

Evaluation & automated correction of historical newspaper OCR using Deep Learning

Delpher Newspaper Corpus

- 11 million pages digitised newspapers
- 1618 – 1995
- Images + metadata + OCR
- www.delpher.nl/kranten

Project aims

- Insight into quality of OCR
- Insight into methods of automated correction

Project output

- Sample set (2000 pages) with ground-truth
- Prototype for OCR correction using deep learning

Deep learning method

- Character-Level Language models
- Long Short Term Memory (LSTM)

Janneke van der Zwaan
(eScience Center)
@jvdzwa

Lotte Wilms (KB)
@lottewilms
netherlands