

# Benchmarking Neural Machine Translation for Southern African Languages

**Laura Martinus**

Explore Data Science Academy, South Africa  
laura@explore-ai.net

**Jade Z. Abbott**

Retro Rabbit, South Africa  
jabbott@rettorabbit.co.za

## Abstract

Unlike major Western languages, most African languages are very low-resourced. Furthermore, the resources that do exist are often scattered and difficult to obtain and discover. As a result, the data and code for existing research has rarely been shared. This has led a struggle to reproduce reported results, and few publicly available benchmarks for African machine translation models exist. To start to address these problems, we trained neural machine translation models for 5 Southern African languages on publicly-available datasets. Code is provided for training the models and evaluate the models on a newly released evaluation set, with the aim of spur future research in the field for Southern African languages.

## 1 Introduction

Africa has over 2 000 languages across the continent (Eberhard et al., 2019). South Africa itself has 11 official languages, and many more unofficial languages. Unlike major Western languages, the multitude of African languages are very low-resourced. Furthermore, the resources that do exist are often scattered and difficult to obtain or discover. The datasets are often owned by small institutions and research is published in smaller African conferences and never published online. As a result, the data and code for existing research has rarely been shared, meaning researchers struggle to reproduce the results properly, and few publicly available benchmarks for African machine translation models exist.

To begin to solve the highlighted problems, we train neural machine translation (NMT) models for a subset of Southern African languages, for which parallel publicly-available datasets exist and publish the code on GitHub<sup>1</sup>. We evaluate the models on the new Autshumato evaluation set for machine translation of Southern African languages (McKellar, 2017), in an attempt to spur future research in machine translation of Southern African languages.

Translation models are trained for English to Afrikaans, isiZulu, Northern Sotho, Setswana and Xitsonga, using state-of-the-art NMT architectures, namely, Convolutional Sequence-to-Sequence (ConvS2S) and Transformer architectures (Chahuneau et al., 2013).

## 2 Languages

The isiZulu, N. Sotho, Setswana, and Xitsonga languages belong to the Southern Bantu group of African languages (Mesthrie and Rajend, 2002). The Bantu languages are agglutinative and all exhibit a rich noun class system, subject-verb-object word order, and tone (Zerbian, 2007). Northern Sotho and Setswana are closely related and highly mutually-intelligible. Xitsonga is a language of the Vatsonga people (Bill, 1984). The language of isiZulu belongs to the Nguni language family, and is known for its morphological complexity (Keet and Khumalo, 2017). Afrikaans is an analytic West-Germanic language, that descended from Dutch settlers (Roberge, 2002).

## 3 Related Work

Machine translation research for Southern African languages is scarce (Martinus and Abbott, 2019; Abbott and Martinus, 2018; Wilken et al., 2012; McKellar, 2014). Much of the research suffers from the

---

<sup>1</sup>Available online at: <https://github.com/LauraMartinus/ukuxhumana>

problems described above. For example, Van Niekerk report BLEU scores as high as 37.9 for English-to-Northern Sotho translation and 71 for English-to-Afrikaans translation, but neither the datasets nor the code are published (van Niekerk, 2014). Wilken *et al.* use publicly-available parallel datasets, but reports using monolingual datasets which were not released (Wilken et al., 2012).

## 4 Data

The Autshumato parallel corpora are aligned parallel corpora of South African governmental data (Groenewald and Fourie, 2009), for English to Afrikaans, isiZulu, Northern Sotho, Setswana, and Xitsonga.<sup>2</sup>

|           | <b>Afrikaans</b> | <b>isiZulu</b> | <b>Northern Sotho</b> | <b>Setswana</b> | <b>Xitsonga</b> |
|-----------|------------------|----------------|-----------------------|-----------------|-----------------|
| Sentences | 53 172           | 26 728         | 30 777                | 123 868         | 193 587         |

Table 1: Number of parallel sentences for English-to-Target language

Table 1 highlights how low resourced these datasets are. The datasets had many duplicates which we removed to avoid data leakage between training and test sets.

A newly proposed Autshumato evaluation set for machine translation of South African languages has been released (McKellar, 2017). The evaluation set consists of 500 sentences translated separately by four different professional human translators for each of the 11 official South African languages. We evaluated our models on this benchmark.

## 5 Architectures

Limited work has been done using NMT techniques for African languages. We trained translation models for two popular NMT architectures, namely, ConvS2S (Gehring et al., 2017) and Transformer (Vaswani et al., 2017)<sup>3</sup>.

## 6 Results

The BLEU scores for each model on English-to-Target language are presented in Table 2. The Transformer model outperformed the ConvS2S model for all languages. These results serve as initial baseline results for the given languages on the evaluation set.

| <b>Model</b> | <b>Afrikaans</b> | <b>isiZulu</b> | <b>Northern Sotho</b> | <b>Setswana</b> | <b>Xitsonga</b> |
|--------------|------------------|----------------|-----------------------|-----------------|-----------------|
| ConvS2S      | 12.30            | 0.52           | 7.41                  | 10.31           | 10.73           |
| Transformer  | <b>20.60</b>     | <b>1.34</b>    | <b>10.94</b>          | <b>15.60</b>    | <b>17.98</b>    |

Table 2: BLEU Scores for English-to-Target language on evaluation dataset

We supply a few qualitative sample translations in Table 3 and Table 4, as well as multi-head attention visualisations, to demonstrate the performance of the models.

We notice that the performance of techniques on a specific target language is related to corpus size and morphological typology of the language. Afrikaans is not agglutinative, thus despite having less than half the number of sentences as Xitsonga and Setswana, the model achieves better performance than the other languages. Xitsonga and Setswana are both agglutinative, but have much more data, so their models achieve higher performance than Northern Sotho or isiZulu. The models for isiZulu achieved the worst performance. We attribute the bad performance to the morphological complexity of the language (discussed in Section 2), the size of the dataset and poor quality of the data.

<sup>2</sup>Available online at: <https://repo.sadilar.org/handle/20.500.12185/404>

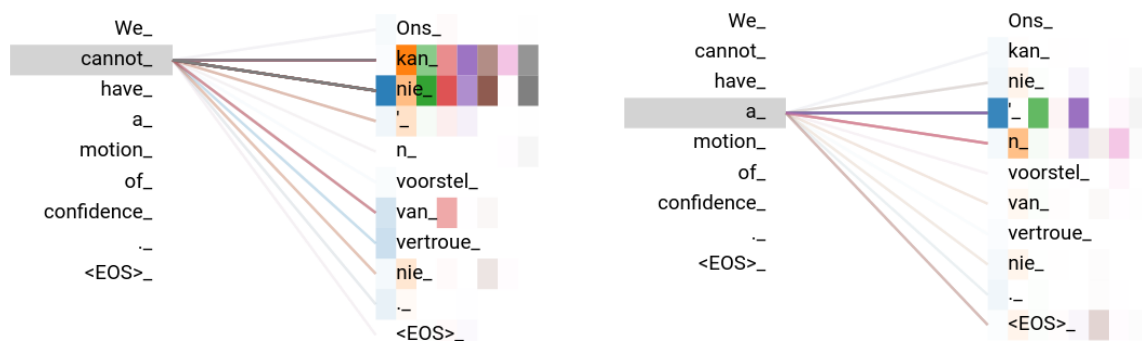
<sup>3</sup>The hyper-parameters used for each architecture are available online here: <https://github.com/LauraMartinus/ukuxhumana>

Table 3: **English to isiZulu Translations:** We show the reference translation, translation by the Transformer model, and translation back to English performed by an isiZulu speaker.

|                  |  |
|------------------|--|
| Source           | Note that the funds will be held against the Vote of the Provincial Treasury pending disbursement to the SMME Fund . |
| Target           | Lemali izohlala emnyangweni wezimali .   |
| Transformer      | Qaphela ukuthi izimali zizobanjwa kweVME esifundazweni saseTreasury zezifo ezithunyelwa ku-MSE .                     |
| Back Translation | Be aware that the money will be held by VME with facilities of Treasury with diseases sent to MSE .                  |

Table 4: **English to Xitsonga Translations:** We show the reference translation, translation by the Transformer model, and translation back to English performed by a Xitsonga speaker.

|                  |  |
|------------------|--|
| Source           | we are concerned that unemployment and poverty persist despite the economic growth experienced in the past 10 years .                                      |
| Target           | hi na swivilelo leswaku mpfumaleko wa mitirho na vusweti swi ya emahlweni hambileswi ku nga va na ku kula ka ikhonomi eka malembe ya 10 lawa ya hundzeke . |
| Transformer      | hi na swivilelo leswaku mpfumaleko wa mitirho na vusweti swi ya emahlweni hambileswi ku nga va na ku kula ka ikhonomi eka malembe ya 10 lawa ya hundzeke . |
| Back Translation | We have concerns that there is still lack of jobs and poverty even though there has been economic growth in the past 10 years.                             |



(a) Visualization of multi-head attention for Layer 1 for the word “cannot”. The coloured bars are individual attention heads. The word “cannot” is translated to “kan nie ... nie” where the second negative “nie” occurs at the end of the sentence.

(b) Visualization of multi-head attention for Layer 2 for the word “a”. The coloured bars are individual attention heads. The word “a” is translated to “n”, as is successfully captured by the attention mechanism.

Figure 1: Visualizations of multi-head attention for an English sentence translated to Afrikaans using the Transformer model.

## 7 Conclusion and Future Work

By publishing the code, datasets and baseline results on a discoverable platform, for machine translation of the above languages and evaluating them on a publicly-available evaluation set, we begin to address the problems of discoverability, reproducibility, and comparability in the research of machine translation of African languages. Future work will include results for other Southern African languages so we can provide baselines for all official languages in South Africa, by using unsupervised and transfer-learning NMT techniques.

## Acknowledgements

The authors would like to thank Mbongiseni Ncube, Reinhard Cromhout, and Vongani Maluleke for assisting us with the back-translations. Research supported with Cloud TPUs from Google’s TensorFlow Research Cloud (TFRC).

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

## References

- Jade Z Abbott and Laura Martinus. 2018. Towards neural machine translation for African languages. *arXiv preprint arXiv:1811.05467*.
- Mary C. Bill. 1984. 100 years of Tsonga publications, 1883-1983. *African Studies*, 43(2):67–81.
- Victor Chahuneau, Eva Schlinger, Noah A Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687.
- David M Eberhard, Gary F. Simons, and Charles D. Fennig. 2019. *Ethnologue: Languages of the worlds*. twenty-second edition.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.
- Hendrik J Groenewald and Wildrich Fourie. 2009. Introducing the Autshumato integrated translation environment. In *Proceedings of the 13th Annual Conference of the EAMT, Barcelona, May*, pages 190–196. Citeseer.
- C Maria Keet and Langa Khumalo. 2017. Grammar rules for the isizulu complex verb. *Southern African Linguistics and Applied Language Studies*, 35(2):183–200.
- Laura Martinus and Jade Z. Abbott. 2019. A Focus on Neural Machine Translation for African Languages. *arXiv e-prints*, page arXiv:1906.05685, Jun.
- Cindy A. McKellar. 2014. An English to Xitsonga statistical machine translation system for the government domain. In *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*, pages 229–233.
- Cindy Arlene McKellar. 2017. Autshumato machine translation evaluation set.
- Rajend Mesthrie and Mesthrie Rajend. 2002. *Language in South Africa*. Cambridge University Press.
- Paul T Roberge. 2002. Afrikaans: considering origins. *Language in South Africa*, pages 79–103.
- Daniel R van Niekerk. 2014. Exploring unsupervised word segmentation for machine translation in the South African context. In *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*, pages 202–206.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Ilana Wilken, Marissa Griesel, and Cindy McKellar. 2012. Developing and improving a statistical machine translation system for English to Setswana: a linguistically-motivated approach. In *Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, page 114.
- Sabine Zerbian. 2007. A first approach to information structuring in Xitsonga/Xichangana. *Research in African Languages and Linguistics*, 7(2005-2006):1–22.