

# Low-resource Machine Translation for Code-switched Kazakh-Russian Language Pair

Maxim Borisov<sup>1</sup>, Zhanibek Kozhirbayev<sup>2</sup>, Valentin Malykh<sup>3,4,5</sup>

<sup>1</sup>ITMO University, Saint-Petersburg, Russia;

<sup>2</sup>Nazarbayev University, Astana, Kazakhstan;

<sup>3</sup>MTS AI, Moscow, Russia;

<sup>4</sup>ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia;

<sup>5</sup>IITU University, Almaty, Kazakhstan

## Abstract

Machine translation for low-resource language pairs is a challenging task. This task could become extremely difficult once a speaker uses code switching. We present the first code-switching Kazakh-Russian parallel corpus. Additionally, we propose a method to build a machine translation model for code-switched Kazakh-Russian language pair with no labeled data. Our method is basing on generation of synthetic data. This method results in a model beating an existing commercial system by human evaluation.

## 1 Introduction

Code-switching presents a significant challenge in Natural Language Processing due to its unpredictability, variability, and the lack of available corpora, especially for low-resource languages. There were no publicly available code-switched Kazakh-Russian parallel dataset, thus we present one in this work. The sample from the dataset is presented in Tab. 1. This dataset contains only 618 parallel sentences, so it can be used only for evaluation and not for training. We propose a method for training a machine translation model for code-switching task. In our method we use several publicly available Kazakh-Russian datasets, but since these datasets do not address code-switching problem, we generate additional training data by translating relevant monolingual corpus and show the effectiveness of this approach. We augment the data to address challenge of code-switching. To do so we developed a novel text transformation method based on SimAlign (Sabet et al., 2020). We train several machine translation models on the augmented dataset resulting in 3.09 Likert score for the best baseline model, while Yandex commercial model shows 2.80 Likert score. These experimental

results suggest that our method is able to improve the performance of machine translation systems on real code-switching data and jump start for those language pairs that do not have collected code-switched data.

The following paper is structured as follows: section 2 describes the work on code-switching done for other language pairs alongside with studies devoted to Russian-Kazakh language pair; section 3 presents the description of the existing public datasets for the mentioned language pair and the description of a newly introduced dataset with code-switching phenomenon captured; section 4 contains the details regarding our proposed augmentation method; section 5 describes the baselines, their training process, and the achieved results, while section 9 concludes the paper.

The contribution of this work is three-fold: (i) we present the first Kazakh-Russian code-switching dataset;<sup>1</sup> (ii) we present an evaluation of the existing models on this dataset; (iii) we propose a novel data augmentation for not code-switched datasets, which allowed us to fine-tune the existing open models achieving almost on par performance with an available commercial system.

## 2 Related Work

Recent progress in NLP has spurred the development of technologies capable of handling code-switched data. Despite the initiation of Code-Switching research several years ago, progress within the research community has been sluggish. The primary challenge to address this issue arises from the insufficient availability of data (Winata et al., 2023). A limited number of languages, such as Spanish-English

<sup>1</sup>KRCS dataset could be accessed here: <https://github.com/madrugado/KRCS>.

Original	қазақстанский гендерлік теңдік саласындағы 12 халықаралық құжаттарды бекітті .	фискал көзқарастан гөрі либералдандыру жақсы
Corrected	Қазақстан гендерлік теңдік саласындағы 12 халықаралық құжаттарды бекітті .	Фискалдық көзқарастан гөрі либералдандыру жақсы
Russian	Казахстаном ратифицировано 12 международных документов в сфере гендерного равенства .	Лучше либерализация , чем фискальный подход
Augmented	Қазақстан гендерного теңдік саласындағы 12 халықаралық құжаттарды бекітті .	Фискалдық подход гөрі либералдандыру

Table 1: Sample sentence triplet from KRCS dataset accompanied with cs-5 augmentation of a Russian one.

(Weller et al., 2022; Xu and Yvon, 2021), Hindi-English (Appicharla et al., 2021; Jadhav et al., 2022), or Chinese-English (Li et al., 2012), dominate research and resources in code switching. Nevertheless numerous countries and cultures that extensively use code switching remain underrepresented in NLP research.

A common feature of natural interactions among bilingual speakers is the spontaneous and continuous switching between the Kazakh and Russian languages. It is worth noting that the field still faces challenges, particularly due to the scarcity of code-switched data and the colloquial characteristic of code-switching. To our knowledge, only a few research papers have been published on this matter. In the context of Kazakh-Russian code-switching, a study by Ubskii et al. (2020) attempted to determine the benefit of bilingual training on matrix language (Kazakh) and embedded language (Russian) monolingual data (Myers-Scotton, 1997), as opposed to training on code-switched data only. The study made use of two datasets: Kazakh speech with code-switching and Russian speech with no code-switching. The main objective of the experiments was to compare the performance of a model trained on code-switched speech with that of a model trained on full utterances in both languages. Experimental results suggested that bilingual training improves the model’s performance on matrix words, and greatly improves its performance on embedded words. Another study by Zharkynbekova and Chernyavskaya (2022) discussed the ethnic bilingual practice in Kazakhstan. The focus was on code-switching or, in other term, code-mixing in the Kazakh-Russian and Russian-Kazakh bilingualism. The bi- and mul-

lingualism is characteristic for Kazakhstan and is caused by multi-ethnicity of the republic. The study analyzed 300 contexts that show the Kazakh-Russian code-mixing in everyday and internet communication, and in modern Kazakh films reflecting the typical code-mixing practice.

### 3 Datasets

**Training Datasets** consist of a dataset collected by Nazarbayev University and described in (Kozhimbayev and Islamgozhayev, 2023), we refer to this dataset as NU below; a dataset collected by Al Farabi University and described in (Balzhan et al., 2015) (KazNU); translated domain adaptation dataset, which is based on Russian tweet corpus described in (Рубцова, 2012) (RTC). We provide more details on domain adaptation in section 8. These three datasets are the main sources of training data, in addition we use several smaller datasets. To acquire these datasets we used MTData tool described in (Gowda et al., 2021). We combine all the datasets in a single one and apply deduplication. We call this dataset “all data” below. We provide the statistics for all the training datasets in Appendix B.

**Evaluation Dataset** We use Kazakh-Russian Code-Switching dataset (KRCS) as our evaluation dataset. The KRCS dataset consists of 618 colloquial Kazakh sentences from social media which include some Russian phrases with corresponding ground truth translations to grammatically correct Kazakh and Russian labeled by annotators. We had two annotators, both of them were natively bilingual in Kazakh and Russian, both of them

are working in academia. The annotation were done as part of their academic duties.

Number of sentences	618
# in an original Kazakh sentence	11.95
Russian # in an original sentence	2.77
# in a corrected Kazakh sentence	12.27
# in a Russian sentence	13.64

Table 2: KRCS dataset statistics. # stands for average number of tokens.

The descriptive statistics of the collected corpus is provided in Tab. 2. In Tab. 1 we provide a sample from KRCS dataset.

## 4 Dataset Augmentation

**Code-Switching Emulation Method** In the previous section we described the training datasets, nevertheless we need to state clearly that that datasets are not consider code-switching phenomenon and thus cannot be used effectively in our setup. Therefore we decided to make code-switching data artificially, using specific techniques for data augmentation.

First, we prepare the data. For it we follow the M2M100 recipe provided in [fairseq repository](#) which is an official implementation of (Ott et al., 2019). Namely, we filter out sentences with more than 50% of punctuation, remove the duplicates, and discard sentences with more than 50% of symbols that are not common for a given language.

Next, we take Kazakh processed sentences and augment them. We chose cs-5 method for augmentation: Replace a Kazakh word with a Russian word aligned using SimAlign (Sabet et al., 2020). Preliminary, we tried several augmentation techniques, their description and evaluation can be found in section 7.

For cs-5 Minimal Aligned Units (MAU) are extracted following an approach described in (Xu and Yvon, 2021): the small bilingual phrase pairs  $(a, b)$  extracted from symmetrical alignment such that for every word in  $a$  there exists a link to word in  $b$  and vice versa.

Next, we replace 15% of tokens/MAUs in the Kazakh sentence at random<sup>2</sup>. Sentences with length of less than 7 tokens have one replacement following (Anwar, 2023). We provide a

<sup>2</sup>The exact percentage is inspired by Masked Language Modeling approach firstly introduced in (Devlin et al., 2018)

sample of augmented sentence in Tab. 1. We also provide additional linguistic analysis and justification for each method in Appendix D.

## 5 Evaluation

**Baselines** There are several baselines which are used in our experiments. We use **identity** baseline, which simply copying its input to the output. This baseline is obviously not trained.

There are two trained from scratch baselines, namely, the first one is **transformer-600**, which is described below. The architecture of the model follows NLLB one, specifically the 600M parameters variant. The details of implementation can be found in Appendix A.

The second trained from scratch baseline is a reproduced approach from (Kozhirbayev and Islamgozhayev, 2023). We call this baseline **transformer-NU**.

The next three baselines are using pre-trained machine translation models and fine-tune them on our training data. These baselines are **mBART**, a model family described in (Liu et al., 2020), we use specifically mbart-large-50-many-to-many-mmt variant; **M2M100**, a model family described in (Fan et al., 2020), specifically facebook/m2m100\_1.2B; and **NLLB-600**, a model family described in (Costa-jussà et al., 2022), specifically facebook/nllb-200-distilled-600M.

The last fine-tuned baseline is **NLLB-3.3B** from the same model family as the previous one, but it is facebook/nllb-200-3.3B variant. We do not fully fine-tune this model, instead we use PiSSA (Meng et al., 2024), a PEFT approach.

**Metrics** In our work we are using three standard metrics: BLEU score (Papineni et al., 2002), which is basically a token accuracy; ChrF++ score (Popović, 2017), which is character level F-score; and COMET score (Rei et al., 2020), which is a Transformer-based model trained to compare translations. For the last metric we use specifically Unbabel/wmt22-cometkiwi-da model, described in (Rei et al., 2022).

## 6 Results

For this evaluation we use all the baselines with cs-5 augmentation, since it is the best in our setup as it shown in previous section. In

Model	w/o training	trained
identity	7.55 / 25.10 / <b>0.56</b>	N/A
transformer-NU	7.87 / 31.99 / 0.50	11.31 / 35.35 / 0.53
transformer-600	N/A	12.49 / 36.44 / 0.54
mBART	4.62 / 17.83 / <b>0.56</b>	12.08 / 34.31 / 0.53
M2M100	5.37 / 21.59 / 0.42	12.50 / 36.44 / 0.53
NLLB-600	12.26 / 36.67 / 0.53	12.95 / 36.44 / 0.54
NLLB-3.3B	<b>15.23 / 39.68 / 0.56</b>	<b>16.48 / 42.27 / 0.56</b>
Commercial APIs		
Yandex MT <sup>2</sup>	22.24 / 47.13 / 0.67	N/A
Google MT	24.14 / 47.84 / 0.64	N/A

Table 3: The comparison of baseline models in BLEU / ChrF++ / COMET on KRCS dataset.

addition, we provide results for two commercial machine translation systems, namely Yandex MT and Google MT. The results are provided in Tab. 3. As one can see, the best results are achieved by NLLB-3.3B model. This is not surprising, once it is the biggest model in comparison. What is interesting in this setup is that our approach allows to achieve good results with all the trained models, and the best trained model once achieved a score close to Yandex MT system<sup>3</sup>. Another point worth mentioning that COMET scores are close for identity baseline, mBART model, and NLLB-3.3B model.

**Human Evaluation** We have done human evaluation for our best model (chosen by BLEU score) and two commercial APIs. We asked our assessors to use Likert scale and averaged their scores for 100 random sentences from KRCS. The results are provided in Tab. 4. As can be seen, the results are a bit unexpected. Despite the automatic metrics scoring the Yandex MT system higher than NLLB-3.3B model, human evaluation showed the opposite. Also, it is worth noting that even the best commercial system is pretty far from ground truth translation in this domain.

We also evaluated the naturalness of augmentation in Kazakh. We chose 100 random sentences with cs-5 augmentation and asked our assessors again to use Likert scale. The achieved result is 2.62, which could be consid-

<sup>3</sup>We provide current scores for Yandex MT system at 15th of June. When the work has been started the score for Yandex MT was **16.72** BLEU.

	Mean	Std.
Ground Truth	4.75	0.68
NLLB-3.3B	3.09	1.13
Yandex MT	2.80	1.17
Google MT	3.49	1.14

Table 4: The human evaluation results.

ered acceptable.

## 7 Augmentation Study

We experiment with 5 augmentation types, namely: **cs-1**: Replace a Kazakh word with a Russian one in normal form; **cs-2**: Replace a Kazakh word with a Russian one’s stem with Kazakh ending, extracted from a Kazakh word by excluding stem from it; **cs-3**: Replace a Kazakh word with a Russian one in random form; **cs-4**: Replace a Kazakh word with a Russian word aligned using fastalign (Dyer et al., 2013); **cs-5**: Replace a Kazakh word with a Russian word aligned using SimAlign (Sabet et al., 2020).

For cs-1, cs-2, and cs-3 we employ a publicly available Kazakh-Russian dictionary from work (Rakhimova, 2020). For cs-4 Minimal Aligned Units are extracted as for cs-5. For all augmentation methods, the replacement is done as for cs-5. We provide samples for all the augmentation types in Tab. 5.

### 7.1 Augmentation Evaluation

In this section we provide a comparison for the models trained on different augmentation types. We train our transformer-600 model on cs-1, cs-

kk	Қазақстан гендерлік теңдік саласындағы 12 халықаралық құжаттарды бекітті .	Фискалдық көзқарастан гөрі либералдандыру жақсы
cs-1	казахстанский гендерлік теңдік саласындағы 12 халықаралық құжаттарды бекітті .	фискал көзқарастан гөрі либералдандыру жақсы
cs-2	казахстансктан гендерлік теңдік саласындағы 12 халықаралық құжаттарды бекітті .	фискадық көзқарастан гөрі либералдандыру жақсы
cs-3	Қазақстан гендерлік теңдік саласындағы 12 международной құжаттарды бекітті .	Фискалдық көзқарастан скорейших либералдандыру жақсы
cs-4	Қазақстан гендерлік теңдік саласындағы 12 халықаралық құжаттарды бекітті .	Фискалдық көзқарастан гөрі либерализация жақсы
cs-5	Қазақстан гендерного теңдік саласындағы 12 халықаралық құжаттарды бекітті .	Фискалдық подход гөрі либералдандыру
ru	Казахстаном ратифицировано 12 международных документов в сфере гендерного равенства .	Лучше либерализация , чем фискальный подход

Table 5: Examples of code-switching augmentations.

Data	NU	CS-1	CS-2	CS-3	CS-4	CS-5	KRCS
all data (AD)	36.03	33.10	33.41	33.00	29.64	35.16	12.25
AD + cs-1	35.07	<b>34.34</b>	33.60	<b>33.67</b>	30.51	34.25	10.20
AD + cs-2	35.54	33.78	<b>35.17</b>	33.49	30.03	34.52	11.65
AD + cs-3	34.24	33.37	32.94	33.25	29.53	33.42	10.22
AD + cs-4	35.58	32.87	33.10	32.74	<b>33.69</b>	37.03	11.38
AD + cs-5	<b>36.83</b>	33.68	34.18	33.63	32.96	<b>39.05</b>	<b>12.49</b>

Table 6: The BLEU scores for transformer-600 model on differently augmented datasets.

2, cs-3, cs-4 and cs-5 augmented datasets. We evaluate the trained models on testing subset of NU dataset, and its augmented versions. A version of NU test set augmented with cs-1 is called CS-1, the other types are called in the same manner. More importantly we evaluate the models on KRCS dataset. The results are presented in Tables 6.

Interesting, that the only augmentation type which helps to improve the baseline results is cs-5. All other types are leading to decrease in quality. For all the types, except cs-3, the evaluation on corresponding augmented testset is the best. For cs-3 the best result is achieved by a model trained on CS-1, this result is not surprising since the cs-3 augmentation is just a random choice between cs-1 and cs-2 augmentations. Another interesting point is that cs-5 augmentation allowed a model to achieve the best performance on the original testset. We hypothesize that this augmentation produces the closest data distribution to the spoken Kazakh

language, thus effectively extending the train-set.

## 8 Domain Adaptation

As one can conclude from section 3, there is a domain mismatch for the available training data and collected evaluation data. We provide a visualization of this mismatch in Fig. 1. It is a tSNE projection of LaBSE embeddings (Feng et al., 2020) of the Kazakh sentences from the training datasets and Russian sentences from Russian Tweet Corpus. One can see that centroid of Russian Tweet Corpus is closer to the centroid of KRCS dataset than any other one of another dataset. This observation drove us to conclusion that we might need a domain adaptation.

Since Russian Tweet Corpus is a monolingual Russian language dataset, we translated it to Kazakh using publicly available machine translation model nllb-200-distilled-600M from

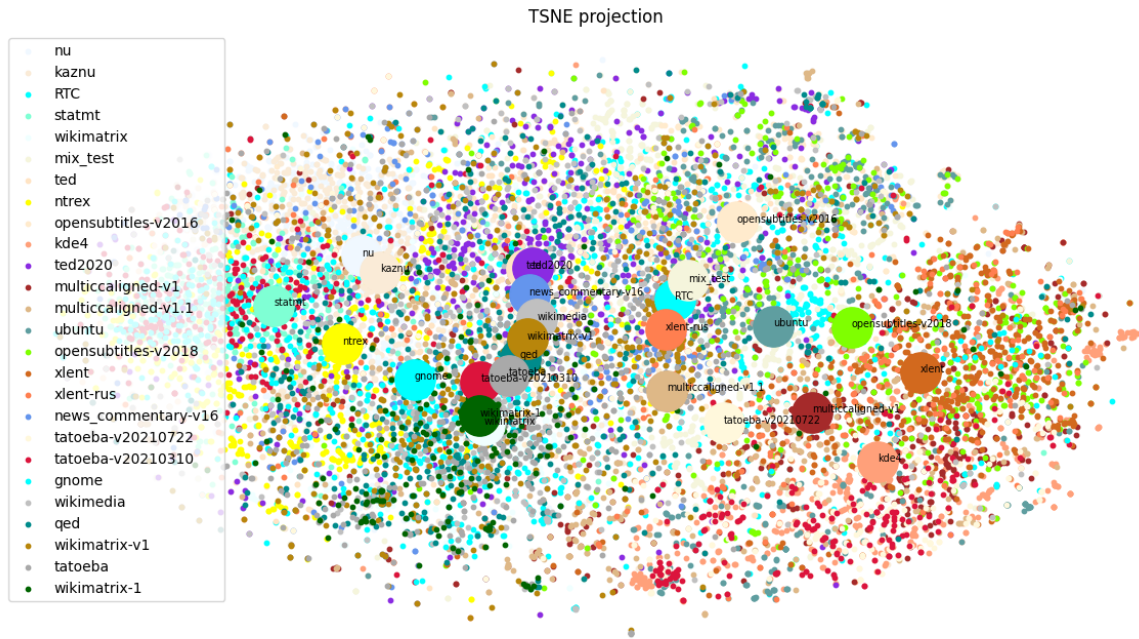


Figure 1: Sentence embedding visualization with dataset centroids.

NLLB model family described in (Costa-jussà et al., 2022). Our choice of the model was driven by the fact that it shows the best quality in standard Russian-Kazakh translation.

### 8.1 Domain Adaptation Evaluation

We decided to evaluate the importance of domain adaptation corpus which is extend our training dataset. We trained our transformer baseline model in three setups, namely: whole training data, including RTC, whole training data, *excluding* RTC, and RTC only. The experiments show that domain adaptation is indeed important, but the single domain adaptation data is not enough to achieve high performance in code switching task. These results are in Tab. 7.

Data	KRCS
all data	12.25 / 37.10 / 0.52
all data w/o RTC	11.64 / 35.58 / 0.49
RTC only	10.86 / 34.76 / 0.52

Table 7: The results of training on different datasets.

## 9 Conclusion

In conclusion, the proposed method demonstrates a viable approach to tackling machine

translation challenges for low-resource, code-switched language pairs, specifically Kazakh-Russian. By utilizing synthetic data generation, the method circumvents the need for labeled training data, which is typically scarce for such language pairs.

Furthermore, the introduction of the first code-switching Kazakh-Russian parallel corpus represents a significant contribution to the field, providing a valuable resource for future research and development. The empirical results indicate that the system’s performance surpasses that of an existing commercial translation system, as evidenced by superior human evaluation outcomes. This highlights the effectiveness and potential of the proposed approach for improving machine translation in similar low-resource, code-switched contexts.

## 10 Limitations

**Synthetic Data Dependence:** The approach relies heavily on the generation of synthetic data, which may not perfectly capture the nuances and complexities of natural code-switching in Kazakh-Russian speech.

**Evaluation Scope:** While achieving a BLEU score of 16.48 is promising, the evaluation is limited to specific criteria and doesn’t necessarily account for all aspects of translation quality, such as fluency and contextual accuracy.

**Corpus Size and Diversity:** The newly presented code-switching Kazakh-Russian parallel corpus may still be limited in size and diversity, potentially impacting the generalizability of the model to broader linguistic contexts or different dialects.

**Commercial System Comparison:** The performance comparison to an existing commercial system is based on certain benchmarks and human evaluations, which might not cover all practical use cases and scenarios where the commercial system might excel.

**Scalability and Adaptability:** The method’s scalability to other low-resource, code-switched language pairs is not addressed, raising questions about its broader applicability and adaptability to different linguistic environments.

**Long-term Sustainability:** There is no discussion on the long-term sustainability and maintenance of the synthetic data generation process and how it might evolve with changes in the language pair dynamics or increased data availability.

By acknowledging these limitations, future research can focus on addressing these gaps to further enhance the robustness and applicability of machine translation models for code-switched languages.

## Acknowledgments

The authors are grateful to Zhannura Manapbekova who did the initial collection and annotation of KRCS corpus. The work of Valentin Malykh was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

## References

Mohamed Anwar. 2023. The effect of alignment objectives on code-switching translation. *arXiv preprint arXiv:2309.05044*.

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT at CALCS2021: English to Hinglish

neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 31–35, Online. Association for Computational Linguistics.

- Abduali Balzhan, Zhadyra Akhmadieva, Saule Zholdybekova, Ualsher Tukeyev, and Diana Rakhimova. 2015. Study of the problem of creating structural transfer rules and lexical selection for the kazakh-russian machine translation system on apertium platform. In *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE "TURKIC LANGUAGES PROCESSING" TurkLang-2015*, pages 5–9.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.
- Ishali Jadhav, Aditi Kanade, Vishesh Waghmare, Sahej Singh Chandok, and Ashwini Jarali. 2022. Code-mixed hinglish to english language translation framework. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 684–688.

- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Zhanibek Kozhimbayev and Talgat Islamgozhayev. 2023. Cascade speech translation for the kazakh language. *Applied Sciences*, 13(15):8900.
- Ying Li, Yue Yu, and Pascale Fung. 2012. [A Mandarin-English code-switching corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2515–2519, Istanbul, Turkey. European Language Resources Association (ELRA).
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*.
- Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.
- Diana R Rakhimova. 2020. Normalization of kazakh language words. *Journal Scientific and Technical Of Information Technologies, Mechanics and Optics*, 128(4):545–551.
- Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *EMNLP 2020*, pages 1627–1643.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Dmitrii Ubskii, Yuri Matveev, and Wolfgang Minker. 2020. Impact of using a bilingual model on kazakh-russian code-switching speech recognition. In *CEUR Workshop Proceedings*, pages 1–6.
- Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. [End-to-end speech translation for code switched speech](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1435–1448, Dublin, Ireland. Association for Computational Linguistics.
- Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Tamar Solorio. 2023. [The decades progress on code-switching research in NLP: A systematic survey on trends and challenges](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.
- Jitao Xu and François Yvon. 2021. [Can you traduir this? machine translation for code-switched input](#). In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94, Online. Association for Computational Linguistics.
- Sholpan K Zharkynbekova and Valeria E Chernyavskaya. 2022. Kazakh-russian bilingual practice: Code-mixing as a resource in communicative interaction. *RUDN Journal of Language Studies, Semiotics and Semantics*, 13(2):468–482.
- Ю Рубцова. 2012. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора. *Инженерия знаний и технологии семантического веба*, 1:109–116.



## A Baseline Implementation Details

Transformer-600 is implemented in fairseq framework (Ott et al., 2019). The model has 6 encoder layers and 6 decoder layers with hidden size of 512. Feed forward network hidden dimension is 4096, there are 8 attention heads for encoder and for decoder. Layer normalization before each encoder and decoder block is applied. For regularization we apply dropout of 0.3, Attention dropout of 0.2 and ReLU dropout of 0.2 (which in a dropout probability after ReLU in FFN). The embedding matrices for encoder input, decoder input and decoder output are all shared. The model was optimized using Adam (Kingma and Ba, 2014) with betas of (0.9, 0.98) and epsilon  $1e^{-0.6}$ . Scheduler is inverse square root with initial learning rate of  $3e^{-0.5}$  and warmup of 2500 updates. Max tokens per batch is 2048. Maximum number of updates is 500000. Criterion is label smoothed cross entropy with smoothing factor of 0.2 following (Szegedy et al., 2016). The hyperparameters of the Transformer-600 model are presented in Tab. 8.

Number of layers	6
Hidden size	512
FFN hidden dimension	4096
Attention heads	8
LN before blocks	True
Max Tokens	2048
Criterion	label smoothed CE
Label smoothing	0.2
Optimizer	adam
Adam epsilon	1e-06
Adam betas	(0.9, 0.98)
Lr scheduler	inverse sqrt
Lr	3e-05
Warmup updates	2500
Dropout	0.3
ReLU dropout	0.2
Attention dropout	0.2
Share all embeddings	True
Max update	500000

Table 8: Model Hyperparameters. LN stands for Layer Normalization. CE stands for Cross-Entropy.

## B Train Datasets

The statistics for the training datasets is presented in Tab. 9. For Russian Tweet Corpus we report number of Kazakh tokens for the generated translation.

## C Additional Scores

The additional statistics for the baseline evaluation on augmented datasets is presented in Tab. 10.

## D Augmentation Analysis

**cs-1:** Replace a Kazakh word with a Russian one in normal form Linguistic

*Soundness:* This approach is straightforward and resembles natural code-switching seen in everyday speech, where speakers often insert words from another language in their base form, especially nouns and technical terms.

*Examples:* In Kazakh media and daily conversations, you might hear sentences like “Мен жаңа ручка сатып алдым” (“I bought a new pen”), where “ручка” is a Russian-origin word used in its normal form.

*Usage Contexts:* Such patterns are common in informal speech, especially when referring to modern or technical terms for which there might be no direct equivalent in Kazakh.

**cs-2:** Replace a Kazakh word with a Russian word’s stem with Kazakh ending

*Linguistic Soundness:* This is somewhat less natural, as it involves morphologically adapting Russian stems with Kazakh endings, which does not always fit the natural phonological or morphological rules of Kazakh. However, speakers often perform such blending to maintain grammatical consistency within a sentence.

*Examples:* This is occasionally seen in youth slang or creative language use in social media where Kazakh speakers playfully adapt Russian words. For instance, “жазаты” (from Russian “писать” but adapted to sound more Kazakh) might appear in informal texts, though not formally accepted.

*Usage Contexts:* This type of adaptation is mostly informal, often perceived as a playful or creative linguistic exercise rather than standard usage.

Dataset Name	#Sentences	#Ave. Tokens	Domain
NU (Kozhirbayev and Islamgozhayev, 2023)	895372	20.58	Juridical docs
KazNU (Balzhan et al., 2015)	80627	20.74	Off. press-releases
Russian tweet corpus (Рубцова, 2012)	12752816	7.88	Social media
Statmt-news_commentary-15-kaz-rus	11735	19.43	News
Statmt-news_commentary-14-kaz-rus	9204	19.15	News
Statmt-news_commentary-16-kaz-rus	13224	19.42	News
Facebook-wikimatrix-1-kaz-rus	165109	10.09	Web docs
OPUS-tatoeba-v2-kaz-rus	2010	8.59	General
OPUS-wikimatrix-v1-kaz-rus	32807	10.47	Wikipedia
OPUS-tatoeba-v20190709-kaz-rus	2390	8.27	General
OPUS-tatoeba-v20210310-kaz-rus	2401	8.26	General
OPUS-tatoeba-v20210722-kaz-rus	2417	8.24	General
OPUS-multiccaligned-v1-kaz-rus	1841440	4.94	Web docs
OPUS-xlent-v1.1-kaz-rus	87167	2.05	Software doc-n
OPUS-kde4-v2-kaz-rus	68014	4.70	Software doc-n
OPUS-qed-v2.0a-kaz-rus	5125	10.74	Software doc-n
OPUS-opensubtitles-v2016-kaz-rus	1246	4.55	Subtitles
OPUS-ubuntu-v14.10-kaz-rus	235	4.13	Software doc-n
OPUS-wikimedia-v20210402-kaz-rus	40714	16.41	Wikipedia
OPUS-tatoeba-v20200531-kaz-rus	2400	8.26	General
OPUS-multiccaligned-v1.1-kaz-rus	431952	12.04	Web docs
OPUS-ted2020-v1-kaz-rus	9484	12.05	Subtitles
OPUS-opensubtitles-v2018-kaz-rus	2223	4.21	Subtitles
OPUS-news_commentary-v14-kaz-rus	9163	19.12	News
OPUS-news_commentary-v16-kaz-rus	9163	19.03	News
OPUS-tatoeba-v20220303-kaz-rus	2418	8.59	General
OPUS-xlent-v1-kaz-rus	307929	2.05	Software doc-n
OPUS-gnome-v1-kaz-rus	20550	3.07	Software doc-n
OPUS-tatoeba-v20201109-kaz-rus	2401	8.26	General
all data (dedup.)	20424090		Mixed

Table 9: Train datasets statistics.

Data	NU	CS-1	CS-2	CS-3	CS-4	CS-5	KRCS
all data (AD)	61.28 / 0.82	59.58 / 0.76	59.44 / 0.76	58.96 / 0.75	56.73 / 0.69	61.78 / 0.78	<b>37.10</b> / 0.52
AD + cs-1	60.64 / 0.81	<b>60.13</b> / <b>0.77</b>	59.67 / 0.77	59.51 / 0.76	56.95 / 0.69	60.47 / 0.77	34.52 / 0.51
AD + cs-2	61.09 / 0.82	59.67 / 0.77	<b>60.85</b> / <b>0.78</b>	59.53 / 0.76	56.76 / 0.69	61.12 / 0.77	36.02 / 0.52
AD + cs-3	60.33 / 0.81	59.62 / 0.77	59.50 / 0.77	<b>59.59</b> / <b>0.76</b>	56.18 / 0.69	59.82 / 0.77	33.72 / 0.50
AD + cs-4	59.81 / 0.81	58.16 / 0.74	58.33 / 0.74	58.16 / 0.74	<b>59.15</b> / <b>0.69</b>	62.56 / 0.77	34.81 / 0.51
AD + cs-5	<b>61.63</b> / <b>0.82</b>	59.22 / 0.75	59.68 / 0.75	59.20 / 0.74	58.81 / 0.69	<b>64.27</b> / <b>0.79</b>	36.44 / <b>0.54</b>

Table 10: The ChrF++ and COMET scores for transformer-600 model on differently augmented datasets.

**cs-3:** Replace a Kazakh word with a Russian one in random form

*Linguistic Soundness:* This approach might lack naturalness as it disregards context, grammar, and sentence flow. The randomness can introduce syntactic or morphological anomalies.

*Examples:* You might hear mismatched forms in spontaneous bilingual speech, particularly among less proficient speakers who switch languages mid-sentence without full grammat-

ical integration. For example, “Мен пошел домой” (“I went home” mixing Kazakh and Russian), where the Russian verb form is not conjugated correctly according to Kazakh syntax.

*Usage Contexts:* Common in highly informal settings, such as among bilingual children or learners who are not fully competent in both languages.

**cs-4:** Replace a Kazakh word with a Russian word aligned using FastAlign *Linguistic Soundness*: Using statistical alignments like FastAlign generally improves the naturalness of word replacements because it considers contextual word pairs frequently appearing together in parallel corpora.

*Examples:* News broadcasts or bilingual podcasts often use consistent patterns of switching, aligning with how FastAlign might map Kazakh-Russian sentence structures. For example, “Менің ойымша, это не совсем правильно” (“I think this is not quite right”) frequently occurs.

*Usage Contexts:* Seen in media content where consistent patterns in code-switching reflect translation or repeated bilingual interactions.

**cs-5:** Replace a Kazakh word with a Russian word aligned using SimAlign

*Linguistic Soundness:* SimAlign uses contextual embeddings, making this approach more linguistically sound as it considers sentence-level semantics for alignment. This tends to produce contextually appropriate and grammatically fitting replacements.

*Examples:* In digital content, such as YouTube videos or podcasts with bilingual speakers, there are instances like “Бұл өте интересно тақырып” (“This is a very interesting topic”), where alignment mirrors natural bilingual communication.

*Usage Contexts:* Common in both formal and informal settings, particularly where speakers frequently shift between languages without disrupting the overall meaning.