Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese

Abstract

Multilingual language models have pushed state-of-the-art in cross-lingual NLP transfer. The majority of zero-shot cross-lingual transfer, however, use *one and the same* massively multilingual transformer (e.g., mBERT or XLM-R) to transfer to *all* target languages, irrespective of their typological, etymological, and phylogenetic relations to other languages. In particular, readily available data and models of resource-rich sibling languages are often ignored. In this work, we empirically show, in a case study for Faroese – a low-resource language from a high-resource language family – that by leveraging the phylogenetic information and departing from the 'one-size-fits-all' paradigm, one can improve cross-lingual transfer to low-resource languages. In particular, we leverage abundant resources of other Scandinavian languages (i.e., Danish, Norwegian, Swedish, and Icelandic) for the benefit of Faroese. Our evaluation results show that we can substantially improve the transfer performance to Faroese by exploiting data and models of closely-related high-resource languages. Further, we release a new web corpus of Faroese and Faroese datasets for named entity recognition (NER) and semantic text similarity (STS), in addition to new language models trained on all Scandinavian languages.

1 Introduction

Massively multilingual Transformer-based language models (MMTs) such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) and mT5 (Xue et al., 2021) have been the driving force of modern multilingual NLP, allowing for rapid bootstrapping of language technology for a wide range of low(er)-resource languages by means of (zero-shot or few-shot) cross-lingual transfer from high(er)-resource languages (Lauscher et al., 2020; Hu et al., 2020; Xu and Murray, 2022; Schmidt et al., 2022). Cross-lingual transfer with MMTs is not without drawbacks. MMTs' representation spaces are heavily skewed in favor of high-resource languages, for which they have been exposed to much more data in pretraining (Joshi et al., 2020; Wu and Dredze, 2020); combined with the 'curse of multilinguality' – i.e., limited per-language representation quality stemming from a limited capacity of the model (Conneau et al., 2020; Pfeiffer et al., 2022) – this leads to lower representational quality for languages under-represented in MMTs' pretraining. Cross-lingual transfer with MMTs thus fails exactly in settings in which it is needed the most: for low-resource languages with small digital footprint. Despite these proven practical limitations, the vast majority of work on cross-lingual transfer still relies on MMTs due to their appealing conceptual generality: in theory, they support transfer between any two languages seen in their pretraining. Such strict reliance on MMTs effectively ignores the linguistic phylogenetics and fails to directly leverage resources of resource-rich languages that are closely related to a target language of interest.

In this work, we attempt to mitigate the above limitations for a particular group of languages, departing from the 'one-size-fits-all' paradigm based on MMTs. We focus on a frequent and realistic setup in which the target language is a low-resource languages but from a high-resource language family, i.e., with closely related resource-rich languages. A recent comprehensive evaluation of the languages used in Europe[1] scores languages

[1]The *Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap* published by the European Language Equality Programme (ELE), https://european-language-equality.eu/agenda/.

based on the available resources. Languages such as German and Spanish score at around 0.5 of the English scores, and more than half of the languages are scored below 0.02 of the English score. Many, including almost all regional and minority languages such as Faroese, Scottish Gaelic, Occitan, Luxembourgish, Romani languages, Sicilian and Meänkieli have the score of (almost) 0. However, what differentiates these languages from low-resource languages from Africa (e.g., Niger-Congo family) or indigenous languages of Latin America (e.g., Tupian family) is the fact that they typically have *closely related high-resource languages*. In this case, we believe, LMs of closely related high-resource languages promise more effective transfer compared to using MMTs, plagued by the 'curse of multilinguality', as the vehicle of transfer.

In this proof-of-concept case study, we focus on Faroese as the target language and demonstrate the benefits of *linguistically informed* transfer. We take advantage of available data and resources from the closely related but much more 'NLP-developed' other Scandinavian languages.[2] We show that using "Scandinavian" LMs brings substantial gains in downstream transfer to Faroese compared to using XLM-R as widely used off-the-shelf MMT. The gains are particularly pronounced for the task of semantic text similarity (STS), the only high-level semantic task in our evaluation. We further show that adding limited-size target-language corpus to LM's pretraining corpora brings further gains in downstream transfer. As another contribution of this work, we collect and release: (1) a corpus of web-scraped monolingual Faroese, (2) multiple LMs suitable for Faroese, including those trained on all five Scandinavian languages, and (3) two new task-specific datasets for Faroese labeled by native speakers: for NER and STS.

2 Background and Related Work

Cross-Lingual Transfer Learning with MMTs and Beyond. A common approach to cross-lingual transfer learning involves pretrained MMTs (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2021). These models can be further pretrained for specific languages or directly adapted for downstream tasks. A major downside of the MMTs has been dubbed the *curse of multilinguality* (Conneau et al., 2020), where the model becomes saturated and performance can not be improved further for one language without a sacrifice elsewhere, something which continued pretraining for a given language alleviates (Pfeiffer et al., 2020). Adapter training, such as in (Pfeiffer et al., 2020; Üstün et al., 2022), where small adapter modules are added to pretrained models, has also enabled cost-efficient adaptation of these models. The adapters can then be used to fine-tune for specific languages and tasks without incurring catastrophic forgetting.

Other methods involve translation-based transfer (Hu et al., 2020; Ponti et al., 2021), and transfer from monolingual language models (Artetxe et al., 2020; Gogoulou et al., 2022; Minixhofer et al., 2022). Bilingual lexical induction (BLI) is the method of mapping properties, in particular embeddings, from one language to another via some means such as supervised embedding alignment, unsupervised distribution matching or using an orthogonality constraint (Lample et al., 2018; Søgaard et al., 2018; Patra et al., 2019), and has also been used to build language tools in low-resource languages (Wang et al., 2022).

Attempts to alleviate the abovementioned issues have been made, such as vocabulary extension methods (Pfeiffer et al., 2021), which add missing tokens and their configurations to the embedding matrix. Phylogeny-inspired methods have also been used where adapters have been trained for multiple languages and stacked to align with the language family of the language of interest (Faisal and Anastasopoulos, 2022). Inspired by this line of research, in this work, we focus on improving MMT-based cross-lingual transfer for a particular group of languages, those that have sibling languages with more abundant data and resources.

NLP Resources in Scandinavian Languages. A fair amount of language resources have been developed for the Scandinavian languages, particularly when considered in total. It is also worth mentioning that some Danish, Icelandic, Norwegian and Swedish exist in large crawled datasets, both monolingual ones such as mC4 (Xue et al., 2021) and parallel corpora such as (Schwenk et al., 2021; Agić and Vulić, 2019). Large multilingual language models have been trained over these datasets (Devlin et al., 2019; Liu et al., 2020; Xue et al., 2021)

[2]The Scandinavian languages are a family of Indo-European languages that form the North Germanic branch of the Germanic languages. The largest languages of the family are: (1) Danish (population 5.8M), Norwegian (5.4M) and Swedish (10.4M) – the Mainland Scandinavian languages, and (2) Icelandic (373K) and Faroese (54K) – the Insular Scandinavian languages.

but have been shown to have limited capacity for low resources languages. Faroese is not included (at least not correctly labelled) in these crawled corpora.This may be in part due to the low amount of Faroese found online, and also due to its similarity to related languages (Haas and Derczynski, 2021). A short summary of prior work in relation to transfer learning and Faroese is given in Appendix D.

In this work, we use the open language resources for the Scandinavian languages listed below.

Danish: The Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021) is a billion-word corpus containing a wide variety of text.We also use a NER resource, the DaNE corpus (Hvingelby et al., 2020).

Icelandic: As Icelandic is the most closely related language to Faroese, we experiment with an Icelandic language model, IceBERT (Snæbjarnarson et al., 2022). For the NER experiment, we make use of the MIM-GOLD-NER corpus (Ingólfsdóttir et al., 2020).

Norwegian: The Norwegian Colossal Corpus (NCC) (Kummervold et al., 2022) contains 49GB of clean Norwegian data from a variety of sources, making it the largest such public collection in the Nordics. We also make use of the NorNE (Jørgensen et al., 2020) NER corpus (both for Bokmål and Nynorsk).

Swedish: The Swedish Gigaword Corpus (Eide et al., 2016) is contains text from between 1950 and 2015. The latest NE-corpus for Swedish is SweNERC (Ahrenberg et al., 2020), where the authors include more modern texts than in earlier corpora.

Faroese: A POS corpus, the Sosiualurin corpus is an annotated Newspaper corpus with 102k words (Hansen et al., 2004). The Faroese Wikipedia has also been used to create a tree bank (Tyers et al., 2018), which has a Universal Dependencies (UD) mapping. We use this corpus along with the FarPaHc (Ingason et al., 2012), which also has a UD mapping.

3 New Faroese Datasets

3.1 Faroese Common Crawl Corpus (FC3)

Faroese mono-lingual data is scarce, and there is an explicit upper bound to the amount available due to the small size of the Faroese-speaking population. Despite this, we manage to extract a decent amount of varied Faroese text from the Common Crawl corpus (FC3) following the methods used by (Snæbjarnarson et al., 2022) for Icelandic, by targeting the `.fo` top-level domain. It amounts to 9 m words in Faroese after clean-up and de-duplication, from 98k paragraphs. While relatively small, this allows for significant gains, as shown in §5.

3.2 Named Entity Recognition (FoNE)

We annotate the Sosialurin corpus (6286 lines, 102k words) with named entities following the CoNLL schema using an Icelandic NER-tagger trained using the ScandiBERT model, see §4. The annotation was then manually reviewed. Out of the 118533 tokens (including punctuation), 9001 are annotated using the Date (546), Location (1774), Miscellaneous (332), Money (514), Organization (2585), Percent (115), Person (2947) and Time (188) tags. We refer to this new dataset as *FoNE*.

3.3 Semantic Similarity (FoSTS)

The Stanford Sentiment Treebank (STS) (Socher et al., 2013) is an annotated sentiment corpus containing pairs of sentences along with a real-valued score between 1 and 5 to indicate how similar in meaning two phrases are. 394 sentence pairs from the test set were translated to preserve the score by a native speaker.

4 Training of Models

We train new language models, *ScandiBERT* on all the Scandinavian languages, *ScandiBERT-no-fo* without Faroese, and *DanskBERT* over only the Danish data. The Danish model is trained to be comparable with the IceBERT model. We also consider models that have been further pre-trained on the FC3 corpus. See Appendix A for an overview of the training datasets and hyperparameters used.

5 Experiments

5.1 Evaluation of Models over Faroese Tasks

In addition to the models presented in §4, we make use of the monolingual Icelandic model IceBERT and the MMT XLM-RoBERTa-base (XLM-R). Downstream performance is evaluated over four Faroese tasks; the POS and UD datasets introduced in §2, and the new NER and STS datasets introduced in §3. All models are fine-tuned ten times with different seeds, for 5 epochs using a batch size of 16. The Fo-STS models are the except where performance was degrading at those settings, runs are over 3 epochs using batch size 8. Ten-fold cross-

Model	POS		NER		UD FP		UD oft		STS
	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	Acc.
IceBERT	85.5 ± 0.6	85.2 ± 0.5	87.9 ± 1.7	96.4 ± 0.3	93.6 ± 0.2	94.6 ± 0.1	92.7 ± 1.0	94.2 ± 0.8	74.9 ± 4.0
IceBERT-fc3	90.9 ± 0.2	90.4 ± 0.2	90.9 ± 1.3	98.9 ± 0.1	96.6 ± 0.2	97.1 ± 0.2	95.3 ± 1.2	96.1 ± 1.0	77.5 ± 2.6
DanskBERT	73.4 ± 0.6	74.3 ± 0.5	85.6 ± 1.4	98.4 ± 0.2	86.2 ± 0.5	87.7 ± 0.3	84.8 ± 1.8	88.7 ± 1.4	71.5 ± 4.5
DanskBERT-fc3	87.1 ± 0.4	86.4 ± 0.4	89.7 ± 1.7	98.8 ± 0.2	96.0 ± 0.2	96.6 ± 0.1	94.2 ± 0.9	95.7 ± 0.6	76.2 ± 4.0
XLM-R	84.6 ± 0.9	85.0 ± 0.9	87.8 ± 1.5	96.3 ± 0.2	93.5 ± 0.2	94.3 ± 0.1	91.5 ± 1.4	93.6 ± 1.1	71.7 ± 4.5
XLM-R-fc3	91.2 ± 0.3	91.2 ± 0.3	90.9 ± 1.3	98.9 ± 0.2	97.3 ± 0.2	97.7 ± 0.1	95.7 ± 0.7	96.8 ± 0.6	75.8 ± 5.2
ScandiBERT-no-fo	88.4 ± 0.3	88.1 ± 0.3	89.9 ± 0.8	96.7 ± 0.5	95.9 ± 0.2	96.4 ± 0.2	93.8 ± 1.1	95.0 ± 1.0	79.8 ± 1.8
ScandiBERT-no-fo-fc3	91.5 ± 0.3	91.2 ± 0.3	**91.4 ± 1.1**	**98.8 ± 0.2**	**97.4 ± 0.1**	**97.8 ± 0.1**	**96.3 ± 0.7**	**96.8 ± 0.6**	**82.4 ± 2.9**
ScandiBERT	90.3 ± 0.3	90.0 ± 0.4	90.2 ± 0.9	99.0 ± 0.2	96.5 ± 0.2	97.1 ± 0.1	95.2 ± 1.0	96.2 ± 0.8	70.4 ± 5.3
ScandiBERT-fc3	**91.6 ± 0.2**	**91.3 ± 0.3**	91.0 ± 1.1	99.0 ± 0.1	97.3 ± 0.2	97.7 ± 0.2	95.9 ± 0.8	96.7 ± 0.7	64.8 ± 9.5

Table 1: Results for all downstream tasks in Faroese using the different base models, with and without continued Faroese pre-training. The -fc3 postfix indicates models that were further pre-trained on FC3.

validation is used to report the numbers. Results are shown in Table 1.

POS: The best-performing model for POS (evaluated over the Sosialurin POS corpus) is ScandiBERT-fc3, outperforming the ScandiBERT model for more than a whole F1 point. We note, though, that the ScandiBERT-no-fo-fc3 model is very close in performance.

NER, UD, STS: The best-performing model for NER (evaluated over FoNE), UD (evaluated over the FarPaHC and OFT UD corpora), and Fo-STS is the ScandiBERT-no-fo-fc3 model. Interestingly, it is better not to include Faroese in the mix, to begin with (and not use Faroese to train the tokenizing vocabulary)We hypothesize that this is since including Faroese in the vocabulary results in less subword overlap with the other Scandinavian languages, limiting the potential for transfer.

5.2 Additional Experiments

Transfer with Wechsel. To put our work in further context, beyond comparison to MMTs, we consider an alternative transfer learning approach, the Wechsel method (Minixhofer et al., 2022), a recent well-performing method for transferring monolingual Transformers to a new language. Further details and results are presented in Appendix B: they all show far worse performance than those presented in Table 1. We hypothesize this is due to how closely related the languages we consider are, as opposed to the distant languages considered in the original Wechsel work.

Task-Specific Transfer. To explore the potential for task-specific transfer between closely related languages, we consider if labelled Scandinavian datasets can be combined to benefit Faroese. In particular, we look at NER as there is an easy way to map between labels of the different languages. See Appendix C for more details. The best result is achieved when training directly from the IceBERT model, which has been trained on the large MIM-GOLD-NER dataset, showing that given enough resources and a close enough language model, such a direct approach can be the most effective.

6 Discussion

Some of the results in Table 1 are as expected. Starting from the closest relative, the Icelandic model, IceBERT, results in better performance for all downstream tasks than starting with the Danish model DanskBERT. The ScandiBERT model performs better than the massively multilingual XLM-R on all tasks, bar the more semantic FO-STS task, though the standard deviation is quite high for both models.

What is more interesting is that the ScandiBERT-no-fo model that is not trained on Faroese outperforms the model that has Faroese included, when fine-tuned further on the FC3 dataset. In particular, for the higher level Fo-STS task. We hypothesize that this forces the Faroese adaptation to use the word segmentations from the related languages for a higher transfer benefit, as the tokenizing vocabulary was trained without Faroese. This is something we hope to see explored in future work.

7 Conclusion

We have shown that by leveraging phylogenetic information and departing from the 'one-size-fits-all' paradigm, one can improve cross-lingual transfer to low-resource languages. Our evaluation results show that we can substantially improve the transfer performance to Faroese by exploiting data and models of closely-related high-resource languages, as opposed to relying on MMTs.

We make the models *ScandiBERT*, *DanskBERT*, *ScandiBERT-no-fo*, with and without continued training over FC3 available. As well as the new datasets *FC3*, *FoNE* and *Fo-STS*.

References

Željko Agić and Ivan Vulić. 2019. https://doi.org/10.18653/v1/P19-1310 JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Lars Ahrenberg, Johan Frid, and Leif-Jöran Olsson. 2020. A new resource for swedish named-entity recognition.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. https://doi.org/10.18653/v1/2020.acl-main.421 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. https://doi.org/10.18653/v1/2020.acl-main.747 Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. https://doi.org/10.18653/v1/N19-1423 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.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The swedish culturomics gigaword corpus: A one billion word swedish reference dataset for nlp.

Fahim Faisal and Antonios Anastasopoulos. 2022. https://aclanthology.org/2022.aacl-main.34 Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Mikel L. Forcada and Francis M. Tyers. 2016. https://aclanthology.org/2016.eamt-2.4 Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2022. https://aclanthology.org/2022.lrec-1.100 Cross-lingual transfer of monolingual models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 948–955, Marseille, France. European Language Resources Association.

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

René Haas and Leon Derczynski. 2021. https://aclanthology.org/2021.vardial-1.8 Discriminating between similar nordic languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75, Kiyv, Ukraine. Association for Computational Linguistics.

Hinrik Hafsteinsson and Anton Karl Ingason. 2021. https://aclanthology.org/2021.nodalida-main.33 Towards cross-lingual application of language-specific PoS tagging schemes. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 321–325, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Zakaris Svabo Hansen, Heini Justinussen, and Mortan Ólason. 2004. Marking av teldutøkum tekstsavni [tagging of a digital text corpus].

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

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. https://aclanthology.org/2020.lrec-1.565 DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Anton Karl Ingason, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel C. Wallenberg. 2012. http://hdl.handle.net/20.500.12537/92 Faroese parsed historical corpus (FarPaHC) 0.1. CLARIN-IS.

Svanhvít L. Ingólfsdóttir, Ásmundur A. Guðjónsson, and Hrafn Loftsson. 2020. Named entity recognition for icelandic: Annotated corpus and models. In *Statistical Language and Speech Processing*, pages 46–57, Cham. Springer International Publishing.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. https://aclanthology.org/2020.lrec-1.559 NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. https://doi.org/10.18653/v1/2020.acl-main.560 The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. https://doi.org/10.18653/v1/D18-2012 SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. https://aclanthology.org/2022.lrec-1.410 The Norwegian colossal corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. https://openreview.net/forum?id=H196sainb Word translation without parallel data. In *International Conference on Learning Representations*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.363 From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. https://doi.org/10.1162/tacl_a_00343 Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. http://arxiv.org/abs/1907.11692 Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. https://doi.org/10.18653/v1/2022.naacl-main.293 WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. https://doi.org/10.18653/v1/N19-4009 fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. https://doi.org/10.18653/v1/P19-1018 Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. https://doi.org/10.18653/v1/2022.naacl-main.255 Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.617 MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. https://doi.org/10.18653/v1/2021.emnlp-main.800 UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulic, and Siva Reddy. 2021. http://arxiv.org/abs/2107.11353

Modelling latent translations for cross-lingual transfer. *CoRR*, abs/2107.11353.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. https://aclanthology.org/2022.emnlp-main.736 Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. https://doi.org/10.18653/v1/2021.acl-long.507 CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. https://aclanthology.org/2022.lrec-1.464 A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. https://aclanthology.org/D13-1170 Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. https://doi.org/10.18653/v1/P18-1072 On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. https://aclanthology.org/2021.nodalida-main.46 The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. https://doi.org/10.18653/v1/W18-6017 Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150, Brussels, Belgium. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. https://aclanthology.org/2022.emnlp-main.541 Hyper-X: A unified hypernetwork for multi-task multilingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. https://doi.org/10.18653/v1/2022.acl-long.61 Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6 Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. https://doi.org/10.18653/v1/2020.repl4nlp-1.16 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.

Haoran Xu and Kenton Murray. 2022. https://doi.org/10.18653/v1/2022.findings-naacl.157 Por qué não utiliser alla språk? mixed training with gradient optimization in few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2043–2059, Seattle, United States. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021.

https://doi.org/10.18653/v1/2021.naacl-main.41
mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Training of language models

Language	Datasets	Size
Icelandic	IGC / IC3 / Skemman / Hirslan	16 GB
Danish	Danish Gigaword Corpus (incl. Twitter)	4,7 GB
Norwegian	NCC corpus	42 GB
Swedish	Swedish Gigaword Corpus	3,4 GB
Faroese	FC3 + Sosialurinn + Bible	69 MB

Table 2: Datasets used to train ScandiBERT, ScandiBERT-no-fo and DanskBERT

We train new BPE vocabularies for all the new models we train, ScandiBERT, ScandiBERT-no-fo and DanskBERT. All models use the same vocabulary size of 50k. The ScandiBERT vocabulary is trained using all the languages, the ScandiBERT-no-fo vocabulary is trained without the Faroese data, and the DanskBERT vocabulary is only trained on the danish text. Vocabularies are trained using the SentencePiece software (Kudo and Richardson, 2018), and character coverage is set to 99.995 %.

Pre-training of the new language models is done using `fairseq` (Ott et al., 2019) using the RoBERTa-base (Liu et al., 2019) configuration, fine-tuning is done using the `transformers` (Wolf et al., 2020) library. ScandiBERT and ScandiBERT-no-fo were trained for 72 epochs, using a batch size of 8.8k sequences on 24 NVIDIA V100 cards for approximately 14 days each. Initial testing showed that the larger batch size showed better performance than going for around 2k sequences, possibly due to the mixture of differing languages. DanskBERT, on the other hand, similar to IceBERT and RoBERTa showed better performance at the smaller batch size. DanskBERT was trained to convergence for 500k steps using 16 V100 cards for approximately 14 days.

All *-fc* models are further trained for 50 epochs, with an effective batch size of 100k tokens for 12k updates, over the FC3 dataset for Faroese adaptation.

An overview of the data used to train the language models is shown in Table 2. For details on the Icelandic data, we refer to (Snæbjarnarson et al., 2022). For the other datasets, we refer to §2.

B Wechsel results

We compare our method to another transfer learning approach presented by Minixhofer et al. (2022). The FC3 dataset is used to train fastText embeddings for Faroese, and the Icelandic datasets are used to train fastText embeddings for Icelandic. These embeddings are then used to convert the multilingual models to Faroese using the Wechsel approach. We confirm the quality of the Icelandic embeddings by running an Icelandic semantic evaluation suite adapted from `https://github.com/stofnun-arna-magnussonar/ordgreypingar_embeddings`, showing our embeddings are comparable or of higher quality than those released by Meta (Grave et al., 2018).

The experiments in Table 3 all show sub-par performance compared to the results in non-Wechsel results in Table 1. The Wechsel work considers transfer from English-dominant models, GPT2 and RoBERTa to French, German, Chinese, Swahili, Sundanese, Scottish Gaelic, Uyghur and Malagasy. None of which are closely related to English. One reason for the discrepancy in the results could be that the shuffling of the embedding matrix to convert it is more catastrophic when considering close languages. Another reason could be that both Faroese and Icelandic are morphologically rich and that all variants of the words were not properly mapped during the conversion of the embedding matrix.

C Mapping NER datasets

The datasets used to create a Scandinavian NER-corpus are DaNE (Danish), FoNE (Faroese), MIM-GOLD-NER (Icelandic), NorNE (Norwegian) and SWE-Nerc (Swedish), presented in §2. The results in Table 4 show that the best result is obtained when training directly from the IceBERT model. The ScandiBERT model has a higher variance when pre-fine-tuned on the combined NER corpora. This approach could also be made directly for the UD corpus, POS (in particular using the Icelandic POS data) and other corpora as they become available for training or evaluation in Faroese. This demonstrates how resources from a related language can substantially benefit a low-resource language.

To combine the NER datasets, we map the tags to the CoNLL schema used by the Icelandic MIM-

	Model	POS F1	POS Acc.	NER F1	NER Acc.	UD FP F1	UD FP Acc.	UD oft F1	UD oft Acc.	STS Acc.
Wechsel	IceBERT	74.4 ± 0.5	75.7 ± 0.5	67.7 ± 3.8	98.7 ± 0.2	83.6 ± 1.1	84.6 ± 1.2	66.6 ± 28.5	75.7 ± 18.6	37.6 ± 13.4
	IceBERT-fc3	89.0 ± 0.2	89.4 ± 0.3	88.5 ± 1.5	96.4 ± 0.3	96.3 ± 0.1	96.5 ± 0.1	95.6 ± 0.9	96.2 ± 0.8	56.4 ± 19.0
	XLM-R	68.9 ± 0.5	73.5 ± 0.4	59.7 ± 2.9	99.0 ± 0.2	81.0 ± 0.6	84.5 ± 0.4	71.8 ± 2.3	79.7 ± 1.4	29.0 ± 4.9
	XLM-R-fc3	86.8 ± 0.3	88.7 ± 0.3	88.8 ± 1.3	98.4 ± 0.2	96.3 ± 0.1	96.7 ± 0.1	95.6 ± 0.8	96.5 ± 0.6	60.6 ± 4.2
	ScandiBERT-no-fo	71.3 ± 0.5	72.5 ± 0.5	65.1 ± 1.7	98.8 ± 0.1	82.0 ± 0.6	83.4 ± 0.6	75.0 ± 2.1	81.0 ± 1.8	43.2 ± 6.8
	ScandiBERT-no-fo-fc3	89.2 ± 0.2	89.6 ± 0.2	89.2 ± 1.7	99.0 ± 0.1	96.8 ± 0.2	97.1 ± 0.2	96.1 ± 0.9	96.8 ± 0.7	74.2 ± 3.2
	ScandiBERT	72.6 ± 0.9	73.8 ± 0.9	65.7 ± 1.7	98.8 ± 0.1	83.0 ± 1.2	84.0 ± 0.9	76.8 ± 1.6	82.5 ± 1.1	29.3 ± 11.5
	ScandiBERT-fc3	89.3 ± 0.3	89.7 ± 0.3	88.8 ± 1.7	98.7 ± 0.2	96.8 ± 0.1	97.1 ± 0.1	96.0 ± 0.8	96.7 ± 0.8	57.6 ± 18.1

Table 3: Results for all downstream tasks using different base models after Wechsel adaptation, with and without continued Faroese pre-training. The results are significantly worse than without Wechsel adaptations.

GOLD-NER and the Faroese FoNE datasets. The Danish DaNE dataset uses a subset of the tags used for Icelandic and Faroese, so the mapping is purely nominal. The mapping for Norwegian (NorNE) and Swedish (SweNERC) datasets is shown in Table 5.

D Prior Work on Transfer learning for Faroese

We know of two prior works that consider transfer learning for Faroese from the Scandinavian languages. In (Tyers et al., 2018), a rule-based translation system (Apertium (Forcada and Tyers, 2016)) is used to translate the Faroese Wikipedia into Swedish, Norwegian Bokmål, and Norwegian Nynorsk. The translations are then aligned, and the translations dependency-parsed. The resulting trees are then mapped to the original Faroese sentences and used for POS-tagging and annotating morphological features. The second work is a mapping between Faroese and Icelandic POS-tags (Hafsteinsson and Ingason, 2021); while not a direct application, the authors suggest the mapping may be of use for transfer learning between the languages.

Model	Pre-ft.	Ft.	F1	Acc.
SB-no-fo-fc3	None	Yes	91.4 ± 1.1	98.8 ± 0.2
ScandiBERT	Icel.	Yes	**92.0 ± 1.0**	98.8 ± 0.2
ScandiBERT	All	No	91.5 ± 1.6	98.9 ± 0.2
ScandiBERT	All	Yes	91.8 ± 1.6	99.0 ± 0.2
XLM-R	All	No	90.6 ± 0.6	99.0 ± 0.1
XLM-R	All	Yes	90.8 ± 1.5	99.0 ± 0.2

Table 4: NER performance when models are pre-finetuned on all Scandinavian datasets and then fine-tuned on FoNER.

Language	Original	Mapped
Norwegian	O	O
Norwegian	PER	Person
Norwegian	ORG	Organization
Norwegian	GPE_LOC	Location
Norwegian	PROD	Miscellaneous
Norwegian	LOC	Location
Norwegian	GPE_ORG	Organization
Norwegian	DRV	O
Norwegian	EVT	Miscellaneous
Norwegian	MISC	Miscellaneous
Swedish	O	O
Swedish	EVN	Miscellaneous
Swedish	GRO	Organization
Swedish	LOC	Location
Swedish	MNT	Miscellaneous
Swedish	PRS	Person
Swedish	TME	Time
Swedish	WRK	Miscellaneous
Swedish	SMP	Miscellaneous

Table 5: Mapping of tags to create a unified NER dataset for the Scandinavian languages.