Approaches to investigating language variation and supporting minority languages

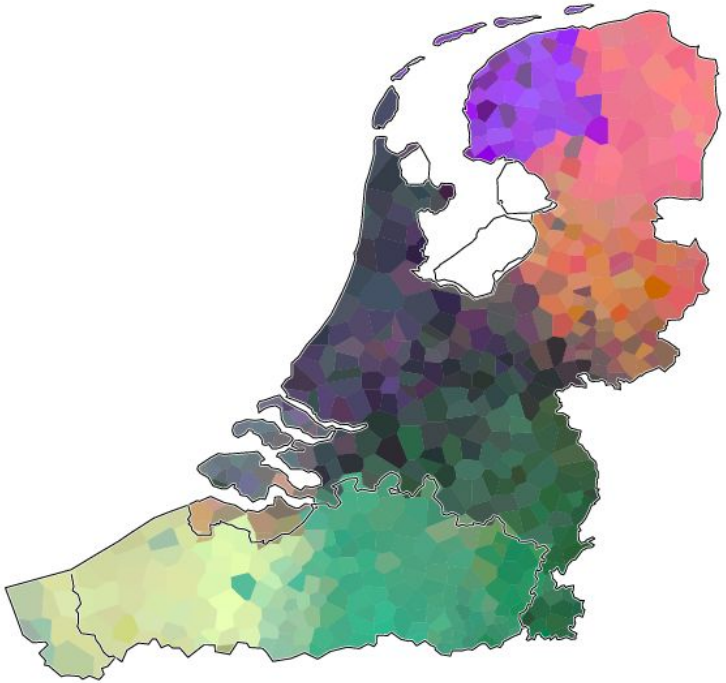
Martijn Wieling
Raoul Buurke
Martijn Bartelds
Wietse de Vries

University of Groningen

Dialectometry: quantifying language variation (using transcriptions)

Background

- Dialects in the Netherlands
- Dialect change (phonetic level)
- Research questions
  - How much have varieties changed over time?
  - Which varieties have changed most over time?



Methodology: datasets (1)

- Phonetic transcription corpora
  - Dialect speakers were asked to translate Standard Dutch words into local dialect
  - Use of existing corpora
- Dataset 1: Reeks Nederlandse Dialectatlassen (RND)
  - Recordings between 1925 and 1982 in Netherlands and Belgium
  - 347 locations
  - 166 words
  - 16 different transcribers

Methodology: datasets (2)

- Dataset 2: Goeman-Taeldeman-Van Reenen Project (GTRP)
  - Recordings between 1980 and 1995
  - Approximately same area as RND
  - 613 locations
  - 562 words
  - 23 different transcribers
- Only overlapping locations can be used
  - 192 locations
  - 62 words

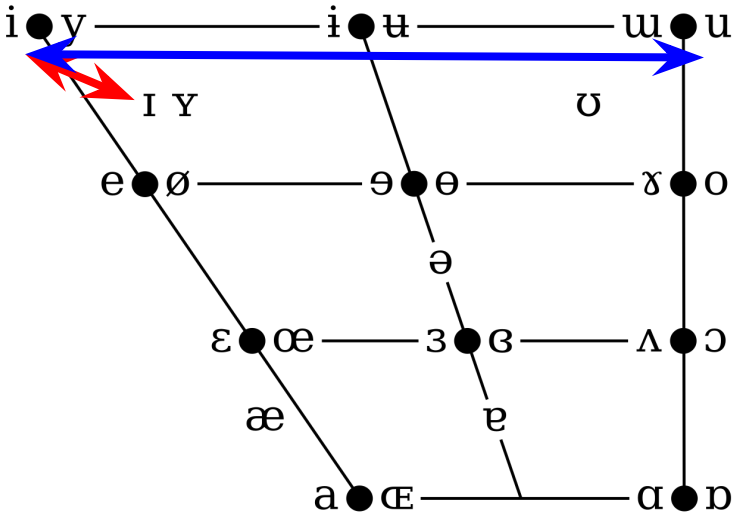
Methodology: phonetic string comparisons (1)

- Count the number of different sounds between transcriptions
- Use of Levenshtein distance
  - Minimal number of **operations** necessary to transform one string into another
  - Operations: insertions / deletions / substitutions
  - Example with binary costs: 0 or 1

	s	t	ʀ	o		d	ə
	s	t	ʀ	ɔ	ə	t	
Operation	-	-	-	**subst.**	**insertion**	**subst.**	**deletion**
Cost	0	0	0	1	1	1	1

Methodology: phonetic string comparisons (2)

- Phonetically sensitive alignments using induced distances
  - [i]-[ɪ] substitution is less ‘costly’ than [i]-[u] substitution
  - PMI-based method using corpus of transcriptions
  - Results in values between 0 and 1
- Linguistic rules also enforced
  - Vowels and consonants cannot align
    - Define these substitutions to have high costs
  - Normalize to account for different word lengths
    - Divide Levenshtein distance by alignment length
    - ‘Streets’ example:
      distance / alignment length
      = 4 / 7
      = 0.57



Methodology: reducing transcriber effects

- Transcriber effects are pervasive when transcriptions are used
  - RND and GTRP are both known to have them
  - Also differences in how fine-grained transcriptions are; number of phonetic symbols:
    - RND: 43
    - GTRP: 76
- Solution: merge transcription inventories
  - Replace characters that occur in the GTRP only with those that occur in both
  - Use the minimal distance from its alternatives
  - Example: [e] is replaced with [ɛ]

	e	ɛ	i	ø
e	0			
ɛ	0.019	0		
i	0.020	0.020	0	
ø	0.023	0.030	0.030	0

Methodology: phonetic change

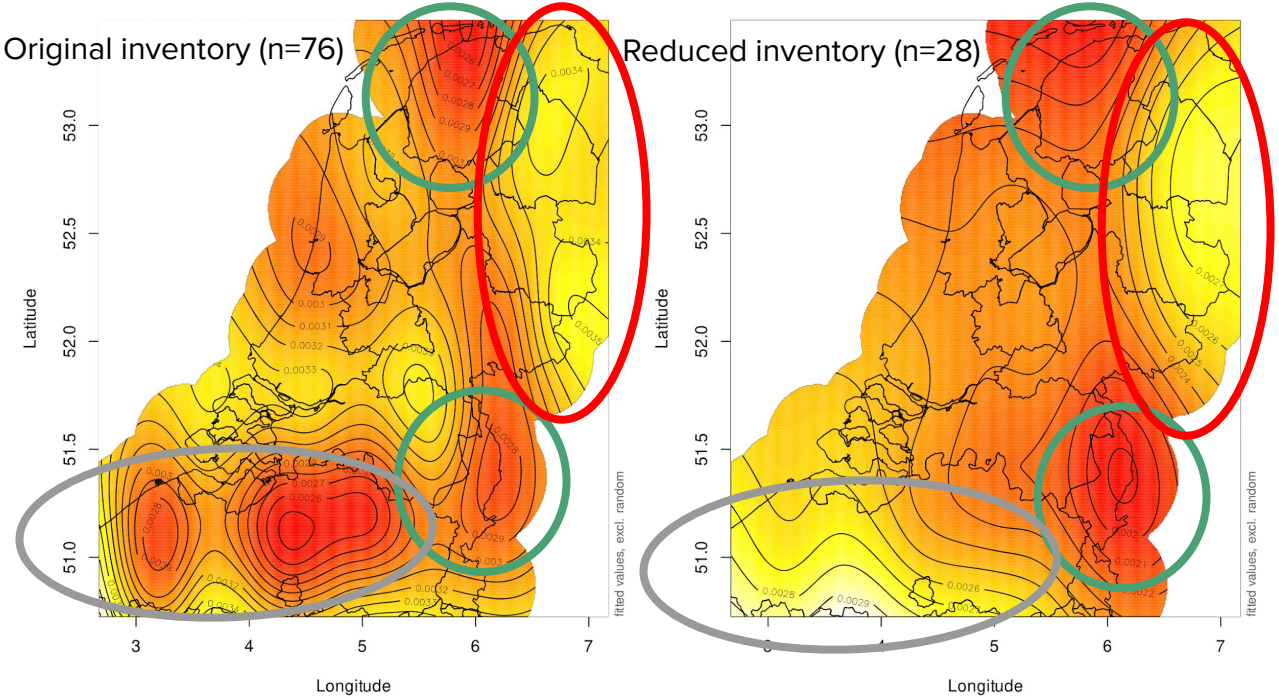
- Compute distance (after reduction) between RND and GTRP for all word-location combinations
- Normalize values between 0 and 1
- For example: Dutch ‘huis’
  - RND transcription for location in 1955: [hus]
  - GTRP transcription for location in 1985: [hys]
  - Levenshtein distance is 1
  - Normalized distance
    = 1/3
    = 0.33
  - 33% change over 30 years

Results

- Reduced to 38 symbol common inventory (originally GTRP: 74, RND: 43)
  - Manually checked for crucial dialect differences based on literature
- Further reduction to 28 symbols
  - Constraint based on low frequency (< 1% of summed total token frequency)
  - Low frequency defined as occurring less than 1% of total sum of characters occurrences
- Final inventory: 10 vowels, 18 consonants

Results

- Model phonetic change by geography
  - Generalized Additive Model using non-linear interaction of latitude and longitude



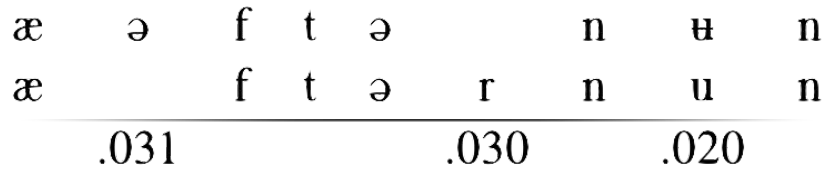
Discussion

- Most stability for Frisian and Limburgish dialects
  - In line with expectations about their speaker populations
- Most change in Low Saxon and West Flemish areas
- Phonetic inventory reduction
  - Fewer characters used gives a ‘cleaned up’ view, but what is going too far?
- Current work: in what direction do they change?
  - Towards Standard Dutch?
  - Towards neighboring dialects?

Moving past transcriptions: using acoustic methods to quantify pronunciation variation

Background

- Levenshtein distance



- Time consuming
- Prone to errors

Motivation

- Can we create fully automatic acoustic-only methods for investigating pronunciation differences?
- Specifically, we would like to quantify how different an L2 speaker’s pronunciation is from native speech

Data

- Speech Accent Archive: data set of non-native and native American English pronunciations
- Our native-speaker reference set has samples from 115 native AE speakers
- We evaluate against human judgements of nativelikeness for 286 speakers (Wieling et al., 2014)

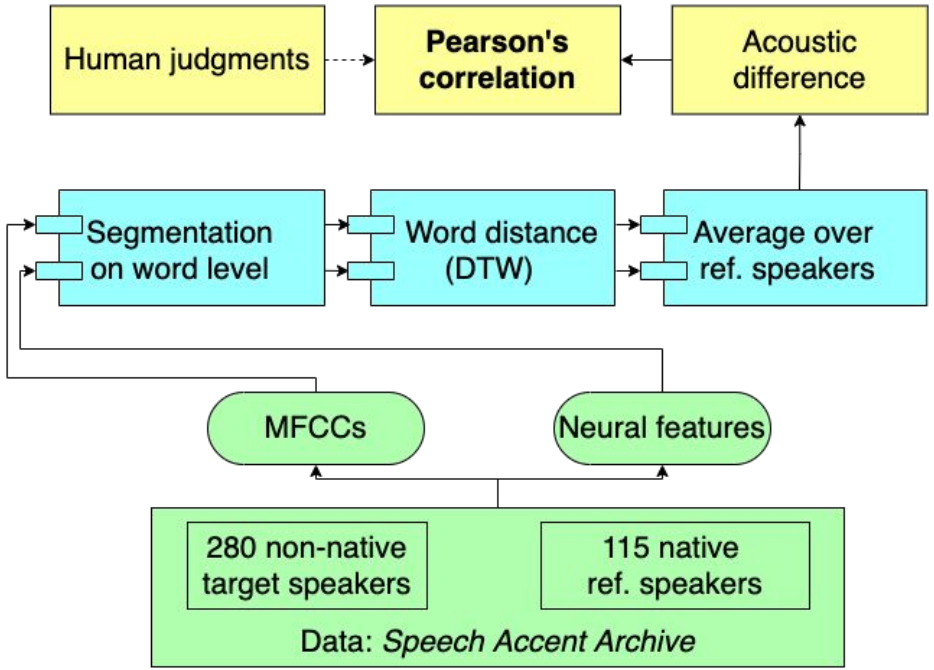
Feature extraction (from models)

- Transform raw audio into feature representations
  - MFCCs
  - Pre-trained neural models (English)
    - Wav2vec
    - Vq-wav2vec + BERT
    - Wav2vec 2.0
    - DeCoAR

Features

- 39-dimensional MFCCs
  - 12 cepstral coefficients
  - 1 energy coefficient
  - first and second order derivatives
- Wav2vec
  - Extract features from the 7 layer encoder network
- Vq-wav2vec + BERT
  - Quantisation layer enables use of BERT
  - Extract features from BERT’s layers
- Wav2vec 2.0
  - End-to-end model
  - Extract features from the Transform layers
- DeCoAR
  - Extract features from the bi-directional LSTM network

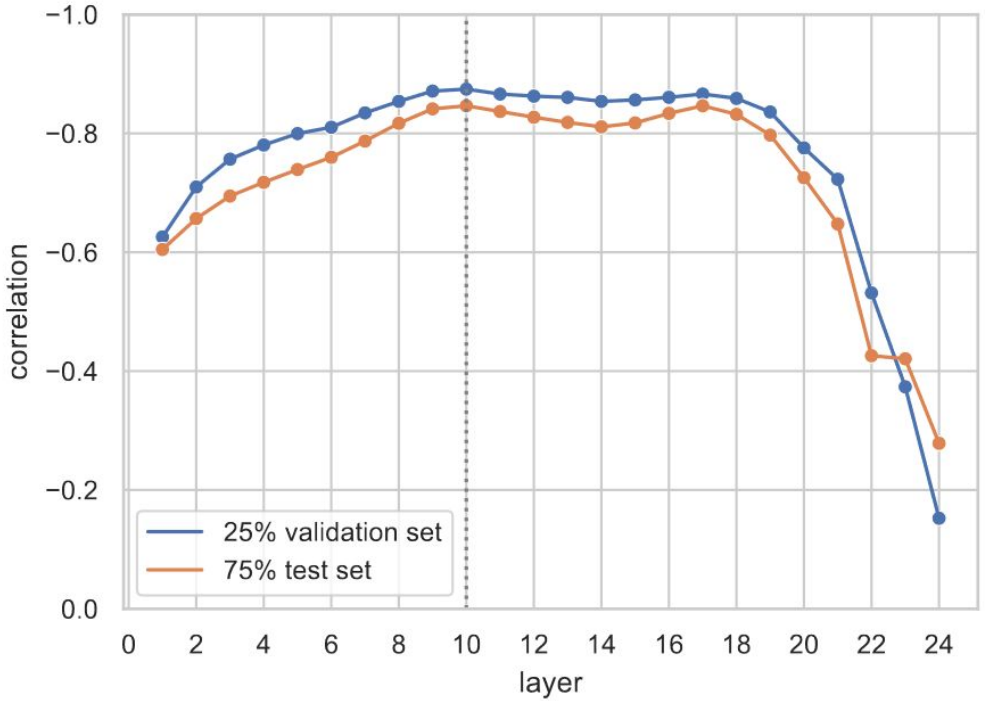
Methodology



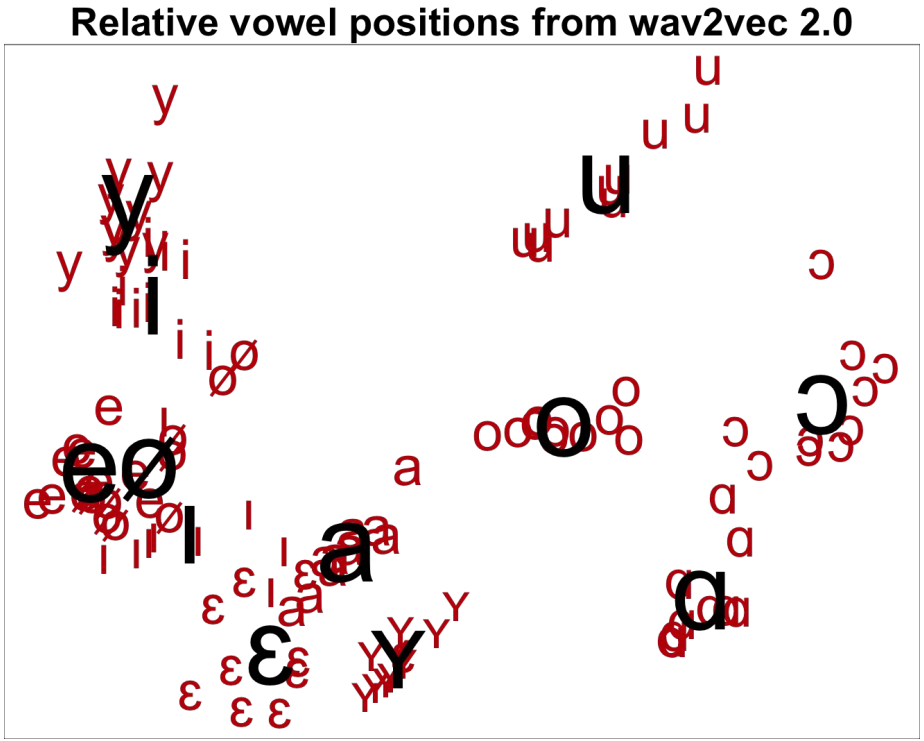
Results: evaluated against human judgements

Model	***r***
Wav2vec	-0.69
Vq-wav2vec + BERT	-0.79
Wav2vec 2.0 (layer 10)	**-0.85**
DeCoAR	-0.72
MFCCs	-0.71
PMI-weighted Levenshtein distance (previous state-of-the-art)	-0.77

Results: Wav2vec 2.0 layer-specific performance



Results: vowel differences



Conclusion and discussion

- Determining nativelikeness can be done without transcriptions
- (or human raters, as the acoustic pronunciation distances match perceptual differences well)
- Wav2vec 2.0 performs best, but large amounts of data and computing resources are needed to create models for other languages
- Current work: examining how well this model works for other languages

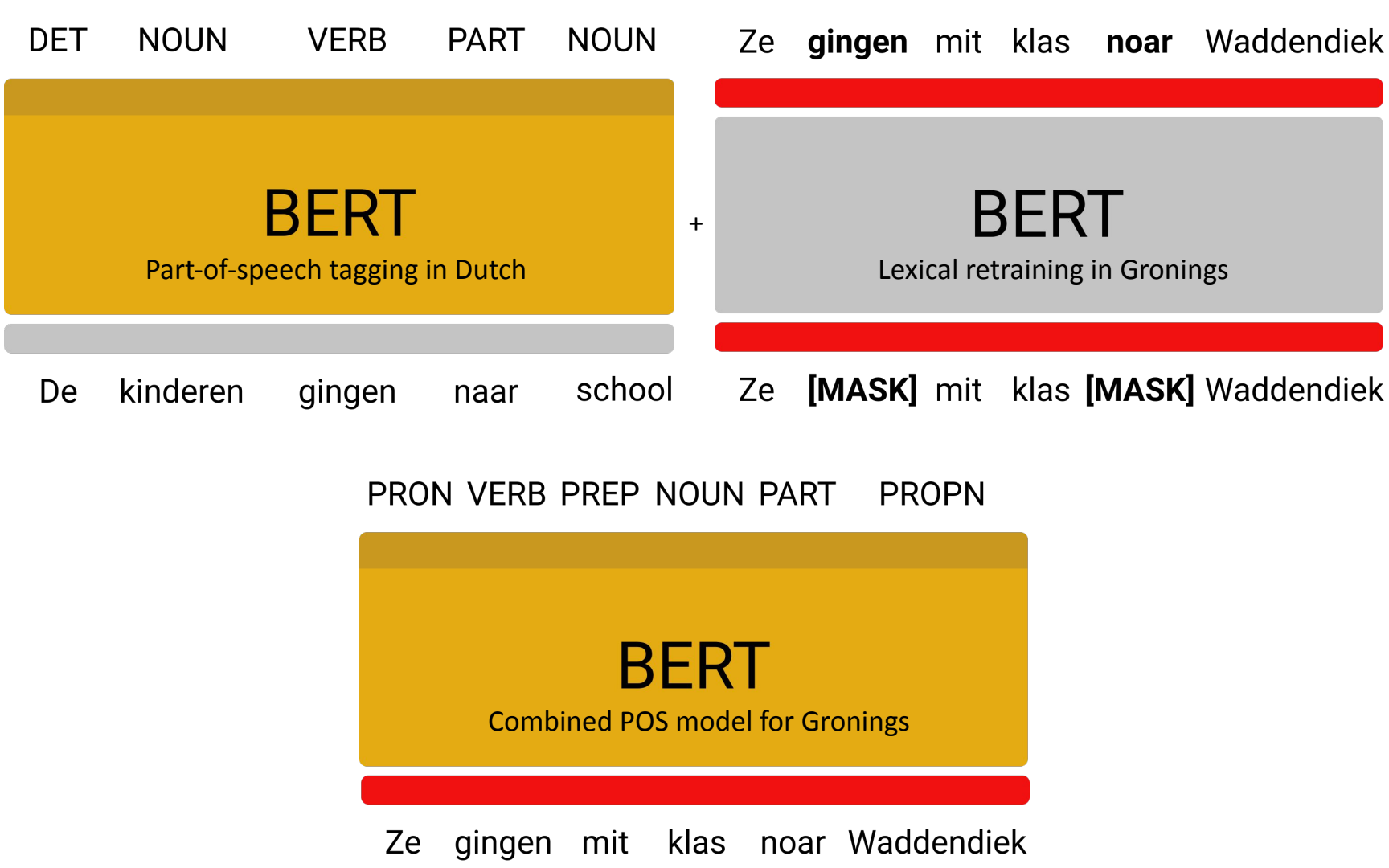
NLP for minority languages

Part-of-speech tagging

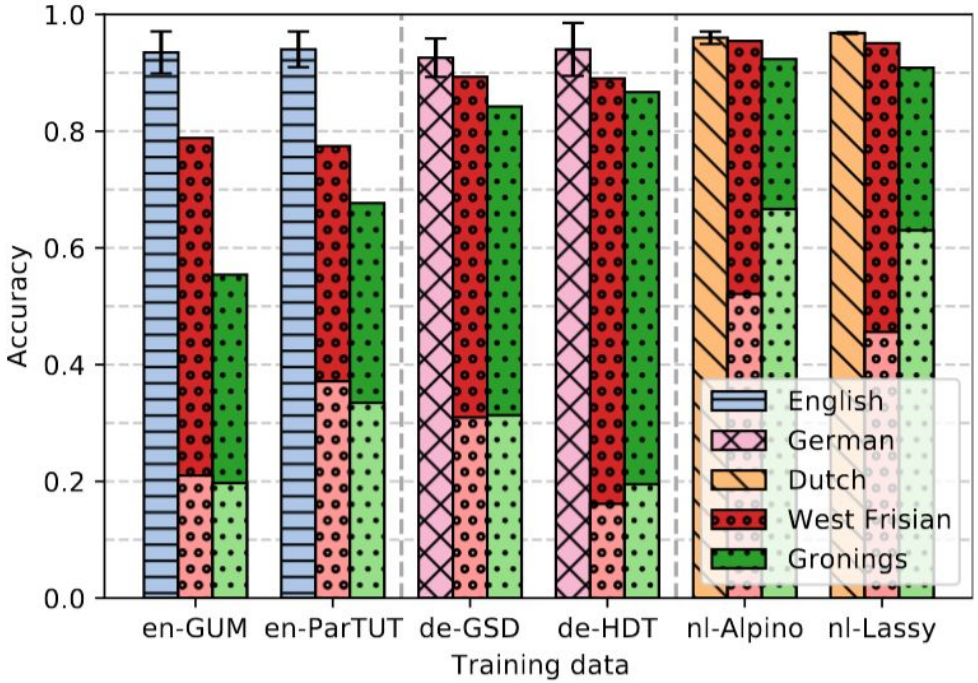
- Requires training data: Universal Dependencies
- If you fine-tune BERT, an accuracy of 95% is generally achieved
- Our method: adapt fine-tuned models for other languages for low-resource languages
  - Gronings
  - West Frisian

Fine-tuning monolingual and multilingual BERT

- Source languages
  - Dutch
  - German
  - English
  - Multilingual (= 104 languages)
- Method:
  - Fine-tune for source language
  - Retrain lexical layer for target language
  - Combine fine-tuned transformer layers With retrained lexical layers

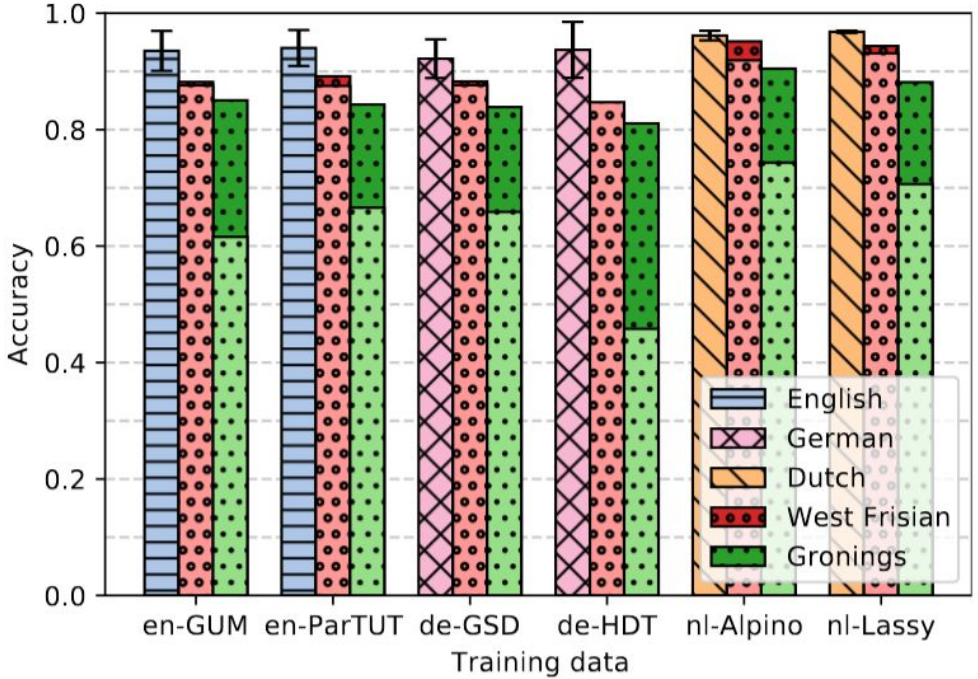

DET NOUN VERB PART NOUN
Ze gingen mit klas noar Waddendiek
BERT
Part-of-speech tagging in Dutch
+
BERT
Lexical retraining in Gronings
De kinderen gingen naar school
Ze [MASK] mit klas [MASK] Waddendiek
PRON VERB PREP NOUN PART PROPN
BERT
Combined POS model for Gronings
Ze gingen mit klas noar Waddendiek

Results: high performance for Gronings and West Frisian



- Light/dark colours: performance without/with retraining lexical layer

Results: lower mBERT performance for Gronings



- Gronings not in mBERT language set

Results: how much unlabeled data?

		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 2: 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.

Discussion

- Multilingual BERT has much higher transfer performance for West Frisian than for Gronings
  - West Frisian was included in mBERT pre-training, so true low-resource languages that were not included do not benefit from mBERT usage
- Our transfer method is most effective for a source language that is lexically similar
  - And only little *unlabeled* data is necessary
- This approach is not limited to only POS tagging or only these languages

Thank you for your attention!

Martijn Wieling
m.b.wieling@rug.nl
www.martijnwieling.nl
www.speechlabgroningen.nl
@martijnwieling