

Continued Pretraining Learning Dynamics of Encoder-Decoder Models on isiXhosa

Rakeen Ramdeen
RMDRAK003@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Abstract

Continued pretraining (CPT) adapts pretrained language models to new languages without the cost of pretraining from scratch. Although widely used for low-resource languages, models are evaluated only after CPT is complete. This creates a gap in understanding how capabilities change throughout adaptation. The study of learning dynamics examines how models acquire knowledge throughout training, however, to the best of our knowledge, no work has examined encoder-decoder CPT learning dynamics for isiXhosa. We investigate the CPT learning dynamics of three encoder-decoder models adapted to isiXhosa: T5, ByT5, and Nguni-ByT5. The models differ in their prior exposure to isiXhosa. We continue pretraining each model for 10,000 steps and evaluate 20 checkpoints on machine translation, data-to-text generation, English-isiXhosa representation alignment, and part-of-speech probing. We compare monolingual isiXhosa CPT with bilingual English-isiXhosa CPT. Finally, we perform ablations by varying the downstream finetuning data budget. We find that CPT does not produce consistent downstream improvement. Nguni-ByT5 remains stable across every evaluation, while ByT5 change more substantially, suggesting that prior adaptation to related languages stabilises further CPT. POS knowledge remains most accessible in the middle encoder layers throughout adaptation. Including English during CPT preserves isiXhosa POS accuracy, most clearly for ByT5. We extend prior work on capability preservation during CPT to grammatical knowledge in the target language. We are the first to provide insights on the learning dynamics of encoder-decoder models for isiXhosa adaptation.

Keywords

Natural Language Processing, Learning Dynamics, Continued Pretraining, Low-Resource Languages, South African Languages, Natural Language Generation, AI Interpretability

1 Introduction

Pretrained Language Models (PLMs) have shown strong performance across many natural language processing (NLP) tasks, specifically generation [49] and understanding [13] tasks. Encoder-decoder models are a type of PLM that uses an encoder to represent an input sequence and a decoder to generate the corresponding output, making them well suited to tasks such as translation and summarisation [49]. PLMs success is primarily dependent on being trained on large datasets, which are often composed of high-resource languages such as English. As a result low-resource languages such as isiXhosa remain underrepresented in NLP due to digital text scarcity and unevenly distributed resources [2, 40]. Researchers address these issues by using adaptation methods such as continued pretraining

[24] and finetuning [29] to adapt PLMS to new languages. CPT further trains an existing model on a new language using a smaller corpus, allowing it to specialise for a target language without training from scratch [24], while finetuning adapts the model to specific downstream tasks using labelled data [29]. Prior work has shown that CPT can improve downstream performance for African languages [2, 40]. However, these studies evaluate the model only after adaptation, treating the adaptation process as a black box. This leaves it unclear how and when capabilities change during CPT.

The study of learning dynamics addresses this gap by examining model behaviour and representations at intermediate training checkpoints [42, 51]. Previous studies have shown that syntactic information develops early during pretraining, while factual knowledge develops more gradually [33, 42]. However, these studies focus on pretraining from scratch. CPT begins with existing linguistic representations, so adaptation may involve preserving or reorganising knowledge rather than acquiring it from scratch. Recent work has investigated these dynamics in decoder-only models, finding that capability changes during language adaptation are not necessarily reflected in validation loss [17]. The learning dynamics of CPT for encoder-decoder models and African languages have not been explored.

We investigate whether downstream performance improves consistently during isiXhosa CPT, how adaptation trajectories differ between models with varying prior isiXhosa exposure, and whether including English helps preserve prior linguistic knowledge. We study T5, ByT5 and Nguni-ByT5. T5 begins with no isiXhosa pretraining [49], ByT5 has multilingual exposure that includes isiXhosa [57], and Nguni-ByT5 has already undergone adaptation to the Nguni languages [40]. We evaluate 20 checkpoints across 10,000 CPT steps through machine translation (MT), data-to-text generation (D2T), cross-lingual alignment, and part-of-speech (POS) probing. We compare monolingual isiXhosa CPT with bilingual English-isiXhosa CPT. Finally, we perform ablations by varying the downstream finetuning data budget.

Our results show that further isiXhosa CPT does not lead to consistent downstream improvement. Performance gains are limited, and the observed trajectories depend on the amount of finetuning data. Nguni-ByT5 remains comparatively stable across evaluations, while ByT5 change more substantially, suggesting that prior adaptation may help stabilise further CPT. Including English helps preserve isiXhosa POS probing accuracy, particularly for ByT5. Mean pooling performs best for POS probing in the base models tested, and POS information remains most accessible in middle encoder layers throughout adaptation. Our findings show that modest changes in generation performance can accompany substantial representational changes. We also extend research on capability preservation during CPT [17] by demonstrating that bilingual adaptation can

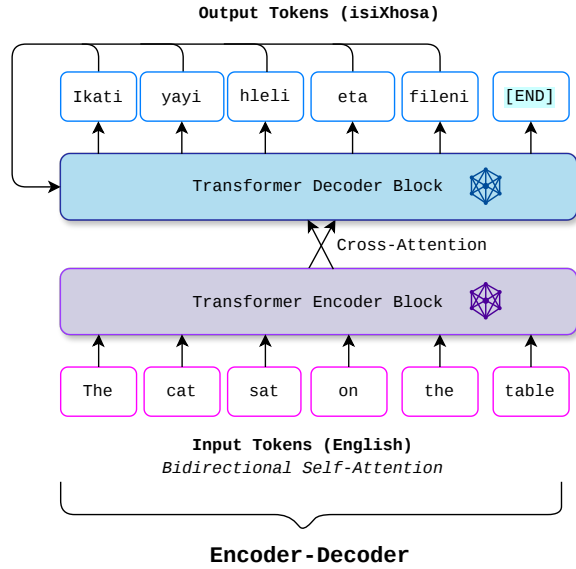


Figure 1: The encoder-decoder architecture, simplified from Vaswani et al. [2017].

help preserve target-language grammatical knowledge. To the best of our knowledge, this is the first study of CPT learning dynamics in encoder-decoder models adapted to isiXhosa, a low-resource South African language.

2 Background & Related Work

In this section, we begin by introducing encoder-decoder models used in our study. Then we explore the background of continued pretraining in African language contexts. Finally we discuss the field of learning dynamics and the gaps in the literature.

2.1 Encoder-Decoder Models

Encoder-decoder models excel in natural language generation tasks such as summarisation, machine translation (MT) and data-to-text (D2T) generation [40, 49]. Their architecture is well-suited for sequence-to-sequence tasks. The encoder maps an input sequence into contextualised vector representations, which are then used by the decoder to autoregressively generate the output sequence one symbol at a time [53]. This process can be visualised by referring to Figure 1. T5 [49] introduced a unified text-to-text framework which involved framing every NLP task as consuming input text and generating output text, achieving state-of-the-art results on English-centric benchmarks. mT5 [58] is a massively multilingual variant of T5, trained on 101 languages which includes isiXhosa. It addressed the poor translation performance of T5. ByT5 [57] extends mT5 by replacing subword tokenisation with raw bytes. This makes it robust to spelling variation and morphology, although the longer byte sequences increase computational cost [15]. Nguni-ByT5 [40] is a version of ByT5 that underwent continued pretraining on the

Nguni languages (isiXhosa, isiZulu, isiNdebele, Siswati). It outperformed other African-adapted language models on Nguni-specific generation tasks.

2.2 Continued Pretraining

CPT is a transfer learning approach [29] that adapts an existing pretrained language model to a new language or domain through further unsupervised training on additional data [24]. Cross-lingual transfer extends transfer learning to low-resource languages where a model is able to leverage knowledge it has learned in a high-resource language and apply it to a low-resource language. This approach is important for low-resource languages, where the limited availability of large corpora makes training language models from scratch impractical.

To adapt pretrained models to African languages, researchers have applied both monolingual and multilingual CPT. Monolingual CPT further trains a model on data from a single target language, allowing it to acquire language-specific knowledge. Adelani et al. [2021] demonstrated that adapting multilingual models to individual African languages through monolingual CPT can improve downstream performance. However, adapting a separate model to each language is resource-intensive and limits the potential for cross-lingual transfer [5].

Multilingual CPT instead adapts a model using a combined corpus of multiple languages. This allows a single model to support several languages while improving cross-lingual transfer between them. Alabi et al. [2022] and Adelani et al. [2022a] applied this approach to African languages, adapting encoder-only and encoder-decoder models through CPT, respectively. However, the benefits of multilingual adaptation depend on the languages included during training. Meyer et al. [2024] showed that restricting multilingual CPT to the linguistically related Nguni languages improved performance and cross-lingual transfer, especially for isiNdebele and Siswati which are extremely under-resourced. Their adapted Nguni-ByT5 model demonstrated the effectiveness of this approach for generative tasks. In all this previous work, the adapted models are evaluated only at the end of training, leaving questions around how and when capability develops during CPT unexplained.

2.3 Learning Dynamics

The study of learning dynamics examines how models acquire knowledge throughout training by evaluating their behaviour at multiple checkpoints. This makes it possible to observe when different capabilities emerge and how their development changes over time.

Learning dynamics have been studied through several approaches, including vector analysis [51], structural probing [10], zero-shot evaluation [33], and downstream finetuning [42]. Across these approaches, prior research reveals a hierarchical timeline. Models acquire basic syntactic rules rapidly in the earliest stages of training [33, 42, 51]. An intermediate phase follows where facts and world knowledge are acquired more slowly [7, 33, 42]. Reasoning emerges last, as the model requires longer training to develop long-range contextualisation [42]. This trajectory is consistent regardless of the model’s size or initialisation [11], and capabilities progress in parallel at different speeds [19].

Nearly all this work is based on pretraining from scratch and encoder-decoder models are absent from it. We cannot assume that findings on pretraining learning dynamics can be generalised to CPT, because pretraining starts from randomly initialised weights and CPT starts from a pretrained model. Elhady et al. [2025] is the closest work that mirrors our research. They investigated learning dynamics during CPT by adapting decoder-only models to Basque, Arabic and Indonesian. Their results identified a critical period early in CPT where models experienced large parameter changes, losing previously acquired capabilities before these changes were reflected in accuracy on downstream tasks. This demonstrates that the behaviour of models during language adaptation cannot necessarily be inferred from training loss alone. The learning dynamics for encoder-decoder models adapted through CPT remain unexplored.

3 Methodology

In this section we discuss the methodology and setup for our experiments. We discuss the models selected, the CPT corpus and setup, the finetuning and evaluation protocol, cross-lingual alignment, structural probes and lastly, bilingual CPT. Our methodology involves continued pretraining three classes of models, extracting intermediate checkpoints and finetuning these checkpoints on downstream tasks to evaluate how downstream performance changes during CPT.

3.1 Model Selection

We compare three encoder-decoder models with varying prior exposure to isiXhosa. T5 has no isiXhosa exposure as it was pretrained on English only. ByT5 underwent massively multilingual pretraining on 101 languages, including isiXhosa. However, its exposure to isiXhosa was proportionally limited because isiXhosa accounted for only a small proportion of its pretraining data. It was favoured over mT5 because character-level models have been shown to outperform subword models in low-resource settings [2, 15, 16]. Nguni-ByT5 was adapted to the Nguni languages and was trained on the highest proportion of isiXhosa data. It has the greatest exposure to isiXhosa. This selection allows us to observe the effect of prior linguistic knowledge during CPT. We use the large variants of all three models, which are available on HuggingFace.

We also note that T5 differs from the byte models in its tokenisation. Tokenisation divides text into smaller segments called tokens. A subword tokeniser might split *playing* into two distinct tokens: *play* and *ing*, while a byte tokeniser would represent each letter as a separate token. Subword tokenisation usually requires fewer tokens for the same text, however this depends on how well the vocabulary represents the language [46]. We expect T5 to process more raw text at each training step than the byte models.

3.2 Continued Pretraining Setup

CPT Corpus Selection. We continue pretraining on the isiXhosa WURA dataset [43] which is publicly available on Huggingface¹. WURA is a curated multilingual corpus with 16 African languages that was produced by filtering mC4 [58] of low-quality text. We choose it for three reasons:

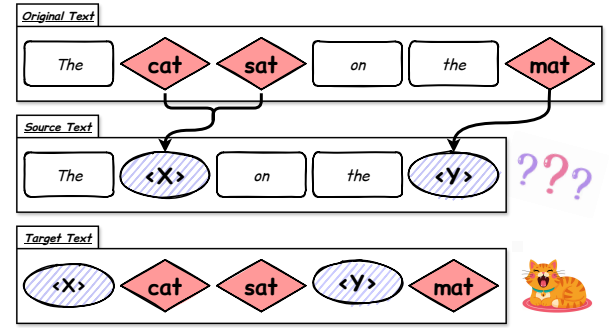


Figure 2: Span corruption process

- (1) The filtering process created cleaner, high-quality text which reduces the need for complex data cleaning pipelines.
- (2) It has been used in prior pretraining work for isiXhosa [35, 36].
- (3) It contains approximately 13 million words [39], which is smaller than other alternatives but is sufficient for CPT and suited for our compute budget.

We use the train and validation split of WURA and perform minimal cleaning, with only white space normalisation. We do not apply any content filtering or deduplication, leaving the dataset as published.

Training Objective And WURA Preprocessing. We use the original span corruption [49] training objective that all three encoder-decoder models were originally pretrained on. Each passage is first tokenised using the model’s pretrained tokeniser and divided into consecutive segments of at most 512 tokens. Each segment forms one training example. Within each training example, 15 percent of token positions are selected for corruption. This involves replacing selected tokens with a sentinel token that acts as a mask. Adjacent selected tokens are grouped into a contiguous span and replaced with one sentinel token in the source sequence. The target sequence contains the removed tokens with the same sentinel token prepended to mark the start of each corrupted span. Figure 2 illustrates this process for an example passage. The corruption process is performed offline with source and target files written to disk before training starts. We follow and adapt the pipeline used to train Afri-ByT5 [2].

Training Protocol and Hyperparameters. We continue pretraining all three encoder-decoder models using the Huggingface Transformers library [56] and Pytorch [45]. Each model was trained on a separate L40 GPU on the UCT HPC. Hyperparameters are listed in Table 1 in Appendix A and are identical across each model. We adopt these settings from Meyer et al. [2024] who established them for adapting ByT5 into Nguni-ByT5. We performed preliminary runs to verify the highest batch size that could fit into memory. Thereafter, we ran smaller candidate batch size configurations for 24 hours to compare loss curves. The different configurations produced indistinguishable loss curves so the fastest configuration in terms of step completion was kept. We use a batch size of 8 with 128 gradient accumulation for T5 and a batch size of 4 with 256

¹<https://huggingface.co/datasets/castorini/wura>

gradient accumulation for both byte models. These settings produce an effective batch size of 1,024 for all three models, which keeps the number of optimiser updates constant between models.

Checkpoint Schedule. To observe the effects of CPT on all models, we use a two-phase checkpoint saving schedule, adapted from Elhady et al. [2025]. During the first 10% of training steps, we save a checkpoint every 1% of steps. For the remaining 90% of training we save a checkpoint at every 10% of steps. We base our decision on prior research which has shown rapid changes in a model’s behaviour early in training while more gradual changes occur in later stages [42]. This results in 20 checkpoints per model, with the unadapted base model taken as step 0.

3.3 Finetuning & Downstream Task Evaluation

Downstream Tasks and Datasets. Data-to-text is a natural language generation task that constructs a text description from a structured input such as table data, tuples or even images [20]. It evaluates a model’s ability to generate faithful and fluent text [55]. We train and evaluate D2T on the Triples-to-isiXhosa (T2X) dataset [38]. Each T2X example contains a triple consisting of a subject, relation and object paired with an isiXhosa verbalisation. For example, the triple (*Abilene Regional Airport*, *cityServed*, *Abilene Texas*) is verbalised as *Isikhululo seenqwelo moya sase-Abilene sisebenzela isixeko sase-Abilene eTexas*.

Machine translation is the automatic translation of written text from one language into another [52]. It evaluates a model’s ability to produce fluent and adequate target-language output [25]. We train all checkpoints using the WMT22 dataset [3] which contains mined parallel translation sentences extracted by using LASER. LASER uses sentence encoders that represent different languages in a shared vector space, allowing potential translations to be identified based on their similarity. We use the FLORES-200 dataset to evaluate each finetuned checkpoint’s performance on MT. FLORES-200 was developed by professional translators under quality control and is the standard evaluation set for low-resource languages.

The complex morphology of agglutinative languages like isiXhosa makes both tasks challenging for text generation [8, 37, 38]. We choose these datasets because they are the only structured text generation tasks available for isiXhosa text-to-text generation. This allows us to evaluate how isiXhosa text generation capabilities evolve during CPT.

Dataset Processing. We filter and remove duplicate entries in WMT22 as it was LASER mined rather than professionally translated. We keep the top 50k examples with the highest LASER score. We ensure this selection is deterministic so that every checkpoint is finetuned with the same data. FLORES-200 is used as published with its dev and devtest split used for validation and testing respectively. T2X is used as published with its training, validation and test set. For MT we also prepend a task string "Translate English to isiXhosa:", following the task prefix format the T5 family was trained on [49]. D2T uses no prefix. All data is tokenised with each model’s pretrained tokeniser.

Finetuning Protocol. We finetune and evaluate each of the 20 checkpoints per model on MT and D2T separately. Finetuning hyperparameters are listed in Table 2 in Appendix A and are based

on hyperparameter values reported in prior work for CPT [40]. We deviate from the cited hyperparameters in two places. We include a 20% learning rate warm ratio for D2T as we encountered unstable training and prior work shows that adding warmup can reduce instabilities [22, 23]. We also increased the number of epochs for MT from three to five because the models did not converge after three epochs. These deviations are applied across all runs making comparisons fair.

Evaluation. We report BLEU [44] and chrF [47] for both D2T and MT. BLEU measures the precision of word n-grams in the model’s generated output with the actual reference text, with a penalty for output that is too short. chrF computes the F-score over character n-grams. This differs from BLEU as complete words do not need to match and has been argued to suit morphologically complex languages more [25]. Using both allows for better analysis as a high chrF with a low BLEU score might point towards morphological accuracy without exact whole word or sequence accuracy. We use the sacreBLEU [48] library to compute both metrics. For each finetuning run, we report the evaluation scores from the model after the final training epoch.

3.4 Cross-Lingual Alignment

Multilingual models represent words and sentences from different languages in a shared vector space. Cross-lingual alignment describes how semantically equivalent sentences are represented across languages [26]. Prior work has shown languages that exhibit stronger alignment with English correlate with better performance on downstream tasks [32]. Languages with similar representations enable cross-lingual transfer. This makes greater alignment desirable for low-resource languages such as isiXhosa. We measure how alignment between English and isiXhosa translated sentences changes over the course of CPT.

We follow the procedure of Idris et al. [2026] using the FLORES-200 devtest split which contains 1,012 English and isiXhosa translated pairs. For each sentence we obtain a sentence representation by passing it through the encoder and averaging its token representations. These representation vectors are L2-normalised so that their dot product computes cosine similarity:

$$SC(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}| |\vec{v}|} \quad (1)$$

where $\vec{u}$ and $\vec{v}$ are the sentence representations being compared. Higher cosine similarity indicates that two sentences have more similar representations in encoder latent space. Cosine similarity is computed between every English and isiXhosa sentence resulting in a 1012×1012 matrix. The diagonal contains the true translation pairs while the off-diagonal entries represent unrelated sentence pairs. We report mean cosine and cosine gap from this matrix. Mean cosine similarity is computed across the diagonal:

$$\text{CosineMean} = \frac{1}{N} \sum_{i=1}^N SC(\vec{u}_i, \vec{v}_i) \quad (2)$$

Cosine gap compares the mean similarity with the total average of all possible sentence pairs:

$$\text{CosineGap} = \text{CosineMean} - \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{SC}(\vec{u}_i, \vec{v}_j) \quad (3)$$

A higher cosine gap indicates that true translation pairs are more distinguishable from unaligned pairings. This is necessary because language model representations are known to be anisotropic, where representations occupy a narrow region of vector space. This causes representations to have similar direction, inflating cosine similarity even for unrelated sentences and potentially giving a misleading impression of stronger alignment. We compute equation 2 and 3 at every encoder layer across every checkpoint to observe the effect of CPT on representation alignment.

3.5 Structural Probing

Structural probing seeks to localise where linguistic abilities are encoded in a language model [12] by training a classifier on intermediate representations. Prior work has used layer-wise probes to determine the extent to which linguistic features are encoded in a model’s hidden layers [4, 10, 12, 33]. We use probing to investigate how part-of-speech knowledge is encoded across the encoder layers of T5, ByT5 and Nguni-ByT5 and how this organisation changes during CPT.

Task and Dataset. POS tagging assigns each word a grammatical category, such as a noun, verb or determiner. These categories represent the grammatical role of a word, which makes POS tagging a useful task for evaluating how a model encodes syntactic information. We use the isiXhosa portion of the MasakhaPos dataset [14], a collection of news sentences in 20 African languages where every word is labelled with its POS tag.

Word Representations. For each checkpoint, we freeze the encoder and extract the hidden representations of every sentence at each encoder layer. These representations form the input to the probe. POS labels are assigned to whole words, however, a word may be tokenised into multiple subword or byte tokens. Each token has a hidden representation, which we combine into one vector per word using pooling. We can combine these representations into one vector using three different pooling methods. For a word segmented into tokens with representations $h_1, \dots, h_n$, first pooling uses h_1 , last pooling uses h_n , and mean pooling uses their average. Dang et al. [2025] found that last pooling outperformed mean pooling in their morphological probing experiments. However, prefixes can change a word’s grammatical category in isiXhosa [14]. This motivates us to consider multiple pooling strategies. Following Aduah and Meyer [2025], who suggest researchers compare pooling strategies for African languages, we compare all three methods on the base models and use the best method for CPT analysis.

Probe Classifier and Training. We train a probe for every combination of model, checkpoint, layer and pooling method. Each probe is a single linear layer that maps a word vector to one of the 17 POS tags, and only the probe is updated during training. We use a linear probe to test whether POS information is linearly separable in the representations. Dang et al. [12] describe the distinction between linear and MLP probes and report similar performance for the two

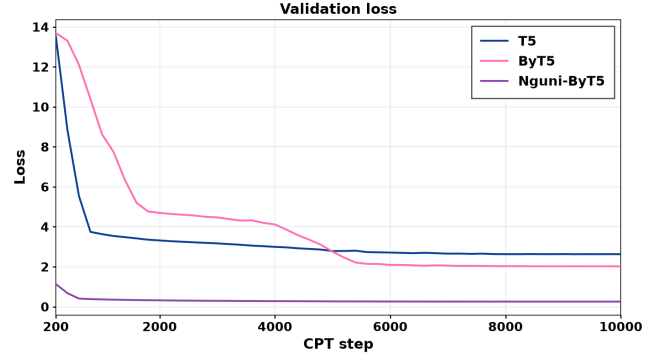


Figure 3: Validation loss across all models, evaluated at every 200 steps.

approaches on morphological tasks. Our training code is adapted from Hewitt and Liang [2019], and our hyperparameters, listed in Table 3 in Appendix A, follow Müller-Eberstein et al. [2023].

Evaluation. We probe the same 20 checkpoints used in the downstream evaluation. Our main metric is POS accuracy on the test split, which is the proportion of words assigned the correct tag.

3.6 Bilingual Continued Pretraining

We compare monolingual and bilingual CPT to examine whether including English benefits isiXhosa adaptation and changes its learning trajectories. Prior work has found that knowledge learned from English can support tasks in other languages through cross-lingual transfer [6]. Including English during CPT can help preserve previously acquired capabilities [17] and improve target-language performance [18]. These findings motivate our comparison. We supplement the isiXhosa corpus with an equal number of English passages from WURA and continue pretraining all three encoder-decoder models, keeping the same training settings. We follow the same evaluation procedures as for monolingual CPT: full finetuning on MT and D2T, cross-lingual alignment analysis, and structural probing.

4 Results

We present our findings from our experiments and discuss their implications. We first discuss CPT trajectories, then downstream task performance for MT and D2T, followed cross-lingual alignment analysis, POS probing and finally bilingual CPT dynamics.

4.1 CPT Trajectories

We note that absolute loss values between models are not directly comparable because of differing tokenisation and vocabularies. We instead focus on the rate of convergence. As shown in Figure 3, validation loss for Nguni-ByT5 converges the fastest at around step 800. This is expected given that it has already been exposed to the most isiXhosa data during its adaptation. T5 also converges at a rapid rate, stabilising at step 1,000, despite having no prior isiXhosa exposure. In contrast ByT5 continues to slowly improve until step 6,000. This slower convergence may be related to the inefficiencies

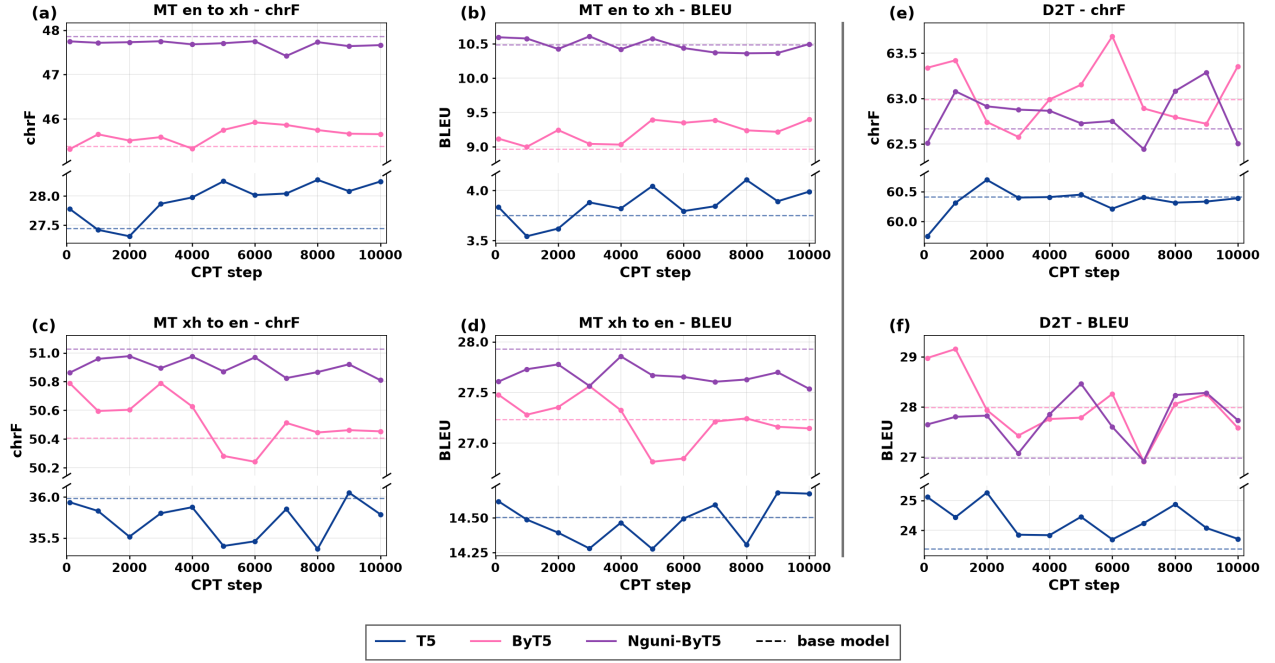


Figure 4: chrF and BLEU reported for all three models across 20 checkpoints for Translation and Data-to-Text tasks.

of byte-level tokenisation [16, 57]. Byte-level models process individual characters rather than contiguous spans of characters like subword models. This results in byte models processing less raw text at each training step. This interpretation is consistent with prior work, which found that differences in convergence between byte and subword models were reduced when sample throughput was matched [21]. We observed that T5 trained at roughly 63 epochs while the byte models only trained at 40 epochs at the same step count budget.

4.2 Translation Dynamics

We report the English-to-isiXhosa direction as our primary translation result. The reverse direction, isiXhosa-to-English is used to check if English capability degrades as CPT progresses and is discussed separately. Figure 4 reports chrF and BLEU across all 20 checkpoints for both translation directions and for D2T.

English-To-isiXhosa. Nguni-ByT5 outperforms ByT5, which in turn outperforms T5, across the entire CPT trajectory for both chrF and BLEU. This ordering aligns with our expectations given the model’s prior isiXhosa exposure. Nguni-ByT5’s performance remains relatively stable throughout CPT. Its chrF shows hardly any change and BLEU oscillates around the base model’s score. This suggests Nguni-ByT5 does not benefit from further adaptation. ByT5 and T5 follow a different pattern where both decrease below the base model’s performance before recovering later in training. These patterns are consistent across chrF and BLEU. T5 shows an overall upward trend, suggesting it has the most to gain from further isiXhosa adaptation. Both models appear to achieve most of their performance gains around halfway through training. These results

suggest that most performance gains from CPT occur primarily in the first half of training, while subsequent training provides minimal additional benefit. The chrF and BLEU trajectories generally agree with each other, however, there are small mismatches which appear to be metric-level variation and noise rather than evidence of morphological behaviour divergence.

isiXhosa-To-English. Translation requires both source and target language understanding, therefore it may be plausible that isiXhosa CPT trades English understanding for better isiXhosa generation. We evaluate isiXhosa to English translation to determine if CPT on isiXhosa degrades English capabilities.

ByT5 exhibits a distinct decline for both metrics around halfway through training. We note that this decline happens in the same range where English-to-isiXhosa performance peaks. The checkpoint at step 6,000 achieves the lowest chrF score for English generation, however, this is the same checkpoint that scores the highest chrF score for isiXhosa generation. The same pattern can be seen for checkpoint 5,000 for BLEU. T5 shows a similar dynamic for chrF where at checkpoint 8,000 it experiences the worst English generation performance while achieving its best isiXhosa generation performance. This suggests that further isiXhosa adaptation does gradually impair English generation, however, the degradation is minimal in scale. Some checkpoints favouring isiXhosa generation appear to do so partly at the expense of English generation. Nguni-ByT5 experiences a different dynamic and does not show the same potential trade-off. Every adapted checkpoint performs below the baseline across both metrics. Its English generation performance drifts downwards with no corresponding gain in isiXhosa generation.

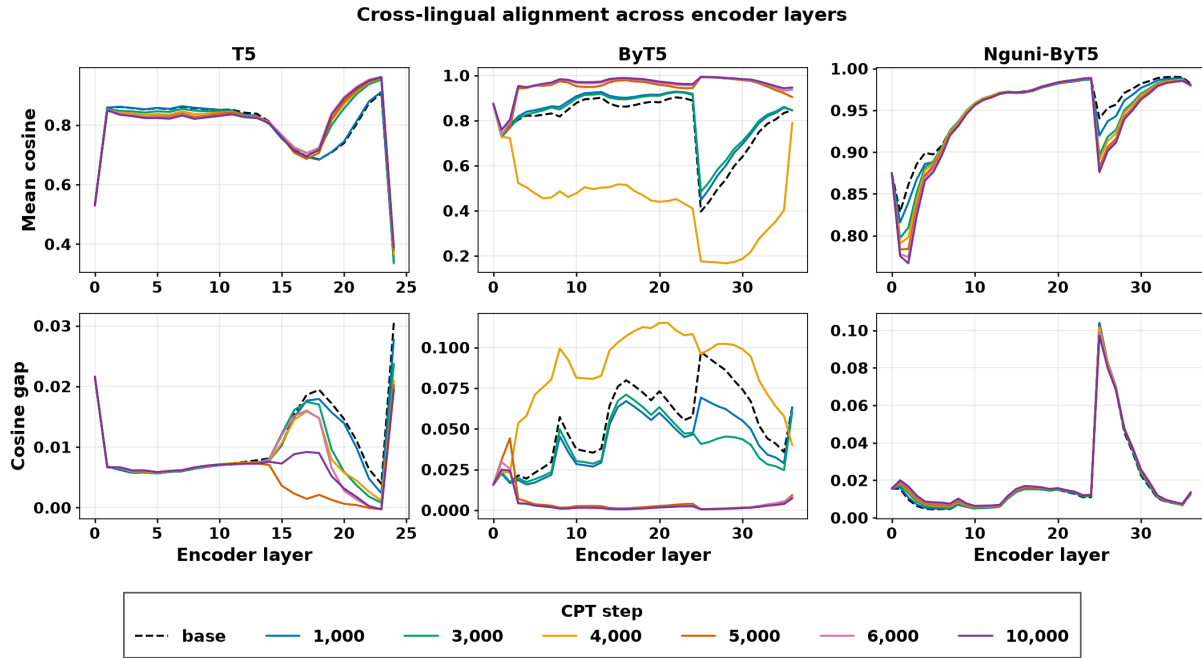


Figure 5: Mean cosine and cosine gap reported for the final encoder layer of each model across 7 checkpoints.

4.3 Data-to-Text Dynamics

As shown in Figure 4, Nguni-ByT5 and ByT5 perform similarly across the D2T trajectory, while T5 remains below both byte models. The ByT5 base model outperforms the Nguni-ByT5 base model on both metrics, although checkpoint performance does not maintain a consistent ordering. Despite its greater isiXhosa exposure, Nguni-ByT5 does not show a clear performance advantage over ByT5 on D2T. The models differ in how their performance changes relative to their base model’s score during CPT. Nguni-ByT5 exceeds its base model’s BLEU score at most checkpoints, while T5 does so at every checkpoint. In contrast, ByT5’s checkpoint performance fluctuates around the base model’s score. However, none of the models show evidence of sustained increase in performance as CPT progresses. T5’s BLEU trajectory fluctuates with no clear upward trend. Furthermore, its chrF score remains in a narrow range, hardly deviating from its base model’s score. The BLEU and chrF trends exhibited by T5 differ, which may indicate that whole-word matching improvement is more apparent than character overlap. For the byte models, both metrics follow similar trajectories with small disagreements that do not indicate a clear morphological divergence. The relatively small size of WURA may explain why more isiXhosa pretraining does not actualise into better D2T performance. A larger corpus may be needed to provide the models with more varied training data to support consistent improvement.

4.4 Alignment Dynamics

The layer-wise results in Figure 5 suggest that English-isiXhosa alignment is stronger in the deeper encoder layers. All three models exhibit a decrease in cosine mean alongside an increase in cosine

gap in the deeper encoder layers. This implies similarity between translation pairs decreases while their distinguishability increases, indicating greater cross-lingual alignment. These patterns indicate that the deeper layers reorganise representations for greater alignment.

CPT affects the three models alignment differently depending on their architecture and prior adaptation. T5’s early and middle layers remain unchanged, while its deeper layers show a gradual declining cosine gap, with an increasing cosine mean. This suggests CPT weakens T5’s cross-lingual alignment up until step 5,000. At step 6,000, cosine gap partly recovers in these deeper layers, suggesting a realignment phase where true translation pairs become more distinguishable again. ByT5 undergoes the most reorganisation across its entire encoder. At step 4,000, cosine gap increases across almost every layer with cosine mean decreasing. At step 5,000, this trend reverses, with cosine gap decreasing across most layers. These results show that ByT5 experiences a transient state of increased sentence distinguishability with lower overall sentence similarity which is then followed by a collapse in cross-lingual alignment. Therefore, CPT may reorganise the representation space which changes both the strength of alignment and where alignment is most expressed. Nguni-ByT5 is distinct in its alignment dynamics. Its cosine gap and mean trends are largely unchanged throughout further adaptation. Its stability contrasts with ByT5’s broader changes despite their shared architecture and tokenisation. Therefore, prior language adaptation may result in more stable representation changes during CPT.

These results highlight the importance of considering cosine gap when interpreting alignment. During CPT, ByT5 and T5 experience increases to cosine mean that, considered alone, suggest improved

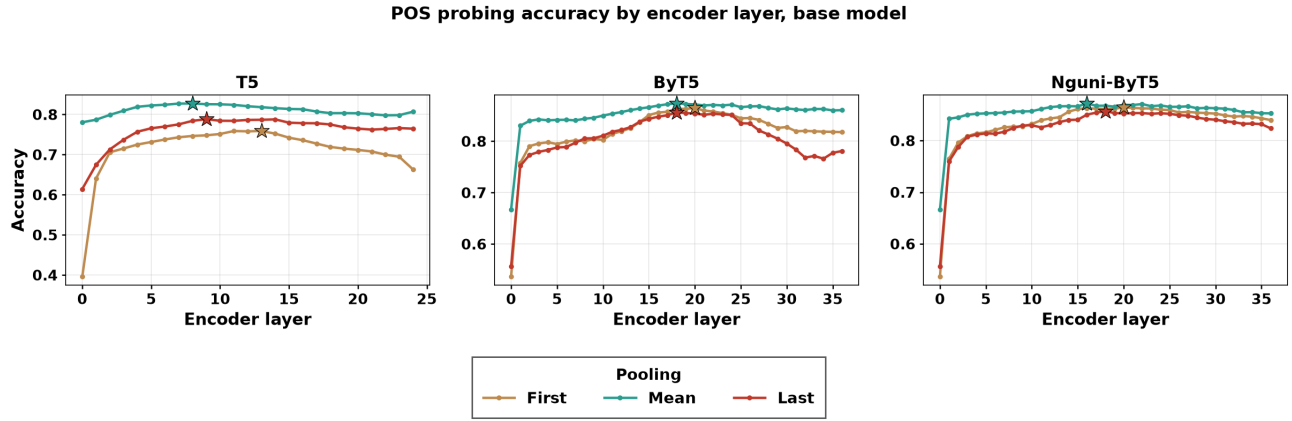


Figure 6: POS accuracy across all encoder layers for the base models comparing mean, first and last pooling.

cross-lingual alignment. However, the decreasing cosine gap may reflect increasing anisotropy, with representations becoming more similar in general.

4.5 Probing Dynamics

Pooling Strategy and Embedding Layer. Mean pooling achieves the highest POS accuracy across the encoder layers of all three models, as shown in Figure 6. This suggests that aggregating representations across a word provides more linguistic information for POS prediction than selecting only its first or last token position. This finding is consistent with isiXhosa’s agglutinative morphology, where multiple parts of a word can contribute grammatical information.

Both byte models show substantially lower accuracy at the embedding layer than after the first encoder layer, across all pooling methods. At the embedding layer, representations do not contain sentence context from self-attention. The sharp increase after the first encoder layer suggests that early layers make POS linguistic information more accessible by combining context across the sentence. For T5, mean pooling substantially improves embedding-layer accuracy over first and last pooling, indicating that aggregating subword embeddings provides useful POS cues even before contextual processing.

CPT Trends. Our findings in Figure 7 show POS probing accuracy peaks in the intermediate encoder layers and decreases towards the output. CPT largely preserves this pattern, suggesting that adaptation changes the strength and organisation of syntactic knowledge while retaining its concentration in the middle of the encoder.

Previous studies of encoder-only models find that middle layers encode strong syntactic knowledge [50]. Probes using these layers perform best at reconstructing syntactic trees [28] and identifying several grammatical properties [31]. Middle layer representations also support strong performance across a wide range of linguistic probing tasks [34]. We observe a similar pattern in the encoders of T5, ByT5 and Nguni-ByT5. One possible explanation for this phenomenon is that middle layers combine lexical and contextual information into grammatical knowledge, while deeper layers reshape these representations to optimise the pretraining objective

[54]. In our models, the ability to reconstruct corrupted spans may be favoured over disambiguating grammatical categories.

Within this shared trend, T5’s peak accuracy initially shifts towards shallower layers before moving deeper. Its peak accuracy remains comparatively stable, suggesting reorganisation within the middle layers. This is consistent with Müller-Eberstein et al. [2023], who find that linguistic representations continue reorganising after performance stabilises. ByT5 instead shows declining accuracy across the encoder while retaining its middle layer peak. Its grammatical representations appear to weaken rather than shift to another depth. This could suggest a total loss of syntactic knowledge or simply that the linear decodability of syntactic knowledge is weakening. Nguni-ByT5 preserves both accuracy and its layer-wise pattern, suggesting that prior Nguni adaptation has established grammatical representations that remain stable during further isiXhosa CPT.

4.6 Bilingual CPT Dynamics

We find that full finetuning follows similar trends in both settings, so we focus on the clearer differences in POS probing and cross-lingual alignment. Figure 8 reports POS probing accuracy, mean cosine similarity and cosine gap across encoder layers for all three models under bilingual CPT.

POS Probing. Bilingual CPT preserves isiXhosa POS accuracy, reducing degradation and keeping adapted checkpoints closer to base model performance. Our results imply that including English allows more isiXhosa grammatical knowledge to be encoded in the models representations. This is consistent with the findings from prior work for cross-lingual transfer discussed in Section 3.6.

ByT5 exhibits the most benefit, with smaller accuracy declines across the encoder than under monolingual CPT. T5’s peak accuracy also shifts within a narrower layer range. This suggests less reorganisation than under monolingual adaptation, where the peak shifts to deeper layers at later checkpoints. Nguni-ByT5 remains stable in both settings, showing minimal benefit from English inclusion.

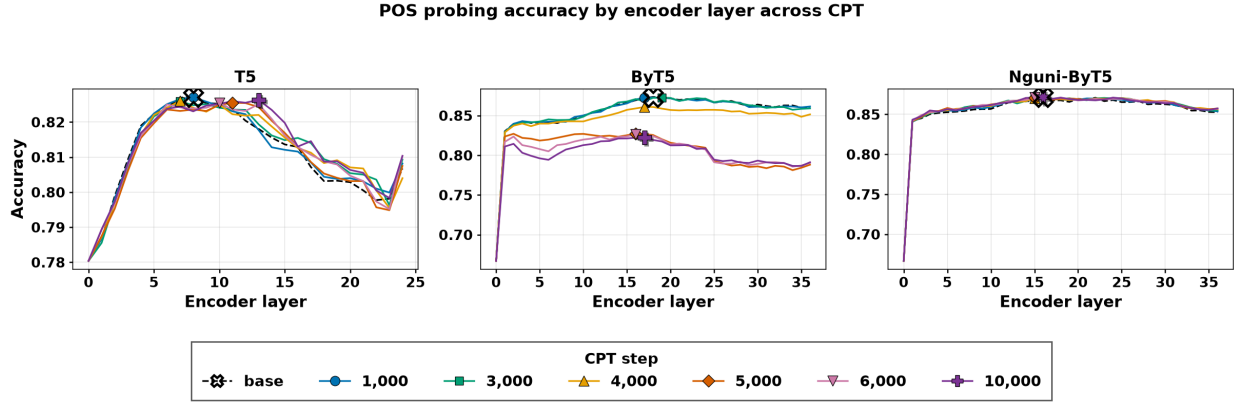


Figure 7: POS accuracy across every encoder layer for T5, ByT5 and Nguni-ByT5 showing selected checkpoint trends

Cross-lingual Alignment. Bilingual CPT preserves greater alignment but does not fully prevent the broader deterioration trends seen in monolingual CPT.

ByT5 still exhibits a decline in cosine gap around the middle of training, accompanied by increasing cosine similarity. However, some cosine gap is preserved in the middle layers, unlike the collapse across the majority of layers at later monolingual checkpoints. T5 shows temporary alignment gains before its deeper layer cosine gap decreases. Its realignment phase now occurs at step 5,000. Nguni-ByT5’s alignment remains largely unchanged.

5 Discussion

In this section we interpret our results. We first consider why further CPT produces limited downstream gains, then how prior isiXhosa exposure relates to adaptation stability, and finally how including English during CPT preserves isiXhosa grammatical knowledge.

5.1 Limited Gains From CPT

Our results display small downstream changes that make it difficult to interpret learning dynamics. Further isiXhosa pretraining does not show significantly improved generation. One possible explanation is that repeated exposure to a relatively small CPT corpus improves familiarity with isiXhosa without providing enough new information to benefit downstream tasks. More varied pretraining data may be more useful than additional passes through the same corpus.

However, further CPT does not necessarily imply significantly greater downstream task performance. CPT trains encoder-decoder models to reconstruct isiXhosa spans, while machine translation and D2T require them to generate text from another language or structured inputs. Our ablation results (Appendix B) show clearer gains from increasing finetuning data. Similarly, Adelani et al. [2022a] found that a few thousand high-quality translation pairs were enough to increase performance, while monolingual CPT provided relatively smaller performance gains. Further CPT may provide limited benefit when the task requirements become the main difficulty.

Our ablations also found that performance and trends fluctuate with less finetuning data, making individual checkpoint differences difficult to attribute to CPT. Based on our findings, there is clearer

evidence for increasing task-specific data than for extending CPT on the same corpus.

5.2 Prior Exposure and Stability

Prior isiXhosa exposure helps explain both the opportunity for improvement and adaptation stability. T5 begins with no isiXhosa exposure so it has the most to gain and more opportunity to improve. Nguni-ByT5 has already undergone Nguni specialisation, therefore further training provides hardly any new information which explains its small gains. Its stability across generation, alignment and probing suggests that prior adaptation reduces the need to reorganise existing representations. ByT5 occupies a middle ground. It has multilingual exposure but less concentrated Nguni exposure. Its more volatile changes suggest that massively multilingual pretraining does not offer the same stability as specialised Nguni adaptation. Prior adaptation may reduce both the need for further learning and the disruption caused by it.

5.3 Preserving Grammatical Knowledge During CPT

We find that including English during CPT helps preserve isiXhosa grammatical knowledge. Cross-lingual alignment may offer an explanation. Multilingual models represent some grammatical relationships similarly across languages [9] and cross-lingual alignment has been found to support transfer on grammatical tasks [41]. Therefore, continued exposure to English may keep these shared grammatical representations aligned between languages, preserving grammatical knowledge.

We hypothesize that layer-specific changes in alignment influence grammatical knowledge. In our models, isiXhosa POS knowledge is concentrated primarily in the middle layers. During monolingual CPT, ByT5 exhibits a collapse in alignment across its encoder layers, accompanied by a deterioration in POS performance. In contrast, during bilingual CPT, alignment in these middle layers is preserved even as deeper layers collapse, with overall POS performance remaining intact. T5 displays a similar trend where under both monolingual and bilingual CPT, alignment deteriorates only in the deeper layers while overall POS performance stays relatively

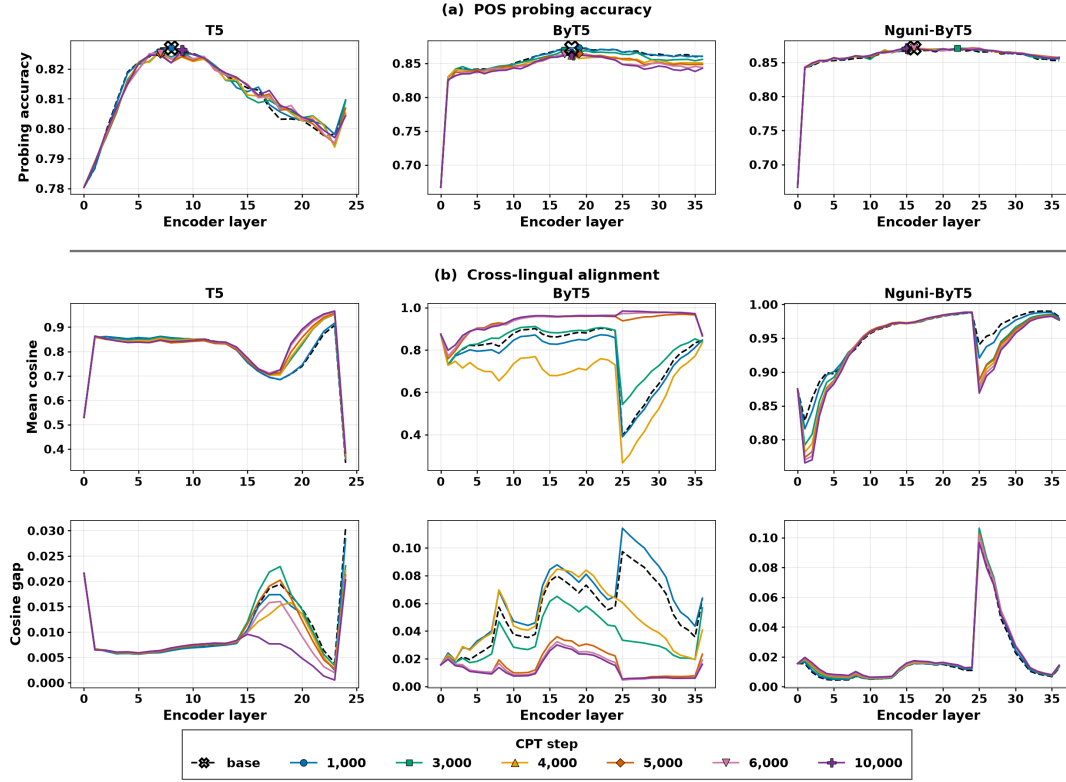


Figure 8: POS probing accuracy and cross-lingual alignment across encoder layers under bilingual CPT for T5, ByT5 and Nguni-ByT5 showing selected checkpoint trends.

the same. This suggests that deep-layer alignment is not strictly necessary for retaining POS knowledge. Meanwhile, Nguni-ByT5’s stability across both monolingual and bilingual CPT indicates that incorporating English is most effective when target-language adaptation would disrupt existing multilingual representations.

We extend the finding of Elhady et al. [2025] that including English during CPT preserves a model’s existing capabilities to also preserve grammatical knowledge in the target language.

6 Conclusion

We investigated isiXhosa CPT dynamics in T5, ByT5 and Nguni-ByT5 through downstream task evaluation, cross-lingual alignment and POS probing. CPT produced limited downstream gains. Ablating the downstream finetuning budget showed that varying the amount of task-specific data produced larger performance differences than varying CPT duration. Furthermore, the observed CPT trajectories depend on the budget used. Across these evaluations, Nguni-ByT5 remained comparatively stable, while ByT5 exhibited greater changes, suggesting that prior adaptation may help stabilise CPT. Mean pooling performed best for isiXhosa POS probing in the base models tested, while POS information was most accessible in middle encoder layers. CPT largely preserved this pattern. Including English helped preserve isiXhosa POS probing accuracy, particularly for ByT5. We extend the finding of Elhady et al. [2025] that including English preserves a model’s existing capabilities, showing

that it also preserves grammatical knowledge in the target language. These findings demonstrate that modest changes in downstream generation can coexist with substantial representational changes, highlighting the value of studying learning dynamics to understand language adaptation.

7 Limitations

Our POS probing experiments do not include a control baseline to distinguish information encoded in the model from patterns learned by the probe classifier. We cannot interpret absolute probing accuracies as all grammatical knowledge the model might possess. However, comparisons across checkpoints and layers under the same probing setup remain informative about the relative performance and accessibility of POS information and how it changes during CPT.

8 AI Usage Statement

All interpretations, analysis, and writing in this paper are strictly our original work. However, Claude was used to generate and refine plotting scripts for data visualization.

Acknowledgments

We would like to thank our supervisor Francois Meyer for his continuous support and the UCT HPC cluster for allowing us access to the GPUs we used to run these experiments.

References

- [1] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics* 9 (10 2021), 1116–1131. doi:10.1162/TACL_A_00416
- [2] David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3053–3070. doi:10.18653/v1/2022.naacl-main.223
- [3] David Ifeoluwa 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. 2022. 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)*, Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Nèveol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 773–800. doi:10.18653/v1/2022.wmt-1.72
- [4] Wisdom Aduah and Francois Meyer. 2025. Designing and Contextualising Probes for African Languages. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, Constantine Lignos, Idris Abdulmumin, and David Adelani (Eds.). Association for Computational Linguistics, Vienna, Austria, 39–51. doi:10.18653/v1/2025.africanlp-1.7
- [5] Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4336–4349. https://aclanthology.org/2022.coling-1.382/
- [6] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4623–4637. doi:10.18653/v1/2020.acl-main.421
- [7] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Proceedings of Machine Learning Research* 202 (5 2023), 2397–2430. http://arxiv.org/abs/2304.01373
- [8] William Chen and Brett Fazio. 2021. Morphologically-Guided Segmentation For Translation of Agglutinative Low-Resource Languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu (Eds.). Association for Machine Translation in the Americas, Virtual, 20–31. https://aclanthology.org/2021.mtsummit-loresmt.3/
- [9] Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5564–5577. doi:10.18653/v1/2020.acl-main.493
- [10] Cheng-Han Chiang, Sung-Feng Huang, and Hung yi Lee. 2020. Pretrained Language Model Embryology: The Birth of ALBERT. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (10 2020), 6813–6828. http://arxiv.org/abs/2010.02480
- [11] Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The Grammar-Learning Trajectories of Neural Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (4 2022), 8281–8297. http://arxiv.org/abs/2109.06096
- [12] Thao Anh Dang, Limor Raviv, and Lukas Galke. 2025. Tokenization and Morphology in Multilingual Language Models: A Comparative Analysis of mT5 and ByT5. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, Mourad Abbas, Tariq Yousef, and Lukas Galke (Eds.). Association for Computational Linguistics, Southern Denmark University, Odense, Denmark, 242–257. https://aclanthology.org/2025.icnlsp-1.24/
- [13] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (10 2018), 4171–4186. https://arxiv.org/pdf/1810.04805
- [14] Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba O. Alabi, Thapelo Sindane, Happy Buzaaba, et al. 2023. MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (2023), 10883–10900. doi:10.18653/v1/2023.ACL-LONG.609
- [15] Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. 2024. Are Character-level Translations Worth the Wait? Comparing ByT5 and mT5 for Machine Translation. *Transactions of the Association for Computational Linguistics* 12 (2024), 392–410.
- [16] Lukas Edman, Antonio Toral, and Gertjan van Noord. 2022. Subword-delimited downsampling for better character-level translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 981–992.
- [17] Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent Abilities of Large Language Models under Continued Pretraining for Language Adaptation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (9 2025), 32174–32186. http://arxiv.org/abs/2506.00288
- [18] Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Sorroa. [n.d.]. Latxa: An open language model and evaluation suite for basque, 2024. URL: https://arxiv.org/abs/2403.02266 [n.d.].
- [19] Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2023), 12205–12218. doi:10.18653/v1/2023.findings-acl.773
- [20] Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. arXiv:1703.09902 [cs.CL] https://arxiv.org/abs/1703.09902
- [21] Théo Gigant, Bowen Peng, and Jeffrey Quesnelle. 2026. Decoupling the Benefits of Subword Tokenization for Language Model Training via Byte-level Simulation. arXiv:2604.27263 [cs.CL] https://arxiv.org/abs/2604.27263
- [22] Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation. arXiv:1810.13243 [cs.LG] https://arxiv.org/abs/1810.13243
- [23] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2018. Accurate, Large MiniBatch SGD: Training ImageNet in 1 Hour. arXiv:1706.02677 [cs.CV] https://arxiv.org/abs/1706.02677
- [24] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (4 2020), 8342–8360. doi:10.18653/v1/2020.acl-main.740
- [25] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli-Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics* 48, 3 (2022), 673–732.
- [26] Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding Cross-Lingual Alignment—A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10922–10943. doi:10.18653/v1/2024.findings-acl.649
- [27] John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2733–2743. doi:10.18653/v1/D19-1275
- [28] John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4129–4138. doi:10.18653/v1/N19-1419
- [29] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (5 2018), 328–339. http://arxiv.org/abs/1801.06146
- [30] Tewodros Kederah Idris, Prasenjit Mitra, and Roald Eiselen. 2026. Can Embedding Similarity Predict Cross-Lingual Transfer? A Systematic Study on African Languages. arXiv:2601.03168 [cs.CL] https://arxiv.org/abs/2601.03168
- [31] Ganesh Jawahar, Benoît Sagot, and Djamel Seddah. 2019. What Does BERT Learn about the Structure of Language?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3651–3657. doi:10.18653/v1/P19-1356
- [32] Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schuetz. 2025. MEXA: Multilingual Evaluation

- of English-Centric LLMs via Cross-Lingual Alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 27001–27023. doi:10.18653/v1/2025.findings-acl.1385
- [33] Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing Across Time: What Does RoBERTa Know and When? *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021* (9 2021), 820–842. <http://arxiv.org/abs/2104.07885>
- [34] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. 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)*, Jill Burstein, Christy Doran, and Tamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 1073–1094. doi:10.18653/v1/N19-1112
- [35] Anri Lombard, Simbarashe Mawere, Temi Aina, Ethan Wolff, Sbonelo Gumede, Elan Novick, Francois Meyer, and Jan Buys. 2026. MzansiText and MzansiLM: An Open Corpus and Decoder-Only Language Model for South African Languages. (3 2026). <https://arxiv.org/pdf/2603.20732>
- [36] Alexis Matzopoulos, Charl Hendriks, Hishaam Mahomed, and Francois Meyer. 2025. BabyLMs for isiXhosa: Data-Efficient Language Modelling in a Low-Resource Context. (1 2025). <https://arxiv.org/pdf/2501.03855>
- [37] Francois Meyer and Jan Buys. 2023. Subword Segmental Machine Translation: Unifying Segmentation and Target Sentence Generation. arXiv:2305.07005 [cs.CL] <https://arxiv.org/abs/2305.07005>
- [38] Francois Meyer and Jan Buys. 2024. Triples-to-isiXhosa (T2X): Addressing the Challenges of Low-Resource Agglutinative Data-to-Text Generation. 16841-16854 pages. <https://aclanthology.org/2024.lrec-main.1464/>
- [39] Francois Meyer and Jan Buys. 2025. The Learning Dynamics of Subword Segmentation for Morphologically Diverse Languages. (11 2025). <http://arxiv.org/abs/2511.09197>
- [40] Francois Meyer, Haiyue Song, Abhisek Chakrabarty, Jan Buys, Raj Dabre, and Hideki Tanaka. 2024. NGLUEni: Benchmarking and Adapting Pretrained Language Models for Nguni Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, Torino, Italia, 12247–12258. <https://aclanthology.org/2024.lrec-main.1071/>
- [41] Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamel Seddah. 2021. First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2214–2231. doi:10.18653/v1/2021.eacl-main.189
- [42] Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace Chronicles: How Linguistic Information Emerges, Shifts and Interacts during Language Model Training. *Findings of the Association for Computational Linguistics: EMNLP 2023* (10 2023), 13190–13208. <http://arxiv.org/abs/2310.16484>
- [43] Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Odunayo Ogundepo, Abraham Toluwalase Owodunni, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better Quality Pre-training Data and T5 Models for African Languages. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings* (2023), 158–168. doi:10.18653/v1/2023.emnlp-main.11
- [44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135
- [45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32 (12 2019). <https://arxiv.org/pdf/1912.01703>
- [46] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. *Advances in Neural Information Processing Systems* 36 (10 2023). <http://arxiv.org/abs/2305.15425>
- [47] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 392–395. doi:10.18653/v1/W15-3049
- [48] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névél, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (Eds.). Association for Computational Linguistics, Brussels, Belgium, 186–191. doi:10.18653/v1/W18-6319
- [49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (9 2023), 1–67. <http://arxiv.org/abs/1910.10683>
- [50] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8 (2020), 842–866. doi:10.1162/tacl_a_00349
- [51] Naomi Saphra and Adam Lopez. 2019. Understanding Learning Dynamics Of Language Models with SVCCA. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (4 2019), 3257–3267. doi:10.18653/v1/N19-1329
- [52] Felix Stahlberg. 2019. Neural machine translation: A review and survey. *arXiv preprint arXiv:1912.02047* (2019).
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017). <https://arxiv.org/pdf/1706.03762>
- [54] Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4396–4406. doi:10.18653/v1/D19-1448
- [55] Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2253–2263. doi:10.18653/v1/D17-1239
- [56] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *Journal of Machine Learning Research* (10 2019). <https://arxiv.org/pdf/1910.03771>
- [57] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics* 10 (3 2022), 291–306. <http://arxiv.org/abs/2105.13626>
- [58] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference* (3 2021), 483–498. <http://arxiv.org/abs/2010.11934>

A Hyperparameters

Table 1: Continued pretraining hyperparameters, following Meyer et al. [2024] and held identical across all three models.

Parameter	Value
Total steps	10,000
Warmup	5,000 steps
Schedule after warmup	Linear decay to 0
Learning rate	1×10^{-4}
Optimiser	AdamW
Effective batch	1,024
Sequence length	512

Table 2: Finetuning hyperparameters, held identical across all checkpoints

Parameter	D2T	MT
Optimiser	Adafactor	Adafactor
Learning rate	1×10^{-4}	1×10^{-5}
Schedule	Linear decay	Linear decay
Warmup	20% of steps	None
Epochs	5	5
Effective batch	4	16

Table 3: Probe training hyperparameters, following Müller-Eberstein et al. [2023] and held identical across all probes.

Parameter	Value
Optimiser	Adam
Learning rate	1×10^{-3}
Batch size	64
Maximum epochs	30
Early stopping	Validation loss, patience 3

B Ablations

We investigate whether the downstream effects of CPT observed with full finetuning data also hold when fewer task-specific examples are available. This is relevant for low-resource settings, where downstream training data may be limited. For D2T we compare training sets of size 1,000, 2,000 and the full 3,859 examples. For MT we compare 10,000, 25,000 and the full 50,000 parallel sentences. The hyperparameters remain fixed for comparison.

Data-to-Text. Nguni-ByT5’s advantage over ByT5 is clearer when finetuning data is limited. It achieves higher chrF at every checkpoint with 1,000 examples and at almost every checkpoint at 2,000 examples. As can be seen in Figure 4, neither byte model consistently leads at full finetuning. This suggests that prior Nguni adaptation results in better performance when fewer examples are available. T5 has the most to gain from more examples. It shows the largest chrF improvement as the dataset increases. Nguni-ByT5 shows the smallest improvement, with ByT5 falling between the two. This ordering is consistent with their prior isiXhosa exposure.

With 1,000 finetuning examples, as shown in Figure 9, T5’s chrF trajectory displays a non-monotonic decline, while Nguni-ByT5’s exhibits a non-monotonic increase. ByT5 displays a much sharper decline around step 5,000. With 2,000 examples, these patterns become less pronounced, and with full data, chrF remains close to each model’s base score without a consistent directional trend.

Machine Translation. With 10,000 examples, as shown in Figure 10, T5 and ByT5’s trajectories show a trend of decreasing performance followed by recovery. This trend changes as finetuning data increases. At 25,000 examples, they show a trend of decreasing performance followed by consistent increasing performance. For these models, CPT improves final translation performance with the larger finetuning sets, but not with the smallest set.

ByT5 shows a different pattern for isiXhosa-To-English translation. With 10,000 examples, its trajectory displays an initial decrease in performance followed by an overall upward trend. With 25,000 examples, this trend inverts. Therefore, increasing the finetuning data does not consistently show the benefit of CPT across translation directions.

These ablations show that some full data results do not hold to limited-data finetuning. The visible effects of CPT depend on the task and amount of data available. Evaluations done with less data might conclude different trends and results. We also find that increasing task-specific examples provides a more noticeable effect on performance as absolute performance values across the different finetuning budgets show a clear increase.

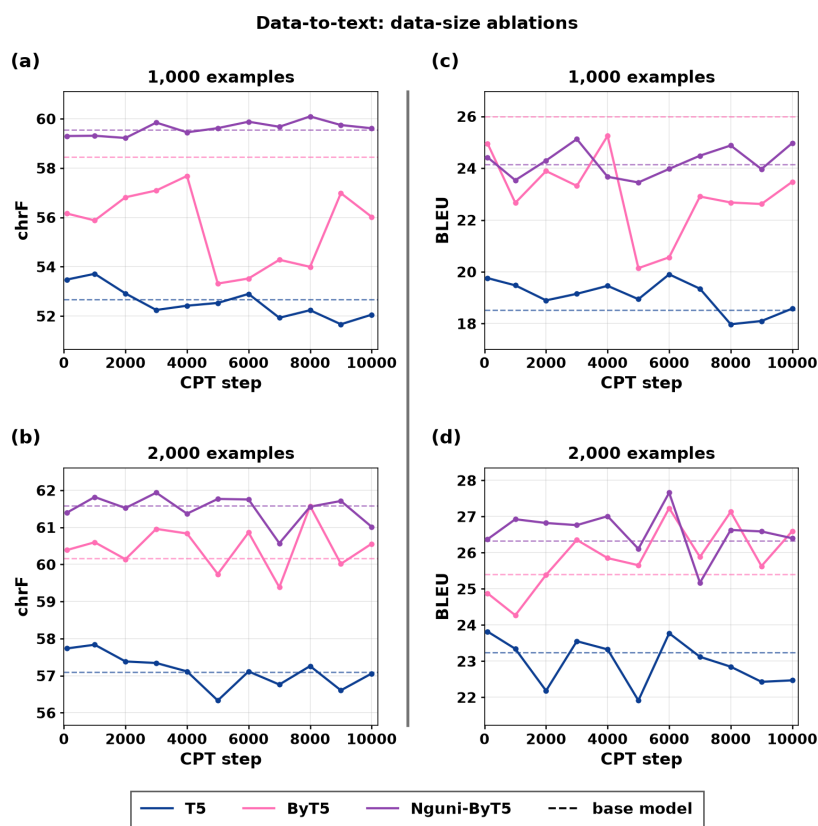


Figure 9: D2T chrF and BLEU trajectories for T5, ByT5 and Nguni-ByT5 at finetuning budgets of 1,000, 2,000 and 3,859 examples.

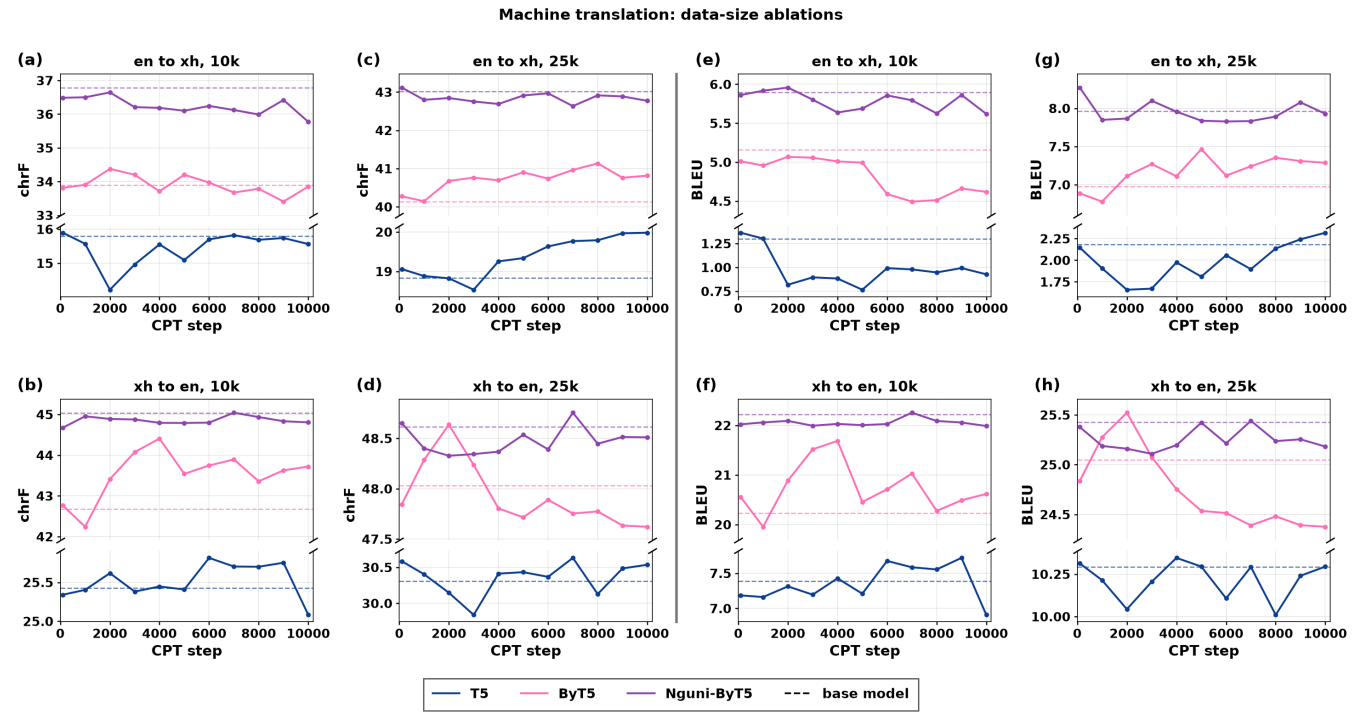


Figure 10: Machine translation chrF and BLEU trajectories for both translation directions at finetuning budgets of 10,000, 25,000 and 50,000 parallel sentences.