Adapting Multilingual NMT to Language Isolates: The Role of Proxy Language Selection and Dialect Handling for Nivkh

Eleonora Izmailova[1,2], **Alexey Sorokin**[2,3], **Pavel Grashchenkov**[1,4]

[1]RCC MSU, [2]MSU Center of Artificial Intelligence,
[3]Yandex, [4]Philological Faculty of MSU

Correspondence: eleon.izm@gmail.com

Abstract

Neural machine translation has achieved remarkable results for high-resource languages, yet language isolates – those with no demonstrated genetic relatives – remain severely underserved, as they cannot benefit from cross-lingual transfer with related languages. We present the first NMT system for Nivkh, a critically endangered language isolate spoken by fewer than 100 fluent speakers in the Russian Far East. Working with approximately 9.5k parallel sentences – expanded through fine-tuned LaBSE sentence alignment – we adapt NLLB-200 to Nivkh-Russian translation. Since Nivkh is absent from NLLB's language inventory, we investigate proxy language token selection, comparing six languages varying in word order, morphological type, and script: Bashkir, Kazakh, Halh Mongolian, Turkish, Tajik, and French. We find that using any proxy substantially outperforms random token initialization (18.00–19.02 vs. 15.44 for rus→niv; BLEU 20.72–21.23 vs. 19.05 for niv→rus), confirming the value of proxy-based transfer. However, the choice of proxy has minimal impact, with all six achieving comparable results despite spanning four language groups and two scripts. This suggests that for language isolates, practitioners can select any typologically reasonable proxy without significant performance penalty. We additionally present preliminary experiments on dialect-specific models for Amur and Sakhalin Nivkh. Our findings establish baseline results for future Nivkh NLP research and provide practical guidance for adapting multilingual models to other language isolates.

1 Introduction

Neural machine translation (NMT) has achieved remarkable results for high-resource languages, yet the majority of the world's approximately 7,000 languages remain underserved due to insufficient parallel data (Joshi et al., 2020). This disparity is particularly acute for language isolates – languages with no demonstrated genetic relationship to any other language – which cannot benefit from transfer learning strategies that exploit similarities with related languages (Zoph et al., 2016).

Nivkh presents an extreme case: a critically endangered isolate with only several dozen remaining speakers – none of whom use the language regularly – all over the age of 70. Spoken in the lower Amur River region and Sakhalin Island, it lacks an established orthographic standard and exhibits significant dialectal variation between its Amur and Sakhalin varieties (Gruzdeva and Bugaeva, 2022). Prior to this work, no neural machine translation system existed for Nivkh.

To illustrate the challenges Nivkh poses for NMT, consider the opening of a traditional narrative, shown here with interlinear glossing:

(1) Ӈа ху-ла нивх,
animal kill-ATR person
нам-нама-ғыт-ӯ.
be.good-be.good-COMPL-CONV.3SG
'[Жил-был] хороший охотник.'
(Eng. [There lived] a good hunter; the existential meaning is implied – Nivkh has no overt copula here.)

(2) Ӈа ӈыӈ-р ви-ке,
animal hunt-CONV.3SG go-CONV:SIM
лаю-дь.
be.stormy-IND
'Пока он был на охоте, погода испортилась.'
(Eng. While he was out hunting, the weather turned foul.)

The glosses reveal several properties that make Nivkh challenging for NMT. There is no grammatical gender: Russian requires masculine agreement (хороший охотник 'good.M hunter') with no source-side cue. Nivkh lacks an adjective category; property concepts are expressed by stative verbs, so нам-нама-ғыт-ӯ ('be.good-be.good-COMPL-CONV.3SG') simultaneously conveys the

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 127–137

property ‘good’ and, through its verbal morphology, the existential meaning ‘there was’ – which Russian renders with the stylistically formulaic жил-был. Reduplication marks intensification (нам-нама ‘be.good-be.good’), with the full form reflecting productive agglutinative verb morphology. Most notably, converbal chaining in (2) – where two non-finite forms (CONV.3SG, CONV:SIM) are followed by a finite form (IND), encoding a temporal sequence – must be restructured into a Russian finite subordinate clause (Пока он был..., погода испортилась).

Massively multilingual models such as NLLB-200 (NLLB Team et al., 2022) have demonstrated capacity for transfer to unseen languages through fine-tuning. However, these models require a language token to identify the source language during encoding and the target language during generation. For languages absent from the model’s inventory, practitioners must select a *proxy token* – but principled guidance for this selection remains scarce, particularly for isolates that lack genetic relatives among supported languages.

We address two research questions:

1. For language isolates absent from multilingual models, does proxy language selection significantly impact translation quality, and what factors predict effectiveness?

2. How should dialect variation be handled under severe data constraints?

Our contributions are:

- The first NMT system for Nivkh, achieving BLEU 21.23 (niv→rus) with approximately 7.6k training sentences
- A fine-tuned LaBSE encoder for Nivkh-Russian sentence alignment, enabling corpus expansion by approximately 2k additional parallel sentences
- Systematic comparison of six proxy languages plus random initialization, demonstrating that proxy selection provides substantial benefit over random initialization, but the choice of proxy has minimal impact
- Preliminary analysis of unified versus dialect-specific model training for Amur and Sakhalin Nivkh

1.1 The Nivkh Language

Nivkh is a language isolate indigenous to the lower Amur River basin and northern Sakhalin Island. The language exhibits several typologically unusual features relevant to NMT adaptation.

Phonology and Orthography. The consonant inventory includes 32 phonemes with contrastive aspiration (п/п', т/т', к/к', қ/қ') and palatalization (д/д', т/т', н/н'), as well as a typologically rare voiceless trill р̌, which patterns with obstruents rather than sonorants (Kreinovich, 1937). The vowel system comprises eight phonemes, including the central vowels ы and ъ; the Amur dialect has additionally developed phonemic vowel length through the historical loss of uvular consonants. Nivkh uses Cyrillic script supplemented with characters absent from Russian – including ғ, ӷ, ӈ, ӽ, and р̌ – necessitating vocabulary expansion when adapting NLLB models.

Morphology. Nivkh is predominantly agglutinative with polysynthetic tendencies, including productive noun incorporation, creating challenges for subword tokenization due to morphophonological alternations at morpheme boundaries. The language employs an elaborate system of 26 numeral classifiers requiring semantic categorization of nominal referents (Panfilov, 1962). The case system exhibits syncretism, and the language features associative plurals distinct from simple plurality.

Syntax. Basic word order is SOV, contrasting with Russian SVO order. Nivkh exhibits “paradoxical finiteness”: converbs (typically non-finite) carry person and tense marking, while indicative forms lack such marking. Number agreement is optional, contributing to morphosyntactic variability.

Orthographic variation. Unlike most written languages, Nivkh lacks an established orthographic standard. Sources spanning 1908–2022 exhibit significant spelling variation, compounded by dialectal differences in phoneme inventories (e.g., Sakhalin retains uvular consonant ғ lost in Amur). This introduces noise into training data.

Dialects. Two major varieties exist: Amur and Sakhalin. These differ substantially in phonology, lexicon, and morphology. Some classifications treat them as separate languages.

Genre	Sources	Sents
Folklore	shtrn, pnf	3,439
Literary	sng, gdn, ptmn, etc.	2,132
Periodicals	nd	1,353
Educational	grz, tmn	1,150
Transcribed speech	shrsh, kn	564
Religious	bible	331
Total		**9,496**

Table 1: Corpus composition by genre. Sources include linguistic documentation, native speaker recordings, the periodical *Nivkh Dif*, and Bible translations.

1.2 Related Work

Low-resource NMT. Transfer learning from related languages has proven effective for low-resource translation (Zoph et al., 2016; Neubig and Hu, 2018). However, this approach assumes availability of a related high-resource language, unavailable for isolates by definition.

Multilingual models for extremely low-resource languages. NLLB-200 (NLLB Team et al., 2022) supports 200 languages but many endangered languages remain absent. Prior work on adapting NLLB to unseen languages includes Dale (2022) achieving BLEU 19.7/17.7 (myv↔rus) for Erzya with 57k sentence pairs, Tuvan reaching BLEU 25.2/23.2 (tyv↔rus) with 50k pairs (Dale, 2023), and Jerpelea et al. (2025) reporting BLEU 31.0/17.3 (rup↔ron) for Aromanian with 79k sentences. Notably, these languages belong to families represented in NLLB (Uralic, Turkic, Romance), providing transfer advantages unavailable to isolates.

Proxy language selection. For languages absent from multilingual models, practitioners must select a proxy token. Dale (2023) provides practical guidance but does not systematically investigate which factors predict effectiveness. We hypothesized that typological similarity – particularly word order and morphological type – might predict transfer effectiveness independently of genetic relatedness.

2 Data

2.1 Corpus Sources

Our corpus derives from heterogeneous sources spanning both dialects and multiple genres (Table 1).

Linguistic documentation. Glossed texts from Shternberg (1908), Kreinovich (1934), Panfilov (1965), and Gruzdeva and Bugaeva (2022) provided high-quality parallel material with interlinear translations. An existing corpus of approximately 3k verified parallel sentences (Gusev and Idrisov, 2019) partially compiled from these sources served as our starting point. We expanded this to approximately 7k verified pairs by incorporating and manually aligning material from the additional sources described below, and this expanded set was then used for encoder fine-tuning (Section 2.2).

Native speaker recordings. Transcribed audio collections by Shiraishi and collaborators (2002–2015)[1] contributed naturalistic speech data.

Periodicals. The newspaper *Nivkh Dif* (est. 1990), the only Nivkh-language periodical, provided contemporary text. Russian and Nivkh content required alignment as it is not strictly parallel.

Religious texts. Bible translations into both dialects offered structurally aligned material.

2.2 LaBSE Fine-tuning for Sentence Alignment

To expand the parallel corpus from loosely aligned texts, we fine-tuned LaBSE (Feng et al., 2022), a language-agnostic BERT-based sentence encoder. Since Nivkh is absent from LaBSE's training data, zero-shot alignment quality was poor. Before fine-tuning, we applied a mapping of over 50 character-level normalization rules to reconcile the divergent transcription conventions used across sources spanning 1908–2022: e.g., historical variants of the retroflex trill (p̃, p̌, p̀ → p̌), alternate velar nasal forms (ᴎ, ᴎ́ → ŋ), macron vowels (ā, ō, ī) → plain vowels, and various dash and quotation mark variants to ASCII equivalents. This normalization was applied to both the fine-tuning data and the texts submitted for alignment.

We fine-tuned on 7,064 verified parallel sentences using `MultipleNegativesRankingLoss` (Henderson et al., 2019), reserving 409 sentences for evaluation. This contrastive objective treats other sentences in each batch as negative examples, teaching the model to recognize Nivkh–Russian translation equivalence. Training hyperparameters are shown in Table 2. Model selection used a task-specific metric rather than validation loss: at every 100 training steps, the evaluator re-ran the full

[1] http://ext-web.edu.sgu.ac.jp/hidetos/HTML/SMNStitle.html

Parameter	Value
Base model	LaBSE
Training sentences	7,064
Test sentences	409
Batch size	8
Learning rate	2e-5
Scheduler	Warmup cosine
Warmup steps	1,000
Epochs	2
Optimizer	AdamW
Mixed precision	Yes (AMP)

Table 2: LaBSE fine-tuning hyperparameters.

alignment pipeline on a held-out text pair and computed the *chain score* – the proportion of aligned sentence pairs forming unbroken monotonic chains (a score of 1.0 indicates fully monotonic alignment with no crossing or missing links). The checkpoint with the highest chain score was saved.

Alignment procedure. For alignment we used the `lingtrain-aligner` library[2] with the fine-tuned LaBSE model. The library computes pairwise cosine similarities between sentence embeddings within a sliding window, builds an initial monotonic alignment path through the similarity matrix, and then resolves conflicts – regions where the alignment chain is broken – by re-embedding and re-scoring candidate pairs. We used a window size of 10, batch size of 500, and conflict resolution with minimum chain length 2 and maximum conflict length 6. Given the inability to manually verify large portions of the output, we additionally applied strict filtering: minimum cosine similarity threshold of 0.6 (versus the default 0.5) and chain score validation requiring consistent sequential alignment patterns.

Results. From 9,348 Nivkh and 9,128 Russian sentences in unaligned source texts, filtering retained approximately 80% in both languages (7,421 Nivkh, 7,201 Russian). The alignment procedure yielded 1,927 new high-confidence parallel sentence pairs. Combined with the original verified corpus, this produced 9,496 total parallel sentences. Chain score on the held-out evaluation texts improved from 0.35 at initialization to 0.96 after fine-tuning, indicating substantially more reliable cross-lingual sentence matching for Nivkh despite its complete absence from LaBSE's pretraining data.

[2] https://github.com/averkij/lingtrain-aligner

Dataset	Train	Dev	Test
Unified (all data)	7,599	948	949
Amur dialect	3,595	449	449
Sakhalin dialect	4,004	499	500

Table 3: Dataset splits. Dialect-specific splits are subsets of the unified data, stratified by source to maintain genre distribution.

2.3 Data Splits

Table 3 shows the data splits. The unified dataset was split 80/10/10 stratified by source to maintain genre distribution. Mean sentence length is 11.27 words (std: 9.82, range: 1–120).

Dialect distribution. Sakhalin sources comprise 5,003 sentences; Amur sources comprise 4,493 sentences. The dialects differ in mean sentence length (Sakhalin: 9.53 words; Amur: 13.22 words), reflecting substantial genre imbalance: Sakhalin data is predominantly folklore and literary texts (93%), while Amur data spans periodicals (30%), folklore (26%), educational materials (18%), and transcribed speech (16%). This compositional difference may contribute to performance variation beyond purely linguistic factors.

3 Experimental Setup

3.1 Model Architecture

We use NLLB-200-distilled-600M (NLLB Team et al., 2022), a 600M parameter encoder-decoder model supporting 200 languages. The model employs SentencePiece tokenization with a vocabulary of 256k tokens.

3.2 Vocabulary Expansion

Nivkh uses Cyrillic characters absent from standard Russian. We trained a SentencePiece model on Nivkh texts and merged novel subword tokens into NLLB's vocabulary, adding approximately 7k tokens. New token embeddings were initialized as the mean of their constituent characters' embeddings in the original tokenizer.

3.3 Proxy Language Selection

Since Nivkh is absent from NLLB-200, we must select an existing language token to represent Nivkh during fine-tuning. We compared six proxy languages chosen to vary along dimensions of typological similarity, script, and language family (Table 4).

Proxy	Group	Script	Order	Morph.
Bashkir	Turkic	Cyrillic	SOV	Agglut.
Kazakh	Turkic	Cyrillic	SOV	Agglut.
Mongolian	Mongolic	Cyrillic	SOV	Agglut.
Turkish	Turkic	Latin	SOV	Agglut.
Tajik	Iranian	Cyrillic	SOV	Fusional
French	Romance	Latin	SVO	Fusional

Table 4: Proxy languages compared. Nivkh is SOV with agglutinative morphology and uses Cyrillic script. Bashkir, Kazakh, Mongolian, and Turkish share SOV order and agglutinative morphology; Tajik shares SOV order but has fusional morphology; French differs on all typological dimensions.

Typologically similar proxies. Bashkir (`bak_Cyrl`), Kazakh (`kaz_Cyrl`), Mongolian (`khk_Cyrl`), and Turkish (`tur_Latn`) share SOV word order and agglutinative morphology with Nivkh. Mongolian has documented lexical overlap with Nivkh, possibly reflecting historical contact (Müller et al., 2013).

Partially similar proxy. Tajik (`tgk_Cyrl`) shares SOV order and Cyrillic script but has fusional rather than agglutinative morphology, testing whether word order alone predicts effectiveness.

Dissimilar control. French (`fra_Latn`) differs from Nivkh on all dimensions (SVO, fusional, Latin script), serving as a negative control.

Excluded proxies. We avoided Slavic proxies (e.g., Bulgarian, Ukrainian) given that Russian is our target language, reasoning that target-adjacent proxies might behave differently than target-distant ones – though we leave this comparison to future work. Among Cyrillic-script languages in NLLB-200, most are either Slavic or Turkic; Tajik is a notable exception as an Iranian language with Cyrillic script, making it useful for separating script effects from language family effects.

3.4 Training Configuration

All models were trained with identical hyperparameters using the Adafactor optimizer (Shazeer and Stern, 2018) with learning rate 2e-4, weight decay 1e-3, gradient clipping threshold 1.0, and constant learning rate schedule with 500 warmup steps. Batch size was 64 with maximum sequence length 128. Training was bidirectional (niv↔rus) with evaluation every 500 steps and early stopping with patience 3 based on development set chrF++. Maximum training steps was 10k. All experiments used a single NVIDIA A100 80GB GPU.

Proxy	niv→rus		rus→niv	
	BLEU	chrF++	BLEU	chrF++
`bak_Cyrl`	**21.23**	41.44	18.57	42.70
`kaz_Cyrl`	21.08	41.39	18.22	42.55
`tgk_Cyrl`	21.09	**42.20**	18.00	42.97
`khk_Cyrl`	20.75	40.91	18.03	42.68
`tur_Latn`	20.80	41.03	18.08	42.74
`fra_Latn`	20.72	41.17	**19.02**	**43.26**
Random init	19.05	39.05	15.44	37.43

Table 5: Proxy language comparison on unified test set. Best results in bold. The bottom row shows performance with a randomly initialized `nivkh_Cyrl` token (no proxy).

Convergence behavior. Models using proxy languages converged faster than random initialization: Bashkir at 4k steps, Tajik at 4.5k, Kazakh at 5.5k, French at 5k, Turkish at 6.5k, and Mongolian at 7.5k steps. The randomly initialized model required the full 10k steps, suggesting that proxy tokens provide useful initialization that accelerates learning.

Generation parameters. For evaluation, we used greedy decoding with maximum output length set dynamically as $16 + 1.5 \times |x|$ where $|x|$ is the input sequence length.

3.5 Evaluation

We report BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) computed with SacreBLEU (Post, 2018) on the held-out test set (949 sentences).

4 Results

4.1 Proxy Language Comparison

Table 5 presents translation quality across all six proxy languages, plus a baseline using a randomly initialized Nivkh token without any proxy.

All proxies substantially outperform random initialization, but differences between proxies are minimal (within 0.5 BLEU for niv→rus and approximately 1 BLEU for rus→niv). We analyze these patterns in Section 5.

4.2 Dialect-Specific Results

We selected Bashkir as the proxy for dialect experiments based on its numerically highest BLEU score in the unified comparison. The cross-dialect evaluation tests whether dialect-specific models outperform the unified model on their respective test sets, and whether models transfer across dialects.

	niv→rus		rus→niv	
Training Data	BLEU	chrF++	BLEU	chrF++
Evaluated on Amur test set (449 sentences)				
Unified	18.83	40.62	19.19	42.94
Amur only	17.99	39.44	13.92	40.49
Sakhalin only	2.51	21.08	1.85	13.53
Evaluated on Sakhalin test set (500 sentences)				
Unified	23.67	42.32	17.56	42.27
Amur only	1.41	15.25	1.70	13.76
Sakhalin only	23.16	41.30	19.12	44.16

Table 6: Cross-dialect evaluation using Bashkir proxy. Cross-dialect transfer fails almost completely (BLEU <3), but the unified model performs well on both varieties.

Cross-dialect evaluation reveals near-zero transfer between Amur and Sakhalin varieties (BLEU 1.41–2.51), confirming their status as highly divergent. However, the unified model matches or exceeds dialect-specific models on both test sets (Amur: 18.83 vs 17.99; Sakhalin: 23.67 vs 23.16), suggesting that data pooling provides benefits without introducing harmful interference. We recommend the unified approach for practical deployment.

5 Discussion

5.1 Proxy vs. Random Initialization

The key findings are: (1) **using any proxy substantially outperforms random initialization** (18.00–19.02 vs. 15.44 for rus→niv; BLEU 20.72–21.23 vs. 19.05 for niv→rus), confirming that proxy language selection provides meaningful transfer; but (2) **the choice of proxy has minimal impact**, with scores ranging within 1 BLEU across all six proxies.

Bashkir achieves numerically highest BLEU for niv→rus, while French surprisingly achieves highest BLEU for rus→niv. Tajik achieves highest chrF++ in both directions. However, these differences are minimal and likely within noise margins for a test set of this size. Notably, French – our typologically dissimilar control – performs comparably to the SOV/agglutinative proxies, suggesting that NLLB's multilingual pretraining creates sufficiently language-agnostic representations that proxy selection has minimal impact after fine-tuning.

5.2 Why Proxy Choice Doesn't Matter (Much)

Given the clear benefit of using *any* proxy over random initialization, it is surprising that the *choice* of proxy has minimal impact. Several factors may explain this:

Fine-tuning dominates. With 7.6k training sentences and 4k–7.5k optimization steps, the model adapts sufficiently to Nivkh that initial proxy differences become irrelevant. The faster convergence of proxy-based models (4k–7.5k steps vs. 10k for random init) suggests proxies provide a better starting point, but all converge to similar final performance.

NLLB's language-agnostic representations. NLLB-200 was trained on 200 languages with the explicit goal of learning cross-lingually transferable representations. This multilingual pretraining may create a representation space where any proxy provides a reasonable initialization for adapting to a new language.

Script handling via vocabulary expansion. Turkish (Latin script) performs comparably to Cyrillic-script proxies, suggesting that our vocabulary expansion for Nivkh-specific characters adequately handles script differences, eliminating this potential source of variation.

5.3 Implications for Practitioners

These findings have practical implications for researchers working with other language isolates:

1. **Use a proxy, not random initialization**: The 2–3 BLEU improvement and faster convergence justify proxy selection over creating a new language token.

2. **Proxy choice is flexible**: Any typologically reasonable proxy appears sufficient. This reduces the burden of identifying an "optimal" proxy.

3. **Vocabulary expansion is essential**: Adding tokens for characters absent from the base model is critical for handling novel scripts.

4. **Data quality matters more**: Effort is better spent on corpus refinement than proxy optimization.

5.4 Comparison to Prior Work

Table 7 contextualizes our results against prior NLLB adaptation work.

Nivkh achieves BLEU 21.2 with only 7.6k training sentences – competitive with or exceeding Erzya (BLEU 19.7 with 57k sentences) despite

			X→Tgt		Tgt→X	
Lang.	**Family**	**Train**	BLEU	chrF++	BLEU	chrF++
Tuvan	Turkic	50k	25.2	49.9	23.2	49.9
Erzya	Uralic	57k	19.7	38.6	17.7	41.2
Arom.	Romance	79k	31.0	51.0	17.3	45.0
Nivkh	**Isolate**	**7.6k**	**21.2**	**41.4**	**18.6**	**42.7**

Table 7: Comparison with prior NLLB adaptation. Target: Russian for Tuvan, Erzya, Nivkh; Romanian for Aromanian. Nivkh achieves competitive scores with 7–10× smaller corpus and no related languages in NLLB.

having 7× smaller corpus and no related languages in NLLB. Tuvan's higher scores (BLEU 25.2) likely reflect both larger corpus size and the presence of related Turkic languages in NLLB. This suggests that NLLB adaptation can be effective even for isolates with severely limited data, though we note that direct comparison is complicated by differences in language pairs, evaluation sets, and domains.

5.5 Error Analysis

Manual inspection of model outputs on the test set reveals systematic error patterns that correlate with Nivkh's typological properties. We summarise these here with word-level illustrations; full sentence-level examples appear in Appendix D.

Successful translations. On formulaic and structurally simple sentences, the model achieves near-perfect output. Representative cases include converb chains such as Кузиӈ вифке, тав-наӽӷтох виӈ йугыӈ ('went out, walked far, approached a yurt and entered'), kinship introductions like Ыма, ни чхыф наныгын виндра ('Mother, I will go hunt a bear'), and simple predicates such as Нӷак иф муғиӈ видь ('Once he went by boat') – all translated without error (Table 10). These successes cluster in the `shtrn` (Shternberg) subcorpus, where literal, linguist-produced reference translations closely match the model's output register.

Converb Chain Truncation. Nivkh expresses action sequences through chains of converbs (non-finite verb forms), often encoding up to 4–6 sequential actions in a single sentence. The model frequently compresses or reorders these chains, which reflects the challenge of mapping Nivkh's serializing verb strategy onto Russian's preference for finite clause coordination. For instance, the four-step sequence изроӈ йетот инифке, сик ниихарт (skin → cook → eat.at.length → eat.all; 'having skinned and cooked, ate for a long time, ate everything up') is reduced to three steps, dropping the cooking event entirely (Table 11, Ex. c). Similarly, озиӈ кныӈкӈ виӈ вифке (rise → depart → go → go.far) has the second converb misidentified: кныӈкӈ ('departing') → насытив ('having fed'), substituting an unrelated verbal root.

Paradoxical Finiteness Mishandling. Nivkh exhibits "paradoxical finiteness": converbs carry person and tense marking (typically properties of finite verbs), while indicative forms lack such marking. The model occasionally inverts this pattern, producing Russian translations with incorrect tense or person agreement. In folklore texts, third-person narratives sometimes shift unexpectedly to first-person mid-sentence, indicating the model struggles to track discourse participants across converb boundaries.

Kinship and Gender Confusion. Nivkh lacks grammatical gender but employs distinct kinship terminology with semantic gender. Translation errors frequently conflate sibling terms (ийасх 'sister' → старший брат 'elder brother'), and lineal terms (ола 'daughter' → сын 'son'). These errors are particularly common in folklore narratives where characters are introduced through kinship relations rather than proper names. The absence of morphological gender agreement in Nivkh source text provides no redundant cues for disambiguation.

Classifier-Numeral Constructions. Nivkh's 26-class numeral classifier system presents persistent challenges. When counting entities, speakers must select the appropriate classifier based on semantic properties (e.g., animacy, shape, size). The model occasionally produces semantically incongruent translations, or classifier information is simply dropped, yielding bare numerals without the specificity encoded in the source. Besides, it can conflate proximal and distal demonstratives (ӽуӈ 'this' ↔ ту/тот 'that') by confusing the dual with cardinals: менымыӈ ('both') surfaces as две ('two'), additionally violating Russian gender agreement with masculine referents.

Polysynthetic Boundary Errors. Nivkh's productive noun incorporation creates long polymorphemic words where noun roots incorporate into verb stems. Subword tokenization sometimes segments these forms at semantically inappropriate

boundaries, yielding translations where incorporated objects are either lost entirely or rendered as separate (syntactically incorrect) constituents. This is most evident with culturally specific compounds involving traditional activities (fishing, hunting) where the incorporated noun carries crucial semantic content. For instance, the noun пхӈа ('skin') in ӿуӈ чхыф пхӈа сивуў ('this bear took off its skin') is rendered as нож ('knife'), yielding 'this bear took off its knife' – grammatically sound but semantically wrong at a single morpheme boundary (Table 11, Ex. d). This is preferable to the catastrophic failures observed in longer polysynthetic complexes, where entire clauses disintegrate.

Synonym and register variation. A substantial fraction of apparent errors reflect legitimate lexical variation penalised by reference based metrics. Three pervasive patterns are юрта ↔ дом ('yurt' ↔ 'house'), старик ↔ муж ('old man' ↔ 'husband'), and сказала ↔ ответила ('said' ↔ 'answered'). In each case the model's output is semantically equivalent to the reference but uses a different Russian lexeme, deflating BLEU and chrF++ points (Table 10).

Genre effects. Error rates vary substantially by genre: linguistic documentation with literal translations achieves BLEU scores 3–4× higher than literary texts with freer translation styles. Folklore texts, despite their formulaic structures, suffer from cultural concept gaps (e.g., shamanic terminology, traditional dwelling types) that produce semantic drift even when syntactic structure is preserved.

6 Limitations

Automatic metrics only. We rely on BLEU and chrF++; human evaluation with native speakers would strengthen conclusions but was infeasible given the critically endangered status of the language.

Statistical significance. We report single-run results without confidence intervals. The small differences between proxies may reflect noise rather than meaningful distinctions. Future work should include multiple runs with different random seeds.

Limited proxy selection. We tested six proxies; other languages (e.g., Tungusic languages, which some hypotheses link to Nivkh) are absent from NLLB and could not be evaluated.

Corpus limitations. Despite expansion, the corpus remains small by NMT standards. Genre imbalance (predominantly folklore) may limit generalization. Orthographic variation across sources spanning 1908–2022 introduces noise.

7 Conclusion

We presented the first neural machine translation system for Nivkh, a critically endangered language isolate. Our systematic comparison of six proxy languages reveals that proxy selection has minimal impact on translation quality when fine-tuning NLLB-200 – a finding with practical implications for researchers working with other isolates.

The key insight is that NLLB's multilingual pretraining creates sufficiently language-agnostic representations that any reasonable proxy suffices, freeing practitioners to focus on corpus development rather than proxy optimization. We achieve competitive results (BLEU 21.23) with only 7.6k training sentences, demonstrating that multilingual model adaptation is viable even for isolates with severely limited data.

Our dialect experiments additionally show that Amur and Sakhalin Nivkh exhibit near-zero cross-dialect transfer, yet a unified model trained on pooled data matches or exceeds dialect-specific models on both varieties, thus suggesting that data aggregation is preferable to dialect-specific training under low-resource conditions.

Future work should investigate whether this proxy-agnostic pattern holds for other isolates, explore larger NLLB model variants (1.3B, 3.3B), and conduct human evaluation to assess real-world translation utility for language documentation and revitalization efforts.

Ethics Statement

This work aims to support language documentation and revitalization efforts for Nivkh. Data sources include published linguistic materials, publicly available periodicals, and copyrighted literary texts. Some literary sources remain under copyright protection and were provided for research purposes by the Nogliki District Library (Sakhalin Oblast, Russia) with permission for use in this study. Due to these copyright restrictions, we cannot release the full parallel corpus publicly. However, the subset derived from public domain

sources will be made available[3] to support future research. We acknowledge that machine translation systems for endangered languages should complement rather than replace human language transmission and community-led revitalization efforts.

Acknowledgments

This research is supported by Russian Science Foundation, RSF project 25-28-00552 "Digitalization of the data of an endangered language: Nivkh", realized at Lomonosov Moscow State University. We thank Daria Savina, Eva Gogua, and Sergei Shevelev for their substantial contributions to the original corpus and continued support throughout this work; Maria Medvedeva for her meticulous manual data correction, which ensured the quality of the final dataset; and Sergei Averkiev for early-stage consultation that helped shape the direction of this research. We also thank the Nogliki District Library (Sakhalin Oblast) for providing access to copyrighted Nivkh literary materials essential for this research.

References

David Dale. 2022. The first neural machine translation system for the Erzya language. In *Proceedings of the First Workshop on NLP Applications to Field Linguistics*, pages 45–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

David Dale. 2023. How to fine-tune a NLLB-200 model for translating a new language. https://cointegrated.medium.com/how-to-fine-tune-a-nllb-200-model-for-translating-a-new-language-a37fc706b865. Accessed: 2025-12-26.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ekaterina Gruzdeva and Anna Bugaeva. 2022. Nivkh. In Martine Robbeets and Alexander Savelyev, editors, *The Oxford Guide to the Transeurasian Languages*. Oxford University Press, Oxford.

Valentin Yurievich Gusev and Ruslan Ildarovich Idrisov. 2019. Nivkh corpus. RNF project №17-18-01649. Available at: http://nivkh.web-corpora.net.

[3] https://github.com/grapaul/NivkhKurng

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerber, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. Efficient natural language response suggestion for smart reply. In *Proceedings of the First Workshop on NLP for Conversational AI*. Association for Computational Linguistics.

Andrei-Ionuţ Jerpelea, Alin Radoi, and Sergiu Nisioi. 2025. Dialectal and low resource machine translation for Aromanian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7209–7228, Abu Dhabi, UAE. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Buber, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Erukhim Abramovich Kreinovich. 1934. Nivkhskie teksty [nivkh texts]. Manuscript materials.

Erukhim Abramovich Kreinovich. 1937. *Fonetika nivkhskogo (gilyatskogo) yazyka [Phonetics of the Nivkh (Gilyak) Language]*. Izdatel'stvo AN SSSR, Moscow-Leningrad.

André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Eric W. Holman, and 1 others. 2013. ASJP world language tree of lexical similarity: Version 4. Automated Similarity Judgment Program.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Celebi, 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 Baez, Gabriel Battber, Shruti Bhosale, and 28 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Vladimir Zinov'evich Panfilov. 1962. *Grammatika nivkhskogo yazyka. Chast' 1 [Grammar of the Nivkh Language. Part 1]*. Izdatel'stvo AN SSSR, Moscow-Leningrad.

Vladimir Zinov'evich Panfilov. 1965. *Grammatika nivkhskogo yazyka. Chast' 2 [Grammar of the Nivkh Language. Part 2]*. Nauka, Moscow-Leningrad.

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. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4596–4604. PMLR.

Lev Yakovlevich Shternberg. 1908. *Materialy po izucheniyu gilyatskogo yazyka i fol'klora [Materials for the Study of the Gilyak Language and Folklore]*. Imperatorskaya Akademiya Nauk, St. Petersburg.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575. Association for Computational Linguistics.

A Nivkh Alphabet

Table 8 presents the Nivkh Cyrillic alphabet as used in the corpus. Characters unique to Nivkh or with Nivkh-specific phonetic values are shown in **bold**.

Table 8: The Nivkh Cyrillic alphabet. The apostrophe marks aspiration. Characters marked † are Sakhalin dialect only.

А а	Б б	В в	Г г	**Ԍ ԍ**†	**Ғ ғ**	**Ӷ ӷ**	Д д
Е е	Ё ё	Ж ж	З з	И и	Й й	К к	К' к'
Ҟ ҟ	Ҟ' ҟ'	Л л	М м	Н н	**Ӈ ӈ**	О о	П п
П' п'	Р р	**Ҏ ҏ**	С с	Т т	Т' т'	У у	**Ӯ ӯ**†
Ф ф	Х х	**Ҳ ҳ**	**Ӿ ӿ**	Ц ц	Ч ч	Ш ш	Щ щ
Ъ ъ	Ы ы	Ь ь	Э э	Ю ю	Я я		

B Glossing Abbreviations

3SG	third person singular
ATR	attributive
COMPL	completive
CONV	converb
IND	indicative
SIM	simultaneity

C Corpus Examples

Table 9 shows parallel sentences from the training corpus across all genres.

Table 9: Training corpus examples. **S** = Nivkh, **R** = Russian, **E** = English.

Educational

S Ныух ньрывух Москваух хоғдьра.
R Сегодня в моем доме в Москве холодно.
E 'It's cold in my house in Moscow today.'

Folklore

S Кэрӈ лырр к'ныкчаӷртох виӈан к'ныкығуин выч ғой хумдьра.
R Когда, пойдя вдоль моря, к трем мысам подойдешь, на конце мыса будет находиться железная лиственница.
E 'When you walk along the sea and reach the three capes, you will find an iron larch at the end of the cape.'

Periodical

S Оғлағу хыскла Север п'иӈгу культура поделкағу ныдьныты, ағағску ныдьныты хадьғу.
R Так, ребята учились мастерить поделки, связанные с культурой и бытом коренных малочисленных народов Севера.
E 'The children learned to make handicrafts related to the culture and way of life of the indigenous peoples of the North.'

Literary

S Ньҏаӷ, хонвайво, уткуоғлагу мен к'еӷ нонӷ пойныт тыӈздьғу.
R Случилось так, что однажды по весне два мальчугана пытались поймать маленького лисенка.
E 'It so happened that one day in the spring, two boys were trying to catch a little fox cub.'

Religious

S Урла ны ӈығар, янҏ наға тёӈҏ ғотть?
R Если ты делаешь добро, то тогда почему поник головой?
E 'If you're doing good, then why are you so down?'

Transcribed speech

S Хоғот наф хыры хоӈгут п'оғлагуах ты хоӈгут наф п'и пандь ӈалакут, ыҥркы нивғгу ӈалакут, п'иннывдь выкскуй керт.
R Поэтому, сейчас своих детей приучаю, как сама воспитывалась, по старинным обычаям нивхов, чтобы свою традиционную кухню не забывали.
E 'That's why I'm teaching my children, as I was taught, to follow the old Nivkh customs, so that they don't forget their traditional cuisine.'

D Translation Examples

Tables 10 and 11 show representative model outputs (**niv**→**rus**) from the test set. **S** = Nivkh source, **R** = reference, **M** = model output, **E** = English translation. Differences between R and M are underlined.

Table 10: Successful translations.

(a) Exact match – folklore

S Ыма, ни чхыф наныгын вииндра.
R Матушка, я медведя искать пойду.
M Матушка, я медведя искать пойду.
E ‘Mother, I will go to hunt a bear.’

(b) Exact match – periodical

S Иф трибунароҳ п'уӈан, зал оздь.
R Когда он вышел к трибуне, зал встал.
M Когда он вышел к трибуне, зал встал.
E ‘When he came to the podium, the hall rose.’

(c) Word order – educational

S Ни ёлкароҳ игрушкағу хрот ҳадь.
R Я на елку вешаю игрушки.
M Я на елку игрушки вешаю.
E ‘I hang ornaments on the tree.’

(d) Synonym: муж ‘husband’ → старик ‘old man’

S Мызн ычих муў лахчиндра.
R Наш муж совсем умер.
M Наш старик уже умер.
E ‘Our husband / old man has died.’

(e) Synonym: закурил ‘lit up’ → покурил табаку ‘smoked tobacco’

S Мыз-ниғывың прафтох йугыў, тамх таў.
R Наш Гиляк в свою юрту вошел, закурил.
M Наш Гиляк в свою юрту вошел, табаку покурил.
E ‘Our Gilyak entered his yurt and smoked.’

(f) Deixis: та ‘that’ → эта ‘this’

S Ҳуӈ чхаў-чхнай итў: Кеў ивндра.
R Та деревянная фигура сказала: Новости есть.
M Эта деревянная фигура сказала: новости есть.
E ‘That / this wooden figure said: There is news.’

Table 11: Characteristic errors. Error locus underlined.

(a) Kinship: дочь ‘daughter’ → сын ‘son’

S Ола, рунд ниғывың-ӈа шоў-пшынд.
R Дочь, что за человека ты привела.
M Сын, какой-то человек пришел.
E ‘Daughter, what person did you bring.’

(b) Kinship: сестра ‘sister’ → брат ‘brother’

S Ҳуных ехыў, пфорох пшыў, ийанғей ҳунывыў, ийасх уигиў.
R ...его жена была, его сестры не было.
M ...его жена была, его брата не оказалось.
E ‘...his wife was there, his sister was gone.’

(c) Converb chain: cooking step dropped

S Изроў йетот инифке, сик ниихарт.
R освежевав, сварив, долго ели, все съели.
M освежевав, поели, все съела.
E ‘skinned, cooked, ate long, ate all up’

(d) Root: шкуру ‘skin’ → нож ‘knife’

S Ҳуӈ чхыф пхӈа сивуў, ознкуў.
R Этот медведь свою шкуру снял, выпрямился.
M Этот медведь свой нож снял, поднял.
E ‘This bear took off his skin, straightened up.’

(e) Tense: present → past

S Т'улф ньың воух оғлағу лаӄхир вардьғу.
R ...дети на лыжах соревнуются.
M ...дети на лыжах соревновались.
E ‘In winter children compete on skis.’

(f) Deixis: оттуда ‘from there’ → отсюда ‘from here’

S Мыз-ниғывың ҳуных вифке, малхолаӈ-вонаҳўтох виў йугыў.
R ...оттуда далеко пошел, в селение пришел...
M ...отсюда далеко пошел, в селение пришел...
E ‘...went far from there, came to a village...’