



# Multi-task learning for historical text normalization: Size matters

Marcel Bollmann, Anders Søgaard, Joachim Bingel

@mmbollmann

marcel@di.ku.dk

## What is normalization?

- Mapping words in historical text to modern equivalent
- Example:

*erthely*  
↓  
*earthly*

## Why normalization?

- Reducing spelling variation
- Typical use cases:
  - Preprocessing step for NLP tools
  - Preprocessing for linguistic analyses
  - Improve search queries
- Increasingly relevant as more historical texts are digitized!

Dataset/Language	Time Period	Tokens	
		Train	Test
DE <sub>A</sub> German (Anselm)	14 <sup>th</sup> -16 <sup>th</sup> c.	233,947	45,999
DE <sub>R</sub> German (RIDGES)	1482-1652	41,857	9,587
EN English	1386-1698	147,826	17,644
ES Spanish	15 <sup>th</sup> -19 <sup>th</sup> c.	97,320	12,479
HU Hungarian	1440-1541	134,028	16,779
IS Icelandic	15 <sup>th</sup> c.	49,633	6,037
PT Portuguese	15 <sup>th</sup> -19 <sup>th</sup> c.	222,525	27,078
SL <sub>B</sub> Slovene (Bohorič)	1750-1840s	50,023	5,969
SL <sub>G</sub> Slovene (Gaj)	1840s-1899	161,211	21,493
SV Swedish	1527-1812	24,458	29,184

Datasets used in our experiments

## Neural sequence-to-sequence model

- **Input:** Single word form, represented as sequence of characters
- **Encoder:** bi-directional LSTM
- **Decoder:** uni-directional LSTM with attention mechanism
- Generate output sequence character by character, using greedy decoding

## Main research questions

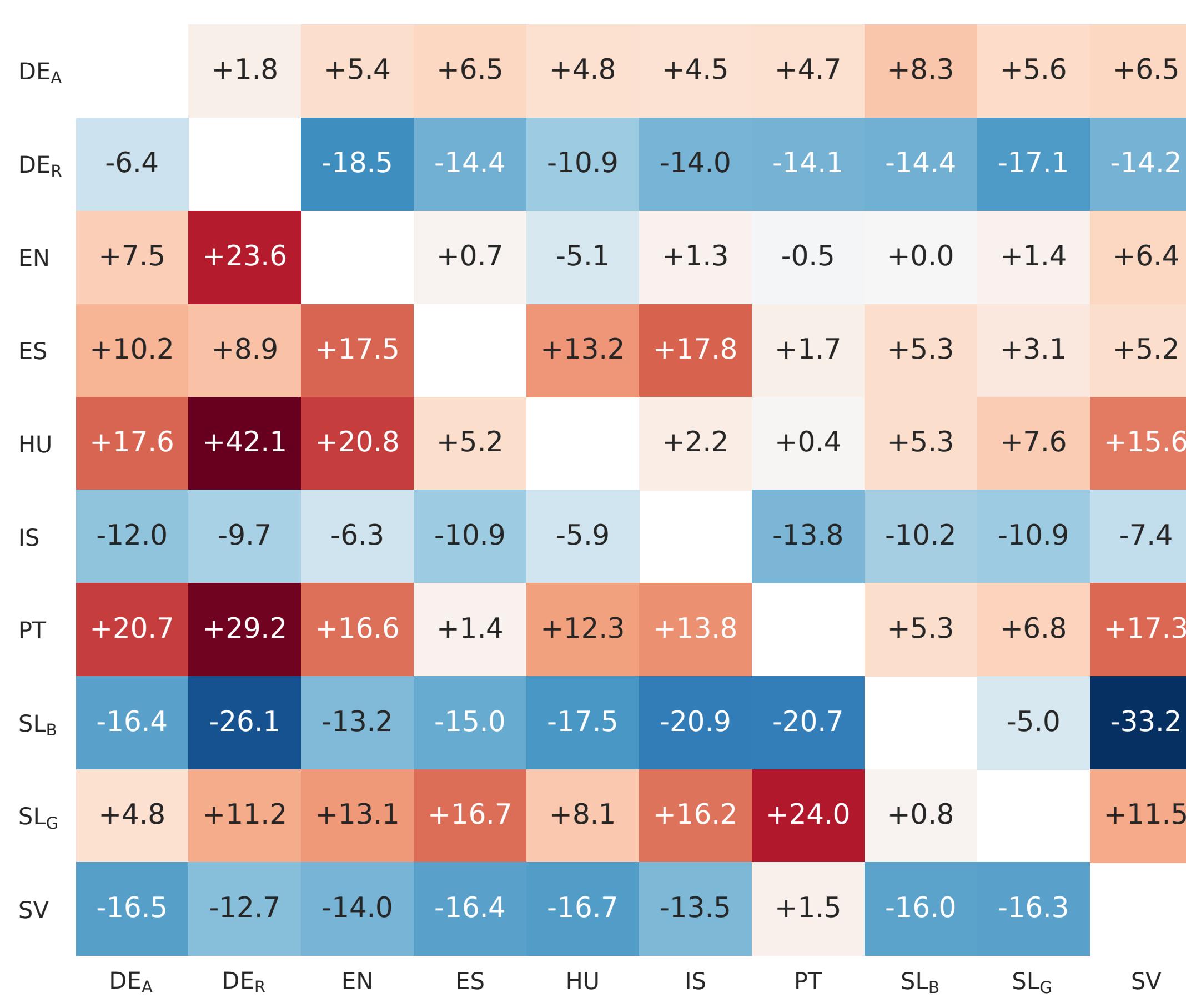
- ★ Can we improve normalization with cross-lingual learning?
- ★ Do some dataset pairings work better than others?
- ★ What dataset properties are most predictive of this?

## Multi-task learning

- Training on **pairs of datasets A and B**
- Share all model components except for the final prediction layer
- Mini-batch training with 50 samples from both A and B
- Validation on dev sets after every 50,000 samples, keeping only the best models for A and B

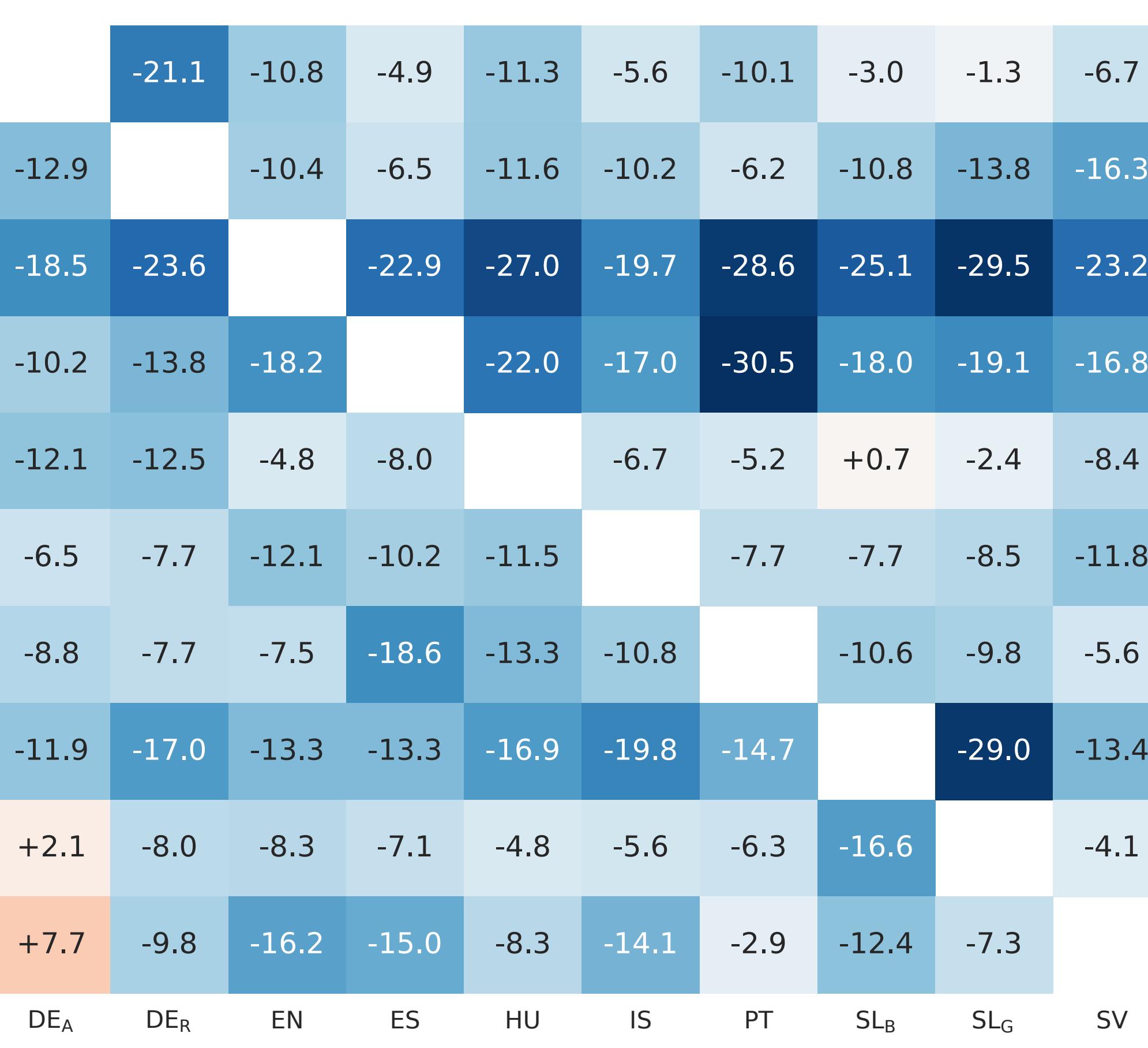
## Full data scenario

Train on full training sets for both datasets



## Sparse data scenario

Train on 5,000 tokens for target dataset (rows), jointly with full auxiliary dataset (columns)



Percentage change of error of multi-task learning over single-task models

(blue = improvements, red = error increases)

Rows are target datasets, columns are auxiliary datasets

## Results

- Multi-task learning helps most **when target dataset is small**
- Multi-task learning can even be **detrimental** when target datasets are already (sufficiently) large
- Size of target training set is more important than choice of auxiliary dataset

## Takeaways

- ★ Multi-task learning can help a lot when you don't have a lot of training data!
- ★ Always consider the size of the training set when evaluating multi-task learning approaches!