UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

A Large-Scale Comparison of Historical Text Normalization Systems

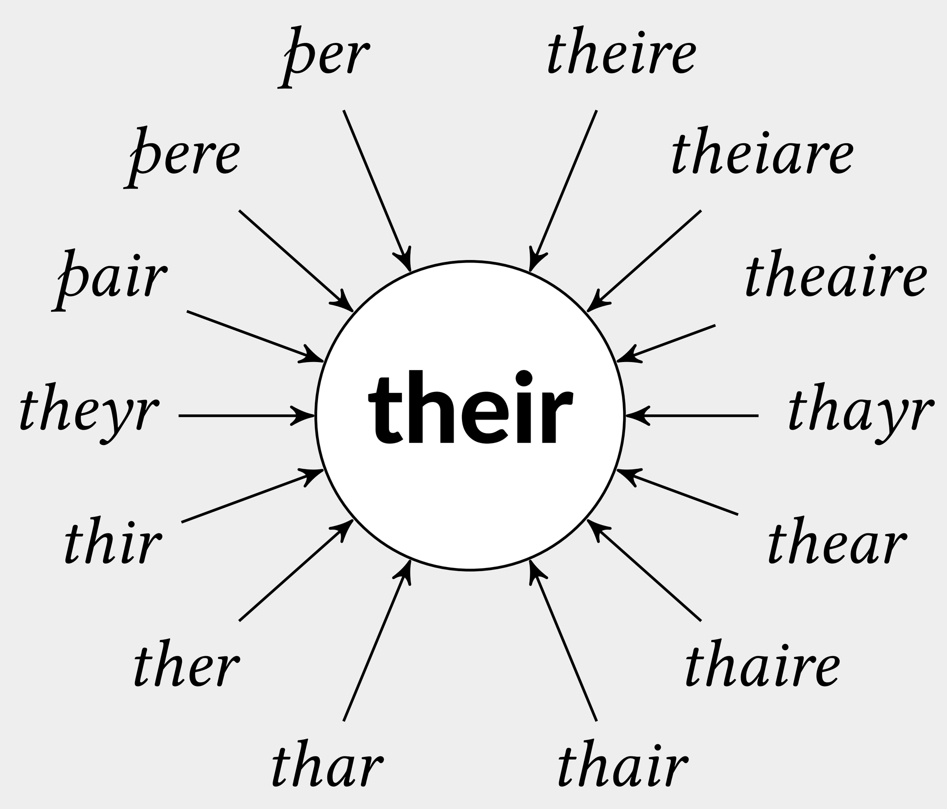
Marcel Bollmann marcel@di.ku.dk @mmbollmann

The Data

- Historical corpora from **eight languages**
- Texts written between the 1300s and 1899

Dataset/Language		Time Period	Genre	Size (Tokens)
DE_A	German (Anselm)	14^{th}–16^{th} c.	Religious	325,942
DE_R	German (RIDGES)	1482–1652	Science	61,156
EN	English	1386–1698	Letters	181,804
ES	Spanish	15^{th}–19^{th} c.	Letters	121,449
HU	Hungarian	1440–1541	Religious	167,514
IS	Icelandic	15^{th} c.	Religious	61,779
PT	Portuguese	15^{th}–19^{th} c.	Letters	276,352
SL_B	Slovene (Bohorič)	1750–1840s	Diverse	61,833
SL_G	Slovene (Gaj)	1840s–1899	Diverse	203,582
SV	Swedish	1527–1812	Diverse	55,887

Normalization
is the mapping of historical (spelling) variants to a canonical (modern) form.



The Systems

All systems are **token-level** approaches with **supervised** learning.

1 Norma (Bollmann, 2012)

- Implements **wordlist lookup**
- Has a **rule-based** component

  þ → th / #_e

- Has a component based on (weighted) **Levenshtein distance** to lexicon entries

2 Statistical MT

- Character-level "translation" of tokens

  Input: <w> þ e r </w>

  Output: <w> t h e i r </w>

- Uses **Moses** toolkit with GIZA++
- Implementation and settings provided by **cSMTiser** toolkit (Ljubešić et al., 2016)

3 Neural MT (seq2seq)

- Character-level encoder-decoder models
- **NMT-1** (Bollmann, 2018):
  - LSTMs with dimensionality 300
  - Implemented with XNMT
- **NMT-2** (Tang et al., 2018):
  - RNNs with dimensionality 1024
  - Implemented with Marian

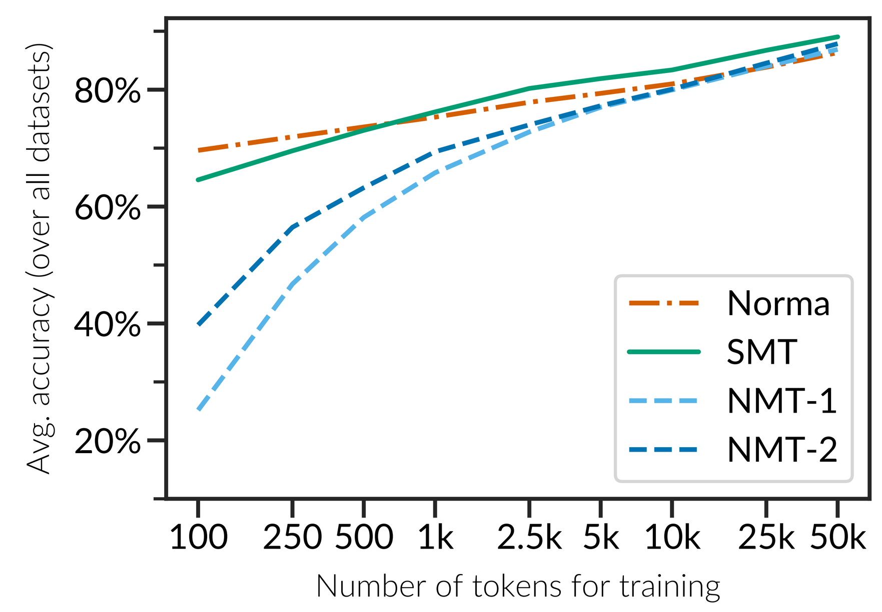
- Marcel Bollmann. 2012. *(Semi-)automatic normalization of historical texts using distance measures and the Norma tool*. In Proceedings of ACRH-2, Lisbon, Portugal.
- Marcel Bollmann. 2018. *Normalization of historical texts with neural network models*. Bochumer Linguistische Arbeitsberichte, 22. (Revised and updated version of PhD thesis.)
- Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. *Normalising Slovene data: historical texts vs. user-generated content*. In Proceedings of KONVENS 2016, Bochum, Germany.
- Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. *An evaluation of neural machine translation models on historical spelling normalization*. In Proceedings of COLING 2018.

Datasets, code, instructions, etc.:
github.com/coastalcph/histnorm

Results

System	Dataset									
	DE_A	DE_R	EN	ES	HU	IS	PT	SL_B	SL_G	SV
Norma (Bollmann, 2012)	88.0	86.6	94.6	94.4	86.8	*86.8	94.2	89.4	91.4	87.1
SMT (Ljubešić et al., 2016)	86.7	*88.2	**95.2**	***95.0**	***91.7**	*86.8	**95.2**	**93.3**	***96.0**	***91.1**
NMT-1 (Bollmann, 2018)	89.2	*88.1	94.8	*94.8	91.2	86.4	94.6	91.6	95.2	90.3
NMT-2 (Tang et al., 2018)	**89.6**	***88.2**	95.0	*94.8	*91.6	***87.3**	94.5	92.6	*95.8	90.4

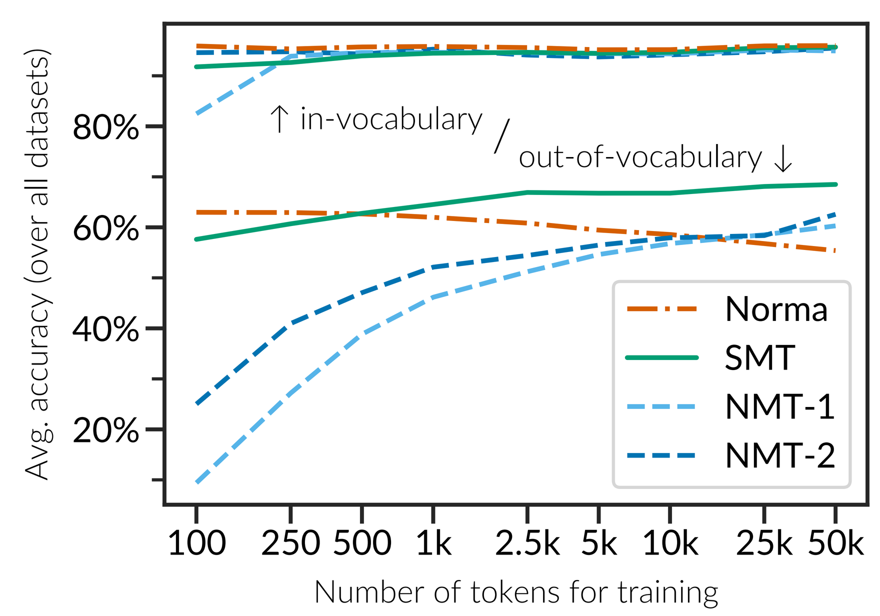
Normalization accuracy on test data (* = difference not statistically significant)



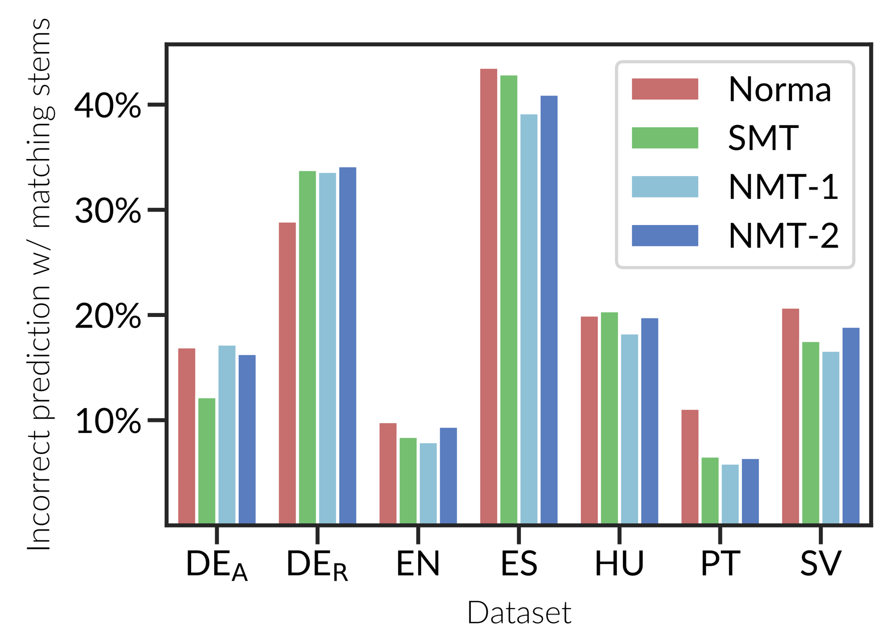
★ **SMT** is best on average

★ **Norma** is a good option with little training data

(NMT might need more data?)



★ Best strategy is **lookup** for IV tokens and trained model (e.g. SMT) only for **OOV tokens**



★ **Stemming** can provide more nuanced view of datasets and prediction errors