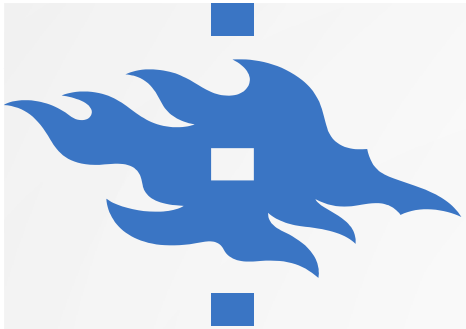


# National Library Newspaper Archive Re-OCR and Reprocessed

Senka Drobac

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Paper at Nodalida 2017:

OCR and post-correction of historical Finnish texts - Senka Drobac and Pekka Kauppinen and Krister Lindén

http://www.ep.liu.se/ecp/131/009/ecp17131009.pdf

Data: https://github.com/sdrobac/nodalida2017

# What did we do?

Optical character recognition (OCR)

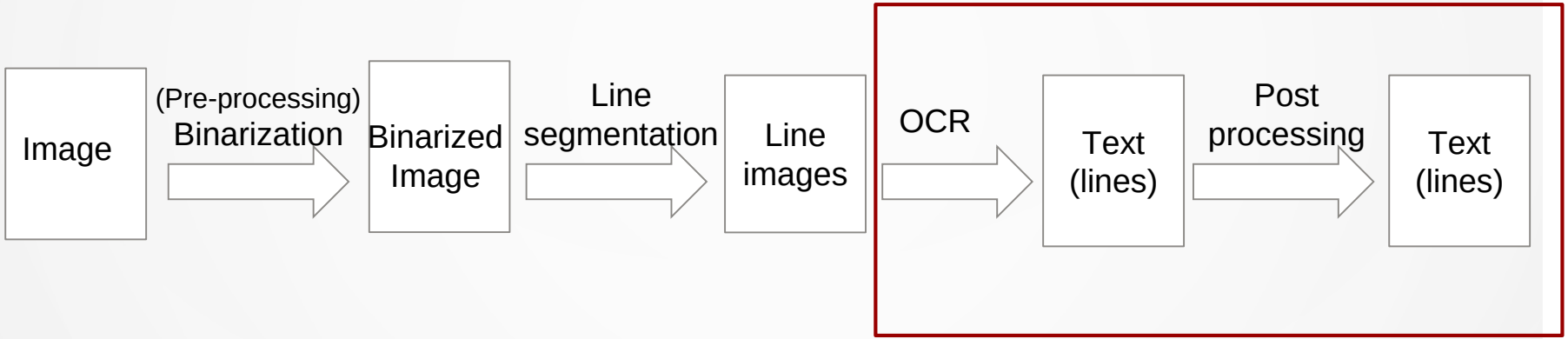
Ocropy software

+

Data-driven spelling correction that uses Weighted Finite-State Methods

Data: Finnish corpora of historical newspaper text

# OCR workflow



# Post-correction – our approach

- Error model = unstructured classifier (Silfverberg et al. 2016)
- Essentially performs weighted context-sensitive parallel substitutions on the input string
  - Contexts only refer to surrounding input level symbols
  - Requires pairs of OCR strings and manually corrected versions as training data
  - Note: only applied on individual strings/words

# Results

- Best result: **95.21%** (CAR)
- Simple post-correction method improved OCR results in all test cases
  - possible to further improve OCR
  - use better post-correction

# Now + future

- Apply this OCR method to all newspapers, add to Kielipankki
- Challenges:
  - Huge amount of data
  - Different fonts, languages
  - Mistakes in meta-data (i.e. wrong line coordinates, language marks)