Adapting Monolingual Models: Data can be Scarce when Language Similarity is High

Wietse de Vries*, **Martijn Bartelds***, **Malvina Nissim** and **Martijn Wieling**

University of Groningen
The Netherlands

{wietse.de.vries, m.bartelds, m.nissim, m.b.wieling}@rug.nl

Abstract

For many (minority) languages, the resources needed to train large models are not available. We investigate the performance of zero-shot transfer learning with as little data as possible, and the influence of language similarity in this process. We retrain the lexical layers of four BERT-based models using data from two low-resource target language varieties, while the Transformer layers are independently fine-tuned on a POS-tagging task in the model's source language. By combining the new lexical layers and fine-tuned Transformer layers, we achieve high task performance for both target languages. With high language similarity, 10MB of data appears sufficient to achieve substantial monolingual transfer performance. Monolingual BERT-based models generally achieve higher downstream task performance after retraining the lexical layer than multilingual BERT, even when the target language is included in the multilingual model.

1 Introduction

Large pre-trained language models are the dominant approach for solving many tasks in natural language processing. These models represent linguistic structure on the basis of large corpora that exist for high-resource languages, such as English. However, for the majority of the world's languages, these large corpora are not available.

Past work on multilingual learning has found that multilingual BERT (mBERT; Devlin et al. 2019a) generalizes across languages with high zero-shot transfer performance on a variety of tasks (Pires et al., 2019; Wu and Dredze, 2019). However, it has also been observed that high-resource languages included in mBERT pre-training often have a better-performing monolingual model, and low-resource languages that are not included in mBERT pre-training usually show poor performance (Nozza et al., 2020; Wu and Dredze, 2020).

An alternative to multilingual transfer learning is the adaptation of existing monolingual models to other languages. Zoph et al. (2016) introduce a method for transferring a pre-trained machine translation model to lower-resource languages by only fine-tuning the lexical layer. This method has also been applied to BERT (Artetxe et al., 2020) and GPT-2 (de Vries and Nissim, 2020). Artetxe et al. (2020) also show that BERT models with retrained lexical layers perform well in downstream tasks, but comparatively high performance has only been demonstrated for languages for which at least 400MB of data is available.

To test if this procedure is also effective for low- to zero-resource languages, we consider two regional language varieties spoken in the North of the Netherlands, namely Gronings (Low Saxon language variant) and West Frisian.

Figure 1: Geographical areas where Gronings (in green) and West Frisian (in red) are spoken. Image modified from https://en.wikipedia.org/wiki/Low_German.

*These authors contributed equally.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4901–4907
August 1–6, 2021.

Figure 1 visualizes the geographical areas where these regional language variants are spoken. The regional Low Saxon language is spoken in the northeastern provinces of the Netherlands and in the North of Germany (shown in yellow). As part of the Low Saxon language, Gronings is spoken in the province of Groningen (highlighted in green). The West Frisian language is spoken in the province of Friesland (shown in red), and it is the second official language of the Netherlands, next to Dutch. Dutch is the national language of the Netherlands, and it is spoken in every province of the Netherlands and in Flanders (North of Belgium).

For both Gronings and West Frisian limited data is available. In addition to unlabeled data, for both target languages we have a small collection of annotated part-of-speech (POS) tagging data, which we use for evaluating zero-shot model transfer. We use three monolingual BERT models (source languages English, German, Dutch) and mBERT to investigate if linguistic structure can be transferred to Gronings and West Frisian by learning new subword embeddings. Our model source and target languages are closely related West Germanic languages (Eberhard et al., 2020). In Table 1, we show parallel sentences in Gronings, West Frisian, Dutch, German, and English to illustrate the lexical similarity between these languages. Additionally, the examples show that there are some lexical and syntactic differences.

We also evaluate to what extent the similarity between each source language of the monolingual models and the target languages is relevant for transferring monolingual representations, and assess the minimum amount of data necessary to adapt these models.

Gronings	Tom is n jong en Mary is n wicht.
West Frisian	Tom is in jonge en Mary is in famke.
Dutch	Tom is een jongen en Mary is een meisje.
German	Tom ist ein Junge und Mary ist ein Mädchen.
English	Tom is a boy and Mary is a girl.
Gronings	Zie haar n bloum ien heur haand.
West Frisian	Se hie in blom yn har hân.
Dutch	Ze had een bloem in haar hand.
German	Sie hatte eine Blume in der Hand.
English	She had a flower in her hand.
Gronings	Dat was n poar joar leden.
West Frisian	Dat wie in pear jier lyn.
Dutch	Dat was een paar jaar geleden.
German	Das war vor ein paar Jahren.
English	That was a couple of years ago.

Table 1: Translations of three sentences in Gronings, West Frisian, Dutch, German, and English.

Our pre-trained models for Gronings and West Frisian (which did not yet exist) are released. Additionally, our code is publicly available for bringing language models to other low-resource languages at `https://github.com/wietsedv/low-resource-adapt`.

2 Materials

Models We use monolingual BERT-based models of the source languages, and multilingual BERT (mBERT; Devlin et al. 2019a). Specifically, we use BERT (Devlin et al., 2019b) for English, German BERT (gBERT; DBMDZ 2019) for German, and BERTje (de Vries et al., 2019) for Dutch. Each model shares the same architecture as the original base-sized (12 layers) BERT model of Devlin et al. (2019b). The lexical layer weights are shared between the first and last layer of the model to transform discrete tokens into distributed vector representations and vice versa.

Each monolingual model has a vocabulary of 30K capitalized tokens, while mBERT has a vocabulary of 120K tokens shared between the 104 languages it is pre-trained on. These languages include English, German, Dutch and West Frisian, but not Gronings. The monolingual BERT models contain 110M parameters, with 24M being part of the lexical embeddings. Due to its larger vocabulary size, mBERT contains 180M parameters, with 92M part of the lexical embeddings.

Labeled data We use POS-annotated treebanks from the Universal Dependencies (UD) project (Zeman et al., 2020), corresponding to the languages of the monolingual BERT models. For English, we use `GUM` (6.0K sentences; 113.4K tokens) and `ParTUT` (2.1K sentences; 49.6K tokens). In addition, `HDT` (189.9K sentences; 3.4M tokens) and `GSD` (15.6K sentences; 287.7K tokens) are used for German. Finally, `Alpino` (13.6K sentences; 208.5K tokens) and `LassySmall` (7.3K sentences; 98.0K tokens) are used for Dutch. All treebanks are based on various text types from a diverse set of sources. The standard data splits for each of the annotated treebanks are used for training, validation and testing.

We evaluate the performance of our language models on POS-annotated data of Gronings and West Frisian. Manually annotated texts from the

Klunderloa[1] project are used for Gronings (3.8K sentences, 49.0K tokens; fiction, poetry, and songs for children). Annotations follow the UD guidelines. West Frisian is under development in the UD project, and we consider all currently available annotations (1.0K sentences, 15.9K tokens; mainly fiction and news). For both treebanks, 25% is used for development, and 75% is used as a test set.

Unlabeled data The new sub-word embeddings are learned from texts written in Gronings and West Frisian. In total, we have 43MB (8.3M tokens) of plain text available for Gronings. These texts are derived from the Bible, fiction and non-fiction texts, poetry, and Low Saxon Wikipedia. The West Frisian data collection consists of 59MB (10.8M tokens) of plain text extracted from fiction and non-fiction texts, and the multilingual OSCAR corpus (Ortiz Suárez et al., 2020).

Language similarity To quantify language similarity, we use the (lexical-phonetic) LDND measure (Wichmann et al., 2010) on the basis of the 40-item word lists from the ASJP database (Wichmann et al., 2010). While a syntax-based measure may be preferred, this is not available for the included language varieties. We use the LDND as a proxy, given that linguistic distance measures between different linguistic levels are correlated (Spruit et al., 2009). Figure 2 visualizes the relative linguistic distances between the five language varieties using multidimensional scaling (MDS; Torgerson, 1952). If cross-lingual transfer benefits from language similarity, we expect Gronings and West Frisian to profit most from a monolingual Dutch model and least from a monolingual English model, with a German model performing in-between.

Figure 2: MDS plot with the relative positions of English, German, Dutch, Gronings, and West Frisian based on the ASJP-based lexical-phonetic distances.

[1] http://www.klunderloa.nl

Source		Src.	Gronings		W. Frisian	
		orig.	orig.	gro.	orig.	fri.
EN	BERT	93.8	26.6	61.6	29.1	78.1
	mBERT	93.8	64.1	84.7	87.1	88.7
DE	gBERT	93.3	25.4	85.5	22.7	89.2
	mBERT	93.0	55.9	82.5	86.1	87.7
NL	BERTje	96.4	64.9	**91.7**	48.0	**95.3**
	mBERT	96.6	**72.4**	89.3	**92.2**	94.7

Table 2: Accuracies for the target languages (columns) with the original and retrained lexical layers (sub-columns), which are averaged per source language.

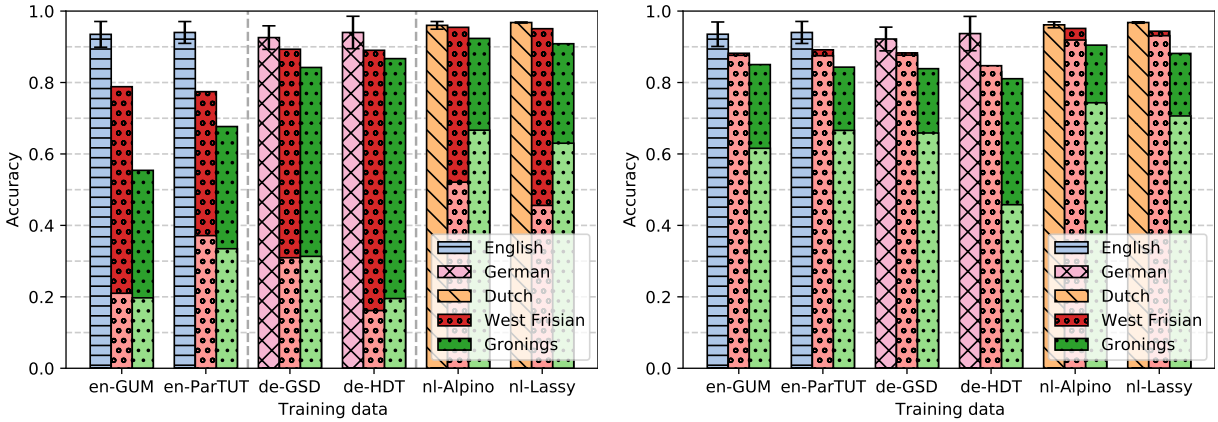
3 Model Training

Our training procedure consists of two separate fine-tuning steps. The Transformer layers in the three monolingual BERT models and mBERT are fine-tuned for the POS-tagging task. Independently, new lexical layers for each BERT model are trained for the two target languages with a masked language modeling pre-training objective. Afterwards, the retrained lexical layer and the fine-tuned Transformer layers are combined to yield a POS-tagging model that is now adapted to the target language. Optimal checkpoint combinations of retrained lexical layers and fine-tuned Transformer layers are based on their performance on the development data for each target language.

POS-tagging The BERT-based models are fine-tuned for POS-tagging with the UD datasets. The task-specific model consists of BERT's layers with an additional linear classification layer that yields predictions for each of the 16 possible POS tags. During training, the lexical layer of BERT is frozen such that the fine-tuned Transformer layers rely on unchanged token representations from pre-training.

The described model is trained with the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}{-8}$ and a linearly decreasing learning rate starting at $lr = 1\text{e}{-5}$. Each model is trained until validation loss stops decreasing.

Lexical layer retraining We retrain lexical layers for each BERT model using Gronings and West Frisian data. First, sub-word vocabularies of 10K tokens are created for Gronings and West Frisian using the WordPiece method (Devlin et al., 2019b) where each token occurs at least 100 times in the data. This vocabulary size is chosen conservatively,



(a) Monolingual model accuracy (BERT, gBERT and BERTje).

(b) Multilingual model accuracy (mBERT).

Figure 3: POS-tagging accuracy for source languages (English, German and Dutch) and target languages (West Frisian and Gronings). Light colors correspond to the accuracy with the original lexical layer and dark colors show improvements with retrained lexical layers. Source language accuracy was averaged across the two source language test sets. Error bars show the upper and lower test set performance for the source language.

		Gronings						West Frisian					
		1MB	5MB	10MB	20MB	40MB	43MB	1MB	5MB	10MB	20MB	40MB	59MB
EN	BERT	32.2	50.5	68.2	69.4	63.3	61.6	51.8	70.6	76.7	78.8	79.1	78.1
	mBERT	25.3	75.4	84.1	84.3	84.4	84.7	72.5	88.0	88.6	89.1	89.2	88.7
DE	gBERT	39.8	83.5	85.5	85.8	85.4	85.5	**76.0**	87.3	87.7	88.0	88.4	89.2
	mBERT	14.1	59.6	79.7	78.0	81.9	82.5	54.9	80.9	84.3	84.5	85.8	85.7
NL	BERTje	**70.2**	**89.5**	**91.4**	**91.4**	**91.4**	**91.7**	44.7	**94.6**	**95.0**	**95.2**	**95.1**	**95.3**
	mBERT	23.8	70.0	87.6	87.6	88.5	89.3	72.2	92.7	93.9	94.4	94.5	94.8

Table 3: POS-tagging accuracy for Gronings and West Frisian with subsets of the unlabeled lexical layer retraining data. Results are averaged per source language for each of the two source language datasets.

as we have limited data to train the lexical layer. Preliminary experiments with 30K tokens showed poor performance on the development data.

The Gronings and West Frisian unlabeled documents are split into sequences of 128 tokens. Then, the models are trained with a masked language modeling objective where 15% of the input tokens are masked. The Adam optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}{-8}$ and a linearly decreasing learning rate starting at $lr = 1\text{e}{-4}$. After retraining, we have three (original, Gronings and West Frisian) interchangeable lexical layers for each base model.

4 Results and Discussion

We summarize our results in Table 2 (details per dataset in Appendix A). The monolingual language models perform poorly on Gronings and West Frisian POS-tagging when the original lexical layers are used, even though Gronings is quite similar to Dutch (see Figure 2). mBERT with its original lexical layer achieves better results than the monolingual models, but only West Frisian performance is comparable to the source language performance. Since West Frisian was included in mBERT pre-training, these results suggest that mBERT might serve languages included in pre-training well, whereas it may be less suitable for those not included (e.g., Gronings).

For all monolingual models, task performance greatly improves by retraining the lexical layer for Gronings and West Frisian (Figure 3a). Best results are obtained by (Dutch) BERTje fine-tuned on the `Alpino` dataset (92.4% for Gronings, 95.4% for West Frisian). In contrast, (English) BERT yields the worst performance. We find that performance scores and the linguistic distance from Gronings and West Frisian to the source languages (Figure 2) strongly correlate ($r = -0.85$, $p < 0.05$). This suggests that measures of linguistic distance can

guide the optimal choice of monolingual models to transfer to low-resource languages. Retraining mBERT's lexical layer also improves performance, especially for Gronings (Figure 3b), but with smaller gains than for monolingual models.

To estimate how our zero-shot approach compares with supervised learning, we train UDPipe (Straka et al., 2016) with five-fold cross-validation on the Gronings and West Frisian POS-tagging data. UDPipe achieves an accuracy of 91.85 ($\sigma = 0.81$) for Gronings and 90.60 ($\sigma = 0.58$) for West Frisian. These results do not indicate out-of-domain performance, since training and test data are from the same source. Also, labeled data for Gronings comes from a corpus with a specific target audience (i.e. children). Therefore, these results can be seen as an upper-bound. Our adapted models perform on par (Gronings) or better (West Frisian) with no need for labeled data in the target language.

Data size Our zero-shot transfer method relies on the availability of unlabeled Gronings and West Frisian data. Other low-resource languages may have even smaller amounts of data available than we have for West Frisian (59MB) and Gronings (43MB). We therefore assess how little data is sufficient for adequate performance by retraining the lexical layer with subsets of (independently randomly sampled) unlabeled data.

Table 3 shows POS-tagging accuracies for each subset. Results are consistent across both target languages and show that ca. 10MB of data (1.9M tokens) is sufficient to achieve almost optimal performance for the monolingual models. By contrast, mBERT shows a steadier improvement with more data, suggesting that it might further improve if even more data is available than we have for Gronings and West Frisian. BERT's POS-tagging accuracy is very low compared to the other monolingual models and performance decreases with more data. These fluctuations suggest that the retrained lexical layer fits BERT poorly and it is unclear if using more data will impact performance positively.

5 Conclusion

We adapted three monolingual BERT models and mBERT to two low-resource languages, Gronings and West Frisian, by retraining the lexical layers with new vocabularies. We found that the adaptability of mBERT is limited, suggesting that a model trained on a large amount of languages may not facilitate transfer to low-resource languages. Instead, monolingual BERT models are transferable to languages with very little data if the source and target languages are relatively similar. In such case, 10MB of unlabeled data, and no task-specific labeled data, is sufficient to achieve high ($> 90\%$ accuracy) downstream task performance.

Acknowledgments

We gratefully acknowledge the support of the Dutch Research Council (NWO Aspasia grant for M. Nissim) and the financial support of the Center for Groningen Language and Culture (CGTC). Finally, we thank the anonymous reviewers for their insightful feedback. Any mistakes remain our own.

References

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.

DBMDZ. 2019. German BERT. https://github.com/dbmdz/berts#german-bert. [Online].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT-base Multilingual cased. https://github.com/google-research/bert/blob/master/multilingual.md. [Online].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. Ethnologue: Languages of the World. Twenty-third edition. SIL International.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

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.

Marco René Spruit, Wilbert Heeringa, and John Nerbonne. 2009. Associations among linguistic levels. *Lingua*, 119(11):1624 – 1642. The Forests behind the Trees.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Warren S Torgerson. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Wietse de Vries and Malvina Nissim. 2020. As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages. *arXiv preprint arXiv:2012.05628*.

Søren Wichmann, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389(17):3632 – 3639.

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.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

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

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Detailed Results

Table 4 shows results per adapted model per training dataset. Dutch POS-tagging accuracy is still relatively high after lexical layer replacement. Similarly, Table 5 shows the POS-tagging performance with subsets of the lexical layer retraining data per training dataset. Training on the Dutch `Alpino` dataset instead of `LassySmall` results in consistently higher performance for both Gronings and West Frisian.

Test language:			**Source**			**Gronings**		**West Frisian**	
Train language:			orig.	gro.	fri.	orig.	gro.	orig.	fri.
EN	GUM	BERT	93.5	13.5	23.5	19.7	55.4	21.0	78.8
		mBERT	93.5	22.0	22.2	61.6	85.0	87.5	88.2
	ParTUT	BERT	**94.0**	16.6	26.4	33.5	67.7	37.1	77.4
		mBERT	**94.0**	41.3	47.6	66.6	84.3	86.7	89.2
DE	GSD	gBERT	92.6	23.3	22.4	31.3	84.2	28.4	89.3
		mBERT	92.2	25.1	22.2	65.9	83.9	87.5	88.3
	HDT	gBERT	**94.0**	28.5	26.2	19.5	86.7	16.9	89.0
		mBERT	93.7	26.1	22.1	45.8	81.1	84.7	83.0
NL	Alpino	BERTje	96.0	90.8	78.1	66.7	**92.4**	50.0	**95.4**
		mBERT	96.2	87.8	82.8	**74.3**	90.5	91.9	95.1
	LassySmall	BERTje	**96.8**	89.6	70.3	63.0	90.9	45.9	95.1
		mBERT	**96.8**	80.4	51.3	70.6	88.1	**92.7**	94.4

Table 4: Accuracy per target language variety (columns) per lexical layer (sub-columns). This is an extended version of Table 1 in the main paper with accuracies separated by POS-tagging training dataset. This table shows that not all datasets are equally effective for transfer to Gronings and West Frisian.

			Gronings						**West Frisian**					
			1MB	5MB	10MB	20MB	40MB	43MB	1MB	5MB	10MB	20MB	40MB	59MB
EN	BERT	GUM	29.2	47.8	66.1	67.1	58.9	55.4	48.0	69.5	76.6	79.8	79.4	78.5
		ParTUT	37.8	55.1	70.4	72.0	67.8	85.0	53.1	70.4	75.9	78.1	77.8	88.7
	mBERT	GUM	19.6	73.5	84.8	84.9	84.8	67.7	69.7	87.1	88.0	88.4	88.5	77.0
		ParTUT	30.0	76.7	84.0	84.2	84.1	84.3	74.3	88.1	88.4	89.7	89.4	89.3
DE	gBERT	GSD	48.8	82.3	83.9	84.0	83.8	84.2	77.7	87.3	88.8	88.5	88.7	89.1
		HDT	30.9	84.5	86.5	87.0	86.3	83.9	73.8	86.3	86.6	87.6	87.1	88.0
	mBERT	GSD	24.0	74.0	82.4	82.4	82.7	86.7	71.1	87.1	87.3	88.1	88.1	89.3
		HDT	03.7	44.2	75.1	72.2	79.5	81.1	34.4	72.0	79.1	78.7	81.2	83.5
NL	BERTje	Alpino	**73.2**	**90.3**	**92.0**	**91.9**	**92.0**	**92.4**	43.5	**94.2**	94.8	**95.1**	**94.9**	**95.4**
		LassySmall	67.0	88.3	90.0	90.2	89.9	90.5	**44.3**	93.6	**94.9**	94.4	94.6	95.0
	mBERT	Alpino	31.0	79.6	89.1	88.5	89.3	90.9	74.9	93.7	93.8	94.5	94.7	94.9
		LassySmall	15.9	57.4	85.0	85.7	86.7	88.1	67.8	91.6	93.0	93.7	94.1	94.2

Table 5: POS-tagging accuracy for Gronings and West Frisian with subsets of the unlabeled lexical layer retraining data. This is an extended version of Table 2 in the main paper with accuracies separated by POS-tagging training dataset.