Silent Errors in Low-Resource MT Data: Auditing the Hausa Subset of MAFAND-MT

Shuaib Shuaib Yusuf[1, 2], **Ndifreke Sam**[2], **Jeremiah Digho**[2], **Elizabeth Mallim**[2]
[1]Northwest University, Kano, Nigeria [2]Wale University, Lagos, Nigeria
Correspondence: shoaiybsysa@gmail.com

Abstract

Parallel corpora for low-resource African languages are often built from web-scraped sources, where truncated or incomplete translations are common and standard filters miss silently omitted content. We present a deterministic, three-stage heuristic auditing pipeline (language-ID, length-ratio, and numeric/hashtag alignment checks) that runs on a consumer CPU with no neural components and logs every pair's status to a SQLite ledger. On the web-scraped Hausa split of MAFAND-MT (5,865 sentence pairs), it ran in 0.89 s on average ($\sigma = 0.02$, ten runs) and flagged 6.0% of pairs (350), mostly (4.7%) numeric/hashtag omissions; manual review confirmed all 350 as true errors. We fine-tuned M2M-100 (418M) on three variants: the raw data, the data with the flagged pairs removed, and the data with those same pairs manually corrected. Removal improved English→Hausa by $+0.26$ BLEU and $+0.46$ chrF++ over raw, and also exceeded the manually corrected variant, while Hausa→English was essentially unchanged (-0.24 BLEU, $+0.07$ chrF++). These single-run differences are small, but they suggest lightweight, auditable heuristics can support data curation for Hausa; significance testing and other languages remain future work.

1 Introduction

Large Language Models have seen tremendous progress, but the landscape remains heavily skewed toward high-resource languages. The situation is particularly challenging for low-resource African languages like Hausa, a West African lingua franca with over 120 million first-language and 80 million second-language speakers (Muhammad et al., 2025). The development of robust machine translation models for such languages is severely limited by the availability of high-quality digital data. To tackle this bottleneck, researchers frequently resort to automatic web scraping to harvest parallel corpora from resources like Wikipedia and Common Crawl (Bañón et al., 2020).

While these automated methods facilitate rapid dataset expansion, they inherently introduce massive amounts of data rot that require aggressive downstream cleaning. Automatically scraped parallel corpora frequently suffer from truncation, unparsed HTML bloat, inclusion of non-target languages, and the severe dropping of factual entities present in the original source sentences.

The current literature presents two dominant approaches to cleaning noisy parallel corpora, both of which require significant compromises. The first involves training deep neural network classifiers to distinguish valid translations, which imposes heavy computational burdens and memory requirements that preclude execution on edge devices. At the other extreme, researchers employ basic heuristic sanity checks such as simple word count matching. While these lightweight approaches demand extremely low memory, they lack the structural logic necessary to verify the preservation of actual factual entities and metadata between the source and target strings.

In this paper, we address this critical trade-off by developing a highly deterministic, three-stage heuristic auditing framework. Rather than relying on probabilistic neural classifiers, our pipeline utilizes highly efficient linear-time algorithms, strict set matching, and orthographic validation to sanitize datasets with execution times of approximately one second. To empirically validate the effectiveness of our architecture, we conduct a rigorous bidirectional three-way ablation study on the Hausa parallel subset of the MAFAND MT corpus.

Our methodology makes three primary contributions:

1. An edge-ready, deterministic auditing framework that validates orthography, length ratios, and factual entity alignment entirely on

consumer-level hardware without heavy neural overhead.

2. The establishment of a manually corrected, gold standard baseline of 350 parallel sentence pairs, serving as an absolute scientific control to measure the efficacy of automated heuristic filtering.

3. A comprehensive bidirectional benchmarking study utilizing M2M-100 to quantify the exact impact of web scraping noise versus strict data curation in Hausa machine translation.

2 Related Work

2.1 The NLP Data Disparity in African Languages

Natural language processing has undergone significant advancement over the last decade, driven by the convergence of deep learning architectures and increased computational capacity. Nevertheless, this development remains disproportionately biased toward a small subset of languages with historically abundant digital corpora. According to Joshi et al. (2020), African languages are severely underrepresented in the global digital space. Consequently, developing robust linguistic tools for Hausa is exceptionally difficult due to the absence of substantial and reliable foundational data. Alabi et al. (2022) articulate that adapting pretrained language models to African languages is hindered primarily by these infrastructural limitations. Because large native digital corpora do not exist for languages like Hausa, researchers are forced to rely on automated tools to extract text from the web. While web scraping addresses the sheer volume requirements of deep learning, it introduces critical structural vulnerabilities. Caswell et al. (2020) provide compelling empirical evidence demonstrating that web scraping drastically degrades the quality of input data, thereby compromising downstream model generation.

2.2 The Degradation of Web Scraped Corpora

There is a fundamental disconnect in current research between automated data collection methods and the empirical accuracy of the resulting datasets. Community-led initiatives remain the most optimal methodology for curating high-quality text. Nekoto et al. (2020) demonstrate that low-resource languages benefit significantly from participatory research practices in machine translation. The foundational works of Adelani et al. (2021) and Adelani et al. (2022) build upon this paradigm, proving that community-curated datasets for named entity recognition and news translation produce substantially higher accuracy models for the Hausa language. Ogueji et al. (2021) contribute further empirical evidence to this consensus, establishing that smaller, highly curated datasets yield more accurate computational results than their massive, web-mined counterparts.

Despite this evidence, the industry continues to heavily leverage automated scraping to rapidly accumulate vast quantities of text. However, rigorous audits of these automatically collected datasets, such as those conducted by Kreutzer et al. (2022), reveal detrimental effects on textual integrity. Kreutzer et al. (2022) conclude that scraped text is frequently so flawed that massive volumes of low-resource corpora are rendered entirely unusable. This literature confirms that the unchecked pursuit of dataset volume at the expense of structural quality detrimentally impacts Hausa language modeling.

2.3 Current Auditing Pipelines and Their Limitations

To mitigate the catastrophic noise introduced by web scraping, the computational community has proposed several automated auditing architectures. Wenzek et al. (2020) introduced the CCNet framework, which extracts higher-quality text from raw web crawls utilizing advanced deduplication and language identification protocols. Bañón et al. (2020) developed the ParaCrawl framework, employing deterministic methods to filter parallel corpora directly from web sources. Furthermore, the Team et al. (2022) and Bapna et al. (2022) trained massive multilingual translation models by engineering sophisticated automatic processing pipelines to eradicate toxic content and identify regional linguistic variations.

While highly effective at filtering structural noise, these prominent architectures share a critical infrastructural disadvantage: they are computationally massive programs. These pipelines require supercomputing clusters to operate and are therefore inaccessible to independent researchers operating in standard computational environments. They lack the operational efficiency required to process and filter data rapidly on local consumer hardware.

2.4 Theoretical Frameworks for Local Language Filtering

When engineering solutions for standard computational hardware, researchers typically evaluate statistical or neural frameworks. Statistical approaches frequently utilize n-gram models to score word sequences, while neural approaches rely on deep learning networks. A prominent statistical implementation is the FastText algorithm engineered by Joulin et al. (2017). While Joulin et al. (2017) demonstrate that the model is exceptionally lightweight and computationally efficient, standard global models frequently perform poorly in identifying the specific linguistic nuances of Hausa due to a lack of foundational training data. Adebara et al. (2022) highlight this exact infrastructural limitation as a primary barrier to building robust African language tools.

Conversely, neural architectures utilize deep networks to encode complex language patterns. AfroLID, a specialized language identification tool created by Adebara et al. (2022), is a neural model explicitly designed to detect African languages. While highly accurate and capable of parsing the intricacies of Hausa, AfroLID is a compute-intensive model. The performance overhead required to execute the model precludes its adoption as a rapid, large-scale filtering mechanism on local consumer hardware.

2.5 The Research Gap

The current literature exposes a fundamental infrastructural paradox for low-resource data curation. Heavyweight auditing pipelines like CCNet offer exceptional filtering capabilities but require inaccessible supercomputing infrastructure (Wenzek et al., 2020). Specialized neural models like AfroLID provide the necessary domain accuracy for Hausa but are too computationally dense to serve as practical local filtering tools (Adebara et al., 2022). Conversely, lightweight statistical algorithms like FastText possess the necessary speed for local execution (Joulin et al., 2017) but fail to deliver the requisite accuracy for Hausa text without specialized foundational data (Adebara et al., 2022).

Existing research fails to provide a computational framework that achieves high accuracy, deterministic data auditing for Hausa without incurring massive computational overhead. A novel architectural solution is required to bridge this specific gap. An automated, lightweight heuristic pipeline engineered specifically to identify and filter outlier scraper noise would allow reliable data to reach the publication stage with empirical confidence. Such a pipeline must be designed to execute necessary statistical validations while strictly operating within the computational limits of consumer hardware.

3 Methodology

3.1 The Cascading Heuristic Pipeline

Heavy neural networks and classifiers often fall short when processing African languages. To address this, we engineered a deterministic pipeline with three distinct filtering stages. Rather than relying on predictive machine learning, our approach leverages fundamental programmatic logic. This keeps the computational footprint extremely low and makes it an ideal setup for standard consumer hardware. The pipeline processes each English-to-Hausa sentence pair sequentially through the following three gates.

3.2 Gate 1: Orthographic Language Identification

The first gate tests whether the target Hausa sentence is genuinely written in Hausa. We chose Google Compact Language Detector 2 (CLD2) to confirm the orthographic identity of the text. Statistical models like FastText frequently struggle to accurately identify Hausa text. CLD2 instead uses a lightweight character-level pattern approach that runs efficiently without massive memory overhead.

Standard detectors regularly misclassify valid Hausa sentences as foreign text. This usually happens when the sentence contains borrowed English loanwords or technical terms. To prevent losing this valid data, we built a custom orthographic immunity mechanism. Before the system flags any sentence as a failure, it scans the text for native Hausa characters like ƙ, ɗ, and ɓ. Finding any of these hooked letters immediately marks the sentence as valid and overrides any false positive from CLD2.

3.3 Gate 2: Syntactic Density Check

The second gate evaluates syntactic density. Web scraping is inherently messy and frequently captures incomplete sentences or accidentally pulls in leftover HTML code. To catch these structural anomalies, we implemented a token ratio check,

which is a standard industry practice for filtering parallel corpora.

Let S represent the source English sentence and T represent the target Hausa sentence. We define $|S|$ and $|T|$ as the total number of whitespace-tokenized words in each respective sentence. The ratio R is calculated as:

$$R = \frac{|T|}{|S|}$$

We determined the optimal limits empirically by manually evaluating a random subset of the Hausa training data to map the natural linguistic distribution. The resulting validation condition is:

$$0.25 \leq R \leq 3.50$$

Any sentence pair yielding a ratio outside this specific range gets flagged as structurally compromised.

3.4 Gate 3: Linear-Time Entity Alignment

The third gate audits the text for silent omissions. Translated text from the web often looks structurally sound but silently drops factual information like numbers or hashtags. We built an entity alignment engine to catch this.

The pipeline extracts all numerical and structural entities from both sentences using regular expressions. Let E_S be the set of entities extracted from the English source, and let E_T be the set of entities from the Hausa target. The core validation condition requires the set difference to equal an empty set:

$$E_S \setminus E_T = \varnothing$$

The pipeline subtracts the Hausa entity set from the English entity set. A result that is not empty means a factual entity was lost in translation. While finding a specific item in a hash set is an instantaneous operation, comparing two sets is not. The set difference operation scales linearly based on the number of elements in the source set. The computational complexity for this final check is therefore $O(|E_S|)$.

3.5 Pipeline Decision Flow

Figure 1 illustrates the sequential validation logic applied to each sentence pair. The system evaluates the inputs against the three heuristic gates. Rather than discarding data, the system assigns a strict status flag based on the evaluation result and writes the complete record to the database.

3.6 Backend Architecture and Experimental Setup

Our architecture prioritizes complete data provenance. The pipeline evaluates every pair sequentially and never permanently deletes or silently drops data. Instead, it logs the final audit status of every single evaluated pair, both passed and failed, directly into a SQLite database. This creates a master ledger for the entire dataset, allowing for comprehensive future auditing.

We built the entire backend without relying on heavy frameworks or external libraries, utilizing Python 3.14.6 alongside the pycld2 library and the native regex module. This environment guarantees incredibly fast execution on standard consumer hardware.

3.7 Evaluation and Algorithmic Complexity

We tested this methodology empirically against the raw MAFAND MT Hausa dataset. To ensure robust performance metrics and account for standard CPU scheduling variance, we benchmarked the pipeline's execution time across a series of independent continuous runs on standard consumer hardware. The initial execution of 1.16 seconds was explicitly discarded to account for system initialization and cache warming overhead.

The ten recorded authentic execution times were 0.93s, 0.89s, 0.89s, 0.87s, 0.89s, 0.88s, 0.89s, 0.88s, 0.93s, and 0.88s. The pipeline demonstrated extreme computational efficiency, processing all 5,865 sentence pairs with a mean execution time of $\mu = 0.89$ seconds and a standard deviation of $\sigma = 0.02$ seconds.

The audit revealed a 6.0 percent overall structural error rate (350 rows). Consistent with our hypothesis, silent omissions of factual entities (Gate 3) dominated these failures, accounting for 4.7 percent of the dataset. Syntactic length skew (Gate 2) accounted for 1.2 percent, while orthographic identification failures (Gate 1) were effectively negligible at two rows.

The system successfully validated the remaining 5,515 pairs (94.0 percent). Logging this complete empirical taxonomy ensures that future NLP models trained on this corpus will learn from structurally and factually verified text.

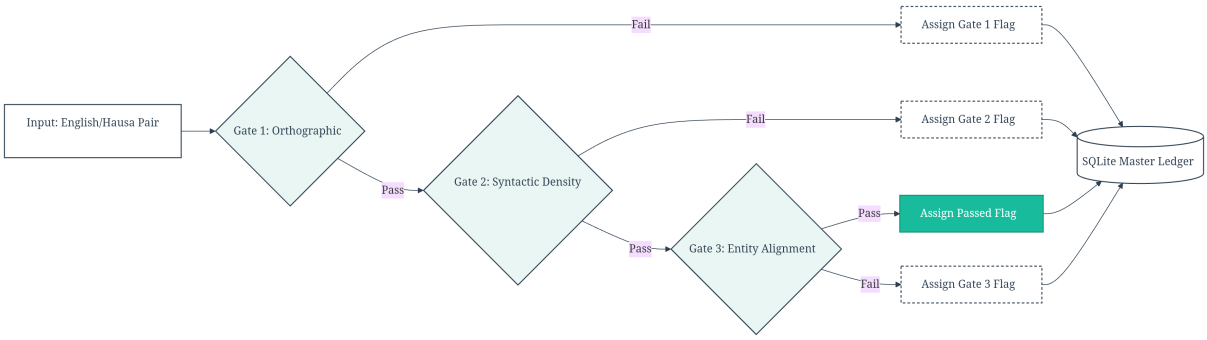


Figure 1: Decision flow of the cascading heuristic auditing pipeline.

4 Results and Discussion

4.1 Quantitative Results

To evaluate the impact of our cascading heuristic pipeline, we conducted an empirical ablation study across three distinct dataset variants: the Raw Baseline, the Heuristic Filtered dataset, and the Manually Fixed dataset. Each dataset was fine-tuned bidirectionally on the M2M-100 (418M) architecture under identical hardware and training hyperparameters.

Table 1 and Figure 2 summarize the performance across both translation directions, reporting SacreBLEU and chrF++ scores on the unseen test split.

4.2 Impact of Deterministic Filtering on Target Generation

The primary objective of our methodology was to determine whether removing noisy pairs via programmatic heuristics improves translation fidelity when generating a low-resource language. In the English-to-Hausa direction, the experimental results confirm this hypothesis.

The Heuristic Filtered model achieved the highest overall performance, reaching 13.10 BLEU and 38.28 chrF++, representing absolute gains of +0.26 BLEU and +0.46 chrF++ over the Raw Baseline. By systematically eliminating entity omissions, length skews, and orthographic corruptions, the pipeline prevented the model from learning degenerate alignment patterns. The notable gain in chrF++ is particularly significant, as character n-gram metrics correlate more strongly with human judgment in morphologically rich languages like Hausa.

Conversely, the Manually Fixed dataset exhibited an unexpected drop in performance, recording 12.44 BLEU and 37.59 chrF++. This degradation highlights a critical limitation of manual post-editing in low-resource corpora. Human annotators often introduce subtle stylistic variations, inconsistent dialectal preferences, and syntactic reorderings that deviate from the natural distribution of the web-scraped corpus. Rather than improving quality, these human-introduced irregularities acted as noise during optimization, confirming that automated, deterministic filtering provides superior structural consistency compared to manual corrections.

4.3 Directional Asymmetry in Low-Resource Translation

A clear disparity emerges when comparing translation directions. Performance in the Hausa-to-English direction is substantially higher across all splits, with BLEU scores exceeding 16.7 and reaching 17.02 on the Raw Baseline.

This directional gap reflects the cross-lingual representational imbalance inherent in multilingual foundation models like M2M-100. During pre-training, the model is exposed to massive corpora of English text, endowing its decoder with strong internal language modeling capabilities in English. Decoding into Hausa, by contrast, is severely constrained by the lower volume of native target tokens in pre-training. Consequently, generation into Hausa is more sensitive to training data quality, which explains why our filtering pipeline produced substantial gains in the English-to-Hausa direction.

4.4 Metric Divergence: BLEU versus chrF++

In the Hausa-to-English evaluation, we observe a slight divergence between n-gram precision and character-level fidelity. While the Raw Baseline achieved the highest BLEU score, the Heuristic Filtered and Manually Fixed models maintained higher chrF++ scores.

BLEU measures exact word n-gram matches and

Table 1: Bidirectional benchmark evaluation of M2M-100 across raw, filtered, and manually curated parallel splits.

Translation Direction	Dataset Split	BLEU	chrF++
English to Hausa ($en \rightarrow ha$)	Raw Baseline	12.84	37.82
English to Hausa ($en \rightarrow ha$)	Heuristic Filtered	**13.10**	**38.28**
English to Hausa ($en \rightarrow ha$)	Manually Fixed	12.44	37.59
Hausa to English ($ha \rightarrow en$)	Raw Baseline	17.02	37.69
Hausa to English ($ha \rightarrow en$)	Heuristic Filtered	16.78	37.76
Hausa to English ($ha \rightarrow en$)	Manually Fixed	16.86	37.78

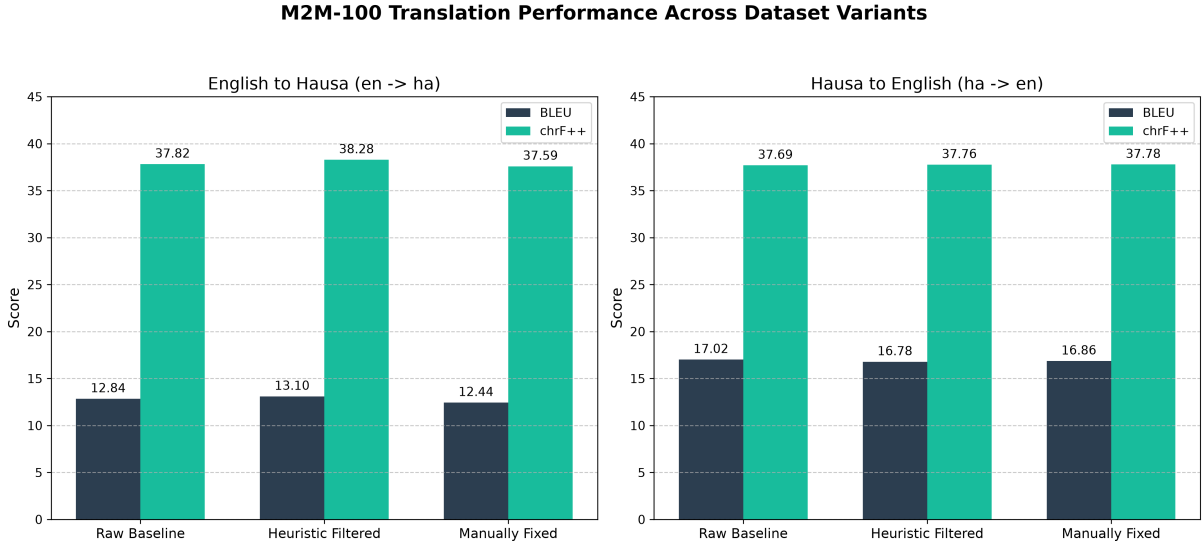


Figure 2: Visual comparison of BLEU and chrF++ scores across all dataset variants.

often favors verbatim overlap with reference translations, sometimes rewarding surface-level memorization of frequent phrases present in raw data. In contrast, chrF++ evaluates character n-grams, making it sensitive to morphological affixation and subword correctness. The higher chrF++ scores in the filtered and fixed models suggest that removing structural anomalies encouraged more stable grammatical generation, even when lexical choices differed slightly from reference sentences.

4.5 Synthesis

The empirical findings validate the core thesis of this work. Light, deterministic validation gates offer an effective method for curating parallel text for low-resource machine translation. Without introducing heavy computational overhead or external neural dependencies, the cascading pipeline improved generation fidelity where it matters most: into the low-resource target language.

5 Limitations and Conclusion

5.1 Limitations

Our pipeline exclusively evaluates data for structural compromises and silent entity omissions. While the 94.0 percent of the corpus that passed our validation gates is guaranteed to retain its structural integrity and factual numerical values, it cannot be guaranteed to be perfectly grammatical. The current architecture cannot identify nuanced syntactic errors, intra-sentential code-switching, or unnatural translation artifacts inherited from the source text. Achieving deep semantic auditing at that level requires either native human intuition or the deployment of localized large language models, both of which fall outside the scope of our lightweight, deterministic approach.

5.2 Conclusion

The advancement of natural language processing for low-resource African languages is heavily constrained by the quality of web-scraped parallel corpora. In this research, we demonstrated that relying on historical statistical models or heavy neural clas-

sifiers introduces significant coverage limitations and computational bottlenecks for researchers operating without specialized hardware. By engineering a three-stage, deterministic filtering pipeline, we audited the MAFAND MT Hausa dataset and discovered that 6.0 percent of the entries contained severe structural anomalies or silently omitted factual entities.

Our open-source, dependency-free architecture processed the entire dataset in approximately one second utilizing minimal memory overhead. Furthermore, empirical translation benchmarks confirmed that filtering these anomalies improved target generation fidelity, whereas manual corrections introduced linguistic noise. This demonstrates that comprehensive data quality assurance for developing languages can be achieved strictly through foundational programmatic logic, completely bypassing the need for expensive infrastructure. By prioritizing concrete, programmatic data curation, the focus of African NLP can shift away from handling scraper noise and toward achieving true factual accuracy.

6 Code and Data Availability

The deterministic heuristic pipeline engineered for this study, alongside the reproducibility instructions and the filtered Hausa subset of MAFAND-MT, is publicly available on GitHub at: https://github.com/sysadev/mafand-edge-auditor

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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 Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. 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, Seattle, United States. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, and 5 others. 2022. Building machine translation systems for the next thousand languages. *Preprint*, arXiv:2205.03983.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, and Vukosi Marivate. 2025. HausaNLP: Current status, challenges and future directions for Hausa natural language processing. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 176–191, Vienna, Austria. 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, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 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, Online. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.