# Neural Machine Translation for English–isiXhosa Medical Communication: A Literature Review

Malibongwe Makhonza
Department of Computer Science
University of Cape Town
Cape Town, Western Cape, South Africa
mkhmal024@myuct.ac.za

## ABSTRACT

Machine translation (MT) between English and isiXhosa is essential in multilingual healthcare, yet both the language pair and the medical domain pose major challenges. This literature review surveys recent progress in neural machine translation (NMT), highlighting how advanced architectures like the Transformer and subword segmentation bolster low-resource languages. It then examines the specific hurdles of isiXhosa translation — primarily data scarcity — and discusses large-scale multilingual models that have made significant strides. Finally, it reviews domain-specific MT techniques, including fine-tuning on in-domain corpora, generating synthetic data, and prompting large language models. These methods aim to incorporate crucial medical terminology and style into MT outputs. Although English–isiXhosa medical MT has lagged behind high-resource pairs, a recently released high-quality bilingual medical dataset now enables rigorous domain adaptation. By combining low-resource translation strategies with domain-focused techniques, researchers can move toward more accurate and context-relevant English–isiXhosa medical MT, reducing language barriers in critical healthcare settings.

## CCS CONCEPTS

• **Computing methodologies** → **Machine translation**; *Transfer learning*; • **Applied computing** → Life and medical sciences.

## KEYWORDS

Neural Machine Translation, Low-Resource Languages, isiXhosa, Domain Adaptation, Medical Translation

**ACM Reference Format:**
Malibongwe Makhonza. 2025. Neural Machine Translation for English–isiXhosa Medical Communication: A Literature Review. In *African AI Conference'25: Proceedings of the ....* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1234567.1234567


*African AI Conference'25, March 28–30, 2025, Cape Town, South Africa*

ACM ISBN 978-1-4503-0000-0/25/xx
https://doi.org/10.1145/1234567.1234567

## 1 INTRODUCTION

Healthcare delivery in multilingual contexts increasingly demands effective language technologies to ensure patient safety and equitable access to information. In South Africa, many clinical resources are available only in English, placing patients who speak isiXhosa or other indigenous languages at a disadvantage [1]. Recent advances in neural machine translation (NMT) have dramatically improved translation accuracy for high-resource language pairs [2]. However, limited parallel corpora and the complexity of medical terminology remain key obstacles for English–isiXhosa translation. This shortage of domain-specific data stands out as a critical bottleneck, as large NMT models generally rely on extensive corpora to learn specialized vocabulary.

Until recently, the absence of English–isiXhosa healthcare texts has hindered the use of domain adaptation methods. Such approaches typically involve fine-tuning a baseline translator on in-domain medical data, thus teaching the model to handle specialized terms like disease names, medication instructions, and clinical procedures [3]. Blocker et al. [2025] have addressed this gap by releasing a dataset of audio and text from primary healthcare consultations in English and isiXhosa, offering an unprecedented opportunity to test adaptation techniques on real-world clinical interactions. Their corpus not only provides aligned translations but also includes medical annotations, enabling researchers to analyze how effectively NMT systems can handle important health-related concepts.

While isiXhosa is sometimes described as "low-resourced" due to its agglutinative morphology, it is more accurate to note that the language's true challenge arises from limited digitized data rather than its grammatical structure [4]. Morphological complexity can pose technical hurdles—for example, subword segmentation plays a larger role in capturing prefixes and suffixes—but does not itself determine resource status. The real bottleneck is that, for decades, only a small amount of isiXhosa text has been digitized and aligned with English. Consequently, models trained on general text rarely see specialized health terms and often mis-translate or skip unknown words.

Developers of NMT for low-resource medical translation thus face a dual challenge. First, they must secure enough bilingual data to represent the terminology of clinical practice. Second, they must adapt NMT architectures to handle specialized text, where accuracy is paramount and errors in dosage instructions or symptom descriptions can jeopardize patient well-being [5]. Techniques like back-translation permit the generation of synthetic in-domain examples from monolingual isiXhosa medical documents, alleviating the data shortage to an extent. However, stable results typically require at least a small parallel "seed" corpus for fine-tuning. The dataset introduced by Blocker et al. [2025] fulfills this need by providing actual patient–provider dialogues. Early results suggest that even a small quantity of domain-specific text can significantly improve model handling of clinical terms.

This literature review surveys the evolution of NMT approaches for low-resource languages and explores how recent breakthroughs—like subword segmentation and domain adaptation—can be applied to English–isiXhosa medical translation. We begin by contrasting older phrase-based MT systems with modern NMT solutions and discuss the importance of data availability for isiXhosa. Next, we examine specialized approaches, including the new isiXhosa healthcare corpus, that offer avenues for adapting models to medical terminology. Throughout, we highlight the potential for back-translation and fine-tuning to overcome data scarcity, while acknowledging the real-world constraints of patient safety and the medical context. By synthesizing existing research, we aim to show how advanced NMT architectures, combined with newly available healthcare corpora, can bring reliable English–isiXhosa translation closer to reality.

# 2 NEURAL MACHINE TRANSLATION

Neural Machine Translation (NMT) has emerged as the dominant paradigm in MT, gradually superseding earlier statistical approaches such as phrase-based systems. Unlike phrase-based SMT, which translates small segments and often yields disjointed output, NMT employs end-to-end neural models that consider the entire input sentence, resulting in more fluent and contextually coherent translations [7]. This transition was motivated by NMT's ability to better handle long-range dependencies and capture nuanced context, thereby improving the fluency and grammaticality of translations [2]. NMT models also adapt more easily to new data or domains through fine-tuning, enhancing their versatility compared to the rigid pipelines of statistical MT. Additionally, modern NMT workflows leverage subword segmentation techniques (e.g., Byte Pair Encoding [17]) to address vocabulary limitations. By splitting rare or morphologically complex words into subword units, NMT can effectively translate languages like isiXhosa without suffering from out-of-vocabulary issues.

## 2.1 Encoder–decoder models

Early NMT architectures adopted a sequence-to-sequence design using recurrent neural networks (RNNs) [18]. An encoder RNN processes the source sentence into a continuous representation, and a decoder RNN generates the target sentence from that representation. Bahdanau et al. [2015] introduced attention mechanisms to overcome the fixed-length bottleneck, letting the decoder attend to different parts of the source at each time step [2]. This enabled significant quality gains, especially on longer sentences. However, RNN-based models still incurred high computational costs, motivating further exploration of parallelizable architectures.

The Transformer [19] replaced recurrence with self-attention. It uses multi-head self-attention layers to learn complex alignments between tokens, along with positional encodings to retain sequence order. The encoder is a stack of layers that each apply multi-head attention to the source embeddings, while the decoder similarly uses self-attention for its own tokens, plus cross-attention over the encoder's output. This design efficiently captures long-range dependencies and trains faster on GPUs, quickly becoming the state-of-the-art for many language pairs. Importantly, the Transformer's subword-based approach can better handle morphologically rich languages by modeling partial word units. This success in general MT sets the stage for applying Transformer-based architectures to isiXhosa translation and domain-specific tasks.

## 2.2 LLMs for Machine Translation

Large language models (LLMs), such as GPT or mT5, represent a related but distinct shift in translation. Unlike a dedicated NMT system supervised on parallel corpora, an LLM is typically pretrained on massive monolingual text across multiple languages, learning broad linguistic patterns that can then be repurposed for translation. GPT-3, for instance, exhibits zero-shot or few-shot translation by leveraging in-context learning [8]. In principle, LLMs can handle English–isiXhosa translation if enough isiXhosa data is included in pretraining, and domain content is present or can be prompted. This capability makes LLM-based translation highly flexible, as a single model can adapt to many domains through careful prompting. However, the approach can suffer if the LLM lacks exposure to specialized medical terms or if isiXhosa is underrepresented in its training. Additionally, inference can be resource-intensive and prone to "hallucinations" or inconsistent terminology usage. Researchers have thus explored prompting techniques that incorporate domain glossaries or exemplars to mitigate these shortcomings [10]. As LLMs continue to expand in scale and coverage, their role in low-resource and domain-specific MT is set to increase, complementing or even supplanting conventional NMT in some cases.

# 3 ISIXHOSA TRANSLATION

As an officially recognized South African language, isiXhosa has historically been low-resourced in MT research. Recent efforts, however, have accelerated progress by both curating new data and leveraging multilingual models. This section reviews the state of isiXhosa MT, covering newly available datasets and benchmarks, as well as the advances in multilingual neural MT systems that include isiXhosa.

## 3.1 Datasets and Benchmarks for isiXhosa MT

The scarcity of parallel corpora has long impeded MT development for isiXhosa. To address this, the WMT 2022 shared task on African MT included English–isiXhosa translation as one of its target language pairs, marking the first time isiXhosa featured in a large MT benchmark. This addition spurred the creation and use of new datasets. Each low-resource African language (including isiXhosa) received between roughly 1.4k and 35k parallel sentence pairs, representing a substantial expansion over previously limited text resources.

Standard evaluation benchmarks have also been extended to cover isiXhosa. Goyal et al. [2022] introduced FLORES-101, a many-to-many evaluation benchmark with 3,001 sentences, explicitly including isiXhosa (language code "xho"). More recently, FLORES-200 further increased coverage to 200 languages, adding about 3k translations per language, providing a high-quality test bed for consistent comparisons of English–isiXhosa systems [11]. Beyond parallel data, large-scale monolingual corpora have also become available to support techniques such as back-translation and pre-training. Notably, the CC100 corpus compiled by Wenzek et al. [2020] contains on the order of 25 million Xhosa tokens gleaned from web-crawled text, which can be leveraged for unsupervised MT or language modeling [4]. Altogether, these combined efforts— from WMT tasks to FLORES-based evaluations— equip researchers with both training corpora and standardized test sets for rigorous development and assessment of isiXhosa MT quality.

### 3.2 Multilingual NMT Models Including isiXhosa

In parallel with data collection, multilingual neural MT systems have rapidly expanded to include isiXhosa. Rather than relying exclusively on scarce Xhosa–English data, these models learn from dozens or even hundreds of languages, sharing cross-lingual representations that prove beneficial for low-resource tongues like isiXhosa. A key milestone was M2M-100 [7], introduced by Fan et al. [2021] as "the first single model that could translate between any pair of 100 languages" without pivoting through English. Crucially, isiXhosa is among the supported languages, facilitating direct XhosaEnglish translations via one unified model. While M2M-100 demonstrated the viability of massively multilingual MT, its performance for very low-resource languages remained somewhat constrained by the available training data.

Building on that foundation, Meta AI released the NLLB-200 [9] model in 2022, significantly advancing isiXhosa translation alongside other low-resource languages. NLLB-200 (No Language Left Behind) is a Mixture-of-Experts Transformer covering 202 languages, including all South African languages such as isiXhosa. By training on large-scale curated parallel data and carefully balancing low-resource language coverage, NLLB-200 achieved strong gains in quality compared to prior systems [9]. FLORES-200 was introduced concurrently to validate these improvements for isiXhosa and many other languages, with human-translated references ensuring reliable test conditions. The outcome is that isiXhosa translation, once nearly absent from mainstream MT, now benefits from large-scale architectures that approach a level of fluency and accuracy deemed unthinkable just a few years ago.

Beyond these high-profile Meta-backed models, smaller-scale multilingual approaches have also shown promise. Emezue and Dossou [2021] successfully fine-tuned an mT5 model on a parallel corpus covering six African languages (including isiXhosa) plus English and French, yielding decent translation results from only moderate data [20]. This approach exemplifies cross-lingual transfer: by training a single multilingual model that includes multiple African languages, the system leverages interlingual similarities to compensate for isiXhosa's sparse resources. Likewise, other regional adaptations of mBART have confirmed the value of domain-focused pre-training and multilingual data to bolster isiXhosa translation quality. Together, these initiatives illustrate how newly enriched datasets and advanced multilingual Transformers can dramatically lift the performance of English–isiXhosa systems. M2M-100 [7] and NLLB-200 [9] highlight the transformative impact of large-scale many-to-many models, while more focused solutions (like African-languages fine-tuning) underscore the versatility of cross-lingual NMT for this historically under-resourced language.

## 4 DOMAIN SPECIFIC TRANSLATION

Domain adaptation is often necessary to achieve high-quality translation in specialized fields like medicine. General-purpose MT models trained on broad datasets tend to falter on domain-specific terminology and styles [5]. For example, a phrase that is common in medical English may have a very specific equivalent in isiXhosa that a generic model has never learned. In the context of English–isiXhosa medical MT (a low-resource scenario), the problem is compounded by the scarcity of parallel medical corpora. Researchers have thus investigated techniques that infuse domain expertise into NMT systems, ranging from fine-tuning on in-domain data to creating synthetic domain examples and prompting large language models with medical knowledge.

### 4.1 Fine-Tuning for Domain Adaptation

A straightforward approach to domain-specific translation is fine-tuning: taking a general model and further training it on in-domain parallel data. Even a small specialized corpus can significantly improve terminology usage, as the model directly "sees" correct translations for medical expressions [6, 15]. However, fine-tuning also risks overfitting if the domain data is too limited. A heavily specialized model may degrade on out-of-domain text or misapply domain vocabulary where it does not belong. In English–isiXhosa medical MT, this trade-off is especially delicate: the parallel data is scarce, yet medical texts demand precision. Nonetheless, when carefully managed with regularization or domain tagging, fine-tuning remains one of the most effective means to incorporate specialized knowledge into an NMT system. It can be combined with a general bilingual corpus or a multilingual model as a base, yielding a strong domain-adapted translator for medical usage.

### 4.2 Training on Synthetic Domain Data

When in-domain parallel corpora are missing or too small, generating synthetic domain data becomes a powerful alternative. Techniques such as back-translation create pseudo-parallel sentences by translating monolingual in-domain text with an existing model, then pairing that output with the original. Sennrich et al. first popularized back-translation for low-resource settings [17], and subsequent work applied it to domain adaptation, specifically targeting missing terminology [5]. In the medical domain, one might use English health guidelines or isiXhosa monolingual brochures to generate new parallel examples. Other approaches rely on dictionary-based replacement of domain terms in generic sentences, artificially expanding coverage of specialized vocabulary [15]. Although synthetic examples can be noisy, they substantially enhance model exposure to rare words or phrases, a crucial factor for English–isiXhosa medical MT. Iterative strategies can refine the quality of synthetic data, gradually improving the model's capacity to handle complex domain terms. A caveat is that an inaccurate initial model might produce flawed translations, so filtering or repeated translation passes are often recommended.

### 4.3 Domain-Specific Prompting of Large Language Models

Large language models (LLMs) add another dimension to domain adaptation. Rather than (or in addition to) training a specialized model, one can prompt an LLM with instructions or examples relevant to the medical field. Ghazvininejad et al. [2023] propose providing domain glossaries and short definitions in the prompt, enabling the model to select correct translations for specialized terms [10]. This method requires no re-training, making it especially attractive for resource-limited setups. If the LLM's pretraining covered enough isiXhosa text or similar Bantu languages, it may already have an internal representation of medical concepts in isiXhosa. By specifying a domain style, we can ensure more formal or consistent translations. However, prompting alone does not guarantee stable performance if the LLM occasionally ignores instructions or has limited knowledge of certain terms. Another concern is cost and accessibility: calling a large API-based model for each translation may be expensive. Even so, domain-specific prompting stands out as a promising strategy, particularly for newly emergent medical vocabulary or small-scale specialized tasks, where building a fully fine-tuned system might be less practical.

## 5 DISCUSSION

At the intersection of general NMT advances, isiXhosa-specific translation research, and domain-focused MT, a complex interplay of challenges and techniques emerges. Over the past decade, general neural MT has seen dramatic improvements in architecture and training methodology. The field progressed from the first sequence-to-sequence RNN models [18] to architectures augmented with attention, and eventually to the Transformer model that now underpins state-of-the-art systems [19]. Notably, Transformers coupled with subword segmentation have proven exceptionally effective even for relatively low-resource settings, by handling long-range dependencies and open-vocabulary translation through byte-pair encoding [17]. These advances provide a strong foundation for English–isiXhosa translation: a modern MT system can better capture context and generate fluent outputs than earlier phrase-based approaches. However, powerful architecture alone is not a panacea. The case of isiXhosa highlights that data sparsity and linguistic complexity can still bottleneck performance. Without sufficient training examples, even Transformers struggle to accurately translate rare medical terms or colloquial expressions. While general NMT improvements offer the tools, success in this context depends critically on how they are adapted to a low-resource, domain-specific scenario.

Progress in isiXhosa machine translation has underscored issues that go beyond generic model architecture. As a low-resource language, isiXhosa presents unique linguistic hurdles that general NMT improvements only partially address. One major challenge is the lack of large, high-quality digitized corpora in specialized fields such as healthcare. Even powerful multilingual models like M2M-100 [7] or NLLB-200 [9] may fail to handle domain-specific expressions in isiXhosa if those expressions were never seen in the training data. Fine-tuning on small medical parallel sets can help, but overfitting risks remain, and coverage of specialized phrases may still be incomplete. Synthetic data generation, via back-translation or dictionary-based insertion of terms, stands out as a key means of expanding coverage, yet it must be used judiciously to avoid introducing inaccurate or unnatural translations. Larger language models bring the intriguing possibility of domain-specific prompting, but the approach hinges on the LLM's latent knowledge of isiXhosa and medicine, which may be uneven or missing. These complexities become particularly stark in a medical context, where incorrectly rendered terms can have serious consequences for patient care.

Combining low-resource language challenges with domain specificity creates a uniquely difficult task for English–isiXhosa medical MT. No single technique can guarantee robust coverage. Instead, the literature suggests that a hybrid approach is needed: start with a strong multilingual or bilingual baseline, fine-tune on the available medical texts, augment with carefully curated synthetic sentences, and possibly leverage an LLM for difficult segments. This synergy is now more feasible thanks to the newly available dataset from Blocker et al. [2025], which provides aligned English–isiXhosa medical conversations. For the first time, the field can systematically evaluate and adapt these methods using genuine clinical content. Researchers can measure how well each adaptation strategy handles domain terms, assess if repeated back-translation truly boosts coverage for specialized words, and see if prompting an LLM for unknown terms can further reduce errors. The dataset thus serves as a testbed where the theoretical connections among NMT, isiXhosa low-resource translation, and domain adaptation converge into real-world validation.

With these tools and data in hand, future work can explore diverse strategies to refine English–isiXhosa medical MT. The immediate priority is to test domain adaptation solutions—fine-tuning, synthetic data augmentation, prompting—directly on the new corpus, quantifying which methods yield the best balance of accuracy and generalization. Another research avenue is to investigate more sophisticated filtering or iterative updating of synthetic data to maintain quality as the model improves. Additionally, one could examine whether related languages, like isiZulu, can be tapped via multilingual transfer to bolster isiXhosa performance in medicine. Large language models offer flexible, quickly configurable translation via prompt engineering, but verifying their reliability in life-critical contexts demands thorough evaluation and possibly domain-constraint methods (like forced dictionary usage). Taken collectively, these directions underscore how the interplay of modern NMT architectures, specialized data acquisition, and advanced adaptation techniques can push English–isiXhosa medical translation to a level that genuinely meets healthcare needs. Although significant obstacles remain—particularly the scarcity of professional translations for the vast array of medical concepts—ongoing progress makes the goal of accurate, domain-aware MT for isiXhosa ever more attainable.

# 6 CONCLUSION

This literature review shows that bridging the language gap in healthcare requires a threefold synthesis of general NMT innovations, isiXhosa-specific insights, and domain adaptation strategies. Modern Transformer-based approaches and subword segmentation have proven indispensable for capturing complex linguistic phenomena, yet they do not inherently resolve the data and vocabulary gaps that plague low-resource languages. Efforts to assemble parallel and monolingual isiXhosa corpora have begun closing that gap, especially via multilingual models like M2M-100 [7] and NLLB-200 [9], which leverage cross-lingual transfer and large training sets.

Nevertheless, the medical domain remains a specialized subset for which generic models often fail to handle crucial terms or style. Approaches such as fine-tuning on in-domain data, synthetic data generation, and domain-specific prompting of large language models all hold promise for infusing a system with the medical knowledge needed to accurately handle clinical text. The recent English–isiXhosa medical corpus provided by Blocker et al. [2025] finally offers a high-quality platform for testing these methods in a real-world setting. By combining domain adaptation, low-resource translation techniques, and state-of-the-art architectures, researchers can target the unique demands of English–isiXhosa medical translation and significantly improve patient access to critical healthcare information. Looking ahead, this synergy of strategies—supported by newly available data—offers a practical route toward robust, domain-aware MT systems that serve isiXhosa speakers effectively in clinical and public health contexts.

# REFERENCES

[1] Levin, M. E. 2006. Language as a barrier to care for Xhosa-speaking patients at a South African paediatric teaching hospital. *South African Medical Journal* 96, 10, 1076–1079. https://pubmed.ncbi.nlm.nih.gov/17164939/ (Accessed Mar 2025).

[2] Bahdanau, D., Cho, K., and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations* (ICLR).

[3] Blocker, A., Meyer, F., Biyabani, A., Mwangama, J., Datay, M. I., and Malila, B. 2025. Benchmarking isiXhosa automatic speech recognition and machine translation for digital health provision. In *Proceedings of the Workshop on Patient-oriented Language Processing*. 1–12.

[4] Martinus, L., and Abbott, J. Z. 2019. A Focus on Neural Machine Translation for African Languages. *arXiv preprint* arXiv:1906.05685.

[5] Mehandru, N., Robertson, S., and Salehi, N. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 1–10. https://dl.acm.org/doi/pdf/10.1145/3531146.3533244 (Accessed Mar 2025).

[6] Sennrich, R., Haddow, B., and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. *arXiv preprint* arXiv:1511.06709.

[7] Fan, A., Bhosale, S., Schwenk, H., Chaudhary, V., Goyal, N., et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research* 22, 1, 748–783.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (NeurIPS). 1877–1901.

[9] Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint* arXiv:2207.04672.

[10] Ghazvininejad, M., et al. 2023. Dictionary-based prompting for domain adaptation in neural machine translation. *arXiv preprint* arXiv:2302.07856.

[11] Goyal, N., Chen, M., et al. 2022. FLORES-101: Evaluating multilingual translation at scale. In *Proceedings of ACL 2022*. 1152–1163.

[12] Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Abbott, J. Z., and Marivate, V. 2020. Masakhane – Machine Translation for Africa. *arXiv preprint* arXiv:2003.11529. Available at: https://arxiv.org/abs/2003.11529.

[13] Hu, J., Xia, M., Neubig, G., and Carbonell, J. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of EMNLP 2019*. 2988–2994. Available at:https://aclanthology.org/P19-1286/ (Accessed Mar 2025).

[14] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the ACL* 5, 339–351. Available at:https://aclanthology.org/Q17-1024/ (Accessed Mar 2025).

[15] Peng, X., Yeh, T., Stefan, W., and Wilkins, D. 2020. Dictionary-based data augmentation for domain adaptation in NMT. *arXiv preprint* arXiv:2004.02577.

[16] Sennrich, R., Haddow, B., and Birch, A. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*. 1715–1725.

[17] Sutskever, I., Vinyals, O., and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in NIPS 2014*. 3104–3112.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. Attention is all you need. In *Advances in NIPS 2017*. 5998–6008.

[19] Wenzek, G., Lachaux, M.-A., Conneau, A., et al. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of LREC 2020*. 4003–4012.

[20] Emezue, C. C., and Dossou, B. F. P. 2021. MMTAfrica: Multilingual Machine Translation for African Languages. In *Proceedings of the 2021 Workshop on Machine Translation (WMT)*. 398–411. Available at:https://www.researchgate.net/publication/359891433_MMTAfrica_Multilingual_Machine_Translation_for_African_Languages (Accessed March 2025).