Abstract for a 15-minute individual paper for consideration for the 2021 NEWSEYE conference

Contributors:

Julie M. Birkholz, Sally Chambers, Michal Hradis, Pavel Smrz, on behalf of the EU funded OCCAM Project on OCR, Classification & Machine Translation.

Short Biographies:

Julie M. Birkholz is Assistant Professor Digital Humanities at Ghent University and Lead of the Royal Library of Belgium's Digital Research Lab. Her research expertise is in historical social network analysis. From 2017 – 2020 she was a DH Fellow on the ERC Agents of Change Research project WeChangEd, investigating the historical networks of women editors, periodicals and organizations in Europe, as well as the research data manager for the linked open data of the bibliographic information of these editors. From 2014 – 2017 she was a Postdoctoral Researcher at the Centre for Higher Education Governance Ghent, researching the identification of social networks through web data. She holds a doctorate in Organization Sciences from the VU University Amsterdam, the Netherlands. Given that the study of networks, both the theory and methods, crosses disciplines her research is inherently interdisciplinary. Her most recent research explores a computational method for extracting social networks from historical newspapers. Email: Julie.Birkholz@ugent.be

Sally Chambers is Digital Humanities Research Coordinator at the Ghent Centre for Digital Humanities, Ghent University, Belgium and National Coordinator for DARIAH in Belgium. Sally initially started working in academic libraries in the UK in the mid-1990s before joining The European Library (the predecessor of Europeana) at the National Library of the Netherlands in The Hague in 2005. Since then she has worked for DARIAH-EU, based in the Göttingen Centre for Digital Humanities, Germany before joining the GhentCDH in early 2015. Since late 2020, she divides her time between GhentCDH and the KBR, Royal Library of Belgium, where she coordinates the DATA-KBR-BE project to facilitate data-level access to KBR's digitised and born-digital collections for digital humanities research. She is an active participant in the international Galleries, Libraries, Archives and Museums (GLAM) Labs community, and a co-author of Open a GLAM Lab, a practical guide for setting up, running and maintaining a Digital Cultural Heritage Innovation Lab. Email: Sally.Chambers@ugent.be

Michal Hradis is a researcher and a project leader at the Faculty of Information Technology, Brno University of Technology, Czech Republic. He specialises in visual intelligence, artificial neural networks, OCR and handwritten character recognition, and digital humanities. He is a principal investigator in the national project PERO and the European CEF project OCCAM. He

published research papers on various computer vision topics, with a special focus on digital humanities applications. Email: ihradis@fit.vutbr.cz

Pavel Smrz is an associate professor and research project leader at the Faculty of Information Technology, Brno University of Technology, Czech Republic. His research interests include advanced machine learning and artificial intelligence, knowledge systems, natural language processing, digital humanities, and human-computer interaction. He leads the Knowledge Technology (KnoT) Research Group which has participated in more than 20 Horizon2020 projects and has been involved in various national and industrial projects too. Pavel authored more than 70 papers in scientific journals and in conference proceedings. He led the Czech mission in ISO/TC 37 and was a member of the W3C Uncertainty Reasoning for the World Wide Web Group. Email: smrz@fit.vut.cz

Title: Evaluating the multilingual capabilities of PERO-OCR with digitised historical newspapers: A Belgian case study

Historical newspapers are increasingly being analyzed in humanities research and thus digitized by cultural heritage institutions. In this process of digitization and OCRing, automatic classification, such as article segmentation, and full text indexing of the extracted text is completed most often with a language model to theoretically increase the quality of the results. In the case of multilingual sources, and further automating this process for large collections over different time periods, the efficiency of this approach is put to question. In this presentation we will present results of the OCCAM (OCR, ClassificAtion & Machine Translation) project's digital humanities case which implements a pipeline utilizing PERO-OCR, resulting in high-quality OCR of multiple languages of historical newspapers.

PERO is a novel, learning-based, fully adaptable and customizable open-source recognition engine that aims to improve accessibility of digitized historic documents. It is based on state-of-the-art methods from computer vision, machine learning (esp. deep neural networks), and language modeling. PERO extends automation and capabilities of digitization pipeline by providing tools for automated quality assessment and control, quality improvement, automated text transcription of historic printed documents, semi-automated handwritten text transcription, and automatic extraction of semantic information from semi-structured documents. Particular attention is paid to low-quality historic printed and handwritten documents that cannot be automatically processed by the currently available tools.

We explain how images of textual sources in multiple languages can efficiently be OCRed using a machine learning based model, enabling both the annotation of corrections and automatic text recognition of a set of multilingual historical newspapers from KBR- the Royal Library of Belgium's historical newspaper collections: BelgicaPress (Figure 1.). Through examples from a set of Dutch and French language newspapers from the early 1900s, we present the results obtained using PERO-OCR and compare this with the original OCR. This results in high-quality

OCR, that is flexible to diverse layouts of historical newspapers of varying quality (Figure 2.), with adaptation needed for line detection / layout identification (Figure 3.); and various combinations of printed and handwritten text (e.g. signatures in newspapers). In generating a PAGE XML format the PERO-OCR platform also affords an adaptable output for post-processing. This suggests that a general model in comparison to the novel learning-based recognition engine of PERO-OCR results in a flexible and adaptable tool for full text results for large multilingual collections.

Figure 1. Multilingual example from within one newspaper article of De Standaard from 1919.

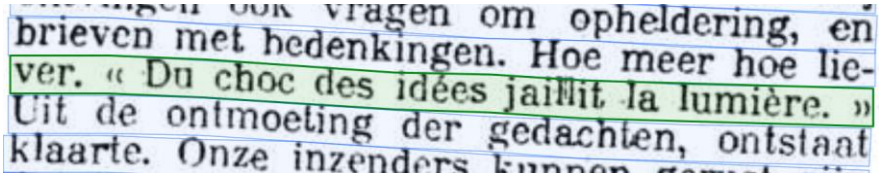

Figure 2. Signaling potential errors due to image quality issues

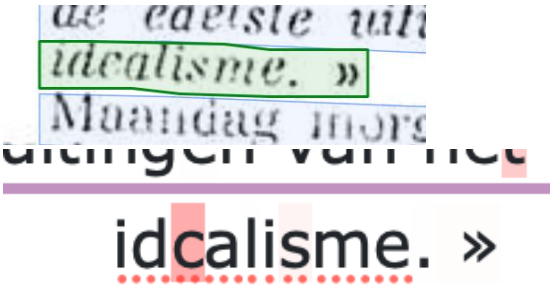

Figure 3. Layout identification