Geographic Adaptation of Pretrained Language Models

Valentin Hofmann[1,2,3] **Goran Glavaš**[4] **Nikola Ljubešić**[5,6]
Janet B. Pierrehumbert[2] **Hinrich Schütze**[3]
[1]Allen Institute for AI, US [2]University of Oxford, UK
[3]LMU Munich, Germany [4]CAIDAS, University of Würzburg, Germany
[5]Jožef Stefan Institute, Slovenia [6]University of Ljubljana, Slovenia
valentinh@allenai.org

Abstract

While pretrained language models (PLMs) have been shown to possess a plethora of linguistic knowledge, the existing body of research has largely neglected extralinguistic knowledge, which is generally difficult to obtain by pretraining on text alone. Here, we contribute to closing this gap by examining *geolinguistic* knowledge, i.e., knowledge about geographic variation in language. We introduce *geoadaptation*, an intermediate training step that couples language modeling with geolocation prediction in a multi-task learning setup. We geoadapt four PLMs, covering language groups from three geographic areas, and evaluate them on five different tasks: fine-tuned (i.e., supervised) geolocation prediction, zero-shot (i.e., unsupervised) geolocation prediction, fine-tuned language identification, zero-shot language identification, and zero-shot prediction of dialect features. Geoadaptation is very successful at injecting geolinguistic knowledge into the PLMs: The geoadapted PLMs consistently outperform PLMs adapted using only language modeling (by especially wide margins on zero-shot prediction tasks), and we obtain new state-of-the-art results on two benchmarks for geolocation prediction and language identification. Furthermore, we show that the effectiveness of geoadaptation stems from its ability to *geographically retrofit* the representation space of the PLMs.

1 Introduction

The default tool for the majority of NLP tasks is now *de facto* pretrained language models (PLMs; Devlin et al., 2019; Liu et al., 2019b; Radford et al., 2019; Brown et al., 2020; Clark et al., 2020; Raffel et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022; Touvron et al., 2023, *inter alia*), which are trained using language modeling objectives on large text corpora. Despite the conceptual simplicity of language modeling, pretraining induces complex forms of linguistic knowledge in PLMs, at various levels (Rogers et al., 2020; Mahowald et al., 2023): morphological (Edmiston, 2020; Hofmann et al., 2020; Weissweiler et al., 2023), lexical (Ethayarajh, 2019; Vulić et al., 2020), syntactic (Hewitt and Manning, 2019; Jawahar et al., 2019; Wei et al., 2021; Weissweiler et al., 2022), and semantic (Wiedemann et al., 2019; Ettinger, 2020). This general linguistic knowledge is then (re-)shaped for concrete tasks via fine-tuning, i.e., supervised training on task-specific labeled data.

Humans, however, additionally make use of a rich spectrum of *extralinguistic* features when they learn and process language, including gender (Lass et al., 1979), ethnicity (Trent, 1995), and geography (Clopper and Pisoni, 2004). Despite the growing awareness for the importance of such factors in NLP (Hovy and Yang, 2021), extralinguistic features have been typically introduced in the fine-tuning phase so far, i.e., when specializing PLMs for a concrete task (e.g., Rosin et al., 2022). This prevents PLMs from forming generalizable representations the way humans do, impeding the exploitation of extralinguistic knowledge for tasks other than the fine-tuning task itself.

In this work, we focus on geographic knowledge, and more specifically *geolinguistic* knowledge, i.e., knowledge about geographic variation in language—the most salient type of extralinguistic variation in language (Wieling and Nerbonne, 2015). We present what we believe to be the first attempt to incorporate geolinguistic knowledge into PLMs in a pretraining step, i.e., *before task-specific fine-tuning*, making it possible to exploit it in any task for which it is expected to be useful. Specifically, we conduct an intermediate training step (Glavaš and Vulić, 2021) in the form of task-agnostic *adaptation*—dubbed

Transactions of the Association for Computational Linguistics, vol. 12, pp. 411–431, 2024. https://doi.org/10.1162/tacl_a_00652
Action Editor: Trevor Cohn. Submission batch: 7/2023; Revision batch: 11/2023; Published 4/2024.

geoadaptation—that couples language modeling with predicting the geographic location (i.e., longitude and latitude) on geolocated texts. We choose adaptation as opposed to pretraining from scratch for three reasons: (i) intermediate training on language modeling (i.e., adaptation) before task-specific fine-tuning has proved beneficial for many NLP tasks (Gururangan et al., 2020), (ii) adaptation has a lower computational cost than pretraining (Strubell et al., 2019), and (iii) PLMs encoding general-purpose linguistic knowledge are readily available (Wolf et al., 2020).[1] The specific method we introduce for geoadaptation combines language modeling with token-level geolocation prediction via multi-task learning, with task weights based on the homoscedastic uncertainties of the task losses (Kendall et al., 2018).

We evaluate our geoadaptation framework on three groups of closely related languages, each with a corresponding PLM: (i) the German dialects spoken in Austria, Germany, and Switzerland (AGS) and GermanBERT; (ii) Bosnian-Croatian-Montenegrin-Serbian (BCMS) and BERTić; and (iii) Danish, Norwegian, and Swedish (DNS) and ScandiBERT. These groups exhibit strong geographic differences, providing an ideal testbed for geoadaptation.[2] We further test geoadaptation at scale by adapting mBERT, a multilingual PLM, on the union of AGS, BCMS, and DNS.

We evaluate the effectiveness of geoadaptation on five downstream tasks expected to benefit from geolinguistic knowledge: (i) fine-tuned (i.e., supervised) geolocation prediction, (ii) zero-shot (i.e., unsupervised) geolocation prediction, (iii) fine-tuned language identification, (iv) zero-shot language identification, and (v) zero-shot prediction of dialect features. Geoadaptation leads to consistent performance gains compared to baseline models adapted on the same data using only language modeling, with particularly striking improvements on all zero-shot tasks. On two popular benchmarks for geolocation prediction and language identification, geoadaptation establishes a new state of the art. Furthermore, we show that geoadaptation *geographically retrofits* the representation space of the PLMs. Overall, we see our study as an exciting step towards grounding PLMs in geography.[3]

2 Related Work

Adaptation of PLMs. Continued language modeling training (i.e., adaptation) on data that comes from a similar distribution as the task-specific target data has been shown to improve the performance of PLMs for many NLP tasks (Glavaš et al., 2020; Gururangan et al., 2020) as well as in various language (Pfeiffer et al., 2020; Parović et al., 2022) and domain adaptation scenarios (Chronopoulou et al., 2021; Hung et al., 2022). Adaptation can be seen as a special case of *intermediate training*, which aims at improving the target-task performance of PLMs by carrying out additional training between pretraining and fine-tuning (Phang et al., 2018; Vu et al., 2020; Glavaš and Vulić, 2021). Intermediate training has also been conducted in a multi-task fashion, encompassing two or more training objectives (Liu et al., 2019a; Aghajanyan et al., 2021). Our work differs from these efforts in that it injects geolinguistic knowledge—a type of extralinguistic knowledge—into PLMs.

Extralinguistic Knowledge. Leaving aside the large body of work on injecting visual (e.g., Bugliarello et al., 2022) and structured knowledge (e.g., Lauscher et al., 2020) into PLMs, a few studies have examined the interplay of PLM adaptation and extralinguistic factors (Luu et al., 2021; Röttger and Pierrehumbert, 2021). However, they focus on *time* and adapt PLMs to *individual* extralinguistic contexts (i.e., time points). In contrast, we inject *geographic* information from *all* contexts into the PLM, forcing it to learn links between linguistic variability and a language-external variable—in our case, geography. This is fundamentally different from adapting the PLM only to certain realizations of the language-external variable.

Most other studies introduce the extralinguistic information during task-specific fine-tuning (Dhingra et al., 2021; Hofmann et al., 2021; Karpov and Kartashev, 2021; Kulkarni et al., 2021; Rosin et al., 2022). In contrast, we leverage

[1] Notice that for the language areas we consider, there is currently also not enough geotagged data that would allow us to geographically pretrain models from scratch.

[2] Our focus on AGS, BCMS, and DNS also contributes to the recent call for more work on languages other than English in NLP (Joshi et al., 2020; Razumovskaia et al., 2022).

[3] We make our code available at `https://github.com/valentinhofmann/geoadaptation`.

geographic information only in the task-agnostic adaptation step. In task fine-tuning, the geoadapted PLM does not require any extralinguistic signal and is fine-tuned in the same manner as standard PLMs.

Geography in NLP. We also build upon the long line of NLP research on geography, which roughly falls into two camps. On the one hand, many studies model geographically conditioned differences in language, pointing to *lexical variation* as the most conspicuous manifestation (Eisenstein et al., 2010; Eisenstein et al., 2011; Doyle, 2014; Eisenstein et al., 2014; Huang et al., 2016; Hovy and Purschke, 2018; Hovy et al., 2020), although phonological (Hulden et al., 2011; Blodgett et al., 2016), syntactic (Dunn, 2019; Demszky et al., 2021), and semantic properties (Bamman et al., 2014; Kulkarni et al., 2016) have been shown to exhibit geographic variation as well. On the other hand, there exists a large body of work on predicting geographic location from text, a task referred to as *geolocation prediction* (Rahimi et al., 2015a,b, 2017; Salehi et al., 2017; Rahimi et al., 2018; Scherrer and Ljubešić, 2020, 2021). To the best of our knowledge, we are the first to geographically adapt PLMs in a task-agnostic fashion, making them more effective for any downstream task for which geolinguistic knowledge is relevant, from geolocation prediction to dialect-related tasks and language identification.

3 Geoadaptation

Let $\mathcal{D}$ be a geotagged dataset consisting of sequences of tokens $X = (x_1, \ldots, x_n)$ and corresponding geotags $T = (t_{\text{lon}}, t_{\text{lat}})$, where t_{lon} and t_{lat} denote the geographic longitude and latitude. We want to adapt a PLM in such a way that it encodes the geographically conditioned linguistic variability in $\mathcal{D}$. Acknowledging the prominence of lexical variation among geographic differences in language (see §2), we accomplish this by combining masked language modeling (i.e., the pretraining objective) with token-level geolocation prediction in a multi-task setup that pushes the PLM to learn associations between linguistic phenomena and geolocations *on the lexical level*.[4]

[4]In this work, we focus on PLMs pretrained via masked language modeling. However, geoadaptation can in principle also be applied to autoregressive PLMs.

Masked Language Modeling. We replace some tokens x_i in X with masked tokens $\tilde{x}_i$. Following Devlin et al. (2019), $\tilde{x}_i$ can be a special mask token (`[MASK]`), a random vocabulary token, or the original token itself. X is fed into the PLM, which outputs a sequence of representations $E = (\mathbf{e}(x_1), \ldots, \mathbf{e}(x_n))$. The representations of the masked tokens $\mathbf{e}(\tilde{x}_i)$ are then fed into a classification head. We compute the masked language modeling loss $\mathcal{L}_{\text{mlm}}$ as the negative log-likelihood of the probability assigned to the true token.

Geolocation Prediction. We additionally feed the vectors of masked tokens $\mathbf{e}(\tilde{x}_i)$ into a feed-forward regression head that predicts two real-values: longitude and latitude. The geolocation prediction loss $\mathcal{L}_{\text{geo}}$ is the mean of the absolute prediction errors for longitude and latitude. Note that the gold geolocation is the same for all masked tokens from the same input sequence. We inject geographic information at the token level because lexical variation represents the most prominent type of geographic language variation (see §2).

Composite Multi-task Loss. We experiment with two different ways to compute the composite multi-task loss $\mathcal{L}_{\text{mt}}$. First, we straightforwardly sum the two task-specific losses: $\mathcal{L}_{\text{mt}} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{geo}}$. In multi-task training, however, a simple sum of the losses can be a suboptimal choice, especially if the losses are not of the same order of magnitude. In our case, $\mathcal{L}_{\text{mlm}}$ and $\mathcal{L}_{\text{geo}}$ are measured on different scales and relatively small values of $\mathcal{L}_{\text{geo}}$ may still be multiples of relatively large values of $\mathcal{L}_{\text{mlm}}$ (or vice versa). In a similar vein, the model might be more confident about one task than about the other (e.g., associating contextual token representations with geolocations may be easier than language modeling—i.e., predicting the correct token). To account for both factors, as a second method we compute the weights with which $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{mlm}}$ contribute to the joint loss based on their homoscedastic (i.e., task-dependent) uncertainties σ_{mlm} and σ_{geo} (Kendall and Gal, 2017). σ_{mlm} and σ_{geo} are learned as part of the model training. The dynamic weighting ensures that the objectives are given equal importance with respect to the overall optimization. Defining $l \in \{\text{mlm}, \text{geo}\}$, we follow Kendall et al. (2018) and replace $\mathcal{L}_l$ with:

$$\tilde{\mathcal{L}}_l = \frac{1}{2\sigma_l^2}\mathcal{L}_l + \log \sigma_l. \quad (1)$$

(a) AGS (b) BCMS (c) DNS

Figure 1: Geographic distribution of the data for AGS (left), BCMS (middle), and DNS (right). Each point represents a Jodel post (AGS) or tweet (BCMS, DNS). Point density correlates with population density, with the densest areas corresponding to urban centers. For DNS, we exclude the Svalbard islands, which do not have any points.

		FT-Geoloc				FT-Lang				ZS-Dialect	
Language	Adaptation	Train	Dev	Test	ZS-Geoloc	Train	Dev	Test	ZS-Lang	Phon	Lex
AGS	15,000	343,748	31,538	33,953	1,600	45,000	4,500	4,500	–	–	–
BCMS	80,000	353,953	38,013	4,189	1,400	60,000	6,000	6,000	6,000	640	610
DNS	300,000	150,000	75,000	75,000	3,900	45,000	4,500	4,500	4,500	–	–
EUR	50,000	100,000	10,000	10,000	4,500	100,000	10,000	10,000	–	–	–

Table 1: Data statistics. The table provides the number of Jodel posts (AGS), tweets (BCMS, DNS), or both (EUR) used for (geo-)adaptation and the five evaluation tasks (FT-Geoloc, ZS-Geoloc, FT-Lang, ZS-Lang, ZS-Dialect). There is no overlap between the Jodel posts/tweets used for (geo-)adaptation and the ones used for evaluation. The FT-Geoloc splits for AGS and BCMS are the original VarDial (Chakravarthi et al., 2021) splits.

Equation 1 holds for both regression (e.g., mean absolute error as for $\mathcal{L}_{\text{geo}}$) and classification losses (e.g., categorical cross-entropy as for $\mathcal{L}_{\text{mlm}}$) and can be derived from their Bayesian formulations (Kendall et al., 2018). Notice that $\tilde{\mathcal{L}}_l$ is smoothly differentiable and well-formed: $\log \sigma_l$ ensures that the task weight $1/\sigma_l^2$ does not converge to zero (or σ_l^2 diverges to infinity), which is the trivial solution to minimizing $1/(2\sigma_l^2)\mathcal{L}_l$. For numerical stability, we set $\eta_l = 2 \log \sigma_l$ and compute $\tilde{\mathcal{L}}_l$ as:

$$\tilde{\mathcal{L}}_l = \frac{1}{2}(e^{-\eta_l}\mathcal{L}_l + \eta_l). \quad (2)$$

The final multi-task loss is the sum of the two uncertainty losses: $\tilde{\mathcal{L}}_{\text{mt}} = \tilde{\mathcal{L}}_{\text{mlm}} + \tilde{\mathcal{L}}_{\text{geo}}$.

4 Experimental Setup

Models. We examine four PLMs in this paper. For AGS, we use GermanBERT, a German BERT (Devlin et al., 2019) model.[5] For BCMS, we use BERTić (Ljubešić and Lauc, 2021), a BCMS ELECTRA (Clark et al., 2020) model.[6] We specifically use the generator, i.e., a BERT model. For DNS, we resort to ScandiBERT, an XLM-Roberta (Conneau et al., 2020) model pretrained on corpora from five Scandinavian languages.[7] Since we are interested to see whether geoadaptation can be expanded to a larger geographical area (e.g., an entire continent), we also geoadapt mBERT, a multilingual BERT (Devlin et al., 2019) model, on the union of the AGS, BCMS, and DNS areas.[8] We refer to this setting as EUR.

Data. We start with a general overview of the data used for the experiments. Details about data splits are provided when describing the setup for geoadaptation as well as the evaluation tasks. Figure 1 shows the geographic distribution of the data. Tables 1 and 2 list summary statistics.

[5] https://huggingface.co/dbmdz/bert-base-german-cased.

[6] https://huggingface.co/classla/bcms-bertic.

[7] https://huggingface.co/vesteinn/ScandiBERT.

[8] https://huggingface.co/bert-base-multilingual-cased.

	FT-Lang			
Language	Train	Dev	Test	ZS-Lang
BCMS	7,374	963	921	921
DNS	22,796	5,699	1,497	1,497

Table 2: Out-of-domain data statistics. The table provides the number of news articles (BCMS) and Wikipedia snippets (DNS) used for out-of-domain FT-Lang and ZS-Lang. The FT-Lang splits are the original SETimes (Rupnik et al., 2023) and NordicDSL (Haas and Derczynski, 2021) splits.

For AGS, we use the German data of the 2021 VarDial shared task on geolocation prediction (Chakravarthi et al., 2021), which consist of geotagged Jodel posts from the AGS area. We merge the Austrian/German and Swiss portions of the data. For BCMS, we use the BCMS data of the 2021 VarDial shared task on geolocation prediction (Chakravarthi et al., 2021), which consist of geotagged tweets from the BCMS area. To remedy the sparsity of the data for some regions, we retrieve an additional set of geotagged tweets from the BCMS area posted between 2008 and 2021 using the Twitter API, ensuring that there is no overlap with the VarDial data. For evaluation, we additionally draw upon SETimes, a news dataset for discriminating between Bosnian, Croatian, and Serbian (Rupnik et al., 2023). For DNS, we use geotagged tweets from the Nordic Tweet Stream (Laitinen et al., 2018), confining geotags to the DNS area.[9] For evaluation, we additionally use the DNS portion of NordicDSL, a dataset of Wikipedia snippets for discriminating between Nordic languages (Haas and Derczynski, 2021). For EUR, we mix the AGS, BCMS, and DNS data.

Geoadaptation. For AGS, we create a balanced subset of the VarDial train posts (5,000 per country).[10] For BCMS, we draw upon the union of the VarDial train posts and the newly collected posts to create a balanced subset (20,000 per country). For DNS, we similarly create a balanced subset of the posts (100,000 per country). For EUR, we sample balanced subsets of the AGS, BCMS, and DNS geoadaptation data (5,000 per country). Using these four datasets, we adapt the PLMs via the proposed multi-task learning approach (see §3). We geoadapt the PLMs for 25 epochs and save the model snaphots after each epoch. To track progress, we measure perplexity and token-level median distance on the VarDial development sets for AGS and BCMS, a separate set of 75,000 posts for DNS, and a separate set of 10,000 posts for EUR.

Evaluation Tasks. Inspired by existing NLP research on geography (see §2), we evaluate the geoadapted PLMs on five tasks that probe different aspects of the learned associations between linguistic phenomena and geography.

Fine-tuned Geolocation Prediction (FT-Geoloc). We fine-tune the geoadapted PLMs for geolocation prediction. For AGS and BCMS, we use the train, dev, and test splits from VarDial. For DNS, we create separate sets of train, dev, and test posts; we do the same for EUR, drawing train, dev, and test posts from the union of the AGS, BCMS, and DNS data (see Table 1). We make sure that there is no overlap between the geoadaptation posts and dev and test posts of any of the downstream evaluation tasks. Following prior work by Scherrer and Ljubešić (2021), we cast geolocation prediction as a multi-class classification task: We first map all geolocations in the train sets into k clusters using k-means and assign each geotagged post to its closest cluster.[11] Concretely, we pass the contextualized vector of the `[CLS]` token to a single-layer softmax classifier that outputs probability distributions over the k geographic clusters.

In line with prior work, we use the median of the Euclidean distance between the predicted and true geolocation as the evaluation metric. Note that FT-Geoloc is *different* from geolocation prediction in geoadaptation (see §3): there, we (i) cast geolocation prediction as a regression task (i.e., predict the exact longitude and latitude) and (ii) predict the geolocation from the masked tokens, rather than the representation of the whole post.

Zero-shot Geolocation Prediction (ZS-Geoloc). Given the central objective of geoadaptation (i.e.,

[9] For the sake of simplicity, in the following we will refer to both Jodel posts and tweets as *posts*.

[10] In preliminary experiments, we found that geographically balanced sampling is beneficial for geoadaptation.

[11] We standardize longitude and latitude values and use the Euclidean distance as the clustering metric. Following Scherrer and Ljubešić (2021), we choose $k = 75$.

to induce mappings between linguistic variation and geography), we next test if the geoadapted models can predict geographic information from text without any fine-tuning. To this end, we directly probe the PLMs for geolinguistic associations: With the help of prompts, we ask the PLMs to generate the correct toponym corresponding to a post's geolocation using their language modeling head, which has not been trained on geolocation prediction in any way (see §3). We do this on the most fine-grained geographic resolution possible, i.e., cities for BCMS/DNS and states for AGS.[12] For EUR, we draw upon the union of AGS, BCMS, and DNS, resulting in a mix of cities and states.

To create the data for ZS-Geoloc, we start by reverse-geocoding all posts and then select cities/states that contain at least 100 posts and have names existing in the PLM vocabulary. We randomly sample 100 posts from each of these cities/states (AGS: *Bayern*, *Bern*, *Brandenburg*, *Bremen*, *Hessen*, *Kärnten*, *Luzern*, *Niedersachsen*, *Oberösterreich*, *Saarland*, *Sachsen*, *Salzburg*, *Steiermark*, *Thüringen*, *Tirol*, *Zürich*; BCMS: *Bar*, *Beograd*, *Bor*, *Dubrovnik*, *Kragujevac*, *Niš*, *Podgorica*, *Pula*, *Rijeka*, *Sarajevo*, *Split*, *Tuzla*, *Zagreb*, *Zenica*; DNS: *Aalborg*, *Aarhus*, *Arendal*, *Bergen*, *Drammen*, *Fredrikstad*, *Göteborg*, *Halmstad*, *Haugesund*, *Helsingborg*, *Kalmar*, *Karlstad*, *Kristiansand*, *København*, *Linköping*, *Luleå*, *Lund*, *Moss*, *Nora*, *Norrköping*, *Odense*, *Oslo*, *Porsgrunn*, *Roskilde*, *Sala*, *Sandefjord*, *Sarpsborg*, *Skien*, *Stavanger*, *Stockholm*, *Södertälje*, *Tromsø*, *Trondheim*, *Tønsberg*, *Uddevalla*, *Umeå*, *Uppsala*, *Ålesund*, *Örebro*; EUR: 45 underlined cities/states above, which are in the mBERT vocabulary).

For zero-shot prediction, we append prompts with the meaning 'This is [MASK]' to the post (AGS: *Das ist [MASK]*; BCMS: *To je [MASK]*; DNS: *Dette er [MASK]*).[13] For EUR, we just append *[MASK]* to the post. We pass the whole sequence to the PLM and forward the output representation of the `[MASK]` token into the language modeling head. Following common practice (Xiong et al., 2020), we restrict the output vocabulary to the set of candidate labels, i.e., we select the city or state name with the highest logit. We measure the performance in terms of accuracy.

Fine-tuned Language Identification (FT-Lang). Next, we consider language identification, a task of great importance for many applications that is particularly challenging in the case of closely related languages (Zampieri et al., 2014; Haas and Derczynski, 2021). While arguably less directly tied to geography than geolocation prediction, we believe that language identification should also benefit from geoadaptation since one or (in the case of multilingual communities) few languages are used at any given location—having knowledge about geolinguistic variation should thus make it easier to distinguish different languages.

We start by *fine-tuning* the PLMs for language identification. For AGS, BCMS, and DNS, we reuse the respective FT-Geoloc datasets and sample 15,000 train, 1,500 dev, and 1,500 test posts per language (determined based on their geolocation). For EUR, we reuse the exact FT-Geoloc train, dev, and test split. To test how well the effects of geoadaptation generalize to out-of-domain data, we also fine-tune BERTić on SETimes (i.e., news articles) and ScandiBERT on NordicDSL (i.e., Wikipedia snippets). In terms of modeling, we formulate language identification as a multi-class classification task, with three classes for AGS/DNS, four classes for BCMS, and 10 classes for EUR. We again pass the contextualized vector of the `[CLS]` token to a single-layer softmax classifier that outputs probability distributions over the languages. We measure the performance in terms of accuracy.

Zero-shot Language Identification (ZS-Lang). Similarly to geolocation prediction, we are interested to see how well the geoadapted PLMs can identify the language of a text without fine-tuning. We reuse the FT-Lang test sets for this task. The setup follows ZS-Geoloc, i.e., we append the same prompts to the posts, pass the full sequences through the PLMs, and feed the output representations of the `[MASK]` token into the language modeling head. However, instead of city/state names, we now consider language names, specifically, *bosanski* ('Bosnian'), *crnogorski* ('Montenegrin'), *hrvatski* ('Croatian'), and *srpski* ('Serbian') in the case of BCMS, and *dansk* ('Danish'), *norsk* ('Norwegian'), and *svensk* ('Swedish') in the case of DNS.[14] We select the

[12] Most posts in the AGS data come from rural areas.

[13] We experimented with other prompts (e.g., 'This is in [MASK]') and obtained similar results.

[14] We do not conduct ZS-Lang for AGS and EUR since the names of the German dialects (e.g., *Schweizerdeutsch*) are not in the GermanBERT and mBERT vocabularies.

language name with the highest logit and measure the performance in terms of accuracy.

Zero-shot Dialect Feature Prediction (ZS-Dialect). The fifth evaluation tests whether geoadaptation increases the PLMs' awareness of dialectal variation. We only conduct this task for BCMS, which exhibits many well-documented dialectal variants that exist as tokens in the BERTić vocabulary.

We consider two subtasks. In the first subtask (Phon), we test whether BERTić can select the correct variant for a phonological variable, specifically the reflex of the Old Slavic vowel *ě*. This feature exhibits geographic variation in BCMS: In the (north-)west, the reflexes *ije* and *je* are predominately used, whereas the (south-)east mostly uses *e* (Ljubešić et al., 2018), e.g., *lijepo* vs. *lepo* ('nice'). Drawing upon words for which both *ije*/*je* and *e* variants exist in the BERTić vocabulary, we filter out words that appear in fewer than 10 posts in the merged VarDial dev and test data, resulting in a set of 64 words (i.e., 32 pairs). Subsequently, we randomly sample 10 posts for each of the words. For the second subtask (Lex), we evaluate the recognition of lexical variation that is not tied to a phonological feature (Alexander, 2006), e.g., *porodica* vs. *obitelj* ('family'). Based on a Croatian-Serbian comparative dictionary,[15] we select all pairs for which both words are in the BERTić vocabulary. We remove words that occur in fewer than 10 VarDial dev and test posts and sample 10 posts for each of the remaining 61 words.

For prediction, we mask out the phonological/lexical variant and follow the same approach as for ZS-Geoloc and ZS-Lang, with the difference that we restrict the vocabulary to the two relevant variants (e.g., *porodica* vs. *obitelj*). We measure the performance in terms of accuracy.

Model Variants. We evaluate the two geoadaptation variants, minimizing the simple sum of $\mathcal{L}_{\text{mlm}}$ and $\mathcal{L}_{\text{geo}}$ (GeoAda-S) and the weighted sum based on homoscedastic uncertainty (GeoAda-W). To quantify the effects of geoadaptation compared to standard adaptation, we adapt the PLMs on the same data using only $\mathcal{L}_{\text{mlm}}$ as the primary baseline (MLMAda), i.e., the MLMAda models are adapted *on the exact same text data* as GeoAda-S and GeoAda-W, but using continued language modeling training *without* geolocation prediction. Where possible (i.e., BCMS FT-Geoloc and out-of-domain BCMS FT-Lang), we compare against the current state-of-the-art (SotA) performances (Scherrer and Ljubešić, 2021; Rupnik et al., 2023)—BERTić fine-tuned on the train data. On the zero-shot tasks, we also report random performance (Rand).

Language identification is a task that is not typically addressed using PLMs. Instead, most state-of-the-art systems are less expensive models trained on character n-grams (Zampieri et al., 2017; Haas and Derczynski, 2021; Rupnik et al., 2023). To get a sense of whether PLMs in general and geoadapted PLMs in particular are competitive with such custom-built systems, we evaluate GlotLID (Kargaran et al., 2023), a strong language identification tool based on FastText (Bojanowski et al., 2017; Joulin et al., 2017), on FT-Lang. Since GlotLID was not specifically trained on the domains examined in FT-Lang, we also train new FastText models on the data used to fine-tune the PLMs.

Hyperparameters. For geoadaptation, we use a batch size of 32 (16 for mBERT) and perform grid search for the learning rate $r \in \{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}\}$. We always geoadapt the PLMs for 25 epochs. For FT-Geoloc, we use a batch size of 32 (16 for mBERT) and perform grid search for the number of epochs $n \in \{1, \ldots, 10\}$ and the learning rate $r \in \{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}\}$. For FT-Lang, we use a batch size of 32 (16 for mBERT) and perform grid search for the number of epochs $n \in \{1, \ldots, 5\}$ and the learning rate $r \in \{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}\}$. For all training settings (geoadaptation, FT-Geoloc, FT-Lang) we tune r for MLMAda only and use the best configuration for GeoAda-W and GeoAda-S. This means that the overall number of hyperparameter trials is three times larger for MLMAda than GeoAda-W and GeoAda-S, i.e., we are giving a substantial advantage to the models that serve as a baseline. We use Adam (Kingma and Ba, 2015) as the optimizer. All experiments are performed on a GeForce GTX 1080 Ti GPU (11GB). For the FastText models trained on FT-Lang, we perform grid search for the number of epochs $n \in \{5, 10, 15, 20, 25\}$, the minimum length of included character n-grams $l_{\text{min}} \in \{1, 2, 3\}$, and the

[15] https://hr.wiktionary.org/wiki/Razlikovni_rje%C4%8Dnik_hrvatskog_jezika_i_srpskog_jezika.

	FT-Geoloc ↓								ZS-Geoloc ↑			
	AGS		BCMS		DNS		EUR					
Method	Dev	Test	Dev	Test	Dev	Test	Dev	Test	AGS	BCMS	DNS	EUR
SotA / Rand	—	—	?30.11	?15.49	—	—	—	—	‡.071	‡.070	‡.026	‡.021
MLMAda	‡193.51	‡196.18	‡29.36	‡16.72	‡101.15	‡101.15	‡107.20	‡107.41	‡.142	‡.144	‡.106	‡.108
GeoAda-S	†190.21	**193.18**	†26.02	†13.98	†98.82	†97.63	†98.00	†101.76	.192	†.287	†.135	†.159
GeoAda-W	**189.06**	†194.85	**23.90**	**12.13**	**95.80**	**97.06**	**97.18**	**97.18**	**.193**	**.319**	**.149**	**.191**

Table 3: Results on fine-tuned geolocation prediction (FT-Geoloc) and zero-shot geolocation prediction (ZS-Geoloc). Measure for FT-Geoloc: median distance (in km); measure for ZS-Geoloc: prediction accuracy. For FT-Geoloc and BCMS, the first row shows the current state-of-the-art performance (Scherrer and Ljubešić, 2021). For ZS-Geoloc, the first row shows random performance. **Bold**: best score in each column; underline: second best score. We highlight scores that are significantly ($p < .05$) worse than the best score with a † and scores that are significantly ($p < .05$) worse than the two best scores with a ‡. We indicate with a ? scores for which we cannot test for statistical significance since we do not have access to the distribution of output predictions.

maximum length of included character n-grams $l_{\max} \in \{4, 5, 6\}$.

5 Results and Analysis

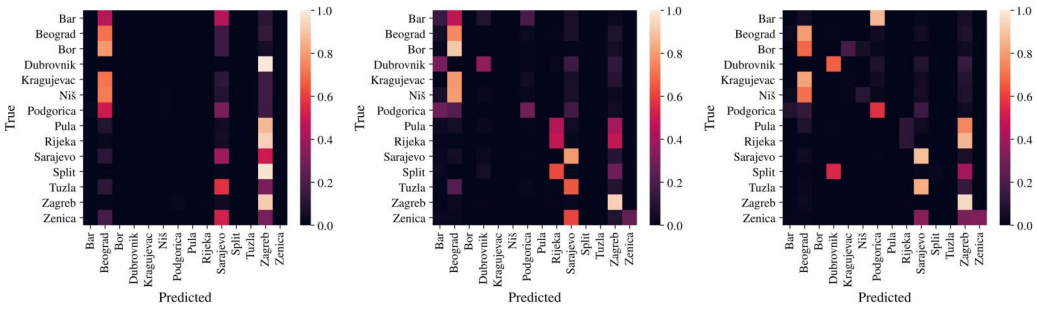
Tables 3, 5, 6, and 7 compare the performance of the geoadapted PLMs against the baselines. To test for statistical significance of the performance differences, we use paired, two-sided Student's t-tests in the case of FT-Geoloc and McNemar's tests for binary data (McNemar, 1947) in the case of ZS-Geoloc, FT-Lang, ZS-Lang, and ZS-Dialect, as recommended by Dror et al. (2018). We correct the resulting p-values for each evaluation using the Holm-Bonferroni method (Holm, 1979).

Overall, the geoadapted models consistently and substantially outperform the baselines—out of the 30 main evaluations, it is *always* one of the two geoadapted models that achieves the best score, a result that is highly unlikely to occur by chance if there is no underlying performance difference between the geoadapted and non-geoadapted models.[16] Furthermore, in the two cases where we can directly compare to a prior state of the art, one or both geoadapted models outperform it. These findings strongly suggest that geoadaptation successfully induces associations between language variation and geographic location.

[16] Assuming equal underlying performance for MLMAda, GeoAda-S, and GeoAda-W (and ignoring other baselines), the probability for this result is $p = (2/3)^{30} < 10^{-5}$.

Fine-tuned Geolocation Prediction. PLMs geoadapted with uncertainty weighting (GeoAda-W) predict the geolocation most precisely (see Table 3). On BCMS, GeoAda-W improves the previous state of the art—achieved by a directly fine-tuned BERTić model—by 3.3 km on test and by over 6 km on dev. On EUR (arguably the most challenging setting), GeoAda-W improves upon MLMAda (i.e., a model adapted without geographic signal) by more than 10 km on both dev and test. MLMAda always performs worse than the two geoadapted models, despite the fact that task-specific fine-tuning likely compensates for some of the geographic knowledge GeoAda-W and GeoAda-S obtain in geoadaptation. This shows that *geoadaptation* drives the performance improvements, and that language modeling adaptation alone does not suffice. Loss weighting based on homoscedastic uncertainties seems beneficial for FT-Geoloc: While GeoAda-S already outperforms the baselines, GeoAda-W in seven out of eight cases brings further significant gains. We also observe that all models reach peak performance in the first few fine-tuning epochs (not shown), and that geoadaptation is useful even when the geoadaptation data are a subset of the fine-tuning data (as is the case for AGS). This confirms that the performance gains come from the geoadaptation and are not merely the result of longer training on geolocation prediction.

Zero-shot Geolocation Prediction. In this task, the PLMs have to predict the *token* of the correct



(a) MLMAda (b) GeoAda-S (c) GeoAda-W

Figure 2: Confusion matrices for MLMAda (a), GeoAda-S (b), and GeoAda-W (c) on ZS-Geoloc (BCMS). While MLMAda always predicts one of the three most frequent city tokens (*Beograd*, *Sarajevo*, or *Zagreb*), the predictions of GeoAda-S and GeoAda-W are much more diverse and less tied to frequency.

toponym (i.e., city or state). Notice that the PLMs receive information about exact geolocations during geoadaptation and do not leverage toponym tokens in any direct way. ZS-Geoloc is thus an ideal litmus test as it shows how well the link between language variation and geography, injected into the PLMs via geoadaptation, generalizes. The results (see Table 3) strongly suggest that geoadaptation leads to such generalization: Both geoadapted model variants bring massive and statistically significant gains in prediction accuracy over MLMAda (e.g., GeoAda-W vs. MLMAda: +17.5% on BCMS, +8.3% on EUR). As on FT-Geoloc, uncertainty weighting (GeoAda-W) overall outperforms simple loss summation (GeoAda-S).

Figure 2 shows the confusion matrices for the three methods on BCMS, offering further insights. MLMAda assigns most posts from a country to the corresponding capital (e.g., posts from Croatian cities to *Zagreb*). These tokens are the most frequent ones out of all considered cities, which seems to heavily affect MLMAda. In contrast, predictions of GeoAda-S and GeoAda-W are much more nuanced, i.e., more diverse and less tied to the frequency of the toponym tokens: The geoadapted models are not only able to correctly assign posts from smaller, less frequently mentioned cities (e.g., *Dubrovnik*, *Zenica*), but their errors also reflect regional linguistic consistency and geographic proximity. For example, GeoAda-S predicts *Rijeka* as the origin of many *Pula* posts, and *Bar* as the origin of many *Dubrovnik* posts; similarly, GeoAda-W assigns posts from *Split* to *Dubrovnik* and posts from *Bar* to *Podgorica*.[17]

[17]Note that *Bar* and *Dubrovnik* are not in the same country.

One common method to alleviate the impact of different prior probabilities in the zero-shot setting (a potential reason for the bad performance of MLMAda) is to *calibrate* the PLM predictions (Holtzman et al., 2021; Zhao et al., 2021). Following Zhao et al. (2021), we measure the prior probabilities of all toponym tokens using a neutral prompt (specifically, 'This is [MASK]' for AGS/BCMS/DNS and a `[MASK]` token for EUR) and repeat the ZS-Geoloc evaluation, dividing the output probabilities by the prior probabilities (Table 4). We find that all models (both geoadapted and non-geoadapted) improve as a result of calibration, i.e., the output probabilities seem to be miscalibrated if not specifically adjusted by means of the prior probabilities. However, refuting the hypothesis that miscalibration causes the inferior performance of MLMAda, the average gain due to calibration is larger for the geoadapted models (GeoAda-S: +4.8%, GeoAda-W: +3.0%) than for the non-geoadapted models (MLMAda: +1.9%). This suggests that a miscalibration of the toponym probabilities—rather than disproportionately affecting the non-geoadapted models—generally impairs the geolinguistic capabilities of a PLM. The consequences of such an impairment seem to be the more detrimental the more profound the underlying geolinguistic knowledge.

Taken together, these observations indicate that GeoAda-S and GeoAda-W possess detailed knowledge of geographic variation in language. Since geoadaptation provides no supervision in the form of toponym names, this implies an impressive generalization, i.e., the association of linguistic constructs to toponyms, with geolocations (specifically, scalar longitude-latitude pairs) as the intermediary signal driving the generalization.

Method	ZS-Geoloc ↑			
	AGS	BCMS	DNS	EUR
MLMAda	‡.156 ↑.014	‡.150 ↑.006	‡*.131 ↑.025	‡*.139 ↑.031
GeoAda-S	*.**229** ↑.036	*.**386** ↑.099	.147 ↑.012	†*.195 ↑.036
GeoAda-W	*.**229** ↑.036	*.373 ↑.054	**.152** ↑.003	*.**219** ↑.028

Table 4: Results on zero-shot geolocation prediction (ZS-Geoloc) with calibration (Zhao et al., 2021). Measure: prediction accuracy. Besides the results, we give the changes compared to vanilla ZS-Geoloc and indicate with a * if they are significant ($p < .05$). See Table 3 for an explanation of the other symbols used in the table.

	FT-Lang ↑								ZS-Lang ↑	
	AGS		BCMS		DNS		EUR			
Method	Dev	Test	Dev	Test	Dev	Test	Dev	Test	BCMS	DNS
Rand	–	–	–	–	–	–	–	–	‡.245	‡.339
GlotLID	–	–	‡.323	‡.316	‡.927	‡.931	–	–	–	–
FastText	‡.843	‡.840	‡.598	‡.588	‡.948	‡.959	‡.757	‡.762	–	–
MLMAda	.851	.855	‡.693	‡.694	‡.964	‡.966	‡.776	‡.777	‡.417	‡.885
GeoAda-S	**.861**	.856	.734	.726	.972	.975	.789	†.786	**.553**	†.896
GeoAda-W	**.861**	**.858**	**.743**	**.734**	**.973**	**.976**	**.792**	**.796**	†.543	**.927**

Table 5: Results on fine-tuned language identification (FT-Lang) and zero-shot language identification (ZS-Lang). Measure: Prediction accuracy. See Table 3 for an explanation of the symbols used in the table.

Fine-tuned Language Identification. The geoadapted PLMs are best at identifying the language in which a text is written: Both GeoAda-S and GeoAda-W consistently show a higher accuracy than MLMAda (e.g., GeoAda-W vs. MLMAda: +5% on BCMS dev, +1.9% on EUR test), and the difference in performance is statistically significant in six out of eight cases (see Table 5). As opposed to the two geolocation tasks where uncertainty weighting (GeoAda-W) clearly leads to better results than summing the losses (GeoAda-S), the difference is less pronounced for FT-Lang and significant only in one case (EUR test), even though GeoAda-W numerically outperforms GeoAda-S overall. Compared to the language identification models operating on the level of character n-grams (GlotLID, FastText), geoadaptation always brings statistically significant performance gains. Even MLMAda outperforms GlotLID and FastText in all cases, indicating that PLMs are generally competitive with more traditional systems on this task. We further notice that the relative disadvantage is particularly pronounced for GlotLID on BCMS. Upon inspection, we find that GlotLID's inferior performance on BCMS is due to the fact that it predicts more than 80% of the examples as Croatian. This imbalance can be explained as a result of the domain difference between GlotLID's training data and the FT-Lang evaluation data: While GlotLID was mostly trained on formal texts such as Wikipedia articles and government documents (Kargaran et al., 2023), we test it on data from Twitter. Crucially, while Croatian is the only BCMS language that consistently uses Latin script in formal contexts, with Cyrillic script being preferred especially in Serbian, Latin script is everywhere much more common on social media, even in Serbia (George, 2019). GlotLID seems to be heavily affected by this script mismatch and is only very rarely able to correctly predict the language of non-Croatian posts written in Latin script.

These trends are also reflected by the results on the out-of-domain language identification benchmarks: Geoadaptation always outperforms adaptation based on language modeling alone as well as models operating on the level of character

	FT-Lang ↑					
	BCMS		DNS		ZS-Lang ↑	
Method	Dev	Test	Dev	Test	BCMS	DNS
SotA / Rand	–	?.995	–	–	‡.311	‡.351
GlotLID	‡.692	‡.697	‡.932	‡.931	–	–
FastText	.992	†.983	‡.957	‡.949	–	–
MLMAda	.992	.992	.962	†.957	‡.604	†.822
GeoAda-S	.993	.995	**.964**	**.962**	**.640**	†.826
GeoAda-W	**.994**	**.997**	**.964**	.961	†.631	**.875**

Table 6: Results on out-of-domain fine-tuned language identification (FT-Lang) and zero-shot language identification (ZS-Lang). Measure for FT-Lang and BCMS: macro-average F1-score (for comparability); measure elsewhere: prediction accuracy. For FT-Lang and BCMS, the first row shows the current state-of-the-art performance (Rupnik et al., 2023). For ZS-Lang, the first row shows random performance. See Table 3 for an explanation of the symbols used in the table.

	ZS-Dialect ↑	
Method	Phon	Lex
Rand	‡.501	‡.499
MLMAda	‡.784	‡.872
GeoAda-S	**.870**	.910
GeoAda-W	.858	**.913**

Table 7: Results on zero-shot dialect feature prediction (ZS-Dialect), which is only conducted for BCMS. Measure: prediction accuracy. See Table 3 for an explanation of the symbols used in the table.

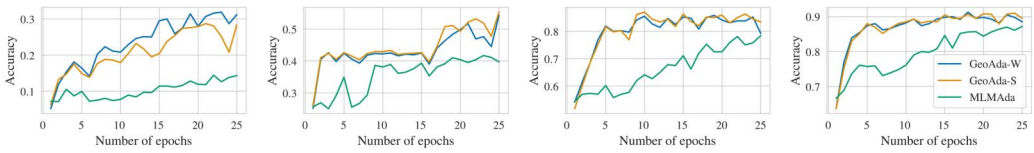
n-gram (see Table 6). On the SETimes benchmark (BCMS), GeoAda-W further establishes a new state of the art, almost halving the error rate from 0.5% to 0.3%. Similarly to in-domain FT-Lang, the two geoadaptation variants perform similarly. GlotLID again predicts many non-Croatian examples in Latin script as Croatian, leading to a substantially worse performance on BCMS.

The superior performance of the geoadapted models in language identification—a task that is distinct from geolocation prediction and not typically addressed by means of PLMs—suggests that the geolinguistic knowledge acquired during geoadaptation is highly generalizable, making it beneficial for a broader set of tasks with a connection to geography, and not only the task used as an auxiliary objective for geoadaptation itself.

Zero-shot Language Identification. Here, the PLMs have to predict the *token* corresponding to the language in which a text is written, e.g., *hrvatski* (‘Croatian’). This task requires generalization on two levels: First (similarly to FT-Lang), the PLMs have not been trained on language identification and are thus required to draw upon the geolinguistic knowledge they have formed during geoadaptation; second (similarly to ZS-Geoloc), the geolinguistic knowledge has not been provided to them in a form that would make it readily usable in a zero-shot setting—recall that the geographic information is presented in the form of longitude-latitude pairs (i.e., two scalars), whereas the language modeling head (which is used for the zero-shot predictions) is not trained differently than for vanilla adaptation (MLMAda). Despite these challenges, we find that geoadaptation substantially improves the performance of the PLMs on ZS-Lang (see Tables 5 and 6). The fact that the performance gains are equally pronounced on in-domain (e.g., GeoAda-W vs. MLMAda: +4.2% on DNS) and out-of-domain examples (e.g., GeoAda-W vs. MLMAda: +5.3% on DNS) highlights again that geoadaptation endows PLMs with knowledge that allows for a high degree of generalization.

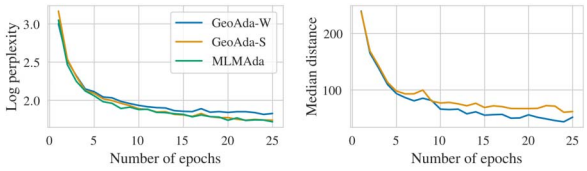
Zero-shot Dialect Feature Prediction. The results on ZS-Dialect—phonological (Phon) and lexical (Lex)—generally follow the trends from the other four tasks (see Table 7): The geoadapted PLMs clearly (and statistically significantly) outperform MLMAda, albeit with overall narrower margins than in most other zero-shot tasks for BCMS (e.g., GeoAda-S vs. MLMAda: +8.6% on Phon, GeoAda-W vs. MLMAda: +4.1% on Lex). MLMAda is expectedly more competitive here: Selecting the word variant that better fits into the linguistic context is essentially a language modeling task, for which additional language modeling training intuitively helps. For example, typical future tense constructions in Serbian vs. Croatian (*ja ću da okupim* vs. *ja ću okupiti*, ‘I’ll gather’) have strong selectional preferences on subsequent lexical units (Alexander, 2006; e.g., *porodicu* vs. *obitelj* for ‘family’).

We further verify this by comparing the zero-shot performance on BCMS for different model checkpoints obtained during training. The



(a) ZS-Geoloc (b) ZS-Lang (c) ZS-Dialect Phon (d) ZS-Dialect Lex

Figure 3: Performance on BCMS ZS-Geoloc (a), ZS-Lang (b), and ZS-Dialect (c, d) for different number of epochs. In stark contrast to geoadaptation (GeoAda-S, GeoAda-W), language modeling adaptation alone (MLMAda) barely helps in acquiring geographic knowledge (a), which is also reflected by the consistently worse performance on ZS-Lang (b). MLMAda does form dialectal associations after several epochs, but the inductive bias of geoadaptation allows GeoAda-S and GeoAda-W to establish those associations more quickly (c, d).



(a) Language modeling (b) Geolocation prediction

Figure 4: (Geo-)adaptation diagnostics. The figure illustrates how log perplexity of language modeling (a) and median distance of token-level geolocation prediction (b) change on dev during BCMS geoadaptation.

performance curves over 25 (geo-)adaptation epochs, shown in Figure 3, confirm our hypothesis that longer language modeling adaptation substantially improves the performance of MLMAda on predicting dialect features, but its benefits for geolocation prediction and language identification remain limited. While prolonged language modeling adaptation allows MLMAda to eventually learn the dialectal associations, the inductive bias of the knowledge injected via geoadaptation allows GeoAda-S and GeoAda-W to reach high performance much sooner, after merely two to three epochs.

Effects of Loss Weighting. The dynamic weighting of $\mathcal{L}_{\text{mlm}}$ and $\mathcal{L}_{\text{geo}}$ (i.e., GeoAda-W) clearly outperforms the simple summation of the losses (i.e., GeoAda-S) on the geolocation prediction tasks (FT-Geoloc, ZS-Geoloc), but the difference between the two geoadaptation variants is less pronounced for FT-Lang, ZS-Lang, and ZS-Dialect. While geographic knowledge is beneficial for all five tasks, geolocation prediction arguably demands a more direct exploitation of that knowledge. Comparing the model variants in terms of the two task losses, we observe that GeoAda-S reaches lower $\mathcal{L}_{\text{mlm}}$ levels, whereas GeoAda-W ends with lower $\mathcal{L}_{\text{geo}}$ levels (see Figure 4 for the example of BCMS), which would explain the differences in their performance. We inspect GeoAda-W's task uncertainty weights after geoadaptation and observe $\eta_{\text{mlm}} = 0.29$ and $\eta_{\text{geo}} = -0.35$ for AGS, $\eta_{\text{mlm}} = 1.12$ and $\eta_{\text{geo}} = -1.22$ for BCMS, $\eta_{\text{mlm}} = 0.84$ and $\eta_{\text{geo}} = -1.23$ for DNS, and $\eta_{\text{mlm}} = 0.90$ and $\eta_{\text{geo}} = -1.95$ for EUR. Thus, GeoAda-W consistently assigns more importance to $\mathcal{L}_{\text{geo}}$.[18] The fact that the divergence of the task uncertainty weights is smallest for AGS explains why the difference between GeoAda-S and GeoAda-W on FT-Geoloc/ZS-Geoloc is least pronounced for that language group.

Sequence-level Geoadaptation. The decision to inject geographical information at the level of tokens was motivated by the central importance of the lexicon for geographically conditioned linguistic variability (see §2). A plausible alternative—one less tied to lexical variation alone—is to geoadapt the PLMs by predicting the geolocation from the representation of the whole input text, i.e., to feed the contextualized representation of the `[CLS]` token to the regressor that predicts longitude and latitude. For comparison, we evaluate this variant too (GeoAda-Seq) and compare it against the best token-level geoadapted model (GeoAda-Tok; e.g., GeoAda-W for BCMS FT-Geoloc) on all PLMs and tasks. For reasons of space, we only present BCMS here, but the overall trends for AGS, DNS, and EUR are very similar.

Sequence-level geoadaptation trails token-level geoadaptation on all tasks except for fine-tuned geolocation prediction (see Table 8). In general, while the difference is small for the fine-tuned

[18] Because $\tilde{\mathcal{L}}_l \propto -\eta_l$ (see Equation 2), the smaller the value of η_l, the larger the emphasis on task l.

	FT-Geoloc ↓			FT-Lang ↑			ZS-Dialect ↑	
Model	Dev	Test	ZS-Geoloc ↑	Dev	Test	ZS-Lang ↑	Phon	Lex
GeoAda-Seq	†27.35	**12.13**	†.188	.737	.730	†.542	†.844	†.885
GeoAda-Tok	**23.90**	**12.13**	**.319**	**.743**	**.734**	**.553**	**.870**	**.913**

Table 8: Comparison between sequence-level geoadaptation (GeoAda-Seq) and token-level geoadaptation (GeoAda-Tok) for BCMS. GeoAda-Tok stands for the better-performing model between GeoAda-S and GeoAda-W on each task (see Tables 3, 5, and 7). See Table 3 for an explanation of the symbols used in the table.

tasks, it is large (and always significant) for the zero-shot tasks—for example, GeoAda-Seq performs only slightly better than MLMAda on ZS-Geoloc (see Table 3). This suggests that injecting geographic information on the level of tokens allows the PLMs to acquire more nuanced geolinguistic knowledge. Nonetheless, sequence-level geoadaptation still outperforms the non-geoadapted baselines.

Geoadaptation as Geographic Retrofitting. Even though it makes intuitive sense that minimizing $\mathcal{L}_{\text{geo}}$ improves the geolinguistic knowledge of PLMs, we want to determine the exact mechanism by which it does so. Based on the results described so far, we make the following hypothesis: Geoadaptation changes the representation space of the PLMs in such a way that tokens indicative of a certain location are brought close to each other, i.e., it has the effect of *geographic retrofitting* (Hovy and Purschke, 2018). We examine this hypothesis by analyzing (i) how the representations of toponyms and lexical variants change in relation to each other, and (ii) how the representations of toponyms change internally. We examine the PLM output embeddings (which directly impact the zero-shot predictions) and focus on BCMS.

For the first question, we use the geoadaptation data to compute type-level embeddings for the five largest Croatian (*Zagreb*, *Split*, *Rijeka*, *Osijek*, *Zadar*) and Serbian (*Beograd*, *Niš*, *Kragujevac*, *Subotica*, *Pančevo*) cities as well as the *ije*/*e* variants used for ZS-Dialect. Following established practice (e.g., Vulić et al., 2020; Litschko et al., 2022), we obtain type-level vectors for words (i.e., city name or phonological variant) by averaging the contextualized output representations of their token occurrences. We then resort to WEAT (Caliskan et al., 2017), a measure that quantifies the difference in association strength between a word (in our case, a city name) and two word sets (in our case, *ije* vs. *e* phonological variants). A positive or negative score indicates that a city name is associated more strongly with the *ije* or *e* variants, respectively. Figure 5 shows that during geoadaptation (GeoAda-W), the Croatian city names develop a strong association with the *ije* variants (i.e., positive WEAT scores), whereas the Serbian city names develop a strong association with the *e* variants (i.e., negative WEAT scores), which is exactly in line with their geographic distribution (Alexander, 2006). By contrast, the associations created during adaptation based on language modeling alone (MLMAda) are substantially weaker.

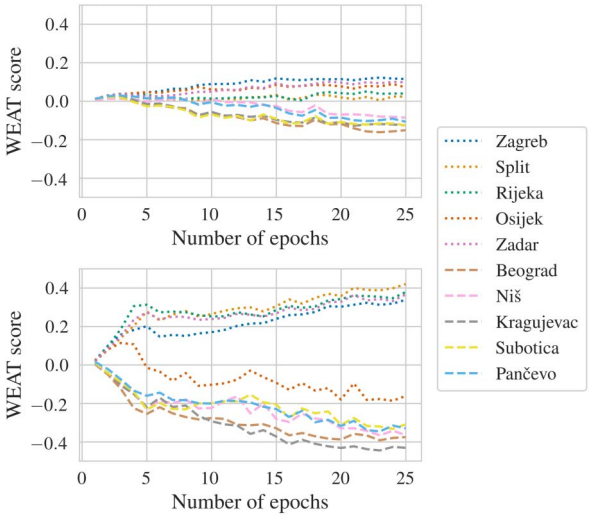


Figure 5: Association strength between the BERTić embeddings of Croatian/Serbian cities and *ije*/*e* variants for MLMAda (top) and GeoAda-W (bottom), measured using WEAT (Caliskan et al., 2017). A positive or negative score indicates that a city is associated more strongly with the *ije* or *e* variants, respectively.

We then use the same set of 10 Croatian and Serbian cities and compare their pairwise geodesic distances against the pairwise cosine distances of the city name embeddings, at the end of geoadaptation. The correlation between the two

Figure 6: The first two principal components of the city output embeddings (points), plotted on top of a geographic map of Croatia and Serbia. The x-marks indicate the actual geographic locations of the cities.

sets of distances (Pearson's r) is only 0.577 for MLMAda, but 0.881 for GeoAda-W, indicating almost perfect correlation. Furthermore, after only five epochs, the correlation is already 0.845 for GeoAda-W (vs. only 0.124 for MLMAda). This striking correspondence between real-world geography and the topology of the embedding space of geoadapted PLMs can also be seen by plotting the first two principal components of the city name embeddings on top of a geographic map, where we use orthogonal Procrustes (Schönemann, 1966; Hamilton et al., 2016) to align the points (see Figure 6).

These results are strong evidence that geoadaptation indeed acts as a form of geographic retrofitting. The geographically restructured representation space of the PLMs can then be further refined via fine-tuning (as in FT-Geoloc and FT-Lang) or directly probed in a zero-shot manner (as in ZS-Geoloc, ZS-Lang and ZS-Dialect).

6 Conclusion

We introduce geoadaptation, an approach for task-agnostic continued pretraining of PLMs that forces them to learn associations between linguistic phenomena and geographic locations. The method we propose for geoadaptation couples language modeling and token-level geolocation prediction via multi-task learning. While we focus on PLMs pretrained via masked language modeling, geoadaptation can in principle be applied to autoregressive PLMs as well. We geoadapt four PLMs and obtain consistent gains on five tasks, establishing a new state of the art on established benchmarks. We further show that geoadaptation acts as a form of geographic retrofitting. Overall, we see our study as an exciting step towards NLP technology that takes into account extralinguistic aspects in general and geographic aspects in particular.

Acknowledgments

This work was funded by the European Research Council (grant #740516 awarded to LMU Munich) and the Engineering and Physical Sciences Research Council (grant EP/T023333/1 awarded to University of Oxford). Valentin Hofmann was also supported by the German Academic Scholarship Foundation. Goran Glavaš was supported by the EUINACTION grant from NORFACE Governance and German Science Foundation (462-19-010, GL950/2-1). Nikola Ljubešić was supported by the Slovenian Research and Innovation Agency (P6-0411, J7-4642, L2-50070). We thank the reviewers and action editor for their very helpful comments.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. arXiv preprint *arXiv 2101.11038*. https://doi.org/10.18653/v1/2021.emnlp-main.468

Ronelle Alexander. 2006. *Bosnian, Croatian, Serbian: A Grammar with Sociolinguistic Commentary*. University of Wisconsin, Madison, WI.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Annual Meeting of the Association for Computational Linguistics (ACL) 52*. https://doi.org/10.3115/v1/P14-2134

Su L. Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016*. https://doi.org/10.18653/v1/D16-1120

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*,

5:135–146. https://doi.org/10.1162/tacl_a_00051

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

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *arXiv preprint arXiv:2201.11732*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. https://doi.org/10.1126/science.aal4230, PubMed: 28408601

Bharathi R. Chakravarthi, Mihaela Găman, Radu T. Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 8*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. arXiv preprint *arXiv 2204.02311*.

Alexandra Chronopoulou, Matthew E. Peters, and Jesse Dodge. 2021. Efficient hierarchical domain adaptation for pretrained language models. *arXiv preprint arXiv:2112.08786*. https://doi.org/10.18653/v1/2022.naacl-main.96

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR) 8*.

Cynthia G. Clopper and David B. Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32(1):111–140. https://doi.org/10.1016/S0095-4470(03)00009-3, PubMed: 21451736

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 *Annual Meeting of the Association for Computational Linguistics (ACL) 58*. https://doi.org/10.18653/v1/2020.acl-main.747

Dorottya Demszky, Devyani Sharma, Jonathan H. Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*. https://doi.org/10.18653/v1/2021.naacl-main.184

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Bhuwan Dhingra, Jeremy R. Cole, Julian M. Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2021. Time-aware language models as temporal knowledge bases. arXiv preprint *arXiv 2106.15110*.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) 14*. https://doi.org/10.3115/v1/E14-1011

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*. https://doi.org/10.18653/v1/P18-1128

Jonathan Dunn. 2019. Modeling global syntactic variation in English using dialect classification. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 6*.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. arXiv preprint *arXiv 2004.03032*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric Xing. 2010. A latent variable model for geographic lexical variation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2010*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric Xing. 2014. Diffusion of lexical change in social media. *PloS ONE*, 9(11):e113114. https://doi.org/10.1371/journal.pone.0113114, PubMed: 25409166

Jacob Eisenstein, Noah A. Smith, and Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Annual Meeting of the Association for Computational Linguistics (ACL) 49*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*. https://doi.org/10.18653/v1/D19-1006

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl_a_00298

Rachel George. 2019. Simultaneity and the refusal to choose: The semiotics of Serbian youth identity on Facebook. *Language in Society*, 49(2):399–423. https://doi.org/10.1017/S004740451900099X

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. Xhate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365. https://doi.org/10.18653/v1/2020.coling-main.559

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? An empirical investigation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) 16*. https://doi.org/10.18653/v1/2021.eacl-main.270

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL) 58*. https://doi.org/10.18653/v1/2020.acl-main.740

René Haas and Leon Derczynski. 2021. Discriminating between similar Nordic languages. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 8*.

William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Annual Meeting of the Association for Computational Linguistics (ACL) 54*. https://doi.org/10.18653/v1/P16-1141

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word

representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. arXiv preprint *arXiv 2203.15556*.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*. https://doi.org/10.18653/v1/2020.emnlp-main.316

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2021. Dynamic contextualized word embeddings. In *Annual Meeting of the Association for Computational Linguistics (ACL) 59*. https://doi.org/10.18653/v1/2021.acl-long.542

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021*. https://doi.org/10.18653/v1/2021.emnlp-main.564

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*. https://doi.org/10.18653/v1/D18-1469

Dirk Hovy, Afshin Rahimi, Timothy Baldwin, and Julian Brooke. 2020. Visualizing regional language variation across Europe on Twitter. In Stanley D. Brunn and Roland Kehrein, editors, *Handbook of the Changing World Language Map*, pages 3719–3742. Springer, Cham. https://doi.org/10.1007/978-3-030-02438-3_175

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*. https://doi.org/10.18653/v1/2021.naacl-main.49

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59(9):244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Mans Hulden, Iñaki Alegria, Izaskun Etxeberria, and Montse Maritxalar. 2011. Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2011*.

Chia-Chien Hung, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Ds-tod: Efficient domain specialization for task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904.

Ganesh Jawahar, Benoit Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Annual Meeting of the Association for Computational Linguistics (ACL) 57*. https://doi.org/10.18653/v1/P19-1356

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. https://doi.org/10.18653/v1/2020.acl-main.560

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) 15*. https://doi.org/10.18653/v1/E17-2068

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023*. https://doi.org/10.18653/v1/2023.findings-emnlp.410

Ilia Karpov and Nick Kartashev. 2021. SocialBERT: Transformers for online social network language modelling. arXiv preprint *arXiv 2111.07148*.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NIPS) 31*.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Conference on Computer Vision and Pattern Recognition (CVPR) 31*.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR) 3*.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 10*. https://doi.org/10.1609/icwsm.v10i1.14798

Vivek Kulkarni, Mishra Shubhanshu, and Aria Haghighi. 2021. LMSOC: An approach for socially sensitive pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. https://doi.org/10.18653/v1/2021.findings-emnlp.254

Mikko Laitinen, Jonas Lundberg, Magnus Levin, and Rafael Martins. 2018. The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. In *Digital Humanities in the Nordic Countries (DHN) 3*.

Norman J. Lass, John E. Tecca, Robert A. Mancuso, and Wanda I. Black. 1979. The effect of phonetic complexity on speaker race and sex identifications. *Journal of Phonetics*, 7(2):105–118. https://doi.org/10.1016/S0095-4470(19)31044-7

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? Investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49. https://doi.org/10.18653/v1/2020.deelio-1.5

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183. https://doi.org/10.1007/s10791-022-09406-x

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL) 57*.

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

Nikola Ljubešić and Davor Lauc. 2021. Bertić: The transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Workshop on Balto-Slavic Natural Language Processing (BSNLP) 8*.

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, 6(2):100–124. https://doi.org/10.1017/jlg.2018.9

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2021. Time waits for no one! Analysis and challenges of temporal misalignment. arXiv preprint *arXiv 2111.07408*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: A cognitive perspective. arXiv preprint *arXiv 2301.06627*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. https://doi.org/10.1007/BF02295996, PubMed: 20254758

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. Bad-x: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799. https://doi.org/10.18653/v1/2022.naacl-main.130

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673. https://doi.org/10.18653/v1/2020.emnlp-main.617

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. In *arXiv 1811.01088*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015a. Twitter user geolocation using a unified text and network prediction model. In *Annual Meeting of the Association for Computational Linguistics (ACL) 53*. https://doi.org/10.3115/v1/P15-2104

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. In *Annual Meeting of the Association for Computational Linguistics (ACL) 55*. https://doi.org/10.18653/v1/P17-2033

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*. https://doi.org/10.18653/v1/P18-1187

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015b. Exploiting text and network context for geolocation of social media users. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 2015*. https://doi.org/10.3115/v1/N15-1153

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo M. Ponti, Anna Korhonen, and Ivan Vulić. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402. https://doi.org/10.1613/jair.1.13083

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *International Conference on Web Search and Data Mining (WSDM) 15*. https://doi.org/10.1145/3488560.3498529

Paul Röttger and Janet B. Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. https://doi.org/10.18653/v1/2021.findings-emnlp.206

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 10*. https://doi.org/10.18653/v1/2023.vardial-1.11

Bahar Salehi, Dirk Hovy, Eduard Hovy, and Anders Søgaard. 2017. Huntsville, hospitals, and

hockey teams: Names can reveal your location. In *Workshop on Noisy User-generated Text (WNUT) 3*. https://doi.org/10.18653/v1/W17-4415

Yves Scherrer and Nikola Ljubešić. 2020. HeLju@VarDial 2020: Social media variety geolocation with BERT models. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 7*.

Yves Scherrer and Nikola Ljubešić. 2021. Social media variety geolocation with geoBERT. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 8*.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 36(1). https://doi.org/10.1007/BF02289451

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL) 57*. https://doi.org/10.18653/v1/P19-1355

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. In *arXiv 2302.13971*.

Sonja A. Trent. 1995. Voice quality: Listener identification of African–American versus Caucasian speakers. *The Journal of the Acoustical Society of America*, 98(5):2936. https://doi.org/10.1121/1.414099

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*. https://doi.org/10.18653/v1/2020.emnlp-main.635

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240. https://doi.org/10.18653/v1/2020.emnlp-main.586

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021*. https://doi.org/10.18653/v1/2021.emnlp-main.72

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schütze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023*. https://doi.org/10.18653/v1/2023.emnlp-main.401

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022*. https://doi.org/10.18653/v1/2022.emnlp-main.746

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. arXiv preprint *arXiv 1909.10430*.

Martijn Wieling and John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics*, 1(1):243–264. https://doi.org/10.1146/annurev-linguist-030514-124930

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

in Natural Language Processing (EMNLP) 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wenhan Xiong, Jingfei Du, William Y. Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations (ICLR) 8*.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 4*. https://doi.org/10.18653/v1/W17-1201

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) 1*. https://doi.org/10.3115/v1/W14-5307

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning (ICML) 38*.