Don't Stop Fine-Tuning: On Training Regimes for Few-Shot Cross-Lingual Transfer with Multilingual Language Models

Fabian David Schmidt[1], Ivan Vulić[2,3], Goran Glavaš[1]
[1] Center For Artificial Intelligence and Data Science, University of Würzburg, Germany
[2] Language Technology Lab, University of Cambridge, UK [3] PolyAI Ltd., UK
{fabian.schmidt, goran.glavas}@uni-wuerzburg.de
iv250@cam.ac.uk

Abstract

A large body of recent work highlights the fallacies of zero-shot cross-lingual transfer (ZS-XLT) with large multilingual language models. Namely, their performance varies substantially for different target languages and is the weakest where needed the most: for low-resource languages distant to the source language. One remedy is *few-shot transfer* (FS-XLT), where leveraging only a few task-annotated instances in the target language(s) may yield sizable performance gains. However, FS-XLT also succumbs to large variation, as models easily overfit to the small datasets. In this work, we present a systematic study focused on a spectrum of FS-XLT fine-tuning regimes, analyzing key properties such as *effectiveness*, *(in)stability*, and *modularity*. We conduct extensive experiments on both higher-level (NLI, paraphrasing) and lower-level tasks (NER, POS), presenting new FS-XLT strategies that yield both improved and more stable FS-XLT across the board. Our findings challenge established FS-XLT methods: e.g., we propose to replace sequential fine-tuning with joint fine-tuning on source and target language instances, offering consistent gains with different number of shots (including resource-rich scenarios). We also show that further gains can be achieved with multi-stage FS-XLT training in which joint multilingual fine-tuning precedes the bilingual source-target specialization.

1 Introduction and Motivation

Successful fine-tuning of mainstream pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020) for various NLP tasks requires a sizeable set of labeled task-specific instances. While such abundant task data are available for many tasks in English and a few high-resource languages, annotated examples are much scarcer for low-resource languages (Joshi et al., 2020). A large body of recent work thus focused on zero-shot cross-lingual transfer (ZS-XLT), for which no labeled instances are available in the target language (Pires et al., 2019; Cao et al., 2020). Catalyzed by pretrained massively multilingual transformers (MMT) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), or mT5 (Xue et al., 2021), ZS-XLT has achieved impressive results on a wide variety of tasks (Hu et al., 2020; Ruder et al., 2021). The MMT-driven ZS-XLT, however, exhibits dramatic performance drops when transferring to low-resource languages and/or languages distant from the source language (Lauscher et al., 2020; Ebrahimi et al., 2021; Adelani et al., 2021, *inter alia*). In contrast, recent work highlights that language models are excellent few-shot learners (Brown et al., 2020; Gao et al., 2021): they adapt well to new tasks or languages when exposed to only on a handful of labeled instances.

For cross-lingual transfer in particular, *sequential* few-shot transfer (FS-XLT) – in which large(r)-scale fine-tuning in the source language is followed by the secondary fine-tuning on a few target language instances – has been rendered particularly effective, with massive performance gains reported for some tasks with as little as 10 target language instances (Lauscher et al., 2020; Zhao et al., 2021). However, the effectiveness of sequential FS-XLT crucially depends on the shot selection (Zhao et al., 2021). Even more concerning, as we show in §3, is the sensitivity of FS-XLT to hyperparameter values, most notably the duration (number of epochs) of few-shot target language training: such fluctuations are problematic for *true* few-shot learning (Perez et al., 2021), where target language validation data, to be leveraged for model selection, does not exist.

Contributions. In this work, we shed new light on FS-XLT and seek to remedy the above pitfalls of current FS-XLT method. We depart from the established sequential FS-XLT paradigm and propose new training regimes for FS-XLT, comparing them across the dimensions of effectiveness, stability, and modularity. Concretely, we propose training regimes that jointly exploit source and target

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10725–10742
December 7-11, 2022 ©2022 Association for Computational Linguistics

language instances, and allow to model their interaction. **1)** We demonstrate, both for higher-level semantic tasks (e.g., NLI) and lower-level token-level tasks (NER, POS tagging), that joint source and target language training 'feeds two birds with one scone': (i) it consistently improves FS-XLT performance, even in setups with a larger number of target-language shots (e.g., $N = 500$), and (ii) makes the training procedure much more stable and robust, allowing for a reliable selection of the model checkpoint in true few-shot transfer setups without a target-language validation set. **2)** We find that preceding the joint bilingual fine-tuning with a multilingual training step, in which we combine the shots from multiple target languages, brings further performance gains. We also show that such multi-stage training regime improves the computational efficiency in multilingual FS-XLT setups, i.e., when the model transfer to multiple target languages is required. **3)** Finally, we validate that benefits of the new FS-XLT training regimes are not limited to English as the source language. Our work fundamentally challenges the status quo in FS-XLT and introduces and compares training paradigms that enable more effective, more efficient, and much more robust few-shot cross-lingual transfer.

Concurrent (closely related) effort. The concurrent work of Xu and Murray (2022) similarly demonstrates the utility of joint multilingual FS-XLT: although their joint fine-tuning approach differs from ours – they employ gradient surgery (Yu et al., 2020), an approach that harmonizes competing gradients originating from instances of different languages in a training batch – it yields the same two main benefits: (1) improved target language performance and (2) more stable training that facilitates models selection (i.e., alleviates the need for target-language validation data).

2 Background and Related Work

MMTs like mBERT and XLM(-R) (Lample and Conneau, 2019; Conneau et al., 2020) have become the main vehicles of cross-lingual transfer. Pretrained on multilingual corpora covering 100+ languages, MMTs conceptually enable zero-shot cross-lingual transfer (ZS-XLT) between any two languages seen in pretraining (Hu et al., 2020) or even to unseen languages (Ansell et al., 2021). The (extent of) ZS-XLT success depends on the quality and alignment of the representation subspaces of individual languages (Cao et al., 2020; Hu et al., 2021; Wu and Dredze, 2020). Accordingly, ZS-XLT with MMTs tends to be ineffective in transfers to target languages that are (i) linguistically distant from the source language and especially those (ii) un(der)represented in MMT's pretraining (Hedderich et al., 2020; Lauscher et al., 2020; Ruder et al., 2021; Ebrahimi et al., 2021).

One line of work boosts ZS-XLT by improving semantic alignment between the representation subspaces of individual languages, exploiting to this end available word or sentence translations (Hu et al., 2021; Wu and Dredze, 2020; Yang et al., 2022). Another, complementary line of work improves ZS-XLT through increasing the MMT's capacity for individual languages (Pfeiffer et al., 2020, 2022; Ansell et al., 2021, 2022). It attempts to remedy for the "curse of multilinguality" (Conneau et al., 2020) – an effect where, for a fixed model capacity, the quality of representations of individual languages at some point starts degrading with the addition of more languages.

Unlike the above efforts, which improve the MMTs' representation space in a task-agnostic fashion, FS-XLT assumes a handful of labeled task-specific examples in the target language(s) (Hedderich et al., 2020; Lauscher et al., 2020; Zhao et al., 2021). Lauscher et al. (2020) propose sequential FS-XLT: fine-tuning on few target-language instances follows the initial fine-tuning on sizable source language data. They show that FS-XLT brings the largest gains exactly where ZS-XLT fails the most: for target languages distant from the source and underrepresented in pretraining. In follow-up work, Zhao et al. (2021) demonstrate that FS-XLT is highly sensitive to the choice of shots. Both studies show the effectiveness of few-shot transfer to be subject to nature of the task: lower-level syntactic and token-level tasks (e.g., POS-tagging, NER) benefit much more from few annotated target language instances than higher-level semantic tasks (e.g., NLI).

The evaluation protocols of both Lauscher et al. (2020) and Zhao et al. (2021), however, do not reflect a true few-shot setup: they assume that substantial validation data in the target language exists and utilize it to guide model selection (hyperparameter optimization and early stopping). As such, these works overestimate the effectiveness of *true* FS-XLT: while focused only on monolingual setups, Perez et al. (2021) demonstrate that model selection criteria based on training data alone yield

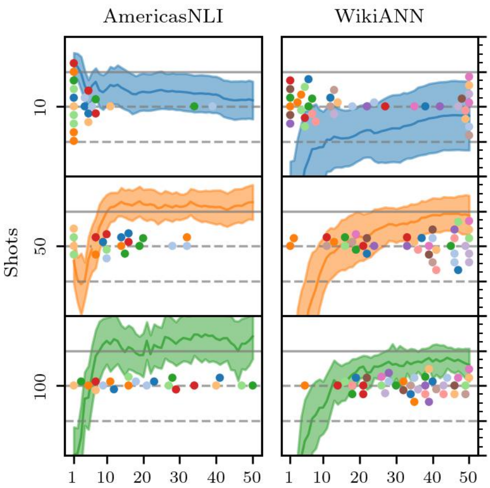

Figure 1: FS-XLT to AmericasNLI and WikiANN with $\{10, 50, 100\}$ shots after training on English data (cf. §4). The line plots the mean (incl. $\pm 1\sigma$) test set spread (in %) of best validation and current checkpoint. Runs across 3 seeds by language are grouped by colored dots that mark epochs scoring best on validation sets.

consistently worse few-shot task performance than model selection based on an extra validation set.

In this work, we rethink FS-XLT and propose novel FS-XLT paradigms that jointly leverage both (sizable) source and (few-shot) target language data in multi-task fashion or via mix-up (Zhang et al., 2018), and demonstrate their effectiveness as well as robustness in realistic (i.e., true) FS-XLT setups.

3 Methodology

Issues with Current FS-XLT Methods. Figure 1 illustrates the main issues of current FS-XLT techniques, adopting the established sequential approach (Lauscher et al., 2020; Zhao et al., 2021; Üstün et al., 2022). In this experiment, we adapt models fine-tuned on sizable English task-specific data with $\{10, 50, 100\}$ target-language shots to AmericasNLI (Ebrahimi et al., 2021) and WikiANN (NER) (Rahimi et al., 2019) (see §4). We execute three FS-XLT runs for each target language with different randomly selected shots and examine the test performance over time, displaying the mean and deviation ($\pm 1\sigma$) across all languages and runs for different training duration (i.e., for $\{1, \ldots, 50\}$ epochs of target language training). The gray horizontal line denotes the optimal performance (average across all languages and runs) in the presence of a target language validation set (i.e., *'not-true'* few-shot learning): for each run, we select the checkpoint that yields the best validation performance. Individual runs are denoted with colored dots, each color indicating one target language. Each dot is vertically aligned with the epoch/checkpoint of the respective run (x-axis) that yields the best validation performance.

The figure reveals the instability of sequential FS-XLT. **1)** The optimal epoch/checkpoint varies across all dimensions of analysis: number of shots, tasks, and languages. Besides the expected result that, on average, with more shots we benefit from longer training,[1] no discernible pattern emerges. **2)** The optimal training duration substantially varies even across different runs of the same language, that is, for different random selections of N shots (and even for larger number of shots, $N = 500$, cf. Figure 2 later in §5.1). These observations render sequential FS-XLT highly unreliable for the *true* FS-XLT setups without target validation data.

New FS-XLT Training Methods. Motivated by these empirical insights, we explore new FS-XLT paradigms, aiming to increase robustness and effectiveness in true FS-XLT setups. Our hypothesis is that combining abundant source-language task examples with scarce target examples *in a joint fashion* will **1)** prevent the models to overfit to source-language features (see Figure 1), **2)** also prevent overfitting to an (extremely) small set of target-language shots (Zhao et al., 2021), and **3)** result in the models that are better *calibrated* for a particular source-target transfer direction. The FS-XLT methods should model the *interaction* between source and target examples, rather than performing source-language fine-tuning which is fully agnostic of the target language (and vice versa).

The first approach, dubbed *'macro-averaging FS-XLT'* (MACRO), conducts bilingual or multilingual fine-tuning in a joint (i.e., multi-task) setup. In particular, we compute the total loss $\mathcal{L} = \delta\mathcal{L}_S + (1-\delta)\mathcal{L}_T$ as a weighted sum of $\mathcal{L}_S$ and $\mathcal{L}_T$, where $\mathcal{L}_S$ and $\mathcal{L}_T$ are monolingual losses associated with the examples from the source language S and the target language T, respectively. δ is a standard interpolation hyper-parameter that adjusts the relative weight between the two losses. The two individual losses operate over the dedicated mini-batches $B_S = \{x_i^s, y_i^s\}_{i=1,\ldots,N}$ and $B_T = \{x_j^t, t_j^t\}_{j=1,\ldots,M}$, which are sampled from the respective source and target language datasets

[1]With as little as 10 shots, longer training, intuitively, leads to overfitting. Figure 1 proves this for AmericasNLI and WikiANN, showing that the first checkpoint yields the best performance for most runs (i.e., the majority of dots are grouped most to the left of the plot).

D_S and D_T. N and M in combination determine the size $|B|$ of the entire mini-batch, as well as the relative share of samples for each language within the mini-batch. The generalization of the bilingual MACRO FS-XLT method (MACRO-BI) to its multilingual variant (MACRO-MULTI) is straightforward: each multilingual batch B would simply comprise examples from more than 2 languages, and the joint loss will span more than 2 language-specific losses.

The second paradigm is based on the standard mix-up technique (Zhang et al., 2018). It has been proven beneficial for improving task performance and robustness in monolingual tasks; here, we extend it to the cross-lingual FS-XLT scenario. This method, termed MIX-UP, linearly interpolates between pairs of annotated instances from the source and the target language as follows:

$$\tilde{x}_{s,t} = \lambda x_i^s * (1-\lambda) x_j^t; \quad \tilde{y}_{s,t} = \lambda y_i^s * (1-\lambda) y_j^t$$

$\lambda \sim Beta(\alpha)$ weighs the contribution between instances (x_i^s, y_i^s) and (x_j^t, y_j^t). Each instance $(x_b, y_b) \in B$ can be paired with any other instance with varying λ. We opt to randomly pair instances in B_S and B_T to be 'mixed', and keep α constant. The fine-tuning loss $\mathcal{L}$ is then computed via soft cross-entropy: $\sum_b^{|B|/2} \tilde{y}_b \log \tilde{y}_b$. Cross-lingual MIX-UP can be interpreted as 'soft' *code switching*, occurring in the latent representation space: it should enhance FS-XLT by further tying, in a task-specific fashion, the representation subspaces of the two languages, as the model is trained for the task on 'mixed' representations, rather than independent language-specific distributions (Cao et al., 2020; Yang et al., 2022).

Overview of FS-XLT Training Methods. Besides introducing novel methods, the main goal of this work is a comprehensive empirical comparative study of different FS-XLT training methods/regimes. For clarity, we provide a quick overview of the wide spectrum of evaluated regimes and configurations. First, models may be trained on target language shots *after* training on the source language data. This approach, termed TARGET, is the standard sequential FS-XLT from prior work (Lauscher et al., 2020; Zhao et al., 2021).[2] The alternative is the regime that combines source-language and target-language data instances, termed SOURCE-TARGET, which comes in two different flavors: our proposed joint MACRO and MIX-UP paradigms. The second axis of difference is the starting point of TARGET or SOURCE-TARGET FS-XLT: we can start fine-tuning from **1)** the original PLM (termed LM henceforth), or **2)** from the final/last checkpoint of source-language task fine-tuning (termed LAST), or **3)** the ORACLE checkpoint. ORACLE violates the true FS-XLT: it refers to the model checkpoint that achieves the best performance on the target language validation set, measured after each epoch of source language training (Keung et al., 2020). We include ORACLE for analysis purposes.

4 Experimental Setup

Tasks and Languages. Following prior studies focused on FS-XLT (Lauscher et al., 2020; Zhao et al., 2021), we evaluate all the methods in a representative set of tasks that require varying degrees of semantic and syntactic understanding for successful cross-lingual transfer.

Natural Language Inference (NLI). NLI experiments are conducted on AmericasNLI (AmNLI) (Ebrahimi et al., 2021): it encompasses indigenous target languages from the Americas, with data carefully translated from the Spanish XNLI dataset (Conneau et al., 2018).[3] Unless stated otherwise, the source is English, and we transfer to the following 7 target languages with sizable NLI data available: Aymara (AYM), Bribri (BZD), Guarani (GN), Quechua (QU), Raramuri (TAR), Shipibo-Konibo (SHP), Wixarika (HCH). For NLI, we jointly embed the hypothesis-premise sentence-pair, obtain the [CLS] token and feed it into the classifier.

Paraphrasing. The paraphrasing task is conducted on the PAWS-X dataset (Yang et al., 2019), spanning parallel evaluation data for 6 high-resource languages: German (DE), Spanish (ES), French (FR), Korean (KO), Japanese (JA), and Chinese (ZH). We train classifiers in the same fashion as classifiers for NLI, now only with paraphrase pairs.

Named Entity Recognition (NER). We use the WikiANN dataset of Pan et al. (2017), and evaluate cross-lingual transfer between English and the following 13 languages: Arabic (AR), Afrikaans (AF), German (DE), Japanese (JA), Quechuan (QU), Russian (RU), Kinyarwanda (RW), Swahili (SW), Tamil (TA), Urdu (UR), Vietnamese (UR), Yoruba

[2] A variant that bypasses source-language fine-tuning and operates only on the few target shots yields massive and consistent drops (Zhao et al., 2021); we thus do not include this variant in our evaluations.

[3] ZS-XLT typically fails in transfer to these languages, as they are unseen during MMT pretraining and are typologically very distant from English.

(YO), Mandarin (ZH). For NER, we feed output representations of each token into the classifier.

Part-Of-Speech Tagging (POS). We use the POS tags of the UD treebanks (Zeman et al., 2020) and transfer from English to the following 12 target languages: Afrikaans (AF), Arabic (AR), Basque (EU), Chinese (ZH), German (DE), Hindi (HI), Hungarian (HU), Indonesian (ID), Japanese (JA), Russian (RU), Tamil (TA), Urdu (UR). The model architecture is identical to NER experiments.

Data Sampling and Shots. For AmNLI and PAWS-X, we subsample training and validation subsets from the provided validation splits.[4] WikiANN and the Universal Dependencies treebank comprise sufficiently large training and validation splits; we subsample shots from the training data. We follow Lauscher et al. (2020) and train models with $k \in \{10, 50, 100, 250, 500\}$ target-language shots, fixed by task and language.[5]

Training Details. The main MMT is the base variant of XLM-R from the `transformers` library (Wolf et al., 2020) with mixed precision. For all tasks, we train models with AdamW (Loshchilov and Hutter, 2019) with the learning rate fixed to $2e^{-5}$ and weight decay of 0.05. All models apply 10% dropout to the output representations prior to the classification layer at training time. The maximal input sequence length is set to 256 subwords for AmNLI and PAWS-X, and 512 for NER and POS.[6] ZS-XLT and SOURCE-TARGET variants are trained for 10 epochs with the linear warm-up rate of 0.1 and linear decay.[7] We fine-tune TARGET regimes for 50 epochs with a constant learning rate. We train in mini-batches of size 32: the SOURCE-TARGET regimes balance instances from source and target languages – for MACRO-BI, we sample 16 instances per language choose the language-balanced loss ($\delta = 0.5$); MIX-UP interpolates between 32 pairs of instances between the languages, resulting with 32 'mixed' bilingual examples. For MIX-UP, we keep α fixed to 0.4.[8] We run all experiments over three (fixed) random seeds. Further details on reproducibility are provided in Appendix A.1.

Evaluation Details. We measure performance with accuracy on AmNLI and PAWS-X. For WikiANN and POS, we report the token-level F_1 score. We report both performance of final/last (L) and oracle (O) checkpoints to provide appropriate bounds on expected and ideal transfer performance.[9]

5 Results and Discussion

The main results are listed in Table 1. Full results per individual target languages in each task are available in the Appendix. First, we corroborate the findings from prior work (Lauscher et al., 2020; Zhao et al., 2021), and report considerable gains with FS-XLT over ZS-XLT across the board and with different FS-XLT methods. We now dissect the results across multiple axes of comparison.

Joint versus Sequential FS-XLT. In general, the joint (i.e., SOURCE-TARGET) FS-XLT variants score on-par or outperform the sequential (i.e., TARGET) variants, and the gains are observed both at Last and Oracle checkpoints. Moreover, we note that the scores taken at the L checkpoint with the joint variants across all setups are typically higher than the scores taken at the O checkpoint. This renders them more suitable for *true* FS-XLT scenarios, and clearly suggests that the proposed joint approaches remedy the issues with overfitting and allow for a more stable fine-tuning. We attribute this finding exactly to bilingual regularization and transfer calibration (see §3).

Joint Methods: MACRO versus MIX-UP. The two joint methods typically yield very similar performance when all other components are kept equal, and fine-tuning starts from the LAST or the ORACLE checkpoint. MIX-UP data augmentation insignificantly affects performance. The effect is most apparent when comparing SOURCE-TARGET setups on the higher-level semantic tasks (AmNLI and PAWS-X), where the model must learn to embed sentence-pair semantics in the [CLS] token. To this end, both tasks require initial source-language fine-tuning as the LM variants lag substantially behind LAST and ORACLE which rely on the initial

[4] This is also why we evaluate AmNLI on the subset of 7 languages which come with enough validation instances.

[5] Unlike Zhao et al. (2021), we operate in a more general unconstrained setup, and do not guarantee an equal number of shots per each class in a task.

[6] As a sanity check, we verified that our ZS-XLT implementation scores comparably to other ZS-XLT work with similar hyperparameters (Wu and Dredze, 2020; Hu et al., 2021).

[7] Note that for SOURCE-TARGET setups the source language datasets dictate training times, as target language shots are continously resampled. SOURCE-TARGET for AmNLI is trained for 5 epochs to reduce computational overhead due to the large size of English MNLI (Williams et al., 2018).

[8] We did not observe significant differences in results with $\alpha \in \{0.1, 0.4, 0.7, 1.0\}$ in preliminary experiments.

[9] Prior work typically reported only the O performance which, depending on the target language and downstream task, can heavily overestimate true FS-XLT performance.

		SOURCE		TARGET				SOURCE-TARGET											
		Zero-Shot		Few-Shot				MACRO						MIX-UP					
	Shots	LM		LAST		ORACLE		LM		LAST		ORACLE		LM		LAST		ORACLE	
		L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
AmNLI	*10*	39.6	40.0	38.3	39.9	38.4	41.2	34.9	36.0	38.0	38.1	37.4	38.6	35.1	35.4	37.9	39.4	37.2	38.3
	50	39.6	40.0	43.8	43.3	44.0	43.6	40.6	42.5	44.4	44.4	44.4	45.0	39.8	40.6	44.0	44.5	44.8	45.0
	100	39.6	40.0	45.8	45.0	46.3	46.2	44.1	44.9	46.8	46.6	47.9	47.5	43.8	44.3	47.4	47.0	47.7	47.7
	250	39.6	40.0	49.7	49.5	49.8	49.4	48.4	49.2	51.0	51.2	51.4	51.0	48.4	49.0	51.5	50.6	51.7	51.3
	500	39.6	40.0	51.7	52.0	52.0	51.2	51.8	52.5	53.3	52.9	53.8	53.4	52.3	51.6	53.2	53.2	53.1	53.1
PAWS-X	*10*	83.8	84.0	81.0	84.2	80.0	84.4	81.1	81.8	84.5	84.5	84.7	84.6	77.3	80.6	84.0	84.1	83.8	84.2
	50	83.8	84.0	83.5	84.2	83.4	84.4	79.9	81.2	84.4	84.3	84.6	84.5	74.4	76.6	84.6	84.4	84.7	84.3
	100	83.8	84.0	84.0	84.3	83.5	84.3	79.9	80.2	84.6	84.5	84.6	84.4	75.2	77.8	84.6	84.4	84.7	84.7
	250	83.8	84.0	83.2	84.9	83.2	84.4	81.2	81.8	84.6	84.6	84.9	84.8	78.4	79.2	84.5	84.5	84.5	84.3
	500	83.8	84.0	83.8	85.3	83.6	85.0	82.8	82.9	85.3	85.0	85.5	85.3	81.9	81.9	85.2	85.0	85.1	85.0
NER	*10*	52.5	60.0	60.7	63.3	61.0	64.3	63.9	65.1	64.9	65.8	64.9	66.2	64.2	65.1	63.9	65.1	64.3	65.6
	50	52.5	60.0	72.0	72.3	72.6	73.1	72.8	73.5	73.1	73.1	73.1	73.6	73.2	73.6	72.9	73.4	73.2	73.5
	100	52.5	60.0	73.6	74.5	74.4	74.7	75.5	75.7	75.4	75.5	75.3	75.2	75.8	75.8	74.9	75.4	75.4	75.5
	250	52.5	60.0	75.6	76.5	76.0	76.7	77.1	77.3	77.0	77.1	76.9	77.1	77.4	77.4	76.9	76.9	77.0	77.1
	500	52.5	60.0	77.4	78.6	77.6	78.7	79.2	79.3	79.0	79.0	79.2	79.2	79.5	79.5	78.9	78.9	79.0	79.0
POS	*10*	62.6	63.8	79.9	79.9	80.2	80.2	80.5	80.6	79.9	80.0	80.1	80.2	80.0	80.2	79.9	80.0	80.1	80.2
	50	62.6	63.8	84.9	84.7	85.1	85.1	85.4	85.4	85.1	85.2	85.3	85.3	85.4	85.3	85.3	85.3	85.5	85.5
	100	62.6	63.8	86.6	86.6	86.7	86.9	87.3	87.3	87.1	87.1	87.2	87.2	87.1	87.2	87.1	87.1	87.2	87.2
	250	62.6	63.8	88.7	88.7	88.8	88.9	89.3	89.2	89.1	89.1	89.2	89.2	89.2	89.2	89.1	89.1	89.2	89.2
	500	62.6	63.8	90.1	90.2	90.2	90.2	90.5	90.4	90.4	90.4	90.5	90.5	90.4	90.4	90.4	90.4	90.5	90.5

Table 1: Benchmarking a spectrum of FS-XLT regimes (see §3). The results are averages over three random seeds, aggregated over all target languages represented in each task (see §4) Training and evaluation data are identical across all regimes in the evaluation. L (O) denote performance measured at last (oracle) checkpoint, see §4.

source fine-tuning. MIXUP-LM is most beneficial for the token-level NER task, but does not yield sizeable gains on average over the arguably conceptually simpler MACRO paradigm.

Starting Point of FS-XLT. Expectedly, starting FS-XLT from the ORACLE checkpoint typically yields better performance than starting from the LAST checkpoint. ORACLE, however, violates the assumption of a true FS-XLT setup: it uses the validation set in the target language to select a better checkpoint for additional FS-XLT fine-tuning, which is organically better-aligned with the target language. We note that the gap in performance between these two initializations slightly decreases in case of joint SOURCE-TARGET FS-XLT variants: this again points to improved robustness compared to sequential FS-XLT.

Performance over Languages and Tasks. Performance benefits with different FS-XLT regimes, naturally, depend on the task and target languages at hand. AmNLI starts profiting from FS-XLT only with $k \geq 50$ shots. The target languages in AmNLI are extremely low-resource and unseen in MMT pretraining: the model thus must see more target-language data points than, e.g., in NLI transfer to higher-resource languages from the XNLI benchmark (Lauscher et al., 2020). Our new SOURCE-TARGET variants again substantially outperform currently established FS-XLT methods, and we observe increasing returns with more shots. In contrast, performance on PAWS-X – which comprises only high-resource languages (see §4) – primarily benefits from the more robust joint FS-XLT regimes rather than from the increased number of shots. For NER and POS, we observe strong performance also with the LM initialization. We speculate that this is because class-conditional token representations align well with the representations from the original MMT pretraining; on the other hand, the models for NLI and paraphrasing must capture higher-level sentence semantics (via source-language fine-tuning) before the FS-XLT step.

5.1 Further Analyses

We base our further analyses and comparisons between sequential and joint approaches on the following two representative variants: TARGET-LAST and MACRO-LAST. They operate in the 'real-life' true FS-XLT scenarios without any validation data to guide few-shot learning (Perez et al., 2021).

Stability of Transfer. Figure 2 compares stability of the two variants for $\{10, 50, 500\}$ shots (cf, Appendix A.2). It demonstrates that joint training substantially reduces instability and variance of FS-XLT fine-tuning across the board: we observe its increased robustness and stability across different tasks, languages, and the numbers of shots. The plots also illustrate that the joint regime in the true FS-XLT setup offers performance which is competitive and comes substantially closer to performance achieved when exploiting target-language validation set: this directly indicates that, with joint bilingual fine-tuning (MACRO) in place, any additional labeled instances in the target language

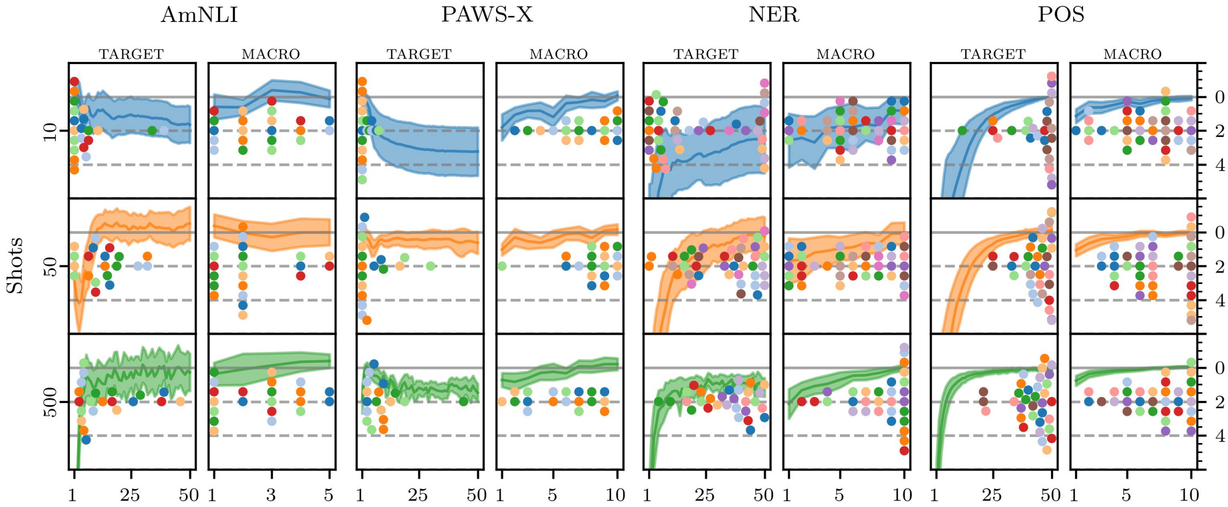

Figure 2: FS-XLT regimes (joint MACRO versus sequential TARGET) starting from the LAST checkpoint of the initial source language fine-tuning step. The colored dots group runs for each seed by language and mark the checkpoints that transfer best to target-language validation data. The line plots the mean (incl. $\pm 1\sigma$) test set spread (in %) of best validation and current checkpoint.

would be better "spent" if used for training than for validation. Relying on the joint MACRO variant, the best-performing checkpoints generally shift closer to the end of the training, which is a desired behavior in the absence of the validation set. In other words, the joint FS-XLT variants not only improve but also consistently make FS-XLT fine-tuning more stable and more predictable, that is, less prone to language- and task-dependent variations.

Notes on Efficiency and Modularity. While the joint FS-XLT regimes improve final transfer performance, they are less modular by design and might incur larger computational costs than the sequential regimes. Namely, they require combining source-language and target-language instances for each individual source-target transfer direction, which is not the case in the sequential regimes. In what follows, we thus delve deeper into studying efficiency- and modularity-related research questions.

Joint Multilingual and Multilingual-Bilingual MACRO. Given N_T target languages, instead of fine-tuning N_T separate bilingual models (MACRO-BI), we can, similar to Xu and Murray (2022), train a single joint multilingual model (MACRO-MULTI, see §3) which serves all N_T at once. Such FS-XLT variant, besides potentially reducing computational and memory costs, might also profit from increased task data provided in multiple languages (Ansell et al., 2021). What is more, we can use the LAST checkpoint of the MACRO-MULTI as the starting point of the additional subsequent bilingual FS-XLT specialization (i.e., MACRO-BI). We denote this novel modular variant, where both steps are based on the joint FS-XLT paradigm, as MULTI→BI.

Furthermore, we conduct another experiment, again focused on efficiency of joint FS-XLT fine-tuning, which includes all the different MACRO variants: (i) the original MACRO-BI, (ii) MACRO-MULTI, and (iii) MACRO-MULTI→BI. The goal is to investigate how the different joint paradigms perform under different computational budget constraints. To this end, we train those MACRO variants with $\{1, 2, 5, 10\}\times$ the number of steps of the sequential TARGET variant.

For the multilingual step, training is always conducted by including 8 instances for each language in a mini-batch: this is done to provide sufficient language-specific examples per mini-batch without dramatically increasing the mini-batch size. For AmNLI and PAWS-X, we include all available languages in training. For NER, we train on {DE, EN, SW, TA,VI, ZH}, and for POS on {AR, EN, EU, HU, ID, JA, UR}. We now evaluate all the MACRO and TARGET variants on the following languages: for AmNLI, AYM, QUY, and TAR; for PAWS-X, DE, KO, JA; for NER, SW, VI, ZH; for POS, EU, UR, JA.

Table 2 presents the complete results of this set of experiments, averaged over the three target languages of each task. First, MACRO-MULTI is on-par or better than TARGET throughout almost all setups, but, with the exception of token-level tasks, does not consistently match the performance of MACRO-BI, which fine-tunes for a particular source-target direction. The highest overall performance is ob-

		TARGET				SOURCE-TARGET (MACRO)																					
						TARGET BUDGET																					
						1×				2×				5×				10×				FT					
				MULTI→BI		BI		MULTI→BI		BI		MULTI→BI		BI		MULTI→BI		BI		MULTI→BI		BI		MULTI		MULTI→BI	
	Shots	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
AmNLI	*10*	36.8	38.8	36.3	36.9	37.6	38.5	36.7	37.7	38.1	38.3	36.2	36.7	37.1	37.9	36.3	36.8	36.5	37.6	36.3	37.7	36.6	36.5	37.1	37.4	35.9	36.7
	50	43.4	42.6	45.6	46.2	42.6	41.9	44.4	44.9	42.4	42.1	44.8	44.7	42.7	43.1	45.4	45.7	42.6	43.4	45.2	45.2	44.4	45.1	43.7	44.7	45.8	45.5
	100	45.7	45.8	48.2	48.4	45.7	45.3	48.7	48.0	46.0	45.6	48.5	48.8	46.3	45.7	48.7	49.2	45.9	46.2	49.2	49.1	46.7	47.2	47.2	47.3	49.0	48.4
	250	50.4	50.2	52.3	52.3	48.4	47.6	52.2	52.4	49.4	49.3	52.7	53.0	49.7	48.9	52.9	52.6	50.5	50.4	53.0	52.6	52.0	51.9	51.0	50.7	52.8	52.6
	500	51.7	52.5	52.3	52.7	52.2	51.1	53.5	53.5	52.8	51.5	53.2	53.0	53.4	52.5	53.3	53.3	53.3	53.1	52.7	53.8	54.0	53.7	54.0	53.7	53.7	53.5
PAWS-X	*10*	77.5	81.5	80.6	80.8	80.8	81.2	81.0	81.3	80.6	81.3	80.5	81.1	80.7	81.3	80.2	81.6	80.6	81.5	80.2	81.7	81.7	81.5	81.1	81.0	80.7	81.3
	50	81.1	81.2	80.9	81.4	81.6	81.3	80.8	81.0	82.0	82.1	80.6	80.8	81.6	81.7	80.8	81.1	81.8	82.0	80.9	81.0	81.7	81.6	81.6	81.5	81.6	81.8
	100	81.6	81.6	82.2	82.4	81.8	82.1	82.5	82.8	81.5	81.9	82.7	82.9	81.6	81.9	82.7	82.8	81.7	81.7	82.8	83.0	81.8	82.1	81.7	81.6	82.0	82.3
	250	80.4	82.4	82.8	83.1	82.2	82.2	82.4	82.8	82.3	82.4	82.5	82.7	82.1	82.0	82.5	82.6	82.0	82.0	82.4	82.8	82.0	81.8	81.9	81.8	82.7	82.7
	500	81.3	82.7	82.9	83.5	83.0	82.9	83.1	83.5	83.0	82.5	83.3	83.4	82.5	82.7	83.6	83.6	83.1	83.0	83.9	83.7	82.7	82.5	83.1	83.0	83.6	83.1
NER	*10*	56.0	58.4	62.2	66.3	61.4	61.8	65.2	66.2	61.8	62.6	65.9	65.9	62.2	62.9	66.5	67.1	62.1	63.6	67.0	67.8	62.6	63.6	62.1	63.9	67.7	68.4
	50	71.4	71.8	73.0	73.8	69.9	70.0	73.5	73.8	71.0	71.3	73.5	73.9	71.1	71.7	73.5	74.2	71.6	72.3	73.3	74.0	71.8	72.2	72.4	72.8	74.0	74.8
	100	72.7	73.8	75.2	75.7	73.2	73.2	75.5	76.0	73.5	73.9	75.7	75.8	73.4	74.0	75.8	76.2	74.3	74.7	76.1	76.4	74.7	74.8	74.7	74.9	76.1	76.5
	250	77.4	78.4	78.7	79.6	76.7	77.0	78.7	79.0	77.1	77.4	78.7	79.0	77.6	78.0	78.9	78.9	78.2	78.4	79.3	79.4	78.2	78.2	78.3	78.3	79.4	79.7
	500	79.3	80.0	80.9	81.4	79.1	79.1	80.7	80.8	79.3	79.5	81.1	81.1	79.8	80.2	81.4	81.3	80.2	80.4	81.1	81.5	80.2	80.3	80.4	80.4	81.4	81.4
POS	*10*	77.5	77.5	80.6	80.7	76.4	76.4	80.7	80.6	77.6	77.7	80.9	80.9	77.9	78.2	81.0	81.0	78.0	78.2	81.0	81.1	78.4	78.3	79.2	79.3	81.2	81.4
	50	83.4	83.3	85.6	85.8	81.2	81.2	85.6	85.5	82.4	82.4	85.7	85.8	83.3	83.3	85.7	85.8	83.5	83.6	85.8	85.8	84.4	84.4	84.8	84.8	86.0	86.0
	100	85.6	85.6	87.5	87.8	84.5	84.4	87.6	87.6	85.3	85.2	87.6	87.7	85.7	85.7	87.8	87.8	86.0	86.0	87.7	87.8	86.6	86.5	87.0	86.9	87.9	88.0
	250	88.0	88.2	89.1	89.4	87.3	87.3	89.4	89.5	87.9	87.8	89.6	89.6	88.4	88.4	89.7	89.6	88.6	88.6	89.6	89.6	88.9	88.8	89.1	89.1	89.7	89.7
	500	89.6	89.8	90.2	90.4	89.1	89.1	90.3	90.4	89.5	89.5	90.5	90.5	90.0	89.9	90.5	90.6	90.1	90.0	90.7	90.6	90.1	90.0	90.2	90.1	90.5	90.5

Table 2: FS-XLT results where each fine-tuning regimes commences from the final checkpoint of English fine-tuning. All tasks comprise three target languages, and the scores are averaged over three fixed random seeds, with training and validation subsets being the same for each seed.

	AmNLI				PAWS-X				NER				POS			
	T		S-T		T		S-T		T		S-T		T		S-T	
Shots	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
10	36.5	36.5	35.8	36.3	78.0	83.6	83.4	83.1	50.8	53.8	54.0	55.3	80.8	80.8	78.5	78.8
100	43.8	44.1	46.9	46.3	81.3	83.0	83.5	82.8	66.9	68.2	69.7	70.2	88.1	88.1	88.6	88.6
500	50.1	49.7	52.9	52.4	82.2	83.0	84.5	84.3	74.2	75.0	76.7	76.6	91.0	91.1	91.3	91.3

Table 3: FS-XLT with Chinese as the source language. S=SOURCE, S-T=SOURCE-TARGET (MACRO is used).

tained with the hybrid MACRO-MULTI→BI, which reaps the best of both worlds: **1)** multilingual fine-tuning prevents overfitting to a single source language and provides a better initialization point for **2)** the more specialized bilingual fine-tuning for a particular source-target direction. Note that the two-stage MULTI→BI fine-tuning also improves the TARGET variant quite consistently. We report increase in performance both for L and O checkpoints. Nevertheless, MACRO still outperforms TARGET.

The results over different computational budgets reveal that longer training is beneficial for the MACRO variants. As expected, the setups with more shots typically require fewer steps to converge. A general finding is that 1) the bilingual SOURCE-TARGET variants do trade off some of the computational efficiency for enhanced performance, but 2) bilingual fine-tuning times can be decreased by starting from a better (i.e., multilingual) initialization: cf., the MULTI→BI columns.

Another Source Language. Cross-lingual transfer predominantly focuses on English as the source language (Hu et al., 2020; Lauscher et al., 2020), mostly because of the wide availability and abundance of annotated task data in English. In order to verify that our main findings generalise and reach beyond English as the source language, we conduct another set of experiments relying on Chinese as the source language.[10] The results for the TARGET-LAST and MACRO-LAST variants are presented in Table 3. The observed patterns largely follow the general trends we reported with English as the source language; what is more, the gains of SOURCE-TARGET over TARGET even widen for AmNLI and PAWS-X. We speculate that this might be due to a lower quality of the source Chinese instances. Namely, except for POS, the task annotations for Chinese were either automatically translated (AmNLI, PAWS-X) or induced via some heuristics (WikiANN). Joint bilingual fine-tuning then provides increased robustness against such noisy source annotations.

6 Conclusion

Recent work demonstrated large benefits of few-shot cross-lingual transfer (FS-XLT) with multilingual language models, where a handful of annotated examples in the target language exist, over its zero-shot counterpart (ZS-XLT). However, as we have proven in this paper, prior work overestimated

[10] For AmLI and PAWS-X, we experiment with the same three languages as in joint multilingual experiments. For NER, we transfer to AR, UR, and JA, and to AR, DE, and UR for POS.

FS-XLT performance, relying on an unrealistic assumption of having a dedicated validation set in the target language to guide model selection. In this work, we have performed an extensive comparative study of a wide variety of FS-XLT approaches, challenging the status quo in FS-XLT. Our detailed analyses have rendered established FS-XLT largely unstable and performing sub-par in true FS-XLT setups without the target validation data. We have thus proposed novel FS-XLT fine-tuning regimes that take into account interaction between source-language and target-language data instances, yielding improved, more stable, and more predictable FS-XLT performance across different tasks, languages, and numbers of target-language shots. We hope that our study will inspire better FS-XLT training and evaluation practices in future work, and guide new developments for true FS-XLT setups.

7 Limitations

While we have striven to present a comprehensive and wide study of a large spectrum of FS-XLT fine-tuning regimes, several additional factors must be taken into consideration. First, few-shot learning naturally comes with high variance, as demonstrated by our work (where we set out to decrease the variance) and a body of prior research in monolingual and cross-lingual transfer contexts. This study demanded an extremely large computational budget (see Appendix A.1), so we constrained experiments to independent runs with three seeds. Ideally, more independent runs (5-10) might yield even more consistent estimates.

Furthermore, due to computational constraints, our work largely focuses on cross-lingual natural language understanding (NLU) and sequence-labeling tasks. In addition, the community might find a similar set of experiments insightful for cross-lingual transfer in other areas such as (i) task-oriented dialogue systems, or (ii) long-range tasks like document classification. Moreover, while we keep hyper-parameters constant throughout different regimes, it is highly likely that they can be further adapted and fine-tuned for a particular task, language, and selection of shots. However, our core findings demonstrate that the novel joint FS-XLT fine-tuning regimes consistently match or exceed oracle performance while requiring no substantial hyper-parameter tuning or checkpoint selection.

Acknowledgements

We thank the state of Baden-Württemberg for its support through access to the bwHPC. Fabian David Schmidt and Goran Glavaš were supported by the EUINACTION grant from NORFACE Governance (462-19-010, GL950/2-1). Ivan Vulić is supported by a personal Royal Society University Research Fellowship (no 221137; 2022-2027) as well as a Huawei research donation to the Language Technology Lab.

References

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*, 9:1116–1131.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 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, Gustavo A. Gim énez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *CoRR*, abs/2104.08726.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. *CoRR*, abs/2205.06266.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-x: A unified hypernetwork for multi-task multilingual transfer. *CoRR*, abs/2205.12148.

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, New Orleans, Louisiana. Association for Computational Linguistics.

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

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Haoran Xu and Kenton Murray. 2022. Por qué não utiliser alla språk? mixed training with gradient optimization in few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2043–2059, Seattle, United States. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.

Daniel Zeman, Joakim Nivre, et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer:

The choice of shots matters. 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 5751–5767, Online. Association for Computational Linguistics.

A Appendix

A.1 Reproducibility

Infrastructure and Compute. We train our models on a cluster that provides virtual machines on which each model was trained on a single NVIDIA Tesla V100 32GB GPU. We evaluate 7 setups with three seeds for $k \in \{10, 50, 100, 250, 500\}$ shots across 4 tasks in our base experiments, amounting to 5,145 models trained for 3,756 GPU hours for our main results. Therein, AmNLI alone takes up 2,170 hours (57.8%).

Datasets. We access all datasets via the Huggingface `datasets` library (Lhoest et al., 2021). Whenever we subsample data, we initially shuffle the dataset with one of seed $s \in \{42, 43, 44\}$ built-in `datasets` method and subsequently extract the first k required instances for our experiments. In case we require a validation subset from the same dataset, we extract the $|N_D| - 500$ last available observations after shuffling to evaluate our models during training (i.e., to measure ORACLE performance). We manually verified that our approach yields consistent subsamples by seed.

Code. Our code is available at: `https://github.com/fdschmidt93/fsxlt`

A.2 Stability Of Few-Shot Cross-Lingual Transfer

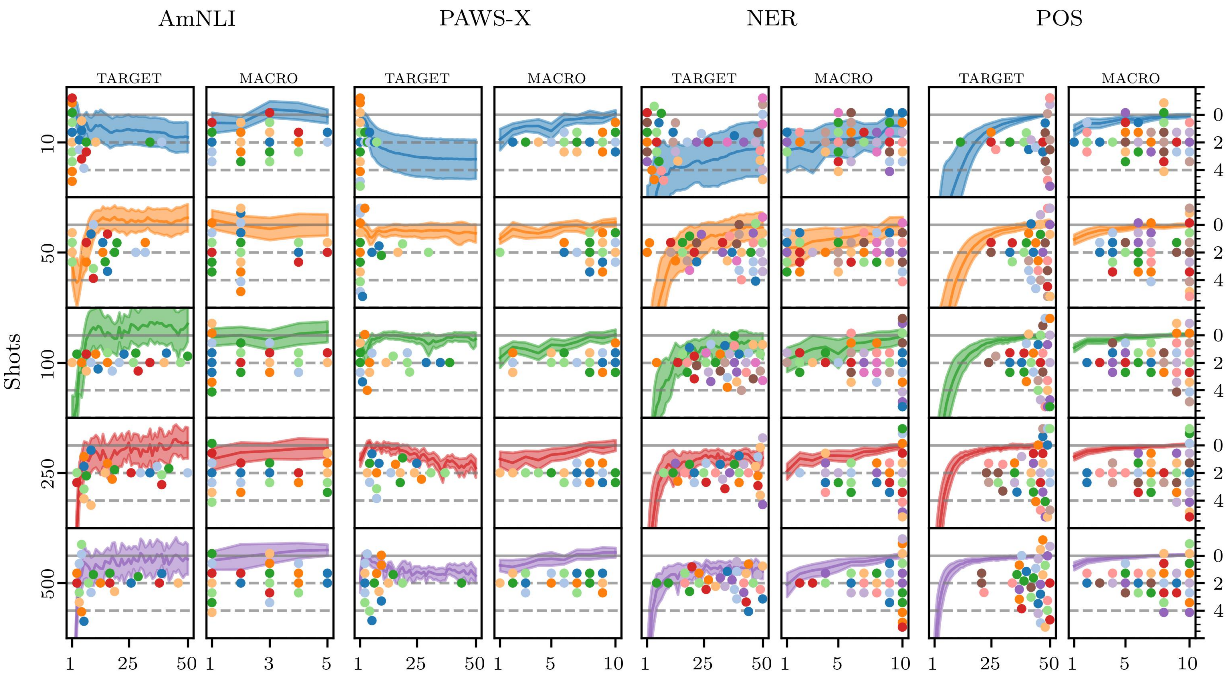

Figure 3: FS-XLT regimes (joint MACRO versus sequential TARGET) starting from the LAST checkpoint of the initial source language fine-tuning step. The colored dots group runs for each seed by language and mark the checkpoints that transfer best to target-language validation data. The line plots the mean (incl. $\pm 1\sigma$) test set spread (in %) of best validation and current checkpoint.

A.3 Full Results over Individual Target Languages

A.4 AmericasNLI

		SOURCE		TARGET				SOURCE-TARGET											
		Zero-Shot		Few-Shot				MACRO						MIX-UP					
Weights	k Shots	LM		LAST		ORACLE		LM		LAST		ORACLE		LM		LAST		ORACLE	
Metric		L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
Aymara	10	39.7	39.5	38.0	39.0	39.3	41.8	34.5	37.0	36.9	37.1	34.8	38.0	34.5	35.9	36.8	38.1	36.7	38.3
AYM	50	39.7	39.5	44.9	45.0	44.7	46.1	40.9	42.1	45.4	45.6	42.2	42.8	37.1	40.0	44.5	44.8	45.6	45.5
	100	39.7	39.5	45.7	45.5	46.9	48.7	45.9	46.4	46.0	46.7	47.7	46.9	43.0	43.6	49.0	47.9	47.3	47.9
	250	39.7	39.5	49.3	50.0	50.1	50.9	50.0	51.4	51.6	51.4	53.3	51.5	50.5	51.6	52.0	51.1	52.0	52.2
	500	39.7	39.5	52.0	52.5	51.6	50.8	51.6	51.6	54.6	54.8	55.2	56.1	52.5	50.9	53.7	54.3	53.5	54.4
Bribri	10	40.8	40.4	39.1	40.0	39.0	40.7	36.4	36.3	39.9	39.7	41.0	41.6	36.1	35.9	40.3	40.2	39.1	39.8
BZD	50	40.8	40.4	44.7	45.2	44.5	44.2	44.1	45.7	47.6	45.8	49.0	49.4	42.5	43.0	47.0	48.6	47.5	48.7
	100	40.8	40.4	48.6	46.8	49.2	48.2	49.2	49.6	51.5	50.5	52.5	51.2	48.2	48.1	51.7	51.2	52.5	50.7
	250	40.8	40.4	52.6	52.0	54.9	55.1	52.0	51.4	54.2	54.6	53.4	53.7	51.3	52.3	55.2	54.8	55.3	54.6
	500	40.8	40.4	56.6	56.4	56.8	56.7	54.6	56.1	57.9	57.3	57.7	58.2	56.4	55.0	56.5	56.7	57.2	56.8
Guarani	10	41.1	42.1	40.3	41.7	39.3	44.0	35.7	35.7	38.6	39.4	38.3	40.0	34.4	34.5	40.6	42.5	39.4	39.1
GN	50	41.1	42.1	46.8	47.1	45.2	44.9	40.8	42.5	45.3	45.4	45.5	46.1	40.7	40.2	47.1	46.7	45.6	46.7
	100	41.1	42.1	47.6	46.6	48.8	47.9	45.0	45.3	49.2	47.6	49.3	49.7	44.8	46.3	49.6	48.8	49.6	50.4
	250	41.1	42.1	52.1	51.4	49.7	49.7	49.8	51.7	51.6	52.2	51.6	51.8	48.3	50.2	51.4	50.8	51.7	50.5
	500	41.1	42.1	54.2	52.8	53.4	51.3	51.3	52.4	53.3	52.7	54.0	52.9	53.5	52.8	52.6	52.5	53.5	52.6
Wixarika	10	38.4	37.5	36.9	38.4	37.5	39.3	33.8	35.5	37.6	38.4	36.4	37.3	34.4	34.3	36.2	38.0	35.5	36.6
HCH	50	38.4	37.5	40.3	39.9	40.6	39.8	35.9	38.1	40.3	40.3	40.1	40.2	36.4	35.8	39.9	39.5	39.8	40.8
	100	38.4	37.5	41.8	39.9	41.0	40.8	37.0	38.3	40.8	40.5	41.6	42.0	37.9	37.2	41.8	42.6	42.3	41.7
	250	38.4	37.5	44.2	44.5	43.4	41.5	40.3	40.3	45.2	45.9	44.8	43.1	39.5	39.3	45.6	45.5	44.0	44.7
	500	38.4	37.5	44.8	46.1	44.5	43.8	45.7	46.8	46.4	45.7	45.5	45.2	44.7	44.0	46.9	46.5	45.7	46.0
Quechua	10	37.3	38.3	37.2	40.1	38.6	42.6	33.6	33.6	37.2	36.0	36.9	37.6	34.4	34.5	37.1	39.1	37.4	38.8
QUY	50	37.3	38.3	43.9	43.1	46.1	46.0	42.3	45.1	44.4	46.0	46.7	48.2	42.5	42.4	43.2	44.6	46.6	46.0
	100	37.3	38.3	47.8	46.2	48.2	48.1	41.3	41.5	47.5	48.1	50.4	49.0	45.3	45.6	47.7	46.1	49.8	49.4
	250	37.3	38.3	52.2	51.7	51.5	52.1	50.3	49.7	52.8	52.1	54.5	54.5	50.1	51.2	52.8	52.8	54.8	54.0
	500	37.3	38.3	52.7	53.3	54.0	53.6	52.1	52.7	55.6	55.1	55.3	54.9	52.7	53.1	54.8	55.4	54.7	55.0
Shipibo	10	41.0	42.9	41.0	42.8	40.3	42.0	36.4	38.7	40.0	39.8	39.2	40.0	36.7	37.5	40.2	42.4	37.9	40.9
SHP	50	41.0	42.9	44.4	43.0	44.4	42.8	39.6	41.9	44.5	44.2	42.9	43.0	40.1	42.4	44.4	44.4	44.3	44.2
	100	41.0	42.9	45.9	44.4	44.4	43.9	45.2	47.4	45.9	45.6	44.6	45.9	43.8	44.2	46.2	46.7	45.2	45.3
	250	41.0	42.9	47.7	48.0	48.2	47.8	48.7	50.8	50.0	50.2	50.5	50.4	50.5	49.6	51.5	50.1	50.3	50.0
	500	41.0	42.9	51.3	51.2	51.5	50.6	55.0	55.2	53.8	53.7	54.3	52.8	54.1	53.4	54.5	54.3	54.4	52.6
Raramuri	10	39.1	39.2	35.3	37.3	34.9	37.6	33.9	35.4	35.7	36.4	35.1	35.9	34.9	34.9	34.2	35.4	34.8	35.0
TAR	50	39.1	39.2	41.5	39.7	42.5	41.1	40.4	42.1	43.4	43.6	44.4	45.1	39.0	40.4	42.0	42.9	44.1	42.8
	100	39.1	39.2	43.5	45.8	45.7	46.1	45.0	45.5	46.7	46.9	49.0	47.8	43.3	45.2	45.7	45.6	46.8	48.4
	250	39.1	39.2	49.5	48.8	50.8	48.8	48.0	48.8	51.7	52.2	51.8	52.4	48.6	49.1	51.9	49.3	53.9	53.2
	500	39.1	39.2	50.4	51.7	52.4	51.4	52.6	52.6	51.6	51.2	54.3	53.9	52.1	51.9	53.3	52.9	52.9	54.3

A.5 PAWS-X

		SOURCE		TARGET				SOURCE-TARGET											
		Zero-Shot		Few-Shot				MACRO						MIX-UP					
Weights	k Shots	LM		LAST		ORACLE		LM		LAST		ORACLE		LM		LAST		ORACLE	
Metric		L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
German	10	88.7	88.7	84.4	88.7	82.9	88.9	87.4	87.2	89.5	89.5	89.2	89.1	84.6	86.0	88.8	88.7	89.1	89.2
DE	50	88.7	88.7	88.5	88.8	88.5	88.8	86.2	87.1	89.0	89.1	89.6	89.3	80.4	83.7	89.0	89.0	89.0	88.5
	100	88.7	88.7	88.9	88.9	88.9	89.1	86.8	86.6	89.4	89.4	89.2	88.8	82.2	83.4	89.1	89.0	89.3	88.9
	250	88.7	88.7	87.2	88.8	87.4	88.7	87.4	86.6	89.2	88.9	89.4	89.4	84.6	84.7	89.0	89.2	89.1	89.1
	500	88.7	88.7	87.6	89.0	86.5	88.9	87.2	87.0	89.4	89.5	89.6	89.4	86.4	87.2	89.3	89.2	89.5	89.2
Spanish	10	89.5	89.7	85.9	89.7	85.9	89.7	88.3	88.4	89.8	89.4	89.8	89.4	85.9	88.2	89.8	89.8	89.8	89.8
ES	50	89.5	89.7	88.9	89.7	88.9	89.7	88.5	88.0	90.0	89.8	90.0	89.8	84.7	85.5	89.7	89.2	89.7	89.2
	100	89.5	89.7	89.0	89.3	89.0	89.3	87.2	86.8	89.6	89.7	89.6	89.7	82.5	85.2	90.0	90.1	90.0	90.1
	250	89.5	89.7	88.7	89.7	88.7	89.7	88.1	87.6	89.9	89.7	89.9	89.7	85.0	85.6	89.5	89.0	89.5	89.0
	500	89.5	89.7	88.2	90.3	88.2	90.3	88.9	88.8	90.3	89.2	90.3	89.2	87.9	87.8	89.8	89.5	89.8	89.5
French	10	89.6	90.2	87.7	89.8	86.5	90.4	89.1	89.0	90.0	90.4	90.5	90.3	87.3	88.2	90.0	89.4	89.8	89.6
FR	50	89.6	90.2	88.7	90.0	89.2	90.4	84.1	87.8	90.0	89.7	90.1	90.1	80.1	84.4	90.3	90.2	90.2	90.0
	100	89.6	90.2	89.4	89.6	89.4	90.2	87.2	87.6	90.2	89.9	90.6	90.5	84.5	86.4	90.2	90.2	90.4	90.1
	250	89.6	90.2	88.9	90.1	89.0	90.1	87.3	88.0	90.0	90.4	90.5	90.7	86.0	86.4	90.5	90.4	90.2	89.9
	500	89.6	90.2	89.3	90.2	88.7	90.1	89.2	89.1	91.0	91.0	91.0	91.0	88.6	88.5	90.6	90.3	90.4	90.7
Japanese	10	77.1	77.1	75.7	77.2	74.5	77.1	72.7	73.5	77.3	77.1	77.2	77.5	68.3	71.6	76.7	77.5	76.1	77.0
JA	50	77.1	77.1	76.6	77.1	74.7	76.8	71.3	73.9	77.6	77.5	76.7	76.9	62.8	64.5	78.0	77.8	78.0	77.3
	100	77.1	77.1	77.2	77.5	74.0	76.3	72.3	73.0	77.2	78.0	77.2	77.2	67.2	70.9	77.8	77.5	77.1	77.4
	250	77.1	77.1	76.9	78.8	76.4	77.3	74.4	75.4	78.4	78.2	78.2	78.0	69.9	71.2	77.5	77.8	78.0	77.6
	500	77.1	77.1	77.7	79.6	77.4	78.9	76.4	76.9	79.2	79.0	79.4	79.3	75.8	75.2	79.4	79.4	78.8	79.1
Korean	10	76.7	77.2	72.2	78.6	72.2	78.2	71.1	73.4	78.3	78.0	78.7	78.4	62.9	69.9	77.3	77.6	77.2	77.6
KO	50	76.7	77.2	78.2	77.5	77.6	78.6	71.4	72.6	78.5	78.2	78.9	79.5	65.0	67.0	78.9	78.3	78.9	78.6
	100	76.7	77.2	78.6	78.6	78.6	78.5	70.6	71.0	78.7	78.8	78.6	78.7	65.8	67.9	78.6	78.1	79.2	79.3
	250	76.7	77.2	77.2	79.6	77.1	78.8	72.9	74.1	78.4	78.3	78.6	78.4	68.8	70.6	78.6	78.3	78.8	78.2
	500	76.7	77.2	78.8	79.7	79.6	79.7	75.4	75.5	79.5	79.1	79.4	79.8	73.8	74.2	80.1	79.6	79.6	79.3
Chinese	10	81.1	81.3	80.0	81.6	78.2	82.2	78.0	79.4	82.3	82.5	82.8	82.8	74.6	79.6	81.3	81.5	81.0	81.9
ZH	50	81.1	81.3	80.3	81.7	81.3	82.2	77.8	78.0	81.4	81.4	82.3	81.5	73.2	74.4	81.7	82.0	82.4	82.4
	100	81.1	81.3	81.2	82.2	81.2	82.3	75.4	76.0	82.2	81.4	82.4	81.4	68.8	73.2	81.7	81.6	82.1	82.7
	250	81.1	81.3	80.2	82.1	80.4	82.1	77.4	78.8	81.6	82.2	82.8	82.8	75.7	76.8	82.0	82.0	81.6	81.9
	500	81.1	81.3	81.6	83.2	81.1	82.4	79.7	80.2	82.1	82.5	83.1	83.0	78.7	78.7	82.3	82.2	82.2	82.1

A.6 WikiANN

		SOURCE		TARGET				SOURCE-TARGET											
		Zero-Shot		Few-Shot				MACRO						MIX-UP					
Weights	k Shots	LM		LAST		ORACLE		LM		LAST		ORACLE		LM		LAST		ORACLE	
Metric		L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
Afrikaans	10	72.9	73.3	72.9	75.6	72.6	75.5	75.0	75.2	75.5	76.3	76.0	75.8	75.4	75.9	75.6	74.5	75.4	75.3
AF	50	72.9	73.3	76.9	76.9	77.8	77.9	78.7	78.6	78.8	77.4	78.2	78.2	77.7	77.8	78.2	78.2	78.8	77.9
	100	72.9	73.3	78.6	79.8	79.3	79.2	80.1	80.3	79.6	79.6	80.0	79.3	79.7	79.5	80.1	80.3	80.5	80.2
	250	72.9	73.3	81.2	81.2	81.0	81.9	81.6	82.0	82.2	82.1	82.0	81.9	82.0	81.8	82.3	82.0	82.2	81.9
	500	72.9	73.3	82.8	83.7	83.3	83.7	83.7	83.5	83.6	83.5	83.5	83.4	84.3	84.4	83.7	83.7	83.7	83.7
Arabic	10	43.2	49.0	66.6	66.7	64.2	66.7	69.1	69.6	69.1	69.9	68.8	70.6	69.0	70.6	68.0	69.8	70.0	71.6
AR	50	43.2	49.0	72.5	73.0	72.1	72.5	73.2	73.8	73.9	74.4	74.3	74.6	73.8	73.9	74.0	74.6	74.3	74.8
	100	43.2	49.0	73.4	73.8	73.2	73.9	74.5	74.9	75.4	75.9	75.5	76.0	75.5	76.0	75.4	76.0	75.7	76.2
	250	43.2	49.0	74.8	76.7	75.5	76.7	77.1	77.3	77.4	77.6	77.8	77.8	77.7	77.8	78.0	78.0	78.1	78.1
	500	43.2	49.0	76.6	79.1	76.8	78.4	79.7	80.0	79.7	79.7	79.8	79.8	80.0	79.9	79.8	79.9	79.8	79.7
German	10	70.6	71.6	68.3	72.3	68.2	73.5	72.6	73.9	71.7	72.5	71.8	72.5	73.5	73.6	71.6	72.3	72.2	73.1
DE	50	70.6	71.6	72.1	73.1	72.6	73.8	74.8	75.0	73.9	74.5	74.3	74.8	75.2	75.4	74.1	74.4	74.5	74.6
	100	70.6	71.6	73.1	73.6	73.2	73.7	75.6	76.2	75.0	75.1	75.1	75.7	75.5	75.6	74.9	75.6	75.9	76.0
	250	70.6	71.6	75.5	76.6	75.7	76.7	77.1	77.5	76.9	76.9	76.7	76.9	77.3	77.4	76.7	76.7	77.1	77.2
	500	70.6	71.6	76.4	77.9	77.1	78.4	78.7	78.7	78.3	78.3	78.5	78.5	78.7	78.8	78.1	78.3	78.4	78.5
Japanese	10	17.1	17.9	32.0	32.8	32.6	33.1	31.5	32.8	34.9	36.2	35.3	36.8	31.5	33.0	32.2	34.7	33.5	34.4
JA	50	17.1	17.9	43.5	44.0	44.7	45.3	46.5	47.4	46.9	47.1	47.4	47.2	47.2	47.5	44.9	46.2	46.2	47.8
	100	17.1	17.9	47.8	48.1	49.2	49.7	51.4	52.2	50.2	50.9	50.8	51.3	51.1	51.5	50.3	50.6	49.6	50.5
	250	17.1	17.9	52.7	53.4	54.2	54.4	56.3	56.4	55.7	55.6	55.0	55.3	56.2	56.3	55.0	55.0	55.2	55.4
	500	17.1	17.9	55.8	57.6	58.0	58.0	59.7	59.7	59.1	59.1	59.5	59.5	59.7	59.9	58.8	58.8	58.8	59.2
Quechuan	10	54.8	55.3	61.1	61.5	59.8	63.9	58.8	62.9	62.2	60.1	62.2	63.2	63.7	64.0	63.2	64.7	63.8	62.9
QU	50	54.8	55.3	70.5	69.2	74.6	73.1	69.6	71.9	68.9	68.3	69.3	70.2	71.2	72.4	69.9	71.7	70.0	69.5
	100	54.8	55.3	71.4	72.9	74.9	73.2	76.3	76.2	74.4	73.3	75.2	73.2	78.0	76.4	70.9	72.0	74.6	74.2
Russian	10	65.7	66.5	64.7	72.0	64.7	72.0	71.6	73.3	73.0	73.8	73.0	73.8	73.1	73.4	72.1	73.8	72.1	73.8
RU	50	65.7	66.5	78.1	78.4	78.1	78.4	78.1	78.4	78.0	78.4	78.0	78.4	78.8	78.8	77.9	78.2	77.9	78.2
	100	65.7	66.5	80.3	80.2	80.3	80.2	78.7	78.5	79.2	79.5	79.2	79.5	79.1	79.2	79.4	79.6	79.4	79.6
	250	65.7	66.5	80.5	82.0	80.5	82.0	81.3	81.4	81.6	81.6	81.6	81.6	81.4	81.5	81.4	81.5	81.4	81.5
	500	65.7	66.5	82.3	83.3	82.3	83.3	83.2	83.1	83.3	83.3	83.3	83.3	83.1	83.1	83.1	83.2	83.1	83.2
Rwanda	10	57.6	57.3	57.6	62.8	59.0	63.5	64.0	61.1	62.2	64.0	60.4	62.9	60.6	59.0	63.3	63.0	59.6	60.4
RW	50	57.6	57.3	75.9	73.2	74.5	76.5	76.6	75.8	73.3	74.0	75.0	75.4	75.1	73.9	73.2	73.2	75.5	74.4
	100	57.6	57.3	75.5	77.6	75.2	76.8	78.3	76.6	76.3	78.2	75.8	75.3	78.4	77.0	76.4	77.4	77.8	76.9
Swahili	10	61.1	63.8	70.6	70.5	70.8	72.6	73.8	74.1	74.7	74.8	73.9	74.9	74.7	74.8	73.2	74.6	74.4	74.6
SW	50	61.1	63.8	84.3	84.2	84.4	84.5	84.3	84.3	83.8	84.8	84.2	83.9	84.2	84.1	84.1	84.3	84.2	83.6
	100	61.1	63.8	84.6	85.0	85.5	85.3	85.4	86.0	86.5	86.3	85.3	85.8	85.9	85.4	85.8	86.5	85.1	85.1
	250	61.1	63.8	87.3	88.2	87.8	87.5	87.7	88.1	88.1	87.9	87.8	87.6	88.4	88.8	88.1	88.3	87.9	87.9
	500	61.1	63.8	89.0	89.8	88.7	89.6	89.5	89.2	89.5	89.6	89.4	89.9	89.6	89.2	89.6	89.5	89.6	89.6
Tamil	10	58.6	61.4	62.8	62.9	63.2	64.8	63.0	63.7	66.4	66.9	66.3	66.9	62.7	64.1	64.5	66.0	65.4	66.7
TA	50	58.6	61.4	70.6	71.2	71.4	71.2	72.3	71.9	72.3	72.1	72.8	72.7	72.0	72.0	72.9	72.4	72.7	72.8
	100	58.6	61.4	73.6	73.4	73.0	72.7	74.1	73.9	74.3	74.4	74.1	74.1	74.0	74.0	73.7	74.3	74.5	74.4
	250	58.6	61.4	74.9	76.1	75.7	76.1	77.0	76.9	77.0	77.1	77.0	77.0	77.3	76.9	76.7	76.5	77.3	77.1
	500	58.6	61.4	76.7	77.7	76.5	78.3	78.4	79.2	78.7	78.6	79.0	78.6	79.6	79.3	77.9	78.1	78.5	78.3
Urdu	10	56.9	64.0	74.6	75.1	75.0	77.5	77.5	78.0	75.9	77.4	76.8	77.2	77.3	78.6	73.6	76.2	76.8	78.2
UR	50	56.9	64.0	79.7	79.6	80.5	81.4	80.8	80.6	81.3	81.5	80.6	81.3	81.1	81.9	81.0	82.8	81.3	82.2
	100	56.9	64.0	80.5	81.8	80.9	83.3	82.9	82.7	83.2	82.4	82.2	81.7	82.0	82.5	83.3	83.6	82.7	83.2
	250	56.9	64.0	83.8	83.9	84.5	85.4	85.3	85.4	85.0	85.3	84.9	85.4	85.1	85.1	85.1	85.3	85.3	85.8
	500	56.9	64.0	85.7	86.5	85.3	86.6	87.3	87.2	87.2	86.7	87.8	87.5	87.6	87.6	87.4	87.4	87.9	87.7
Vietnamese	10	70.7	70.8	64.2	71.6	64.9	72.0	73.2	75.5	74.1	75.5	75.0	76.1	74.0	74.5	74.9	75.1	74.8	75.6
VI	50	70.7	70.8	78.3	78.9	77.8	78.5	78.4	79.0	78.8	78.7	78.8	78.9	79.5	80.1	78.7	79.2	78.7	79.3
	100	70.7	70.8	79.9	80.1	79.4	79.6	79.8	80.0	79.5	79.6	79.7	80.1	80.7	81.0	79.9	80.6	80.0	80.4
	250	70.7	70.8	82.2	83.0	81.9	82.5	82.3	82.4	82.0	82.0	82.0	82.2	83.0	83.1	82.1	82.1	81.7	82.0
	500	70.7	70.8	83.0	83.2	82.2	83.5	83.4	83.8	83.0	83.1	83.0	82.9	83.8	83.9	82.9	83.1	82.5	82.8
Yoruba	10	28.1	48.7	61.3	65.9	65.5	67.3	61.8	65.8	65.0	66.8	65.2	69.0	63.1	65.7	61.7	62.5	61.2	67.4
YO	50	28.1	48.7	82.0	85.5	83.5	85.1	79.2	83.2	86.8	86.3	84.7	87.7	80.9	83.7	86.4	85.6	84.7	86.1
	100	28.1	48.7	85.1	85.5	89.3	87.2	86.4	87.0	88.1	87.3	88.0	86.7	86.7	87.7	86.7	86.5	87.1	87.3
Chinese	10	25.6	27.8	33.0	33.0	33.0	33.0	38.7	40.6	39.2	40.6	39.2	40.6	35.8	38.9	36.7	39.0	36.7	39.0
ZH	50	25.6	27.8	51.7	52.3	51.7	52.3	53.4	55.0	52.9	53.2	52.9	53.2	54.4	55.2	52.9	53.7	52.9	53.7
	100	25.6	27.8	53.6	56.3	53.6	56.3	58.5	59.3	58.0	58.5	58.0	58.5	58.6	59.1	56.8	57.8	56.8	57.8
	250	25.6	27.8	62.8	63.9	62.8	63.9	65.4	65.5	64.3	64.8	64.3	64.8	65.0	65.3	63.8	64.0	63.8	64.0
	500	25.6	27.8	66.0	67.0	66.0	67.0	68.7	68.9	68.0	68.3	68.0	68.3	68.4	68.6	67.6	67.5	67.6	67.5

A.7 Part-Of-Speech Tagging

		SOURCE		TARGET				SOURCE-TARGET											
		Zero-Shot		Few-Shot				MACRO						MIX-UP					
Weights	k Shots	LM		LAST		ORACLE		LM		LAST		ORACLE		LM		LAST		ORACLE	
Metric		L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O	L	O
Afrikaans	10	86.5	86.5	91.6	91.4	92.8	92.8	91.2	91.1	90.2	90.2	90.9	90.6	91.5	91.5	90.5	90.4	91.0	91.2
AF	50	86.5	86.5	94.5	94.4	94.9	94.7	94.0	94.0	93.0	93.0	93.3	93.5	94.2	94.1	93.0	93.1	93.7	93.7
	100	86.5	86.5	95.3	95.3	95.6	95.6	95.1	95.1	94.7	94.6	94.6	94.6	95.5	95.4	94.7	94.5	94.9	94.6
	250	86.5	86.5	96.8	96.7	96.9	96.9	97.0	96.9	96.3	96.3	96.5	96.6	97.0	97.0	96.4	96.3	96.4	96.4
	500	86.5	86.5	97.3	97.3	97.4	97.5	97.6	97.5	97.4	97.3	97.6	97.5	97.7	97.7	97.2	97.2	97.5	97.4
Arabic	10	70.6	71.4	83.2	83.1	83.2	83.2	83.4	83.4	82.8	82.8	82.7	82.8	82.7	82.8	82.9	82.9	83.1	83.4
AR	50	70.6	71.4	84.8	84.9	85.0	85.1	85.2	85.2	85.4	85.4	85.3	85.2	85.1	85.1	85.3	85.4	85.4	85.3
	100	70.6	71.4	85.4	85.6	85.5	85.6	86.1	86.1	86.2	86.2	86.2	86.2	85.8	85.8	86.2	86.2	86.4	86.4
	250	70.6	71.4	86.7	86.7	86.8	86.8	87.2	87.2	87.3	87.3	87.4	87.3	87.0	87.0	87.2	87.2	87.3	87.3
	500	70.6	71.4	87.4	87.4	87.4	87.5	87.7	87.7	87.8	87.7	87.8	87.8	87.5	87.6	87.7	87.7	87.7	87.7
Basque	10	54.5	55.2	73.7	73.8	74.1	74.1	73.9	74.0	73.4	73.6	73.9	74.2	73.7	74.0	74.2	74.2	74.3	74.4
EU	50	54.5	55.2	81.6	81.5	81.9	81.9	81.9	82.0	81.7	81.9	81.7	81.9	82.0	82.1	82.3	82.3	82.6	82.5
	100	54.5	55.2	84.8	84.8	84.9	85.1	85.4	85.3	85.7	85.7	85.5	85.5	85.4	85.4	85.6	85.7	85.9	85.9
	250	54.5	55.2	88.0	88.4	88.5	88.9	89.0	89.1	89.0	89.1	89.1	89.2	89.0	89.0	89.2	89.2	89.3	89.4
	500	54.5	55.2	90.4	90.6	90.7	90.8	91.1	91.0	91.0	91.0	91.0	91.0	91.0	90.9	91.0	91.0	91.0	91.1
Chinese	10	34.2	40.8	64.9	64.9	65.0	65.2	67.8	68.0	67.3	67.3	67.1	67.4	66.1	66.9	66.9	67.1	67.3	67.2
ZH	50	34.2	40.8	74.9	74.9	75.4	75.6	77.6	77.4	76.8	76.8	77.2	77.1	78.0	77.8	77.0	77.0	77.6	77.6
	100	34.2	40.8	78.7	78.6	79.1	79.2	81.7	81.7	80.8	80.7	81.2	81.3	81.5	81.5	80.8	80.8	81.1	81.1
	250	34.2	40.8	82.9	82.9	83.2	83.1	84.7	84.6	84.2	84.2	84.5	84.5	84.6	84.6	84.1	84.1	84.6	84.5
	500	34.2	40.8	85.5	85.5	85.6	85.7	86.8	86.7	86.5	86.4	86.6	86.6	86.7	86.7	86.3	86.3	86.5	86.5
German	10	86.1	86.3	90.0	90.0	90.0	90.0	90.0	90.1	89.2	89.4	89.1	89.5	90.1	90.1	89.2	89.6	89.3	89.6
DE	50	86.1	86.3	92.4	92.4	92.4	92.4	92.3	92.3	91.6	91.8	91.6	91.7	92.2	92.3	91.7	91.8	91.6	91.9
	100	86.1	86.3	93.4	93.5	93.5	93.5	93.4	93.4	92.9	93.0	92.9	92.9	93.5	93.4	92.9	93.0	92.9	92.9
	250	86.1	86.3	94.6	94.6	94.5	94.7	94.7	94.8	94.4	94.4	94.4	94.4	94.8	94.8	94.5	94.5	94.5	94.5
	500	86.1	86.3	95.2	95.3	95.1	95.3	95.4	95.4	95.2	95.2	95.2	95.2	95.4	95.4	95.2	95.3	95.3	95.3
Hindi	10	66.7	67.2	84.3	84.3	84.5	84.7	84.7	84.8	83.6	83.9	84.2	84.1	84.1	84.3	83.8	83.8	84.0	84.1
HI	50	66.7	67.2	88.4	88.4	88.3	88.3	88.6	88.4	88.3	88.4	88.5	88.5	88.2	88.1	88.4	88.5	88.5	88.6
	100	66.7	67.2	89.1	89.3	89.4	89.3	89.6	89.5	89.6	89.6	89.5	89.6	89.2	89.3	89.5	89.4	89.7	89.6
	250	66.7	67.2	90.5	90.7	90.4	90.6	90.9	90.9	90.9	90.9	90.9	91.0	90.8	90.9	90.9	90.8	90.9	90.9
	500	66.7	67.2	91.1	91.2	91.2	91.2	91.5	91.5	91.5	91.4	91.5	91.5	91.4	91.4	91.5	91.5	91.5	91.5
Hungarian	10	75.0	75.3	86.9	86.9	87.2	87.2	85.2	85.3	84.9	85.1	84.8	85.4	85.6	85.7	84.8	84.9	84.7	85.0
HU	50	75.0	75.3	91.7	91.7	91.8	91.8	92.0	91.8	91.5	91.3	91.3	91.1	91.7	91.7	91.5	91.7	91.5	91.5
	100	75.0	75.3	93.0	93.0	93.2	93.2	93.3	93.2	93.0	92.9	93.1	93.1	93.3	93.3	93.0	92.9	93.1	93.1
	250	75.0	75.3	94.8	94.6	94.9	94.9	94.9	94.8	94.8	94.8	94.9	94.9	95.1	95.0	95.0	94.9	95.0	95.0
	500	75.0	75.3	95.8	95.8	95.9	95.9	95.9	95.9	95.9	95.9	95.9	95.8	95.8	95.9	95.9	95.8	95.9	95.8
Indonesian	10	71.6	71.6	74.1	74.1	74.6	74.6	74.4	74.4	73.4	73.6	73.7	73.8	74.2	74.2	73.9	73.9	74.1	74.0
ID	50	71.6	71.6	76.3	76.3	76.5	76.3	75.7	75.7	75.8	75.9	75.9	75.9	75.7	75.9	76.2	76.3	76.4	76.5
	100	71.6	71.6	76.5	76.5	76.4	76.4	76.1	76.0	76.3	76.3	76.4	76.2	76.0	76.0	76.7	76.7	76.6	76.6
	250	71.6	71.6	77.0	76.9	76.8	76.9	76.8	76.7	77.0	76.9	77.0	77.1	76.7	76.7	77.1	77.1	77.1	77.0
	500	71.6	71.6	76.9	77.0	76.7	76.8	76.8	76.7	77.0	77.0	77.1	77.1	76.9	76.6	77.2	77.1	77.3	77.2
Japanese	10	24.7	28.3	75.2	75.2	75.6	75.5	78.9	79.0	78.2	78.0	78.3	78.5	77.4	77.5	77.3	77.5	77.2	77.2
JA	50	24.7	28.3	81.2	81.2	81.5	81.6	83.8	83.7	83.6	83.6	83.5	83.3	83.6	83.6	83.3	83.1	83.1	82.8
	100	24.7	28.3	83.3	83.3	83.4	83.7	85.1	85.0	85.1	84.8	85.4	85.3	84.7	84.9	84.5	84.6	84.6	84.5
	250	24.7	28.3	86.1	86.1	86.2	86.3	87.3	87.2	87.1	87.0	87.3	87.0	87.1	87.1	86.7	86.6	86.8	86.7
	500	24.7	28.3	87.4	87.8	87.7	87.7	88.1	88.2	88.2	88.1	88.2	88.2	87.8	87.8	88.1	88.0	88.2	88.1
Russian	10	82.8	83.1	85.3	85.3	85.2	85.2	86.2	86.4	85.5	85.6	85.6	85.7	86.4	86.4	85.5	85.6	86.1	86.1
RU	50	82.8	83.1	88.4	88.4	88.8	88.8	88.9	88.9	87.8	87.9	88.1	88.2	89.3	89.3	88.4	88.5	88.8	88.8
	100	82.8	83.1	90.2	90.2	90.3	90.4	90.5	90.6	89.6	89.7	90.1	90.1	90.6	90.6	90.0	90.0	90.3	90.4
	250	82.8	83.1	91.8	91.9	91.9	92.1	92.3	92.3	91.7	91.7	91.9	91.9	92.3	92.3	92.0	92.0	92.1	92.1
	500	82.8	83.1	93.0	93.1	93.1	93.2	93.4	93.4	93.0	93.1	93.3	93.3	93.3	93.3	93.2	93.2	93.4	93.4
Tamil	10	43.5	44.0	66.5	66.3	66.9	66.3	66.2	66.7	67.2	67.1	68.1	67.1	65.5	65.5	67.1	66.7	67.3	66.7
TA	50	43.5	44.0	76.7	75.5	77.7	77.6	78.3	78.7	78.5	78.1	79.7	78.8	77.3	77.2	78.6	78.4	78.7	78.8
	100	43.5	44.0	80.9	80.7	81.0	81.8	82.9	82.7	82.6	82.3	82.9	82.1	81.4	81.4	82.1	82.4	82.4	82.5
	250	43.5	44.0	85.4	84.3	85.6	85.4	86.1	86.0	86.1	85.8	86.3	86.5	85.7	85.8	86.1	86.0	86.0	85.9
Urdu	10	55.6	55.9	83.7	83.7	83.8	83.8	83.6	83.6	83.6	83.4	83.2	83.3	82.9	82.9	83.2	83.3	83.1	83.2
UR	50	55.6	55.9	87.4	87.3	87.5	87.5	87.2	87.2	87.8	87.8	87.5	87.7	87.1	87.1	87.5	87.4	87.8	87.7
	100	55.6	55.9	88.9	88.8	88.7	88.7	88.9	88.7	89.1	89.2	89.0	89.0	88.9	88.9	89.1	88.9	89.0	88.9
	250	55.6	55.9	90.0	90.2	90.0	90.0	90.4	90.2	90.4	90.4	90.6	90.6	89.9	89.9	90.3	90.2	90.3	90.4
	500	55.6	55.9	90.9	90.9	90.8	90.9	90.9	90.9	91.1	91.0	91.2	91.2	90.7	90.7	90.9	91.0	91.0	91.2

A.8 Multilingual Results

		TARGET		S-T (MACRO-LAST)		MULTI (MACRO-LAST)	
		L	O	L	O	L	O
ANLI	*10*	38.3	39.9	38.0	38.1	38.0	38.2
	50	43.8	43.3	44.4	44.4	43.9	44.7
	100	45.8	45.0	46.8	46.6	46.8	46.8
	250	49.7	49.5	51.0	51.2	50.1	50.1
	500	51.7	52.0	53.3	52.9	52.6	52.4
PAWS-X	*10*	81.0	84.2	84.5	84.5	84.4	84.2
	50	83.5	84.2	84.4	84.3	84.4	84.3
	100	84.0	84.3	84.6	84.5	84.2	84.0
	250	83.2	84.9	84.6	84.6	84.3	84.2
	500	83.8	85.3	85.3	85.0	85.2	85.1
NER	*10*	59.8	62.1	65.2	66.1	64.5	66.1
	50	71.4	71.9	72.4	72.7	72.9	73.2
	100	73.0	73.7	74.7	74.8	74.8	74.9
	250	76.6	77.6	77.7	77.7	77.9	77.9
	500	78.2	79.1	79.5	79.6	79.7	79.6
POS	*10*	79.4	79.4	79.4	79.4	80.1	80.2
	50	84.3	84.3	84.3	84.3	84.7	84.7
	100	85.9	85.8	85.9	85.8	86.3	86.2
	250	87.6	87.6	87.6	87.6	87.9	87.9
	500	88.5	88.4	88.5	88.4	88.6	88.6

Table 4: Multilingual FS-XLT transfer results. Please refer to §5 for details.