# Evaluation and Adaptation of Language Models for Under-Resourced Languages

Wietse de Vries

university of groningen

# Overview

- Part 1: Do language models actually model Dutch during pre-training?
- Part 2: How well do language models perform on various Dutch tasks?
- Part 3: Can we adapt English models to Dutch and Italian with little training?
- Part 4: Can we adapt models to low-resource languages without labeled data?
- Part 5: How does cross-lingual training work with any source and target language?

# Probing BERT's layers for a Dutch NLP pipeline

**de Vries, W.**, van Cranenburgh, A., and Nissim, M. (2020). What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

# Introduction

- Diagnostic probing has revealed a pipeline-like behaviour for English BERT (Tenney et al. 2019)
  - Simple models trained on hidden transformer layer representations
- E.g. low level tasks like POS tagging can be found in early layers and higher-level tasks like coreference resolution at later layers
- Is this pipeline actually this neat?
- Can this behaviour be found for other languages such as Dutch?

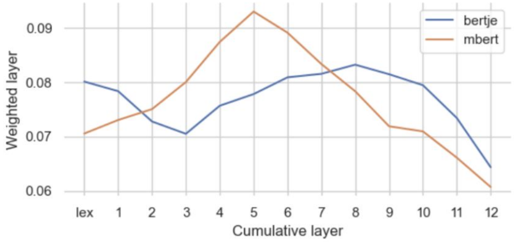
# Methodology

- **Simple probes**: token label prediction with a linear model using hidden layer representations
- **Scalar mixing probes**: use a weighted sum of all hidden layers and evaluate the learned layer weights
- **Models**: BERTje (Dutch) and mBERT

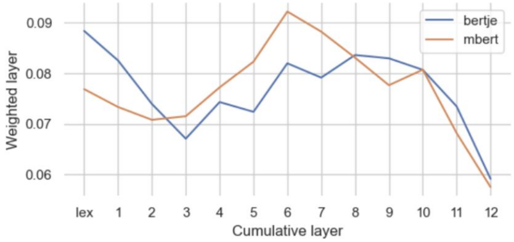
- **POS tagging (POS)**
  - Lassy Small corpus
  - Alpino corpus
- **Dependency edge labeling (DEP)**
  - Lassy Small corpus
  - Alpino corpus
- **Named Entity Recognition (NER)**
  - CoNLL-2002
- **Coreference resolution (Coref)**
  - SoNaR-1

# Scalar mixing results

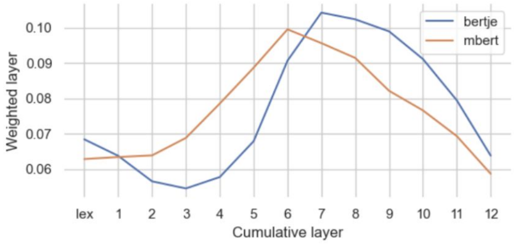
- Scalar mixing probes show higher accuracies than single-layer probes
- mBERT most informative layers are more central
- Word embeddings are more informative for BERTje
- Final layer is relatively uninformative



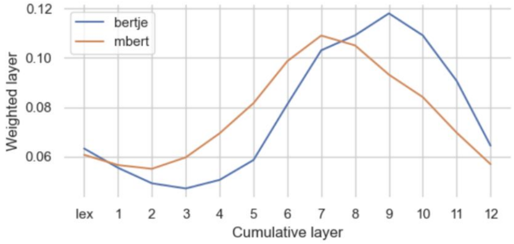

**(a)** UDLassy POS



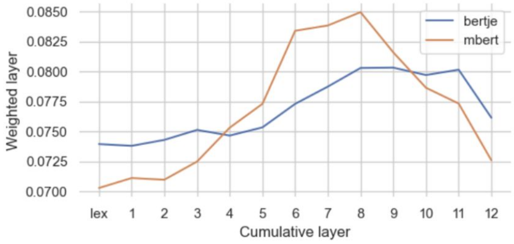

**(b)** UDAlpino POS



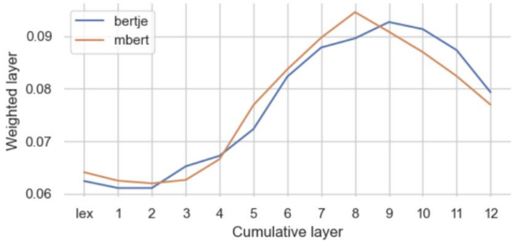

**(c)** UDLassy DEP




**(d)** UDAlpino DEP

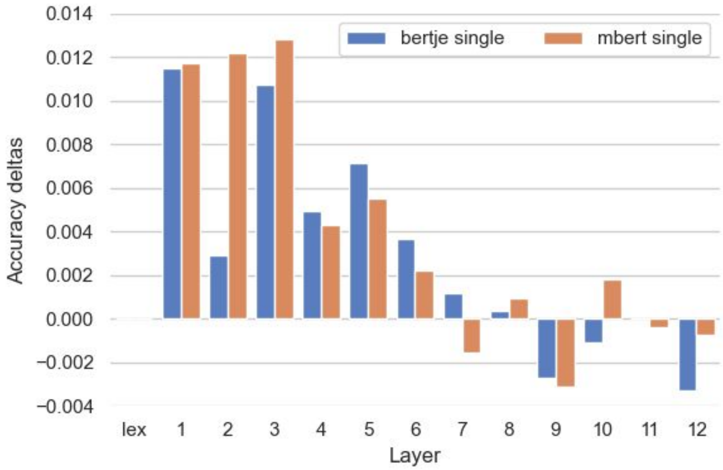


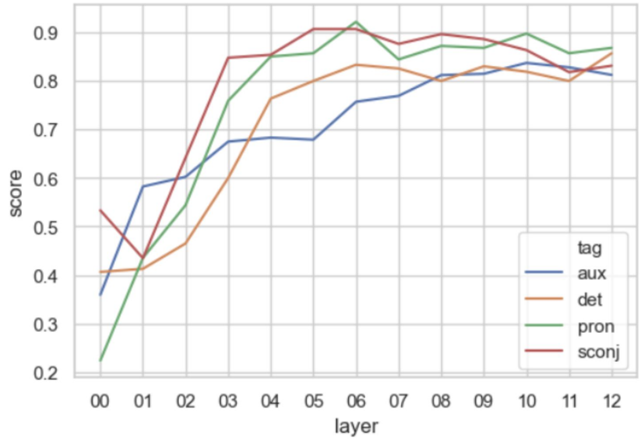

**(e)** CoNLL-2002 NER

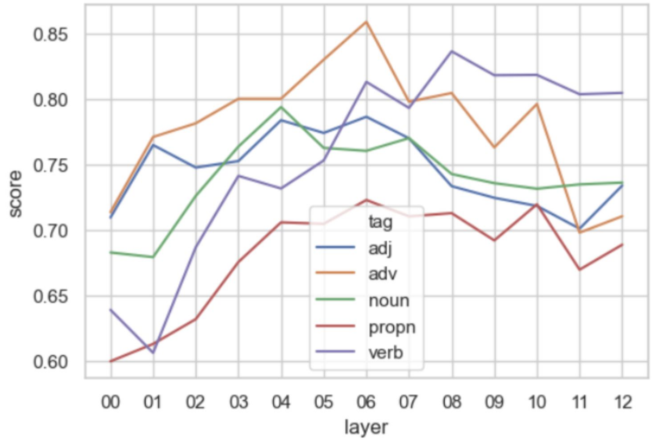



**(f)** SoNaR Coref

# Label differences within one task: POS tagging (BERTje; single layers)



# Conclusions

- BERTje and mBERT show a similar pipeline structure for Dutch as BERT for English but task differences are not very strong
- The most informative mBERT layers are earlier layers than those of BERTje
- Task information is spread out over multiple layers
  - Rule of thumb: the word embeddings and the layers at 2/3 of the model may be most informative
- BERTje shows consistent results across datasets
- More general: task-specific information is learned during pre-training

EMNLP 2023

# DUMB: A Dutch Model Benchmark

**de Vries**, **W.**, Wieling, M., and Nissim, M. (2023). DUMB: A benchmark for smart evaluation of Dutch models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7221–7241, Singapore. Association for Computational Linguistics.

# Introduction

- There are multiple Dutch and multilingual pre-trained language models
  - Unclear which model is most useful for which task
  - New models tend to be re-trained models of the same type: RobBERT-v1, RobBERT-v2, RobBERT-2022, RobBERT-2023…
- English (and other monolingual) benchmarks such as GLUE are not perfect:
  - Task duplication (e.g. 4/10 tasks in GLUE are just Natural Language Inference)
  - Averaging absolute scores undervalue improvement of already high scores
- Our benchmark:
  - 9 tasks of which 4 not previously available in Dutch
  - A different task scoring method: Relative Error Reduction

# Tasks

- Word tasks:
  - **Part-Of-Speech tagging (POS)**: New standardized train/dev/test splits with Lassy Small corpus
  - **Named Entity Recognition (NER)**: New standardized train/dev/test splits with SoNaR-1 corpus
- Word pair tasks:
  - **Word Sense Disambiguation (WSD)**: New Words in Context (WiC) task based on DutchSemCor
  - **Pronoun Resolution (PR)**: New task data based on coreference annotations in SemEval 2010 Task 1
- Sentence pair tasks:
  - **Causal Reasoning**: Choice of Plausible Alternatives (COPA) translated from English to Dutch
  - **Natural Language Inference (NLI)**: Existing SICK-NL dataset (translated SICK from English)
- Document tasks:
  - **Sentiment Analysis (SA)**: Existing Dutch Book Reviews Dataset (DBRD)
  - **Abusive Language Detection (ALD)**: Existing Dutch Abusive Language Corpus (DALC)
  - **Question Answering (QA)**: Translated SQuAD (v2) from English to Dutch

# Evaluation metric: Relative Error Reduction

- Problem with normal averaging:
  - Absolute score differences are weighted equally for every task
  - An accuracy improvement from 50% to 55% has the same effect on the average as 90% to 95%
  - My assumption: a small absolute improvement on a high score can be very meaningful
- Solution: Evaluate on Relative Error Reduction
  - E.g. 50% to 55% is only a 10% error reduction while 90% to 95% is a 50% error reduction
- In our benchmark, we use the BERTje model as a baseline for all other models

# Models

- Only transformer encoder models
- Three model types:
  - BERT (MLM + Sentence pair task)
  - RoBERTa (MLM)
  - DeBERTaV3 (ELECTRA-style generator-discriminator)
- Two model sizes:
  - Base: 12 layers (768 dimensions)
  - Large: 24 layers (1024 dimensions)
- Three pre-training language groups:
  - Dutch
  - Multilingual (including Dutch)
  - English

| **Model** |
| --- |
| BERTje |
| $\text{RobBERT}_{\text{V1}}$ |
| $\text{RobBERT}_{\text{V2}}$ |
| $\text{RobBERT}_{\text{2022}}$ |
| $\text{mBERT}_{\text{cased}}$ |
| $\text{XLM-R}_{\text{base}}$ |
| $\text{mDeBERTaV3}_{\text{base}}$ |
| $\text{XLM-R}_{\text{large}}$ |
| $\text{BERT}_{\text{base}}$ |
| $\text{RoBERTa}_{\text{base}}$ |
| $\text{DeBERTaV3}_{\text{base}}$ |
| $\text{BERT}_{\text{large}}$ |
| $\text{RoBERTa}_{\text{large}}$ |
| $\text{DeBERTaV3}_{\text{large}}$ |

# Results

| | | Word | | Word Pair | | Sent. Pair | | Document | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Avg** | **POS** | **NER** | **WSD** | **PR** | **CR** | **NLI** | **SA** | **ALD** | **QA** |
| BERTje | 0 | 0 | 0 | 0 | **0** | 0 | 0 | 0 | 0 | 0 |
| $\text{RobBERT}_{\text{V1}}$ | -16.3 | 12.5 | -19.4 | -15.3 | -24.0 | -14.7 | -12.7 | -58.2 | 4.8 | -19.4 |
| $\text{RobBERT}_{\text{V2}}$ | 1.6 | 16.2 | 4.1 | -5.3 | **0.1** | -10.2 | -3.8 | -0.5 | 12.0 | 2.2 |
| $\text{RobBERT}_{2022}$ | 3.6 | 17.3 | 7.6 | -6.4 | **-1.8** | -10.1 | 3.1 | 4.0 | 18.9 | -0.2 |
| $\text{mBERT}_{\text{cased}}$ | -5.8 | 6.2 | 9.2 | 7.7 | -11.0 | -18.4 | -6.2 | -41.7 | -4.5 | 6.9 |
| $\text{XLM-R}_{\text{base}}$ | -0.3 | 13.9 | 10.8 | 1.9 | -16.2 | -26.8 | 2.0 | -3.6 | 3.4 | 12.3 |
| $\text{mDeBERTaV3}_{\text{base}}$ | 12.8 | 18.2 | 17.2 | 10.8 | -20.8 | 19.7 | **25.2** | 3.3 | 12.4 | 29.2 |
| $\text{XLM-R}_{\text{large}}$ | 14.4 | **26.5** | **29.7** | **21.3** | -15.8 | -25.8 | 24.4 | **13.2** | **19.0** | 37.2 |
| $\text{BERT}_{\text{base}}$ | -42.8 | -19.8 | -30.8 | -22.4 | -18.7 | -28.0 | -19.2 | -203.9 | -16.1 | -26.2 |
| $\text{RoBERTa}_{\text{base}}$ | -25.6 | -6.5 | -27.3 | -14.0 | -20.4 | -24.1 | -19.7 | -99.9 | -16.0 | -2.1 |
| $\text{DeBERTaV3}_{\text{base}}$ | -1.6 | 6.5 | 1.7 | -4.2 | -25.3 | -20.5 | 8.6 | -14.6 | 3.5 | 29.7 |
| $\text{BERT}_{\text{large}}$ | -35.1 | -12.0 | -25.9 | -25.4 | -29.3 | -31.2 | -15.4 | -158.7 | -7.8 | -10.4 |
| $\text{RoBERTa}_{\text{large}}$ | -14.1 | 6.4 | -12.3 | -19.8 | -23.3 | -26.1 | -8.5 | -63.8 | 1.2 | 19.7 |
| $\text{DeBERTaV3}_{\text{large}}$ | **15.7** | 17.9 | 10.9 | 12.7 | -14.4 | **35.4** | 24.1 | -6.4 | 12.5 | **48.4** |

# Correlations between tasks

| | POS | NER | WSD | PR | CR | NLI | SA | ALD | QA |
|---|---|---|---|---|---|---|---|---|---|
| **POS** | - | 0.85 | 0.75 | 0.31 | 0.43 | 0.77 | **0.89** | **0.93** | 0.66 |
| **NER** | 0.85 | - | **0.92** | 0.41 | 0.42 | **0.88** | 0.87 | 0.81 | 0.75 |
| **WSD** | 0.75 | **0.92** | - | 0.35 | 0.52 | 0.86 | 0.77 | 0.64 | 0.75 |
| **PR** | 0.31 | 0.41 | 0.35 | - | 0.29 | 0.15 | 0.50 | 0.38 | -0.03 |
| **CR** | 0.43 | 0.42 | 0.52 | 0.29 | - | 0.64 | 0.48 | 0.47 | 0.51 |
| **NLI** | 0.77 | 0.88 | 0.86 | 0.15 | **0.64** | - | 0.74 | 0.79 | **0.87** |
| **SA** | 0.89 | 0.87 | 0.77 | 0.50 | 0.48 | 0.74 | - | 0.82 | 0.66 |
| **ALD** | **0.93** | 0.81 | 0.64 | 0.38 | 0.47 | 0.79 | 0.82 | - | 0.59 |
| **QA** | 0.66 | 0.75 | 0.75 | -0.03 | 0.51 | 0.87 | 0.66 | 0.59 | - |
| | 0.70 | 0.74 | 0.70 | 0.30 | 0.47 | 0.71 | 0.72 | 0.68 | 0.59 |

# Missing models: A lot of room for improvement

- Dutch pre-training is better than multilingual, which is better than English
- Large models perform better than smaller
- DeBERTaV3 models are better than RoBERTa and BERT
- More information and a leaderboard can be found on dumbench.nl

| | Dutch | | Multilingual | | English | |
|---|---|---|---|---|---|---|
| | *base* | *large* | *base* | *large* | *base* | *large* |
| BERT | 0 | 4.3 $^{9.6}$ | -5.8 | 2.8 $^{8.1}$ | -42.8 | -35.1 |
| RoBERTa | 3.6 | 13.4 $^{7.8}$ | -0.3 | 14.4 | -25.6 | -14.1 |
| DeBERTaV3 | 24.1 $^{8.1}$ | 38.0 $^{10.8}$ | 12.8 | 36.4 $^{8.6}$ | -1.6 | 15.7 |

Findings of ACL 2021

# Recycling GPT-2 for Dutch and Italian

**de Vries, W.** and Nissim, M. (2021). As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

# Introduction

- English models can be effective for Dutch
- At the time of this research, there was no generative Dutch model
- Can GPT-2 generate Dutch and Italian without training the transformer layers?
- Word embedding / Lexical layer retraining for Dutch and Italian
    - The lexical layer is the layer that maps hidden representations to the byte pair encoding vocabulary

# Method

- Unlabeled data from Wikipedia, web scraped data, newspapers and books
- Train GPT-2 (small) with randomly initialized word embeddings and frozen transformer layers
- Result: separate new word embeddings for Dutch and Italian that should be compatible with the English transformer model

# Sanity check: word embedding alignment

- Dutch/Italian word embeddings should have similar embeddings as literal translations in English
- This is actually true!

| English | Italian | Dutch |
|---|---|---|
| while | mentre | terwijl |
| genes | geni | genen |
| clothes | vestiti | kleren |
| musicians | composi[...] | artiesten |
| permitted | ammessa | toegelaten |
| Finally | infine | Eindelijk |
| satisfied | soddisfatto | tevreden |
| *Accuracy:* | 85% | 89% |

**Table 4.1** | Alignment of closest tokens in the lexical embeddings of $\texttt{sml}_{\texttt{rle}}$ for Italian and Dutch. Accuracy scores are based on a manual evaluation by the authors of 200 randomly selected aligned tokens.

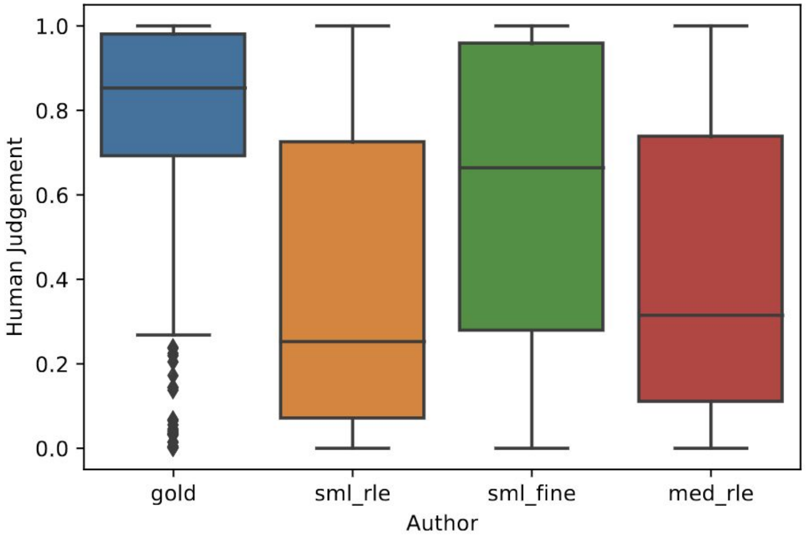
# Scaling to larger models by using alignments

- We have aligned GPT-2 word embeddings for English/Dutch/Italian
- A transformation that converts GPT-2 small to GPT-2 medium embeddings can be applied to the Dutch/Italian embeddings
- Transformation strategies:
  - Linear regression (lstsq; least-squares regression)
  - Orthogonal Procrustes (proc)
  - Weighted K-Nearest Neighbors (knn)

| | Italian | | Dutch | | |
|---|---|---|---|---|---|
| **Model** | Int@1k | PPL | Int@1k | PPL | PPL (1 epoch) |
| $\mathtt{med}_{\mathtt{rle}}$ (1 epoch) | 0.38 | - | 185.02 | - | - |
| $\mathtt{sml}_{\mathtt{rle}} \xrightarrow{proc} \mathtt{med}$ | **0.61** | $8.12 \times 10^{12}$ | **0.61** | $5.02 \times 10^{12}$ | 52.69 |
| $\mathtt{sml}_{\mathtt{rle}} \xrightarrow{lstsq} \mathtt{med}$ | 0.56 | **364.06** | 0.56 | **293.61** | **47.57** |
| $\mathtt{sml}_{\mathtt{rle}} \xrightarrow{1-nn} \mathtt{med}$ | 0.37 | 2,764.19 | 0.36 | 1,101.59 | 50.25 |
| $\mathtt{sml}_{\mathtt{rle}} \xrightarrow{10-nn} \mathtt{med}$ | 0.37 | 20,715.80 | 0.35 | 11,871.66 | 56.88 |

**Table 4.3** | Scores for different transformation methods. Int@1K are the average 1k nearest English neighbors intersection (int) fractions between `sml` and transformed `med` embeddings. *PPL* is the perplexity on the test sets for Italian and Dutch. *PPL (1 epoch)* indicates the perplexity after one epoch of training, which is low if the transformed embeddings were close to a good local optimum.

# Quality: Quite good but with anglicisms




**(b)** Human judgment scores for Dutch texts.

| **Italian** | **Literal English translation** |
|---|---|
| La prima parte del film venne *distribuito* in Giappone con l'aggiunta della colonna sonora. | The first part of the film was *distributed* in Japan with the addition of the soundtrack. |
| L'unico motivo *di la* mia insoddisfazione fu il fatto che l'inizio della sua attività [...] | The only reason *of the* my unsatisfaction was the fact that the beginning of-the his/her activity [...] |
| Il suo nome deriva da un vocabolo arabo. | The his/her name derives from a word Arabic. |
| **Dutch** | **Literal English translation** |
| In een artikel in de Journal of Economicologie (1998), *The New York Times schrijft*: | In an article in the Journal of Economicology (1998), *The New York Times writes*: |
| Ik kan me niet voorstellen dat mensen van mijn generatie *zijn zo boos op mij te wachten*. | I can me not imagine that people of my generation *are so mad at me to wait*. |
| Ik heb niets gedaan om mijn moeder te helpen. | I have nothing done to my mother to help. |

**Table 4.2** | A selection of generated sentences by the `sml` model with Italian and Dutch lexical embeddings. Words or phrases marked in italics are ungrammatical in the target language.

# Conclusion

- GPT-2 can be adapted to Dutch and Italian with only word embedding retraining
- However, extra full model fine-tuning is needed for better performance
- This cheaper adaptation generates the same quality of Italian as an Italian model of the same size trained from scratch (with more data and much longer training)
- We did not find a meaningful difference between Dutch and Italian as target languages

Findings of ACL 2021

# Adapting monolingual models to low-resource languages

**de Vries*, W.**, Bartelds*, M., Nissim, M., and Wieling, M. (2021). Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

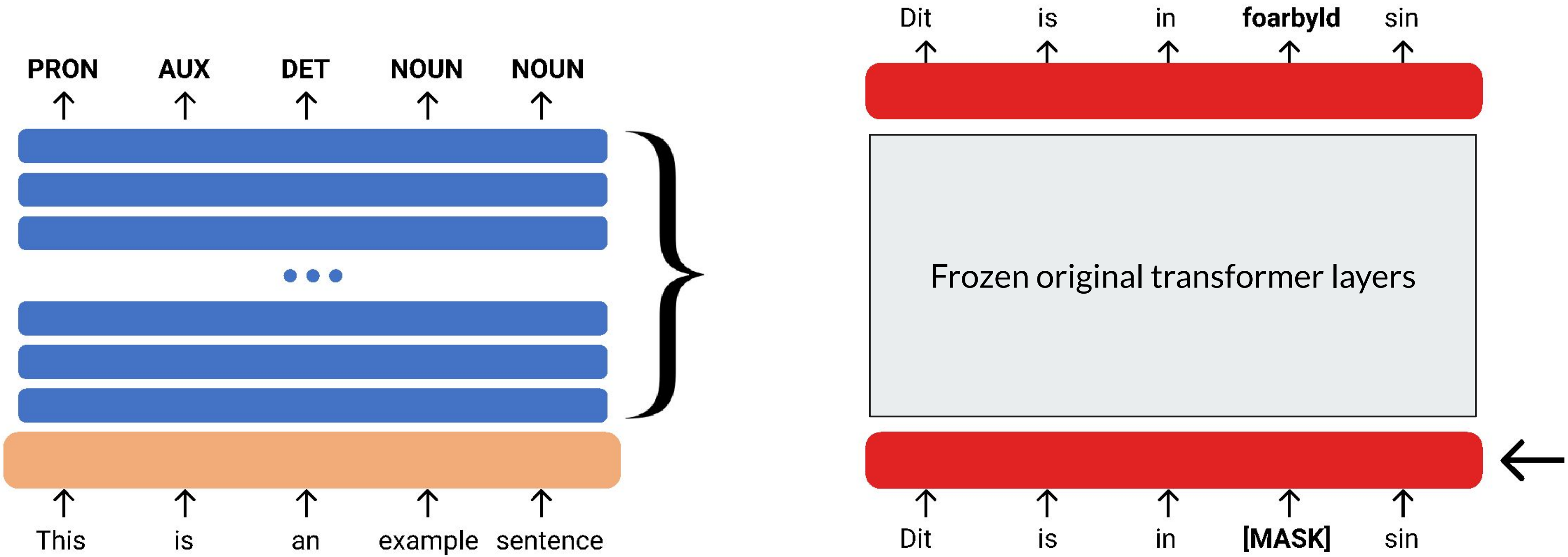
** equal contribution*

# Introduction

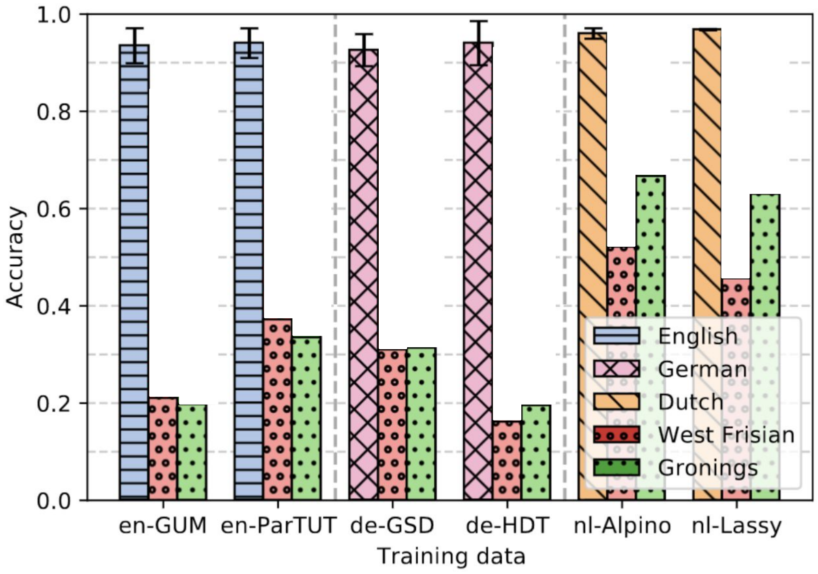
- Word embedding retraining can be effective, but can we use that for real low-resource languages?
- Target languages: **Gronings** (Low Saxon) and **Frisian**
- Source languages: **Dutch**, **German** and **English**
  - All languages are germanic languages, Frisian and Gronings are most similar to Dutch
- Independent word embedding retraining and Transformer layer fine-tuning
- Tested with monolingual BERT models and mBERT

| | |
|---|---|
| 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. |

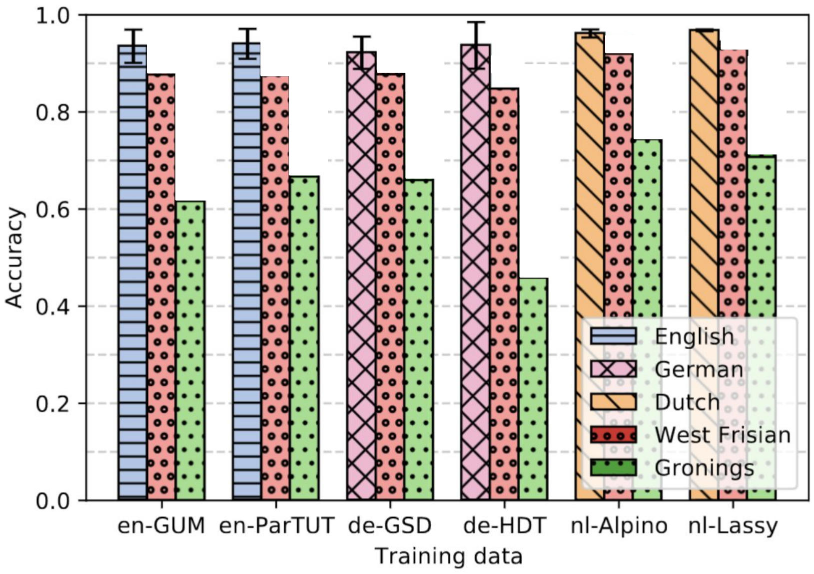
# Separate fine-tuning and word embedding retraining



# Results: original word embeddings

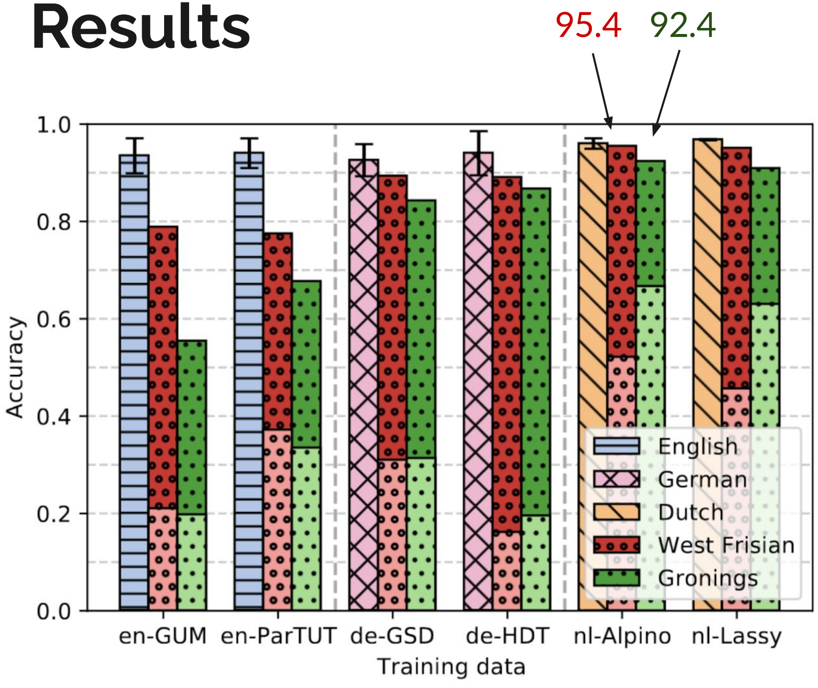



**(a)** Monolingual POS accuracies for BERT, gBERT and BERTje.

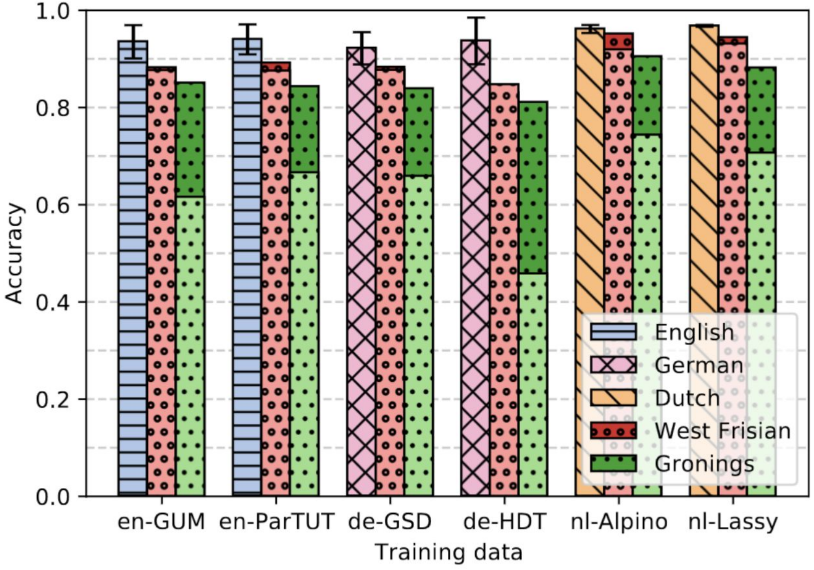



**(b)** Multilingual POS accuracies for mBERT.

# Results




**(a)** Monolingual POS accuracies for BERT, gBERT and BERTje.




**(b)** Multilingual POS accuracies for mBERT.

# Results

| 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 5.2** | Accuracy per target language variety (columns) per lexical layer (sub-columns). This table shows that not all datasets are equally effective for transfer to Gronings and West Frisian.

# How much data is needed for word embeddings

| | | | 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 5.3**| POS-tagging accuracy for Gronings and West Frisian with subsets of the unlabeled lexical layer retraining data.

# Conclusion

- Word embedding retraining is an extremely effective way to adapt task-specific models!
- Only 10mb of data (~1.9 million tokens) is enough to adapt from a very similar language
- Monolingual models outperform mBERT cross-lingually
- How important is language similarity in general?

ACL 2022

# Cross-lingual training with over 100 languages

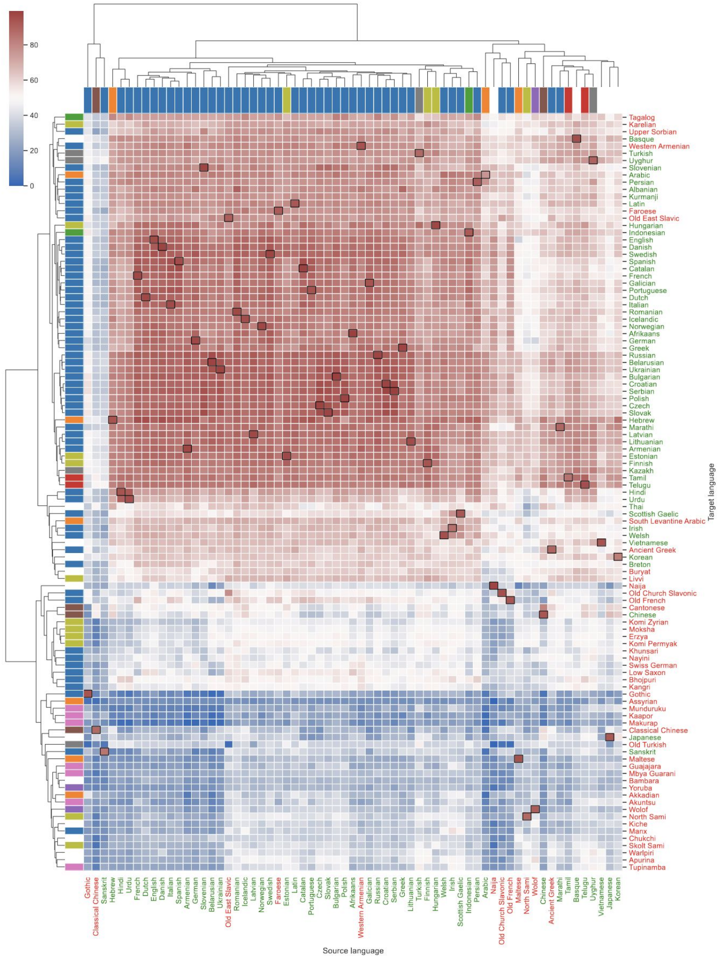
**de Vries, W.**, Wieling, M., and Nissim, M. (2022). Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

# Introduction

- Previous 2 papers: adaptation from English or from highly similar source languages
- How does this generalize to other languages and language families?
- Simple setup: Fine-tune XLM-RoBERTa for POS tagging with all languages in Universal Dependencies v2.8
  - 65 languages with (enough) training data
  - 114 languages with test data
- 65 x 114 = 7410 test scores (!)

# Introduction

- Previous 2 papers: adaptation from English or from highly similar source languages
- How does this generalize to other languages and language families?
- Simple setup: Fine-tune XLM-RoBERTa for POS tagging with all languages in Universal Dependencies v2.8
  - 65 languages with (enough) training data
  - 114 languages with test data
- 65 x 114 = 7410 test scores (!)



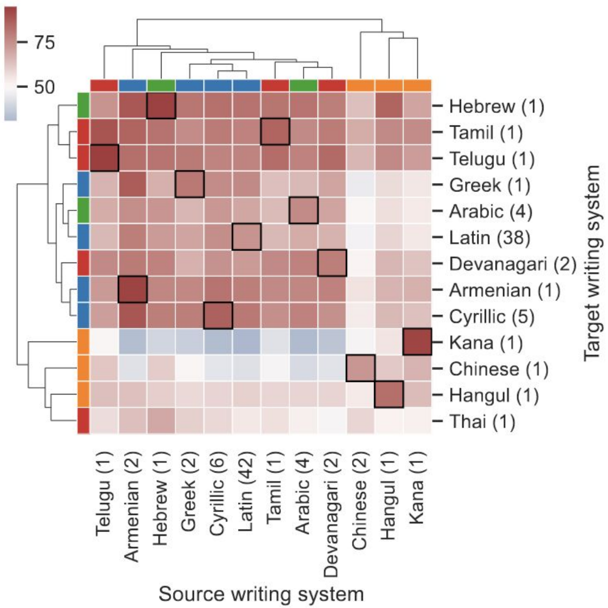
# Analysis

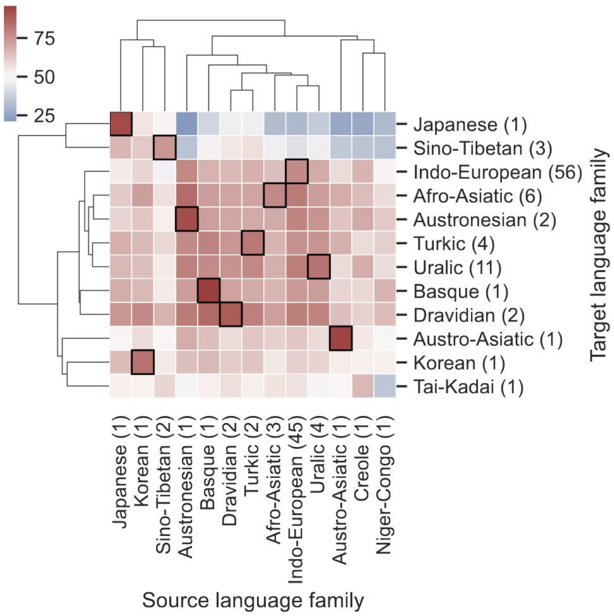
- What are the effects of:
  - Inclusion in pre-training
  - Language similarity
    - automatic LDND measure for lexical similarity)
  - Language families
  - Writing systems
  - Word order
- Mixed effects regression analysis
  - Random effects for source and target languages (no interactions)

| **Predictor** | **Coef.** | **Std. Err.** |
|---|---|---|
| (Intercept) | 42.2 | 3.3 |
| Target pre-trained | 19.2 | 2.5 |
| LDND distance | −12.7 | 1.0 |
| Both pre-trained | 7.4 | 7.4 |
| Same family | 6.8 | 6.8 |
| Source pre-trained | 5.6 | 2.0 |
| Same writing system type | 3.6 | 0.4 |
| Same writing system | 1.4 | 0.3 |
| Same SOV word order | 1.3 | 0.2 |

**Table 6.1**| Coefficients and standard errors of predictors in the final mixed-effects regression model with Accuracy as the dependent variable. All predictors were significant at the $p < 0.01$ level. LDND distances were scaled between 0 (minimum) and 1 (maximum). The predictors are sorted in order of decreasing importance.

# Effects of writing systems and language families



# Source/Target symmetry

- Estonian and Finnish
- Icelandic and Faroese
- French and Italian
- Chinese and Japanese
- Irish and Scottish Gaelic
- Croatian and Serbian
- Catalan and Spanish
- Belarusian and Ukrainian
- Hindi and Urdu
- Armenian and Western Armenian
- English and Swedish

- From same or neighbouring countries
  - Exceptions: English-Swedish
- Genetically closest siblings (or actually two variants of the same language)
  - Exceptions: English-Swedish, Chinese-Japanese, Catalan-Spanish

# What is the best source language?

- Real answer: pick the highest resource language that is closely related to the target language
- Our experiments contain multiple language families and writing systems, but Indo-European languages are still overrepresented. Therefore, aggregates are biased

# What is the best source language?

- Anyway: **Romanian** and **Swedish** are the best for most target languages (**10** and **7** respectively)
- They also achieve the highest global average accuracy: **67.2%** and **65.9%**
- **English** is only the **19th** best source language (out of **65**)
- English is even just the 5th best Germanic Indo-European language...

# Conclusion

- Languages need to be included in pre-training (can be overcome with the strategy of the previous paper)
- Cross-writing system performance is good for alphabetic writing systems but not for logo-syllabic systems
- Any cross-lingual experiment that you will see does **not** show how good a multilingual model works for a target language, but how good it will transfer from English to that target language

# Conclusions

# Every language except English is under-resourced

- Dutch is not considered a low-resource language, but we show that other model types and larger sizes would yield much better results than current models
- Smarter transfer strategies such as word embedding retraining or using adapters work better than just fine-tuning a multilingual model. Especially with monolingual models
- Cross-lingual performance of multilingual models is highly dependent on the relationship between source and target languages
- The models that I used are small by today's standards. How this affects huge generative models is an open question

# Thanks for your attention!

- Please get in touch if you have any questions
  - Only via email: wietse.de.vries@rug.nl