Kumatigi: Quality-Driven Data Augmentation for Low-Resource Machine Translation

Cheick Tidiani Cissé
Orange Research
cheicktidiani.cisse@orange.com
*

Abstract

Neural machine translation for extremely low-resource languages faces compounding challenges: scarce parallel data, orthographic inconsistency, and absence of quality metadata for principled training. We present Kumatigi, a quality-annotated French-Bambara corpus combining systematic curation with data augmentation strategies tailored to Bambara. We provide 67k quality-scored pairs that enable targeted data filtering and address pervasive orthographic normalization issues in existing resources. Our dual-dataset generation framework strategically exploits round-trip translation, producing synthetic pairs for fluency reinforcement alongside back-translated pairs that preserve authentic vocabulary for coverage expansion. We further introduce linguistically-motivated augmentation techniques addressing Bambara's orthographic variability, improving model robustness for real-world text. Experiments with LoRA-based fine-tuning demonstrate consistent improvements across automatic metrics, with our full system achieving up to +3–4 BLEU over strong baselines. Data generation and augmentation strategies contribute +1-2 BLEU beyond high-quality parallel data alone. Human evaluation by native speakers confirms these automatic improvements align with substantial gains in translation adequacy and fluency, with our best model approaching human reference translation quality. Our methodology provides a reproducible framework applicable to other under-resourced languages facing similar data challenges.

1 Introduction

Extremely low-resource languages present unique challenges for neural machine translation (NMT): limited parallel corpora, scarce monolingual text, orthographic inconsistency, and restrictive licensing. These challenges are particularly acute for African languages, where data scarcity intersects with limited digitization, variable annotation quality, and inconsistent standardization efforts. Bambara, a morphologically rich Mande language spoken by 30-40 million people across Mali and neighboring regions, exemplifies these constraints. Despite initiatives from Masakhane (Orife et al., 2020) and institutions like AMALAN and INALCO[1], parallel data remains severely limited. Bambara additionally faces multiple scripts (Latin, Adjami, N'ko), special characters (e.g. ɛ, ɔ, ŋ) frequently replaced by ASCII approximations, and restrictive content licensing (Costa-Jussà et al., 2022) that limits resource distribution.

Recent work by (Tapo et al., 2025) contributed the first substantial parallel corpus (47k French-Bambara pairs), yet critical gaps remain: existing resources lack quality metadata forcing practitioners to treat all samples equally, systematic exploitation of abundant monolingual data through principled augmentation remains underexplored, and evaluation focuses generally on one, in-domain dataset.

We address these challenges through Kumatigi, a systematic methodology for corpus development in extremely low-resource settings. Our work answers the following research questions:

- RQ1 How can we create quality-annotated corpora that enable principled data selection and prioritized annotation?
- RQ2 Which augmentation strategies can be adopted to effectively leverage abundant monolingual data for low-resource NMT?
- RQ3 How do we ensure robustness to orthographic variability and real-world noise?
- RQ4 Does quality improvement from data scaling manifest consistently across both automatic and human evaluation?

*This work was conducted independently outside of professional duties. Computational resources were graciously provided by Orange Research.

[1]Institut national des langues et civilisations orientales: https://www.inalco.fr/

Findings of the Association for Computational Linguistics: ACL 2026, pages 24433–24446

The contributions of the paper includes: a new quality-annotated French-Bambara corpus (67k), A quality-driven dual-dataset generation and selection framework, Orthographic-aware, linguistically-motivated augmentation tailored to Bambara's morphology and orthographic challenges, new empirical results on untagged or selectively tagged (hybrid) back-translation and a comprehensive evaluation through automated metrics and human evaluation.

2 Related Work

2.1 African Language NLP and Bambara Resources

African language NLP has gained substantial momentum through initiatives such as MasakhaNER (Adelani et al., 2021), AfriMTE (Wang et al., 2024), and AfroMT (Reid et al., 2021), yet multilingual models (Costa-Jussà et al., 2022) continue to underperform on Bambara due to limited training data. Recently, Tapo et al. (2025) contributed the first substantial parallel corpus with 47k French-Bambara pairs (36k in training), while the Bambara Reference Corpus (Vydrin et al., 2011) provides extensive monolingual texts across multiple domains. Despite progress, challenges persist: orthographic inconsistency, limited domain coverage, and unsustainable annotation pipelines (Nekoto et al., 2020). Critically, existing datasets lack quality metadata, forcing practitioners to treat all samples equally or apply crude filtering. Furthermore, systematic data augmentation and exploitation of monolingual data through back-translation remains underexplored for Bambara (Sennrich et al., 2016), despite its proven effectiveness when parallel data is scarce.

2.2 Data Augmentation for Low-Resource NMT

Bitext mining methods (Schwenk, 2018; Schwenk et al., 2021) have enabled large-scale parallel corpus construction, but they remain challenging to apply effectively to extremely low-resource languages such as Bambara. In our case, Bambara has a very limited web presence and frequent code-switching with French, which reduces the effectiveness of web-based mining. Nevertheless, seed-based mining approaches based on LASER (Language-Agnostic SEntence Representations) could still be useful, for example, to filter back-translated pairs or align available corpora. Back-translation (BT) (Sennrich et al., 2016) generates synthetic parallel pairs by translating target monolingual text back to the source language, creating pairs with authentic targets and model generated sources. While effective up to a 1:1 synthetic-to-authentic ratio (Poncelas et al., 2018), BT depends heavily on backward model quality—out-of-domain data can trigger hallucinations (Wang and Sennrich, 2020), leading to noisy source texts. BT Refinements include sampling or noising strategies (Edunov et al., 2018), quality filtering, and tagged back-translation (Caswell et al., 2019) which showed improvements. Other augmentation methods include lexical substitution (Fadaee et al., 2017) and data generation using Generative Language models (Oh et al., 2023) by paraphrasing for instance. However most of these augmentations are language-agnostic and ignore morphological properties. For morphological rich languages like Bambara, language specific augmentations (elision, synonyms etc.) could offers untapped potential as presented in this study.

2.3 Data Quality Assessment and Selection

Data quality is particularly critical in low-resource settings where each training sample disproportionately influences model behavior. Quality estimation (QE) approaches (Zerva et al., 2022) predict translation quality without reference translations, enabling filtering of noisy parallel data (Koehn et al., 2018), with cross-entropy-based filtering yielding substantially improved BLEU scores (Junczys-Dowmunt, 2018). Recent neural QE metrics fine-tune pretrained cross-lingual models to predict quality scores. COMET (Rei et al., 2020) established this paradigm, with AfriCOMET (Wang et al., 2024) extending coverage to 21 African languages but notably excluding Bambara. Alternative uncertainty-based scoring (Wang et al., 2019) enables quality-aware training and active learning strategies (Settles, 2009). While these advances provide valuable frameworks, practical methodologies for creating quality-annotated corpora remain limited for extremely low-resource languages. We address these gaps through Kumatigi and its data curation process, described below.

3 The Kumatigi Dataset

This section presents Kumatigi, our quality-annotated French-Bambara corpus, detailing data sources, licensing considerations, and creation methodologies. We distinguish between publicly

releasable data and training-only data to ensure compliance with licensing requirements while maximizing research utility.

3.1 Data Collection Strategy and Licensing

Our corpus development followed a two-tier approach based on licensing constraints:

Tier 1: Publicly Releasable Data includes materials with permissive licenses that allow redistribution. This tier forms the core of our public release.

Tier 2: Training-Only Data consists of materials under restrictive licenses. These are used exclusively for model training and language-pair adaptation but are not redistributed.

3.1.1 Foundation: The Bayelemabaga Corpus

Our work builds upon Bayelemabaga (Tapo et al., 2025), the largest prior French-Bambara corpus with 47k sentence pairs (train, dev, test). While Bayelemabaga represented a significant milestone, our analysis revealed critical limitations: (1) frequent orthographic errors in special character representation (ɛ→e, ɔ→o, ɲ→ny,), and (2) absence of quality metadata making it difficult to distinguish high-quality pairs from noisy examples, with some misalignment likely due to automatic extraction.

3.1.2 Corpus Correction and Normalization

Bambara uses special characters essential for lexical distinction—e.g., *fɛɛrɛ* (strategy) vs. *feere* (sell); *nɔgɔ* (dirt) vs. *nogo* (bowel). We applied systematic correction using: (1) automated pattern-based rules for high-confidence cases, and (2) manual verification for ambiguous contexts where native speaker judgment was required. Moreover, we manually refined approximately 1k pairs with semantic issues.

For the test set, we added 245 alternative Bambara translations where multiple valid equivalents exist, creating the **Kumatigi test set**—an enhanced version of the original Bayelemabaga test set with improved coverage of translation variation. The corrected Bayelemabaga corpus forms part of the Kumatigi dataset and will be released under CC BY-SA, consistent with the original corpus license.

3.1.3 Publicly Releasable Data Sources

Project Gutenberg Translations: We leveraged French literary texts from Project Gutenberg, which are in the public domain and free of copyright restrictions. Using a domain-adapted NLLB-200-600M model, we translated approximately 280k French sentences to Bambara, forming partially synthetic data, each receiving a quality score as described in Section 3.2.4. A subset of 8k pairs underwent manual quality assessment by native speakers. This forms the primary component of our public release, providing researchers with high-quality, legally distributable parallel data spanning literary and narrative domains.

Contemporary Community Content: We collected approximately 1,250 pairs from daily observations of written and spoken Bambara, focusing on culturally significant expressions, contemporary vocabulary, and challenging translation cases. Emphasis was placed on sentences with cultural references or emerging terminology absent from existing corpora. Additionally, we generated 1,640 pairs by extracting entries from a French-Bambara dictionary and creating contextually appropriate example sentences, ensuring coverage of core vocabulary with natural usage patterns.

Historical Documents: We digitized Bambara manuscripts, newspapers, and educational materials from the Bambara Reference Corpus[2] using a Mistral-based OCR model. Raw extraction yielded over 110k monolingual sentences with 105k kept after filtering (duplicates, short sentences, etc). These texts (1970–2020) contain culturally specific vocabulary, idiomatic expressions, and proverbs that represent authentic linguistic patterns increasingly rare in contemporary usage, making them valuable but challenging to translate even for native speakers. OCR outputs exhibited systematic character recognition errors, particularly for special characters (e.g. ɛ→c, ɔ→o etc). We distinguished between explicit and implicit errors during post-processing. Explicit errors correspond to systematic, predictable OCR mistakes or ASCII approximations caused by keyboard limitations, which follow consistent patterns and could therefore be corrected globally in high-confidence cases. Implicit errors are context-dependent ambiguities, where a form may reflect either a legitimate orthographic variant or an OCR mistake. These cases required manual verification by native speakers to decide whether the variation should be retained as

[2]Part of the Corpora Mandeica initiative: http://cormand.huma-num.fr/

a dialectal/orthographic form or corrected. Overall, the extracted texts underwent extensive and time-consuming manual post-processing, substantially reducing OCR errors.

3.1.4 Training-Only Data Sources

Web-Mined Content: We collected approximately 12k Bambara monolingual sentences from Bambara-language websites, government publications (Mali), and news sources. Due to copyright restrictions or lack of explicit authorization for redistribution, these sentences are used only for model training and language-pair adaptation. We provide source URLs and collection timestamps as metadata to enable others to negotiate access or reconstruct similar datasets independently.

NLLB Dataset Integration: We incorporated French-Bambara pairs from the NLLB dataset (Costa-Jussà et al., 2022), which permits usage for research and training but not redistribution. The corpus exhibited substantial noise including inconsistent orthography, alignment errors, and mining artifacts, necessitating systematic manual review. Approximately 4k sentences have been human-annotated with quality scores. These are used exclusively for training our models; researchers should access NLLB data directly from Meta AI's official release.

Through these efforts, and beyond the normalized and corrected Bayelemabaga foundation, we added approximately **21k entirely human annotated/reviewed** new sentence pairs and more than **360k synthetic/back-translated** pairs.

3.2 Strategic Data Augmentation via Knowledge Extraction

Our data collection yielded approximately 67k high-quality parallel pairs alongside over 360k monolingual sentences. This imbalance, abundant monolingual data but scarce parallel supervision, is characteristic of low-resource scenarios and presents both a challenge and an opportunity for principled corpus expansion.

3.2.1 Motivation and Framework

Our framework leverages model confidence, failures as informative signals to identify knowledge gaps and optimize annotation effort. We systematically exploit these signals through three stages: (1) language-pair adaptation to create a knowledge-bearing base model, (2) dual-dataset generation via round-trip translation, and (3) quality-based filtering and prioritization for targeted human annotation.

3.2.2 Stage 1: Language-Pair Adaptation

We fine-tune NLLB-200-600M (Costa-Jussà et al., 2022) on our 57k high-quality training pairs, creating a language-pair-adapted model that encodes lexical, syntactic, and cultural patterns specific to French-Bambara translation. This adapted model serves dual purposes: a translator for data augmentation and a knowledge extractor revealing vocabulary gaps through its translation behavior.

3.2.3 Stage 2: Dual-Dataset Generation

Given monolingual Bambara text BM_{orig}, we perform round-trip translation (RTT): forward translation produces French ($\text{BM}_{\text{orig}} \rightarrow \text{FR}_{\text{trans}}$), followed by back-translation to Bambara ($\text{FR}_{\text{trans}} \rightarrow \text{BM}_{\text{back}}$). This process yields two distinct training pair types:

Synthetic pairs $(\text{FR}_{\text{trans}}, \text{BM}_{\text{back}})$: Both sides are model-generated, representing the model's learned translation patterns and internal consistency. These pairs reinforce existing knowledge and provide fluent, coherent examples within the model's learned distribution. In many cases, they represent paraphrases of the original content, contributing to target-side fluency and robustness to lexical variation. Moreover, synthetic pairs serve as a regularization mechanism during fine-tuning on semi-supervised data, helping to mitigate catastrophic forgetting of the model's prior knowledge.

Back-translated pairs $(\text{FR}_{\text{trans}}, \text{BM}_{\text{orig}})$: The target preserves authentic monolingual text while the source is model-generated. These pairs are particularly valuable because they expose vocabulary, expressions, and patterns potentially absent from the model's learned distribution. Critically, when FR_{trans} is semantically misaligned with BM_{orig}, the divergence between BM_{back} and BM_{orig} signals where the model's knowledge is incomplete, likely where supervision is most needed.

For French monolingual sentences, we apply forward translation $\text{FR}_{\text{orig}} \rightarrow \text{BM}_{\text{trans}}$, creating pairs with confidence scores as described in Section 3.2.4. This asymmetric treatment reflects our primary goal of expanding Bambara vocabulary coverage and improving French-to-Bambara translation. More details in appendix A.

3.2.4 Stage 3: Quality Estimation and Prioritization

We assign scores to synthetic and back-translated pairs by estimating confidence scores based on pair type and model internal knowledge.

Synthetic pairs receive scores based on Monte Carlo dropout sampling (Wang et al., 2019): we generate K translations with dropout enabled and compute uncertainty metrics including mean confidence (expectation), variance, coefficient of variation, and min-max spread using token probabilities. These metrics are combined into a confidence score as follows:

$$s_{\text{conf}} = \alpha \cdot \text{expectation} + \beta \cdot \text{S} + (1-\alpha-\beta) \cdot \text{C} \quad (1)$$

where S = 1/(1 + CoV) and C = 1 - spread.

Back-translated pairs combine confidence scores ($\text{BM}_{\text{orig}} \rightarrow \text{FR}_{\text{trans}}$) with round-trip similarity comparing BM_{orig} to BM_{back} via character similarity (difflib python library). The final score blends both signals:

$$q = \alpha \cdot s_{\text{RTT}} + (1-\alpha) \cdot s_{\text{conf}} \quad (2)$$

α is set empirically to 0.6 to balances scores.
All automatic scores are scaled to $[0.25, 0.68]$ to distinguish them from human-reviewed data (≥ 0.7). This quality scoring comes with three advantages: (1) *quality filtering* to retain only pairs above threshold for training, (2) *uncertainty sampling* to prioritize low-scoring pairs for human correction where learning value is highest, and (3) *progressive refinement* as annotation resources allow.

3.3 Linguistic Data Augmentation

To improve robustness to real-world noise, typographical errors, OCR artifacts, orthographic variations, we apply data augmentation with probability p_i during training. All augmentations target the source text only, preserving clean target.

Bambara-Specific Augmentation Three common phenomena in natural Bambara text are simulated: (1) **elision** (ka a → k'a, ye a → y'a), (2) **special character substitution** (ɛ → {e, è, ë}, ɔ → {o, ò, ô}, ɲ → {ny, ni}, ŋ → {ng, n}), reflecting keyboard limitations and encoding errors and (3) **morphological variants** (mɔgɔ ↔ maa "person"), leveraging Bambara's agglutinative nature.

French-Specific Augmentation We randomly remove or corrupt accent marks (à → a, é → e, ç → c), simulating informal text, common in French digital content.

Language-Agnostic Augmentation Four mutually exclusive categories simulate general text degradation: (1) **character-level**: typos, adjacent key swaps, keyboard noise; (2) **visual confusion**: OCR-like substitutions (O↔0, l↔1, I↔1); (3) **word-level**: dropout of content words (preserving function words); (4) **formatting**: punctuation removal, case changes, whitespace variation. Additionally, case transformation is applied to both source and target with low probability after observing lower prediction quality on upper case inputs.

Example

```
Original:   Muso ka kan ka a kɛ sugu la.
          (The woman must do it at the market.)
Augmented:  Muso ka kan k'a ke sugu la
(elision: ka a→k'a, char subst: ɛ→e, no punct)
```

This augmentation strategy combines linguistic knowledge (Bambara/French-specific) with language-agnostic noise, preparing models for the orthographic variability inherent in low-resource language texts where standardization is incomplete.

3.4 Corpus Analysis

Table 1 and 2 presents comprehensive statistics comparing Kumatigi with existing French-Bambara resources. For fair comparison, statistics are computed only on high-quality, human-reviewed pairs ($q \geq 0.7$) from our corpus. The advantages of Kumatigi become even more pronounced when including the full dataset with synthetic and back-translated pairs ($q < 0.7$), which substantially expand vocabulary coverage and domain diversity.

3.4.1 Statistics and Comparison

Table 1 shows corpus comparison, including vocabularies computed using NLLB-200 tokenizer for tokenization. Kumatigi provides 57k high-quality training pairs, a 58% increase over Bayelemabaga's 36k pairs. Both Kumatigi and Bayelemabaga exhibit similar average sentence lengths (18-23 tokens), while FLORES-200+ contains substantially longer sentences (38-40 tokens), reflecting its focus on formal, professional written text. This length disparity has important implications for evaluation: longer sentences typically pose greater translation

Dataset	Train	Dev	Test	Total	FR Vocab	BM Vocab	FR Avg Len	BM Avg Len
FLORES-200+	–	997	1,012	2,009	6,075	4,809	**40.0**	**37.7**
Bayelemabaga	36,556	4,549	4,669	45,784	12,282	8,238	23.0	20.9
Kumatigi HQ	**57,748**	4,549	**4,832**	**67,129**	**15,945**	**13,098**	20.3	18.6

Table 1: Comparison of FR-BM parallel corpora. BM=Bambara, FR=French.

challenges due to increased syntactic complexity and lexical diversity.

3.4.2 Vocabulary Coverage Analysis

To assess complementarity between datasets, we compute Out-of-Vocabulary (OOV) rates measuring how much vocabulary in test datasets is unseen in reference datasets (Table 2). Both corpora were Unicode-normalized, lowercased, and tokenized at the word level. Boundary punctuation and numeric forms were normalized to reduce orthographic variation. Word-level OOV rates were computed to quantify lexical novelty introduced by the new corpus. This reveals whether the new data introduces novel vocabulary or just consolidates existing ones.

Reference	Test	BM OOV%	FR OOV%
FLORES-200+	Bayelemabaga	90.54	81.50
FLORES-200+	Kumatigi	91.40	84.04
Bayelemabaga	FLORES-200+	74.44	41.71
Bayelemabaga	Kumatigi	27.32	22.48
Kumatigi	FLORES-200+	68.36	35.16
Kumatigi	Bayelemabaga	1.04	0.04

Table 2: Out-of-Vocabulary rates between datasets.

Table 2 shows high OOV rates when using Bayelemabaga as reference and Kumatigi as test (27% BM, 22% FR). This demonstrate that our corpus introduces substantial novel vocabulary beyond orthographic normalization. Conversely, Kumatigi's low OOV rate on Bayelemabaga (1%) indicates comprehensive coverage of prior work. This analysis confirms that Kumatigi provides complementary rather than redundant content. On the other hand, one can observe high OOV rates when FLORES-200+ is used as test set for both dataset, giving an indication on domain shift. Furthermore higher OOV with Bambara confirms the morphological rich nature of Bambara.

4 Experiments

To demonstrate the corpus's utility and the effectiveness of the data generation/augmentation techniques, we conducted comprehensive experiments by comparing translation models trained on different portion of Kumatigi compared to other existing datasets.

4.1 Experimental Setup

We fine-tune NLLB-200-distilled-600M (Costa-Jussà et al., 2022), a 600M-parameter model supporting Bambara among 200 languages, balancing performance and computational efficiency. We employ LoRA (Hu et al., 2022) with rank $r = 32$ applied to all attention and feed-forward layers, enabling efficient adaptation while preserving the pretrained model's multilingual capabilities. It is worth noting that model is trained on both directions before final evaluation. We chose NLLB-200 based on both prior evidence and practical considerations. Previous comparative work (Tapo et al., 2025) showed that NLLB-200 outperforms alternatives such as mBART, mT5, and M2M-100 for Bambara MT in both zero-shot and fine-tuned settings. Its explicit support for Bambara is also important, as many multilingual models do not cover the language. In addition, training from scratch is impractical in our low-resource setting, and our preliminary experiments confirmed that NLLB-200 provides strong performance for both data synthesis and final translation. Exploring other architectures remains future work.

To ensure robust evaluation, all experiments are conducted with three independent runs using different random seeds, enabling statistical significance testing via paired t-test. We select checkpoints based on best validation BLEU and report mean scores with standard deviation. Table 3 summarizes our training configuration. The computations have been conducted on one L40S (48GB) GPU.

4.1.1 Evaluation framework

The evaluation framework is divided into 2 parts: automated metrics-based evaluation and human evaluation. To account for in-domain, out-domain, different data distribution, we used 3 test sets: The Kumatigi test set which is in domain, the FLORES-200+ test set an Out-of domain datasets used for

Hyperparameter	Value
Learning rate	3e-5 (linear schedule)
Warmup steps	10% of total steps
Batch size	32 (4 gradient accum.)
Optimizer	AdamW (β_1=0.9, β_2=0.98)
Weight decay	1e-3
Gradient clipping	1.0
Max epochs	10 (early stopping patience=2)
Checkpoint selection	Best validation BLEU

Table 3: Training hyperparameters for all experiments.

benchmarks and the Bayelemabaga test set used in prior work. As in existing work, we employ BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), TER (Snover et al., 2006) as automatic metrics to establish quantitative results. BLEU is a precision-oriented metric that measures n-gram overlap (n=1..4). Despite known limitations, BLEU remains standard for comparability. chrF++ (Character-level F-score) incorporates character n-grams and word n-grams. Likely more robust to morphological variation than BLEU, particularly relevant for Bambara's agglutinative properties. TER (Translation Edit Rate) measures the minimum edits (insertions, deletions, substitutions, shifts) to transform hypothesis into reference. Table 4 summarizes the evaluation framework.

Test Set	Size	Purpose
Kumatigi	4,832	In-domain held-out
FLORES-200+	1,012	Out-of-domain benchmark
Bayelemabaga	4,669	Prior work comparison

Table 4: Evaluation test sets and metrics.

4.2 Results

This section is divided into two parts: global performance with baselines and human evaluation.

4.2.1 Overall Translation Performance

Results in Table 5 demonstrate consistent improvements across all metrics and test sets when training on Kumatigi's pairs. More specifically, one can observe gains of up to +3–4 BLEU over models trained on the Bayelemabaga corpus. Incorporating synthetic and back-translated pairs yields an additional +2 BLEU, confirming that strategic exploitation of monolingual data can boost translation performance at relatively low cost. The impact of linguistic data augmentation is mostly visible on the Bambara-to-French direction. Indeed, it comes with +1 BLEU improvements on Kumatigi, Bayelemabaga test set. However, this is not the case for FLORES-200+ test set, suggesting less variations. In real-world deployment scenarios, where inputs exhibit substantial character inconsistencies, variability and transcription noise, we observe qualitatively better translation robustness.

On the other hand, metrics on FLORES-200+ test set are notably lower across all configurations. As discussed in Section 3.4.2, this dataset is out-of-domain with higher OOV rates and longer sentences, potentially explaining the reduced scores. Nevertheless, even on this challenging test set, performance has increased significantly compared to baselines. Another important point is that metrics are better (+2 BLEU) on Kumatigi's test set when compared to Bayelemabaga's one, even on zero-shot. This is mainly due to orthographic correction and normalization applied on Bayelemabaga test set to produce Kumatigi test set, revealing the impact of such process.

4.2.2 Human Evaluation

To validate that quantitative improvements reflect genuine translation quality gains, three native Bambara speakers independently evaluated 120 translations stratified by test set (60 from Kumatigi, 60 from FLORES-200+) and translation divergence. We used quartile-based sampling (25% from each divergence quartile), ensuring coverage across varying difficulty levels. Divergence was computed using TF-IDF cosine distance between system outputs. Following Direct Assessment guidelines (Barrault et al., 2020), annotators rated adequacy (meaning preservation) and fluency (grammatical naturalness) on 3-point scales (1=poor, 2=adequate, 3=perfect), blinded to system identity. Reference translations were included as an additional system to contextualize MT performance. We compared four systems: Bayelemabaga (baseline from prior work), Kumatigi HQ (trained on human-labeled pairs only), Kumatigi SB (human-labeled plus synthetic and back-translated pairs), and the reference.

Results (Table 6) **confirm that automatic metric improvements correlate with human judgment**: Kumatigi SB achieves the highest scores across both test sets, with substantial gains over Bayelemabaga of +15% on Kumatigi (+0.34 adequacy, +0.22 fluency) and +23% on FLORES (+0.35 adequacy, +0.35 fluency). **This alignment between automatic and human evaluation validates synthetic and back-translated pairs comes with real gains**. The ranking of sys-

Test Set	System	Bambara→French			French→Bambara		
		BLEU↑	chrF++↑	TER↓	BLEU↑	chrF++↑	TER↓
Kumatigi	NLLB-200 (zero-shot)	6.69	25.27	93.27	10.90	31.27	93.24
	Bayelemabaga baseline	11.07±0.02‡	30.11±0.04‡	83.12±0.07‡	11.13±0.05†	31.23±0.06	96.32±0.22
	+ L Augmentation	12.10±0.06‡	31.62±0.05‡	80.71±0.01‡	11.15±0.16	31.53±0.09∗	92.19±0.93
	Kumatigi (57k HQ)	13.12±0.06‡	32.48±0.06‡	79.70±0.14‡	13.05±0.01‡	34.50±0.05‡	85.43±0.41‡
	+ SB Augmentation	**14.00±0.14**‡	**34.28±0.12**‡	**79.11±0.21**‡	**15.54±0.07**‡	**36.86±0.03**‡	**77.81±0.09**‡
FLORES-200+	NLLB-200 (zero-shot)	11.79	34.52	83.14	4.84	28.43	104.81
	Bayelemabaga baseline	11.17±0.08	34.25±0.04	85.64±0.35	4.57±0.06	27.48±0.19	102.31±0.59†
	+ L Augmentation	10.38±0.21	33.96±0.31	86.18±0.50	4.51±0.05	27.78±0.04	99.78±0.26‡
	Kumatigi (57k HQ)	11.77±0.11	35.90±0.03‡	84.88±0.44	6.09±0.02‡	31.59±0.06‡	94.95±0.24‡
	+ SB Augmentation	**12.25±0.12**∗	**36.60±0.01**‡	**83.37±0.42**	**6.88±0.10**‡	**33.56±0.03**‡	**90.63±0.22**‡
Bayelemabaga	NLLB-200 (zero-shot)	5.24	23.30	95.70	6.24	26.05	100.84
	Bayelemabaga baseline	8.85±0.04‡	27.54±0.04‡	86.86±0.09‡	7.45±0.05‡	26.84±0.06‡	102.04±0.23
	+ L Augmentation	10.35±0.03‡	29.72±0.01‡	83.65±0.13‡	8.68±0.12‡	28.41±0.06‡	96.84±0.80†
	Kumatigi (57k HQ)	10.98±0.06‡	30.17±0.05‡	83.11±0.17‡	9.12±0.02‡	29.65±0.05‡	92.71±0.46‡
	+ SB Augmentation	**12.04±0.21**‡	**32.02±0.20**‡	**82.04±0.32**‡	**11.01±0.04**‡	**31.82±0.03**‡	**84.28±0.10**‡

Table 5: Main results across test sets. "SB Augmentation" includes synthetic and back-translated pairs. "L Augmentation" is for linguistically-motivated augmentation. Bold indicates best performance. Statistical significance computed via paired t-test: $*$ $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$.

Test Set	System	Adequacy	Fluency
Kumatigi	Bayelemabaga	2.25±0.90	2.42±0.79
	Kumatigi HQ	2.44±0.81	2.56±0.67
	Kumatigi SB	**2.59±0.61**	**2.63±0.60**
	Reference	2.34±0.90	2.44±0.89
FLORES-200+	Bayelemabaga	1.51±0.78	1.59±0.83
	Kumatigi HQ	1.68±0.77	1.81±0.76
	Kumatigi SB	**1.86±0.72**	**1.94±0.75**
	Reference	1.76±0.90	1.74±0.91

Table 6: Human evaluation scores (mean±std). Inter-annotator agreement: Krippendorff's α=0.39 (adequacy), 0.40 (fluency)

Divergence	System	Adequacy	Fluency
Q1 (Low)	Bayelemabaga	2.30	2.41
	Kumatigi HQ	**2.33**	2.42
	Kumatigi SB	2.32	**2.44**
	Reference	2.32	2.30
Q2	Bayelemabaga	1.98	2.06
	Kumatigi HQ	2.10	2.28
	Kumatigi SB	**2.32**	**2.32**
	Reference	1.99	2.03
Q3	Bayelemabaga	1.81	1.96
	Kumatigi HQ	2.00	2.04
	Kumatigi SB	**2.17**	**2.14**
	Reference	1.92	1.97
Q4 (High)	Bayelemabaga	1.43	1.60
	Kumatigi HQ	1.81	1.99
	Kumatigi SB	**2.08**	**2.24**
	Reference	1.96	2.06

Table 7: Human evaluation stratified by translation divergence.

tems by human evaluation perfectly mirrors automatic metrics: Kumatigi SB > Kumatigi HQ > Bayelemabaga across both adequacy and fluency. Notably, Kumatigi SB achieves scores comparable to or exceeding human reference translations on in-domain data (2.59 vs 2.34 adequacy on Kumatigi). In pairwise preference judgments, the reference was preferred most frequently (23.1%), followed by Kumatigi SB (16.7%), Kumatigi HQ (11.4%), and Bayelemabaga (10.3%). When directly compared to reference translations, Kumatigi SB demonstrated superior adequacy in 33.6% of cases versus 30.7% for Kumatigi HQ and 25.8% for Bayelemabaga, with similar trends for fluency. From quick observation of systems outputs, Reference surpass the other systems mostly on scientific, technical registers with OOV. This observation concerns mainly FLORES-200+ test set and was expected as analyzed in 3.4.2.

Table 7 reveals that Kumatigi SB's advantage grows systematically with translation difficulty: in low-divergence cases (Q1), all systems perform similarly (2.30-2.33 adequacy), but in high-divergence cases (Q4), Kumatigi SB substantially outperforms others (+0.27 adequacy over Kumatigi HQ, +0.65 over Bayelemabaga). **This demonstrates that synthetic data specifically improves robustness on challenging translations where systems diverge most**.

Overall, human evaluation confirms three key findings: (1) automatic metrics alignment with human quality judgments, (2) synthetic data augmen-

tation provides measurable improvements in both adequacy and fluency, and (3) these gains are most pronounced for difficult translations, suggesting our approach successfully addresses the long tail of translation challenges in low-resource settings. Qualitative samples along with explanation can be found in appendix B.

5 Conclusion

This work introduced Kumatigi, the largest French-Bambara corpus ($67k$ high-quality) along with data augmentation/generation approach, achieving state-of-the-art translation performance (+3–4 BLEU over prior work), consistent with human evaluation. We proposed a systematic recipe for low-resource corpus development by leveraging Round-trip translation to generate *two distinct datasets*-complete synthetic pairs for knowledge extraction, correction and back-translated pairs preserving authentic monolingual targets for vocabulary expansion. Our approach offers strategic data mixing (full synthetic, back-translated pairs) with appropriate scores (confidence score, round-trip similarity) and targeted annotation via uncertainty sampling, converting the typical low-resource constraint (abundant monolingual, scarce parallel) into an advantage. Results show clear improvements with both the Kumatigi high quality pairs and the data augmentation component. While our data collection, annotation and expansion methodology has been developed in the context of French to Bambara translation, it remains a methodological blueprint applicable to other under-resourced African languages facing similar data scarcity challenges.

Future works

Future work includes continuing to improve, expand the Kumatigi corpus. We intend to maintain it, release it publicly with versioning as it grows, in a framework of iterative dataset refinement. Furthermore, we believe that quality scores can be leveraged to set up or integrate Bambara to AfriCOMET, which currently does not include Bambara.

Acknowledgments

This work was conducted as an independent research initiative outside of the author's professional responsibilities at Orange. We are deeply grateful to **Orange** management for graciously providing access to GPU infrastructure that made this research possible. Thank you to **Moussa Albert Cissé** who actively participated to the data curation process. Many thanks to colleagues for valuable feedback on manuscript drafts.

Limitations

While our dataset curation, generation framework and quality scoring methodology are language-agnostic, this work focuses exclusively on French-Bambara translation. Generalization to other low-resource language pairs with different typological properties (e.g., non-agglutinative morphology, tonal vs. non-tonal) requires empirical validation. Our quality scores reflect one native-speaker judgment using consistent rubrics. Future work should investigate inter-annotator agreement and multi-annotator score aggregation. The corpus exhibits domain skew toward literary texts, conversational registers, news, religious materials, and historical documents, with limited coverage of technical and scientific registers. Our experiments exclusively employ fine-tuning of NLLB-200, which may reinforce existing biases in the pretrained model rather than learning language-specific patterns from scratch. Finally, our augmentation strategies address orthographic variability but do not resolve underlying standardization challenges in Bambara orthography.

Ethical Considerations

This research was conducted with careful attention to ethical considerations:
Data sourcing: All data sources are kept with URLs and metadata, allowing full distinction and comprehensive analysis.
Benefit to community: Beyond academic contribution, this work aims to enable practical applications for Bambara speakers: educational technology, government communication, cultural preservation, and digital accessibility.
Open access: Public release under permissive licensing ensures Bambara-speaking communities and researchers worldwide can benefit from these resources without financial barriers.
Data governance: Some copyrighted materials were used temporarily for language-pair adaptation during training. These materials won't be redistributed. All released datasets are partially or completely synthetic or derived from non-copyrighted sources

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *EMNLP 2018 Third Conference on Machine Translation (WMT18)*, pages 726–739. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Seokjin Oh, Su Ah Lee, and Woohwan Jung. 2023. Data augmentation for neural machine translation using generative language model. *arXiv preprint arXiv:2307.16833*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, and 1 others. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

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.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. Afromt: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. *arXiv preprint arXiv:2109.04715*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 228–234.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M Mustafa Rafique, and Marcos Zampieri. 2025. Bayelemabaga: Creating resources for bambara nlp. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12060–12070.

Valentin Vydrin, Kirill Maslinsky, Jean-Jacques Méric, and Andrij Rovenchak. 2011. Corpus bambara de référence.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, and 1 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. *arXiv preprint arXiv:1909.00157*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José GC De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orasan, Marina Fomicheva, and 1 others. 2022. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99.

A Annotation Guidelines and Process

The annotation process was conducted by two native Bambara speaker. We employed an iterative annotation framework that evolved from manual to semi-supervised as model quality improved.

A.1 Annotation Workflow

Phase 1: Manual Annotation (Rounds 1–2) Initially, after pre-processing sentences (filtering, normalization etc), annotators were presented with source sentences (French or Bambara) and manually produced translations. This phase established baseline quality standards and generated the initial training set.

Phase 2: Semi-Supervised Annotation (Rounds 3+) After fine-tuning an initial model, we transitioned to a semi-supervised framework. Annotators were presented with source sentences alongside model-generated translations and associated confidence scores. Their task shifted to validation and correction rather than translation from scratch, significantly reducing cognitive load and annotation time.

The model's confidence score served as a prioritization signal and annotators focused on low-confidence predictions, where errors were more frequent. However, the confidence score was advisory only; annotators independently judged quality and corrected high-confidence errors when identified. This created an implicit active learning loop: as annotation progressed, model performance improved on previously problematic cases, and annotators increasingly focused on remaining difficult examples.

Iterative Refinement and Scaling Over three annotation iterations, we observed progressive improvements in model quality and developed intuitions about systematic error patterns (e.g., rare vocabulary, cultural expressions, orthographic variants). When appropriate, annotators used large language models (GPT-4, Claude Haiku 4.5) to paraphrase French source sentences for increased fluency and diversity, expanding coverage of natural language variation.

After completing three iterations with sufficient model quality, we used the best checkpoint to automatically translate all remaining monolingual data via back-translation and forward translation. These model-generated pairs were assigned quality scores in the range $q \in [0.25, 0.68]$ to distinguish them from acceptable human-annotated translations ($q \geq 0.70$). This scoring strategy enables: (1) separation of synthetic from authentic data during training, (2) quality-based filtering experiments, and (3) prioritization for future human review. As annotation continues, model-generated pairs will be progressively reviewed and corrected by annotators, with scores updated to reflect human judgment.

Annotation Infrastructure Annotation was conducted through a custom web application built with Next.js, with our fine-tuned NLLB model serving as the backend for pre-annotation. The interface displayed source text, model predictions, confidence scores, and allowed annotators to edit translations and assign quality scores efficiently. This infrastructure supported rapid iteration and quality control throughout the annotation process.

A.2 Quality Scoring Rubric

Annotators assigned each translation a quality score $q \in [0, 1]$ based on the following criteria:

- **0.00–0.25:** Unusable translation. Major semantic errors or incomprehensible output. Annotator may lack domain knowledge despite understanding the source.
- **0.25–0.50:** Core meaning preserved but significant lexical errors. Requires substantial revision.
- **0.50–0.70:** Acceptable translation with minor lexical or grammatical issues. Usable for training.
- **0.70–0.90:** Good quality translation. Minor stylistic improvements possible but semantically accurate.
- **0.90–1.00:** Near-perfect to perfect translation. Natural, fluent, and accurate.

Multiple valid translations were allowed when acceptable alternatives existed.

B Qualitative Examples

To understand more in details errors pattern, metrics flaws, we analyze translation patterns where metric scores and human judgments may diverge (Table 8). Ref, Bayel, Ku, KuSB correpond respectively to: reference, bayelemaga, kumatigi, kumatigi (Synthetic + Back-translated pairs) systems. Table 8 shows some examples along with chrF scores. One can notice that the automated (chrF++) can be misleading as the best, human preferred system has not always the best score. More importantly, the ranking is corrupted. Beyond this point, globally all the systems tends to produce good or acceptable translation except in OOV cases where errors can be dramatic.

C Ablation Study

To understand the contribution of each component in our data augmentation framework, we conduct systematic ablations varying dataset types (synthetic, back-translated), quality thresholds, and tagging strategies. Following Caswell et al. (2019), we also tested tagged back-translation to help the model distinguish data sources. Back-translated sentences (synthetic source, authentic target) are tagged as `[BT]` SRC_{text} `[/BT]`, while fully synthetic pairs are tagged with `[SYN]`. These tags should enable the model to leverage authentic target-side fluency from back-translated data while recognizing synthetic source characteristics. Our naming convention: $H70\_S70\_BB70\_BF70$ indicates quality thresholds where H = high-quality parallel pairs (bidirectional), S = synthetic pairs, BB = back-translated with authentic Bambara targets, BF = back-translated with authentic French targets. Numbers indicate minimum quality scores (e.g., 70 = score ≥ 0.7). Recall that synthetic and back-translated pairs that has not been reviewed yet scores are included in $[0.25, 0.68]$.

C.1 Key Findings

Back-translation outperforms synthetic data. Comparing lines 1 and 3 (adding synthetic: +0.30 BLEU) versus lines 1 and 5 (relaxing back-translation threshold: +1.13 BLEU) reveals that **back-translated pairs provide substantially larger gains than synthetic pairs**. This suggests that even imperfect back-translations with authentic target-side text are more valuable than fully synthetic pairs, likely because they preserve authentic Bambara linguistic patterns.

Untagged training dramatically outperforms tagged. Strategy II (lines 7-8) achieves +2.49 BLEU over baseline, vastly exceeding Strategy I's +1.17 BLEU (line 6). This counterintuitive result suggests that **for low-resource MT, treating all augmented data as authentic allows better generalization** than explicitly marking synthetic sources. The model appears to benefit from increased data in both directions.

Hybrid strategy yields best results. Strategy III (line 11) achieves +2.56 BLEU, slightly exceeding pure untagged training. By applying tags only to low-quality pairs (score < 0.55) while treating high-quality pairs as authentic, we leverage the benefits of both approaches: **tags help the model handle noisy sources while preserving authentic targets for learning**, whereas high-quality pairs train without artificial markers. This nuanced approach effectively performs implicit quality-based curriculum learning.

Example & Analysis	chrF++↑	Human
FR: Le crocodile hurla **EN:** The crocodile screamed.		
Ref: Bama kulela		Valid variant
Bayel: Donsokɔrɔnin pɛrɛnna.	2.82	**Not valid**
Ku: bama kasira	31.19	**Acceptable**
KuB: bama kulera	59.38	Valid variant
Multiple morphological derivations and Lexical variation *kulela* = *kulera* = *pɛrɛnna* for "screamed" are acceptable but penalized by n-gram metrics. kasira means cry, close but not exact. Donsokɔrɔnin is completely false for crocodile		
FR: Il va causer avec le vieux tout le temps, même si les autres sont à une réjouissance, il laisse la réjouissance... **EN:** He goes and chats with the old man all the time, even when the others are having fun, he leaves the fun behind...		
Ref: A bɛ taa cɛkɔrɔba baro tuma bɛɛ, hali ni tɔw bɛ ɲɛnajɛ la, a bɛ ɲɛnajɛ in to yen...		**Preferred**
Bayel: A bɛ kuma ni cɛkɔrɔba ye tuma bɛɛ, hali ni mɔgɔ tɔw bɛ ɲɛnajɛ la, a bɛ ɲɛnajɛ to.	64.61	Valid/Fluent
Ku: a bɛ baro kɛ ni cɛkɔrɔba ye tuma bɛɛ, hali ni tɔw bɛ nisɔndiya la, a bɛ nisɔndiya to yen...	54.34	Acceptable
KuB: A bɛ taa baro kɛ ni cɛkɔrɔba ye tuma bɛɛ, hali ni tɔw bɛ nisɔndiya la, a bɛ nisɔndiya to...	54.27	**Preferred**
Context: *ɲɛnajɛ* (feast, have fun) and *nisɔndiya* (rejoicing, happiness) are closely related.		
FR: L'enfance est une période agréable. **EN:** Childhood is a pleasant time.		
Ref: *Bilakoroya* ka di.		**Fluent**
Bayel: *Denbaya* ye waati ɲuman ye.	6.56	**Minor error**
Ku: *Denmisɛnni* ye waati ɲuman ye.	6.24	**Minor error**
KuB: *Denmisɛnya* ye waati ɲuman ye.	7.36	**Preferred**
Cultural: Use of *Bilakoroya* (childhood before circumcision) in ref which is tied to cultural practices. *Denmisɛnya* (childhood, more general), *Denmisɛnni* (child), *Denbaya* (Parenthood). *ye waati ɲuman ye* means *is a good time* and *ka di* means *is good*		
FR: Et que tu voudras bien me pardonner », et elle se mit à genoux, les deux mains croisées au dos. **EN:** "And I hope you will forgive me," she said, kneeling with her hands clasped behind her back.		
Ref: I ka yafa n ma: a y'a bolo fila d'a ko.		Natural
Bayel: i ka yafa ne ma, a y'i ɲɔngiri, a bolo fila fara a kɔ kan.	43.61	Acceptable
Ku: ko i be yafa ne ma, a ye i ɲɔngiri, a bolo fila sirilen be a kɔ fɛ.	35.38	**Preferred**
KuB: A y'i ɲɔngiri k'a bolo fila fara a kɔ kan.	37.41	**Missing part**
There is different way to express *hands clasped behind her back.* Ku is prefered here to reference.		

Table 8: Translation examples get insight on systems quality along with chrF scores.

Lower thresholds improve Fr→Bm more than Bm→Fr. Relaxing back-translation thresholds (BB70→BB50, lines 1→5) yields +1.13 BLEU for Fr→Bm but minimal gains for Bm→Fr. This asymmetry indicates that **the French→Bambara direction benefits more from quantity over quality in authentic Bambara targets**, consistent with Bambara being the lower-resource side where exposure to diverse authentic patterns matters most.

C.2 Implications

Our ablations challenge conventional wisdom about tagged back-translation for low-resource MT. While tagging helps in high-resource settings (Caswell et al., 2019), low-resource scenarios benefit from treating augmented data uniformly, with selective tagging only for demonstrably low-quality pairs. This suggests future work should explore adaptive tagging strategies based on resource availability and data quality distributions.

D Linguistic Data Augmentation

Table 10 presents the probabilities p_i used for each augmentation technique in our experiments. Language agnostics augmentations are applied within mutually exclusive groups. In other words, for a source text, only one augmentation is randomly selected for each group. Language specific augmentations are applied independently. All Case augmentation, in particular, is applied on both source and target. More experiments should be done to find the optimal hyper-parameters.

E Data Release

We will publicly release the Kumatigi corpus (CC BY-SA 4.0 license) upon paper acceptance to ensure reproducible research and community benefit. The dataset will be hosted on Hugging Face Hub with comprehensive documentation and metadata.

E.1 Release Strategy

Our release follows a progressive versioning approach:

Training Configuration	Bambara→French			French→Bambara		
	BLEU↑	chrF++↑	TER↓	BLEU↑	chrF++↑	TER↓
NLLB-200 (zero-shot)	6.69	25.27	93.27	10.90	31.27	93.24
Strategy I: Tagged pairs (only authentic targets used for training)						
1. $H70\_S70\_BB70\_BF70$	13.12±0.06‡	32.48±0.06‡	79.70±0.14‡	13.05±0.01‡	34.50±0.05‡	85.43±0.41‡
2. $H70\_S70\_BB70\_BF70$ + aug	13.16±0.08‡	32.55±0.08‡	79.68±0.13‡	13.29±0.15‡	34.63±0.11‡	83.94±0.95†
3. $H70\_S60\_BB70\_BF70$	13.24±0.13‡	32.74±0.15‡	79.80±0.24‡	13.35±0.05‡	34.82±0.05‡	84.08±0.40‡
4. $H70\_S70\_BB55\_BF70$	13.10±0.03‡	32.53±0.03‡	79.89±0.10‡	13.78±0.14‡	35.42±0.07‡	83.96±0.73†
5. $H70\_S70\_BB50\_BF70$	13.11±0.02†	32.74±0.06‡	80.12±0.23‡	14.18±0.04‡	35.74±0.04‡	82.19±0.58‡
6. $H70\_S55\_BB55\_BF55$	13.76±0.24 ‡	33.74±0.27 ‡	78.96±0.19 ‡	14.26±0.15‡	36.07±0.07‡	81.19±0.27‡
7. $H70\_S55\_BB40\_BF55$	13.54±0.15‡	33.33±0.31‡	79.27±0.15‡	14.22±0.09‡	36.01±0.13‡	80.90±0.53‡
Strategy II: Untagged pairs (all pairs treated as clean, bidirectional training)						
7. $H55$	14.00±0.14‡	34.28±0.12‡	79.11±0.21‡	15.54±0.07‡	36.86±0.03‡	77.81±0.09‡
8. $H50$	13.62±0.05‡	33.94±0.15‡	79.49±0.23‡	15.37±0.06‡	36.76±0.04‡	78.77±0.26‡
9. $H40$	13.68±0.22‡	33.95±0.12‡	79.40±0.22‡	15.27±0.09‡	36.86±0.06‡	78.73±0.21‡
Strategy III: Hybrid (untagged for high-quality, tagged for low-quality pairs)						
10. $H55\_S60\_BB40\_BF40$	13.69±0.36‡	33.86±0.43‡	79.09±0.26‡	15.61±0.18‡	36.92±0.17‡	78.34±0.20‡
11. $H55\_S30\_BB30\_BF30$	13.63±0.35‡	33.94±0.39‡	79.30±0.35‡	15.57±0.02‡	36.96±0.10‡	78.23±0.48‡

Table 9: Systematic ablation on Kumatigi test set. Baseline (Zero-shot): BLEU=10.90 (Fr→Bm). Strategy III achieves +2.56 BLEU over baseline (line 1).

Augmentation	Prob	Group
Bambara elision	0.08	None
Bambara character substitution	0.04	none
Bambara synonym substitution	0.1	none
French accent substitution	0.02	none
char typo	0.01	char modification
char swap	0.01	char modification
keyboard noise	0.01	char modification
punctuation	0.01	formatting
case noise	0.01	formatting
whitespace	0.01	formatting
ocr noise	0.02	visual confusion
word dropout	0.01	word level
All Case	0.04	none

Table 10: Probabilities values for all experiments.

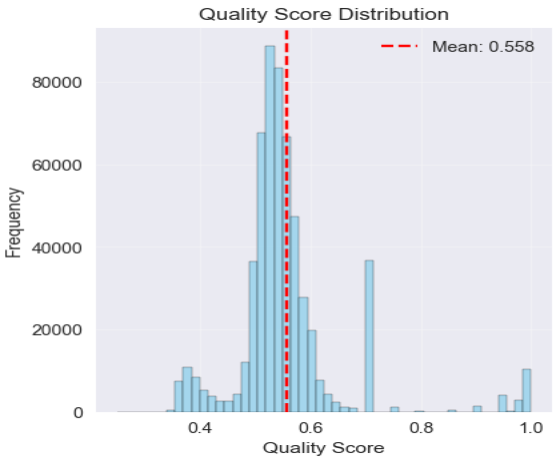

Figure 1: Quality score distribution

- **Phase 1:** Kumatigi HQ (quality scores $q \geq 0.7$) containing 67k high-quality parallel pairs
- **Phase 2-K:** Incremental releases as annotation progresses, with clear version control and changelogs

E.2 Score Distribution

Figure 1 shows the distribution of quality scores across our annotated corpus. Human-reviewed pairs ($q \geq 0.7$) are much less then synthetic/back-translated pairs ($q \in [0.25, 0.68]$). This scoring enables principled data filtering and targeted annotation prioritization.