# Shughni Machine Translation Enhanced by Donor Languages

**Dmitry Novokshanov[1], Innokentiy S. Humonen[2, 3, 4], Ilya Makarov[2,3, 4],**
[1]HSE University, [2]AXXX, [3]iMak Lab [4]Trusted AI Research Center, RAS
**Correspondence author:** danovokshanov@gmail.com

## Abstract

This paper presents the first machine translation system for Shughni, an extremely low-resource Eastern Iranian language spoken in Tajikistan and Afghanistan. We fine-tune NLLB-200 models and explore auxiliary language selection through typological similarity and "super-donor" experiments. Our final Shughni–Russian model achieves a chrF++ score of 36.3 (45.7 on BivalTyp data), establishing the first computational translation resource for this language. Beyond reporting system performance, this work demonstrates a practical path toward supporting languages with virtually no prior MT resources.

Our demo system with Shughni-Russian-English translation (Russian serves as a pivot language for the Shughni-English pair) is available on HuggingFace (https://huggingface.co/spaces/Novokshanov/Shughni-Translator).

## 1 Introduction

Machine translation (MT) has advanced rapidly with the emergence of neural MT (NMT) (Bahdanau et al., 2014), particularly transformer-based architectures (Vaswani et al., 2017). Yet these breakthroughs mostly benefit high-resource language pairs, while low-resource languages remain severely underserved (Koehn and Knowles, 2017).

To mitigate data scarcity, low-resource MT typically leverages transfer learning (Zoph et al., 2016), multilingual models (Johnson et al., 2017), and data augmentation such as back-translation (Sennrich et al., 2015) or pivoting (Cheng et al., 2016; Cheng, 2019; Kim et al., 2019). Multilingual systems show that parameter sharing across related languages can substantially improve translation quality, raising the question of how to select the most beneficial auxiliary languages. A major step in multilingual NMT has been the No Language Left Behind (NLLB) project, which released a single model for 200 languages, including Iranian languages such as Pashto, Tajik, Dari, and Persian (Team et al., 2022), and provided a strong foundation for fine-tuning.

Building on these insights, we target the Shughni–Russian translation pair. Based on our examination of the impact of auxiliary languages — chosen by typological similarity or "super-donor" status — and of isolated factors such as token overlap, we present the first Shughni–Russian translation model, obtaining scores of 39.3 METEOR, 14.6 BLEU, and 36.3 chrF++.

Shughni, spoken by approximately 90,000 people in Tajikistan and Afghanistan (Kalandarov, 2018; Edel'man and Jusufbekov, 2000; Wendtland, 2009), is an Eastern Iranian language with extremely limited written resources. Its lack of standardized orthography and the small number of available texts make it a quintessential low-resource case.

Current large language models perform poorly: in tests with recent proprietary systems[1] via the LLM Arena platform[2], only trivial sentences (e.g., *Karamsho read an interesting book*) are translated from Shughni to Russian with minor errors, while translation into Shughni fails entirely. Even the latest ChatGPT 5.2 translated a simple sentence into Tajik instead of Shughni, as shown in Figure 1. We provide a comparison of our system's metrics with those of different open-source LLMs on our test set in the Results section.

## 2 Data

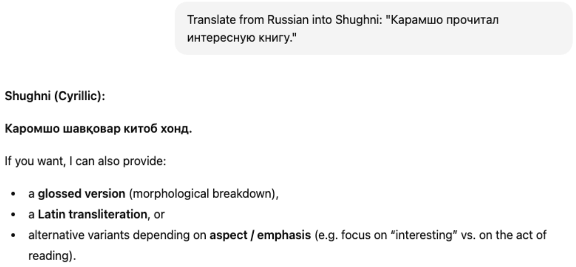
### 2.1 Training and validation

Our parallel Shughni–Russian data comes from two sources: a Shughni corpus (Makarov et al., 2022) and a Shughni–Russian dictionary (Makarov et al., 2022). The language corpus contains oral and writ-

[1] Model names: *claude-sonnet-4-20250514*, *o4-mini-2025-04-16*, *YandexGPT 5 Pro*, *GigaChat 2 Max*.

[2] `LLM Arena` v0.4.2, https://llmarena.ru/.

*Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP)*, pages 114–120
March 29, 2026 ©2026 Association for Computational Linguistics



Translate from Russian into Shughni: "Карамшо прочитал интересную книгу."

**Shughni (Cyrillic):**

**Каромшо шавқовар китоб хонд.**

If you want, I can also provide:

- a **glossed version** (morphological breakdown),
- a **Latin transliteration**, or
- alternative variants depending on **aspect / emphasis** (e.g. focus on “interesting” vs. on the act of reading).

Figure 1: Example of translation into Shughni by ChatGPT 5.2. The correct translation is given in Figure 2.

ten texts of different genres, including fairy tales, stories, prose fiction, poetry, and riddles. Some texts are only annotated via a morphological parser and do not include translations. For our purposes, we take oral texts and fairy tales translated into Russian by native speakers, as well as Gospel fragments aligned with the Russian version (4,256 sentences in total). The Shughni–Russian dictionary contains ordinary entries (21,871) and example sentences (43,300). For dictionary entries, only the first translation equivalent was retained. After preprocessing, we obtained 69,227 training pairs.

Preprocessing included removal of annotations, special symbols, and stress marks; lowercasing and whitespace normalization; and orthographic unification. Since Shughni lacks a standardized script, we normalized all data into one of the accepted Cyrillic scripts using the converter by (Makarov et al., 2022) in order to maximize token overlap with Russian and Tajik and to ensure consistency across resources. In our demo, we also added support for the Latin script for both input and output.

Auxiliary language data were added from OPUS corpora (Tiedemann, 2012), prioritizing similar domains and excluding sources used in NLLB pretraining. For each auxiliary language, we sampled datasets of equal size to the Shughni corpus (69k parallel sentences).

### 2.2 Test sets

In the present study, two test sets are used. The first was obtained from the BivalTyp project (Say, 2020), a typological database of bivalent predicates and their encoding frames, which contains 123 Shughni-English sentence pairs (Chistiakova and Ryzhova, 2023). The sentences are quite short but varied. All sentences from this set include two arguments and a predicate connecting them; at the same time, the predicate meanings are selected variously in order to illustrate different argument encoding strategies. Thus, we also check how well the argument structure of the source sentence is preserved in translation. The sentences were translated into Russian manually from English. Figure 2 presents an example of a parallel sentence from this set.

| **Russian** | **Shughni (Cyrillic)** |
|---|---|
| Карамшо прочитал интересную книгу. | Карамшойи аҷойӣб китоб х̌е̂йд. |

Figure 2: Example parallel sentence from the BivalTyp test set. English translation: *Karamsho read an interesting book.*

For the second test set, we manually selected 110 sentences from NLLB-MD (the NLLB Multi-Domain Dataset) (Team et al., 2022) with diverse syntactic structures. After minimal adaptation to Pamiri realities, we asked several native speakers from Khorog, Gorno-Badakhshan Autonomous Region of the Republic of Tajikistan, to translate them into Shughni. The Shughni portions of both test sets were transliterated into the same uniform Cyrillic orthography used in the training dataset.

## 3 Methodology and Experiments

We framed our experiments around fine-tuning multilingual NMT models with auxiliary languages. The design was as follows: combine Shughni–Russian data with equal-sized corpora from an auxiliary language and train a shared model in both directions.

All models were trained in the same virtual environment and with the same training parameters that were empirically found to be optimal: AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-3, weight decay of 1e-3, and a total batch size of 1024 examples. Full training configurations are available upon request.

### 3.1 Baseline

As a starting point, we compared the performance of three models traditionally used in NMT: NLLB-200 (600M), mT0-small, and ByT5-small, and Gemma 3 (4B) as a solid multilingual baseline LLM. Even though these models are multilingual, we believe they have not seen any Shughni data. For fair comparison, missing Shughni tokens were added to the NLLB-200 and mT0 tokenizers to ensure all of our limited data receive desired representations. ByT5 is a token-free model; it operates directly on raw texts and does not require tokenization. On Shughni→Russian translation, NLLB-200-distilled clearly outperformed the alternatives (see Table 1), and we selected the larger 1.3B distilled version as our baseline due to its greater stability.

| Model | Meteor | chrF++ |
|---|---|---|
| nllb-200 | **0.384** | **33.6** |
| mt0-small | 0.117 | 12.5 |
| byt5-small | 0.246 | 26.6 |
| Gemma 3 (4B) | 0.278 | 25.2 |

Table 1: Base model evaluation metrics on the Shughni → Russian validation set. Best in bold.

### 3.2 Auxiliary languages

Protasov et al. (Protasov et al., 2024) investigated whether morphological similarity enhances cross-lingual transfer in multilingual masked language modeling. They adopted features from The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) as predictors of positive knowledge transfer in mT5 pretraining. Authors identified the most important typological features and overall best donor languages.

To explore transfer learning in machine translation, we experimented with several auxiliary languages. Two were chosen for typological proximity (Pashto, Somali) and two for typological distance (Vietnamese, Zulu). Table A.1 contains WALS (Dryer and Haspelmath, 2013) features of Shughni and donor languages. This illustrates that the most important features for transfer learning in language modeling (discovered in (Protasov et al., 2024)) are similar between Shughni, Pashto, and Somali. For Shughni, Vietnamese, and Zulu, on the other hand, these features are mostly opposite. We also considered two "super donors" reported to benefit many languages (Afrikaans, Slovenian), as discovered by (Protasov et al., 2024).

### 3.3 Final model

For our final system, we use the auxiliary language that performed best in both translation directions and enhance the training data with backtranslation (Sennrich et al., 2015) of a Shughni novel (8,138 unique sentences) and additional Shughni–Tajik Gospel fragments (195 sentences). This configuration yielded the strongest overall results, establishing a usable Shughni–Russian MT system.

### 3.4 Evaluation

To assess the impact of our transfer learning and backtranslation, we report two widely used automatic metrics implemented in the *evaluate*[3] library: **METEOR** and **chrF++**. We report METEOR scores multiplied by 100 for consistency. We compare metrics on our two test sets combined.

Additionally, to verify that automatic gains reflect genuine quality, we conducted a human evaluation using the **XSTS** protocol (AI et al., 2022) on 50 sentences randomly chosen from the combined test set. Native speakers from Khorog (3 university professors and 3 students with high Russian language proficiency) rated the final system against the baseline in both directions.

## 4 Results

### 4.1 Main experiment

We evaluated models both separately and jointly on the BivalTyp and NLLB-MD test sets. Table 2 summarizes key results: the NLLB-200 baseline, systems enhanced by auxiliary languages one at a time (with Pashto yielding the best mean performance), and our final system in two model sizes. The latter is enhanced by Pashto as an auxiliary language, as well as backtranslation and a small amount of Shughni–Tajik data. Pashto brought moderate gains over the baseline in both directions, while Slovenian was useful only when translating into Russian. The final model reached a chrF++ score of 45.7 for translation into Russian on the BivalTyp test set (36.3 on the combined test set), establishing a substantial advance over all alternatives.

The 600-million-parameter model excels at translating into Shughni, where the 1.3B model is better into Russian. We suggest that given fixed training dataset of a very limited size, 600M model has a better parameter-to-data ratio for the decoder to learn a new language. See (Caillaut et al., 2024) for more on translation-focused decoder scaling laws. Encoder, on the other hand, is less dependent on size scaling (Ghorbani et al., 2021). Table 3 compares our systems with the most recent open-source LLMs and shows their superiority. Moreover, NLLB models are much more compact than even the lightest LLMs, which allows them to be run on personal devices. This is especially important in hard-to-reach areas.

[3] evaluate v0.4.2, https://github.com/huggingface/evaluate.

| | **Meteor** | | **chrF++** | |
|---|---|---|---|---|
| **Source** | **sh** | **ru** | **sh** | **ru** |
| **Target** | **ru** | **sh** | **ru** | **sh** |
| Baseline | 35.6 | 23.9 | 33.2 | 23.7 |
| Afrikaans (aux) | 37.5 | 24.0 | 34.4 | 22.6 |
| Vietnamese (aux) | 36.3 | 24.4 | 34.1 | 23.6 |
| Zulu (aux) | 36.1 | 25.2 | 33.6 | 24.3 |
| Somali (aux) | 35.6 | 24.2 | 33.6 | 23.5 |
| Pashto (aux) | *38.3* | *24.1* | *34.6* | *23.9* |
| Slovanian (aux) | 38.9 | 23.0 | 34.9 | 22.8 |
| Final 1.3B | **39.3** | 22.7 | **36.3** | 23.4 |
| Final 600M | 35.6 | **25.8** | 34.0 | **26.2** |

Table 2: Evaluation on the combined test set. Best results in bold. Best performance among auxiliary languages is underlined. Second-best performance among auxiliary languages is italicized.

| | **Meteor** | | **chrF++** | |
|---|---|---|---|---|
| **Source** | **sh** | **ru** | **sh** | **ru** |
| **Target** | **ru** | **sh** | **ru** | **sh** |
| Baseline | 35.6 | 23.9 | 33.2 | 23.7 |
| Qwen 3 (235B) | 15.8 | 10.4 | 17.5 | 14.4 |
| Gemma 3 (27B) | 16.1 | 10.3 | 18.1 | 13.3 |
| GPT-OSS (120B) | 10.1 | 6.5 | 3.9 | 3.1 |
| Final 1.3B | **39.3** | 22.7 | **36.3** | 23.4 |
| Final 600M | 35.6 | **25.8** | 34.0 | **26.2** |

Table 3: Comparison of open-source LLMs and our systems. Evaluation on combined test set.

Human evaluation was conducted comparing the final system (1.3B) and the NLLB-200 baseline (also 1.3B). The final system scored higher than the baseline in both directions, confirming the practical usability of our Shughni–Russian MT system, as shown in Table 4. For inter-annotator agreement, we report weighted Cohen's kappa (Cohen, 1968). Following common interpretive conventions (Landis and Koch, 1977), our results indicate substantial agreement under quadratic weighting ($\kappa = 0.65$). The most common errors fall into two categories: lexical errors and grammatical time expression errors.

## 5 Demo description

We additionaly present a demo translation app via HuggingFace Spaces. The model is optimized for sentence-level translation and the Cyrillic Shughni script. Therefore, the demo application automatically performs segmentation and orthography conversion. Moreover, for convenience and wider use, our demo includes translation into English. Translation in this direction is implemented through Russian as a pivot language due to the almost total absence of Shughni–English data and reaches a chrF++ score of 42.2 on the BivalTyp data.

| | **Mean XSTS** | |
|---|---|---|
| **Source** | **sh** | **ru** |
| **Target** | **ru** | **sh** |
| *Combined test set* | | |
| Baseline | 2.85 | 3.15 |
| Final (1.3B) | **3.21** | **3.61** |

Table 4: Human evaluation on Shughni → Russian and Russian → Shughni translation (Mean XSTS scores).

## 6 Conclusion

We present the first neural MT system for the Shughni language, an endangered Eastern Iranian language with virtually no prior machine translation support. Our experiments show that auxiliary languages can provide meaningful gains, though not always in line with standalone factors: Pashto proved most helpful not only because it is typologically similar but also as a close relative with potentially shared lexical units, despite differing writing systems. By combining Pashto auxiliary data with backtranslation, we achieve the best overall results and provide a translation model in two sizes. Human evaluation confirms that our improvements translate into practical usability.

## 7 Future work

Future research will examine auxiliary language choice more systematically, testing a broader set of candidates and features. Moreover, we will expand the number of language models, including the latest translation-oriented LLMs. We also plan to utilize Shughni–Russian dictionary entries more effectively and to benefit from annotations in the corpus (e.g., morphemes, glosses, part of speech, and general meaning). We believe these two sources can be a valuable source of syntactic data and can be integrated into the translation system itself. Furthermore, while Shughni is the local lingua franca, we aim to extend MT development to other Pamiri languages as part of ongoing efforts in their documentation and digital support. In the meantime, we encourage researchers and native speakers of Pamiri languages to collaborate on the expansion of the corpus and the machine translation project.

## Acknowledgments

The authors thank everyone who participated in digitizing the Shughni–Russian dictionary and collecting the Shughni corpus. We would like to express our special appreciation to the native speakers in Khorog who translated the test dataset and participated in the human evaluation.

## References

Daniel Licht META AI, Cynthia Gao META AI, Janice Lam META AI, Francisco Guzmán META AI, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. *Volume 1: MT Research Track*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Gaëtan Caillaut, Raheel Qader, Mariam Nakhlé, Jingshu Liu, and Jean-Gabriel Barthélemy. 2024. Scaling laws of decoder-only models on the multilingual machine translation task. *Preprint*, arXiv:2409.15051.

Yong Cheng. 2019. Joint training for pivot-based neural machine translation. In *Joint training for neural machine translation*, pages 41–54. Springer.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*.

Daria Chistiakova and Daria Ryzhova. 2023. Bivalent patterns in Shughni. *BivalTyp: Typological database of bivalent verbs and their encoding frames*, (5):(Available online at https://bivaltyp.info, Accessed on 26 May 2025.).

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Edel'man and Jusufbekov. 2000. Shugnanskij jazyk [Shughni Language]. In *Jazyki mira: Iranskie jazyki. III. Vostochnoiranskie jazyki*, page 225.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling laws for neural machine translation. *Preprint*, arXiv:2109.07740.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Tokhir Kalandarov. 2018. Pamirskie narody, ih jazyki i perepis': etnicheskij diskurs [The Peoples of Pamir, Their Languages and the Census: The Ethnic Discourse]. *Etnograficheskoe obozrenie*, (5):162–178.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Yury Makarov, Maksim Melenchenko, and Dmitry Novokshanov. 2022. Digital resources for the Shughni language. In *Proceedings of the workshop on resources and technologies for Indigenous, endangered and lesser-resourced languages in Eurasia within the 13th language resources and evaluation conference*, pages 61–64.

Vitaly Protasov, Elisei Stakovskii, Ekaterina Voloshina, Tatiana Shavrina, and Alexander Panchenko. 2024. Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 94–108.

Sergey Sergeevič Say. 2020. *BivalTyp: Typological database of bivalent verbs and their encoding frames. (Available online at https://bivaltyp.info, Accessed on 1 January 2026.)*. Institute for Linguistic Studies Russian Academy of Sciences.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

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, and

20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Antje Wendtland. 2009. The position of the pamir languages within east iranian. *Orientalia Suecana*, 58:172–188.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

## A WALS Features

Table A.1 contains selected WALS (Dryer and Haspelmath, 2013) features of Shughni and donor languages. The selection was based on (Protasov et al., 2024) and aimed not to show typological distance or similarity, but to illustrate our choice of donor languages in the context of machine translation and transfer learning only. This illustrates that the most important to us features between Shughni, Pashto, and Somali are nearly the same. This is the evidence of typological similarity between Shughni and Somali in one narrow fragment of the system, not in typological profiles of the entire languages.

| WALS feature | Shughni | Pashto | Somali | Vietnamese | Zulu |
|---|---|---|---|---|---|
| Prefixing vs. suffixing in inflectional morphology | Strongly suffixing | Strongly suffixing | Strongly suffixing | Little affixation | Strong prefixing |
| Order of Object and Verb & Order of Adposition and NP | Other | OV & postpositions | Other | VO & prepositions | VO & prepositions |
| Order of Object and Verb | OV | OV | OV | VO | VO |
| Preverbal negative morphemes | [Neg–V] | [Neg–V] | Neg V | Neg V | [Neg–V] |
| Order of demonstrative and noun | Demonstrative–Noun | Demonstrative–Noun | Demonstrative suffix | Noun–Demonstrative | Mixed |
| Order of numeral and noun | Numeral–Noun | Numeral–Noun | Numeral–Noun | Numeral–Noun | — |
| Coding of nominal plurality | Plural suffix | Plural suffix | Plural suffix | Plural word | Plural prefix |

Table A.1: Comparison of selected WALS features across five languages.