# MAD-G: Multilingual Adapter Generation for Efficient Cross-Lingual Transfer

**Alan Ansell**[1], **Edoardo Maria Ponti**[1,2], **Jonas Pfeiffer**[3], **Sebastian Ruder**[4], **Goran Glavaš**[5], **Ivan Vulić**[1], **Anna Korhonen**[1]

[1]Language Technology Lab, University of Cambridge
[2]Mila - Quebec AI Institute and McGill University
[3]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
[4]DeepMind
[5]Data and Web Science Research Group, University of Mannheim

## Abstract

Adapter modules have emerged as a general parameter-efficient means to specialize a pretrained encoder to new domains. Massively multilingual transformers (MMTs) have particularly benefited from additional training of language-specific adapters. However, this approach is not viable for the vast majority of languages, due to limitations in their corpus size or compute budgets. In this work, we propose MAD-G (**M**ultilingual **AD**apter **G**eneration), which contextually generates language adapters from language representations based on typological features. In contrast to prior work, our time- and space-efficient MAD-G approach enables (1) sharing of linguistic knowledge across languages and (2) zero-shot inference by generating language adapters for unseen languages. We thoroughly evaluate MAD-G in zero-shot cross-lingual transfer on part-of-speech tagging, dependency parsing, and named entity recognition. While offering (1) improved fine-tuning efficiency (by a factor of around 50 in our experiments), (2) a smaller parameter budget, and (3) increased language coverage, MAD-G remains competitive with more expensive methods for language-specific adapter training across the board. Moreover, it offers substantial benefits for low-resource languages, particularly on the NER task in low-resource African languages. Finally, we demonstrate that MAD-G's transfer performance can be further improved via: (i) *multi-source training*, i.e., by generating and combining adapters of multiple languages with available task-specific training data; and (ii) by further fine-tuning generated MAD-G adapters for languages with monolingual data.

## 1 Introduction

Multilingual NLP has witnessed large advances, with cross-lingual word embedding spaces (Mikolov et al., 2013; Artetxe et al., 2018; Glavaš et al., 2019) and, more recently, massively multilingual Transformers (MMTs) like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021) as main vehicles of cross-lingual transfer. Although MMTs display impressive (zero-shot) cross-lingual transfer abilities (Pires et al., 2019; Wu and Dredze, 2019), their performance has been shown to drop when the target language is typologically distant to the source language, or the size of its pretraining data is limited (Hu et al., 2020; Lauscher et al., 2020). In addition, their coverage of the world's languages—and consequently the range of language technology applications they can support—remains insufficient.[1]

Adapters (Rebuffi et al., 2017; Houlsby et al., 2019) have been proposed as a parameter-efficient means to extend multilingual models to under-represented languages (Bapna and Firat, 2019; Üstün et al., 2020). The general practice is to train a language adapter on the unlabeled data for each language (Pfeiffer et al., 2020b) via masked language modeling (MLM). However, this generally requires substantial amounts of monolingual data, which prevents adapters from serving under-resourced languages where such additional language-specific capacity would be most useful.

To address this deficiency, we propose *multilingual adapter generation* (MAD-G), a novel paradigm that enables the generation of adapters for low-resource languages by *sharing information across languages*. Instead of learning separate adapters for each language, MAD-G leverages contextual parameter generation (CPG; Platanios et al., 2018a; Ponti et al., 2019b), that is, it learns a single model that can generate a language adapter for an arbitrary target language. At the core of MAD-G is a contextual parameter generator which

[1]mBERT and XLM-R have been trained on corpora from 104 and 100 languages, respectively. According to Glottolog (Hammarström et al., 2017), however, there are over 7,000 languages spoken around the world.



*Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781
November 7–11, 2021. 

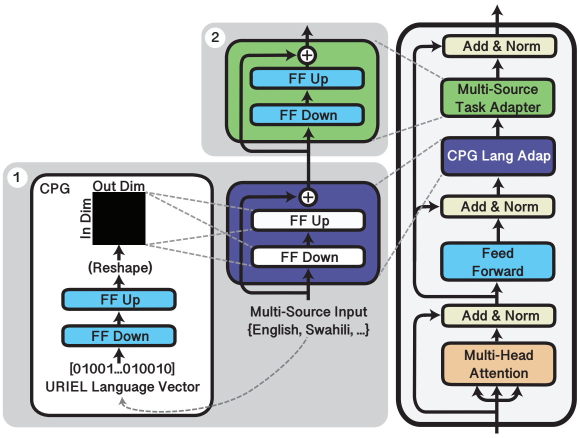


Figure 1: Cross-lingual transfer with MAD-G. ① MAD-G training: the generator component learns to generate language-specific adapters given URIEL vectors of input languages; the parameters of the generator are trained with an MLM objective, where instances of the respective language are passed through the frozen Transformer layers and the generated adapter parameters. ② In the downstream task fine-tuning, both the Transformer weights as well as the weights of the generated source-language adapter are frozen; an additional task adapter with randomly initialized weights is placed on top of the generated source language adapter. During target language downstream inference, the generated source language adapters are replaced with the generated target language adapters.

takes the typological vector of a language as input and outputs the parameters of the language-specific adapter. The generator's parameters are trained via MLM on the Wikipedias of 95 languages, selected to maximize linguistic diversity. Unlike prior CPG work (Platanios et al., 2018a; Üstün et al., 2020), MAD-G generates language adapters that are task-agnostic, thus allowing for an efficient and modular cross-lingual transfer across the board, i.e., the MAD-G language adapters can be leveraged in arbitrary downstream tasks (Pfeiffer et al., 2020b).

MAD-G shares information across languages (i) at the level of hidden representations by sharing the parameters of the adapter generator as well as (ii) at the typological level by conditioning on features from the URIEL database (Littell et al., 2017). The latter additionally enables zero-shot transfer to unseen languages. Further, we propose a variant of MAD-G in which we generate adapters also conditioned on their Transformer layer position (see Section 3.2), allowing MAD-G to be much more parameter-efficient than adapter-based transfer methods of prior work.

In experiments on zero-shot cross-lingual transfer on part-of-speech tagging (POS), dependency parsing (DP), and named entity recognition (NER), MAD-G demonstrates competitive performance to training more expensive language-specific adapters and shows strong performance in low-resource scenarios, e.g., in the NER task for African languages. What is more, we show that transfer performance can be further improved by (a) multilingual training of task adapters and (b) fine-tuning of generated MAD-G adapters, via MLM, on small amounts of monolingual data. Finally, we provide a nuanced analysis of transfer performance to unseen languages, highlighting the importance of the diversity of the language sample selected for pretraining.

## 2 Background

Before introducing MAD-G in detail in Section 3, we recapitulate its key components adopted from previous work. In particular, we discuss language adapters (LA) in Section 2.1 and Contextual Parameter Generation (CPG) in Section 2.2.

### 2.1 (Why) Language Adapters

Massively multilingual models infamously suffer from the 'curse of multilinguality' (Arivazhagan et al., 2019; Conneau et al., 2020): for a fixed model capacity, their performance decreases as they cover more languages. Extending them to underrepresented and unseen languages is far from trivial: additional training (of all model parameters) for such languages can lead to catastrophic forgetting of the previously acquired knowledge (McCloskey and Cohen, 1989; Santoro et al., 2016). A common remedy for both their coverage–performance tradeoff and limited flexibility is to allocate *additional* model parameters for individual languages. This is typically achieved through the use of adapter layers (Houlsby et al., 2019; Pfeiffer et al., 2020b).

In particular, a language adapter is a light-weight component inserted into a MMT such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) with the purpose of specializing the MMT for a particular language, in order to either (a) support a new language not covered by the MMT's original multilingual pretraining (Pfeiffer et al., 2020b; Artetxe et al., 2020) or (b) recover/improve the performance for a particular (resource-rich) language (Bapna and Firat, 2019; Rust et al., 2021). In this work, we adopt the competitive and lightweight (so-called *bottleneck*) adapter variant of Pfeiffer et al. (2021a). There, only one adapter module, consisting of a successive down-projection and up-projection, is injected per Transformer layer, after the feed-forward sublayer (see Figure 1).[2] The language adapter $\text{LA}_b$ at the $b$-th Transformer layer/block performs the following operation:

$$\text{LA}_b(\boldsymbol{h}_b, \boldsymbol{r}_b) = U_b \, \text{a}(D_b \boldsymbol{h}_b) + \boldsymbol{r}_b, \tag{1}$$

where $\boldsymbol{h}_b$ and $\boldsymbol{r}_b$ are the Transformer hidden state and the residual at layer $b$, respectively. $D_b \in \mathbb{R}^{h \times m}$ and $U_b \in \mathbb{R}^{m \times h}$ are the down- and up-projections, respectively ($h$ being the Transformer's hidden layer size, and $m$ the adapter's dimension), and $\text{a}(\cdot)$ is a non-linear activation function. The residual connection $\boldsymbol{r}_b$ is the output of the Transformer's feed-forward layer whereas $\boldsymbol{h}_b$ is the output of the subsequent layer normalisation. The parameters of a language adapter are learned through MLM with the original parameters of the MMT kept frozen (Pfeiffer et al., 2020b).

### 2.2 (Why) Contextual Parameter Generation

Language adapters are an instance of a common design pattern in multilingual NLP: training a separate model or model components for each target language.[3] This approach based on a separate instance per language has two crucial drawbacks: **1**) the total training time and number of parameters learned increase linearly with the number of languages; **2**) a lack of information sharing across languages due to the complete independence of learned parameters, which prevents low-resource languages from benefiting from their typological and genealogical ties to high(er)-resource languages.

CPG is a technique introduced by Platanios et al. (2018a) to address these drawbacks. While originally conceived for neural machine translation (NMT), CPG can be applied to any neural model $f$ parameterized by $\boldsymbol{\theta}$, for which we aim to learn parameterizations for a number of different *contexts*; in multilingual NLP, these "contexts" are languages. In the instance-per-language approach, an independent parameterization $\boldsymbol{\theta}^{(l)}, l \in \{1, \ldots, n_l\}$, is learned for each of the $n_l$ languages of interest. In CPG, the only language-specific parameters that we learn are the low-dimensional *language embeddings* $\boldsymbol{\lambda}^{(l)} \in \mathbb{R}^{d_l}$. These are used by the generator $g$, a hyper-network (Ha et al., 2017) component[4] with its own parameterization $\boldsymbol{\phi}$, to produce the language-specific parameterization of the main model: $\boldsymbol{\theta}^{(l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\lambda}^{(l)})$. While $g$ can in principle be any differentiable function (i.e., arbitrarily deep neural model), in practice it is typically set to a simple linear projection (i.e., $\boldsymbol{\phi} = W$):

$$g_W(\boldsymbol{\lambda}^{(l)}) \triangleq W\boldsymbol{\lambda}^{(l)}, \tag{2}$$

where $W \in \mathbb{R}^{n_p \times d_l}$ is a learnable weight matrix, $n_p$ being the number of parameters of $f$.

The total number of parameters learned when training $n_l$ independent models is $n_l n_p$, whereas the number of parameters in the $W$ matrix is $d_l n_p$. Therefore, neglecting the small number of parameters dedicated to language embeddings, the CPG approach uses fewer parameters when $d_l < n_l$.[5] More importantly, in multilingual training the generator matrix $W$ is shared across all languages, which enables knowledge sharing across languages and leads to improved transfer performance.

Platanios et al. (2018b) and Ponti et al. (2021a) opt for randomly initializing language embeddings $\boldsymbol{\lambda}^{(l)}$ and learning them end-to-end. Specified like this, however, CPG cannot generalize to languages unseen in training, as it would lack embeddings for those languages at inference. To support generalization to arbitrary new languages, one must ground language embeddings in some external language representation, available for many languages. To this end, Ponti et al. (2019b) exploit typological language vectors from the URIEL database (Littell et al., 2017) directly as language embeddings to generate a full set of model parameters. In a similar vein, Üstün et al. (2020) use the typological language vectors from URIEL to generate task- and language-specific adapters for dependency parsing: they learn the parameters $\boldsymbol{\phi}$ of the generator $g$ via multilingual dependency parsing training on 13 languages. In contrast, MAD-G's multilingual MLM training allows the generation of task-agnostic LAs that can support downstream cross-lingual transfer for arbitrary NLP tasks.

[2] According to Pfeiffer et al. (2020a, 2021a) and Rücklé et al. (2021), such an architecture with a single adapter per Transformer layer is more parameter-efficient while performing on par with the architecture of Houlsby et al. (2019) with two adapters per Transformer layer (one after the multi-head attention sublayer and one after the feed-forward sublayer).

[3] Other examples include the training of language-specific pretrained language models (Rust et al., 2021) as well as language pair-specific encoder–decoder models for machine translation (Luong et al., 2016; Firat et al., 2016).

[4] A hyper-network is a neural model that generates the parameters of another (main) neural model.

[5] Training MAD-G on 95 languages with $d_l = 32$ (this work) achieves roughly a threefold saving in parameter size.

## 3 MAD-G: Methodology

MAD-G aims to enable resource-efficient adaptation of MMTs to a wide range of previously unseen, radically resource-poor languages,[6] and contribute in this manner to more sustainable (Strubell et al., 2019; Moosavi et al., 2020) and more inclusive NLP (Joshi et al., 2020). We couple (i) the computational efficiency of the light-weight adapters (cf. Section 2.1) and (ii) knowledge sharing and zero-shot language transfer capabilities of CPG (cf. Section 2.2), with (iii) external linguistic (i.e., typological) knowledge (Ponti et al., 2019a) towards supporting arbitrary NLP tasks for (even radically) resource-poor languages.

MAD-G mitigates important limitations of prior work. Unlike Üstün et al. (2020), we generate *task-agnostic* LAs, (re)usable across NLP tasks. Unlike the MAD-X framework (Pfeiffer et al., 2020b), which trains LAs independently for each language (requiring sufficient monolingual corpora), MAD-G can support unseen and resource-poor languages in downstream tasks by generating LAs from typological vectors. Moreover, MAD-G leverages typological relations between languages. We also show that the two approaches can be successfully combined: monolingual MLM fine-tuning of a MAD-G-generated LA yields further benefits.

### 3.1 Generating Language Adapters

Our input representation for each language is a sparse typological vector $\boldsymbol{t}^{(l)}$ encompassing 289 binary linguistic features (103 syntactic, 28 phonological and 158 phonetic features) from the URIEL language typology database (Littell et al., 2017). We obtain the language embedding $\boldsymbol{\lambda}^{(l)}$ from $\boldsymbol{t}^{(l)}$ using a single-layer linear down-projection: $\boldsymbol{\lambda}^{(l)} = V\boldsymbol{t}^{(l)}$, with the parameter matrix $V \in \mathbb{R}^{d_l \times 289}$. Down-projecting to a dimension $d_l << 289$ prevents $W$ from being impractically large. By grounding language embeddings in external expert linguistic knowledge (i.e., URIEL vectors), we enable generalization to all languages for which such typological vectors exist, regardless of the availability of monolingual text for those languages for generator training. In multilingual MLM training, we generate the adapter parameters $\boldsymbol{\theta}^{(l)}$ for each instance from the embedding of the respective language, as specified in Eq (2).[7] Let $n_b$ be the number of layers in the MMT (e.g., for mBERT (Devlin et al., 2019), $n_b = 12$). The MAD-G parameter matrix $W$ then has $n_b \cdot 2 \cdot h \cdot m \times d_l$ parameters, where $h$ is the hidden size of the Transformer layer and $m$ the bottleneck size of the adapter layer (i.e., a single adapter module has $2 \cdot h \cdot m$ parameters).

### 3.2 Factoring Out Layer Embeddings

By factoring out language-specific embeddings $\boldsymbol{\lambda}^{(l)}$, we force the MAD-G parameters $W$ to share knowledge across languages. The generated language adapters in different Transformer layers are, however, still mutually independent. By additionally factoring out representations of each Transformer layer indices into *layer embeddings* $\boldsymbol{\lambda}^{(b)} \in \mathbb{R}^{d_b}$, $b \in \{1, 2, \ldots, n_b\}$, we can condition the adapter generation not only on languages but also on layers. This has two potential benefits: (i) it allows for information sharing between adapters of different layers, and, more importantly, (ii) it substantially reduces the size of the generator $W$. In this model variant, dubbed **`MAD-G-LS`**, the generator outputs adapters $\boldsymbol{\theta}^{(l,b)}$ for language-layer pairs:

$$\boldsymbol{\theta}^{(l,b)} \triangleq W(\boldsymbol{\lambda}^{(l)} \oplus \boldsymbol{\lambda}^{(\boldsymbol{b})}), \tag{3}$$

with the concatenation of the language embedding $\boldsymbol{\lambda}^{(l)}$ and layer embedding $\boldsymbol{\lambda}^{(\boldsymbol{b})}$ as input. The **`MAD-G-LS`** generator has $2 \cdot h \cdot m \times (d_l + d_b)$ parameters, which is, assuming language and layer embeddings of equal size (i.e., $d_b = d_l$), a parameter reduction by a factor $\frac{n_b}{2}$ compared to the base MAD-G configuration from §3.1.

### 3.3 Multi-Source Task Adapters

Once the multilingual adapter generator has been trained via multilingual MLM, the generated LAs can be used to facilitate downstream cross-lingual transfer. Here, we follow the task-specific fine-tuning setup of MAD-X (Pfeiffer et al., 2020b): we insert and train the task-specific adapter (TA) on top of the language adapter of the source language—the parameters of the LA as well as parameters of the original MMT are kept frozen. In prior work, the TA is trained on data from a *single source* language $l_s$ with the LA for $l_s$ activated (with frozen parameters). At inference time, the LA for the *tar-*

[6]With "radically resource-poor" languages we refer to languages for which even the acquisition of non-negligible amounts of text data is difficult.

[7]An alternative option for adapter generator input would be randomly initialized language embeddings $\boldsymbol{\lambda}^{(l)}$; this would, however, prevent the opportunity of downstream generalization to unseen languages.

*get language* $l_t$ is plugged in instead of $l_s$'s adapter, with the same TA (Pfeiffer et al., 2020b).

In downstream tasks with task data in multiple languages, we can resort to *multi-source* transfer, i.e., multilingual training of the task adapter. This is possible with per-language trained LAs (e.g., MAD-X adapters) as well as without any LAs. We hypothesized that multi-source training would be particularly beneficial with MAD-G because of the knowledge shared by LAs of different languages as a result of their generation with the MAD-G's multilingual generator. In other words, with MAD-G, the multi-source task adapter training is supported by a single LA generator model (see Figure 1), rather than a set of independently trained LAs. However, our experiments show that multi-source training is greatly beneficial regardless of language adapter type; the advantage does not seem larger for MAD-G in particular.

We employ a straightforward approach to TA training on the set of source languages $L_s$: in each step, we (1) randomly select a language $l$ from $L_s$ from which we sample a training batch and (2) in the forward pass – before the task adapter – activate the LA of the language $l$ for that batch. To the best of our knowledge, we are the first to investigate multi-source adapter-based transfer in cross-lingual settings.

## 4 Experimental Setup

**Tasks and Languages.** We evaluate on three downstream tasks which provide sufficient evaluation data for low-resource languages: part-of-speech (POS) tagging, dependency parsing (DP), both on the Universal Dependencies (UD) 2.7 dataset (Zeman et al., 2020), and named entity recognition (NER) on the MasakhaNER dataset for African languages (Adelani et al., 2021). For POS and DP, we evaluate on a substantial subset of all UD languages with available treebanks.[8] We discern between three language groups in evaluation, with some examples in Table 1: (i) `mBERT-seen` languages are those included in mBERT's pretraining; (ii) `MAD-G-seen` languages were not part of mBERT's pretraining but are included in MAD-G training; and (iii) `unseen` languages are those not included in mBERT pretraining nor in MAD-G training.

### 4.1 Baselines and MAD-G Variants

**`mBERT`** is an MMT pretrained on the Wikipedias of 104 languages. We use mBERT as the base MMT for MAD-G. **`XLM-R`** is a state-of-the-art MMT pretrained on the CommonCrawl data of 100 languages (Conneau et al., 2020).[9] We evaluate them in the standard transfer setup with full-model fine-tuning (**`-ft`**).

**`MAD-X`** is the state-of-the-art modular adapter-based framework for cross-lingual transfer (Pfeiffer et al., 2020b) based on independent MLM-training of a dedicated LA for each language. We train our own MAD-X LAs when no pretrained ones are available, notably for the six `MAD-G-seen` UD languages. Training LAs for all other low-resource languages, however, is prohibitively computationally expensive,[10] so during all MAD-X experiments, the pool of languages with available MAD-X adapters consists of the 20 high-resource source languages used in multi-source setups (see Section 4.2) and `MAD-G-seen` languages. When evaluating on a target language without an available MAD-X LA, we instead choose the available MAD-X LA for the language that is *closest* to the target language.[11]

**`MAD-G`** is the base setup of our method from Section 3.1. **`MAD-G-LS`** is the variant of MAD-G in which the adapter generation is additionally conditioned on layer embeddings, as described in Section 3.2. **`MAD-G-en`** uses the English adapter rather than that of the target language during inference on target language instances. The purpose of this baseline is to test if the parameters generated for different languages are actually meaningfully different and able to outperform the English LA.

**`TA-only`** trains the task adapter directly on top of the MMT, i.e., without any language adapter. With

[8] For POS and DP, we omit only (i) languages with scripts unseen in mBERT's pretraining, where mBERT's tokenizer predominantly produces unknown (`UNK`) tokens (Pfeiffer et al., 2021b), (ii) languages lacking any information in URIEL, and (iii) languages whose treebanks have missing fields. For MasakhaNER, we evaluate on all dataset languages except Amharic, as Amharic also uses a script unseen by mBERT.

[9] Although it mostly outperforms mBERT in multilingual and cross-lingual transfer experiments, mBERT was used in prior work as a more robust choice for radically resource-poor languages in general (Pfeiffer et al., 2020b). Our NER experiments on African languages confirm this (Table 3 later). Note that MAD-G can be applied to XLM-R as well.

[10] Note that this efficiency and scalability shortcoming of MAD-X is precisely one of the main motivations for MAD-G, i.e., for language adapter generation for unseen languages.

[11] We quantify the linguistic proximity of languages as the cosine similarity between their respective URIEL-based language vectors (Lauscher et al., 2020).

| group | definition | # with UD treebank | language examples |
|---|---|---|---|
| `mBERT-seen` | seen during mBERT pretraining | 56 | English, Japanese, Chinese |
| `MAD-G-seen` | seen only during MAD-G training | 6 | Buryat, Maltese, Erzya |
| `unseen` | completely unseen | 33 | Bhojpuri, Moksha, Warlpiri |

Table 1: Definitions of three language groups. "# with UD treebank" is the number of languages belonging to each group included in the evaluation of the UD POS-tagging/dependency parsing tasks.

this baseline, we seek to quantify the contribution of dedicated LAs in general.

### 4.2 MAD-G Training Setup

MLM-training of MAD-G's adapter generator is run on Wikipedias of 95 languages. We considered only the languages with at least 1,000 Wikipedia articles and selected them following a greedy process that maximizes typological diversity. At each step, we select the language with the largest number of articles belonging to the language family and its *genus* that are least represented in the current sample of languages (Ponti et al., 2020); see Appendix for a full list.

Following Pfeiffer et al. (2020b), the LA bottleneck size is $m = 384$. Both the language embedding dimension $d_l$ and the layer embedding (if used) dimension $d_b$ are set to 32. At each MLM training step, we randomly sample a batch in a language from an exponentially smoothed distribution with a cap preventing oversampling of high-resource languages: the probability of selecting a language $l$ is proportional to $\min(\text{n\_examples}^{(l)}, 500,000)^{0.5}$. Training runs for 200,000 steps in total over all languages; batch size is 64 and the maximum sequence length is 256. We used a linearly decreasing learning rate, starting at $5e$-5. In contrast, relying on the same batch size and max sequence length, MAD-X was trained for 100,000 steps *for each language*. This makes the average per-language duration of MAD-G training *≈50 times shorter* than for MAD-X. Moreover, **`MAD-G`** and **`MAD-G-LS`** have 226M and 38M parameters respectively, compared to 728M for a hypothetical 95 MAD-X dedicated language adapters.

**Single- and Multi-Source Transfer.** We train task adapters on English data with the English MAD-G adapter. For comparability, we adopt the TA configuration of MAD-X (Pfeiffer et al., 2020b): the bottleneck size is $m = 48$. For POS-tagging and NER we use the standard token-level single-layer multi-class classifier. For DP, we use the shallow variant (Glavaš and Vulić, 2021) of the biaffine dependency parser of Dozat and Manning (2017). For POS tagging and DP, we train on the English EWT treebank. For consistency and comparability with multi-source experiments, we sample 12,000 sentences for training (out of the 12,543 available examples). For NER, we train on the CoNLL 2003 English dataset (Tjong Kim Sang and De Meulder, 2003).[12] For all tasks, we train for 15,000 steps with batch size 8 (roughly 10 epochs) and a linearly decreasing learning rate, starting at $5e$-5.

For multi-source transfer experiments, we select 20 typologically diverse high-resource source languages for POS-tagging and DP using the following process: we iterate over the UD languages in the descending order of treebank size and select a language if it belongs to a genus not already represented in the sample.[13] We again sample a total of 12,000 examples (600 per language).

## 5 Results and Discussion

In what follows, we focus on reporting and analyzing the most important global trends in results with accompanying discussions and side experiments. For completeness, the full results per individual target language are provided in the Appendix.

**Single-Source Transfer.** Relative to all methods which do not employ language adaptation, we find that the use of MAD-G in the primary **`MAD-G`** and **`MAD-G-LS`** settings is greatly beneficial on all tasks for `MAD-G-seen` languages in both the single- and multi-source transfer scenarios (see Tables 2 and 3), with the very parameter-efficient **`MAD-G-LS`** being only slightly weaker than the base **`MAD-G`** variant in general, even slightly outperforming it for some languages and transfer setups. Despite having far less capacity per target language, MAD-G retains much of the performance gain of MAD-X on languages seen during language adapter training, showing that MAD-G

[12] As MasakhaNER does not have the `MISC` category, we replace the `B-MISC` and `I-MISC` token tags with the `O` tag in the CoNLL training set. Similarly, we exclude the `DATE` class (i.e., `B-DATE` and `I-DATE` tags) from the MasakhaNER evaluation, because they do not exist in the CoNLL dataset.

[13] For comparability with single-source experiments, we selected English instead of German as the only exception.

| source | method | Part-of-speech tagging | | | Dependency parsing | | |
|---|---|---|---|---|---|---|---|
| | | mBERT-seen | MAD-G-seen | unseen | mBERT-seen | MAD-G-seen | unseen |
| en | **MAD-G** | 76.7 | 65.9 | 44.4 | 63.9/49.2 | 46.3/28.0 | 34.7/16.8 |
| | **MAD-G-LS** | 77.8 | 65.2 | 43.9 | 64.9/49.9 | 44.4/26.0 | 34.7/16.0 |
| | **MAD-G-en** | 76.5 | 40.5 | 44.9 | 66.4/**51.9** | 27.6/11.0 | 35.4/**18.2** |
| | **TA-only** | 78.4 | 40.8 | **45.5** | **67.0**/51.8 | 29.6/11.4 | **36.0**/18.1 |
| | **MAD-X** | 76.9 | **68.8** | 43.4 | 61.5/46.9 | **48.6/30.8** | 33.1/15.7 |
| | **mBERT-ft** | 76.6 | 38.7 | 43.9 | 66.3/51.3 | 27.8/10.0 | 34.0/16.4 |
| | **XLM-R-ft** | **79.6** | 46.8 | 43.6 | 55.4/42.0 | 30.0/13.4 | 31.9/15.5 |
| multi | **MAD-G** | 86.1 | 71.0 | 50.4 | 75.6/65.4 | 54.4/38.0 | 40.1/23.1 |
| | **MAD-G-LS** | 86.5 | 70.0 | 51.0 | 76.6/66.5 | 53.9/36.9 | **41.6/23.7** |
| | **MAD-G-en** | 85.8 | 45.8 | 50.5 | 75.8/65.6 | 33.1/15.2 | 40.3/23.6 |
| | **TA-only** | 86.8 | 48.8 | 51.2 | 76.9/66.8 | 35.7/17.0 | 41.3/**23.7** |
| | **MAD-X** | 83.7 | **73.8** | 47.3 | 74.7/64.2 | **58.1/42.9** | 39.6/22.5 |
| | **mBERT-ft** | 87.4 | 45.4 | 51.2 | **80.6/70.4** | 35.5/15.6 | 41.3/23.4 |
| | **XLM-R-ft** | **89.4** | 53.9 | **55.0** | 65.5/55.4 | 36.8/19.4 | 36.3/21.4 |

Table 2: UD POS tagging accuracy scores and dependency parsing unlabeled/labeled attachment scores for various language adapter and fine-tuning settings. Values are shown as averages over each of the language groups mBERT-seen, MAD-G-seen and unseen, defined in Table 1. Task adapters are trained only on English data (*en*, upper part) and 20 diverse, high-resource languages (*multi*, lower part). The highest score per column in each of the two setups is in **bold**, the second highest is underlined.

| method | hau MAD-G-seen | ibo MAD-G-seen | kin MAD-G-seen | lug unseen | luo unseen | pcm unseen | swa mBERT-seen | wol unseen | yor mBERT-seen | avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| **MAD-G** | **77.1** | **69.9** | **66.1** | 54.2 | 32.5 | **72.6** | 72.6 | 32.1 | **68.8** | **60.7** |
| **MAD-G-LS** | 72.8 | 67.5 | 63.0 | **55.7** | **33.3** | 72.4 | 71.3 | **36.7** | 68.4 | 60.1 |
| **MAD-G-en** | 44.9 | 54.5 | 51.4 | 50.6 | 32.9 | 70.4 | 69.2 | 36.4 | 63.9 | 52.7 |
| **TA-only** | 43.4 | 55.7 | 52.8 | 47.9 | 32.8 | 72.3 | 68.6 | 32.1 | 65.3 | 52.3 |
| **mBERT-ft** | 43.2 | 45.5 | 49.9 | 49.3 | 31.6 | 70.5 | 65.8 | 28.1 | 54.3 | 48.7 |
| **XLM-R-ft**† | 66.4 | 45.5 | 36.1 | 34.8 | 31.9 | 68.4 | **74.5** | 21.6 | 33.4 | 45.8 |

Table 3: $F_1$ scores on the MasakhaNER dataset for African languages. Task adapter training/model fine-tuning is conducted on the CoNLL 2003 English NER dataset. †**XLM-R-ft** results are as reported by Adelani et al. (2021).

| method | Part-of-speech tagging | | | Dependency parsing | | |
|---|---|---|---|---|---|---|
| | mBERT-genus | MAD-G-genus | unseen-genus | mBERT-genus | MAD-G-genus | unseen-genus |
| **MAD-G** | 49.1 | 40.6 | 34.0 | 38.2/19.7 | **28.4/13.2** | 28.5/11.1 |
| **MAD-G-LS** | 50.0 | **40.8** | 29.4 | 38.7/19.2 | 26.2/11.9 | 28.5/9.8 |
| **MAD-G-en** | 51.1 | 37.5 | 32.2 | 39.7/**21.4** | 24.3/11.1 | 29.8/**13.2** |
| **TA-only** | **51.5** | 37.9 | 33.4 | **40.4**/21.3 | 26.9/11.9 | 29.0/12.7 |
| **MAD-X** | 49.3 | 38.3 | 30.3 | 37.3/18.8 | 23.8/9.0 | 26.5/10.7 |
| **mBERT-ft** | 48.7 | 37.3 | **34.5** | 37.6/19.4 | 23.5/8.6 | **29.9**/12.4 |
| **XLM-R-ft** | 50.8 | 39.1 | 27.1 | 34.7/17.7 | 24.5/10.2 | 28.4/12.5 |

Table 4: UD POS tagging accuracy scores and dependency parsing unlabeled/labeled attachment scores for for various language adapter/fine-tuning settings. Values are shown as averages over each of the language groups mBERT-genus, MAD-G-genus and unseen-genus. The task adapter is trained only on English data.

achieves efficient yet effective language adaptation. The **MAD-G-en** variant does not achieve such gains on MAD-G-seen languages, demonstrating that MAD-G does generate meaningfully different adapter parameters for different languages.

The use of MAD-G is not in general beneficial for mBERT-seen languages; this is unsurprising since it is unrealistic to believe that mBERT's knowledge of languages observed during its own pretraining can be substantially improved through language adaptation on a much smaller amount of data. At first glance there also does not appear to be any benefit to using MAD-G for unseen target languages, except for NER, where gains are substantial. However, averaging the results over all languages in this group does not provide a full picture because it consists of languages whose relationships to those observed during training differ substantially. Therefore, we provide a finer-grained analysis below.

While the use of typological vectors for generating LAs allows MAD-G to learn features which could generalize well to unseen languages, this assumption should mostly hold for unseen languages whose 'typological relatives' are available during training. To investigate the effect the degree of typological relatedness has on MAD-G's generalization ability, we further divide the unseen lan-

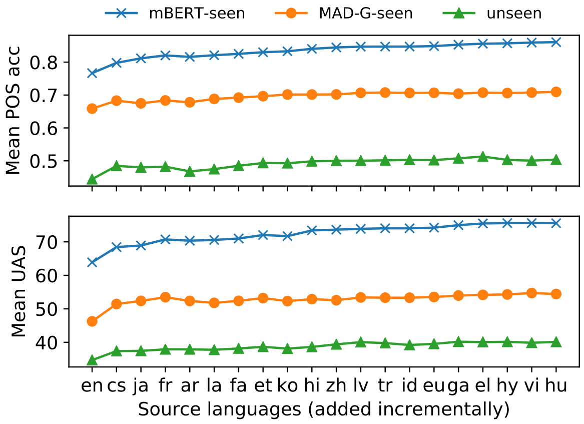


Figure 2: Multi-source transfer with MAD-G. We increase the number of source languages left-to-right from 1 to 20 while keeping the total number of (multi-source) examples constant at each step.

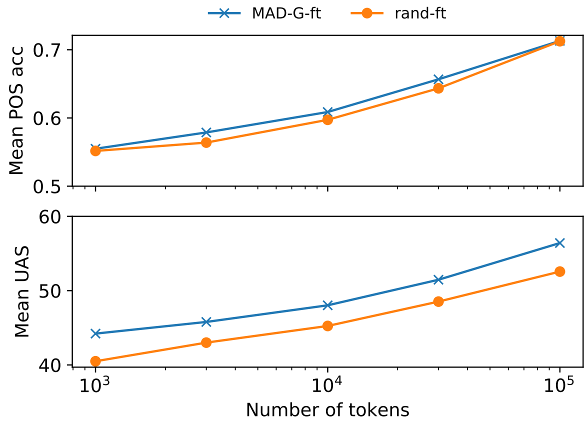


Figure 3: Performance on POS tagging and DP on `unseen` languages when MAD-G-initialized (**`MAD-G-ft`**) or randomly initialized (**`rand-ft`**) language adapters are fine-tuned by MLM on varying amounts of unlabeled text.

guages into three subgroups: `mBERT-genus` (the 21 languages whose genus matches that of at least one language seen during mBERT pretraining); `MAD-G-genus` (the 4 languages whose genus was not seen during mBERT pretraining but was seen during MAD-G training); `unseen-genus` (the 8 languages whose genus is completely unseen). Table 4 shows the POS tagging and DP performance for each of the three `unseen` subgroups. MAD-G is beneficial on the `MAD-G-genus` subgroup, while its benefits do not extend to the other two subgroups. The results for `mBERT-genus` versus `MAD-G-genus` languages mirror those for `mBERT-seen` versus `MAD-G-seen` languages; in general, mBERT's knowledge of a genus (or specific language) can be improved through language adaptation if and only if that genus/language was not observed during mBERT's pretraining. As expected, the scores on `unseen-genus` languages confirm the intuition that the performance on languages typologically unrelated to any language seen during mBERT and/or MAD-G training cannot be recovered solely on the basis of limited external typological information. For cross-lingual generalization, the typological diversity of pretraining languages is thus paramount.

**Multi-Source Transfer.** When training on 20 languages, while maintaining the overall number of training examples, we observe large gains across all settings and language groups for both POS tagging and DP (see Table 2). This suggests that multi-source training yields a more general and language-agnostic representation of the task adapter, thus transferring better to unseen languages. We investigate the effect of multi-source training further in Figure 2, where we gradually add languages to the multi-source pool, while (again) maintaining the overall number of training examples. We find that the transition from one language to two languages in the source-pool results in the largest relative performance increase, but the performance still rises with the addition of more languages. In sum, in line with previous findings (Ponti et al., 2021b), our results indicate that the language diversity of training data has strong positive effects on zero-shot transfer across multiple methods and setups.

**Fine-tuning MAD-G-Initialized Adapters.** Although interesting from a theoretical point of view, the scenario where there is no unannotated data whatsoever available for the target language might be unrealistic. We thus examine a setup where there is a small amount of unannotated data available. In this case, we can still exploit MAD-G by generating an initialization of a language-specific adapter for a target language $l_t$, and then fine-tuning its parameters via MLM on the unannotated data.

We perform POS tagging and DP experiments when fine-tuning MAD-G-initialized language-specific adapters on the 14 `unseen` UD languages which have Wikipedias.[14] We simulate different degrees of resource-poverty by sampling training datasets with 1,000, 3,000, 10,000, 30,000 and 100,000 words from the full Wikipedia. We compare this **`MAD-G-ft`** setting with the results of fine-tuning randomly-initialized LAs on the same data

[14]We do not perform NER experiments because there are only two `unseen` MasakhaNER languages with Wikipedias.

(**rand-ft**).[15] Figure 3 shows that there is a large and consistent improvement on the 14 `unseen` evaluation languages as their language adapters are fine-tuned on increasingly large amounts of unannotated text. For both tasks, the performance is better when the language adapter is initialized with the weights generated by MAD-G than when the weights are randomly initialized. The difference between the two settings is modest for POS tagging, but it is larger for DP and is maintained even when 100,000 training tokens are available.

## 6 Conclusion

We proposed MAD-G, a modular and efficient cross-lingual transfer framework for low-resource languages, that generates task-agnostic adapters for massively multilingual Transformers (e.g., mBERT) from typological language representations. MAD-G performs competitively with a state-of-the-art adapter-based transfer approach MAD-X; yet its training is roughly 50 times more efficient per target language. MAD-G can also be applied to unseen languages, benefiting those belonging to a genus introduced during its training, and it can be used as a better initialization for "radically low-resource languages"; there, its generated language adapters can be further refined on small amounts of text, improving downstream performance. We further show that cross-lingual performance with adapters can be greatly improved by training on multiple source languages. We release the MAD-G code online at: https://github.com/Adapter-Hub/adapter-transformers.

## Acknowledgements

Alan wishes to thank David and Claudia Harding for their generous support via the Harding Distinguished Postgraduate Scholarship Programme. Jonas is supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. Goran is supported by the KI-Innovation grant Multi2ConvAI of the Baden-Württemberg's Ministry of Economics, Labor and Tourism. Anna and Ivan are supported by the ERC Grant LEXICAL (no. 648909) and the ERC PoC Grant MultiConvAI (no. 957356).

[15]We fine-tune both variants for 200 epochs with batch size 4 and learning rate 5$e$-5. We evaluate the fine-tuned language adapters on POS tagging and DP using our better-performing multi-source task adapters trained on 20 languages.

## References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *arXiv preprint*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. 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 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

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. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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 *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 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.

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.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. OpenReview.net.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint*.

Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf, editors. 2020. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Online.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021a. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. 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.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021b. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online, November , 2021*.

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.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018a. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018b. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium. Association for Computational Linguistics.

Edoardo Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021a. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*, 9(0):410–428.

Edoardo Maria Ponti, Rahul Aralikatte, Disha Shrivastava, Siva Reddy, and Anders Søgaard. 2021b. Minimax and neyman–Pearson meta-learning for outlier languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1245–1260, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019b. Towards zero-shot language modeling. 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 2900–2910, Hong Kong, China. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 506–516.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the Efficiency of Adapters in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online, November , 2021*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021, Online, August 1-6, 2021*, pages 3118–3135. Association for Computational Linguistics.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1842–1850, New York, New York, USA. PMLR.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th*

*Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. 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.

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, Online. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Ĥórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinh͡ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 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.

# A Languages

## A.1 MAD-G training languages

Table 5: Details of languages used for MAD-G training.

| code | name | family | genus |
|---|---|---|---|
| ab | Abkhazian | Northwest Caucasian | - |
| ar | Arabic | Afro-Asiatic | Semitic |
| ary | Moroccan Arabic | Afro-Asiatic | Semitic |
| arz | Egyptian Arabic | Afro-Asiatic | Semitic |
| atj | Atikamekw | Algic | Algonquian |
| av | Avar | Nakh-Daghestanian | Avar-Andic-Tsezic |
| ay | Aymara | Aymaran | - |
| azb | South Azerbaijani | Turkic | Southwestern |
| bo | Tibetan | Sino-Tibetan | Bodic |
| bxr | Buryat | Mongolic | - |
| cdo | Min Dong | Sino-Tibetan | - |
| ce | Chechen | Nakh-Daghestanian | Nakh |
| ceb | Cebuano | Austronesian | Greater Central Philippine |
| cv | Chuvash | Turkic | Oghur |
| cy | Welsh | IE | Celtic |
| el | Greek | IE | Greek |
| en | English | IE | Germanic |
| et | Estonian | Uralic | Finnic |
| eu | Basque | Basque | - |
| fa | Persian | IE | Iranian |
| fi | Finnish | Uralic | Finnic |
| fr | French | IE | Romance |
| gn | Guarani | Tupian | Tupi-Guarani |
| ha | Hausa | Afro-Asiatic | West Chadic |
| hak | Hakka | Sino-Tibetan | - |
| he | Hebrew | Afro-Asiatic | Semitic |
| hu | Hungarian | Uralic | Ugric |
| hy | Armenian | IE | Armenian |
| id | Indonesian | Austronesian | Malayo-Sumbawan |
| ig | Igbo | Niger-Congo | Igboid |
| inh | Ingush | Nakh-Daghestanian | Nakh |
| ja | Japanese | Japanese | - |
| jv | Javanese | Austronesian | Javanese |
| ka | Georgian | Kartvelian | - |
| kab | Kabyle | Afro-Asiatic | Berber |
| kbd | Karbardian Circassian | Northwest Caucasian | - |
| kbp | Kabiye | Niger-Congo | Southern-Central Gur |
| kk | Kazakh | Turkic | Northwestern |
| km | Khmer | Austro-Asiatic | Khmer |
| kn | Kannada | Dravidian | Southern |
| ko | Korean | Korean | - |
| kv | Komi | Uralic | Permic |
| la | Latin | IE | Latin |
| lbe | Lak | Nakh-Daghestanian | Lak-Dargwa |
| lez | Lezgian | Nakh-Daghestanian | Lezgic |
| ln | Lingala | Niger-Congo | Bantoid |
| lo | Lao | Tai-Kadai | - |
| mg | Malagasy | Austronesian | Barito |
| mhr | Meadow Mari | Uralic | Mari |
| min | Minangkabau | Austronesian | Malayo-Sumbawan |
| ml | Malayalam | Dravidian | Southern |
| mn | Mongolian | Mongolic | - |
| mrj | Hill Mari | Uralic | Mari |
| ms | Malay | Austronesian | Malayo-Sumbawan |
| mt | Maltese | Afro-Asiatic | Semitic |
| my | Burmese | Sino-Tibetan | Burmese-Lolo |
| myv | Erzya | Uralic | Mordvin |
| nah | Nahuatl | Uto-Aztecan | Aztecan |
| new | Newar | Sino-Tibetan | Mahakiranti |
| nso | Northern Sotho | Niger-Congo | Bantoid |
| nv | Navajo | Na-Dene | Athapaskan |
| om | Oromo | Afro-Asiatic | Lowland East Cushitic |
| qu | Quechua | Quechuan | - |
| ru | Russian | IE | Slavic |
| rw | Kinyarwanda | Niger-Congo | Bantoid |
| sah | Sakha | Turkic | Northeastern |
| sat | Santali | Austro-Asiatic | Munda |
| se | Northern Sami | Uralic | Sami |
| shn | Shan | Tai-Kadai | - |
| smn | Inari Sami | Uralic | Sami |
| sn | Shona | Niger-Congo | Bantoid |
| so | Somali | Afro-Asiatic | Lowland East Cushitic |
| sq | Albanian | IE | Albanian |
| su | Sundanese | Austronesian | Malayo-Sumbawan |
| sv | Swedish | IE | Germanic |
| sw | Swahili | Niger-Congo | Bantoid |
| ta | Tamil | Dravidian | Southern |
| tcy | Tulu | Dravidian | Southern |
| te | Telugu | Dravidian | South Central |
| th | Thai | Tai-Kadai | - |
| tl | Tagalog | Austronesian | Greater Central Philippine |
| tr | Turkish | Turkic | Southwestern |
| tt | Tatar | Turkic | Northwestern |
| tyv | Tuvan | Turkic | Northeastern |
| ug | Uyghur | Turkic | Southeastern |
| uz | Uzbek | Turkic | Southeastern |
| vi | Vietnamese | Austro-Asiatic | Viet-Muong |
| war | Waray-Waray | Austronesian | Greater Central Philippine |
| wuu | Wu | Sino-Tibetan | - |
| xal | Kalmyk | Mongolic | - |
| xmf | Mingrelian | Kartvelian | - |
| yo | Yoruba | Niger-Congo | Defoid |
| za | Zhuang | Tai-Kadai | - |
| zh | Chinese | Sino-Tibetan | - |
| zu | Zulu | Niger-Congo | Bantoid |

## A.2 Universal Dependencies Evaluation Languages

Table 6: Details of languages used for POS tagging and dependency parsing evaluation. `unseen` languages have their language sub-group (`mBERT-genus`, `MAD-G-genus` or `unseen-genus`) specified.

| **code** | **name** | **group** | **treebank** | **family** | **genus** |
|---|---|---|---|---|---|
| af | Afrikaans | mBERT-seen | UD_Afrikaans-AfriBooms | IE | Germanic |
| ajp | South Levantine Arabic | mBERT-genus | UD_South_Levantine_Arabic-MADAR | Afro-Asiatic | Semitic |
| akk | Akkadian | mBERT-genus | UD_Akkadian-RIAO | Afro-Asiatic | Semitic |
| apu | Apurina | unseen-genus | UD_Apurina-UFPA | Arawakan | - |
| aqz | Akuntsu | unseen-genus | UD_Akuntsu-TuDeT | Tupian | Tupari |
| ar | Arabic | mBERT-seen | UD_Arabic-PUD | Afro-Asiatic | Semitic |
| bam | Bambara | unseen-genus | UD_Bambara-CRB | Mande | - |
| be | Belarusian | mBERT-seen | UD_Belarusian-HSE | IE | Slavic |
| bg | Bulgarian | mBERT-seen | UD_Bulgarian-BTB | IE | Slavic |
| bho | Bhojpuri | mBERT-genus | UD_Bhojpuri-BHTB | IE | Indic |
| br | Breton | mBERT-seen | UD_Breton-KEB | IE | Celtic |
| bxr | Buryat | MAD-G-seen | UD_Buryat-BDT | Mongolic | - |
| ca | Catalan | mBERT-seen | UD_Catalan-AnCora | IE | Romance |
| ckt | Chukchi | unseen-genus | UD_Chukchi-HSE | Chukotko-Kamchatkan | - |
| cs | Czech | mBERT-seen | UD_Czech-PDT | IE | Slavic |
| cu | Old Church Slavonic | mBERT-genus | UD_Old_Church_Slavonic-PROIEL | IE | Slavic |
| cy | Welsh | mBERT-seen | UD_Welsh-CCG | IE | Celtic |
| da | Danish | mBERT-seen | UD_Danish-DDT | IE | Germanic |
| de | German | mBERT-seen | UD_German-HDT | IE | Germanic |
| el | Greek | mBERT-seen | UD_Greek-GDT | IE | Greek |
| en | English | mBERT-seen | UD_English-EWT | IE | Germanic |
| es | Spanish | mBERT-seen | UD_Spanish-AnCora | IE | Romance |
| et | Estonian | mBERT-seen | UD_Estonian-EDT | Uralic | Finnic |
| eu | Basque | mBERT-seen | UD_Basque-BDT | Basque | - |
| fa | Persian | mBERT-seen | UD_Persian-PerDT | IE | Iranian |
| fi | Finnish | mBERT-seen | UD_Finnish-TDT | Uralic | Finnic |
| fo | Faroese | mBERT-genus | UD_Faroese-FarPaHC | IE | Germanic |
| fr | French | mBERT-seen | UD_French-GSD | IE | Romance |
| fro | Old French | mBERT-genus | UD_Old_French-SRCMF | IE | Romance |
| ga | Irish | mBERT-seen | UD_Irish-IDT | IE | Celtic |
| gd | Scottish Gaelic | mBERT-genus | UD_Scottish_Gaelic-ARCOSG | IE | Celtic |
| gl | Galician | mBERT-seen | UD_Galician-TreeGal | IE | Romance |
| got | Gothic | mBERT-genus | UD_Gothic-PROIEL | IE | Germanic |
| gsw | Swiss German | mBERT-genus | UD_Swiss_German-UZH | IE | Germanic |
| gun | Mbya Guarani | MAD-G-genus | UD_Mbya_Guarani-Thomas | Tupian | Tupi-Guarani |
| gv | Manx | mBERT-genus | UD_Manx-Cadhan | IE | Celtic |
| he | Hebrew | mBERT-seen | UD_Hebrew-HTB | Afro-Asiatic | Semitic |
| hi | Hindi | mBERT-seen | UD_Hindi-HDTB | IE | Indic |
| hr | Croatian | mBERT-seen | UD_Croatian-SET | IE | Slavic |
| hsb | Upper Sorbian | mBERT-genus | UD_Upper_Sorbian-UFAL | IE | Slavic |
| hu | Hungarian | mBERT-seen | UD_Hungarian-Szeged | Uralic | Ugric |
| hy | Armenian | mBERT-seen | UD_Armenian-ArmTDP | IE | Armenian |
| id | Indonesian | mBERT-seen | UD_Indonesian-PUD | Austronesian | Malayo-Sumbawan |
| is | Icelandic | mBERT-seen | UD_Icelandic-IcePaHC | IE | Germanic |
| it | Italian | mBERT-seen | UD_Italian-ISDT | IE | Romance |
| ja | Japanese | mBERT-seen | UD_Japanese-GSD | Japanese | - |
| kfm | Khunsari | mBERT-genus | UD_Khunsari-AHA | IE | Iranian |
| kk | Kazakh | mBERT-seen | UD_Kazakh-KTB | Turkic | Northwestern |
| kmr | Kurmanji | mBERT-genus | UD_Kurmanji-MG | IE | Iranian |
| ko | Korean | mBERT-seen | UD_Korean-GSD | Korean | - |
| koi | Komi Permyak | MAD-G-genus | UD_Komi_Permyak-UH | Uralic | Permic |
| kpv | Komi Zyrian | MAD-G-seen | UD_Komi_Zyrian-Lattice | Uralic | Permic |
| krl | Karelian | mBERT-genus | UD_Karelian-KKPP | Uralic | Finnic |
| la | Latin | mBERT-seen | UD_Latin-LLCT | IE | Latin |
| lt | Lithuanian | mBERT-seen | UD_Lithuanian-ALKSNIS | IE | Baltic |
| lv | Latvian | mBERT-seen | UD_Latvian-LVTB | IE | Baltic |
| lzh | Classical Chinese | mBERT-genus | UD_Classical_Chinese-Kyoto | Sino-Tibetan | - |
| mdf | Moksha | MAD-G-genus | UD_Moksha-JR | Uralic | Mordvin |
| mr | Marathi | mBERT-seen | UD_Marathi-UFAL | IE | Indic |
| mt | Maltese | MAD-G-seen | UD_Maltese-MUDT | Afro-Asiatic | Semitic |
| myu | Munduruku | unseen-genus | UD_Munduruku-TuDeT | Tupian | Munduruku |
| myv | Erzya | MAD-G-seen | UD_Erzya-JR | Uralic | Mordvin |

| code | name | group | treebank | family | genus |
|---|---|---|---|---|---|
| nl | Dutch | mBERT-seen | UD_Dutch-Alpino | IE | Germanic |
| no | Norwegian | mBERT-seen | UD_Norwegian-Bokmaal | IE | Germanic |
| nyg | Nayini | mBERT-genus | UD_Nayini-AHA | IE | Iranian |
| olo | Livvi | mBERT-genus | UD_Livvi-KKPP | Uralic | Finnic |
| orv | Old East Slavic | mBERT-genus | UD_Old_Russian-RNC | IE | Slavic |
| pcm | Naija | unseen-genus | UD_Naija-NSC | Creole | - |
| pl | Polish | mBERT-seen | UD_Polish-PDB | IE | Slavic |
| pt | Portuguese | mBERT-seen | UD_Portuguese-GSD | IE | Romance |
| ro | Romanian | mBERT-seen | UD_Romanian-RRT | IE | Romance |
| ru | Russian | mBERT-seen | UD_Russian-GSD | IE | Slavic |
| sa | Sanskrit | mBERT-genus | UD_Sanskrit-UFAL | IE | Indic |
| sk | Slovak | mBERT-seen | UD_Slovak-SNK | IE | Slavic |
| sl | Slovenian | mBERT-seen | UD_Slovenian-SSJ | IE | Slavic |
| sme | North Sami | MAD-G-seen | UD_North_Sami-Giella | Uralic | Sami |
| sms | Skolt Sami | MAD-G-genus | UD_Skolt_Sami-Giellagas | Uralic | Sami |
| soj | Soi | mBERT-genus | UD_Soi-AHA | IE | Iranian |
| sq | Albanian | mBERT-seen | UD_Albanian-TSA | IE | Albanian |
| sr | Serbian | mBERT-seen | UD_Serbian-SET | IE | Slavic |
| sv | Swedish | mBERT-seen | UD_Swedish-Talbanken | IE | Germanic |
| ta | Tamil | mBERT-seen | UD_Tamil-TTB | Dravidian | Southern |
| te | Telugu | mBERT-seen | UD_Telugu-MTG | Dravidian | South Central |
| th | Thai | mBERT-seen | UD_Thai-PUD | Tai-Kadai | - |
| tl | Tagalog | mBERT-seen | UD_Tagalog-TRG | Austronesian | Greater Central Philippine |
| tr | Turkish | mBERT-seen | UD_Turkish-GB | Turkic | Southwestern |
| ug | Uyghur | MAD-G-seen | UD_Uyghur-UDT | Turkic | Southeastern |
| uk | Ukrainian | mBERT-seen | UD_Ukrainian-IU | IE | Slavic |
| ur | Urdu | mBERT-seen | UD_Urdu-UDTB | IE | Indic |
| vi | Vietnamese | mBERT-seen | UD_Vietnamese-VTB | Austro-Asiatic | Viet-Muong |
| wbp | Warlpiri | unseen-genus | UD_Warlpiri-UFAL | Pama-Nyungan | - |
| wo | Wolof | unseen-genus | UD_Wolof-WTB | Niger-Congo | Northern Atlantic |
| yo | Yoruba | mBERT-seen | UD_Yoruba-YTB | Niger-Congo | Defoid |
| yue | Cantonese | mBERT-genus | UD_Cantonese-HK | Sino-Tibetan | - |
| zh | Chinese | mBERT-seen | UD_Chinese-GSD | Sino-Tibetan | - |

# B Full Result Tables

## B.1 Single-source Transfer

Table 7: Full per-language results for single-source zero-shot cross-lingual transfer experiments. POS tagging results are given as accuracy scores, dependency parsing results are unlabeled/labeled attachment scores. **G** = **MAD-G**, **LS** = **MAD-G-LS**, **en** = **MAD-G-en**, **TA** = **TA-only**, **X** = **MAD-X**, **mB** = **mBERT-ft**, **R** = **XLM-R-ft**.

| language | | Part-of-speech tagging | | | | | | | Dependency parsing | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **code** | **group** | **G** | **LS** | **en** | **TA** | **X** | **mB** | **R** | **G** | **LS** | **en** | **TA** | **X** | **mB** | **R** |
| af | mBERT-seen | 81.5 | 84.5 | 86.0 | 86.2 | 82.8 | 85.7 | 88.0 | 61.5/47.2 | 67.4/53.9 | 68.1/55.4 | 68.3/55.2 | 61.6/47.2 | 65.6/52.2 | 63.1/49.6 |
| ajp | mBERT-genus | 55.8 | 58.0 | 56.4 | 58.8 | 56.0 | 54.9 | 63.0 | 48.7/27.9 | 48.4/28.6 | 48.9/30.7 | 50.4/33.0 | 50.4/31.1 | 46.0/28.6 | 26.0/13.2 |
| akk | mBERT-genus | 41.1 | 38.9 | 36.7 | 33.9 | 30.4 | 33.2 | 30.0 | 26.4/5.3 | 25.9/5.5 | 22.6/4.4 | 23.6/4.3 | 23.9/3.5 | 20.3/3.2 | 19.8/3.4 |
| apu | unseen-genus | 48.2 | 29.2 | 37.7 | 41.9 | 37.8 | 43.6 | 32.7 | 18.7/10.3 | 20.3/6.5 | 19.6/8.9 | 17.7/7.5 | 18.0/4.4 | 16.3/5.8 | 15.8/7.2 |
| aqz | unseen-genus | 32.5 | 25.0 | 27.5 | 27.5 | 21.2 | 33.8 | 16.2 | 32.5/6.2 | 26.2/2.5 | 28.8/11.2 | 26.2/11.2 | 28.8/13.8 | 21.2/7.5 | 30.0/11.2 |
| ar | mBERT-seen | 72.5 | 73.1 | 72.8 | 74.0 | 69.1 | 67.5 | 78.1 | 66.0/49.9 | 64.7/48.4 | 66.1/49.6 | 67.4/49.0 | 65.5/50.0 | 69.0/50.7 | 48.2/34.6 |
| bam | unseen-genus | 38.0 | 36.0 | 36.6 | 37.6 | 30.8 | 33.6 | 25.5 | 26.8/8.2 | 26.7/7.1 | 30.8/10.6 | 30.2/9.5 | 28.9/6.6 | 30.4/9.6 | 21.3/5.8 |
| be | mBERT-seen | 83.7 | 84.7 | 84.9 | 84.6 | 84.5 | 84.8 | 88.1 | 68.8/58.3 | 68.5/58.4 | 70.8/60.7 | 70.5/59.2 | 65.4/54.7 | 72.3/62.5 | 65.1/55.4 |
| bg | mBERT-seen | 86.3 | 86.4 | 86.6 | 86.4 | 86.4 | 86.2 | 88.8 | 81.6/66.5 | 80.9/65.7 | 82.1/67.0 | 82.4/67.0 | 77.2/62.0 | 83.1/68.4 | 66.8/52.7 |
| bho | mBERT-genus | 43.5 | 46.9 | 48.7 | 49.4 | 51.2 | 47.2 | 50.4 | 30.2/17.2 | 30.8/15.3 | 31.3/16.7 | 33.0/17.1 | 22.0/10.2 | 31.2/16.4 | 25.5/14.1 |
| br | mBERT-seen | 65.1 | 66.3 | 69.7 | 71.5 | 61.9 | 65.8 | 58.3 | 63.3/42.5 | 64.7/43.8 | 70.9/52.1 | 71.3/52.6 | 60.1/35.6 | 66.2/47.0 | 44.0/27.2 |
| bxr | MAD-G-seen | 68.6 | 66.3 | 58.3 | 59.6 | 70.5 | 55.9 | 59.5 | 41.4/22.3 | 39.4/19.7 | 39.3/19.4 | 41.6/19.9 | 38.3/23.9 | 41.2/19.4 | 35.9/17.1 |
| ca | mBERT-seen | 86.7 | 86.4 | 86.6 | 86.8 | 87.3 | 87.0 | 88.6 | 75.5/63.4 | 75.1/62.8 | 76.5/64.7 | 76.5/63.9 | 72.3/60.0 | 78.1/66.4 | 74.4/63.1 |
| ckt | unseen-genus | 30.7 | 24.8 | 23.5 | 23.6 | 23.2 | 22.6 | 30.3 | 24.9/12.0 | 20.3/9.1 | 18.5/10.9 | 20.4/10.3 | 21.0/12.4 | 17.6/9.1 | 32.4/17.6 |
| cs | mBERT-seen | 83.6 | 84.3 | 84.4 | 84.8 | 84.3 | 84.9 | 86.8 | 72.3/58.6 | 73.4/60.1 | 74.8/61.7 | 74.7/60.1 | 71.5/58.3 | 75.2/61.9 | 60.3/48.1 |
| cu | mBERT-genus | 34.1 | 33.8 | 35.4 | 37.1 | 34.7 | 30.3 | 45.0 | 31.9/12.9 | 30.4/11.2 | 32.3/13.9 | 32.6/14.3 | 27.3/9.4 | 28.6/12.2 | 31.5/15.6 |
| cy | mBERT-seen | 64.9 | 64.7 | 64.4 | 64.7 | 59.6 | 60.7 | 66.4 | 63.9/45.9 | 64.6/45.8 | 64.8/45.3 | 65.6/45.5 | 57.7/33.1 | 62.3/40.1 | 46.1/33.0 |
| da | mBERT-seen | 88.9 | 89.0 | 89.2 | 89.2 | 86.3 | 88.7 | 90.1 | 74.3/66.0 | 74.6/66.4 | 75.8/67.7 | 76.3/67.9 | 70.9/61.7 | 77.1/68.7 | 66.1/56.9 |
| de | mBERT-seen | 84.8 | 85.8 | 85.7 | 86.1 | 86.5 | 85.7 | 87.6 | 71.2/61.8 | 75.4/66.9 | 76.7/68.3 | 76.8/68.6 | 75.3/66.7 | 77.4/69.1 | 62.3/53.7 |
| el | mBERT-seen | 81.5 | 81.6 | 81.4 | 81.5 | 83.2 | 82.8 | 86.4 | 78.0/65.4 | 77.2/64.9 | 78.1/65.4 | 79.0/64.8 | 74.7/62.4 | 82.9/70.5 | 57.1/47.5 |
| en | mBERT-seen | 96.3 | 96.3 | 96.3 | 96.3 | 96.4 | 96.7 | 97.3 | 89.6/87.0 | 89.4/86.8 | 89.6/87.0 | 89.8/87.0 | 89.7/87.1 | 91.8/89.4 | 59.8/53.3 |
| es | mBERT-seen | 87.1 | 87.7 | 87.9 | 88.1 | 88.2 | 87.5 | 89.0 | 73.6/61.9 | 76.0/64.6 | 76.9/65.9 | 77.3/66.0 | 74.7/63.9 | 77.8/67.2 | 72.6/62.0 |
| et | mBERT-seen | 83.4 | 82.9 | 83.1 | 83.3 | 86.4 | 81.4 | 87.8 | 64.1/46.6 | 62.7/45.2 | 64.8/47.1 | 64.9/46.3 | 65.0/49.0 | 64.0/44.8 | 63.1/45.4 |
| eu | mBERT-seen | 69.8 | 69.0 | 69.0 | 68.9 | 73.4 | 67.4 | 71.1 | 52.6/33.4 | 51.3/31.7 | 52.7/33.4 | 54.0/33.8 | 53.8/35.3 | 51.2/31.6 | 41.8/24.6 |
| fa | mBERT-seen | 73.4 | 73.5 | 68.5 | 69.3 | 69.4 | 66.9 | 76.3 | 47.3/34.8 | 46.8/33.3 | 43.7/31.7 | 44.2/31.6 | 42.9/31.1 | 42.5/29.9 | 31.7/22.0 |
| fi | mBERT-seen | 83.8 | 83.7 | 83.9 | 84.2 | 71.6 | 82.2 | 88.2 | 66.4/50.9 | 65.1/49.6 | 66.5/51.1 | 66.4/50.2 | 51.1/32.7 | 68.0/51.1 | 61.4/45.9 |

| language | | Part-of-speech tagging | | | | | | | Dependency parsing | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **code** | **group** | `G` | `LS` | `en` | `TA` | `X` | `mB` | `R` | `G` | `LS` | `en` | `TA` | `X` | `mB` | `R` |
| fo | `mBERT-genus` | 71.0 | 71.7 | 72.7 | 73.2 | 64.4 | 68.7 | 72.7 | 51.2/36.3 | 50.8/35.6 | 52.4/37.5 | 52.3/38.1 | 43.4/26.1 | 49.6/34.6 | 48.2/33.4 |
| fr | `mBERT-seen` | 87.0 | 87.0 | 87.0 | 87.7 | 88.1 | 88.4 | 89.1 | 79.1/71.0 | 78.8/70.6 | 79.2/71.1 | 79.1/70.6 | 78.5/71.2 | 78.5/71.8 | 73.1/65.6 |
| fro | `mBERT-genus` | 57.9 | 57.1 | 60.0 | 60.4 | 57.0 | 55.0 | 43.9 | 58.3/32.2 | 57.6/31.0 | 62.3/37.3 | 61.9/35.7 | 55.5/30.0 | 57.3/30.8 | 45.2/22.3 |
| ga | `mBERT-seen` | 51.1 | 57.8 | 71.2 | 71.4 | 74.0 | 65.1 | 69.4 | 31.0/15.3 | 44.9/22.4 | 63.1/41.3 | 64.2/42.3 | 61.8/44.1 | 60.6/37.0 | 48.4/32.2 |
| gd | `mBERT-genus` | 44.4 | 44.4 | 47.1 | 47.1 | 51.4 | 41.8 | 58.0 | 38.4/14.1 | 38.6/13.5 | 40.5/16.2 | 40.5/16.5 | 43.3/19.8 | 38.0/14.8 | 44.6/24.7 |
| gl | `mBERT-seen` | 85.9 | 86.5 | 86.6 | 86.8 | 84.5 | 86.3 | 87.6 | 77.5/67.1 | 77.9/67.7 | 78.9/69.1 | 78.9/68.4 | 75.1/62.8 | 79.6/69.8 | 69.5/60.4 |
| got | `mBERT-genus` | 23.6 | 24.7 | 22.1 | 22.0 | 21.2 | 18.7 | 11.6 | 27.3/8.7 | 28.3/5.8 | 28.3/8.5 | 28.9/9.0 | 27.5/6.1 | 26.6/7.0 | 22.4/6.2 |
| gsw | `mBERT-genus` | 52.0 | 56.9 | 60.9 | 63.7 | 60.2 | 52.6 | 43.8 | 45.4/29.7 | 52.8/33.7 | 56.7/39.0 | 60.2/42.2 | 54.9/36.7 | 46.6/29.0 | 31.0/14.5 |
| gun | `MAD-G-genus` | 36.2 | 35.2 | 30.9 | 30.0 | 30.0 | 30.9 | 26.0 | 20.7/6.4 | 20.5/6.2 | 14.6/5.2 | 17.0/6.3 | 12.5/3.2 | 11.5/3.3 | 12.0/4.1 |
| gv | `mBERT-genus` | 32.7 | 31.4 | 33.1 | 36.0 | 35.1 | 32.4 | 26.9 | 32.8/8.4 | 31.3/6.3 | 31.0/7.4 | 30.8/7.2 | 37.7/11.9 | 28.7/6.1 | 22.6/4.0 |
| he | `mBERT-seen` | 79.3 | 78.8 | 79.3 | 79.7 | 77.7 | 77.1 | 81.9 | 66.3/48.8 | 65.8/48.4 | 66.3/48.7 | 68.0/50.0 | 61.6/42.9 | 68.3/51.7 | 52.5/38.2 |
| hi | `mBERT-seen` | 40.8 | 67.4 | 68.1 | 68.2 | 70.1 | 67.0 | 69.9 | 16.1/6.9 | 39.3/25.6 | 42.4/29.5 | 44.0/30.6 | 29.0/17.5 | 46.0/31.7 | 35.0/22.4 |
| hr | `mBERT-seen` | 84.6 | 83.9 | 84.4 | 84.3 | 83.9 | 84.7 | 86.7 | 76.3/63.4 | 75.4/63.1 | 77.4/65.0 | 76.9/62.7 | 74.5/61.4 | 79.4/67.1 | 69.9/58.1 |
| hsb | `mBERT-genus` | 69.1 | 70.8 | 71.8 | 72.2 | 69.2 | 69.9 | 71.9 | 46.4/33.2 | 49.8/35.4 | 53.3/39.3 | 53.2/38.5 | 50.3/35.5 | 51.4/37.6 | 44.0/29.4 |
| hu | `mBERT-seen` | 81.4 | 81.5 | 81.5 | 82.1 | 82.3 | 81.8 | 85.1 | 71.0/51.6 | 70.3/50.4 | 70.9/51.4 | 71.1/50.8 | 68.3/49.1 | 73.0/51.9 | 62.7/44.7 |
| hy | `mBERT-seen` | 77.1 | 77.1 | 76.9 | 77.4 | 79.3 | 75.1 | 86.0 | 55.7/36.5 | 55.3/35.5 | 56.3/36.8 | 58.2/37.4 | 54.9/35.9 | 58.2/37.5 | 56.1/37.1 |
| id | `mBERT-seen` | 85.9 | 85.8 | 85.7 | 86.2 | 87.2 | 84.3 | 87.2 | 70.1/59.0 | 68.0/57.4 | 69.6/58.9 | 70.9/59.3 | 67.9/58.1 | 66.8/56.9 | 55.4/45.2 |
| is | `mBERT-seen` | 76.0 | 77.3 | 78.4 | 78.8 | 77.6 | 76.0 | 84.3 | 53.4/36.6 | 55.1/38.5 | 56.8/40.5 | 57.2/40.5 | 56.7/39.9 | 57.5/40.7 | 54.3/39.7 |
| it | `mBERT-seen` | 90.8 | 90.9 | 91.5 | 91.8 | 90.9 | 90.3 | 91.9 | 81.5/73.3 | 81.4/72.8 | 82.9/75.5 | 83.2/75.1 | 77.8/69.2 | 84.4/77.5 | 72.9/64.5 |
| ja | `mBERT-seen` | 49.2 | 49.1 | 49.1 | 49.9 | 52.5 | 47.6 | 33.6 | 33.7/18.5 | 33.8/18.3 | 34.1/18.8 | 32.9/19.0 | 35.2/19.4 | 32.5/17.0 | 33.4/16.4 |
| kfm | `mBERT-genus` | 33.8 | 36.5 | 35.1 | 37.8 | 39.2 | 43.2 | 41.9 | 21.6/4.1 | 17.6/5.4 | 23.0/12.2 | 25.7/6.8 | 18.9/5.4 | 21.6/4.1 | 27.0/13.5 |
| kk | `mBERT-seen` | 77.4 | 77.2 | 76.9 | 76.8 | 70.9 | 75.9 | 81.1 | 59.3/40.0 | 58.4/38.3 | 59.2/40.0 | 60.4/40.8 | 48.4/27.2 | 59.5/37.4 | 43.2/25.9 |
| kmr | `mBERT-genus` | 37.6 | 38.4 | 42.0 | 42.0 | 46.9 | 38.3 | 70.0 | 23.7/6.5 | 25.3/5.7 | 26.8/7.6 | 27.9/8.5 | 25.2/8.8 | 24.5/7.3 | 40.5/25.2 |
| ko | `mBERT-seen` | 64.6 | 64.4 | 64.3 | 64.2 | 64.1 | 63.7 | 67.5 | 41.0/27.5 | 40.1/25.9 | 41.0/27.5 | 43.9/29.4 | 42.3/28.1 | 38.9/24.7 | 30.8/20.4 |
| koi | `MAD-G-genus` | 44.2 | 43.9 | 41.1 | 41.4 | 40.3 | 41.8 | 48.2 | 33.1/17.5 | 26.9/14.9 | 28.2/12.6 | 32.7/15.9 | 27.1/11.0 | 26.9/9.5 | 28.2/13.5 |
| kpv | `MAD-G-seen` | 54.8 | 55.2 | 34.0 | 33.4 | 56.3 | 34.5 | 40.8 | 39.3/19.1 | 38.1/18.3 | 23.6/8.6 | 24.5/8.9 | 42.1/21.5 | 22.8/7.4 | 26.0/10.7 |
| krl | `mBERT-genus` | 65.0 | 66.6 | 66.6 | 67.7 | 53.9 | 62.4 | 68.0 | 48.2/25.4 | 46.0/23.9 | 47.9/27.5 | 45.8/25.4 | 37.4/15.5 | 44.7/23.6 | 40.4/21.8 |
| la | `mBERT-seen` | 73.0 | 71.8 | 70.7 | 69.9 | 76.6 | 62.6 | 76.0 | 47.5/30.6 | 46.6/29.8 | 43.9/28.3 | 45.8/28.8 | 52.1/34.1 | 41.0/24.1 | 47.6/29.4 |
| lt | `mBERT-seen` | 75.1 | 77.3 | 80.7 | 81.1 | 78.9 | 78.1 | 85.8 | 56.3/37.3 | 59.6/40.4 | 64.3/45.9 | 63.8/45.2 | 59.6/40.7 | 62.9/43.4 | 56.2/39.4 |
| lv | `mBERT-seen` | 77.9 | 79.0 | 80.6 | 80.9 | 83.6 | 78.8 | 85.4 | 61.8/42.5 | 65.4/46.1 | 67.7/48.9 | 68.3/48.5 | 65.8/47.5 | 66.2/45.8 | 55.4/38.5 |
| lzh | `mBERT-genus` | 50.0 | 50.4 | 50.3 | 49.7 | 48.7 | 49.0 | 27.7 | 46.7/27.4 | 47.6/27.2 | 48.7/29.8 | 48.0/28.3 | 45.6/27.6 | 49.3/30.2 | 25.4/9.9 |
| mdf | `MAD-G-genus` | 47.2 | 48.5 | 46.7 | 48.9 | 46.4 | 47.1 | 46.2 | 34.0/17.6 | 34.9/17.8 | 32.2/17.4 | 34.2/17.6 | 31.8/13.7 | 33.9/14.3 | 28.2/12.6 |
| mr | `mBERT-seen` | 71.8 | 73.0 | 74.2 | 72.4 | 60.7 | 70.6 | 80.4 | 48.8/28.4 | 48.1/26.7 | 48.1/28.2 | 46.8/27.7 | 25.2/14.8 | 44.2/26.0 | 40.0/23.8 |
| mt | `MAD-G-seen` | 71.7 | 72.1 | 27.4 | 26.3 | 75.6 | 24.6 | 24.6 | 61.8/43.0 | 61.3/43.1 | 29.3/6.9 | 32.7/7.6 | 65.4/49.3 | 28.5/5.6 | 20.7/3.9 |
| myu | `unseen-genus` | 21.4 | 15.5 | 17.3 | 19.9 | 18.8 | 25.1 | 17.3 | 24.0/10.3 | 26.9/9.2 | 26.6/14.4 | 21.8/12.2 | 19.9/11.4 | 28.4/16.6 | 31.7/19.6 |
| myv | `MAD-G-seen` | 71.0 | 68.7 | 46.7 | 49.0 | 76.9 | 49.5 | 49.0 | 53.2/33.3 | 51.5/31.9 | 32.5/15.5 | 33.6/15.4 | 59.3/40.5 | 34.3/13.7 | 26.4/11.4 |
| nl | `mBERT-seen` | 87.7 | 88.3 | 88.8 | 89.0 | 89.0 | 88.4 | 89.1 | 74.1/64.6 | 77.4/69.4 | 78.4/70.9 | 78.5/70.9 | 77.4/69.9 | 77.7/70.5 | 63.8/55.9 |
| no | `mBERT-seen` | 89.9 | 90.4 | 90.7 | 90.9 | 90.9 | 90.5 | 92.1 | 79.6/73.4 | 79.9/73.7 | 80.8/74.9 | 81.0/74.9 | 81.3/75.1 | 82.3/75.8 | 65.7/57.5 |
| nyg | `mBERT-genus` | 33.3 | 29.5 | 39.7 | 37.2 | 29.5 | 38.5 | 41.0 | 29.5/11.5 | 24.4/9.0 | 25.6/11.5 | 25.6/10.3 | 24.4/14.1 | 26.9/10.3 | 41.0/17.9 |
| olo | `mBERT-genus` | 64.9 | 64.4 | 64.7 | 64.7 | 56.5 | 59.6 | 59.8 | 46.0/24.0 | 45.6/22.9 | 44.0/22.4 | 46.0/24.3 | 36.7/16.7 | 43.1/20.0 | 31.8/14.0 |
| orv | `mBERT-genus` | 80.9 | 80.8 | 80.6 | 80.3 | 80.8 | 78.8 | 84.6 | 57.3/41.4 | 57.1/41.3 | 57.8/42.0 | 57.5/40.9 | 54.4/38.8 | 57.6/41.7 | 55.0/41.0 |
| pcm | `unseen-genus` | 45.5 | 45.5 | 45.7 | 46.4 | 43.5 | 44.3 | 45.2 | 49.1/26.7 | 49.3/26.3 | 49.7/27.2 | 52.3/27.5 | 46.4/23.9 | 50.1/27.5 | 31.8/14.5 |
| pl | `mBERT-seen` | 76.1 | 80.9 | 83.4 | 83.2 | 83.0 | 81.3 | 84.9 | 62.1/46.3 | 69.4/54.4 | 76.4/62.1 | 75.9/61.0 | 71.6/57.0 | 76.7/62.5 | 63.1/51.1 |
| pt | `mBERT-seen` | 88.4 | 88.6 | 88.8 | 89.1 | 88.1 | 88.5 | 90.1 | 73.0/61.8 | 74.4/63.2 | 75.4/64.7 | 75.8/64.8 | 72.5/61.1 | 75.5/64.9 | 69.7/59.0 |
| ro | `mBERT-seen` | 81.7 | 82.8 | 83.5 | 83.5 | 81.0 | 82.6 | 86.3 | 70.8/56.0 | 71.5/56.0 | 74.5/59.7 | 75.2/59.4 | 67.5/51.7 | 75.9/60.7 | 68.1/54.5 |
| ru | `mBERT-seen` | 83.3 | 83.6 | 83.4 | 83.6 | 84.3 | 83.2 | 87.1 | 74.5/63.6 | 73.8/62.7 | 74.5/63.4 | 75.2/63.0 | 71.3/60.6 | 77.5/65.9 | 62.2/51.3 |
| sa | `mBERT-genus` | 36.4 | 41.5 | 44.2 | 43.1 | 43.4 | 41.7 | 59.0 | 25.9/12.2 | 32.9/9.7 | 34.7/12.5 | 37.9/14.4 | 25.0/7.4 | 30.1/9.9 | 34.3/15.4 |
| sk | `mBERT-seen` | 84.0 | 85.0 | 84.6 | 85.0 | 83.9 | 83.9 | 86.4 | 79.0/66.4 | 78.9/66.1 | 80.4/68.0 | 80.3/66.8 | 76.1/63.8 | 82.1/70.2 | 64.4/51.7 |
| sl | `mBERT-seen` | 81.2 | 82.7 | 83.1 | 83.1 | 77.3 | 82.8 | 85.6 | 75.3/61.2 | 75.9/62.1 | 78.0/64.9 | 78.5/63.9 | 65.7/49.5 | 78.3/65.2 | 70.2/57.6 |
| sme | `MAD-G-seen` | 71.1 | 68.5 | 41.6 | 42.1 | 75.8 | 39.0 | 33.3 | 48.6/32.7 | 46.2/29.3 | 24.3/9.0 | 23.9/8.6 | 50.4/33.5 | 22.9/6.5 | 20.6/7.0 |
| sms | `MAD-G-genus` | 34.6 | 35.7 | 31.2 | 31.3 | 36.6 | 29.6 | 36.2 | 25.7/11.5 | 22.3/8.9 | 22.0/8.9 | 23.7/8.0 | 23.7/8.4 | 21.5/7.4 | 29.7/10.7 |
| soj | `mBERT-genus` | 41.8 | 45.5 | 43.6 | 41.8 | 43.6 | 43.6 | 43.6 | 21.8/7.3 | 27.3/9.1 | 20.0/5.5 | 20.0/5.5 | 34.5/12.7 | 21.8/12.7 | 40.0/12.7 |
| sq | `mBERT-seen` | 77.8 | 78.9 | 78.6 | 78.3 | 71.6 | 74.7 | 81.1 | 84.8/66.3 | 82.6/64.4 | 83.6/64.8 | 86.9/66.2 | 71.8/50.4 | 86.2/68.5 | 65.3/47.5 |
| sr | `mBERT-seen` | 84.9 | 84.5 | 84.7 | 84.1 | 84.5 | 85.2 | 86.9 | 77.8/66.1 | 76.4/64.8 | 78.1/67.0 | 78.1/64.7 | 75.8/63.4 | 80.7/68.7 | 71.3/60.0 |
| sv | `mBERT-seen` | 90.3 | 90.6 | 90.3 | 90.6 | 90.4 | 90.2 | 92.6 | 80.8/74.6 | 80.4/74.0 | 80.9/74.6 | 81.1/74.7 | 81.3/74.9 | 82.8/76.3 | 70.9/63.0 |
| ta | `mBERT-seen` | 65.4 | 64.5 | 65.5 | 64.7 | 54.1 | 64.9 | 67.9 | 37.9/18.4 | 38.3/17.8 | 38.2/18.4 | 41.1/20.2 | 16.9/5.1 | 43.2/17.5 | 43.3/21.4 |
| te | `mBERT-seen` | 75.6 | 75.7 | 76.0 | 75.7 | 67.0 | 76.0 | 85.4 | 70.3/51.6 | 64.1/46.6 | 70.9/53.8 | 73.0/53.4 | 43.0/29.8 | 59.5/42.4 | 53.3/34.5 |
| th | `mBERT-seen` | 48.7 | 48.5 | 48.6 | 50.0 | 47.9 | 46.4 | 55.1 | 42.4/21.1 | 43.4/21.4 | 43.7/22.3 | 43.5/22.9 | 41.7/19.2 | 39.9/21.7 | 45.8/32.7 |
| tl | `mBERT-seen` | 70.7 | 69.5 | 68.7 | 69.6 | 62.3 | 64.7 | 71.1 | 81.6/51.0 | 77.7/48.6 | 75.9/51.5 | 75.1/54.1 | 64.6/37.3 | 71.7/42.1 | 44.7/26.0 |
| tr | `mBERT-seen` | 74.6 | 74.3 | 74.4 | 74.6 | 78.8 | 70.7 | 80.9 | 64.7/43.9 | 63.1/40.9 | 64.9/43.9 | 67.2/45.3 | 62.5/42.1 | 60.6/37.5 | 43.9/28.1 |
| ug | `MAD-G-seen` | 58.0 | 60.4 | 35.1 | 34.4 | 57.9 | 28.9 | 73.5 | 33.3/17.4 | 29.7/13.6 | 16.5/6.5 | 21.1/7.9 | 36.0/16.2 | 17.1/7.1 | 50.3/30.2 |
| uk | `mBERT-seen` | 82.2 | 83.1 | 83.4 | 82.7 | 83.8 | 83.5 | 85.8 | 73.0/60.6 | 72.6/60.3 | 73.8/61.6 | 73.1/59.9 | 69.9/57.4 | 75.7/63.0 | 66.7/54.4 |
| ur | `mBERT-seen` | 49.9 | 61.2 | 62.2 | 63.5 | 58.7 | 60.4 | 65.6 | 20.6/10.1 | 35.7/21.4 | 36.7/22.7 | 36.7/22.6 | 21.3/10.7 | 35.2/21.9 | 37.9/24.1 |
| vi | `mBERT-seen` | 63.5 | 63.1 | 63.6 | 62.9 | 63.9 | 60.3 | 63.2 | 55.8/39.0 | 54.9/37.7 | 55.9/38.9 | 55.7/38.6 | 54.9/36.6 | 53.5/37.3 | 30.1/18.7 |
| wbp | `unseen-genus` | 25.8 | 27.4 | 32.8 | 32.2 | 33.1 | 37.9 | 22.6 | 24.2/8.9 | 26.8/10.8 | 32.5/13.7 | 30.9/14.3 | 15.9/4.1 | 47.1/17.2 | 44.6/19.7 |
| wo | `unseen-genus` | 30.3 | 32.2 | 36.8 | 38.0 | 34.1 | 35.2 | 27.1 | 28.1/6.3 | 31.5/6.5 | 31.8/8.7 | 32.7/8.9 | 32.9/8.8 | 28.4/6.3 | 19.9/4.5 |

| language | | Part-of-speech tagging | | | | | | | Dependency parsing | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| code | group | G | LS | en | TA | X | mB | R | G | LS | en | TA | X | mB | R |
| yo | mBERT-seen | 64.2 | 63.3 | 60.3 | 59.4 | 56.3 | 47.7 | 26.6 | 46.4/28.0 | 45.3/26.6 | 40.9/23.5 | 41.7/24.0 | 37.0/19.8 | 37.2/19.0 | 11.9/2.4 |
| yue | mBERT-genus | 62.1 | 62.5 | 61.8 | 63.3 | 62.4 | 63.3 | 53.0 | 45.4/27.5 | 44.8/27.4 | 45.1/27.9 | 46.2/27.9 | 45.4/28.3 | 45.2/28.4 | 32.0/18.8 |
| zh | mBERT-seen | 70.9 | 70.9 | 70.6 | 68.9 | 69.8 | 67.4 | 48.3 | 56.9/35.4 | 56.3/34.7 | 57.1/35.5 | 56.9/35.5 | 55.8/34.9 | 59.4/38.0 | 47.9/26.5 |

## B.2 Multi-source Transfer

Table 8: Full per-language results for multi-source zero-shot cross-lingual transfer experiments with 20 languages. POS tagging results are given as accuracy scores, dependency parsing results are unlabeled/labeled attachment scores. **G** = **MAD-G**, **LS** = **MAD-G-LS**, **en** = **MAD-G-en**, **TA** = **TA-only**, **X** = **MAD-X**, **mB** = **mBERT-ft**, **R** = **XLM-R-ft**.

| language | | Part-of-speech tagging | | | | | | | Dependency parsing | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| code | group | G | LS | en | TA | X | mB | R | G | LS | en | TA | X | mB | R |
| af | mBERT-seen | 85.0 | 86.9 | 87.4 | 88.2 | 83.2 | 88.9 | 89.6 | 66.8/54.1 | 68.8/55.9 | 69.4/57.3 | 69.2/57.3 | 66.0/53.2 | 71.8/59.4 | 67.8/55.1 |
| ajp | mBERT-genus | 63.9 | 66.2 | 66.3 | 64.1 | 65.0 | 64.4 | 73.5 | 58.3/41.7 | 53.6/34.9 | 55.6/39.2 | 54.5/36.4 | 55.8/38.3 | 56.7/39.0 | 34.6/21.9 |
| akk | mBERT-genus | 41.8 | 41.2 | 37.9 | 42.7 | 2.9 | 46.1 | 46.4 | 30.8/8.1 | 32.0/9.5 | 29.8/7.1 | 30.2/8.1 | 28.8/6.1 | 31.6/10.1 | 26.5/9.8 |
| apu | unseen-genus | 37.1 | 44.5 | 41.7 | 45.4 | 34.5 | 45.5 | 50.4 | 21.0/17.2 | 27.1/12.1 | 23.8/14.4 | 24.9/13.5 | 19.8/10.2 | 24.5/9.1 | 26.3/11.0 |
| aqz | unseen-genus | 30.0 | 27.5 | 20.0 | 30.0 | 22.5 | 22.5 | 32.5 | 35.0/15.0 | 27.5/10.0 | 23.8/10.0 | 25.0/5.0 | 33.8/12.5 | 30.0/8.8 | 27.5/16.2 |
| ar | mBERT-seen | 80.1 | 79.9 | 80.2 | 80.1 | 80.1 | 80.3 | 80.6 | 76.2/66.1 | 76.4/66.4 | 76.7/66.7 | 76.7/66.5 | 76.4/66.7 | 76.8/66.7 | 55.3/46.2 |
| bam | unseen-genus | 31.6 | 31.8 | 33.0 | 33.3 | 29.4 | 29.8 | 30.5 | 31.7/8.3 | 31.8/7.2 | 32.8/8.8 | 32.0/8.2 | 28.2/6.6 | 30.1/7.3 | 23.7/5.5 |
| be | mBERT-seen | 88.8 | 89.2 | 89.4 | 89.4 | 88.7 | 90.8 | 92.1 | 78.2/71.3 | 78.9/72.2 | 79.4/72.5 | 78.9/72.3 | 79.0/71.7 | 82.5/74.6 | 76.4/67.8 |
| bg | mBERT-seen | 93.5 | 94.1 | 93.9 | 93.6 | 91.3 | 93.2 | 95.3 | 85.2/75.3 | 85.4/75.4 | 85.2/75.3 | 85.9/75.7 | 85.6/75.4 | 87.6/78.5 | 70.9/61.1 |
| bho | mBERT-genus | 59.3 | 61.4 | 61.3 | 61.5 | 61.6 | 61.8 | 63.3 | 44.5/27.5 | 48.9/33.7 | 44.4/28.1 | 48.6/32.7 | 46.9/31.9 | 51.9/35.6 | 32.0/21.1 |
| br | mBERT-seen | 72.0 | 72.1 | 74.9 | 75.2 | 64.8 | 70.2 | 68.8 | 71.7/52.7 | 71.3/53.1 | 75.1/58.8 | 76.0/58.4 | 64.3/43.9 | 73.8/53.9 | 54.1/36.0 |
| bxr | MAD-G-seen | 73.2 | 72.0 | 63.7 | 63.9 | 74.3 | 62.2 | 67.2 | 51.7/32.1 | 52.3/31.5 | 47.0/25.4 | 47.4/26.0 | 54.3/34.0 | 49.4/25.2 | 41.1/22.3 |
| ca | mBERT-seen | 90.1 | 90.0 | 89.7 | 89.4 | 87.6 | 89.9 | 89.9 | 81.0/71.1 | 81.3/71.7 | 81.4/71.2 | 81.6/71.4 | 78.3/68.0 | 85.5/74.7 | 81.2/69.9 |
| ckt | unseen-genus | 34.5 | 33.7 | 25.4 | 28.3 | 32.0 | 26.2 | 34.4 | 25.8/16.5 | 28.8/15.7 | 21.4/12.8 | 28.0/16.3 | 29.3/18.0 | 23.5/11.6 | 33.3/18.1 |
| cs | mBERT-seen | 95.4 | 95.6 | 93.8 | 95.8 | 96.1 | 96.5 | 97.5 | 83.9/79.1 | 84.5/79.8 | 83.7/78.6 | 84.7/80.0 | 85.5/80.9 | 88.4/84.1 | 70.9/64.9 |
| cu | mBERT-genus | 36.1 | 36.0 | 37.3 | 37.8 | 36.3 | 37.3 | 51.2 | 33.7/16.0 | 34.3/15.9 | 33.6/16.6 | 38.7/19.6 | 33.1/16.3 | 34.4/16.0 | 44.2/24.0 |
| cy | mBERT-seen | 68.8 | 69.3 | 68.4 | 70.3 | 66.2 | 69.7 | 73.4 | 69.4/51.9 | 70.6/53.7 | 69.3/51.4 | 69.6/51.1 | 65.8/41.9 | 72.2/50.3 | 57.5/42.4 |
| da | mBERT-seen | 90.3 | 90.1 | 90.5 | 90.8 | 86.8 | 91.2 | 92.9 | 72.7/65.3 | 72.8/65.6 | 73.3/66.1 | 73.4/66.3 | 70.9/62.5 | 77.2/68.6 | 67.4/58.4 |
| de | mBERT-seen | 87.2 | 87.6 | 87.4 | 87.1 | 87.3 | 88.8 | 89.6 | 77.7/71.3 | 81.1/74.7 | 81.3/75.2 | 80.8/74.9 | 81.4/75.2 | 85.2/78.9 | 71.9/63.6 |
| el | mBERT-seen | 96.4 | 96.5 | 96.4 | 96.6 | 97.0 | 97.6 | 98.2 | 89.4/86.3 | 89.6/86.7 | 89.3/86.3 | 89.6/86.7 | 90.3/87.5 | 93.3/90.7 | 63.7/59.4 |
| en | mBERT-seen | 92.2 | 92.3 | 92.2 | 92.4 | 92.3 | 93.5 | 94.7 | 82.6/77.5 | 82.5/77.4 | 82.6/77.5 | 82.4/77.3 | 82.9/78.0 | 87.1/82.5 | 63.5/55.8 |
| es | mBERT-seen | 91.7 | 91.8 | 91.9 | 91.7 | 85.5 | 92.3 | 92.3 | 79.7/71.0 | 81.4/73.2 | 81.6/73.4 | 81.9/73.4 | 82.2/73.0 | 85.4/76.4 | 78.8/69.8 |
| et | mBERT-seen | 91.9 | 91.7 | 91.7 | 91.7 | 93.7 | 92.9 | 95.6 | 76.8/69.4 | 76.7/69.0 | 76.7/69.0 | 76.4/68.6 | 79.4/72.8 | 80.6/73.6 | 74.4/66.9 |
| eu | mBERT-seen | 87.9 | 87.8 | 87.7 | 88.0 | 89.9 | 91.2 | 92.8 | 72.8/65.4 | 72.7/65.2 | 71.9/64.4 | 72.9/65.6 | 75.1/68.5 | 78.0/71.4 | 59.0/50.5 |
| fa | mBERT-seen | 90.2 | 90.6 | 84.0 | 90.1 | 91.4 | 92.6 | 96.0 | 81.0/74.9 | 80.6/74.5 | 65.1/58.7 | 80.0/73.8 | 81.7/75.9 | 85.6/80.0 | 51.8/43.3 |
| fi | mBERT-seen | 87.2 | 87.1 | 87.2 | 86.8 | 74.6 | 86.3 | 91.3 | 74.1/65.0 | 74.2/65.1 | 74.1/64.9 | 74.2/64.6 | 60.3/47.5 | 77.5/68.6 | 64.5/56.0 |
| fo | mBERT-genus | 73.0 | 73.5 | 73.5 | 74.5 | 68.5 | 72.1 | 71.7 | 54.1/39.7 | 53.9/39.6 | 54.9/40.7 | 54.5/40.6 | 47.0/30.9 | 52.2/36.8 | 48.7/34.0 |
| fr | mBERT-seen | 96.5 | 96.4 | 96.5 | 96.3 | 96.8 | 97.2 | 97.7 | 87.3/83.6 | 87.1/83.7 | 87.3/83.6 | 87.4/83.9 | 87.6/83.8 | 91.9/88.8 | 84.3/79.4 |
| fro | mBERT-genus | 63.3 | 64.8 | 66.8 | 66.7 | 62.0 | 66.0 | 64.8 | 60.5/40.4 | 60.6/40.9 | 62.2/43.2 | 61.4/42.1 | 57.8/37.4 | 62.4/42.2 | 50.7/31.6 |
| ga | mBERT-seen | 82.5 | 84.6 | 76.1 | 87.7 | 92.3 | 92.3 | 93.7 | 70.4/57.5 | 73.9/61.5 | 70.8/52.3 | 76.5/65.5 | 79.7/71.2 | 83.3/73.6 | 67.4/59.0 |
| gd | mBERT-genus | 49.5 | 54.0 | 49.0 | 55.7 | 59.9 | 57.6 | 76.6 | 47.2/22.7 | 49.1/25.9 | 48.1/23.7 | 48.4/25.6 | 52.0/30.0 | 49.8/26.3 | 59.9/41.2 |
| gl | mBERT-seen | 91.5 | 91.8 | 91.9 | 91.8 | 87.4 | 91.7 | 92.7 | 80.1/72.8 | 80.8/73.6 | 81.2/74.1 | 80.8/73.7 | 78.6/68.8 | 83.5/76.3 | 73.8/66.0 |
| got | mBERT-genus | 34.4 | 34.6 | 38.0 | 37.7 | 42.1 | 34.6 | 34.8 | 29.0/13.4 | 34.2/13.3 | 32.2/13.2 | 31.2/11.3 | 31.9/12.6 | 34.0/12.7 | 27.5/9.6 |
| gsw | mBERT-genus | 64.5 | 65.7 | 70.0 | 68.2 | 65.4 | 62.6 | 52.7 | 54.5/39.1 | 63.0/44.6 | 63.2/46.3 | 64.2/47.5 | 62.5/46.8 | 56.0/38.2 | 38.9/23.4 |
| gun | MAD-G-genus | 41.4 | 40.7 | 37.8 | 38.4 | 31.5 | 39.8 | 34.5 | 30.4/10.5 | 31.3/10.7 | 26.6/9.0 | 27.2/9.0 | 25.8/7.4 | 29.2/9.2 | 23.8/8.6 |
| gv | mBERT-genus | 42.2 | 42.4 | 42.0 | 45.8 | 46.3 | 45.2 | 45.2 | 40.6/14.9 | 39.8/15.1 | 38.7/13.2 | 39.8/14.5 | 44.2/19.6 | 41.6/15.4 | 35.5/11.1 |
| he | mBERT-seen | 77.3 | 80.3 | 77.7 | 81.1 | 70.5 | 79.0 | 85.5 | 67.1/53.3 | 68.0/54.4 | 66.9/53.3 | 68.3/54.4 | 62.5/46.2 | 73.1/58.6 | 59.9/47.0 |
| hi | mBERT-seen | 86.9 | 89.3 | 81.9 | 89.9 | 91.4 | 92.0 | 94.6 | 74.9/66.3 | 81.0/72.8 | 66.8/53.0 | 81.6/74.2 | 80.4/72.9 | 88.2/80.6 | 42.3/34.2 |
| hr | mBERT-seen | 92.4 | 92.7 | 91.9 | 93.0 | 93.8 | 93.6 | 94.1 | 83.6/75.9 | 83.2/75.8 | 83.6/76.3 | 83.5/76.1 | 84.1/76.2 | 87.3/80.0 | 79.6/71.2 |
| hsb | mBERT-genus | 77.7 | 78.2 | 78.7 | 79.1 | 77.7 | 78.4 | 79.9 | 56.5/47.7 | 57.8/49.4 | 60.1/51.1 | 59.6/51.5 | 59.3/50.6 | 61.3/51.9 | 58.3/48.5 |
| hu | mBERT-seen | 93.8 | 93.8 | 93.8 | 93.8 | 94.1 | 95.9 | 97.0 | 82.6/76.4 | 82.5/76.3 | 82.5/76.2 | 81.7/75.5 | 83.8/77.4 | 88.4/82.4 | 69.4/61.8 |
| hy | mBERT-seen | 90.9 | 90.7 | 90.8 | 91.1 | 92.2 | 93.6 | 95.7 | 77.5/68.6 | 78.0/69.4 | 77.2/68.3 | 76.9/67.8 | 79.4/71.4 | 83.4/75.3 | 73.2/65.1 |
| id | mBERT-seen | 88.8 | 88.7 | 88.8 | 88.8 | 89.1 | 88.5 | 89.3 | 81.8/62.5 | 81.9/62.8 | 81.9/62.8 | 81.8/62.6 | 82.7/63.8 | 82.4/63.4 | 67.7/49.6 |
| is | mBERT-seen | 78.8 | 80.3 | 80.8 | 81.2 | 79.2 | 79.0 | 84.5 | 57.1/41.9 | 58.3/43.4 | 59.3/44.3 | 58.4/43.5 | 58.9/44.0 | 58.3/42.5 | 54.9/40.2 |
| it | mBERT-seen | 94.1 | 94.1 | 94.7 | 94.6 | 92.0 | 94.7 | 94.8 | 83.7/78.4 | 83.4/77.9 | 83.9/78.7 | 84.1/78.8 | 81.5/75.4 | 87.2/82.1 | 77.3/70.3 |
| ja | mBERT-seen | 92.5 | 92.5 | 92.4 | 92.6 | 93.1 | 95.8 | 96.6 | 81.9/77.8 | 82.2/78.0 | 81.7/77.5 | 81.1/77.1 | 82.7/78.3 | 91.0/87.7 | 83.0/78.5 |
| kfm | mBERT-genus | 43.2 | 43.2 | 40.5 | 41.9 | 41.9 | 51.4 | 41.9 | 40.5/20.3 | 37.8/21.6 | 40.5/18.9 | 37.8/18.9 | 29.7/14.9 | 28.4/14.9 | 17.6/9.5 |
| kk | mBERT-seen | 82.6 | 82.7 | 82.4 | 82.7 | 73.7 | 82.9 | 86.6 | 67.5/55.4 | 68.3/55.7 | 67.3/55.1 | 68.1/56.4 | 61.9/47.9 | 70.9/57.2 | 49.9/38.5 |
| kmr | mBERT-genus | 47.7 | 46.2 | 47.1 | 47.1 | 52.1 | 45.4 | 79.9 | 27.3/8.8 | 29.3/9.5 | 29.2/11.3 | 29.9/11.1 | 34.2/14.7 | 28.5/9.9 | 55.9/38.6 |
| ko | mBERT-seen | 87.4 | 87.6 | 87.1 | 87.3 | 88.6 | 93.8 | 95.1 | 74.5/68.5 | 74.7/68.4 | 74.2/68.1 | 74.4/68.2 | 75.6/69.4 | 84.7/79.3 | 58.7/51.5 |
| koi | MAD-G-genus | 48.2 | 48.4 | 45.3 | 47.1 | 44.5 | 47.7 | 52.3 | 36.1/20.6 | 33.8/18.5 | 29.7/14.9 | 37.7/20.7 | 31.7/15.9 | 32.1/16.4 | 34.0/19.2 |

| language | | Part-of-speech tagging | | | | | | | Dependency parsing | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **code** | **group** | **G** | **LS** | **en** | **TA** | **X** | **mB** | **R** | **G** | **LS** | **en** | **TA** | **X** | **mB** | **R** |
| kpv | MAD-G-seen | 57.6 | 56.5 | 37.4 | 38.1 | 61.6 | 36.3 | 43.0 | 45.0/26.6 | 44.9/26.4 | 25.6/10.6 | 28.7/12.5 | 48.5/32.5 | 27.1/10.5 | 29.4/14.6 |
| krl | mBERT-genus | 69.9 | 72.5 | 72.4 | 72.9 | 56.6 | 70.3 | 74.9 | 56.2/37.0 | 57.5/39.7 | 55.5/41.5 | 54.4/40.2 | 44.3/27.5 | 55.6/39.3 | 53.4/38.7 |
| la | mBERT-seen | 95.4 | 94.7 | 93.6 | 94.9 | 96.1 | 97.5 | 98.1 | 74.3/70.1 | 74.5/70.3 | 72.1/67.0 | 74.1/69.5 | 76.6/72.6 | 84.1/80.8 | 79.4/74.5 |
| lt | mBERT-seen | 83.3 | 84.7 | 85.7 | 86.2 | 82.3 | 84.9 | 90.5 | 69.9/56.7 | 72.1/59.8 | 73.1/61.8 | 73.3/60.5 | 72.7/59.8 | 74.4/60.6 | 64.5/52.7 |
| lv | mBERT-seen | 89.0 | 89.4 | 88.1 | 89.8 | 92.3 | 91.8 | 94.6 | 77.0/69.6 | 78.1/70.8 | 77.5/68.6 | 78.5/71.1 | 81.7/75.3 | 82.0/75.4 | 63.4/55.6 |
| lzh | mBERT-genus | 56.1 | 59.8 | 57.1 | 57.5 | 53.3 | 57.8 | 57.4 | 50.8/31.4 | 52.5/33.5 | 52.5/33.0 | 51.7/32.5 | 49.0/29.5 | 52.9/33.2 | 33.1/18.3 |
| mdf | MAD-G-genus | 52.6 | 54.4 | 52.2 | 51.9 | 47.0 | 50.3 | 50.4 | 38.8/21.5 | 38.7/22.9 | 37.5/22.5 | 37.3/22.1 | 35.3/22.3 | 41.7/22.5 | 29.0/16.2 |
| mr | mBERT-seen | 85.9 | 83.4 | 81.6 | 84.0 | 68.7 | 81.0 | 86.5 | 59.5/41.0 | 59.0/44.7 | 57.5/43.2 | 61.9/42.5 | 45.6/27.4 | 59.5/42.2 | 38.6/28.2 |
| mt | MAD-G-seen | 80.2 | 78.8 | 35.4 | 37.1 | 80.4 | 35.7 | 35.9 | 68.6/54.4 | 68.1/54.0 | 37.1/10.8 | 39.0/12.1 | 73.1/60.4 | 37.4/9.8 | 35.9/8.0 |
| myu | unseen-genus | 26.6 | 28.8 | 29.9 | 27.7 | 21.0 | 22.9 | 35.4 | 24.4/8.1 | 29.5/11.8 | 30.3/15.5 | 31.7/12.9 | 25.8/10.0 | 29.2/14.0 | 37.6/19.6 |
| myv | MAD-G-seen | 73.2 | 71.2 | 51.8 | 51.7 | 78.5 | 51.3 | 50.9 | 63.2/46.8 | 62.1/43.7 | 35.9/19.0 | 36.3/19.1 | 67.3/52.0 | 40.0/19.7 | 29.0/15.6 |
| nl | mBERT-seen | 87.9 | 88.4 | 88.8 | 89.0 | 88.3 | 89.2 | 89.3 | 77.8/69.8 | 79.7/72.7 | 81.0/74.3 | 80.1/73.6 | 80.5/73.3 | 84.5/77.4 | 70.2/62.3 |
| no | mBERT-seen | 88.1 | 88.4 | 88.0 | 88.7 | 89.7 | 89.5 | 92.2 | 78.9/72.9 | 80.0/73.9 | 80.4/74.4 | 79.8/73.6 | 80.6/75.3 | 83.9/77.1 | 67.9/58.9 |
| nyg | mBERT-genus | 42.3 | 41.0 | 47.4 | 46.2 | 19.2 | 64.1 | 52.6 | 33.3/20.5 | 30.8/19.2 | 32.1/20.5 | 34.6/20.5 | 24.4/15.4 | 35.9/26.9 | 25.6/16.7 |
| olo | mBERT-genus | 70.2 | 72.8 | 70.5 | 71.2 | 58.9 | 68.6 | 71.8 | 57.7/39.5 | 57.9/41.0 | 54.4/37.5 | 54.2/37.9 | 43.4/27.1 | 57.1/39.5 | 47.7/31.7 |
| orv | mBERT-genus | 87.4 | 87.5 | 87.2 | 87.7 | 87.1 | 87.1 | 91.3 | 65.1/53.0 | 65.0/52.9 | 64.8/53.0 | 65.2/53.0 | 64.9/52.4 | 68.4/56.0 | 66.8/55.1 |
| pcm | unseen-genus | 46.3 | 45.8 | 45.9 | 45.7 | 42.4 | 45.2 | 45.7 | 48.8/25.9 | 49.3/26.5 | 49.8/26.5 | 50.4/26.7 | 45.0/20.6 | 50.3/26.3 | 35.9/16.5 |
| pl | mBERT-seen | 86.3 | 88.3 | 89.2 | 89.8 | 90.2 | 90.7 | 92.5 | 75.5/64.4 | 79.8/68.3 | 83.2/73.5 | 83.5/73.2 | 84.2/73.7 | 87.3/77.0 | 72.2/61.6 |
| pt | mBERT-seen | 90.4 | 90.9 | 90.4 | 90.3 | 88.7 | 90.6 | 91.0 | 79.9/70.7 | 80.8/71.8 | 80.9/71.9 | 81.2/72.0 | 79.8/69.6 | 83.9/74.5 | 76.2/66.5 |
| ro | mBERT-seen | 87.0 | 88.5 | 88.2 | 88.8 | 84.9 | 89.6 | 91.6 | 79.6/67.0 | 80.8/67.8 | 80.8/68.6 | 81.4/69.3 | 77.6/64.2 | 83.5/70.5 | 77.3/64.4 |
| ru | mBERT-seen | 88.6 | 88.9 | 88.7 | 89.2 | 84.8 | 90.3 | 92.6 | 82.6/75.2 | 82.5/74.9 | 82.8/75.3 | 82.4/75.0 | 80.9/73.2 | 87.6/80.3 | 72.4/64.3 |
| sa | mBERT-genus | 49.6 | 48.7 | 50.6 | 49.9 | 45.1 | 44.1 | 63.0 | 28.4/18.3 | 42.4/19.6 | 39.8/22.0 | 43.8/19.6 | 42.3/17.1 | 45.6/19.6 | 30.9/17.5 |
| sk | mBERT-seen | 92.9 | 93.3 | 92.1 | 94.1 | 94.5 | 94.4 | 95.3 | 87.7/82.4 | 88.4/83.6 | 87.7/82.5 | 88.6/84.1 | 88.9/84.5 | 90.8/86.4 | 72.9/66.3 |
| sl | mBERT-seen | 89.1 | 90.1 | 90.1 | 90.6 | 83.0 | 90.6 | 93.1 | 84.1/75.4 | 84.2/75.8 | 84.8/76.5 | 85.5/77.0 | 78.7/66.4 | 87.6/79.1 | 81.1/71.2 |
| sme | MAD-G-seen | 73.8 | 72.8 | 48.1 | 48.5 | 79.2 | 47.9 | 45.7 | 54.8/40.4 | 53.1/38.4 | 28.6/13.3 | 28.5/12.0 | 57.5/45.7 | 28.6/12.5 | 26.3/11.1 |
| sms | MAD-G-genus | 37.4 | 41.8 | 34.9 | 36.3 | 46.8 | 36.5 | 45.8 | 29.4/13.6 | 28.1/12.6 | 24.7/10.8 | 30.7/13.5 | 30.0/16.0 | 26.8/11.1 | 32.8/15.6 |
| soj | mBERT-genus | 52.7 | 45.5 | 47.3 | 47.3 | 52.7 | 56.4 | 43.6 | 27.3/12.7 | 34.5/20.0 | 38.2/23.6 | 30.9/18.2 | 50.9/34.5 | 29.1/18.2 | 18.2/14.5 |
| sq | mBERT-seen | 82.5 | 83.6 | 81.8 | 82.2 | 73.0 | 82.2 | 87.2 | 87.4/72.7 | 86.4/71.6 | 86.9/72.0 | 88.7/74.7 | 77.4/59.1 | 89.9/76.5 | 70.3/52.8 |
| sr | mBERT-seen | 92.9 | 93.3 | 93.0 | 93.6 | 95.1 | 94.9 | 94.1 | 84.5/77.1 | 84.0/76.6 | 84.4/77.5 | 84.2/76.9 | 85.2/77.4 | 87.8/79.7 | 81.1/72.3 |
| sv | mBERT-seen | 91.5 | 91.6 | 91.5 | 91.3 | 91.9 | 92.0 | 95.1 | 79.1/72.5 | 78.6/72.3 | 79.0/72.4 | 78.7/72.2 | 79.5/73.2 | 81.7/75.0 | 71.8/63.6 |
| ta | mBERT-seen | 64.4 | 64.9 | 63.7 | 66.2 | 38.9 | 66.3 | 74.2 | 55.8/39.6 | 57.0/39.1 | 55.9/38.9 | 56.4/38.4 | 36.8/20.0 | 61.6/41.2 | 56.3/39.3 |
| te | mBERT-seen | 80.9 | 81.8 | 81.7 | 82.0 | 59.1 | 81.6 | 86.0 | 82.2/66.9 | 82.8/67.1 | 82.5/67.8 | 83.8/66.2 | 63.7/48.0 | 82.9/67.8 | 59.9/45.2 |
| th | mBERT-seen | 51.4 | 50.4 | 50.6 | 51.4 | 38.0 | 55.5 | 71.9 | 52.6/27.8 | 53.0/26.3 | 53.1/28.3 | 52.7/26.7 | 50.4/26.0 | 56.3/29.1 | 64.6/43.1 |
| tl | mBERT-seen | 73.8 | 73.6 | 74.1 | 74.4 | 67.0 | 67.0 | 76.0 | 81.1/54.1 | 80.7/54.2 | 75.6/51.2 | 78.2/53.7 | 68.8/41.8 | 80.9/54.1 | 47.4/29.7 |
| tr | mBERT-seen | 83.6 | 84.1 | 83.6 | 83.6 | 86.2 | 83.6 | 88.1 | 76.1/64.7 | 76.6/65.0 | 76.1/64.4 | 76.3/64.1 | 77.7/67.4 | 78.5/66.2 | 48.3/37.2 |
| ug | MAD-G-seen | 67.8 | 68.8 | 38.5 | 53.1 | 68.4 | 39.2 | 80.5 | 43.1/27.7 | 42.6/27.4 | 24.5/11.9 | 34.4/20.3 | 48.2/32.9 | 30.7/16.0 | 59.3/44.7 |
| uk | mBERT-seen | 89.8 | 90.6 | 89.9 | 90.9 | 91.9 | 92.2 | 93.2 | 81.2/73.5 | 81.7/74.3 | 81.6/74.0 | 81.5/74.1 | 82.2/74.9 | 86.3/79.2 | 77.3/68.9 |
| ur | mBERT-seen | 74.0 | 80.7 | 76.4 | 83.7 | 77.9 | 83.3 | 89.5 | 41.6/29.9 | 62.7/51.8 | 54.5/40.6 | 65.8/54.3 | 61.1/50.3 | 74.2/62.4 | 52.3/43.1 |
| vi | mBERT-seen | 86.9 | 87.3 | 86.9 | 87.4 | 88.8 | 90.0 | 92.8 | 68.2/58.7 | 68.4/58.9 | 68.1/58.8 | 68.3/58.8 | 68.8/59.5 | 72.7/63.4 | 35.0/26.2 |
| wbp | unseen-genus | 38.2 | 38.2 | 44.3 | 39.2 | 40.1 | 36.9 | 47.1 | 21.3/8.6 | 25.2/10.2 | 15.6/6.4 | 21.3/8.3 | 14.3/6.7 | 21.7/7.6 | 14.3/7.6 |
| wo | unseen-genus | 40.6 | 39.4 | 42.6 | 41.9 | 41.4 | 39.8 | 38.1 | 37.0/11.8 | 39.5/12.5 | 37.2/13.4 | 37.5/12.7 | 36.5/11.1 | 38.5/12.2 | 31.6/9.5 |
| yo | mBERT-seen | 69.3 | 65.4 | 60.4 | 61.2 | 56.2 | 53.9 | 29.2 | 51.9/33.8 | 52.4/32.4 | 48.7/29.5 | 48.1/28.9 | 44.5/24.1 | 45.6/23.5 | 20.6/5.4 |
| yue | mBERT-genus | 73.2 | 73.0 | 72.2 | 69.7 | 72.0 | 74.7 | 81.7 | 47.7/31.4 | 48.1/31.8 | 47.5/31.0 | 47.0/30.3 | 49.0/31.9 | 50.5/33.7 | 42.5/26.3 |
| zh | mBERT-seen | 91.0 | 91.0 | 90.9 | 90.9 | 91.5 | 94.7 | 95.3 | 74.2/68.6 | 74.4/68.6 | 74.1/68.2 | 73.8/68.4 | 74.8/69.1 | 83.6/79.0 | 73.8/67.4 |

### B.3 Fine-tuning MAD-G-Initialized Adapters

Table 9: POS tagging accuracy scores on `unseen` languages when MAD-G-initialised (**`MAD-G-ft`**) or randomly initialised (**`rand-ft`**) language adapters are fine-tuned by MLMing on varying amounts of unlabeled text, specifically 1,000, 3,000, 10,000, 30,000 or 100,000 tokens.

| | 1,000 | | 3,000 | | 10,000 | | 30,000 | | 100,000 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **language** | **MAD-G-ft** | **rand-ft** | **MAD-G-ft** | **rand-ft** | **MAD-G-ft** | **rand-ft** | **MAD-G-ft** | **rand-ft** | **MAD-G-ft** | **rand-ft** |
| bam | 31.9 | 31.7 | 31.8 | 27.9 | 31.4 | 30.8 | 31.7 | 30.8 | 32.7 | 31.8 |
| bho | 63.2 | 62.0 | 65.3 | 62.8 | 67.0 | 66.5 | 68.1 | 68.4 | - | - |
| cu | 36.3 | 40.0 | 41.3 | 37.2 | 42.3 | 41.2 | 44.5 | 43.5 | - | - |
| fo | 75.3 | 75.0 | 79.7 | 78.0 | 81.7 | 81.0 | 84.5 | 83.3 | 86.6 | 86.4 |
| gd | 54.3 | 56.0 | 57.5 | 54.9 | 60.6 | 58.1 | 64.5 | 64.1 | 67.3 | 67.9 |
| got | 32.3 | 33.7 | 34.9 | 36.1 | 33.8 | 33.0 | - | - | - | - |
| gv | 50.4 | 45.6 | 52.0 | 47.2 | 61.3 | 58.7 | 68.8 | 65.8 | 74.1 | 74.2 |
| hsb | 79.5 | 80.1 | 81.3 | 81.9 | 86.0 | 85.8 | 87.9 | 87.6 | 89.7 | 88.8 |
| koi | 53.0 | 51.6 | 56.4 | 52.4 | 59.5 | 54.1 | 60.9 | 56.7 | - | - |
| mdf | 55.8 | 53.3 | 60.9 | 57.9 | 66.1 | 61.2 | - | - | - | - |
| olo | 71.8 | 71.8 | 74.9 | 74.9 | 77.8 | 78.7 | 80.2 | 79.9 | 82.5 | 83.1 |
| sa | 55.9 | 53.1 | 56.1 | 57.7 | 57.9 | 58.9 | 62.6 | 61.3 | 65.3 | 66.8 |
| wo | 43.4 | 47.4 | 45.4 | 48.4 | 55.0 | 56.1 | 62.6 | 60.3 | 69.8 | 69.8 |
| yue | 73.7 | 71.2 | 72.8 | 72.2 | 71.8 | 72.2 | 71.6 | 70.5 | 73.7 | 72.5 |

Table 10: Dependency parsing unlabeled/labeled attachment scores on `unseen` languages when MAD-G-initialized (**`MAD-G-ft`**) or randomly initialized (**`rand-ft`**) language adapters are fine-tuned by MLMing on varying amounts of unlabeled text, specifically 1,000, 3,000, 10,000, 30,000 or 100,000 tokens.

| | 1,000 | | 3,000 | | 10,000 | | 30,000 | | 100,000 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **language** | `MAD-G-ft` | `rand-ft` | `MAD-G-ft` | `rand-ft` | `MAD-G-ft` | `rand-ft` | `MAD-G-ft` | `rand-ft` | `MAD-G-ft` | `rand-ft` |
| bam | 32.1/8.7 | 29.1/7.8 | 31.3/8.2 | 29.0/4.8 | 31.4/8.4 | 29.7/7.8 | 31.1/9.0 | 29.3/7.8 | 31.0/9.3 | 28.9/5.5 |
| bho | 44.8/27.6 | 38.6/24.1 | 43.7/27.4 | 41.1/24.9 | 42.7/27.6 | 42.1/25.3 | 44.4/28.0 | 41.0/23.3 | -/- | -/- |
| cu | 34.0/16.9 | 35.6/18.8 | 34.8/18.2 | 35.5/19.2 | 35.9/18.7 | 35.7/18.6 | 37.8/20.0 | 36.9/19.2 | -/- | -/- |
| fo | 55.9/41.8 | 54.4/39.8 | 58.6/45.1 | 54.6/40.5 | 60.3/47.4 | 58.2/45.2 | 61.9/49.0 | 57.8/45.4 | 62.8/50.9 | 56.7/44.7 |
| gd | 50.5/25.9 | 45.2/22.4 | 51.7/27.4 | 48.8/24.5 | 55.0/31.9 | 52.2/28.2 | 59.8/37.0 | 53.3/29.4 | 61.0/40.8 | 53.7/32.3 |
| got | 29.7/13.2 | 23.8/14.1 | 29.6/13.7 | 27.0/7.4 | 29.5/14.0 | 27.5/6.9 | -/- | -/- | -/- | -/- |
| gv | 42.7/19.8 | 36.8/13.3 | 44.6/22.3 | 38.0/16.5 | 51.4/31.6 | 45.0/25.4 | 53.2/36.7 | 47.1/30.4 | 57.1/41.9 | 50.5/35.0 |
| hsb | 61.4/51.2 | 60.2/49.8 | 66.2/55.5 | 63.6/53.3 | 71.3/61.1 | 64.3/54.4 | 73.8/64.4 | 69.6/60.6 | 75.7/67.2 | 71.3/62.8 |
| koi | 41.7/25.5 | 34.1/19.2 | 40.6/25.0 | 33.6/19.3 | 43.0/28.1 | 37.3/20.4 | 43.5/29.2 | 37.1/24.4 | -/- | -/- |
| mdf | 41.2/25.0 | 33.3/23.2 | 46.4/30.2 | 42.1/26.8 | 50.7/36.1 | 48.2/32.4 | -/- | -/- | -/- | -/- |
| olo | 61.7/43.9 | 56.9/40.9 | 63.4/46.1 | 61.6/43.8 | 66.8/50.9 | 60.1/43.4 | 68.1/54.7 | 65.5/51.4 | 69.8/56.5 | 64.3/50.5 |
| sa | 37.5/20.8 | 40.8/24.4 | 41.9/23.2 | 43.7/24.7 | 42.9/25.0 | 46.6/27.1 | 47.6/29.9 | 48.3/29.1 | 48.0/30.3 | 48.9/31.9 |
| wo | 37.6/12.5 | 34.8/13.3 | 40.4/14.5 | 39.3/16.2 | 44.3/19.1 | 42.8/19.6 | 49.9/24.9 | 51.8/25.4 | 55.0/31.9 | 53.0/29.5 |
| yue | 48.2/31.9 | 43.4/28.0 | 47.9/31.6 | 44.3/28.3 | 47.2/30.9 | 43.8/28.1 | 46.4/31.6 | 44.6/29.2 | 47.2/31.8 | 45.7/30.4 |