Translation or Recitation? Calibrating Evaluation Scores for Machine Translation of Extremely Low-Resource Languages

Anonymous ACL submission

Abstract

The landscape of extremely low-resource machine translation (MT) is characterized by perplexing variability in reported performance, often making results across different language pairs difficult to contextualize. For researchers focused on specific language groups—such as ancient languages—it is nearly impossible to determine if breakthroughs reported in other contexts (e.g., native African or American languages) result from superior methodologies or are merely artifacts of benchmark collection. To address this problem, we introduce the **FRED Difficulty Metrics**, which include the *Fertility Ratio (F)*, *Retrieval Proxy* (R), *Pretraining Exposure* (E), and *Corpus Diversity* (D) and serve as dataset-intrinsic metrics to contextualize reported scores. These metrics reveal that a significant portion of result variability is explained by train-test overlap and pre-training exposure rather than model capability. Additionally, we identify that some languages — particularly extinct and non-Latin indigenous languages — suffer from poor tokenization coverage (high token fertility), highlighting a fundamental limitation of transferring models from high-resource languages that lack a shared vocabulary. By providing these indices alongside performance scores, we enable more transparent evaluation of cross-lingual transfer and provide a more reliable foundation for the XLR MT community.[1]

1 Introduction

Multilingual pre-trained models have significantly advanced machine translation for low-resource languages through cross-lingual transfer learning (NLLB Team et al., 2024). However, performance across extremely low-resource settings exhibits a staggering degree of variation. Recent studies (Haddow et al., 2022) show that while some African

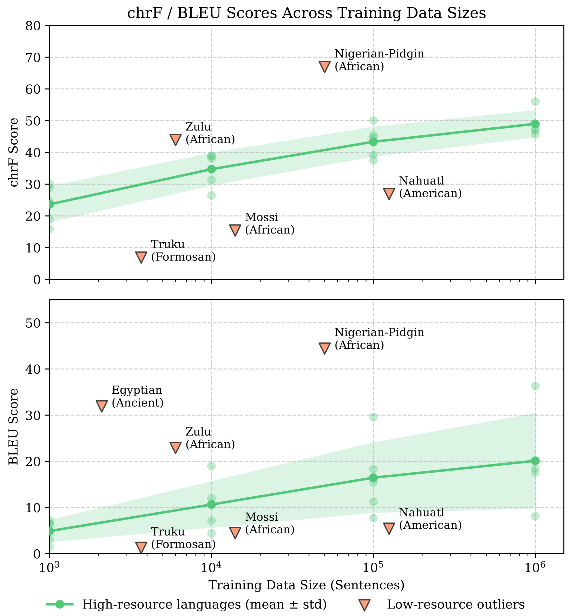

Figure 1: Performance distribution of extremely low-resource (XLR) languages in machine translation, shown as scattered outliers. The green line represents the mean performance of mBART when fine-tuned on varying sizes of data from five high-resource languages (see Table 5 and §2.1), with the shaded region indicating ± 1 standard deviation.

and ancient languages can achieve BLEU scores exceeding 40, certain American or South Asian indigenous languages struggle to reach 5 BLEU under similar settings, i.e., translating to a high-resource language with a similar amount of training data.

This disparity creates a significant barrier for the MT community (Silva et al., 2024). Without a standardized way to provide context for these results, it is difficult to interpret whether a high BLEU score indicates an effective model or a benchmark with unnaturally low complexity. The issue goes beyond metric choices: while BLEU is not suitable enough for morphologically rich languages, we identify a

[1]Code: https://anonymous.4open.science/r/analysis-loresMT-209C

similar trend with ChrF. These observations raise a fundamental question: *Is the variability in MT performance due to inherent linguistic properties, or is it an artifact of how benchmark datasets were collected?*

To answer this question, we consider a specific category: **Extremely Low-Resource (XLR)** languages. We characterize XLR languages as those that exist in the "blind spot" of current multilingual models—lacking both the monolingual data required for effective pre-training and the lexical (subword) overlap necessary for efficient transfer learning (Haddow et al., 2022). To establish a performance baseline, we cap high-resource language datasets to extremely low-resource sizes (ranging from 10^3 to 10^6 sentences). As illustrated in Figure 1, we plot the performance of mBART (Tang et al., 2020) on five typologically diverse high-resource languages as a reference region.

We find that the performances of many XLR languages act as significant outliers: some vastly overperform relative to their size, while others achieve scores much lower than the baseline. We hypothesize that train-test data relationships, not just training size or architecture, drive these discrepancies. To verify the hypothesis, we propose **Difficulty Metrics**—Pre-training Exposure (E), Retrieval Proxy (R) and Fertility Ratio (F)—to contextualize MT benchmarks and identify when high scores correlate with low diversity or contamination.

Our contributions are as follows:

- We provide a systematic analysis of data quality factors driving result variability in **XLR MT**, advancing understanding of when and why transfer learning from high-resource languages fails.
- We establish **high-resource reference baselines** by constraining data to XLR sizes (10^3-10^6 sentences), providing controlled performance references.
- We introduce the **FRED Difficulty Metrics** that quantify task complexity independent of model performance, improving transparency of reported gains.
- We identify that underperforming outliers suffer from poor lexicon overlap with pre-trained models, highlighting structural limits of transfer learning for non-digitized languages.

Metric	Notation	Interpretation
Fertility Ratio	$F = N_{\text{token}}/N_{\text{char}}$	Tokenization efficiency; **higher = harder**
Retrieval Proxy	R	Upper bound via memorization; **lower = harder**
Pre-training Exposure	E	Overlap with pre-training corpus; **lower = harder**
Corpus Diversity	D	Train-test lexical similarity; **lower = harder**

Table 1: Summary of the four core metrics (F, R, E,D) quantify intrinsic task difficulty.

2 Methods

XLR languages exhibit extreme performance variability (BLEU 5–40+) without interpretable context, making it impossible to distinguish genuine translation capability from benchmark artifacts. We address this through two complementary approaches: (1) establishing controlled reference baselines, and (2) introducing difficulty metrics that quantify dataset characteristics.

2.1 Establishing High-Resource Reference Baselines

A key challenge in evaluating XLR performance is the lack of controlled baselines. To address this, we establish reference baselines by artificially restricting high-resource MT systems to extremely low-resource sizes (ranging from 10^3 to 10^6 sentences), providing a "gold standard" for expected behavior under data-constrained conditions.

We select five typologically diverse languages—Finnish (fi), Chinese (zh), Arabic (ar), Japanese (ja), and Hindi (hi)—with details shown in Appendix Table 5. These languages represent distinct language families and writing systems (Alphabetic, Logographic, Abjad, Moraic, and Abugida orthographies), ensuring the baselines account for different morphological and orthographic challenges independent of data availability.

2.2 FRED Difficulty Metric Definition

The performance of MT systems is heavily influenced by the underlying data distribution and benchmark characteristics. To move beyond raw performance scores, we propose four **Difficulty Metrics** that quantify dataset complexity (Table 1).

We describe all metrics below, and refer to appendix E for detailed implementation. By design, these metrics are computationally efficient and non-parametric, with the computational overhead reported in Appendix G.

Token Fertility (F), defined as the ratio of tokens to characters ($N_{\text{token}}/N_{\text{char}}$), which captures the efficiency of tokenization. We calculate the F-score on both the source and target sides and report the larger one. Note that this is not a commonly seen fertility metric in the literature, but we find it is a good indicator of tokenization quality.

Retrieval Proxy (R). The R score simulates the performance of a perfect retrieval-based system, establishing a ceiling on what can be achieved through memorization alone. For each test sentence, we identify its nearest neighbor in the training set and measure target similarity:

$$R(f) = \frac{1}{M} \sum_{i \in D_{\text{te}}} f(y_i, y_{j^*}), \tag{1}$$

where $j^* = \arg\max_{j:(x_j, y_j) \in D_{\text{tr}}} f(x_i, x_j)$, and f denotes a base metric such as BLEU. Higher R scores indicate that simple nearest-neighbor retrieval without cross-lingual understanding can achieve competitive results, signaling low inherent task difficulty.

Pre-training Exposure (E), which quantifies overlap between evaluation data and the model's pre-training corpus. Let G_{te} be the set of unique n-grams in the test set target sentences, and $\text{count}(g, D_{\text{pt}})$ be the frequency of 4-gram[2] g within the pre-training corpus D_{pt}, calculated using *infini-gram* (Liu et al., 2025):

$$E = \frac{1}{|G_{\text{te}}|} \sum_{g \in G_{\text{te}}} \text{count}(g, D_{\text{pt}}). \tag{2}$$

A high E score indicates that test data contains phrases frequently seen during pre-training, suggesting the model may rely on memorization rather than cross-lingual transfer.

Corpus Diversity (D). This metric evaluates lexical diversity by measuring similarity between training and test instances, similar to self-BLEU (Zhu et al., 2018). We compute the average pairwise similarity between all training and test instances:

$$D(f) = \frac{1}{NM} \sum_{i \in D_{\text{te}}} \sum_{j \in D_{\text{tr}}} f(y_i, y_j) \tag{3}$$

A high D score indicates similar vocabulary and phrasal patterns between training and test sets (low lexical diversity), which is a common phenomenon in domain-restricted corpora.

3 Experiments and Analysis

3.1 Dataset Collection and Analysis

We surveyed low-resource workshops and papers at *ACL conferences over the past three years. As

[2] any n-gram size can be used, 4 is out of rule-of-thumb.

Lang	N_{train} (# sent)	N_{token}	F-score	E-Score	D-score	R-score	Model
				4-gram	BLEU	BLEU	BLEU
High-resource languages (Table 7)							
avg.	10k	30.8	0.41	96.9	1.37	3.24	16.27
ja→en	10k	31.4	0.56	114	1.65	2.86	9.03
hi→en	10k	33.1	0.28	86	0.95	2.78	14.02
fi→en	10k	27.3	0.25	85	2.03	3.21	13.46
zh→en	10k	38.0	0.66	90	0.78	2.38	13.32
ar→en	10k	24.0	0.32	109	1.45	4.98	31.54
Ancient (extinct) languages (Chen et al. (2024); De Cao et al. (2024), Table 8)							
akk→en	50k	24.6	<u>1.00</u>	82.2	1.59	32.10	44.41
egy→en/de	10k	24.0	<u>1.00</u>	0.08	3.47	23.43	34.45
Formosan languages (indigenous languages in Taiwan) (Zheng et al. (2024), Table 9)							
avg.	-	15.2	<u>0.92</u>	0.006	5.27	13.81	8.14
tao→zh	5k	11.0	<u>0.91</u>	0.08	5.34	17.80	14.74
Americas Indigenous Languages (De Gibert et al. (2025), Table 10)							
avg.	-	38.8	0.40	1.17	1.83 (11.19)	4.72 (15.06)	5.98 (26.41)
shp→es	14k	20.9	0.33	0.65	3.83 (8.69)	4.86 (14.43)	7.22 (27.33)
hch→es	9k	26.9	0.40	0.65	3.05 (11.54)	3.41 (13.65)	3.69 (23.26)
quy→es	125k	22.3	0.34	0.65	1.27 (13.77)	3.85 (12.79)	8.76 (33.83)
guc→es	59k	35.9	0.40	0.24	0.93 (13.35)	8.15 (23.89)	2.22 (12.58)
African indigenous languages (Adelani et al. (2022a), Table 11)							
avg.	-	51.7	0.39	1.43	5.20	13.2	15.1
hau→en	3k	46.1	0.29	146	1.41	6.28	12.9
zul→en	3k	49.9	0.33	99.6	1.41	32.85	31.1
bam→fr	3k	56.2	0.45	2.65	1.79	7.28	10.0
Indic indigenous Languages (Pal et al. (2023), Table 12)							
mni→en	50k	48.1	0.54	330	1.24	19.91	69.75
kha→en	24k	60.5	0.39	727	2.00	4.43	20.72

Table 2: Overview of automatic metrics to measure the data quality of different languages (translating into high-resource direction). The high-resource language are shown here for reference. E-Score is calculated on the target side. * For Americas Group, we also reported chrF++ score in parenthesis. For a complete data, refer to Appendix (Table 6).

shown in Appendix Table 4, XLR languages fall into three groups: (1) under-represented languages with substantial speakers but limited digital presence (e.g., African and Indic); (2) endangered languages with small speaker communities (e.g., Formosan and Americas indigenous); and (3) ancient (extinct) languages with fixed corpora (e.g., Ancient Egyptian and Akkadian).

For fair comparison, we report numbers without extra training data, primarily using pre-trained models. Table 2 shows metrics for translation into high-resource direction; the reverse direction is in Appendix Table 6.

3.2 Interpreting Metrics Against Baselines

High-resource baselines establish expected ranges. The high-resource group exhibits mean D-BLEU of 1.37 and R-BLEU of 3.24, indicating relatively high sentence diversity and low train-test overlap. The average E-score of 96.9 shows moderate pre-training exposure. **These values provide reference points**: XLR languages significantly deviating from these ranges likely have data quality issues rather than inherent linguistic difficulty.

Not all XLR languages can match baseline diversity. For ancient languages like Akkadian (akk) and Ancient Egyptian (egy), R-scores of 32.10 and 23.43 far exceed the baseline range (3.24). **This is not an error but a real constraint**: with fixed, limited corpora, achieving low train-test similarity is infeasible. However, **these high R-scores explain their unexpectedly high BLEU scores** (44.41 and 34.45)–the task is genuinely easier due to memorization opportunities.

3.3 Metric Correlation Analysis

Feature	R^2 Value	Pred Strength
R-Score	0.5821	Strongest
N_{token}	0.3415	Moderate
F-Score (N_{token}/N_{char})	0.2248	Low-Moderate
D-Score	0.1204	Low
E-Score	0.0142	Negligible
N_{train}	0.0011	None

Table 3: Individual R^2 for Hybrid Regression Model (BLEU/ChrF/chrF++).

We conducted a regression correlation analysis between the proposed metrics, training size, and model performance (details can be found in Appendix A). R-score emerges as the dominant predictor ($R^2 = 0.582$), explaining over 58% of performance variance. This confirms our hypothesis that **train-test relationships matter more than training size**. Additionally, we found E-Score has very little correlation compared to other factors, which suggests in XLR MT, pre-training data contamination is not the most significant issue.

3.4 Specific Findings for Language Groups

Ancient languages: High performance explained by high R-scores. Akkadian and Ancient Egyptian achieve BLEU scores exceeding 40—dramatically above the high-resource baseline (Figure 1). **Our metrics reveal the reason**: their R-scores (32.10 and 23.43) are 6-10× higher than the baseline (3.24), indicating that even perfect retrieval would achieve strong performance. These languages represent the easiest XLR translation tasks due to limited corpus diversity.

Token fertility reveals structural limits. Formosan and ancient languages show fertility approaching 1.0, meaning nearly every character becomes a separate token, **indicating a tokenization failure**: the pre-trained model lacks subword units for efficient representation. Compared to Americas languages with better tokenization (fertility 0.40), these languages show marginal improvement over R-score baselines. Formosan neural models average BLEU 8.14, far below the R score of 13.81, suggesting that these models cannot learn from poorly represented data.

Neural models should outperform retrieval baselines. The R score represents a trivial lower bound, i.e., what nearest-neighbor matching achieves without any learning. **Yet many XLR MT systems fail to surpass it**. For example, Zulu's R-score achieves BLEU 32.85, exceeding the best neural model (21.2). Similarly, some Americas languages (e.g., guc) peak at ChrF++ 12.58, well below their R-score (23.89). This suggests systematic underfitting or poor hyperparameter tuning (Sennrich and Zhang, 2019), as effective and meaningful neural models should surpass simple retrieval.

Pre-training gaps explain Americas underperformance. Despite comparable D scores and R scores to high-resource baselines, Americas languages achieve lower ChrF. Their E scores (0.24-0.65) are 100× lower than high-resource languages (86-114). Without pre-training coverage, these languages rely on cross-lingual transfer from limited parallel data, validating prior observations about monolingual data importance (Mager et al., 2023).

4 Conclusion

We investigated variability in XLR machine translation performance, establishing that much stems from dataset characteristics rather than linguistic properties or model capabilities. Our analysis reveals clear patterns: overperforming outliers benefit from high train-test similarity or pre-training exposure, while underperforming outliers suffer from poor tokenization and minimal pre-training representation. These findings highlight fundamental limitations: pre-trained multilingual models cannot effectively transfer to languages outside their representation space.

We strongly encourage future XLR MT research to report FRED Difficulty Metrics alongside standard metrics, enabling reliable cross-study comparisons and helping distinguish genuine methodological advances from benchmark artifacts.

Limitations

While BLEU or ChrF (and their variants) scores provide a common evaluation metric, their comparability across different languages remains challenging. The automatic metrics we propose are a step toward better evaluation of extremely low resource languages, but there is room for improvement, particularly in the measurement of lexicon overlap.

The performance of mBART on the five high-resource languages, though informative, could be enhanced with more fine-grained approximations. Future work could benefit from more detailed analyses to better distinguish outliers in XLR languages in machine translation.

We wish we have more time to investigate the correlation between the monolingual data and the performance of the models, which we could also apply similar E-score to the monolingual data.

Ethics Statement

This work highlights the challenges faced by extremely low-resource languages, which we define as those with fewer than 1M training examples and without additional unlabeled resources. By emphasizing this definition, we aim to underscore the need for more effective cross-lingual transfer learning approaches that can operate in data-scarce scenarios.

We have made efforts to ensure that the data we examined in this paper represents languages from diverse regions around the world, promoting inclusivity and comprehensiveness in our analysis.

Acknowledgements

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 2022a. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022b. A few thousand translations go a long way! leveraging pre-trained models for african news translation.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Krikpatrick. 2024. Logogramnlp: Comparing visual and textual representations of ancient logographic writing systems for nlp. *ACL*.

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: a hieroglyphic transformer for translating Ancient Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 71–86, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina Von Der Wense, and Manuel Mager. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Gideon George, Olubayo Adekanmbi, and Anthony Soronnadi. 2024. Tangalenlp: Building po tangle to english parallel corpora and machine translation of the tangle (tangale) language. In *5th Workshop on African Natural Language Processing*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based neural unsupervised machine translation.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2025. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the Americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133, Toronto, Canada. Association for Computational Linguistics.

Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondřej Bojar, Akiko Eriguchi, Yusuke Oda, Chenhui Chu, and Sadao Kurohashi. 2023. Overview of the 10th workshop on Asian translation. In *Proceedings of the 10th Workshop on Asian Translation*, pages 1–28, Macau SAR, China. Asia-Pacific Association for Machine Translation.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, editors. 2023. *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*. Association for Computational Linguistics, Dubrovnik, Croatia.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Jonne Sälevä and Constantine Lignos. 2024. Language model priors and data augmentation strategies for low-resource machine translation: A case study using Finnish to Northern Sámi. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12949–12956, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221.

Ana Silva, Nikit Srivastava, Tatiana Moteu Ngoli, Michael Röder, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2024. Benchmarking low-resource machine translation systems. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 175–185, Bangkok, Thailand. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Elena Voita, Rico Senrich, Fedor Ratnikov, and Standa Kuřík. 2019. Moseskit: Train moses phrase-based machine translation without the pain of configuring it. `https://github.com/yandexdataschool/moseskit`.

Mahshar Yahan and Dr. Mohammad Islam. 2025. Leveraging large language models for Spanish-indigenous language machine translation at AmericasNLP 2025. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 126–133, Albuquerque, New Mexico. Association for Computational Linguistics.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11248–11259, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zebiao Zhou, Hui Li, Xiangxun Zhu, and Kangzhen Liu. 2025. TranssionMT's submission to the Indic MT shared task in WMT 2025. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1271–1275, Suzhou, China. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

A Regression Correlation Analysis

We conducted a regression correlation analysis on the X and Y for language pairs listed in Table 2.

X: N_{train}, N_{token}, F-score, R-score, E-score, D-score

Y: neural models' performance in BLEU (except for native Americas languages, we use ChrF++ instead of BLEU, the RED scores are)

B Tables

B.1 Low resource Machine Translation survey

Table 4 shows our survey of venues and publications on XLR languages from the past three years..

B.2 High resource MT statistics

Table 5 shows the High-resource reference baselines (xx→en). BLEU / ChrF scores with capped training data (number of training sentences). Languages are selected from varied language families and writing systems to represent diverse morphological and orthographic challenges.

C Full Overview of Automatic Metrics

Table 1 shows the overview of automatic metrics to measure the data quality of different languages from both low-res to hi-res and hi-res to low-res directions.

D Tables in details for different MT benchmark

D.1 Metrics

For results of Table 7, 8, 9, 10, 11 and 12, we report N_{train} by directly counting number of training line pairs in the datasets. D,R scores are calculated on translating into high-resource direction. E scores are calculated by counting on the high-resource side of the language pairs.

E Implementation Details of Metrics

E.1 Implementation details on Pretraining Exposure (E)

In our implementation, we utilize the **infini-gram** engine (Liu et al., 2025) to index the pretraining dataset for fast 4-gram count retrieval. The implementation details are as follows:

1. **Dataset retrieval:** We retrieved a subset of public bitext which is a part of the pretraining dataset of NLLB from the official NLLB GitHub repository[3].
2. **Indexing:** We utilized infini-gram engine to index all the bitext retrieved, using mBART-50 tokenizer (Tang et al., 2020) for gram separation.
3. **Counting:** For a target test corpus, we tokenized all sentences in target language in the test corpus using the same mBART-50 tokenizer, splited into all possible 4-grams, and used indexed infini-gram engine to retrieve the count of each 4-gram in the pretraining dataset.
4. **Report:** Take the mean value for all possible 4-grams and report as result.

E.2 Tokenization policy on PBSMT training and BLEU calculation

To adapt variability in nature of various languages, we define the following policies in our calculation of BLEU scores:

- All BLEU scores are **average sentence BLEU scores** computed using the `sentence_bleu` implementation from SacreBLEU(Post, 2018). To ensure consistent evaluation, we use the internal default setting of the `sentence_bleu` method with exponential smoothing except changing the tokenizer for calculating BLEU depends on the nature of different languages.
- For most space-separated languages, we utilize the built-in default `13a` tokenizer adapted by sacreBLEU, which is also the WMT standard tokenizer and suitable for space-separated languages.
- For Chinese languages in both high-resource languages group and Formosan Mandarin group, we utilize the built-in `zh` tokenizer adapted by sacreBLEU, which does character-wise separation on Chinese characters but preserve word structures on other space-separated languages.
- For Japanese in high-resource languages, we use built-in `ja-mecab` tokenizer adapted by sacreBLEU.

[3] https://github.com/facebookresearch/fairseq/blob/nllb/examples/nllb/data/download_parallel_corpora.py

Venue	Language	Region/Period	Reference
WMT	Multiple	Global	Kocmi et al. (2023)
AfricaNLP	Tangale	African	George et al. (2024)
AmericasNLP	Multiple	Latin America	Ebrahimi et al. (2024)
WAT/WMT	Indic	South Asia	Nakazawa et al. (2023); Pal et al. (2023); Zhou et al. (2025)
ML4AL	Ancient Egyptian	Ancient	De Cao et al. (2024)
*CL Conf	Akkadian	Ancient	Chen et al. (2024)
LoResMT	Cantonese	East Asia	Ojha et al. (2023)
*CL Conf	Formosan	East Asia	Zheng et al. (2024)
*CL Conf	Northern Sámi	Europe	Sälevä and Lignos (2024)

Table 4: Survey of venues and publications on XLR languages from the past three years.

Lang	Writing	Phonography	Dataset	BLEU / ChrF at Different Data Sizes			
				1k	10k	100k	1M
fi	Latin	Alphabetic	wmt18-fi-en	6.32 / 28.87	11.97 / 38.82	18.31 / 45.39	18.47 / 46.98
zh	Han	Logographic	wmt18-zh-en	6.51 / 30.05	10.77 / 38.10	15.42 / 44.67	17.59 / 47.34
ar	Arabic	Abjad	iwslt2017-ar-en	7.20 / 24.49	18.94 / 38.90	29.61 / 50.10	36.31 / 56.03
ja	Han/Kana	Moraic	iwslt2017-ja-en	3.08 / 18.97	7.17 / 31.37	11.24 / 37.49	–
hi	Devanagari	Abugida	IITB-hi-en	1.42 / 15.88	4.40 / 26.47	7.71 / 39.17	8.11 / 45.66
		avg.		4.91 / 23.65	10.65 / 34.73	16.46 / 43.36	20.12 / 49.01
		std.		2.25 / 5.50	4.93 / 5.00	7.50 / 4.54	10.19 / 4.11

Table 5: High-resource reference baselines (xx→en). BLEU / ChrF scores with capped training data (number of training sentences). Languages are selected from varied language families and writing systems to represent diverse morphological and orthographic challenges.

- For Akkadian ancient language, we utilize the built-in `char` tokenizer adapted by sacreBLEU which does character-wise separation.

We also defined the tokenization policy on training Phrase-based translation systems:

- We used a Docker-powered open-source project called `Moseskit` (Voita et al., 2019) for running Moses PBSMT (Lample et al., 2018). We keep the tunable configurations default to ensure consistent evaluation, including using 5-gram language model and default tokenizer that does space-separation. To maintain evaluation consistency, we use `13a` tokenizer in SacreBLEU to calculate average sentence BLEU between PBSMT prediction and test dataset ground truths.

- For non-space-separating languages like Chinese, we use MBart tokenizer(Tang et al., 2020) to separate words for PBSMT training. We have conducted experiments to test the PBSMT performances on whether we use MBart in non-space-separating language side only, or both source and target languages in Formosan language corpus, table 13 shows that PBSMT generally performs better if apply MBart in one side only. We use `zh` tokenizer for Chinese, `ja-mecab` tokenizer for Japanese, and `char` tokenizer for Akkadian in sentence BLEU evaluation for consistency.

- The default argument `kndiscount` for applying Kneser-Ney Discounting on PBSMT may fail when either the dataset length or the lexical diversity (number of unique characters or words) being too small. In this case, we switch to fallback `wbdiscount` to use Witten-Bell Discounting, and reduce the tuning set size to 50 if the training dataset is small as only a few hundred samples.

F Correlation between PBSMT and cross-BLEU

G Computational Cost

D,R Scores: These metrics run in comparable execution time since they all iterate every possible train-test pairs. For a dataset with approximately 1000 test samples and 15000 training samples, where each sample has an average of 25 tokens, the process runs for roughly 2 minutes with 32 threads, which is 0.9 CPU-hours on two AMD EPYC 7282 16-Core Processors.

PBSMT: In our experiment, the Moses PBSMT training process was configured to run on 24 CPU threads on an AMD Ryzen Threadripper 3960X 24-Core Processor. A training corpus with around 5000 parallel sentences with around 100k total tokens runs in approximately 2 hours.

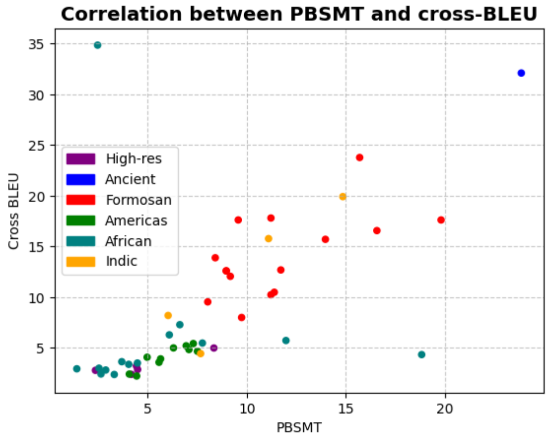

Figure 2: The correlation between PBSMT and R-BLEU of different Languages. A clear proportional relationship between PBSMT and R-BLEU can be observed with a Pearson correlation coefficient of 0.617. If 3 of the outliers from African corpus is removed, the Pearson correlation score rises to 0.913. This shows the high correlation between PBSMT and R-BLEU.

E-score: Utilizing the infinigram engine (Liu et al., 2025), the indexing process on an approximately 30GB bitext pretraining dataset costs roughly around 3 hours. The indexing process is configured to use 16 threads of two AMD EPYC 7282 16-Core Processors CPU, 32GB of memory, and 524288 open-file limit, with the custom MBart tokenizer with unsigned 32-bit integer token datatype. The final index folder occupies around 60GB of storage space. Calculating E-score for a 3000 lines test dataset with roughly 15000 unique 4 grams cost around 30 seconds.

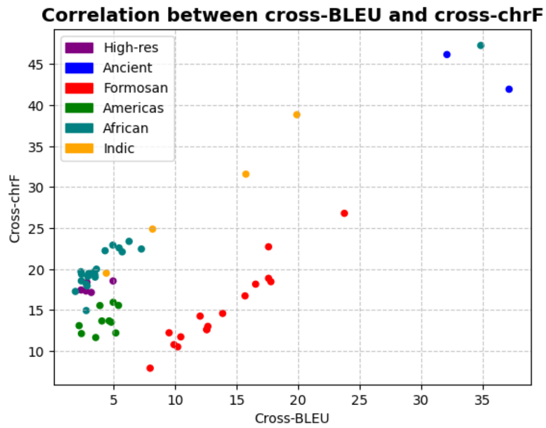

Figure 3: The correlation between R-BLEU and R-chrF of different corpus. A clear proportional relationship between R-BLEU and R-chrF can be observed with a Pearson correlation coefficient of 0.684. Removing all the data from Formosan gives a Pearson correlation coefficient of 0.922, Formosan corpus itself gives a coefficient of 0.969. This shows the high correlation between R-BLEU and R-chrF.

Lang	Parallel (# sent)	Monolingual	N_{token}	N_{char}	$\frac{N_{\text{token}}}{N_{\text{char}}}$	D-score	R-score	PBSMT	E-Score	Model
						BLEU / ChrF	BLEU / ChrF	BLEU / ChrF	4-gram	BLEU / ChrF
High-resource languages (Table 7)										
ja→en	10k	>10M	31.4	56.3	0.56	1.65 / 5.01	2.86 / 18.38	4.49 / 21.52	114.47	9.03 / 31.37
hi→en	10k	>10M	33.1	117.2	0.28	0.95 / 10.08	2.78 / 17.34	2.36 / 11.20	85.98	14.02 / 26.47
fi→en	10k	>10M	27.3	109.8	0.25	2.03 / 16.44	3.21 / 17.16	4.63 / 24.01	85.46	13.46 / 38.82
zh→en	10k	>10M	38.0	59.0	0.66	0.78 / 1.60	2.38 / 17.46	4.18 / 24.87	90.17	13.32 / 38.10
ar→en	10k	>10M	24.0	75.7	0.32	1.45 / 11.36	4.98 / 18.56	8.34 / 22.71	108.57	31.54 / 38.90
en→ja	10k	>10M	31.0	129.9	0.24	1.75 / 16.84	2.88 / 6.46	5.62 / 9.46	0.08	-
en→hi	10k	>10M	27.6	117.6	0.24	1.39 / 12.60	2.28 / 14.92	2.11 / 6.95	20.7	-
en→fi	10k	>10M	25.8	108.9	0.24	1.66 / 15.63	3.74 / 18.06	4.47 / 27.10	0.27	-
en→zh	10k	>10M	37.2	156.5	0.24	1.30 / 15.25	1.95 / 3.09	8.89 / 9.82	0.56	-
en→ar	10k	>10M	22.8	94.0	0.24	2.01 / 13.64	3.58 / 15.09	7.34 / 23.22	0.25	-
Ancient (extinct) languages (Chen et al. (2024); De Cao et al. (2024), Table 8)										
akk→en	50k	0	24.6	24.6	1.00	1.59 / 2.61	32.10 / 46.18	23.86 / 43.25	82.21	44.41
egy→en/de	10k	0	24.0	24.0	1.00	3.47 / 9.22	23.43 / 41.95	3.39 / 13.86	0.08	34.45
en→akk	50k	0	26.2	86.9	0.30	1.60 / 8.50	29.79 / 32.37	31.81 / 33.08	0	-
en/de→egy	10k	0	18.8	63.6	0.30	2.15 / 11.92	31.08 / 41.64	6.69 / 13.83	1.12	-
Formosan languages (indigenous languages in Taiwan) (Zheng et al. (2024), Table 9)										
tao→zh	5k	0	11.0	30.8	0.36	5.34 / 31.06	17.80 / 18.47	11.22 / 9.19	0.08	4.72
zh→tao	5k	0	10.6	11.7	0.91	3.41 / 2.45	19.22 / 34.84	11.73 / 17.86	0.0004	20.32
Americas Indigenous Languages (De Gibert et al. (2025), Yahan and Islam (2025), Table 10)										
shp→es	14k	0	20.9	64.4	0.33	3.98 / 8.69	5.42 / 14.43	7.30 / 21.11	0.65	7.22 / 27.33
hch→es	9k	0	26.9	66.9	0.40	3.18 / 11.54	3.91 / 13.65	5.65 / 18.20	0.65	3.69 / 23.26
quy→es	125k	-	22.3	65.7	0.34	2.60 / 13.77	4.66 / 12.79	7.51 / 20.60	0.65	8.76 / 33.83
guc→es	59k	-	35.9	35.9	0.40	0.93 / 13.35	8.15 / 23.89	9.42 / 23.89	0.24	2.22 / 12.58
es→shp	14k	0	15.1	64.5	0.23	2.77 / 9.19	5.71 / 17.98	5.95 / 22.49	0.14	1.30 / 18.12
es→hch	9k	0	15.2	64.7	0.23	1.87 / 11.08	4.72 / 19.90	6.90 / 23.66	0.37	8.66 / 28.17
es→quy	125k	-	15.1	64.5	0.23	1.94 / 13.68	3.37 / 19.05	5.40 / 29.08	1.05	2.43 / 40.01
es→guc	59k	-	18.0	76.2	0.24	1.22 / 13.62	8.36 / 31.93	5.24 / 28.19	0.15	1.11 / 17.56
African indigenous languages (Adelani et al. (2022a), Table 11)										
hau→en	3k	236k	46.1	159.8	0.29	1.41 / 18.82	6.28 / 23.39	6.09 / 29.69	146	12.9
zul→en	3k	667k	50.0	153.3	0.33	1.41 / 16.67	32.85 / 47.29	2.47 / 15.97	99.6	31.1
bam→fr	3k	-	56.2	123.7	0.45	1.79 / 14.76	7.28 / 22.46	6.62 / 27.78	2.65	10.0
en→hau	3k	236k	31.2	138.9	0.22	1.64 / 16.10	5.56 / 26.20	5.54 / 31.54	0.10	10.4
en→zul	3k	667k	33.2	136.8	0.24	1.34 / 14.72	11.55 / 34.94	2.39 / 15.26	2.81	21.2
fr→bam	3k	-	34.4	131.2	0.26	1.74 / 15.08	6.35 / 23.12	7.09 / 27.88	1.71	18.6
Indic indigenous Languages (Pal et al. (2023), Table 12)										
mni→en	50k	4M	48.2	89.8	0.54	1.24 / 13.07	19.91 / 38.84	14.85 / 42.69	330	69.75
kha→en	24k	910k	60.5	157.0	0.39	2.00 / 19.61	4.43 / 19.51	7.67 / 31.82	727	20.72
en→mni	50k	4M	19.4	89.0	0.22	1.57 / 14.87	17.80 / 37.66	4.86 / 43.61	5.79	29.50
en→kha	24k	910k	31.3	113.8	0.28	1.93 / 17.71	5.39 / 25.41	11.17 / 36.32	2.94	21.63

Table 6: Overview of automatic metrics to measure the data quality of different languages on both low-to-high and high-to-low directions. The high-resource language (ja, hi, fi, zh, ar) are shown here for reference. N_{token}, N_{char} and $\frac{N_{\text{token}}}{N_{\text{char}}}$ are calculated on the test dataset of the source side, TTR is calculated on the train dataset of the target side, and E-score is calculated on the test dataset on the target side. For Americas languages, ChrF++ scores are reported instead of ChrF in Model column.

Lang	N-train	D-chrF	R-chrF	PBSMT-chrF
Japanese (ja)	10000	5.01	18.38	21.52
Hindi (hi)	10000	10.08	17.34	11.20
Finnish (fi)	10000	16.44	17.16	24.01
Chinese (zh)	10000	1.60	17.46	24.87
Arabic (ar)	10000	11.36	18.56	22.71
Lang		**D-chrF++**	**R-chrF++**	**PBSMT-chrF++**
Japanese (ja)		3.76	16.35	19.71
Hindi (hi)		8.37	15.10	11.04
Finnish (fi)		13.66	15.26	21.38
Chinese (zh)		1.32	13.80	22.42
Arabic (ar)		9.30	17.11	22.48

Table 7: Translation Performance Metrics for high-resource language pairs – chrF as similarity function.

Lang	N-train	D-chrF	R-chrF	PBSMT-chrF
Akkadian (akk)	50000	2.61	46.18	43.25
Egyptian (egy)	10000	9.22	41.95	13.86
Lang		**D-chrF++**	**R-chrF++**	**PBSMT-chrF++**
Akkadian (akk)		2.21	45.10	41.48
Egyptian (egy)		9.02	41.34	12.12

Table 8: Translation Performance Metrics – chrF and chrF++ for Ancient language pairs.

Lang	N-train	D-BLEU	R-BLEU	PBSMT-BLEU	E-4-gram	BLEU
Sakizaya (ais)	4590	6.09	12.05	9.17	0	3.11
Amis (ami)	4600	5.51	12.59	8.97	0	3.56
Bunun (bnn)	7180	6.21	15.7	13.97	0.002	5.44
Kavalan (ckv)	6573	5.04	17.61	19.81	0	7.18
Rukai (dru)	8319	5.82	10.48	11.39	0	8.44
Paiwan (pwn)	4126	5.24	9.93	9.77	0.001	3.80
Puyuma (pyu)	5515	4.12	10.23	11.22	0.001	7.86
Seediq (sdq)	4367	3.20	9.53	8.03	0	1.52
Thao (ssf)	5952	4.53	23.77	15.70	0	10.50
Saaroa (sxr)	3839	7.87	13.88	8.41	0	6.03
Yami (tao)	5186	5.34	17.8	11.22	0.082	4.72
Atayal (tay)	4600	5.51	12.59	8.96	0	4.86
Truku (trv)	3678	6.27	7.99	9.74	0.003	1.26
Tsou (tsu)	3550	5.62	17.61	9.57	0	2.07
Kanakanavu (xnb)	5294	3.68	16.56	16.57	0	9.54
Saisiyat (xsy)	4839	4.25	12.68	11.72	0	3.99
Lang		**D-chrF**	**R-chrF**	**PBSMT-chrF**		**chrF**
Sakizaya (ais)		14.52	14.27	8.17		14.38
Amis (ami)		14.79	12.63	6.38		12.08
Bunun (bnn)		16.18	16.74	10.60		17.91
Kavalan (ckv)		15.09	18.88	16.03		24.03
Rukai (dru)		16.01	11.74	7.79		36.66
Paiwan (pwn)		15.89	10.79	9.82		4.86
Puyuma (pyu)		16.34	10.52	10.15		15.69
Seediq (sdq)		13.11	12.24	8.61		13.24
Thao (ssf)		15.27	26.81	12.87		26.66
Saaroa (sxr)		17.55	14.60	6.97		14.06
Yami (tao)		13.06	18.47	9.19		18.27
Atayal (tay)		14.79	12.63	6.39		12.26
Truku (trv)		11.82	7.92	7.95		6.87
Tsou (tsu)		13.23	22.73	9.27		19.50
Kanakanavu (xnb)		15.58	18.17	13.92		20.93
Saisiyat (xsy)		15.15	13.02	10.40		16.07
Lang		**D-chrF++**	**R-chrF++**	**PBSMT-chrF++**		
Sakizaya (ais)		13.58	12.23	10.57		
Amis (ami)		13.92	11.18	9.15		
Bunun (bnn)		15.16	15.24	14.42		
Kavalan (ckv)		13.58	17.00	20.82		
Rukai (dru)		14.32	10.26	11.36		
Paiwan (pwn)		14.53	9.95	12.82		
Puyuma (pyu)		14.39	9.59	13.81		
Seediq (sdq)		11.58	11.10	10.86		
Thao (ssf)		13.66	24.54	17.10		
Saaroa (sxr)		15.44	13.89	9.74		
Yami (tao)		12.23	17.10	12.55		
Atayal (tay)		13.92	11.18	9.15		
Truku (trv)		11.69	7.10	10.72		
Tsou (tsu)		13.04	20.48	11.48		
Kanakanavu (xnb)		13.55	17.12	18.11		
Saisiyat (xsy)		14.32	11.53	13.90		

Table 9: Translation Performance Metrics for Formosan-Chinese (Mandarin) language pairs. Exposure scores are calculated by counting on the high-resource side of the language pairs. The BLEU and chrF scores are taken from Zheng et al. (2024).

lang	N-train	D-BLEU	R-BLEU	PBSMT-BLEU	E-4-gram	BLEU
ashaninka (cni)	3883	3.98	4.94	6.36	0.62	2.35
awajun (agr)	21964	1.23	5.50	5.75	0.87	11.12
aymara (aym)	6531	1.74	3.51	5.17	0.65	8.82
bribri (bzd)	7508	2.01	4.33	6.01	0.65	4.31
chatino (ctp)	357	1.50	9.63	4.69	7.62	-
guarani (gn)	26032	1.89	4.51	6.29	0.65	8.62
nahuatl (nah)	16145	1.21	4.18	5.72	0.62	7.22
otomi (oto)	4889	0.63	2.47	3.78	0.67	1.50
quechua (quy)	125008	1.27	3.85	5.83	0.65	8.76
raramuri (tar)	14720	0.57	2.02	4.06	0.65	-
shipibo (shp)	14592	3.83	4.86	6.18	0.65	7.22
wayuu (guc)	59715	0.93	8.15	9.42	0.24	2.22
wixarika (hch)	8966	3.05	3.41	5.34	0.65	3.69

lang	D-chrF	R-chrF	PBSMT-chrF
ashaninka (cni)	16.21	12.56	16.46
awajun (agr)	16.30	21.34	20.78
aymara (aym)	15.32	13.28	15.51
bribri (bzd)	6.37	16.03	20.40
chatino (ctp)	19.90	29.41	25.63
guarani (gn)	12.61	14.43	21.00
nahuatl (nah)	13.50	14.02	16.96
otomi (oto)	8.75	12.10	14.02
quechua (quy)	17.55	14.55	23.32
raramuri (tar)	7.82	13.92	16.42
shipibo (shp)	9.83	15.64	23.58
wayuu (guc)	17.04	26.33	25.66
wixarika (hch)	13.91	15.95	20.19

lang	D-chrF++	R-chrF++	PBSMT-chrF++	chrF++
ashaninka (cni)	13.52	11.39	15.21	24.24
awajun (agr)	13.26	19.60	19.15	32.80
aymara (aym)	12.49	11.90	14.29	31.72
bribri (bzd)	6.54	13.72	18.32	26.74
chatino (ctp)	17.68	27.18	23.81	-
guarani (gn)	10.70	12.68	19.32	32.07
nahuatl (nah)	10.77	12.51	15.28	26.89
otomi (oto)	6.93	10.42	12.12	19.01
quechua (quy)	13.77	12.79	20.60	33.83
raramuri (tar)	6.19	11.56	14.65	-
shipibo (shp)	8.69	14.43	21.11	27.33
wayuu (guc)	13.35	23.89	23.89	12.58
wixarika (hch)	11.54	13.65	18.20	23.26

Table 10: Translation Performance Metrics for Americas Indigenous Languages - Spanish Language Pairs from 2025's shared tasks. We use dev set as the test set for the evaluation of this corpus since the test set of the dataset this year is not publicly released. Exposure scores are calculated by counting on the high-resource side of the language pairs. The BLEU and chrF++ scores are taken from the dev set performance of the team Syntax Squad in 2025's competition (Yahan and Islam, 2025).

Lang	N-train	D-BLEU	R-BLEU	PBSMT-BLEU	E-4-gram	BLEU
Translate into English						
Amharic (amh)	899	0.29	2.80	4.42	109.65	-
Hausa (hau)	3098	1.41	6.28	6.09	146.42	12.9
Igbo (ibo)	6998	1.14	5.47	7.76	86.69	21.0
Kinyarwanda (kin)	460	1.24	2.95	4.24	94.52	-
Luganda (lug)	4075	1.96	4.96	5.36	84.58	19.8
Luo (luo)	4262	1.50	2.99	2.54	79.44	12.1
Chichewa (nya)	483	1.15	2.40	2.83	91.70	-
Nigerian-Pidgin (pcm)	4790	1.47	4.33	18.83	96.85	44.2
Shona (sna)	556	1.27	2.87	3.30	101.49	-
Swahili (swa)	30782	1.38	5.72	11.98	91.70	29.5
Setswana (tsn)	2100	1.75	3.63	3.69	107.03	18.6
Twi (twi)	3337	1.49	2.37	3.31	98.85	9.8
Xhosa (xho)	486	1.86	3.52	4.07	107.90	-
Yoruba (yor)	6644	1.26	3.50	4.48	88.26	12.3
Zulu (zul)	3500	1.41	34.85	2.47	99.61	31.1
Translate into French						
Bambara (bam)	3013	1.79	7.28	6.62	2.65	10.0
Ghomala (bbj)	2232	1.05	2.93	1.42	2.51	2.7
Ewe (ewe)	2026	2.12	1.92	2.87	4.55	4.1
Fon (fon)	2637	1.49	2.42	2.64	3.75	4.9
Mossi (mos)	2493	1.56	2.81	2.88	3.06	1.5
Wolof (wol)	3360	1.53	3.37	4.04	3.05	7.2

Lang	O-chrF	R-chrF	PBSMT-chrF	chrF
Translate into English				
Amharic (amh)	6.53	14.95	10.34	-
Hausa (hau)	18.82	23.39	29.69	33.2
Igbo (ibo)	12.86	22.59	32.41	46.4
Kinyarwanda (kin)	19.37	19.19	25.83	-
Luganda (lug)	18.15	22.93	30.47	45.4
Luo (luo)	18.83	19.46	22.66	34.1
Chichewa (nya)	20.63	18.57	23.93	-
Nigerian-Pidgin (pcm)	14.89	22.25	55.03	66.9
Shona (sna)	19.87	17.96	23.10	-
Swahili (swa)	15.44	22.11	40.83	53.7
Setswana (tsn)	18.84	20.00	25.00	42.4
Twi (twi)	13.98	19.69	25.89	32.9
Xhosa (xho)	18.06	19.03	22.50	-
Yoruba (yor)	11.85	19.33	25.70	31.4
Zulu (zul)	16.67	47.29	15.97	43.9
Translate into French				
Bambara (bam)	14.76	22.46	27.78	31.2
Ghomala (bbj)	11.01	19.12	16.98	21.8
Ewe (ewe)	12.34	17.28	21.16	24.8
Fon (fon)	11.84	19.38	20.82	20.5
Mossi (mos)	12.29	18.18	15.21	15.4
Wolof (wol)	13.88	19.61	24.77	26.2

Lang	D-chrF++	R-chrF++	PBSMT-chrF++
Translate into English			
Amharic (amh)	5.04	13.13	9.92
Hausa (hau)	16.03	21.30	27.57
Igbo (ibo)	10.66	20.25	28.94
Kinyarwanda (kin)	16.67	19.19	23.06
Luganda (lug)	14.82	20.70	26.96
Luo (luo)	15.91	17.07	19.54
Chichewa (nya)	16.84	16.22	20.72
Nigerian-Pidgin (pcm)	12.86	19.35	50.67
Shona (sna)	15.81	15.65	19.82
Swahili (swa)	12.90	19.72	37.62
Setswana (tsn)	16.47	17.69	22.30
Twi (twi)	12.30	17.20	22.85
Xhosa (xho)	14.67	16.61	19.83
Yoruba (yor)	10.27	17.03	23.40
Zulu (zul)	13.39	46.32	13.15
Translate into French			
Bambara (bam)	13.20	20.35	25.62
Ghomala (bbj)	8.95	16.08	14.06
Ewe (ewe)	10.82	14.67	18.84
Fon (fon)	10.84	16.83	18.96
Mossi (mos)	10.67	15.93	13.64
Wolof (wol)	12.08	16.90	22.40

Table 11: Translation Performance Metrics for African language pairs. Exposure scores are calculated by counting on the high-resource side of the language pairs. BLEU and chrF scores are taken from the best possible BLEU scores from mBART/mT5/T5 of table 4 of the paper (Adelani et al., 2022b).

Lang	N-train	D-BLEU	R-BLEU	PBSMT-BLEU	E-4-gram	BLEU
Assamese (as)	50000	1.44	8.19	6.03	484.92	66.36
Mizo (lus)	50000	2.08	15.77	11.10	1141.87	33.30
Manipuri (mni)	21687	1.24	19.91	14.85	330.78	69.75
Khasi (kha)	24000	2.00	4.43	7.67	727.82	20.72
Lang		**D-chrF**	**R-chrF**	**PBSMT-chrF**		
Assamese (as)		11.59	24.89	26.30		75.88
Mizo (lus)		14.87	31.59	31.11		52.74
Manipuri (mni)		13.07	38.84	42.69		78.16
Khasi (kha)		19.61	19.51	31.82		43.34
Lang		**D-chrF++**	**R-chrF++**	**PBSMT-chrF++**		
Assamese (as)		9.47	22.86	24.39		
Mizo (lus)		13.13	30.46	29.57		
Manipuri (mni)		10.50	36.90	40.46		
Khasi (kha)		17.87	18.05	30.31		

Table 12: Translation Performance Metrics for Indic language pairs. Exposure scores are calculated by counting on the high-resource side of the language pairs. BLEU and chrF scores are taken from the best performance in WMT 2023 shared task(Pal et al., 2023).

Lang	zh only BLEU / chrF	Both side BLEU / chrF
Sakizaya (ais)	9.17 / 8.17	6.78 / 7.43
Amis (ami)	8.97 / 6.38	6.73 / 5.96
Bunun (bnn)	13.97 / 10.60	12.36 / 11.69
Kavalan (ckv)	19.81 / 16.03	16.26 / 14.79
Rukai (dru)	11.39 / 7.79	10.11 / 9.08
Paiwan (pwn)	9.77 / 9.82	9.02 / 9.82
Puyuma (pyu)	11.22 / 10.15	8.99 / 9.35
Seediq (sdq)	8.03 / 8.61	6.62 / 8.20
Thao (ssf)	15.70 / 12.87	14.75 / 13.78
Saaroa (sxr)	8.41 / 6.97	8.08 / 9.21
Yami (tao)	11.22 / 9.19	9.14 / 8.82
Atayal (tay)	8.96 / 6.39	6.69 / 5.92
Truku (trv)	9.74 / 7.95	8.73 / 8.47
Tsou (tsu)	9.57 / 9.27	7.78 / 9.63
Kanakanavu (xnb)	16.57 / 13.92	14.14 / 14.62
Saisiyat (xsy)	11.72 / 10.40	10.45 / 10.64

Table 13: PBSMT Scores for Formosan language pairs: translate to high-resource (zh) direction. The two columns shows the performance difference of using MBart as tokenizer on Chinese side only and using MBart on both sides.