The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 3, June 2025
https://doi.org/10.33022/ijcs.v14i3.4714

A Culture-Aware Bidirectional IsiXhosa-English Neural Machine Translation Model Using MarianMT

Tebatso Gorgina Moape[1], Thuto Siyamthanda Mohale[2], Bester Chimbo[3]
moapetg@unisa.ac.za[1],61501980@mylife.unisa.ac.za[2], chimbb@unisa.ac.za3
[1,2,3]University of South Africa

Article Information

Received : 7 Feb 2025
Revised : 27 May 2025
Accepted : 9 Jun 2025

Keywords

Low resource languages, machine translation, neural machine translation, MarianMT

Abstract

Machine translation for low-resource African languages faces significant challenges due to limited data availability and complex linguistic features such as rich morphology, agglutinative grammar, and rich cultural expressions. This study proposes and develops a culturally aware machine translation model for isiXhosa-English language pairs using the MarianMT transformer-based model. We combine traditional parallel corpora with culturally enriched datasets, addressing the unique challenges of isiXhosa's linguistic intricacies. The proposed model was trained on a carefully curated dataset of 127,690 parallel sentences and used SentencePiece tokenization for handling agglutinative morphology. Our approach achieved a BLEU score of 58.79, marking a substantial improvement over previous methods, typically scoring between 20.9 and 37.11. The results demonstrate that integrating cultural context and linguistic specificities into the translation model substantially improves translation quality for low-resource languages. The study's findings suggest that considering cultural context, combined with appropriate model architecture and data preprocessing strategies, can lead to more accurate and culturally aware machine translation systems.

A. Introduction

In recent years, the field of Natural Language Processing (NLP) has been significantly transformed by the developments in Artificial Intelligence (AI), particularly in the areas of machine and deep learning [1]. Neural machine translation systems such as Google Translate have become integral to global communication and information exchange across diverse cultural contexts and languages. However, the benefits of these developments have not been evenly distributed across all world languages [2]. While widely spoken languages such as English, Mandarin, and Spanish have seen remarkable progress in translation technologies, many African languages, including South African languages like isiXhosa, remain underrepresented in the language technology landscape [3]. These translation systems often struggle with out-of-vocabulary (OOV) words, leading to inaccurate translations. The disparity is mainly due to the limited availability of linguistic resources.

IsiXhosa is a Bantu language spoken by approximately 7 million people in South Africa. The isiXhosa language presents unique linguistic challenges due to its complex morphology, agglutinative grammar, intricate syntactic structures, distinctive click consonants, tonal variations, and rich cultural phrases [4]. More often than not, these phrases cannot be directly translated. For example, “Umntu ngumntu ngabantu,” which directly translates to “A person is a person through other people,” carries a much deeper meaning. It embodies the philosophy of Ubuntu, emphasizing that an individual's identity and humanity are shaped by their interactions and treatment of others. A literal translation fails to capture the full essence and cultural significance of such expressions, highlighting the difficulties faced by machine translation systems in accurately interpreting isiXhosa.

Furthermore, expressions such as “Ukuzila” signify a profound period of mourning rooted in tradition and reverence, and “Imoto Onitsha rhatya” does not merely describe a “new car” but one that exudes a freshness that extends beyond a simple English equivalent. Similarly, words like “Kunjalo ngu” imply agreement with an added tone of finality, while interjections such as “Yhu, iyabanda!” capture both the intensity of the cold and the emotional response to it. Culture-specific words, whether referencing surprise (“Into yokumangalisa”), confusion (“ndixakwe”), or encouragement to stay composed (“Vele nje ukhumbule ukuhlala uzolile”), are integral to the isiXhosa language and culture. However, these culturally rich expressions are often not included in existing training datasets, leading to inaccurate translations as they cannot be effectively translated word-for-word.

Recent advancements in transformer-based models, such as MarianMT [5] and BERT [6], have demonstrated significant improvements in capturing context and enhancing translation quality, particularly for high-resource languages. Given this progress, this study employs the MarianMT transformer-based model to train a bidirectional isiXhosa-English translation system, evaluating its effectiveness in handling a linguistically and culturally rich language like isiXhosa. The model is trained on a small-scale parallel corpus enriched with culturally specific data. The objective is to integrate cultural and contextual intricacies into the translation process to ensure that idiomatic expressions and culturally significant concepts are accurately represented.

This paper is organized as follows: Section B explores the linguistic features and cultural characteristics of isiXhosa. Section C reviews existing research on isiXhosa-English machine translation. Section D details the materials and methods employed in this study. Section E presents the evaluation and experimental results. Finally, Section F concludes the paper.

B. Linguistic Characteristics and Cultural Intricacies of IsiXhosa

IsiXhosa is a Bantu language spoken primarily in South Africa. It is a tonal language where pitch variations can alter words' meaning. One of its most notable phonetic features is the presence of click sounds [7], which were borrowed from Khoisan languages. Another key phonetic feature is vowel harmony, where certain vowels within a word must belong to the same class, ensuring phonological consistency. Morphologically, isiXhosa is an agglutinative language. It forms words by adding prefixes and suffixes to roots [8]. It has a complex noun class system with 15 noun classes, which influence sentence structure and verb agreement. The syntax follows a Subject-Verb-Object (SVO) word order, although word order can sometimes be flexible depending on emphasis [4].

Regarding semantics and pragmatics, isiXhosa is rich in idiomatic expressions, proverbs, honorific forms, and politeness strategies that carry cultural significance [9]. Many of these expressions do not translate literally into English but hold deep philosophical meaning. For example, instead of the casual greeting “Molo”(“Hello”), one would say “Molweni” when addressing older people or multiple people as a sign of respect. Another example is “ukuphuma esikhumbini,” which literally means “step out of the womb,” but it conveys a complex sense of being unskilled or unsophisticated. The English equivalent holds no meaning to the phrase, and it isn't easy to express effectively in English.

These linguistic features present significant challenges for machine translation models. The agglutinative nature of isiXhosa, combined with its complex noun class system, makes it difficult for models to segment and interpret words correctly. Traditional sequence-to-sequence architectures often struggle to maintain proper agreement between noun classes and their corresponding verb forms across long sentences [10]. In addition, the tonal nature of the language adds another layer of complexity, as current text-based models cannot capture these pitch variations that can completely alter word meanings. The cultural and contextual aspects of isiXhosa pose even more significant challenges for translation models. Idiomatic expressions and cultural references lose their meaning when translated literally, and current models lack the cultural knowledge required to provide appropriate equivalents in target languages. This highlights the need for machine translation models incorporating cultural context and pragmatic understanding alongside linguistic rules.

C. Related Works

The challenge of data scarcity for indigenous languages cannot be overstated. In light of this challenge, researchers have worked with the limited resources to develop machine translation models. While some datasets exist, they are often not freely accessible. As a result, researchers frequently rely on publicly available resources, such as biblical texts and government domain data, to train their models.

Authors in [3] highlighted the significant challenge that many African languages have not fully benefited from advancements in neural machine translation due to this challenge. To address this issue, the authors compared three approaches, zero-shot learning, transfer learning, and multilingual learning for Shona, isiXhosa, and isiZulu to English translations. A parallel corpus of 128,342 English-isiXhosa sentences and isiXhosa-isiZulu parallel sentences were used to train the models. To evaluate the performance of the three approaches, the authors trained a many-to-many multilingual model capable of translating between English-isiXhosa, isiXhosa-isiZulu, and English-isiZulu language pairs. The model was trained using English-isiXhosa and isiXhosa-isiZulu language pairs for zero-shot learning. The best-performing model produced a BLEU score of 34.9 for the isiXhosa-isiZulu model and 20.9 for the English-isiXhosa model. The authors attributed the high BLEU score achieved for isiXhosa-isiZulu to the significant vocabulary overlap between these languages, as they belong to the Nguni language group.

Contrary to developing and training separate language models for each language, [11] trained a single multilingual translation model capable of translating between English and eight Southeast African languages. The model utilized overlapping Byte Pair Encoding (BPE), back-translation, synthetic training data generation, and data augmentation with additional translation directions during training. English and isiXhosa were among the language pairs included in the study. The model was trained on a parallel dataset of 8.6 million sentence pairs and achieved a BLEU score of 27.5 for isiXhosa-to-English translations and 12.1 for English-to-isiXhosa translations.

The authors in [3] and [11] included isiXhosa-isiZulu as language pairs due to their shared language family. Interestingly, [3] achieved a significantly higher BLEU score of 34.9, compared to 11.2 in [11]. This outcome is surprising, given the general assumption that larger datasets typically yield better results. Notably, [1] achieved its higher BLEU score using only 1.49% of the dataset size employed in [2]. This discrepancy highlights a compelling area for further research to analyze the relationship between dataset size, translation performance, and the methods used.

A large-scale multilingual machine translation model was trained in [12] using a transformer-based architecture and back-translation and reconstruction techniques. The isiXhosa monolingual dataset comprised 158,660 sentences, paired with 138,111 English sentences. IsiXhosa was also paired with other languages, such as Swahili and French. However, since this paper focuses solely on isiXhosa and English, only the results for this language pair are reported. The model achieved a BLEU score of 30.25. Across all models, the transformer-based approach demonstrated higher performance than the methods employed in [1] and [2].

Due to the limited availability of freely accessible datasets, some researchers utilize public domain data from government institutions or biblical texts. Researchers in [13] used the JW300 dataset, which contains approximately

786,371 parallel sentences, and achieved a BLEU score of 37.11 using neural machine translation techniques.

The relationship between dataset size, methods, and BLEU scores must be studied. Based on the reviewed papers, it is evident that both data quality and the choice of methodology play an essential role in achieving high translation performance. While larger datasets, such as those used in [2], provide extensive training opportunities, they do not always guarantee better results, as demonstrated by the significantly higher BLEU score achieved by [1] using a fraction of the data. This suggests that factors such as vocabulary overlap, linguistic similarity, and the appropriateness of the model architecture can outweigh dataset size in improving translation accuracy.

D. Materials and Methods

To create a culturally and contextually aware bidirectional translation model for isiXhosa and English, the MarianMT model was customized and fine-tuned. The process involved several key stages: data collection and augmentation, data preprocessing, and adaptive model training. Each of these steps is elaborated upon in the following subsections.

4.1 MarianMT

The MarianMT is an open-source transformer model explicitly designed for neural machine translation (MT) [5]. It was developed through collaboration between the University of Edinburgh, Adam Mickiewicz University in Poznań, and Microsoft. Built in C++, the model is highly optimized for machine translation tasks, focusing on efficiency and performance. Unlike general-purpose NLP transformer models prioritizing versatility across various tasks, MarianMT is tailored exclusively for translation, making it particularly effective in this specialized area. Key features of this model include deep recurrent neural networks (RNNs) with deep transition cells. [14], transformer-based architecture [6], multi-source models [15], a combination of RNN and transformer-based language models [16], and additional layer normalization with tied embeddings [16, 17], and residual or skip connections.

The RNNs with deep transition cells improve the model’s ability to capture long-range dependencies in sequences, making it more effective in handling complex sentence structures. This makes it suitable for complex languages such as isiXhosa. The transformer-based architecture enables parallel input data processing, significantly boosting translation speed and performance. The inclusion of the multi-source models allows the system to leverage multiple input languages or contexts, enhancing translation accuracy by providing richer linguistic context, which is important for processing the rich morphology of languages. The combination of RNN and transformer-based language models integrates the strengths of both architectures, balancing sequential dependency modeling with high computational efficiency.

Additional layer normalization stabilizes training and accelerates convergence, leading to more robust translations. Tied embeddings reduce the model's memory footprint by sharing parameters between input and output embeddings, making training more efficient. Lastly, residual/skip connections facilitate the flow of gradients during backpropagation. This prevents vanishing gradients and enables deeper networks to be trained effectively. A combination of these features makes MarianMT a highly optimized and powerful model for neural machine translation.

4.2 Data Collection

Data is integral to the development of any language model. For data collection, the study utilized the isiXhosa-English parallel data freely available from Opus, GitHub, and the South African Center for Digital Language Resources (SADiLaR) and additional culture-specific parallel sentences. These datasets provided a foundation for training and fine-tuning the model to enhance translation accuracy. The composition and distribution of these datasets are presented in Table 1.

Table 1. Datasets

Data	***Number of parallel sentences***
Opus	54
Github	877
SADiLaR	126 709
Additional culture-specific parallel sentences	50
Total Number of sentences	127,690

The Opus dataset comprised 54 sentences, GitHub contained 877 sentences, and SADiLaR contributed 126,709 sentences, resulting in 127,690 sentences. An additional 980 sentences were incorporated to further enhance the dataset, including over 50 culturally significant isiXhosa phrases. This augmentation aimed to bridge gaps in standard bilingual corpora by expanding the model's vocabulary and improving its ability to capture and interpret complex linguistic and cultural contexts.

4.3 Data Pre-processing

Data was preprocessed using three key steps: cleaning, tokenization, and normalization. These steps were essential for handling the linguistic complexities of isiXhosa, an agglutinative language where words are formed by combining more minor morphemes. Data cleaning involved removing noise, ensuring proper sentence alignment, and decomposing contractions to enhance the quality of training data. The inconsistencies were eliminated to ensure high-quality parallel text between isiXhosa and English, allowing the model to learn meaningful translation patterns.

Tokenization played a crucial role in preparing the text for machine translation. To enhance model comprehension, it involved breaking sentences into smaller units, such as words, subwords, or characters. Due to isiXhosa’s complex morphology, subword tokenization was employed, as it allows the model to recognize root words and prefixes while maintaining linguistic structure. Although character-level tokenization preserves linguistic granularity, it often reduces contextual coherence, making subword tokenization the preferred method. Studies like [3] highlight the importance of choosing the appropriate tokenization method to balance computational efficiency and translation accuracy for isiXhosa-English models.

Normalization ensured that all text data followed a standard format, reducing variability and enhancing consistency. This step included converting text to lowercase, removing punctuation, and handling diacritics to minimize inconsistencies in the dataset. For a language like isiXhosa, where tonal and phonetic variations are common, normalization is vital to maintaining translation accuracy. By applying these preprocessing techniques, the machine translation model was better equipped to handle isiXhosa's linguistic challenges, ultimately improving translation quality and model efficiency.

4.3 Model Training

The model was trained on 10 epochs and fine-tuned through iterative hyperparameter adjustments to align it with its specific linguistic objectives. This process involved optimizing key parameters such as learning rate, batch size, dropout rate, and subword tokenization techniques, with each modification guided by continuous performance evaluation. Given isiXhosa's agglutinative morphology, which results in long and morphologically complex words, the SentencePiece tokenizer was employed as a language-agnostic subword segmentation tool. By breaking words into subword units, the model effectively recognized shared roots, improving its handling of rare words and significantly reducing out-of-vocabulary (OOV) issues. This approach minimized the occurrence of <unk> tokens, ensuring better translation fluency and overall model accuracy.

The MarianMT pipeline was implemented on an RTX 3080 GPU with 12 GB VRAM, leveraging PyTorch and Hugging Face Transformers to enable efficient training and fine-tuning. A learning rate of 3e-5 was chosen to balance convergence speed and model stability, ensuring effective training without causing gradient divergence. Additionally, a batch size of 32 was selected to optimize GPU utilization, allowing for smooth training while preventing memory overload.

To improve translation coherence and contextual relevance, beam search decoding was applied with a beam size of 4, refining the model's ability to generate fluent, well-structured, and contextually accurate translations. Furthermore, dropout rates were set between 0.1 and 0.3 to prevent overfitting and mitigate repetitive token generation, a common issue in low-resource translation models. These dropout adjustments ensured that the model learned generalized linguistic patterns without memorizing training data.

Throughout training, validation set metrics played a crucial role in guiding hyperparameter adjustments, allowing for continuous refinements based on the model's performance. This iterative fine-tuning approach helped the model improve its handling of complex sentence structures, cultural expressions, and linguistic generalization, making it more effective for isiXhosa-English bidirectional translation.

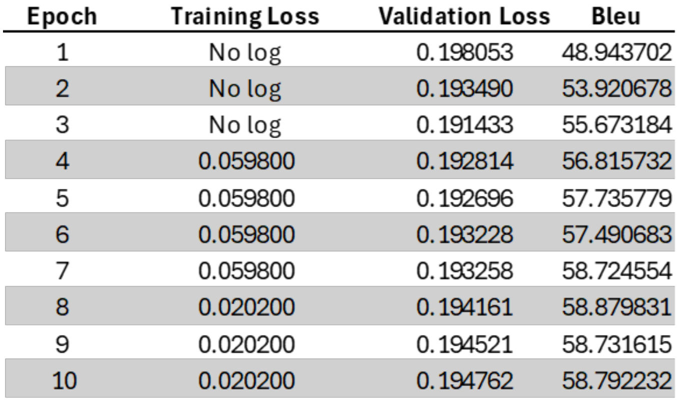
E. Evaluation and Results

Machine translation models are assessed through manual human or automatic evaluation [18]. In this study, automatic evaluation was preferred due to its efficiency in managing multiple iterations of the translation model. Furthermore, automatic metrics offer a standardized and reproducible approach, ensuring consistency in translation quality assessment across various experiments. The Bilingual Evaluation Understudy (BLEU) metric, introduced by [19], was utilized to measure the alignment between model-generated translations. This evaluation method applies the formula in equation (1) to quantify translation accuracy.

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} \frac{1}{N} \cdot \log\left(\text{precision}_{n}\right)\right) \quad (1)$$

The BLEU score is based on precision, which calculates the percentage of n-grams in the machine-generated translation that also appears in the reference translation. The score ranges from 0 to 1, with 1 indicating a perfect match between the model-generated translation and the reference translation. The evaluation was done using the NLTK library, where the output translation of each translation was passed to an evaluation function to calculate individual precision and brevity penalties.

The individual precision measures the accuracy of a translation system for each sentence, whereas the brevity penalty accounts for discrepancies in length between the candidate and reference translations. The brevity penalty penalizes the model for producing translations that are shorter than the reference translation. This ensures that the model does not prioritize shorter outputs to inflate precision scores. The final BLEU score is derived by integrating these two metrics. Figure 1 depicts the model performance.



Epoch	Training Loss	Validation Loss	Bleu
1	No log	0.198053	48.943702
2	No log	0.193490	53.920678
3	No log	0.191433	55.673184
4	0.059800	0.192814	56.815732
5	0.059800	0.192696	57.735779
6	0.059800	0.193228	57.490683
7	0.059800	0.193258	58.724554
8	0.020200	0.194161	58.879831
9	0.020200	0.194521	58.731615
10	0.020200	0.194762	58.792232

Figure1. Model training performance

The model performance displayed a consistent improvement in the BLEU score from epoch 2. The baseline configuration achieved a BLEU score of 53.92 and peaked at 58.79 at 10 epochs, which started stabilizing at epoch 7. There were minor fluctuations with the validation loss, but it remained relatively stable across interactions. This suggests that the adjustments in configuration, particularly the use of finer log-based parameter values, have a notable impact on enhancing translation quality while maintaining overall stability in the model's training process. These results demonstrate the effectiveness of iterative fine-tuning and hyperparameter optimization in improving machine translation outputs. Fig 2 provides a graphical visualization of the model’s performance.

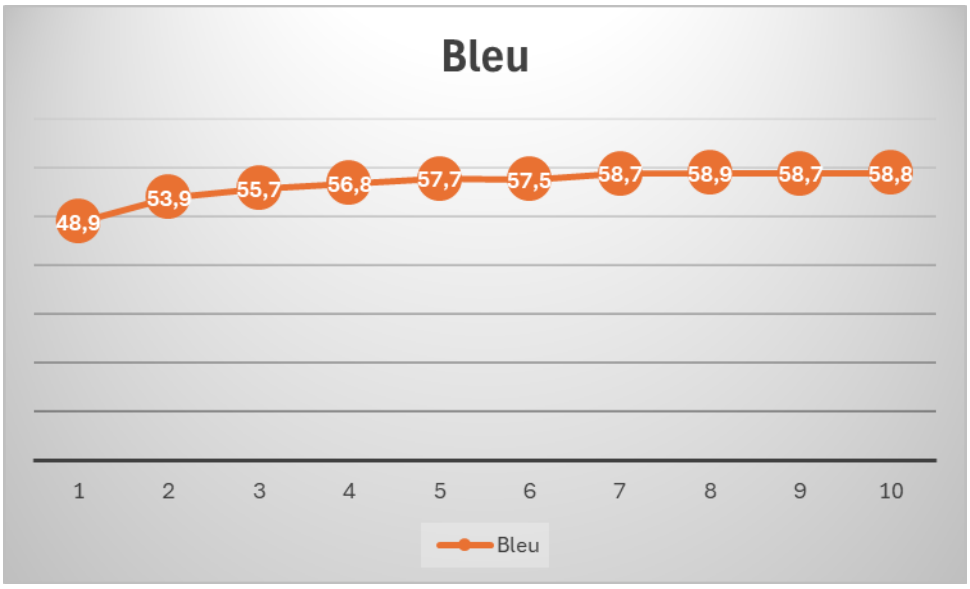


Figure2. Model Bleu score

The overall BLEU score indicates that translations that closely align with the reference translations were produced. This demonstrates improved performance and consistency in producing reliable translations, especially for languages such as isiXhosa with challenging linguistic structures. Table 2 provides a comparative analysis of the current English-isiXhosa machine translation systems and the proposed model.

Table 2. Comparative analysis results

Author	Architecture	Bleu
[3]	Zero-short transfer learning	20.9
[11]	Neural machine translation	27.6
[12]	mT5 model transformer-based	30.25
[13]	Neural machine translation	37.11
Proposed model	MarianMT	**58.79**

The proposed model in the paper, MarianMT, achieves a BLEU score of 58.79, significantly outperforming other models discussed in the literature. Compared to the zero-shot transfer learning model, which attained a BLEU score of 20.9, the proposed model demonstrates nearly three times better translation quality. Similarly, it surpasses neural machine translation models with BLEU scores of 27.6 and 37.11 by wide margins, highlighting its ability to generate accurate translations. The mT5 model, a transformer-based architecture with a BLEU score of 30.25, falls short in comparison. A key contributing factor to the model's success is its incorporation of culturally specific concepts and phrases, which improves coverage and addresses out-of-vocabulary (OOV) challenges. By integrating these elements into the training dataset, the model enhances its capacity to interpret linguistic contexts and reduces the prevalence of unknown tokens. These results underscore the efficacy of MarianMT in handling complex linguistic tasks while adapting to cultural and linguistic diversity.

F. Conclusion

This study introduced a culture-aware bidirectional isiXhosa-English neural machine translation model using MarianMT. The study integrated cultural context into the training dataset and leveraged SentencePiece tokenization to enable the proposed model to effectively address the linguistic complexities of isiXhosa. The model achieved a BLEU score of 58.79, significantly outperforming existing approaches. This demonstrates that incorporating cultural knowledge enhances translation accuracy for low-resource languages. Future work will focus on expanding the dataset with more diverse and contextually rich isiXhosa-English parallel sentences, particularly those covering domain-specific datasets. Future work will also include incorporating self-supervised learning techniques and multilingual pretraining strategies for different South African languages and integrating audio and visual contexts.

G. References

[1] Fanni, S.C., et al., Natural language processing, in Introduction to Artificial Intelligence. 2023, Springer. p. 87-99.

[2] Ranathunga, S., et al., Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 2023. **55**(11): p. 1-37.

[3] Nyoni, E. and B.A. Bassett, Low-resource neural machine translation for southern african languages. arXiv preprint arXiv:2104.00366, 2021.

[4] Oosthuysen, J.C., The grammar of isiXhosa. 2016: African Sun Media.

[5] Junczys-Dowmunt, M., et al., Marian: Fast neural machine translation in C++. arXiv preprint arXiv:1804.00344, 2018.

[6] Vaswani, A., Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[7] Gxowa-Dlayedwa, N., Investigating click clusters in isiXhosa syllables. South African Journal of African Languages, 2018. 38(3): p. 317-325.

[8] Mzamo, L., A. Helberg, and S. Bosch. Towards an unsupervised morphological segmenter for isiXhosa. in 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA). 2019. IEEE.

[9] Diko, M., IsiXhosa as a preservative instrument of culture: A consideration of ethnolinguistics. Southern African Linguistics and Applied Language Studies, 2024. 42(sup1): p. S12-S22.

[10] Kann, K., Neural sequence-to-sequence models for low-resource morphology. 2019, lmu.

[11] Elmadani, K.N., F. Meyer, and J. Buys, University of Cape Town's WMT22 System: Multilingual Machine Translation for Southern African Languages. arXiv preprint arXiv:2210.11757, 2022.

[12] Emezue, C.C. and B.F. Dossou, MMTAfrica: Multilingual machine translation for African languages. arXiv preprint arXiv:2204.04306, 2022.

[13] Martinus, L., et al., Neural machine translation for South Africa's official languages. arXiv preprint arXiv:2005.06609, 2020.

[14] Barone, A.V.M., et al., Deep architectures for neural machine translation. arXiv preprint arXiv:1707.07631, 2017.

[15] Junczys-Dowmunt, M. and R. Grundkiewicz, An exploration of neural sequence-to-sequence architectures for automatic post-editing. arXiv preprint arXiv:1706.04138, 2017.

[16] Ba, J.L., Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[17] Press, O. and L. Wolf, Using the output embedding to improve language models. arXiv preprint arXiv:1608.05859, 2016.

[18] Martin, J.H., Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2009: Pearson/Prentice Hall.

[19] Papineni, K., BLEU: a method for automatic evaluation of MT. Research Report, Computer Science RC22176 (W0109-022), 2001.