# Efficient Unseen Language Adaptation for Multilingual Pre-Trained Language Models

**Po-Heng Chen    Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
r11922044@csie.ntu.edu.tw  y.v.chen@ieee.org

## Abstract

Multilingual pre-trained language models (mPLMs) have demonstrated notable effectiveness in zero-shot cross-lingual transfer tasks. Specifically, they can be fine-tuned solely on tasks in the source language and subsequently applied to tasks in the target language. However, for low-resource languages **unseen** during pre-training, relying solely on zero-shot language transfer often yields sub-optimal results. One common strategy is to continue training PLMs using masked language modeling objectives on the target language. Nonetheless, this approach can be inefficient due to the need to adjust all parameters for language adaptation. In this paper, we propose a more efficient solution: soft-prompt tuning for language adaptation. Our experiments demonstrate that with carefully designed prompts, soft-prompt tuning enables mPLMs to achieve effective zero-shot cross-lingual transfer to downstream tasks in previously unseen languages. Notably, we found that prompt tuning outperforms continuously trained baselines on two text classification benchmarks, encompassing 20 low-resource languages while utilizing a mere 0.28% of the tuned parameters. These results underscore the superior adaptability of mPLMs to previously unseen languages afforded by soft-prompt tuning compared to traditional fine-tuning methods.[1]

## 1 Introduction

The issue of gathering sufficient annotated data for downstream tasks becomes particularly challenging for **low-resource** languages, leading to extensive research on **zero-shot cross-lingual transfer**. This basic approach involves fine-tuning a model using annotated data in a *source* language and evaluating its performance directly on data in a *target* language. Multilingual pre-trained language models (mPLMs) have demonstrated remarkable success in zero-shot cross-lingual transfer across various NLP tasks (Wu and Dredze, 2019; Conneau et al., 2020; Deshpande et al., 2022). Despite significant progress, it is impractical for mPLMs to cover all languages globally due to two primary reasons. First, including additional languages increases the vocabulary size, posing challenges in managing and maintaining. Second, there exists a trade-off between the number of languages covered and the model's capacity (Conneau et al., 2020). Consequently, when mPLMs encounter an unseen target language, the performance of zero-shot cross-lingual transfer often falls short of expectations. Therefore, the task of adapting mPLMs to **unseen** languages has emerged as a crucial concern. A naive approach involves continuing to train PLMs using the masked language modeling (MLM) objective on unlabeled text in the unseen target language, aiming to leverage the language-specific capabilities (Ebrahimi et al., 2022). However, with the increasing number of parameters in PLMs, tuning the *entire* model becomes more resource-intensive and inefficient. Additionally, in the case of low-resource unseen languages, the limited availability of data for continued training may compromise the model's generalizability if full parameters are tuned.

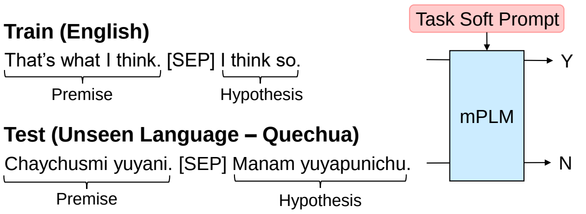
Task Soft Prompt
Train (English)
That's what I think. [SEP] I think so.
Premise
Hypothesis
Test (Unseen Language – Quechua)
Chaychusmi yuyani. [SEP] Manam yuyapunichu.
Premise
Hypothesis
mPLM
Y
N

Figure 1: An example of zero-shot cross-lingual transfer to an unseen language with soft-prompt tuning.

**Prompting** has emerged as a solution to avoid the overhead associated with fine-tuning by leveraging natural language prompts to query pre-trained

[1] The source code is publicly available at https://github.com/MiuLab/UnseenAdapt.

*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18983–18994
November 12-16, 2024 ©2024 Association for Computational Linguistics

|  | Source<br>(EN) | Seen<br>Target | Unseen<br>Target |
|---|---|---|---|
| Fine-tuning | 89.01 | 79.2 | 42.58 |
| Prompt-tuning | 88.94 | 79.9 | 43.35 |

Table 1: Gap between cross-lingual transfer to seen and unseen target languages. The scores of seen target languages are from Tu et al. (2022).

models. Tu et al. (2022) demonstrated the potential of prompt-tuning for zero-shot cross-lingual transfer among languages already included in the training data of mPLMs. Nevertheless, whether this approach sustains comparable performance when confronted with unseen target languages is still questionable. To address this question, we designed preliminary experiments to evaluate the performance of zero-shot cross-lingual transfer for both seen and unseen languages. We utilized XNLI (Conneau et al., 2018b) as the task for seen target languages and AmericasNLI (Ebrahimi et al., 2022) for unseen target language. The results are presented in Table 1. Notably, we observed a significant performance degradation when handling languages unseen by mPLMs, evident in both fine-tuning and prompt-tuning scenarios (79% to 43%). This underscores the necessity of establishing an effective and efficient adaptation mechanism for mPLMs to new languages prior to engaging in cross-lingual transfer.

Building upon prior research and empirical findings, our objective is to adapt mPLMs to previously **unseen low-resource** languages and achieve effective cross-lingual transfer. In this challenging scenario, the model lacks prior exposure to the target language, necessitating the development of novel strategies to accomplish our objective. At the same time, the adaptation process needs to be parameter-efficient, as tuning all parameters becomes impractical with the increasing scale of mPLMs. In this paper, we investigate the effectiveness of soft-prompt tuning for adapting mPLMs to **unseen** languages. Specifically, we keep all parameters in the mPLM frozen and solely focus on tuning the prefix soft prompts within the overall framework. Our results demonstrate that incorporating soft prompts significantly enhances the mPLM's ability to generalize to new languages, leading to a superior zero-shot cross-lingual performance on downstream tasks compared to fine-tuning. Figure 1 illustrates a simplified example of the zero-shot cross-lingual transfer process in our experiments. In summary, our contributions can be summarized in 3-fold:

- We are the first to extend the generalization of mPLMs to **unseen** languages using only soft-prompt tuning.
- We demonstrate that unseen low-resource language adaptation based on soft prompts outperforms fine-tuning in zero-shot cross-lingual transfer, even with only 0.28% of the parameters being tunable.
- Our results are comparable to MAD-X, a strong method for zero-shot cross-lingual transfer while utilizing 17 times fewer parameters.

## 2 Related Work

**Multilingual pre-trained language models (mPLMs)** Multilingual pre-trained models focus on learning language-agnostic embedding for a wide range of NLP downstream tasks. In the beginning, cross-lingual was achieved by aligning word level representation(Conneau et al. (2018a); Grave et al. (2018)). After the rise of the transformer-based pre-trained model, many multilingual pre-trained language models were proposed. mBERT and XLM, introduced by Devlin et al. (2019) and Lample and Conneau (2019) respectively, are both multilingual models trained without supervised cross-lingual alignment objectives. Conneau et al. (2020) proposed XLM-R, trained in one hundred languages solely with the masked language modeling (MLM) objective, leading to notable performance enhancements across various cross-lingual transfer tasks. Additionally, they identify several limitations of mPLMs, such as the transfer-dilution trade-off and the curse of multilinguality.

**Adapter** Applying adapters (Rebuffi et al., 2017) is one of the representative strategies to achieve parameter-efficient fine-tuning. When performing adapter tuning, the original model's weight is untouched and only the newly added adapter layers are tuned. Houlsby et al. (2019) utilize adapter to achieve transfer learning to multiple downstream tasks and attain near fine-tuning performance while having significant trainable-parameter reduction. Pfeiffer et al. (2020) introduced MAD-X, a modular framework designed to perform parameter-efficient cross-lingual transfer through the use of adapters. Their framework trains a language adapter (LA) for each source and target language using MLM, and a task adapter (TA) for the task from the source languages, incorporating the corresponding LA during training. During inference, their method achieves impressive task performance on the target language by stacking the task adapter with the target language adapter, showcasing its great potential for zero-shot cross-lingual transfer.

**Prompt tuning** Prompt tuning emerged as a strategy to leverage knowledge from pre-trained models and avoid the overhead associated with fine-tuning. For instance, Lester et al. (2021) use trainable continuous prompts as input tokens for the pre-trained model and show that prompt-tuning becomes more competitive with scale. Similarly, Liu et al. (2022) incorporate prefix embeddings into each layer of the pre-trained model, resulting in more direct impacts on model predictions and achieving performance on par with fine-tuning across various model scales. Tu et al. (2022) demonstrated the potential of prompt-tuning for zero-shot cross-lingual transfer. However, their experiments primarily concentrated on cross-lingual transfer among languages that were already included in the training data of the mPLMs. In contrast, our objective is to adapt mPLMs to previously **unseen** and **low-resource** languages, which poses a more challenging scenario.

## 3 Soft-Prompt Language Adaptation

In this paper, our focus lies in investigating the effectiveness of soft-prompt tuning in adapting mPLMs to previously unseen low-resource target languages. We aim to evaluate its performance in zero-shot cross-lingual transfer across downstream tasks. The experimental procedure comprises two stages, as illustrated in Figure 2.

### 3.1 MLM on Unlabeled Data

The first stage aims to adapt the mPLM to the target language, which has not been seen before. To ensure the adaptability of soft prompts across both the source language and the unseen target language, we combine unlabeled data from both languages and fine-tune the soft prompts using the masked language model objective. Typically, the source language is relatively high-resource, allowing us to obtain more unlabeled data. However, to prevent the adapted model from being overly biased towards the source language, we adjust the amount of source language unlabeled data used based on the quantity of unlabeled data available for each target language, aiming for a balanced distribution.

Our soft-prompt tuning framework follows the design proposed by Liu et al. (2022), which incorporates tunable prefix tokens into each layer illustrated in the left part of Figure 2. This design offers a more direct influence on the model's predictions by modifying the output of each layer.

### 3.2 Tuning on Source-Language Labeled Data

The second stage involves performing soft-prompt tuning on the downstream task in the source language and subsequently transferring the model to the target language in a zero-shot manner. First, we use the tuned soft-prompt in stage 1 as initialization for this stage. This design aligns with the *Soft Prompt Transfer* proposed by Vu et al. (2022). To further leverage the capacity of masked token prediction from the soft-prompt obtained in stage 1 (see Section 3.1), we employ a **template** and a **verbalizer** (Schick and Schütze, 2021) to transform the input of the downstream task into a masked language modeling problem. For more details about our implementation, please refer to Appendix A.

Since we only tune the soft-prompt on source-language labeled data, we need to ensure the transferability of tuned model and avoid the *catastrophic forgetting* (McCloskey and Cohen, 1989) on the target language. Previous works found that the upper layers of mPLMs are more task-focused and language-independent (Libovický et al., 2020; Foroutan et al., 2022). Based on this observation, we propose to tune only the soft-prompt of **Top-K** layers, shown in the right part of Figure 2. Section 5.4 provides further analysis regarding the selection of $K$. By tuning only the prompts in the upper layers, we can enhance cross-lingual transferability by limiting task-focus capacity to stay in language-dependent top layers and preserving language-dependent information in the lower layers. After training, we evaluate our model directly on the corresponding task in the target language.

**Objective function** Let $M$ be the parameters of the mPLM and $N$ be the number of layers in the mPLM. Furthermore, we denote the parameters of the soft prompts as $\theta = \{\theta_1, \theta_2, ..., \theta_N\}$, where $\theta_i$ is the parameters of the soft prompt of the i-th layer. Finally, we define a verbalizer $v$ to be a function that maps each label to a specific token which can represent the meaning of that label. The probability of classifying input $x$ as label $y$ can be represented

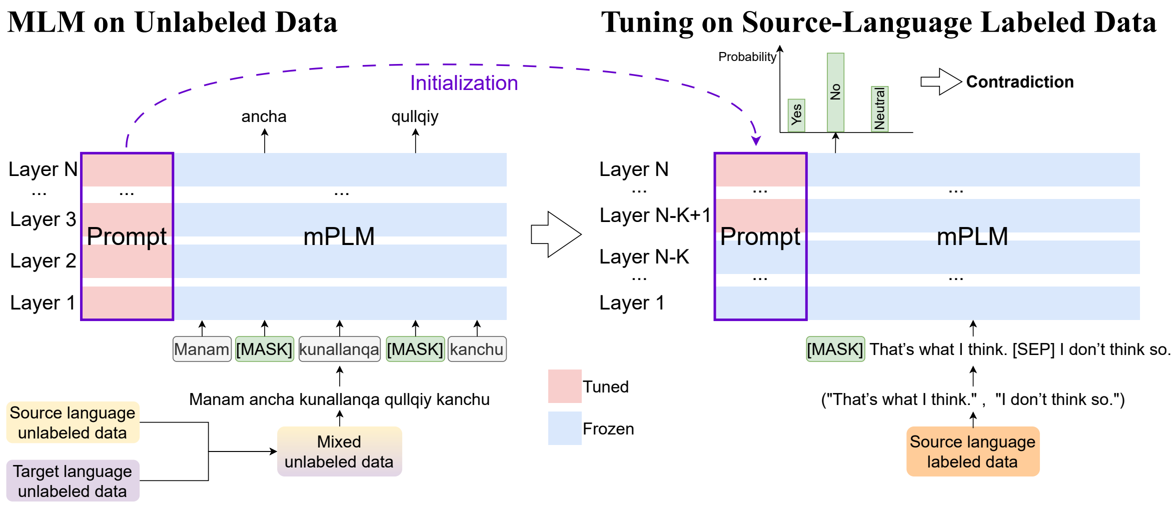
MLM on Unlabeled Data
Tuning on Source-Language Labeled Data
Initialization
ancha
qullqiy
Layer N
...
Layer 3
Layer 2
Layer 1
Prompt
mPLM
Manam
[MASK]
kunallanqa
[MASK]
kanchu
Manam ancha kunallanqa qullqiy kanchu
Source language unlabeled data
Target language unlabeled data
Mixed unlabeled data
Tuned
Frozen
Probability
Yes
No
Neutral
Contradiction
Layer N
Layer N-K+1
Layer N-K
Layer 1
Prompt
mPLM
[MASK] That's what I think. [SEP] I don't think so.
("That's what I think." , "I don't think so.")
Source language labeled data

Figure 2: Illustration of our soft-prompt language adaptation. **Left**: In the first stage(Sec. 3.1), we perform soft-prompt tuning on unlabeled data in both source and unseen target languages via MLM for language adaptation. **Right**: In the second stage(Sec. 3.2), we initial the soft prompts based on the results from the first stage. The soft prompts of selected layers are then fine-tuned using a template and a verbalizer specific to the downstream task in the source language.

as:

$$P(y \mid x, \theta, M) = P(\langle \text{mask} \rangle = v(y) \mid x, \theta, M) \quad (1)$$

The training objective is to maximize the likelihood of verbalized label tokens predicted by the MLM head, tuning only the soft prompts in the last $K$ layers :

$$\underset{\theta_{N-K+1}, \theta_{N-K+2}, \ldots, \theta_N}{\operatorname{argmax}} \sum_x P(y \mid x, \theta, M) \quad (2)$$

After training, we evaluate our model directly on the corresponding task in the target language $t$.

# 4 Experiments

## 4.1 Data

To evaluate the effectiveness of prompt-based language adaptation, we conduct experiments on two text classification datasets including multiple **low-resource** languages.

1. **MasakhaNEWS** (Adelani et al., 2023): This dataset focuses on news topic classification and encompasses 16 languages commonly spoken in Africa. Since our emphasis is on languages **unseen** by the mPLMs, we only utilize eight of these languages, which are unseen by XLM-R, for evaluation. During the first stage of our procedure, we use the news articles of each language in the training set as unlabeled data, as only the testing set is required to assess zero-shot performance. Furthermore, Hausa (hau) is selected as the source language since it covers all the news topics used in MasakhaNEWS.

2. **AmericasNLI** (Ebrahimi et al., 2022): Extending XNLI (Conneau et al., 2018b), this dataset incorporates ten Indigenous languages of the Americas, all of which are characterized by limited linguistic resources and are unseen by the XLM-R model. The unlabeled data for these languages are accessible via the AmericaNLP repository.[2] English serves as the source language, and the MultiNLI dataset (Williams et al., 2018) is employed as labeled data for the second stage of our procedure.

The comprehensive list of languages along with their corresponding unlabeled data sources is provided in Appendix B. As illustrated in Table 4, the quantity of unlabeled data for languages in MasakhaNEWS averages around 1K, whereas for languages in AmericasNLI, it ranges from 4K to 125K.

[2] https://github.com/AmericasNLP/americasnlp2021

### 4.2 Setup

In this study, all experiments are conducted using the XLM-R model (Conneau et al., 2020) of LARGE size as the baseline. We set the length of the soft prompt to 32. For soft-prompt tuning on unlabeled text in stage 1, we ensured that the amount of English data used was equivalent to the amount of data available in the target language. Furthermore, We masked 15% tokens for each input sentence. In this stage, We train soft prompts on unlabeled data for every target language for 100K steps, with a batch size of 32 and a learning rate of 5e-3. For soft-prompt tuning on the downstream task in stage 2, we set the number of trainable layers of soft-prompt $K$ to 18. In this stage, the soft prompts are trained on labeled data in the source language for 10 epochs, with a batch size of 32 and a learning rate of 1e-3.

### 4.3 Baselines

**Fine-tuning based baselines** Our main baseline for zero-shot cross-lingual transfer is fine-tuning the full XLM-R (Conneau et al., 2020). We compare the zero-shot performance of XLM-R with and without adaptation to the target language, following the approach outlined by Ebrahimi et al. (2022). In the *without adaptation* setting, XLM-R is fine-tuned on the training set in the source language and directly evaluated on the testing set in target languages. In the *with adaptation* setting, XLM-R is additionally trained on unlabeled data in the target language using the MLM objective before fine-tuning. For tuning on unlabeled data in the target language, we set batch size to 32 and learning rate to 5e-3. For tuning on the downstream task in the source language, we set batch size to 32 and learning rate to 1e-6.

**Prompt-tuning zero-shot transfer** Tu et al. (2022) demonstrated the effectiveness of prompt-tuning for zero-shot cross-lingual transfer. We adopt a similar experimental setup to theirs, where we keep XLM-R frozen, conduct soft-prompt tuning using the training set of the source language, and evaluate the performance of zero-shot cross-lingual transfer on the testing set in target languages. We set the length of the soft prompt to 32, with a batch size of 32 and a learning rate of 1e-3.

**Adapter-based language adaptation** To further compare with parameter-efficient language adaptation approaches, we use MAD-X, an adapter-based language adaptation method, as an additional benchmark. For a fair comparison, we perform MAD-X experiments using the XLM-R checkpoint of LARGE size. As the original MAD-X paper (Pfeiffer et al., 2020) does not cover the target language we conduct experiments on, we retrain their models on unlabeled data for all target languages. For English, we utilize the checkpoint provided in their AdapterHub [3]. We adopt the adapter reduction factors specified in the original paper, which are 2 for language adapters and 16 for task adapters. For training both types of adapters, we set batch size to 32 and learning rate to 1e-4.

### 4.4 Results

Table 2 presents the results of zero-shot cross-lingual transfer on target languages in MasakhaNews and AmericasNLI, respectively. First, it can be observed that purely zero-shot transfer without adapting to the unseen target language results in a lower average accuracy. All models with adaptation outperform all models without adaptation, highlighting the necessity of adapting to **unseen** languages. Second, our soft-prompt-based adaptation method (Ours) demonstrates comparable zero-shot cross-lingual transfer performance to the best baseline in both datasets, despite introducing relatively *fewer trainable parameters* compared to other baselines. This underscores the effectiveness and parameter-efficiency of soft-prompt-based language adaptation, as well as its generalizability across different types of classification datasets. Further details regarding the comparison of trainable parameter quantities are presented in 5.1.

Finally, the impact of unlabeled data volume discrepancies in the target language on various language adaptation methods can also be observed from Table 2. Adapter-based language adaptation (MAD-X) demonstrates better performance with a higher volume of target language data, as evidenced by the results for AmericasNLI. Conversely, in MasakhaNEWS, where target language data is relatively scarce, MAD-X is less effective compared to fine-tuning. Soft-prompt-based language adaptation shows consistently good average scores across both datasets, indicating its higher versatility and stable performance regardless of the quantity of target language data. This versatility is particularly

[3] https://adapterhub.ml

| Model | ibo | lin | lug | pcm | run | sna | tir | yor | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| *Zero-shot* | | | | | | | | | |
| Fine-tuning | 67.95 | 74.86 | 60.54 | 93.11 | 69.25 | 58.27 | 66.54 | 67.64 | 69.77 |
| Prompt-tuning | 64.36 | 66.86 | 43.50 | 91.15 | 63.35 | 49.32 | 54.41 | 67.40 | 62.54 |
| *Zero-shot w/ adaption* | | | | | | | | | |
| Fine-tuning | 81.03 | **85.71** | 61.43 | **95.41** | **85.71** | 81.03 | 72.43 | 83.45 | 80.78 |
| MAD-X | 78.97 | 78.86 | 56.05 | 86.23 | 76.71 | 73.98 | **73.16** | 77.86 | 75.23 |
| Ours | **81.62** | 82.48 | **71.15** | 91.59 | 85.17 | **86.68** | 72.79 | **85.32** | **82.10** |

(a) MasakhaNEWS.

| Model | aym | bzd | cni | gn | hch | nah | oto | quy | shp | tar | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Zero-shot* | | | | | | | | | | | |
| Fine-tuning | 40.67 | 41.33 | 43.07 | 42.93 | 39.20 | 45.39 | 42.25 | 42.13 | 48.27 | 40.53 | 42.58 |
| Prompt-tuning | 42.13 | 41.47 | 44.67 | 44.53 | 39.07 | 45.93 | 43.45 | 44.40 | 48.00 | 39.86 | 43.35 |
| *Zero-shot w/ adaption* | | | | | | | | | | | |
| Fine-tuning | 48.00 | 44.80 | **44.93** | 56.00 | **42.40** | 47.70 | 42.51 | 49.73 | **46.40** | 42.67 | 46.51 |
| MAD-X | **60.93** | **46.00** | 41.73 | **62.27** | 37.33 | 47.29 | 42.25 | **65.73** | 46.13 | **43.20** | **49.29** |
| Ours | 59.51 | 42.84 | 44.04 | 60.31 | 40.71 | **47.97** | **43.09** | 63.60 | 44.67 | 39.15 | 48.59 |

(b) AmericasNLI.

Table 2: The cross-lingual transfer results for soft prompt language (Ours) adaptation and each baseline. For Ours, the results are averaged across 3 runs.

| Method | Trainable Parameter | Checkpoint Size |
|---|---|---|
| Fine-tuning | 816M | 2.24GB |
| MAD-X | 27M | 103MB |
| Ours | **1.57M** | **6.2MB** |

Table 3: The number of trainable parameters of each language adaptation method and the checkpoint size for one language.

crucial in low-resource language scenarios, where acquiring substantial amounts of high-quality unlabeled data may not be feasible.

## 5 Analysis

### 5.1 Parameter and Storage Efficiency

Table 3 shows the trainable parameters needed by each baseline method and our soft prompt language adaptation (Ours). The XLM-R baseline fine-tunes the *entire* XLM-R Large model, which has 560M parameters. MAD-X with a reduction factor of 2 requires approximately 27M parameters. In contrast, our model has only 1.57M tunable parameters in the soft prompts, accounting for approximately 0.28% of the original model's parameters and 17 times fewer tunable parameters than MAD-X.

In addition, we also showed the disk space needed to store a checkpoint when introducing a new language. The XLM-R baseline needs an *entire* new model when adapted to a new language, which costs the most space to store. MAD-X needs a language adapter for a single language, while soft prompt language adaptation requires a new set of prefix soft prompts, which costs much less space than the adapter. This significant reduction in tunable parameters and storage demonstrates the efficiency and practicality of our framework.

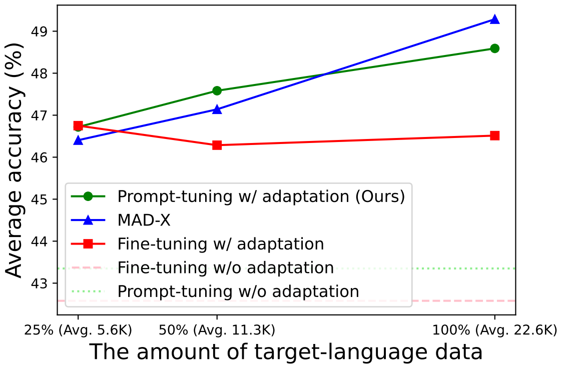

Figure 3: The average performance on AmericasNLI (Ebrahimi et al., 2022) against different amount of target language unlabeled data.

### 5.2 Size of Target Language Unlabeled Data

In this section, we look into analyzing the impact of varying amounts of unlabeled data in the target language on performance. As discussed in Section 4, there exists a notable contrast in the quantity of unlabeled data available between MasakhaNews and AmericasNLI, resulting in performance discrepancies across different language adaptation methods. Hence, our objective is to further examine the influence of data quantity on these language adaptation techniques. In the experiment here, we focus on AmericasNLI and systematically reduce the proportion of unlabeled data in the target language to 25% and 50% for comparative analysis.

Figure 3 shows the zero-shot transfer performance at each data quantity level. The figure indicates that fine-tuning lacks a discernible correlation with the amount of data available. This observation can be attributed to the fact that, in our experiment, even at full data usage (100%), the average dataset size is only 22.6K. This size is insufficient for fine-tuning to achieve stable performance. Conversely, the other two parameter-efficient tuning methods, MAD-X and soft-prompt, show significant improvement with increasing unlabeled data. Previous studies have suggested that parameter-efficient tuning demonstrates better generalizability and achieves superior performance compared to fine-tuning when there is limited labeled data available for downstream tasks (Li and Liang, 2021). Our experiments reveal a similar trend when using unlabeled data for language adaptation. Additionally, soft-prompt tuning outperforms MAD-X when there is a relatively small amount of unlabeled data. This is consistent with observations from MasakhaNEWS, which has only around 1,000 unlabeled data points, where soft-prompt-tuning shows superior performance. These findings suggest that soft-prompt tuning is particularly effective for truly low-resource target languages.

### 5.3 Few-shot Evaluation

We further conduct a few-shot evaluation to see the models' generalizability when encountering extremely few labeled data for downstream tasks. Here, we employ MasakhaNEWS (Adelani et al., 2023) for evaluation. Similar to the previous experiments, we use Hausa (hau) as the source language. However, in this case, we reduce the downstream labeled data to 5, 10, 20, and 50 samples per class. The models are trained on such limited data and then evaluated on the testing set of target languages for zero-shot cross-lingual transfer.

Figure 4 presents our few-shot results. We observe that models without language adaptation perform poorly in few-shot scenarios, and soft-prompt-tuning performs considerably worse than fine-tuning. Without prior training on the target language to initialize soft-prompt effectively, satisfactory performance cannot be achieved with insufficient labeled data. In cases with adaptation, both methods exhibit significant performance improvement, underscoring the importance of language adaptation for enhancing knowledge in an unseen language. Once proficiency in the target language is attained, favorable cross-lingual transfer results can be achieved even with few-shot labeled data. Additionally, we note that across most different shot numbers, prompt tuning-based adaptation yields better cross-lingual transfer performance than fine-tuning. This again aligns with previous findings that soft-prompt-tuning has superior generalizability than fine-tuning when dealing with few-shot labeled data (Li and Liang, 2021).

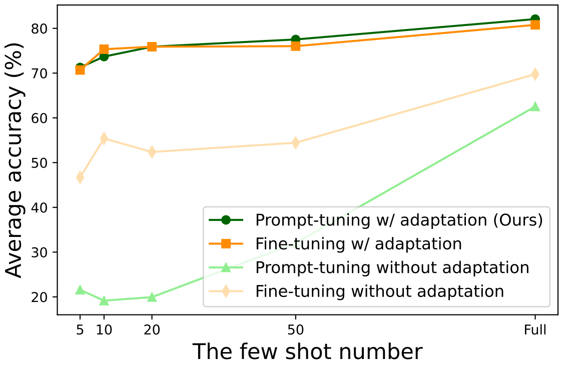

Figure 4: The average few-shot performance on MasakhaNEWS (Adelani et al., 2023).

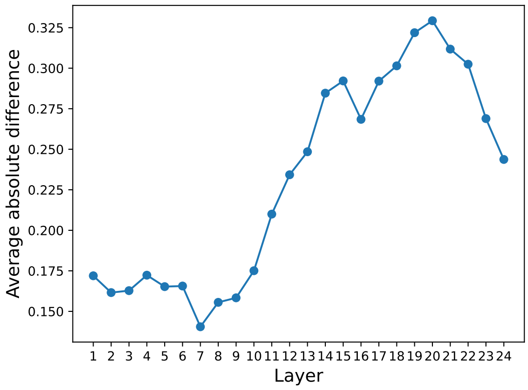

Figure 5: The changes in parameter values of soft prompts at each layer. Experiment is conducted on MasakhaNEWS (Adelani et al., 2023).

### 5.4 Trainable Soft-Prompt Layers for Downstream Tasks

In this section, we delve into the performance analysis of various configurations of trainable soft-prompt layers for downstream tasks (refer to Section 3.2). Firstly, we aim to validate whether the upper layers actually exhibit more task-centric behavior. To achieve this, we fine-tune the soft prompts across all layers ($K = 24$) and quantify the resultant change in parameter values. This change is defined as the absolute difference between the parameter values of the soft prompts before and after tuning. Figure 5 illustrates the changes observed in soft prompts across each layer. The upper layers exhibit larger changes in parameters compared to the lower ones, with layer 20 demonstrating the most significant change. This discovery indicates that the upper layers have more influence and importance during fine-tuning for downstream tasks, thereby corroborating the hypothesis of our design.

Secondly, we vary the value of $K$, representing the number of trainable top layers during the labeled data tuning stage, and evaluate each configuration using the test set from MasakhaNEWS (Adelani et al., 2023). We refer to this setting as **Top K**. Additionally, we conducted experiments where the trainable layers are set to the bottom $K$ layers, referred to as **Bottom K**, and compared their performance. Figure 6 illustrates the average accuracy scores (in percentage) for each $K$ selection, both for the top and bottom layers settings. From the figure, we can see a notable impact on the model's performance based on the selection of trainable layers in the **Top K** setting. When $K$ approaches 24, indicating nearly all layers have trainable soft prompts, there's a risk of overfitting to the source language and loss of target language knowledge acquired during the unlabeled data MLM stage. Conversely, as $K$ decreases towards 1, indicating only a few top layers' prompts are adjustable, the model may lack the capacity to significantly adapt its output. In addition, by comparing the **Top K** and **Bottom K** settings, we find that selecting the top K layers yields better performance compared to the bottom K layers when the number of trainable layers is the same. For instance, there is a significant performance gap (80.53 versus 63.04) between the top 6 layers and the bottom 6 layers. This discovery again validates the hypothesis upon which our design is based: the upper layers are more task-focused and language-independent.

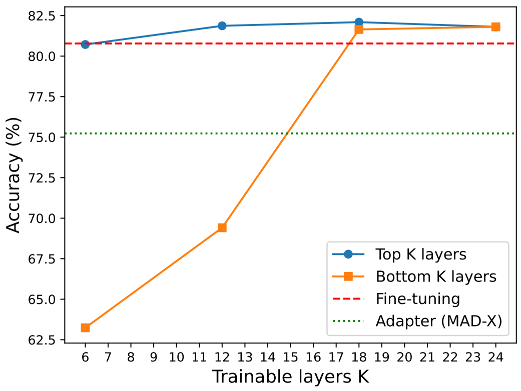

Figure 6: The average performance on MasakhaNEWS (Adelani et al., 2023) with varying trainable layers on source-language labeled data. When the same amount of layers are trainable, setting them on top layers yields better performance than on bottom layers.

Combining the findings above, we can conclude that it's critical to carefully choose layers for prompt tuning after target-language adaptation. Selecting upper layers for prompt adjustment in downstream tasks while preserving target language information in lower layers is essential to prevent catastrophic forgetting and ensure effective zero-shot cross-lingual transfer. Though the selection of the tunable layers affects the performance of zero-shot transfer, the performance of the majority of choices for parameter K for **Top K** setting still outperforms traditional fine-tuning baseline. These findings provide compelling evidence of the efficacy of soft-prompt language adaptation.

## 6 Conclusion

In this paper, our objective is to adapt multilingual pre-trained language models(mPLMs) to previously unseen languages and enhance their cross-lingual transferability. Initially, we demonstrate the necessity of adapting mPLMs to new languages by comparing their performance in cross-lingual transfer between known and unknown languages. Subsequently, we propose using soft-prompt tuning to accomplish efficient language adaptation and effective zero-shot cross-lingual transfer to downstream tasks. Our primary findings indicate that employing soft-prompt tuning for language adaptation can yield comparable performance to baseline methods such as fine-tuning and adapter-based approaches, while utilizing significantly fewer tunable parameters. Furthermore, we conduct various experiments to delve deeper into soft-prompt-based language adaptation, examining factors such as data and prompt settings. Our experiments reveal that the soft prompts in lower layers function as a language-dependent component while tuning only the soft prompts in upper layers for downstream tasks leads to improved results. Additionally, soft-prompt-based language adaptation demonstrates consistent performance even with limited amounts of unlabeled target language data and few-shot downstream data. These findings collectively affirm the superior efficiency of soft-prompt-based language adaptation, both in terms of the number of trainable parameters and the volume of data.

## Limitations

In this work, we only focus on masked language-based models. We leave the application of our framework to the generative-based model as future work. Besides, our current experiment only utilizes XLM-R as the backbone model. If there are a significant number of words or characters in the unseen language that cannot be properly encoded by XLM-R, it may affect the performance of its zero-shot cross-lingual transfer. In our future work, we plan to conduct experiments with byte-level models to address this particular limitation and explore alternative approaches to mitigate this restriction.

Additionally, we observe in the results that different methods perform differently across languages, each with their own strengths. Our experiments suggest that the volume of unlabeled data for each language is one of the factors influencing performance. However, other linguistic factors may also affect the adaptation results, such as language family, text structure, and so on. These potential factors were not explored in this study, but we hope to investigate them further in future work.

## Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 112-2223-E002-012-MY5. We thank to National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

## References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. Masakhanews: News topic classification for african languages.

Feliciano Elizondo Adolfo Constenla and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

David Brambila. 1976. *Diccionario raramuri - castellano: Tarahumar.*

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. http://www.lengamer.org/ publicaciones/diccionarios/.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 2022. 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)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri. http://bribri.net.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Erwin Lauriout James Loriot and Dwight Day. 1993. *Diccionario Shipibo-Castellano*. Instituto Lingüístico de Verano.

Carla Victoria Jara Murillo. 2018a. *Gramática de la Lengua Bribri*. EDigital.

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris, second edition*. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. EDigital.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri, second edition*. Editorial de la Universidad de Costa Rica.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland. European Association for Machine Translation.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

## A Template and verbalizer

For the template and verbalizer, we follow the notation in Schick and Schütze (2021). Let $V$ be mPLM's token vocabulary which contains mask token $[MASK]$, and $L$ be the set of labels of the downstream task. We denote an input of the downstream as a sequence of phrases $x = (s_1, s_2, ..., s_n)$ where $s_i \in V*$ and the corresponding label as $y$. Then We define a template to be a function $T$ that converts an input x to a sequence of tokens $T(x) \in V^*$. Finally, we define a verbalizer $v : L \to V$ to be a function that maps each label to one token.

**Template** The template converts the origin input phrases to another sequence that contains only one mask token. Taking Natural Language Inference (NLI) as an example. The input sequence contains two phrases, the premise $p$ and the hypothesis $h$. The origin input sequence can be represented as $x = (p, h)$. The task involves analyzing the relationship between them. Assume the template concatenate mask token, the premise $p$, question mark, the mask token, and the hypothesis $h$, the actual input $T(x)$ will:

| Source | Target | Data Source(s) | Size |
|---|---|---|---|
| Hausa (hau) | Igbo (ibo) | Adelani et al. (2023) | 1.4K |
| | Lingala (lin) | Adelani et al. (2023) | 0.6K |
| | Luganda (lug) | Adelani et al. (2023) | 0.8K |
| | Naija (pcm) | Adelani et al. (2023) | 1K |
| | Rundi (run) | Adelani et al. (2023) | 1.1K |
| | chiShona (sna) | Adelani et al. (2023) | 1.2K |
| | Tigrinya (tir) | Adelani et al. (2023) | 0.9K |
| | Yorùbá (yor) | Adelani et al. (2023) | 1.4K |

(a) MasakhaNEWS.

| Source | Target | Data Source(s) | Size |
|---|---|---|---|
| English (en) | Aymara (aym) | Tiedemann (2012) | 6.5K |
| | Bribri (bzd) | Feldman and Coto-Solano (2020); Margery (2005); Jara Murillo (2018a); Adolfo Constenla and Pereira (2004); Jara Murillo and García Segura (2013); Jara Murillo (2018b); Flores Solórzano (2017) | 7.5K |
| | Asháninka (cni) | Cushimariano Romano and Sebastián Q (2008) | 3.8K |
| | Guarani (gn) | Chiruzzo et al. (2020) | 26K |
| | Wixarika (hch) | Mager et al. (2018) | 8.9K |
| | Náhuatl (nah) | Gutierrez-Vasques et al. (2016) | 16K |
| | Otomí (oto) | https://tsunkua.elotl.mx | 4.8K |
| | Quechua (quy) | Agić and Vulić (2019) | 125K |
| | Rarámuri (tar) | Galarreta et al. (2017); James Loriot and Day (1993); Montoya et al. (2019) | 14K |
| | Shipibo-Konibo (shp) | Brambila (1976) | 14K |

(b) AmericasNLI.

Table 4: List of the languages and the source of the unlabeled data for each of them used in our experiments.

$$T(x) = T(p, h) = p\,?\,\text{[MASK]}\,h$$

**Verbalizer** We define a specific set of vocabulary tokens for each label, which can consist of tokens from the source language $s$, the target language $t$, or even other languages. After using the MLM head of mPLM to extract the most likely substitute token among the verbalizer's range, we map it back to the corresponding label as the prediction. Take NLI for example, we can define $v(\text{entailment}) = \{\text{Yes}, v(\text{contradiction}) = \{\text{No}\}$, and $v(\text{neutral}) = \{\text{Neutral}\}$. If the label is *entailment*, the model should predict *Yes* on the mask token.

## B Languages

The unlabeled low-resource target languages are detailed in Table 4.