While some languages, such as English, Japanese, Mandarin, and Spanish (Koehn, 2005; Pryzant et al., 2018), benefit from a wealth of parallel corpora, linguistic tools, and pre-trained models, lower-resourced languages are often in a state of resource paucity, yearning for the abundance available to their higher-resourced counterparts.

This paper focuses on NMT from and to Kazakh, a Turkic language that utilises the Cyrillic script and has an estimated 13 million native speakers (Campbell and King, 2020; Johanson, Lars and Csató, Éva Á., 2021). Notwithstanding notable recent advances in Kazakh NLP (Mussakhojayeva et al., 2022; Yeshpanov et al., 2022), the language remains relatively lower-resourced and in need of further research efforts and resource development, with MT, especially in terms of the availability of parallel data, being one of these critical areas.

In this study, we attempt to bridge this source scarcity by presenting a parallel corpus for four languages. The corpus includes parallel data for two Turkic languages, Kazakh and Turkish, belonging to the Kypchak and Oghuz branches, respectively. We also provide parallel data for two Indo-European languages, English and Russian, representing the West-Germanic and Slavic branches, in turn. Furthermore, we introduce an NMT model trained using the aforementioned parallel corpus. The experimental results demonstrate that our model achieves competitive and, in some cases, even superior performance to that of industry giants, when evaluated using standard evaluation metrics, such as bilingual evaluation understudy (BLEU) (Papineni et al., 2002) and character F-score (chrF) (Popović, 2015).

The structure of the paper is as follows: Section 2 offers a review of previous research within the field. Section 3 delves into the details of data sources, collection, pre-processing, partitioning methods, and corpus statistics. Section 4 is comprised of subsections focusing on the experimental design, evaluation metrics, and experimental results. Section 5 provides a discussion of the obtained results. Section 6 concludes the paper and outlines potential areas for future work.

2. Related Work

Kazakhstan implements a trilingual policy, designating Kazakh as its official state language, Russian as the language for interethnic communication, and English as the language essential for effective global economic integration (Sanders, 2016). Consequently, most research in Kazakh MT has predominantly revolved around Russian or English as either source or target translation languages.

Early attempts at Kazakh↔English and Kazakh↔Russian MT involved building structural transfer rules on Apertium (Forcada et al., 2011; Shormakova and Sundetova, 2013; Sundetova et al., 2014), implementing morphological segmentation techniques to address the rich morphology of the Kazakh language (Assylbekov and Nurkas, 2014; Bekbulatov and Kartbayev, 2014), and exploring sentence alignment through Russian lemmatisation and bilingual dictionaries (Assylbekov et al., 2016; Myrzakhmetov and Makazhanov, 2016).

Regarding Kazakh↔Turkish MT, the scarcity of parallel training data has posed a significant limitation, resulting in a small number of research studies dedicated to the development of translation systems for these two Turkic languages (Kessikbayeva and Cicekli, 2021). As an illustration, in the study by Bayatli et al. (2018), efforts were made to address this data deficit by manually translating about a thousand Kazakh treebank sentences (Tyers and Washington, 2015) into Turkish to create an RBMT system. This system achieved a BLEU score of 0.17 and word error rate (WER) of 0.46.

It is worth noting that parallel data for the aforementioned language pairs did exist to some extent (Tiedemann, 2012). However, the prevailing approach in most related studies was to create custom parallel corpora (Kuandykova et al., 2014). This practice was motivated by the numerous issues in the existing data, including recurring repetitions, corrupted text segments, and obvious misalignment between the pairs (Myrzakhmetov and Makazhanov, 2016), which collectively contributed to a substantial reduction in the quality and quantity of the available data.

In Rakhimova and Zhumanov (2017), Kazakh–English (25,000 sentences) and Kazakh–Russian (10,000 sentences) parallel corpora were constructed utilising an open-source tool designed for the extraction of bitexts from multilingual websites. In a separate study by Zhumanov et al. (2017), the researchers collected an additional 73,031 Kazakh–English parallel sentences using the same tool. Importantly, the data collected in both studies were aligned automatically and are not open access.

In Makazhanov et al. (2017), over 890,000 parallel sentences in Russian and Kazakh were extracted from online news articles published on websites related to state institutions, national companies, and other quasi-governmental bodies. An SMT model trained on the parallel data yielded a BLEU score of 0.34. Interestingly, in Tukeyev et al. (2020), the authors achieved BLEU scores of 0.25 and 0.18 for the Kazakh→English and English→Kazakh language pairs, respectively, training an NMT model on a dataset acquired from the same aforementioned sources, although more than eight times smaller in size. In their later study (Tukeyev et al., 2019), a 439,176-sentence-long synthetic corpus using the complete set of Kazakh suffixes was constructed. An NMT model trained on the corpus produced BLEU scores in the range of 0.14 to 0.16 for the Kazakh↔Russian and Kazakh↔English language pairs.

In the study by Khairova et al. (2019), automated alignment was performed to create a Kazakh-Russian parallel corpus. This corpus comprised 3,000 texts that were extracted from four bilingual news websites in Kazakhstan, with a specific focus on criminal-related content. The researchers acknowledged the intricate syntactic structures inherent in both Kazakh and Russian, which posed significant challenges to the automatic alignment process. It was further observed that approximately 40% of the sentences required manual alignment due to these complexities.

The inclusion of the Kazakh↔English language pair as a translation task within the Fourth Conference on Machine Translation (WMT19) sparked several research efforts. Given the limited availability of parallel data for Kazakh–English, these initiatives leveraged the more abundant English–Russian and Kazakh–Russian sentence pairs, which numbered approximately 14 million and 5 million, respectively, using Russian as a pivot language (Casas et al., 2019; Littell et al., 2019; Sánchez-Cartagena et al., 2019). Additional attempts involved transfer learning utilising supplementary parallel data from the Turkish↔English language pair, as Turkish shares linguistic kinship with Kazakh (Toral et al., 2019; Briakou and Carpuat, 2019), albeit being a low-resource language itself. While Briakou and Carpuat (2019) obtained a BLEU score of 0.1 with just over 100 thousand Kazakh–English sentence pairs and another 200 thousand sentence pairs from Turkish–English data, Toral et al. (2019), using English–Russian, Kazakh–Russian, and English–Turkish data, achieved a BLEU of 0.24 for Kazakh→English and a chrF of 0.48 for English→Kazakh.

While recent research efforts in Kazakh↔English and Kazakh↔Russian MT have demonstrated noteworthy advancements, including the development of large-scale crawled parallel corpora (Rakhimova and Karibayeva, 2022; Zhumanov and Tukeyev, 2021), which are publicly accessible and capable of yielding commendable BLEU scores of up to 0.49 (Karyukin et al., 2023), as well as the construction of NMT post-editing models trained on such data (Rakhimova et al., 2021), the majority of textual resources continue to come from governmental sources. This preference is attributed to the perception of governmental

Domain	lines		tokens							
			EN		KK		RU		TR	
	#	%	#	%	#	%	#	%	#	%
Mass media	120,547	32.4	1,817,276	28.3	1,340,346	28.6	1,454,430	29.0	1,311,985	28.5
General	94,988	25.5	844,541	13.1	578,236	12.3	618,960	12.3	608,020	13.2
Legal documents	77,183	20.8	2,650,626	41.3	1,925,561	41.0	1,991,222	39.7	1,880,081	40.8
Education and science	46,252	12.4	522,830	8.1	392,348	8.4	444,786	8.9	376,484	8.2
Fiction	32,932	8.9	589,001	9.2	456,385	9.7	510,168	10.2	433,968	9.4
Total	**371,902**	**100**	**6,424,274**	**100**	**4,692,876**	**100**	**5,019,566**	**100**	**4,610,538**	**100**

Table 1: KazParC domain statistics

texts as subjected to moderation and therefore trustworthy (Karyukin et al., 2023). However, it should be noted that ensuring the quality and alignment of such texts still requires a significant amount of manual intervention (Zhumanov and Tukeyev, 2021). Excessive reliance on sources related to state bodies further harbours the potential to introduce bias into the corpus, thereby constraining the generalisability of models trained on such data. In light of these challenges, our study sought to create an extensive parallel corpus containing texts from diverse sources through the collaborative contributions of human translators, which would hopefully facilitate MT across Kazakh, English, Russian, and Turkish, as elaborated in subsequent sections.

3. Corpus Development

3.1. Data Sources

The data for our **Kaz**akh **Par**allel **C**orpus (hereafter KazParC) were sourced from a wide selection of textual materials, including proverbs and sayings, terminology glossaries, phrasebooks, literary works, periodicals, language learning resources, including the SCoRE corpus (Chujo et al., 2015), educational video subtitle collections, such as QED (Abdelali et al., 2014), news items, such as KazNERD (Yeshpanov et al., 2022) and WMT (Tiedemann, 2012), TED talks[1], governmental and regulatory legal documents from Kazakhstan[2], communications from the official website of the President of the Republic of Kazakhstan[3], United Nations publications[4], and image captions derived from sources, such as COCO (Lin et al., 2014). The data acquired from these sources were subsumed under five broad categories or domains—namely, Education and science, Fiction, General, Legal documents, and Mass media. Table 1 provides information about the number of lines and tokens collected per domain.

3.2. Data Collection

The process of data collection, which involved gathering text materials and their translation, was initiated in July 2021 and persisted until September 2023. Throughout this period, an average of 10 human translators were involved, which equates to 41,600 hours of human effort (26 months x 10 translators x 160 hours/month). The human translators not only engaged in the collection of readily translated publicly available data but also undertook the translation of texts that originally lacked translations in the languages under consideration.

The data collected were screened to remove any information that could potentially identify individuals, as well as to filter out instances of hate speech, discriminatory language, or violence. Subsequently, the data were segmented into sentences, each labelled with a domain identifier. A careful review for grammatical and spelling accuracy was conducted and duplicate sentences removed. Given the common practice of Kazakh-Russian code-switching in Kazakhstan (Pavlenko, 2008), sentences containing both Kazakh and Russian words underwent a modification process, wherein the Russian elements were translated into Kazakh for uniformity, taking care not to compromise the intended meaning of the sentences.

3.3. Data Pre-Processing

All the data collected were subjected to initial pre-processing, which involved segmenting the data into language pairs. Extraneous characters were systematically eliminated and homoglyphs effectively replaced. In addition, the characters responsible for line breaks (\n) and carriage returns (\r) were removed. The pre-processing further entailed the identification and elimination of duplicate entries, filtering out rows with identical text in both language columns. However, in order to enrich the diversity of the corpus and capture a wider range of synonyms for different words and expressions, lines with duplicate text in a single language column were judiciously retained.

In Table 2, we present statistics for language pairs within the corpus. The "# lines" column indicates the number of rows per language pair. In the '# sents", "# tokens", "# types" columns, we provide unique sentence, token, and type (i.e., unique token) counts for each language pair, respectively, with the upper numbers referring to the first language in the pair and the lower numbers to

[1]https://www.ted.com/
[2]https://adilet.zan.kz/
[3]https://www.akorda.kz/
[4]https://www.un.org/

Pair	# lines	# sents	# tokens	# types
KK↔EN	363,594	362,230 361,087	4,670,789 6,393,381	184,258 59,062
KK↔RU	363,482	362,230 362,748	4,670,593 4,996,031	184,258 183,204
KK↔TR	362,150	362,230 361,660	4,668,852 4,586,421	184,258 175,145
EN↔RU	363,456	361,087 362,748	6,392,301 4,994,310	59,062 183,204
EN↔TR	362,392	361,087 361,660	6,380,703 4,579,375	59,062 175,145
RU↔TR	363,324	362,748 361,660	4,999,850 4,591,847	183,204 175,145

Table 2: KazParC pairwise statistics

Pair	# lines	# sents	# tokens	# types
KK↔EN	1,787,050	1,782,192 1,781,019	26,630,960 35,291,705	685,135 300,556
KK↔RU	1,787,448	1,782,192 1,777,500	26,654,195 30,241,895	685,135 672,146
KK↔TR	1,791,425	1,782,192 1,782,257	26,726,439 27,865,860	685,135 656,294
EN↔RU	1,784,513	1,781,019 1,777,500	35,244,800 30,175,611	300,556 672,146
EN↔TR	1,788,564	1,781,019 1,782,257	35,344,188 27,806,708	300,556 656,294
RU↔TR	1,788,027	1,777,500 1,782,257	30,269,083 27,816,210	672,146 656,294

Table 3: SynC pairwise statistics

the second language. The token and type counts were obtained after processing the data with Moses tokeniser 1.2.1[5].

3.4. Data Splitting

We first created a test set. To this end, we conducted a random selection process, curating a set comprising 250 distinct and non-repetitive rows from each of the specified sources in Section 3.1. The remaining data were partitioned pairwise in adherence to an 80/20 ratio, preserving the distribution of domains within the training and validation sets (see Table 4).

3.5. Synthetic Corpus

To expand the scope of our parallel corpus and enhance its data diversity, as well as to investigate the performance characteristics of the developed NMT models when confronted with a combination of human-translated and machine-translated content, we conducted web crawling to acquire a total of 1,797,066 sentences from English-language websites. Subsequently, these sentences underwent an automated translation process into Kazakh, Russian, and Turkish languages utilising the Google Translate service. Within the context of our research, this collection of data will be referred to as "SynC" (**Syn**thetic **C**orpus). Table 3 presents statistics pertaining to the quantity of unique sentences, tokens, and types per each language pair. The synthetic corpus was further partitioned pairwise into training and validation sets at a ratio of 90/10 to facilitate model development and evaluation (see Table 5).

3.6. Corpus Structure

Both KazParC and SynC are openly accessible to the research community through our GitHub repository.[6] The corpora consist of multiple files categorised into two distinct groups based on their file prefixes: Files "01" through "19" bear the "kazparc" prefix, while Files "20" to "32" are denoted by the "sync" prefix.
File "01" contains the original, unprocessed text collected for the four languages considered within KazParC. Files "02" through "19" represent pre-processed texts divided into language pairs to serve as training data (Files "02" to "07"), validation data (Files "08" to "13"), and testing data (Files "14" to "19"). Language pairs are denoted within the filenames through the utilisation of two-letter language codes (e.g., kk_en).
SynC files are organised similarly. File "20" holds raw, unprocessed text data from the four languages. Files "21" to "32" contain pre-processed text split language pairwise for training (Files "21" to "26") and validation (Files "27" to "32") purposes.
In Files "01" and "20", each line comprises distinct components: a unique line identifier (`id`), texts in Kazakh (`kk`), English (`en`), Russian (`ru`), and Turkish (`tr`), along with accompanying domain information (`domain`). As for the remaining files, the data fields are `id`, `source_lang`, `target_lang`, `domain`, and the language `pair` (e.g., kk_en).

4. Experiment

4.1. Experimental Setup

The Transformer architecture (Vaswani et al., 2017) has proven highly effective in various NLP tasks, including MT, text generation, and text classification. In our study, we opted to employ Facebook's No Language Left Behind (NLLB) model (Team et al., 2022). The model supports MT for 202 languages, including Kazakh, English, Russian, and Turkish.
We first tested both the baseline[7] and distilled[8] versions of the model, obtained from the Hugging Face (Wolf

[5] https://pypi.org/project/mosestokenizer/
[6] https://github.com/IS2AI/KazParC
[7] https://huggingface.co/facebook/nllb-200-1.3B
[8] https://huggingface.co/facebook/nllb-200-distilled-1.3B

Pair	Train				Valid				Test			
	# lines	# sents	# tokens	# types	# lines	# sents	# tokens	# types	# lines	# sents	# tokens	# types
KK↔EN	290,877	286,958	3,693,263	164,766	72,719	72,426	920,482	83,057	4,750	4,750	57,044	17,475
		286,197	5,057,687	54,311		72,403	1,259,827	32,063		4,750	75,867	9,729
KK↔RU	290,785	286,943	3,689,799	164,995	72,697	72,413	923,750	82,958	4,750	4,750	57,044	17,475
		287,215	3,945,741	165,882		72,439	988,374	87,519		4,750	61,916	18,804
KK↔TR	289,720	286,694	3,691,751	164,961	72,430	72,211	920,057	82,698	4,750	4,750	57,044	17,475
		286,279	3,626,361	157,460		72,190	904,199	80,885		4,750	55,861	17,284
EN↔RU	290,764	286,185	5,058,530	54,322	72,692	72,377	1,257,904	32,208	4,750	4,750	75,867	9,729
		287,261	3,950,362	165,701		72,427	982,032	87,541		4,750	61,916	18,804
EN↔TR	289,913	285,967	5,048,274	54,224	72,479	72,220	1,256,562	32,269	4,750	4,750	75,867	9,729
		286,288	3,621,531	157,369		72,219	901,983	80,838		4,750	55,861	17,284
RU↔TR	290,899	287,241	3,947,809	165,482	72,725	72,455	990,125	87,831	4,750	4,750	61,916	18,804
		286,475	3,626,436	157,470		72,362	909,550	80,962		4,750	55,861	17,284

Table 4: KazParC training, validation, and test sets (by line, sentence, token, and type)

Pair	Train				Valid			
	# lines	# sents	# tokens	# types	# lines	# sents	# tokens	# types
KK↔EN	1,608,345	1,604,414	23,970,260	650,144	178,705	178,654	2,660,700	208,838
		1,603,426	31,767,617	286,372		178,639	3,524,088	105,517
KK↔RU	1,608,703	1,604,468	23,992,148	650,170	178,745	178,691	2,662,047	209,188
		1,600,643	27,221,583	642,604		178,642	3,020,312	235,642
KK↔TR	1,612,282	1,604,793	24,053,671	650,384	179,143	179,057	2,672,768	209,549
		1,604,822	25,078,688	626,724		179,057	2,787,172	221,773
EN↔RU	1,606,061	1,603,199	31,719,781	286,645	178,452	178,419	3,525,019	104,834
		1,600,372	27,158,101	642,686		178,379	3,017,510	235,069
EN↔TR	1,609,707	1,603,636	31,805,393	286,387	178,857	178,775	3,538,795	105,641
		1,604,545	25,022,782	626,740		178,796	2,783,926	221,372
RU↔TR	1,609,224	1,600,605	27,243,278	642,797	178,803	178,695	3,025,805	235,970
		1,604,521	25,035,274	626,587		178,750	2,780,936	221,792

Table 5: SynC: training and validation sets (by line, sentence, token, and type)

et al., 2020) repository, by fine-tuning them on KazParC. Upon comparison of the results, we observed that the distilled model consistently outperformed the baseline model, albeit by a slight margin of 0.01 BLEU. Therefore, in the subsequent experiments, we focused exclusively on fine-tuning the distilled model.

A total of four models, with each serving a specific purpose, were explored: (1) `base`, the off-the-shelf model, (2) `parc`, fine-tuned on KazParC data, (3) `sync`, fine-tuned on SynC data, and (4) `parsync`, fine-tuned on both KazParC and SynC data.

The `base` model was used as a reference point for evaluating the performance of the NLLB model. The `parc` model was fine-tuned exclusively on clean, manually translated data and was therefore considered suitable for tasks where accurate translation is important, especially in the domains covered by the training set.

The decision to test a model fine-tuned solely on synthetic data pursued the aim of discerning whether the performance of the model is more influenced by the quality or quantity of data within the parallel corpus. As a result, the `sync` model was expected to emphasise the viability of using synthetic data in scenarios where creating a human-translated parallel corpus is not feasible.

To assess the influence of data volume on translation quality, we explored the incorporation of synthetic data into our training set. This investigation aimed not only to evaluate its potential for enhancing translation quality but also to introduce distinctive lexemes absent in the original KazParC. Therefore, the `parsync` model was anticipated to leverage the synthetic and manual corpora and achieve a higher degree of universality and applicability to real-world problems.

The hyperparameters were tuned using the validation sets. Synthetic data were included in the validation sets only when the performance of the `sync` and `parsync` models was assessed. The best-performing models were evaluated on the test sets. Furthermore, we utilised Google Translate[9] and Yandex Translate[10] to translate the test sets, allowing us to make a comparative assessment between the results generated by our models and those produced by industry-leading machine translation services. In addition to the KazParC test set, we used the parallel FLoRes-200 (hereafter FLoRes) dataset (Team et al., 2022). This dataset was created to evaluate translation quality for 204 languages and contains texts from the Wikivoyage, Wikijunior, and Wikinews resources. FLoRes is divided into *dev* and *devtest* sets, but we combined them into one set. We also used the FLoRes test set to evaluate the quality for the language pairs German-French (two Latin-based higher-resourced Indo-European languages), German-Ukrainian (a higher-resourced language and a Cyrillic-based lower-resourced Indo-European language), and French-Uzbek (a higher-resourced language and a Latin-based low-resourced Turkic language) to see whether the translation quality changes for these control pairs after fine-tuning the model.

All the models were fine-tuned using eight GPUs on an NVIDIA DGX A100 machine. An initial learning rate of $2 \cdot 10^{-5}$ was set. The optimization algorithm chosen was AdaFactor. The training spanned across three epochs, with both the training and evaluation batch sizes set to 8.

4.2. Evaluation Metrics

In evaluating the MT models, we employed two widely recognised metrics: BLEU (Papineni et al., 2002) and chrF (Popović, 2015). While BLEU quantifies how closely the machine-produced translation matches human references, by calculating precision in n-grams (4 in our study), chrF evaluates translation quality by considering character n-grams instead of word-based approaches. This makes chrF particularly suitable for agglutinative languages, such as Kazakh and Turkish, which have rich and complex inflectional and derivational morphologies (Stanojević et al., 2015). chrF computes the harmonic mean of character-based precision and recall, providing a robust evaluation of translation performance. Both BLEU and chrF provide a score between 0 and 1, with higher scores indicating better translation quality.

4.3. Experiment Results

Model performance results are presented in Table 6. The table illustrates a notable disparity in bidirectional translation outcomes, particularly between higher-resourced Indo-European languages—English and Russian—and Turkic languages, Kazakh and Turkish. As can be seen from the table, it is apparent that BLEU scores exhibit a strong and positive correlation with chrF scores.

In the "→English" translation direction, Google consistently led on the FLoRes test set, achieving a minimum BLEU score of 0.35. However, on the KazParC test set, the leadership shifted to the `parc` model, which was exclusively trained on our parallel corpus. Notably, `parc` demonstrated an impressive BLEU score of up to 0.43 when translating RU→EN.

In the "→RU" translation, Google achieved the highest BLEU scores on both test sets. The only exception was observed in the EN→RU translation on the FLoRes test set, where Yandex outperformed Google by a margin of 0.01. Interestingly, when translating "→RU", the `parc` model generally exhibited lower performance compared to the `parsync` model, which was trained on a combination of our parallel corpus and synthetic data.

The same pattern was observed for the "→KK" and "→TR" translations. Google obtained the highest BLEU scores in both test sets. What is truly noteworthy is the clear underperformance of `parc` compared to `parsync` in these translation directions. This observation strongly supports the idea that model performance for lower-resourced (Turkic languages) can be substantially enhanced when synthetic data are employed alongside human-translated parallel data.

In the "EN→" translation direction, Google delivered superior translations across both test sets, with exceptions observed where Yandex briefly outperformed in the EN→RU language pair within the FLoRes dataset. It is worth noting that the `parsync` model consistently ranked among the top three performers on both test sets, attaining a commendable BLEU score of 0.20 in the EN→KK language pair within the FLoRes dataset, a result akin to that of Google.

Conversely, in the "KK→" translation direction, Google retained its translation accuracy predominance across both test sets, albeit with occasional instances where `parc` and `parsync` surpassed Google's performance. Notably, both `parc` and `parsync` consistently demonstrated the second-best performance, often matching or surpassing that of Yandex in this specific translation direction.

Within translation pairs involving Russian as the source language, out of the two models trained on our parallel corpus, `parsync` exhibited a consistent presence among the top three performers. Google, on the other hand, occasionally ceded its position to `parc` and Yandex in the RU→EN language pair.

For the "TR→" translation direction, `parsync` achieved noteworthy success, securing a leading BLEU score of 0.38 on the KazParC test set for TR→EN and a commanding BLEU score of 0.13 in the TR→KK language pair on the FLoRes test set, with Google being the frontrunner.

After thoroughly assessing the qualitative and quantitative results, we determined that the `parsync` model, fine-tuned on a combination of the KazParC corpus and synthetic data, displayed the highest results among the three developed models. In the upcoming

[9] https://translate.google.com/

[10] https://translate.yandex.com/

<table>
<thead>
<tr><th rowspan="2">Pair</th><th colspan="6">FLoRes Test Set</th><th colspan="6">KazParC Test Set</th></tr>
<tr><th><code>base</code></th><th><code>parc</code></th><th><code>sync</code></th><th><code>parsync</code></th><th>Yandex</th><th>Google</th><th><code>base</code></th><th><code>parc</code></th><th><code>sync</code></th><th><code>parsync</code></th><th>Yandex</th><th>Google</th></tr>
</thead>
<tbody>
<tr><td>EN→KK</td><td>0.11|0.49</td><td>0.14|0.56</td><td>0.20|0.60</td><td>0.20|0.60</td><td>0.18|0.58</td><td>0.20|0.60</td><td>0.12|0.51</td><td>0.18|0.58</td><td>0.18|0.58</td><td>0.21|0.60</td><td>0.18|0.58</td><td>0.30|0.65</td></tr>
<tr><td>EN→RU</td><td>0.25|0.56</td><td>0.26|0.58</td><td>0.28|0.60</td><td>0.28|0.60</td><td>0.32|0.63</td><td>0.31|0.62</td><td>0.31|0.64</td><td>0.38|0.68</td><td>0.35|0.66</td><td>0.38|0.68</td><td>0.39|0.70</td><td>0.41|0.71</td></tr>
<tr><td>EN→TR</td><td>0.19|0.58</td><td>0.22|0.61</td><td>0.27|0.65</td><td>0.27|0.65</td><td>0.29|0.66</td><td>0.30|0.66</td><td>0.19|0.59</td><td>0.22|0.62</td><td>0.25|0.63</td><td>0.25|0.64</td><td>0.27|0.64</td><td>0.34|0.68</td></tr>
<tr><td>KK→EN</td><td>0.28|0.59</td><td>0.32|0.62</td><td>0.31|0.62</td><td>0.32|0.63</td><td>0.30|0.62</td><td>0.36|0.65</td><td>0.24|0.55</td><td>0.33|0.62</td><td>0.24|0.57</td><td>0.32|0.62</td><td>0.28|0.60</td><td>0.31|0.62</td></tr>
<tr><td>KK→RU</td><td>0.15|0.49</td><td>0.17|0.51</td><td>0.18|0.52</td><td>0.18|0.52</td><td>0.18|0.52</td><td>0.20|0.53</td><td>0.22|0.56</td><td>0.29|0.63</td><td>0.24|0.59</td><td>0.29|0.63</td><td>0.29|0.63</td><td>0.29|0.61</td></tr>
<tr><td>KK→TR</td><td>0.09|0.48</td><td>0.13|0.52</td><td>0.14|0.54</td><td>0.14|0.54</td><td>0.12|0.52</td><td>0.17|0.56</td><td>0.10|0.47</td><td>0.15|0.54</td><td>0.14|0.52</td><td>0.16|0.55</td><td>0.13|0.52</td><td>0.23|0.59</td></tr>
<tr><td>RU→EN</td><td>0.31|0.62</td><td>0.32|0.63</td><td>0.32|0.63</td><td>0.32|0.63</td><td>0.33|0.64</td><td>0.35|0.65</td><td>0.34|0.63</td><td>0.43|0.71</td><td>0.34|0.65</td><td>0.42|0.70</td><td>0.43|0.71</td><td>0.42|0.71</td></tr>
<tr><td>RU→KK</td><td>0.08|0.49</td><td>0.10|0.52</td><td>0.13|0.53</td><td>0.13|0.54</td><td>0.12|0.54</td><td>0.13|0.54</td><td>0.15|0.55</td><td>0.21|0.61</td><td>0.18|0.58</td><td>0.22|0.62</td><td>0.23|0.62</td><td>0.24|0.62</td></tr>
<tr><td>RU→TR</td><td>0.10|0.49</td><td>0.12|0.52</td><td>0.14|0.54</td><td>0.14|0.54</td><td>0.13|0.54</td><td>0.17|0.56</td><td>0.11|0.49</td><td>0.16|0.56</td><td>0.16|0.55</td><td>0.18|0.57</td><td>0.16|0.55</td><td>0.22|0.60</td></tr>
<tr><td>TR→EN</td><td>0.34|0.64</td><td>0.35|0.65</td><td>0.36|0.66</td><td>0.36|0.66</td><td>0.38|0.67</td><td>0.39|0.67</td><td>0.31|0.61</td><td>0.38|0.67</td><td>0.32|0.63</td><td>0.38|0.66</td><td>0.36|0.66</td><td>0.37|0.66</td></tr>
<tr><td>TR→KK</td><td>0.07|0.45</td><td>0.10|0.51</td><td>0.13|0.54</td><td>0.13|0.54</td><td>0.12|0.53</td><td>0.13|0.54</td><td>0.08|0.46</td><td>0.14|0.53</td><td>0.14|0.52</td><td>0.16|0.55</td><td>0.14|0.53</td><td>0.19|0.57</td></tr>
<tr><td>TR→RU</td><td>0.15|0.48</td><td>0.17|0.51</td><td>0.18|0.52</td><td>0.19|0.53</td><td>0.20|0.54</td><td>0.21|0.54</td><td>0.17|0.50</td><td>0.23|0.56</td><td>0.20|0.54</td><td>0.24|0.57</td><td>0.23|0.57</td><td>0.26|0.58</td></tr>
<tr><td>Average</td><td>0.18|0.53</td><td>0.20|0.56</td><td>0.22|0.58</td><td>0.22|0.58</td><td>0.23|0.58</td><td>0.25|0.59</td><td>0.20|0.55</td><td>0.27|0.61</td><td>0.23|0.59</td><td>0.27|0.62</td><td>0.26|0.61</td><td>0.30|0.63</td></tr>
</tbody>
</table>

Table 6: BLEU|chrF scores for models on the FLoRes and KazParC test sets

Discussion section, we will simply refer to this model as “Tilmash” [tɪlˈmɑʃ], a Kazakh term denoting “interpreter”, “translator”.

It is worth noting that when comparing the translation results between `base` and Tilmash on the control language pairs, the latter displayed less favourable results, hinting at a possible decline in translation quality after fine-tuning (see Table 7).

Pair	**BLEU**		**chrF**	
	`base`	Tilmash	`base`	Tilmash
DE→FR	0.33	0.28	0.61	0.58
FR→DE	0.22	0.19	0.55	0.53
DE→UK	0.15	0.04	0.49	0.36
UK→DE	0.19	0.16	0.53	0.50
FR→UZ	0.06	0.02	0.48	0.31
UZ→FR	0.25	0.22	0.56	0.53

Table 7: Results of the `base` and Tilmash models on the control language pairs on the FLoRes test set

The lower BLEU scores for Kazakh and Turkish translations can be attributed to the agglutinative nature of these languages. In agglutinative languages, words are formed by stringing together different morphemes, leading to longer and more complex words. This linguistic characteristic poses a challenge for translation models, as they may have difficulty capturing the complicated morphological structures, resulting in a statistically lower BLEU score.

However, we observed that the chrF score remains relatively stable across language pairs. This suggests that the overall translation quality, measured by chrF, is consistent across all language pairs. The chrF metric considers n-grams at the character level and provides a more robust evaluation that is less sensitive to the structural differences between languages.

We hypothesise that the differences in translation quality between language pairs may be influenced by the resourcefulness of the languages and the training data available for the baseline NLLB model. Languages with richer linguistic resources and diverse training data may demonstrate better translation results.

5. Discussion

The comparison of the results of Tilmash with those of Yandex and Google on the FLoRes and KazParC test sets reveals that the performance of our model is on par with that of the industry giants. It is particularly pleasing to note that Tilmash yields consistent results on the diverse FLoRes test set, spanning a wide range of topics, from rare diseases to long-extinct dinosaurs, which may not be present in KazParC. This further reinforces the versatility of our model in effectively translating texts across various domains. That said, Tilmash appears to struggle with translating figurative expressions, such as proverbs and idioms, where conveying both literal accuracy and the rich cultural, historical, and emotional connotations they hold can be a challenging balance to maintain.

While it is true that the results of Tilmash are not significantly higher than those of `parc`, which was exclusively trained on our parallel corpus and, in some cases, even lower (see, for instance, “→EN”), we must acknowledge that the inclusion of synthetic data in the training set has had a positive impact on the performance of Tilmash, as evident from its strong performance on the FLoRes test set—a feat that the `parc` model cannot claim. The substantial increase in the number of word types, and, consequently, the diversity of vocabulary, introduced by the synthetic data not only appears to enhance translation performance but also suggests the potential of utilising synthetic data in conjunction with much smaller amounts of human-translated parallel data to achieve improved results. However, it is important to remain mindful of the inherent translation inaccuracies and incorrect syntactic structures that can result from MT of large, web-crawled, and uncurated data. For example, Tilmash occasionally stumbles over second-person singular pronouns in Kazakh (*сіз*, *сен*), Russian (*вы*, *ты*),

Pair	Type	Text	BLEU	chrF
KK→EN	source	*Ыстық және желді.* *Ystyq jane jeldi.*		
	reference	*It is hot and windy.*	1.00	1.00
	Tilmash	*It's hot and windy.*	0.55	0.81
	Yandex	*Hot and windy.*	0.00	0.66
	Google	*Hot and windy.*	0.00	0.66
KK→EN	source	*1 қыркүйекте бесінші ана өлімі тіркелді.* *1 qyrkuiekte besinshi ana olimi tirkeldi.*		
	reference	*On September 1, the fifth maternal death was registered.*	1.00	1.00
	Tilmash	*A fifth maternal death was recorded on 1 September.*	0.27	0.63
	Yandex	*On September 1, the fifth maternal death was registered.*	1.00	1.00
	Google	*On September 1, the fifth maternal death was recorded.*	0.81	0.86

Table 8: A selection of translation outputs from Tilmash, Yandex, and Google

and Turkish (*siz*, *sen*) when translating the English "you". This can lead to instances where Tilmash produces informal (*сен*, *ты*, *sen*) pronouns instead of the expected polite (*сіз*, *вы*, *siz*) forms. We attribute this issue to the use of the synthetic corpus, as `parc`, trained solely on KazParC, accurately handles these pronouns.
A thorough examination of the performance of Tilmash, Yandex, and Google across the domains within the KazParC test set reveals the remarkable superiority of Tilmash in legal documents and texts pertaining to the general domain.[11] This notable performance is observed in nine translation directions, as indicated by either BLEU or chrF scores, which we attribute to the extensive presence of well-translated legal documents and everyday social expressions within the parallel corpus (see Table 1). The somewhat lower, yet still comparable, results observed in the mass media domain, despite the majority of texts in KazParC originating from this domain, can be attributed to several factors. It is challenging to rival Google and Yandex in this domain, as their models are likely to have been extensively trained on news articles. Additionally, the presence of numerous proper nouns (e.g., names of individuals, organisations, locations, and more) and abbreviations within news content can pose challenges for MT models in ensuring accurate handling.
Table 8 provides some examples of KK→EN translation. We can see that in the first example Tilmash demonstrated a distinct approach compared to Yandex and Google, which simply translated the adjectives into English. Not only was Tilmash able to correctly detect that the source sentence was an impersonal construction, but it also produced "it", which effectively functions as a placeholder for the weather condition. While the BLEU and chrF scores are not perfect, it is worth emphasising that the difference between the reference sentence and the Tilmash-generated sentence solely lies in the use of the contraction "it's", with both sentences conveying the same information and maintaining identical grammatical structures.
In the second example, we observe that the sentence generated by Tilmash, as well as the reference sentence and those produced by Yandex and Google, convey similar meanings but exhibit differences in sentence structure, word choice influenced by regional date conventions ("September 1" vs. "1 September") and formality ("registered" vs. "recorded"), and the use of articles ("the" vs. "a"). While, in many contexts, these variations in dates and verbs can be used interchangeably, the choice of articles depends on contextual information. Specifically, it hinges on whether one is referring to one of multiple maternal deaths or a specific, previously mentioned, or contextually precise fifth maternal death. Without context, Tilmash may face challenges in determining the appropriate article to use while maintaining proper grammar. Nevertheless, we believe that such cases can be effectively addressed by a human translator during the post-editing phase, if necessary.

6. Conclusion

We have introduced KazParC, a parallel corpus developed for MT of Kazakh, English, Russian and Turkish. It is the first and largest publicly available corpus of its kind and includes 371,902 parallel sentences from different domains created with the help of human translators. In addition, our research has led to the development of the Tilmash NMT model, which has demonstrated remarkable performance, often matching or surpassing Yandex Translate and Google Translate, as evidenced by standard evaluation metrics such as BLEU and chrF. Both KazParC and Tilmash are available for download under the Creative Commons Attribution 4.0 International Licence (CC BY 4.0) from our GitHub repository.[6]
In the future, we are committed to expanding KazParC to cover a wider range of domains and lexica, including figurative expressions, with the aim of improving translation quality. We also plan to conduct further experiments with the NLLB model to preserve the

[11] Due to space constraints, we have published the detailed tables of results per domain on our GitHub page.

original translation quality in non-target language pairs. In addition, we will continue to explore different pre-trained models and training parameters to refine our models.

7. Acknowledgements

We extend our sincere gratitude to Aigerim Baidauletova, Ainagul Akmuldina, Assel Mukhanova, Aizhan Seipanova, Askhat Kenzhegulov, Elmira Nikiforova, Gaukhar Rayanova, Gulim Kabidolda, Gulzhanat Abduldinova, Indira Yerkimbekova, Moldir Orazalinova, Saltanat Kemaliyeva, and Venera Spanbayeva for their invaluable assistance with translation throughout the entirety of the study.

8. Bibliographical References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Zhenisbek Assylbekov, Bagdat Myrzakhmetov, and Aibek Makazhanov. 2016. Experiments with Russian to Kazakh sentence alignment. The 4-th International Conference on Computer Processing of Turkic Languages.

Zhenisbek Assylbekov and Assulan Nurkas. 2014. Initial Explorations in Kazakh to English Statistical Machine Translation. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa*, pages 12–16.

Meritxell Fernández Barrera, Vladimir Popescu, Antonio Toral, Federico Gaspari, and Khalid Choukri. 2016. Enhancing cross-border EU E-commerce through machine translation: needed language resources, challenges and opportunities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4550–4556.

Sevilay Bayatli, Sefer Kurnaz, Ilnar Salimzianov, Jonathan North Washington, and Francis M. Tyers. 2018. Rule-based machine translation from Kazakh to Turkish. In *European Association for Machine Translation Conferences/Workshops*.

Eldar Bekbulatov and Amandyk Kartbayev. 2014. A Study of Certain Morphological Structures of Kazakh and Their Impact on the Machine Translation Quality. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Eleftheria Briakou and Marine Carpuat. 2019. The University of Maryland's Kazakh-English Neural Machine Translation System at WMT19. In *Conference on Machine Translation*.

George L Campbell and Gareth King. 2020. *Compendium of the World's Languages*. Routledge.

Noe Casas, José A. R. Fonollosa, Carlos Escolano, Christine Basta, and Marta R. Costa-jussà. 2019. The TALP-UPC Machine Translation Systems for WMT19 News Translation Task: Pivoting Techniques for Low Resource MT. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 155–162, Florence, Italy. Association for Computational Linguistics.

Kiyomi Chujo, Kathryn Oghigian, and Shiro Akasegawa. 2015. A corpus and grammatical browsing system for remedial EFL learners. *Multiple affordances of language corpora for data-driven learning*, pages 109–128.

Olivia Craciunescu, Constanza Gerding-Salas, and Susan Stringer-O'Keeffe. 2004. Machine Translation and Computer-Assisted Translation: A New Way of Translating? *Machine Translation and Computer-Assisted Translation*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

W. John Hutchins. 1995. Machine Translation: A Brief History. In E.F.K. KOERNER and R.E. ASHER, editors, *Concise History of the Language Sciences*, pages 431–445. Pergamon, Amsterdam.

Johanson, Lars and Csató, Éva Á. 2021. *The Turkic languages (2nd ed.)*. Routledge.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Vladislav Karyukin, Diana Rakhimova, Aidana Karibayeva, Aliya Turganbayeva, and Asem Turarbek. 2023. The neural machine translation models for the low-resource Kazakh–English language pair. *PeerJ Computer Science*, 9:e1224.

Gulshat Kessikbayeva and Ilyas Cicekli. 2021. Impact of Statistical Language Model on Example Based Machine Translation System between Kazakh and Turkish Languages. In *Proceedings of the 4th*

International Conference on Natural Language Processing and Information Retrieval, NLPIR '20, page 112–118, New York, NY, USA. Association for Computing Machinery.

NF Khairova, Anastasiia Kolesnyk, Orken Mamyrbayev, and Kuralay Mukhsina. 2019. *The aligned Kazakh-Russian parallel corpus focused on the criminal theme*. Ph.D. thesis.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Ayana Kuandykova, Amandyk Kartbayev, and Tannur Kaldybekov. 2014. English-Kazakh Parallel Corpus for Statistical Machine Translation. *International Journal on Natural Language Computing (IJNLC)*, 65.

Sangmin-Michelle Lee. 2020. The impact of using machine translation on EFL students' writing. *Computer assisted language learning*, 33(3):157–175.

David Lewis, Alexander O'Connor, Andrzej Zydron, Gerd Sjögren, and Rahzeb Choudhury. 2012. On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market. In *LREC*, pages 1403–1409.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Patrick Littell, Chi-kiu Lo, Samuel Larkin, and Darlene Stewart. 2019. Multi-Source Transformer for Kazakh-Russian-English Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 267–274, Florence, Italy. Association for Computational Linguistics.

Aibek Makazhanov, Bagdat Myrzakhmetov, and Zhanibek Kozhirbayev. 2017. On Various Approaches to Machine Translation from Russian to Kazakh.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. "KazakhTTS2: Extending the open-source Kazakh TTS corpus with more data, speakers, and topics". In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5404–5411, Marseille, France. European Language Resources Association.

Bagdat Myrzakhmetov and Aibek Makazhanov. 2016. Initial Experiments on Russian to Kazakh SMT.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.

Aneta Pavlenko. 2008. Russian in post-Soviet countries. *Russian Linguistics*, 32(1):59–80.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus.

Diana Rakhimova and Aidana Karibayeva. 2022. Aligning and Extending Technologies of Parallel Corpora for the Kazakh Language. *Eastern-European Journal of Enterprise Technologies*, 118(2).

Diana Rakhimova, Kamila Sagat, Kamila Zhakypbaeva, and Aliya Zhunussova. 2021. Development and Study of a Post-editing Model for Russian-Kazakh and English-Kazakh Translation Based on Machine Learning. In *Advances in Computational Collective Intelligence*, pages 525–534, Cham. Springer International Publishing.

Diana Rakhimova and Zhandos Zhumanov. 2017. *Complex Technology of Machine Translation Resources Extension for the Kazakh Language*, pages 297–307. Springer International Publishing, Cham.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2019. The Universitat d'alacant submissions to the English-to-Kazakh news translation task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 356–363, Florence, Italy. Association for Computational Linguistics.

Rita Sanders. 2016. *Staying at Home: Identities, Memories and Social Networks of Kazakhstani Germans*, volume 13. Berghahn Books.

Asem Shormakova and Aida Sundetova. 2013. Machine translation of different systemic languages using a Apertium platform (with an example of English and Kazakh languages). In *2013 International Conference on Computer Applications Technology (ICCAT)*, pages 1–4.

Felix Stahlberg. 2020. Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research*, 69:343–418.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Aida Sundetova, Aidana Karibayeva, and Ualsher Tukeyev. 2014. Structural transfer rules for Kazakh-to-English machine translation in the free/open-source platform Apertium. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 7(2):48–53.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Lukas Edman, Galiya Yeshmagambetova, and Jennifer Spenader. 2019. Neural machine translation for English–Kazakh with morphological segmentation and synthetic data. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 386–392, Florence, Italy. Association for Computational Linguistics.

Ualsher Tukeyev, Aidana Karibayeva, and Balzhan Abduali. 2019. Neural machine translation system for the Kazakh language based on synthetic corpora. *MATEC Web Conf.*, 252:03006.

Ualsher Tukeyev, Aidana Karibayeva, and Zhandos Zhumanov. 2020. Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 7(1):1856500.

Francis M Tyers and Jonathan Washington. 2015. Towards a Free/Open-source Universaldependency Treebank for Kazakh. In *Proceedings of the International Conference "Turkic Languages Processing"TurkLang-2015*, pages 276–289.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Eva Wiesmann. 2019. Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German. *Comparative Legilinguistics*, 37(1):117–153.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Zhandos Zhumanov, Aigerim Madiyeva, and Diana Rakhimova. 2017. New Kazakh Parallel Text Corpora with On-line Access. In *Computational Collective Intelligence*, pages 501–508, Cham. Springer International Publishing.

Zhandos Zhumanov and Ualsher Tukeyev. 2021. Integrated Technology for Creating Quality Parallel Corpora. In *Advances in Computational Collective Intelligence*, pages 511–524, Cham. Springer International Publishing.