Overall impresso pipeline

Main objective: enabling critical text mining to search, extract, process, link, and explore data from print media archives via a unified web interface

INVOLVES
- computational linguists
- digital humanists
- historians
- web designers

THEIR COMMON GOAL
- tackling the challenges of content enrichment and data representation, visualization and analysis, completed by methodological and epistemological reflections

EXPECTED OUTCOMES
- natural language processing (NLP) tools dedicated to historical print media written in French and German
- visualization interfaces for active and goal-oriented exploration and critical analysis of newspaper corpora
- an application of digital history research on resistance to European integration

Surprising success: Researchers report astonishing OCR results on historical newspapers!

OCR QUALITY OF HISTORICAL NEWSPAPERS

- 167 front pages from the Neue Zürcher Zeitung (NZZ)
- we used Transkribus’ to create a gold standard
- manual correction of words and baselines
- training of Handwritten Text
- Recognition (HTR) model within
- Transkribus with 158 pages

EVALUATION AND COMPARISON
- bag-of-words f-measure evaluation with TextEval 1.4
- www.primaresearch.org/tools/PerformanceEvaluation
- compare three different outputs
- original OCR by NZZ (ocr-2005)
- re-OCRised material using ABBYY Finereader 2 (ocr-2017)
- Transkribus’ HTR model (htr-2018)

TACKLING OCR WITH HTR TOOLS

- HTR models significantly increase OCR quality
- relatively small gold standards for training purposes suffice for decent OCR
- better OCR is beneficial for text mining techniques

DISCUSSION & OUTLOOK

Open questions:
- Do the HTR models trained on the NZZ perform equally well on other newspapers?
- How does occasionally occurring text in antiqua affect OCR quality?