# NGLUEni: Benchmarking and Adapting Pretrained Language Models for Nguni Languages

**Francois Meyer**[2], **Haiyue Song**[1], **Abhisek Chakrabarty**[1],
**Jan Buys**[2], **Raj Dabre**[1], **Hideki Tanaka**[1]
[1]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[2]Department of Computer Science, University of Cape Town, South Africa
{francois.meyer, jan.buys}@uct.ac.za
{haiyue.song, abhisek.chakra, raj.dabre,
hideki.tanaka}@nict.go.jp

## Abstract

The Nguni languages have over 20 million home language speakers in South Africa. There has been considerable growth in datasets for Nguni languages, but no analysis of performance of NLP models for these languages has been reported across all languages and tasks. In this paper we study pretrained language models for the 4 Nguni languages - isiXhosa, isiZulu, isiNdebele, and Siswati. We compile all publicly available datasets for natural language understanding and generation, spanning 6 tasks and 11 datasets. This benchmark, which we call NGLUEni, is the first centralised evaluation suite for the Nguni languages, allowing us to systematically evaluate the Nguni-language capabilities of PLMs. Besides evaluating existing PLMs, we develop new PLMs for the Nguni languages through multilingual adaptive finetuning. Our models, Nguni-XLMR and Nguni-ByT5, outperform their base models and large-scale adapted models, showing that performance gains are obtainable through limited language group-based adaptation. We also perform experiments on cross-lingual transfer and machine translation. Our models achieve notable cross-lingual transfer improvements in the lower resourced Nguni languages (isiNdebele and Siswati). To facilitate future use of NGLUEni as a standardised evaluation suite for the Nguni languages, we create a web portal to access the collection of datasets and publicly release our models.

## 1 Introduction

Multilingual pretrained language models (PLMs) have revolutionised NLP for low-resource languages. It enables cross-lingual transfer from high-resource languages, as demonstrated by models like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), and ByT5 (Xue et al., 2022). Models are now pretrained on over 100 languages, reflecting a trend towards increasingly multilingual PLMs. While this has undoubtedly benefited low-resource languages, it also has certain shortcomings.

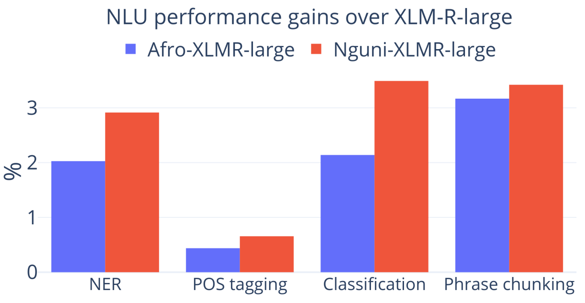

Figure 1: We adapt XLM-R for Nguni languages and plot proportional performance gains over Afro-XLMR and XLM-R averaged across the Nguni languages.

A major limitation is the lack of proper multilingual evaluation. Recent works might train PLMs for over 100 languages, but they do not perform downstream evaluation for all these languages because many lack sufficient datasets. As a result, the true low-resource language capabilities of multilingual PLMs are unknown.

Another limitation of scaling PLMs to more languages is that it sacrifices greater gains on individual languages for gains across more languages. While models like XLM-R and ByT5 achieve impressive results across many languages, PLMs focused on smaller groups of languages can outperform more multilingual models (Alabi et al., 2022; Dabre et al., 2022; Ebrahimi et al., 2022). The strategy of increasing multilinguality is not optimal if we are focused on improving performance for specific low-resource languages.

| Task | Dataset | xh | zu | nr | ss | Size |
|---|---|---|---|---|---|---|
| **Natural language understanding (NLU)** | | | | | | |
| NER | MasakhaNER | ✓ | ✓ | | | 5783 |
| | SADiLaR NER | ✓ | ✓ | ✓ | ✓ | 6520 |
| POS tagging | MasakhaPOS | ✓ | ✓ | | | 753 |
| | NLAPOST | ✓ | ✓ | ✓ | ✓ | 2717 |
| Classification | MasakhaNEWS | ✓ | | | | 1032 |
| | ANTC | ✓ | | | | 2961 |
| | NCHLT Genre | ✓ | ✓ | ✓ | ✓ | 1919 |
| Phrase chunk | NCHLT PC | ✓ | ✓ | ✓ | ✓ | 848 |
| **Natural language generation (NLG)** | | | | | | |
| Data-to-text | T2X | ✓ | | | | 3859 |
| Headline generation | MasakhaNEWS | ✓ | | | | 1032 |
| | Vuk'uzenzele | ✓ | ✓ | ✓ | ✓ | 149 |

Table 1: The NGLUEni evaluation suite language coverage and average training size per language.

In this work we address these limitations of multilingual PLMs for the Nguni languages of South Africa - isiXhosa (xh), isiZulu (zu), isiNdebele (nr), and Siswati (ss). They are a related group of languages with a combined 23.4 million L1 speakers (Eberhard et al., 2019) (see Table 7 for language statistics), comprising over 43% of the population of South Africa. They are linguistically distinct in that they have highly agglutinative morphologies, which can hurt cross-lingual transfer from other languages (Wang et al., 2021; Ács, 2019). While isiXhosa and isiZulu have increasingly more datasets available, isiNdebele and Siswati remain extremely low-resourced. This paper is the first systematic investigation into multilingual PLMs for Nguni languages. Our contributions span benchmark compilation, modeling and evaluation.

We compile NGLUEni, a benchmark for Nguni languages. It covers natural language understanding (NLU) and generation (NLG). For NLU we collect publicly available datasets, covering 4 tasks across 8 datasets. The situation is more challenging for NLG. Besides machine translation (MT) the only dataset for Nguni NLG is T2X (Meyer & Buys, 2024), an isiXhosa data-to-text dataset. To expand NLG we repurpose news article datasets (Lastrucci et al., 2023; Adelani et al., 2023) for the task of headline generation, which provides a starting point for evaluating Nguni text generation models. The full NGLUEni benchmark is summarised in Table 1.

Beyond facilitating standardized evaluation, the other aim of this work is to develop PLMs tailored specifically for the Nguni languages. For this we turn to multilingual adaptive finetuning (Alabi et al., 2022), wherein existing PLMs are subjected to continued training on a Nguni-only corpus. For NLU we adapt XLM-R-large to produce Nguni-XLMR-large, which improves average performance across the 4 Nguni languages for all NLU tasks (as shown in Figure 1). It achieves gains of up to 5.2 F1% on text classification. For NLG we adapt ByT5-large to produce Nguni-ByT5-large, which improves performance on isiXhosa data-to-text and headline generation. It outperforms its baselines across 7 automatic metrics, with BLEU gains of 2.0 on data-to-text. We perform additional experiments on zero-shot cross-lingual transfer and MT. Our Nguni-adapted models achieve particularly large gains in cross-lingual transfer to isiNdebele and Siswati.

Adapting PLMs for the Nguni languages improves performance on a diverse set of tasks. This is achieved by continued training on only the 4 Nguni languages, which is more efficient than pretraining from scratch or multilingual adaptive finetuning on a larger set of languages. In summary, our contributions are as follows:

- We compile **NGLUEni**, an evaluation suite spanning NLU and NLG tasks for evaluating the Nguni-language capabilities of PLMs. The NGLUEni datasets can be accessed through a centralised repository: `https://github.com/francois-meyer/nglueni`.
- We adapt **Nguni-XLMR**[1] and **Nguni-ByT5**[2], which improve Nguni-language performance. We publicly release these adapted models on Hugging Face.
- We evaluate our models, their base PLMs, and existing adaptation-based baselines on NGLUEni, cross-lingual experiments, and machine translation.

[1] `https://huggingface.co/nict-astrec-att/nguni-xlmr-large`

[2] `https://huggingface.co/nict-astrec-att/nguni-byt5-large`

## 2 Related Work

**Nguni benchmarks** Several Nguni NLU datasets exist, but only a few have been used to evaluate PLMs (and only for isiXhosa and isiZulu). Masakhane[3] has produced African language datasets for NER (Adelani et al., 2022c), POS tagging (Dione et al., 2023), and news classification (Adelani et al., 2023). These datasets cover up to 20 African languages, including isiXhosa and/or isiZulu. The respective papers also benchmark many existing multilingual PLMs on the datasets. Likewise, Alabi et al. (2022) evaluate several PLMs on their ANTC dataset, a news topic classification task for 5 African languages including isiZulu. On ANTC and the Masakhane datasets, Afro-XLMR (Alabi et al., 2022) generally emerges as the strongest multilingual PLM overall.

**Multilingual PLMs** A few massively multilingual PLMs (more than 100 languages) include isiXhosa or isiZulu in their pretraining. Among encoder-only PLMs, XLM-R (Conneau et al., 2020) includes isiXhosa. Among encoder-decoder PLMs, mT5 (Xue et al., 2021) and ByT5 (Xue et al., 2022) include isiXhosa and isiZulu. None of these works evaluate on any of the Nguni languages.

**PLMs for African languages** AfroLM (Dossou et al., 2022) is trained from scratch with active learning on 23 African languages, including isiXhosa and isiZulu. Afro-XLMR (Alabi et al., 2022) adapt XLM-R for 17 African languages, including isiXhosa and isiZulu. They achieve this through multilingual adaptive finetuning, which amounts to taking model checkpoints and continuing their pretraining on a corpus of selected languages.

Adelani et al. (2022a) use the same approach for encoder-decoder PLMs, where adaptation amounts to continued span denoising training. They adapt mT5 and ByT5 for 17 African languages, including isiXhosa and isiZulu, to produce respectively AfriMT5 and AfriByT5. Their work is focused on MT (including isiZulu $\leftrightarrow$ English), so they do not evaluate these models on any other NLG tasks. In the MT experiments, they find that multilingual adaptation improves performance and that AfriByT5 outperforms AfriMT5.

## 3 Data

We collect publicly available data for training and evaluating Nguni PLMs. Some of the datasets require cleaning and splitting into train/valid/test sets. We also preprocess news article datasets to use them for headline generation.

### 3.1 Adaptation Corpus

To adapt PLMs we continue their language model training on a Nguni-only corpus, optimising their original pretraining objective. We create a Nguni corpus of 260,950,000 tokens by combining publicly available monolingual corpora for all 4 languages. The sizes of the respective monolingual corpora are shown in the last row of Table 7, which also highlights that our adapted models are the first to be trained on isiNdebele and Siswati.

For isiXhosa and isiZulu we use data from the mC4 corpus (Xue et al., 2021). These corpora are web crawled, so while their large sizes are beneficial to training, it comes with the cost of lower quality data (Kreutzer et al., 2022). For isiNdebele and Siswati we include their NCHLT Text corpora (Eiselen & Puttkammer, 2014), which consist of South African government documents, articles covering various domains, and prose.

Table 7 highlights the imbalance in data availability among the Nguni languages, with isiXhosa and isiZulu having much larger corpora than isiNdebele and Siswati. To alleviate this we upsample data from isiNdebele and Siswati using the same multinomial sampling distribution used to train XLM-R (Conneau et al., 2020) with $\alpha = 0.3$.

[3] https://www.masakhane.io/

### 3.2 Natural Language Understanding (NLU) Evaluation

Our NLU benchmark spans 4 tasks and 8 datasets. Some datasets cover only isiXhosa or isiZulu, but for each task we have at least one dataset covering all 4 languages, enabling cross-lingual experiments. The SADiLaR and NCHLT datasets are publicly available as raw annotated datasets (not separated into training and evaluation sets), so we split these into train/valid/test sets (80%/10%/10%).

**Named entity recognition (NER)** Masakha-NER (Adelani et al., 2022c) covers 20 African languages including isiXhosa and isiZulu. The annotated text is extracted from newspaper articles. SADiLaR NER (Eiselen, 2016a) contains annotated government domain text for 10 South African languages, including all 4 Nguni languages.

**Part-of-speech (POS) tagging** MasakhaPOS (Dione et al., 2023) covers 20 African languages including isiXhosa and isiZulu, for which it contains tagged news articles. NLAPOST is part of a collection of annotated government text in all 4 Nguni languages (Gaustad & Puttkammer, 2022) and was used for a shared task on Nguni POS tagging (Pannach et al., 2022).

**Text classification** MasakhaNEWS (Adelani et al., 2023) and ANTC (Alabi et al., 2022) are news topic classification datasets for respectively 16 and 5 African languages. MasakhaNEWS includes isiXhosa and ANTC includes isiZulu, in each case labelling news articles into one of 5 possible categories (e.g. sports, politics). NCHLT Genre (Snyman et al., 2011) is a collection of government articles in 10 South African languages. For all 4 Nguni languages each article is labelled as one of 3 categories (non-fiction neutral, non-fiction subjective, or non-fiction objective).

**Phrase chunking** Phrase chunking (Abney, 1992) assigns words to non-recursive multi-word segments of major syntactic categories (e.g. noun phrase, verb phrase). It is an intermediary between POS and syntax trees. NCHLT Phrase Chunking (Eiselen, 2016b) is a dataset of phrase-annotated government text for 10 South African languages, including all 4 Nguni languages.

### 3.3 Natural Language Generation (NLG) Evaluation

The options for evaluating Nguni NLG are limited to data-to-text. To improve the situation we adapt 2 datasets for the task of headline generation. Additionally, we finetune models for MT.

**Data-to-text** T2X (Meyer & Buys, 2024) is an isiXhosa data-to-text dataset. It contains triples of (subject, relation, object) mapped to descriptive text in isiXhosa e.g. (France, currency, Euro) → "Imali yaseFransi yi-Euro" ("The currency of France is the Euro"). It was constructed by translating a subset of the English WebNLG dataset (Gardent et al., 2017). It is the only existing text generation task dataset for a Nguni language.

**Headline generation** Generating headlines based on articles can be framed as a sequence-to-sequence summarisation task (Rush et al., 2015). Many existing datasets for Nguni languages contain news articles, so we leverage this to create headline generation datasets. The MasakhaNEWS data (Adelani et al., 2023) contains separate columns for article text and headline, which we extract to form text-headline pairs for an isiXhosa headline generation task.

Vuk'uzenzele (Lastrucci et al., 2023) contains government news articles in all 4 Nguni languages. We automatically extract article-headline pairs and manually remove erroneously processed examples. The datasets are too small for finetuning (around 150 examples per language), so we only use them to evaluate models finetuned on MasakhaNEWS headline generation. MasakhaNEWS and Vuk'uzenzele cover different domains and languages, so this tests cross-domain and zero-shot cross-lingual performance.

**Machine translation (MT)** In the absence of more NLG datasets, MT presents additional opportunities to evaluate Nguni text generation capabilities. We use 2 multilingual MT datasets: translation from English to Nguni languages with the Autshumato dataset (McKellar & Puttkammer, 2020) and translation between Nguni languages with WMT22 (Adelani et al., 2022b).

# 4 Models

We follow the multilingual adaptive finetuning of Alabi et al. (2022), who adapt PLMs for 17 African languages. Our approach is narrower in linguistic scope, focusing on the Nguni languages. Since they are related, we expect a high degree of cross-lingual transfer during adaptation. This could especially benefit the lower resourced languages of isiNdebele and Siswati.

We perform experiments for our Nguni-PLMs, their original unadapted PLMs, as well as variants that have been adapted for a larger set of African languages. This allows us to compare PLMs at various stages: (1) no adaptation, (2) adapted for a diverse set of African languages, and (3) adapted for the Nguni languages.

## 4.1 NLU

### 4.1.1 Baselines

XLM-R (Conneau et al., 2020) scales cross-lingual masked language modelling (MLM) to 100 languages including isiXhosa. We adapt XLM-R-large and use the original model as a baseline. We also use Afro-XLMR-large (Alabi et al., 2022) as a baseline, which adapts XLM-R-large for 17 African languages including isiXhosa and isiZulu.

### 4.1.2 Nguni-XLMR-large

We adapt XLM-R-large for the Nguni languages through continued MLM training. Nguni-XLMR-large expands the multilingual knowledge of XLM-R-large to cover all the Nguni languages. In the original XLM-R training, isiXhosa made up a very small proportion of the multilingual corpus. Our Nguni adaptation corpus contains a much greater proportion of isiXhosa and exposes the model to isiZulu, isiNdebele, and Siswati for the first time.

## 4.2 NLG

### 4.2.1 Baselines

ByT5 (Xue et al., 2022) is a text-to-text PLM for 101 languages, including isiXhosa and isiZulu. It is trained with the span denoising objective of mT5 (Xue et al., 2021), but text is processed as sequence of UTF-8 bytes instead of subwords. We chose to adapt ByT5 because it has been shown to comfortably outperform subword-based NLG models on low-resource languages (Edman et al., 2023; Adelani et al., 2022a). We also evaluate Afri-ByT5-base (Adelani et al., 2022a) as a baseline, which adapts ByT5-base for 17 African languages including isiXhosa and isiZulu.

### 4.2.2 Nguni-ByT5-large

We adapt ByT5-large for the Nguni languages through continued span denoising training. ByT5 includes isiXhosa and isiZulu in its training data, but they make up very small proportions of its multilingual corpus. Nguni-ByT5-large builds on the isiXhosa and isiZulu knowledge of ByT5-large and is the first text-to-text multilingual PLM trained on isiNdebele and Siswati data.

# 5 Experimental Setup

We use Huggingface Transformers (Wolf et al., 2020) to adapt XLM-R-large and ByT5-large. Nguni-XLMR-large is adapted for 600k training steps with a learning rate of 5e-5, no warmup steps, and a batch size of 80. Nguni-ByT5-large is adapted for 10k training steps with a learning rate of 1e-4, 5k warmup steps, and batch size of 1024. Validation performance plateaued after the reported number of training steps. Our training is distributed across across 8 Tesla V100 GPUs.

## 5.1 NLU

We finetune models on the NLU datasets listed in Section 3.2, using finetuning scripts from previous works where available (Adelani et al., 2022c; Dione et al., 2023; Adelani et al., 2023; Alabi et al.,

| Task | Dataset | lang | XLM-R-large | Afro-XLMR-large | **Nguni-XLMR-large** |
|---|---|---|---|---|---|
| NER | MasakhaNER | xh | 88.1 | 89.9 | **90.4**$_{\pm 0.004}$ |
| | | zu | 86.7 | 90.6 | **91.8**$_{\pm 0.006}$ |
| | SADiLaR NER | xh | 74.8$_{\pm 0.7}$ | 76.3$_{\pm 0.9}$ | **77.3**$_{\pm 0.5}$ |
| | | zu | 73.6$_{\pm 0.3}$ | 74.1$_{\pm 0.6}$ | **74.3**$_{\pm 0.4}$ |
| | | nr | 78.6$_{\pm 0.2}$ | **79.4**$_{\pm 0.4}$ | 79.1$_{\pm 0.7}$ |
| | | ss | 71.8$_{\pm 0.6}$ | 72.8$_{\pm 0.4}$ | **74.1**$_{\pm 0.7}$ |
| POS | MasakhanePOS | xh | 88.1 | **88.7** | 88.3$_{\pm 0.1}$ |
| | | zu | 89.4 | **90.1** | **90.1**$_{\pm 0.1}$ |
| | NLAPOST | xh | 97.1$_{\pm 0.1}$ | 97.8$_{\pm 0.1}$ | **97.9**$_{\pm 0.1}$ |
| | | zu | 92.5$_{\pm 0.2}$ | 92.9$_{\pm 0.2}$ | **93.3**$_{\pm 0.1}$ |
| | | nr | 90.3$_{\pm 0.1}$ | 90.5$_{\pm 0.1}$ | **90.6**$_{\pm 0.2}$ |
| | | ss | 90.9$_{\pm 0.3}$ | 91.0$_{\pm 0.1}$ | **91.6**$_{\pm 0.3}$ |
| Classification | MasakhaNEWS | xh | 89.2 | 97.3 | **98.2**$_{\pm 0.5}$ |
| | ANTC | zu | 78.7 | 81.6$_{\pm 1.4}$ | **86.8**$_{\pm 0.6}$ |
| | NCHLT Genre | xh | **89.1**$_{\pm 0.9}$ | 89.0$_{\pm 1.0}$ | 88.8$_{\pm 0.6}$ |
| | | zu | 82.8$_{\pm 1.4}$ | 84.9$_{\pm 1.2}$ | **86.5**$_{\pm 1.7}$ |
| | | nr | **96.4**$_{\pm 2.6}$ | 94.9$_{\pm 0.6}$ | 95.2$_{\pm 0.6}$ |
| | | ss | 96.3$_{\pm 1.4}$ | **96.7**$_{\pm 0.8}$ | 96.0$_{\pm 0.6}$ |
| Phrase chunking | NCHLT PC | xh | 88.2$_{\pm 0.6}$ | 90.1$_{\pm 0.7}$ | **91.0**$_{\pm 0.4}$ |
| | | zu | 87.8$_{\pm 0.7}$ | 90.2$_{\pm 0.1}$ | **90.5**$_{\pm 0.3}$ |
| | | nr | 56.0$_{\pm 1.2}$ | **59.5**$_{\pm 1.0}$ | 58.4$_{\pm 1.0}$ |
| | | ss | 83.4$_{\pm 0.3}$ | 84.3$_{\pm 0.4}$ | **84.8**$_{\pm 0.3}$ |

Table 2: NLU test performance averaged across 5 runs. We include standard deviation for finetuning run ourselves but omit it for results taken from existing papers (since it was not reported). For NER, classification, and phrase chunking we report weighted F1, while for POS we report accuracy.

2022) and adapting these for datasets that have not previously been used for finetuning (SADiLaR NER, NLAPOST, NHCLT Genre & PC). We report test set results averaged across 5 finetuning runs. Our NLU finetuning hyperparameters are specified in Appendix B.1.

**Cross-lingual experiments** To evaluate zero-shot cross-lingual transfer we use the 4 datasets that cover all 4 Nguni languages (SADiLaR NER, NLAPOST, NCHLT Genre, NCHLT PC). For each task we finetune models on the isiXhosa training sets and evaluate them on the test sets of the other languages. We chose to evaluate transfer from isiXhosa to the other languages, since this is the most realistic real-world use case. IsiXhosa is featured in datasets more often than the other Nguni languages, so being able to apply models trained on isiXhosa to other languages would be valuable.

### 5.2 NLG

Our NLG finetuning hyperparameters are specified in Appendix B.2. We are the first to use MasakhaNEWS (Adelani et al., 2023) for headline generation. It is much harder than T2X because of the less structured nature of the task and the smaller training set, so we run a grid search across several hyperparameter combinations to find optimal finetuning settings.

**Cross-lingual experiments** The Vuk'uzenzele (Lastrucci et al., 2023) datasets are too small to split intro train/test sets so instead of using them for finetuning and evaluation, we use the full datasets as evaluation sets. We evaluate our final MasakhaNEWS headline generation model. MasakhaNEWS covers only isiXhosa, while Vuk'uzenzele covers all 4 Nguni languages. Therefore we present the results on Vuk'uzenzele as cross-domain and cross-lingual experiments. It tests the ability of our models to generalise their headline generation capabilities to the previously unseen domain of Vuk'uzenzele (government news articles). With the isiZulu, isiNdebele, and Siswati parts of Vuk'uzenzele, it also tests cross-lingual transfer from isiXhosa.

| Task | Dataset | lang | XLM-R-large | Afro-XLMR-large | **Nguni-XLMR-large** |
|---|---|---|---|---|---|
| NER | SADiLaR NER | zu | 30.3$_{\pm0.5}$ | 32.4$_{\pm0.2}$ | **32.6**$_{\pm0.6}$ |
| | | nr | 31.0$_{\pm0.7}$ | 36.3$_{\pm0.7}$ | **37.0**$_{\pm0.8}$ |
| | | ss | 31.4$_{\pm0.9}$ | 37.3$_{\pm1.6}$ | **41.1**$_{\pm1.1}$ |
| POS | NLAPOST | zu | 87.1$_{\pm0.3}$ | 88.5$_{\pm0.2}$ | **88.6**$_{\pm0.1}$ |
| | | nr | 77.6$_{\pm0.3}$ | 79.8$_{\pm0.2}$ | **79.9**$_{\pm0.2}$ |
| | | ss | 70.7$_{\pm0.3}$ | 83.7$_{\pm0.5}$ | **87.4**$_{\pm0.2}$ |
| Classification | NCHLT Genre | zu | 67.0$_{\pm5.9}$ | **70.4**$_{\pm2.7}$ | 66.0$_{\pm6.1}$ |
| | | nr | 45.7$_{\pm17.2}$ | 80.3$_{\pm8.2}$ | **80.9**$_{\pm7.7}$ |
| | | ss | 69.2$_{\pm16.2}$ | 86.9$_{\pm6.5}$ | **88.8**$_{\pm2.6}$ |
| Phrase chunking | NCHLT PC | zu | 63.4$_{\pm0.7}$ | 65.5$_{\pm0.3}$ | **66.4**$_{\pm0.6}$ |
| | | nr | 35.6$_{\pm0.5}$ | 38.2$_{\pm0.4}$ | **39.0**$_{\pm0.5}$ |
| | | ss | 61.5$_{\pm0.8}$ | 68.0$_{\pm0.6}$ | **68.9**$_{\pm0.8}$ |

Table 3: Zero-shot cross-lingual NLU test performance after finetuning on isiXhosa and evaluating on isiZulu, isiNdebele, and Siswati. Results are averaged across 5 runs (± standard deviation). For NER, classification, and phrase chunking we report weighted F1, while for POS we report accuracy.

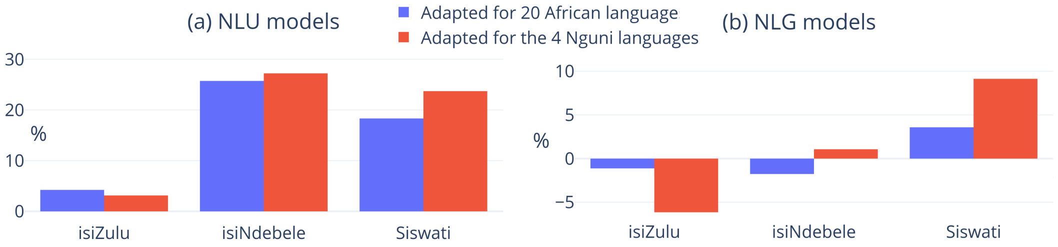

Figure 2: Improvements in zero-shot cross-lingual transfer (xh → zu, nr, ss) after adaptation. Figure (a) compares improvement of Afro-XLMR-large and Nguni-XLMR-large over XLM-R-large, averaged over the 4 tasks in Table 3. Figure (b) shows how chrF++ changes for Vuk'uzenzele headline generation after adapting ByT5-base to Afri-ByT5-base and ByT5-large to Nguni-ByT5-large.

**MT experiments** As an additional text generation task we finetune our NLG models on MT. We finetune each model twice - on multilingual translation from English to (isiZulu, isiNdebele, Siswati) and in all directions between (isiXhosa, isiZulu, Siswati).

## 6 RESULTS

We report the results of our finetuning experiments in Tables 2-6 and plot the improvements obtained through adaptation in Figures 1 and 2.

### 6.1 NLU

Table 2 reports all NLU results. **In most instances Nguni-XLMR outperforms Afro-XLMR, which in turn outperforms XLM-R.** This holds with average task scores over all Nguni languages, as shown in Figure 1. Nguni-XLMR does well on NER, outperforming both baselines in all instances except one. POS tagging sees smaller gains, with XLM-R already achieving accuracies that outperform the best NLAPOST2021 models (Pannach et al., 2022). Classification results vary between datasets. Nguni-XLMR achieves its largest gains on MasakhaneNEWS and ANTC (it outperforms Afro-XLMR by 5.2 percentage points on ANTC). The NCHLT Genre classification datasets are considerably easier, so XLM-R is competitive. Phrase chunking sees the largest average performance gain of any task, suggesting again that harder tasks (phrase chunking is somewhere between POS tagging and syntactic parsing) stand to benefit more from adaptation.

Afro-XLMR outperforms XLM-R across all 4 Nguni languages, even though it is only adapted for isiXhosa and isiZulu. However, the fact that Nguni-XLMR outperforms Afro-XLMR shows that

| Dataset | lang | Model | chrF++ | chrF | BLEU | NIST | MET | ROU | CID |
|---|---|---|---|---|---|---|---|---|---|
| T2X | xh | ByT5-base | 46.17 | 51.16 | 17.62 | 4.32 | 22.60 | 35.69 | 1.21 |
| | | Afri-ByT5-base | 51.95 | 56.67 | 22.56 | 5.21 | 26.11 | 44.67 | 1.64 |
| | | ByT5-large | 44.16 | 48.87 | 15.95 | 4.08 | 21.92 | 34.25 | 1.13 |
| | | Nguni-ByT5-large | **52.96** | **57.71** | **24.56** | **5.34** | **26.61** | **45.09** | **1.71** |
| Masakha–NEWS | xh | ByT5-base | 20.65 | 25.49 | 1.89 | 0.87 | 4.63 | 9.09 | 0.29 |
| | | Afri-ByT5-base | 20.21 | 24.87 | 1.41 | 1.02 | 4.77 | 9.94 | 0.31 |
| | | ByT5-large | 19.36 | 24.23 | 1.41 | 0.83 | 3.98 | 8.14 | 0.26 |
| | | Nguni-ByT5-large | **21.91** | **26.93** | **2.06** | **1.06** | **5.25** | **10.74** | **0.36** |

Table 4: Test performance on isiXhosa NLG tasks (T2X data-to-text and MasakhaNEWS headline generation), measured across different automatic metrics.

| Cross– | | lang | ByT5-base | | Afri-ByT5-base | | ByT5-large | | Nguni-ByT5-large | |
|---|---|---|---|---|---|---|---|---|---|---|
| lingual | domain | | BLEU | chrF | BLEU | chrF | BLEU | chrF | BLEU | chrF |
| ✗ | ✓ | xh | 0.58 | 23.54 | 0.35 | 22.62 | **0.95** | **24.74** | 0.68 | 24.31 |
| ✓ | ✓ | zu | 0.58 | 24.03 | 0.56 | 23.76 | 0.33 | **25.52** | **0.79** | 23.95 |
| ✓ | ✓ | nr | 0.32 | 20.91 | 0.51 | 20.54 | 0.49 | 21.54 | **0.51** | **21.77** |
| ✓ | ✓ | ss | 0.66 | 22.88 | 0.59 | 23.70 | **0.77** | 22.77 | 0.34 | **24.85** |

Table 5: Zero-shot headline generation test performance after finetuning on isiXhosa Masakhane-NEWS and evaluating on Vuk'uzenzele in all 4 Nguni languages.

there is substantial negative interference (Wang et al., 2020) from the other languages in the Afro-XLMR adaptation corpus. **Nguni-XLMR achieves cross-lingual transfer between the Nguni languages without interference from additional languages.**

**Zero-shot cross-lingual** Table 3 shows the results of our zero-shot cross-lingual NLU experiments. Nguni-XLMR achieves the best scores in all instances except one, demonstrating effective cross-lingual transfer capabilities on all NLU tasks. To compare transfer capabilities across the 3 target languages, we plot the relative improvement of Afro-XLMR and Nguni-XLMR over XLM-R averaged over all tasks in Figure 2(a).

**The adapted models greatly improve transfer to isiNdebele and Siswati**, with Nguni-XLMR improving zero-shot test scores in both languages by more than 20%. Nguni-XLMR sees larger gains than Afro-XLMR, which can be attributed to the inclusion of isiNdebele and Siswati in its adaptation corpus. The inclusion of (albeit small) amounts of text from these languages proves better at inducing cross-lingual transfer than the more multilingual adaptation of Afro-XLMR, which covers more languages but excludes isiNdebele and Siswati.

## 6.2 NLG

Table 4 reports isiXhosa NLG results. **On both tasks Nguni-ByT5-large outperforms all baselines across all 7 metrics.** On T2X the gains are substantial, with gains of by 2.0 BLEU points over the second best baseline. MasakhaneNEWS headline generation is a more challenging task than T2X, so the metrics are much lower. In this case the character-based chrF and chrF++ are more informative than BLEU, because they measure subword-level overlap which is more relevant for the agglutinative Nguni languages. Nguni-ByT5-large outperforms all baselines on these metrics and achieves chrF++ scores comparable to the scores of state-of-the-art MT for certain low-resource African languages (Team et al., 2022).

The base variant of ByT5 outperforms the large variant on both tasks. We performed separate hyperparameter tuning for all models, so we do not believe this to be due to suboptimal hyperparameters. It is possible that the large models are overfitting to the small training sets. Comparing Afri-ByT5-base and Nguni-ByT5-large to their base PLMs, we fine that the Nguni-only adaptation leads to greater increases in performance (proportionally and in absolute terms). Nguni-ByT5-large might

| Train/Test | langs | ByT5-base | | Afri-ByT5 -base | | ByT5-large | | Nguni-ByT5 -large | |
|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | chrF | BLEU | chrF | BLEU | chrF | BLEU | chrF |
| Autshumato | en→zu | 8.79 | 44.96 | 9.90 | 45.89 | **10.51** | 47.03 | **10.55** | **47.63** |
| | en→ss | 9.15 | 49.36 | **9.50** | 50.00 | **9.95** | **50.64** | **9.92** | **50.72** |
| | en→nr | 8.94 | 48.01 | 9.25 | 48.71 | **10.00** | **50.04** | **10.41** | **50.21** |
| WMT/ FLORES | xh→zu | 12.59 | 49.68 | 12.65 | 49.64 | **13.34** | **50.30** | **13.43** | **50.31** |
| | zu→xh | 11.50 | 49.29 | **11.64** | **49.39** | **11.90** | **49.58** | **12.00** | **49.60** |
| | xh→ss | 8.25 | 44.64 | 8.49 | 44.85 | 8.75 | **45.55** | **9.01** | **45.73** |
| | ss→xh | **9.40** | 44.82 | **9.21** | 44.49 | **9.63** | **45.24** | **9.42** | **45.35** |
| | zu→ss | **9.30** | 47.60 | **9.48** | 47.54 | **9.37** | **47.87** | **9.50** | **47.91** |
| | ss→zu | 11.23 | 48.01 | **11.47** | 47.86 | **11.76** | **48.45** | **11.77** | **48.45** |

Table 6: MT test set performance of PLMs after finetuning for multilingual MT. Underline indicates best scores, while **bold** indicates scores with differences from the best that are not statistically significant (based on paired bootstrap resampling testing with p-value 0.05).

be more inefficient than Afri-ByT5-base in terms of model size, but it requires a smaller adaptation corpus to achieve superior performance.

**Zero-shot cross-domain/cross-lingual** Table 5 evaluates our MasakhaneNEWS isiXhosa headline generation models on Vuk'uzenzele. As for MasakhaneNEWS, the metrics are quite low. However, given the lack of existing NLG datasets for the Nguni languages, these results should be viewed as a first step towards assessing cross-lingual Nguni text generation.

Adaptation does not help for isiXhosa and isiZulu, as ByT5-large is the best model for these languages. However, for languages not covered by base models (isiNdebele and Siswati), adaptation does improve transfer. Nguni-ByT5-large outperforms all baselines on these languages. Figure 2(b) shows the proportional gains obtained by adapted models over their respective base models. The results are similar to the NLU results in Figure 2(a), except that NLG transfer to isiZulu actually degrades after adaptation, perhaps because some of the transfer capabilities that emerge from greater multilinguality is lost. Nguni-ByT5-large improves cross-lingual transfer to isiNdebele and Siswati (especially the latter), more so than Afri-ByT5-base. This reiterates the findings from our cross-lingual NLU experiments, which showed that **Nguni-only adaptation leads to better cross-lingual transfer to the lower resourced Nguni languages than larger scale adaptation.**

**Multilingual MT** Table 6 contains MT results for 2 finetuned checkpoints per NLG model. One is finetuned to translate in all 3 directions listed for Autshumato, while another is finetuned in those listed for WMT22/FLORES. We perform paired bootstrap resampling (Koehn, 2004) to test for statistical significance. Nguni-ByT5-large obtains the best evaluation scores for most translation directions, although in all but 2 instances its improvements are not statistically significant over ByT5-large. However, the gains achieved by Nguni-ByT5-large across all languages does indicate greater consistency in its translation capabilities. Afri-ByT5-base mostly improves over ByT5-base, but the large variants comfortably outperform both.

## 7 Conclusion

This paper is the first comprehensive study of PLMs for Nguni languages. It provides a detailed overview of the prevailing landscape in datasets and modelling. Overall the state of affairs regarding PLMs for Nguni languages has improved markedly in recent years. Nevertheless, there remains a large gap between the capabilities of PLMs in the Nguni languages, compared to high-resource languages. Nguni-XLMR and Nguni-ByT5 show that this can be partially addressed through simple techniques like multilingual adaptive finetuning. Furthermore, we hope that our NGLUEni benchmark is used by future researchers to standardise evaluation for Nguni PLMs. This could facilitate a more accurate assessment of the true Nguni-language capabilities of PLMs, a subject that presents considerable challenges in the existing literature.

## ACKNOWLEDGEMENTS

This work was done during Francois Meyer's research internship at the National Institute of Information and Communications Technology (NICT) in Kyoto, Japan. Francois Meyer was also supported by the Hasso Plattner Institute for Digital Engineering, through the HPI Research School at the University of Cape Town. This work was also supported in part by the National Research Foundation of South Africa (Grant Number: 129850).

## REFERENCES

Steven P. Abney. *Parsing By Chunks*, pp. 257–278. Springer Netherlands, Dordrecht, 1992.

David Adelani, Jesujoba 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 Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 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*, pp. 3053–3070, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL `https://aclanthology.org/2022.naacl-main.223`.

David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 773–800, Abu Dhabi, United Arab Emirates (Hybrid), December 2022b. Association for Computational Linguistics. URL `https://aclanthology.org/2022.wmt-1.72`.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022c. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.298. URL `https://aclanthology.org/2022.emnlp-main.298`.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-Azzawi, Blessing K. Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Oluwaseyi Ajayi, Tatiana Moteu Ngoli, Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka C. Obiefuna, Shamsuddeen Hassan Muhammad, Saheed Salahudeen Abdullahi, Mesay Gemeda Yigezu, Tajuddeen Rabiu Gwadabe, Idris Abdulmumin, Mahlet Taye Bame, Oluwabusayo Olufunke Awoyomi, Iyanuoluwa Shode, Tolulope Anu Adelani, Habiba Abdulganiy Kailani, Abdul-Hakeem Omotayo, Adetola Adeeko, Afolabi Abeeb, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Raphael Ogbu, Chinedu E. Mbonu, Chiamaka I. Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge, Sakayo Toadoum Sari, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Ussen Abre Kimanuka, Kanda Patrick Tshinu, Thina Diko, Siyanda Nxakama, Abdulmejid Tuni Johar, Sinodos Gebre, Muhidin Mohamed, S. A. Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, and Pontus Stenetorp. Masakhanews: News topic classification for african languages. 2023.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1849–1863, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.145. URL https://aclanthology.org/2022.findings-acl.145.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10883–10900, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.609. URL https://aclanthology.org/2023.acl-long.609.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 52–64, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sustainlp-1.11. URL https://aclanthology.org/2022.sustainlp-1.11.

David M. Eberhard, Gary F. Simons, and Charles D. Fenning. *Ethnologue*. 2019.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6279–6299, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.435. URL https://aclanthology.org/2022.acl-long.435.

Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. Are character-level translations worth the wait? comparing character- and subword-level models for machine translation, 2023.

Roald Eiselen. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3344–3348, Portorož, Slovenia, May 2016a. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1533.

Roald Eiselen. South African language resources: Phrase chunking. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 689–693, Portorož, Slovenia, May 2016b. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1109.

Roald Eiselen and Martin Puttkammer. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3698–3703, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 179–188, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1017. URL https://aclanthology.org/P17-1017.

Tanja Gaustad and Martin J. Puttkammer. Linguistically annotated dataset for four official south african languages with a conjunctive orthography: Isindebele, isixhosa, isizulu, and siswati. *Data in Brief*, 41, 2022.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250.

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, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. doi: 10.1162/tacl_a_00447. URL https://aclanthology.org/2022.tacl-1.4.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007.

Richard Lastrucci, Jenalea Rajab, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pp. 18–25, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.rail-1.3.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Cindy McKellar and Martin Puttkammer. Dataset for comparable evaluation of machine translation between 11 south african languages. *Data in Brief*, 29:105146, 04 2020. doi: 10.1016/j.dib.2020.105146.

Francois Meyer and Jan Buys. Triples-to-isixhosa (t2x): Addressing the challenges of low-resource agglutinative data-to-text generation, 2024.

Franziska Pannach, Francois Meyer, Edgar Jembere, and Sibonelo Zamokuhle Dlamini. Nlapost2021 1st shared task on part-of-speech tagging for nguni languages. *Journal of the Digital Humanities Association of Southern Africa*, 3(01), Feb. 2022. doi: 10.55492/dhasa.v3i01.3865. URL https://upjournals.up.ac.za/index.php/dhasa/article/view/3865.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1044. URL https://aclanthology.org/D15-1044.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4603–4611. PMLR, 2018. URL http://proceedings.mlr.press/v80/shazeer18a.html.

Dirk Snyman, Gerhard Van Huyssteen, and Walter Daelemas. Cross-lingual genre classification for closely related languages. In *Proceedings of the 2011 Conference of the Pattern Recognition Association of South Africa*, 2011.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 473–482, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.40. URL https://aclanthology.org/2021.naacl-main.40.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4438–4450, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.359. URL https://aclanthology.org/2020.emnlp-main.359.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17.

Judit Ács. Exploring bert's vocabulary., 2019. URL http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html.

# A PLM SIZES

Nguni-XLMR-large directly adapts XLM-R-large, so it shares its model size (355M parameters - 24 layers, hidden size 1024, feed-forward hidden size 4096, 16 heads) and its 250k subword vocabulary constructed with ULM Kudo (2018).

Nguni-ByT5-large has the same parameter size as ByT5-large (1.23B - 36 encoder layers, 12 decoder layers, hidden size 1536, feed-forward hidden size 3840, 16 heads) and the same vocabulary of 256 possible byte values.

| **Language** | **xh** | **zu** | **nr** | **ss** |
|---|---|---|---|---|
| **Speaker statistics** | | | | |
| L1 | 8m | 12m | 2.3m | 1.1m |
| L2 | 22m | 16m | 2.4m | 1.4m |
| **Pretraining corpus size (tokens)** | | | | |
| XLM-R | 13m | 0 | 0 | 0 |
| ByT5 | 60m | 200m | 0 | 0 |
| **Adaptation corpus size (tokens)** | | | | |
| Afro-XLMR | 60m | 200m | 0 | 0 |
| Afri-ByT5 | 60m | 200m | 0 | 0 |
| **Nguni-XLMR/ByT5** | **60m** | **200m** | **450k** | **500k** |

Table 7: Language speaker statistics and per-language corpus sizes for different PLMs.

# B FINETUNING DETAILS

## B.1 NLU

We finetune all our models with the AdamW optimizer (Loshchilov & Hutter, 2019) for 20 epochs using a learning rate of 5e-5 (except classification, where we used a learning rate of 2e-5), and no warmup steps. We initially used a batch size of 32 across all tasks, but this led to unstable training in some unbalanced datasets (the model would classify all examples as the majority class). In cases where this was observed, we used a larger batch size (128) for stabler training.

## B.2 NLG

We finetune NLG models on T2X until validation loss increases, which is before 5 epochs for all models. We use the Adafactor optimizer (Shazeer & Stern, 2018) with a learning rate of 1e-4, no warmup, and a batch size of 4.

For MasakhaNEWS (Adelani et al., 2023) headline generation we finetune models with Adafactor and no warmup, with a grid search across learning rates {1e-2, 1e-3, 1e-4} and batch sizes {2, 4, 8}. Validation performance peaks at a learning rate of 1e-3 and a batch size of 2.

For our MT experiments we follow the hyperparameters of Adelani et al. (2022a), using a learning rate of 1e-5, linear decay, no warmup steps, a batch size of 16, and train for 3 epochs of the training corpus. Each source sentence is concatenated with a prefix describing the translation direction e.g. "Translate English to Xhosa" during training and testing.