EVALUATING OCR QUALITY IN FLEMISH NEWSPAPER COLLECTIONS

nieuwetijdingen@vlaamse-erfgoedbibliotheken.be

VLAAMSE ERFGOED BIBLIOTHEKEN

Introduction

What is the OCR quality of digitized Flemish newspaper collections? What level of improvements can we expect with today's technology? And which steps should be taken to improve OCR quality and increase usability of digitized newspapers? A new project evaluates and reprocesses a representative sample of pages in order to answer these questions. This testing is carried out as part of Nieuwe Tijdingen, a three-year initiative by the Flanders Heritage Libraries in collaboration with meemoo and more than 50 partner organisations. Nieuwe Tijdingen lays the groundwork for a large-scale programme for the digitization, presentation and preservation of Flemish newspaper collections.

Ground Truth

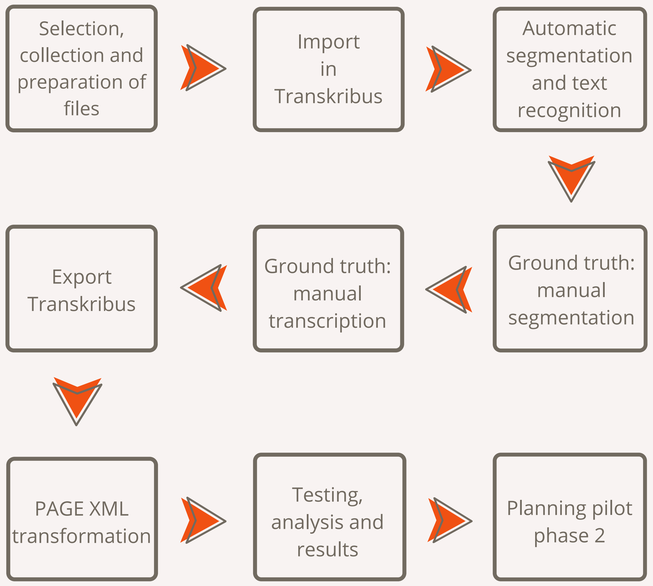
A representative selection of 130 pages from digitized Flemish newspaper collections was made in early 2021. From this corpus, ground truth files are created for 75 to 100 pages for testing, following the OCR-D guidelines for layout analysis and transcription.

Using Transkribus, the regions are manually marked and tagged, and the reading order assigned. The 'CITlab Advanced' method is used for line recognition, which is corrected manually. Colleagues from partner organisations help transcribe the pages using the web version of Transkribus. A transformation is run on the resulting XML files to bring them in line with PAGE XML standards. The ground truth files are then sent to the Staatsbibliothek zu Berlin along with the original scans and OCR files for evaluation.

Testing

The Staatsbibliothek zu Berlin (SBB) performs an evaluation of the files to determine the current error rate. They also run the original scans through their state-of-the-art OCR processing pipeline to determine the quality that is possible to reach with today's technology. Both the quality of the text and the layout analysis (segmentation) is analyzed. The SBB also evaluates the quality of the automatic output that is generated using the Transkribus 'Printed Block Detection' method for the recognition of regions, tagging and reading order, the 'CITlab Advanced' method for line recognition, and finally 'HTR (CITlab HTR+ & PyLaia)' method with the 'Transkribus print 0.3 model' for text recognition. They are additionally testing a small sample that was processed in mid-2021 by a service provider.

Workflow



Next steps

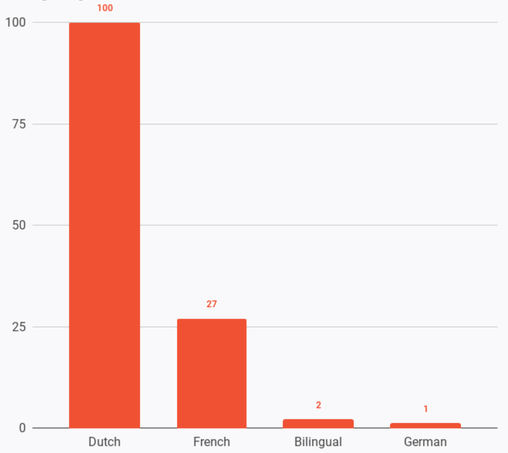
Alongside this testing we are also exploring further collaborations and other opportunities to evaluate and improve the quality of existing OCR texts. The results of the testing and further research will be used to plan a pilot in 2022 to improve the OCR of at least 100,000 pages of Flemish digitized newspaper collections. This project will also inform how we process OCR for new digitization initiatives by the Flanders Heritage Libraries and meemoo.

Partners

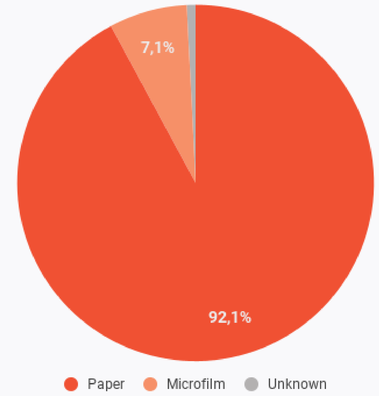
- Meemoo, Vlaams instituut voor het archief
- Staatsbibliothek zu Berlin
- Amsab-Instituut voor Sociale Geschiedenis
- Erfgoedbibliotheek Hendrik Conscience
- Erfgoedcel Waasland
- KADOC-KU Leuven
- Liberas
- Openbare Bibliotheek Brugge
- Stadsarchief Kortrijk
- Stuifzand
- Universiteitsbibliotheek Gent-Boekentoren
- zuidwest

Corpus

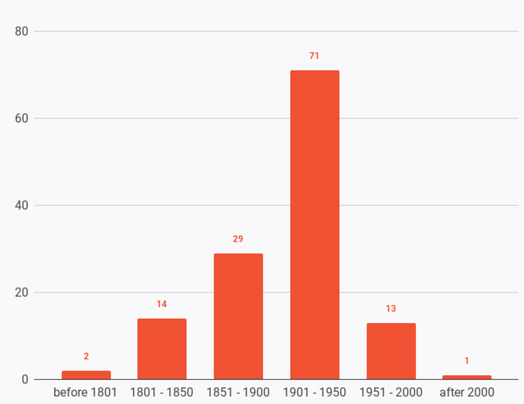
Language



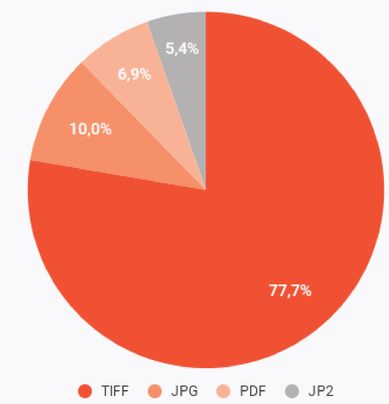
Scan from paper or microfilm



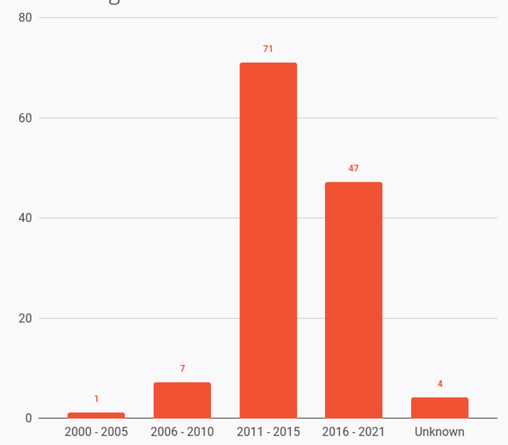
Year of publication



Scan format



Year of digitization



Existing OCR formats