Transkribus: Handwritten Text Recognition technology for historical documents

Louise Seaward
louise.seaward@ucl.ac.uk
University College London, United Kingdom

Maria Kallio
maria.kallio@arkisto.fi
National Archives of Finland, Finland

Topic

Transkribus is a platform for the automated recognition, transcription and searching of handwritten historical documents. Transkribus is part of the EU-funded Recognition and Enrichment of Archival Documents (READ) project. The core mission of the READ project is to make archival material more accessible through the development and dissemination of Handwritten Text Recognition (HTR) and other cutting-edge technologies.

The workshop is aimed at researchers and students who are interested in the transcription, searching and publishing of historical documents. It will introduce participants to the technology behind the READ project and demonstrate the Transkribus transcription platform. Our team has already conducted 30 similar workshops over the course of 2016, including several sessions with digital humanities scholars and students.

Transkribus can be freely downloaded from the Transkribus website. Participants will be instructed to create a Transkribus account and install Transkribus on their laptops in advance of the workshop. They should bring their laptops along to the workshop.

The workshop will consist of five parts:

Introduction to Handwritten Text Recognition (HTR) technology

The introduction to this workshop will explain how new algorithms and technologies are making it possible for computer software to process handwritten text. Handwritten Text Recognition (HTR) technology works differently from Optical Character Recognition (OCR) for printed texts (Leifert et al., 2016). Rather than focusing on individual characters, HTR engines process the entire image of a word or line, scanning it in various directions and then putting this data into a sequence. This introduction will outline the workings of HTR technology and show examples of the successful automatic transcription and searching of historical documents. It will also explain the possibilities of working with different languages and styles of handwriting. The latest experiments demonstrate that Transkribus can automatically generate transcripts with a Character Error Rate of 5-10%. This means that 90-95% of the characters in the transcript would be correct.

Overview of the READ project

This presentation will give an overview of the READ project and the specific tools it is creating. Computer scientists working on READ are developing HTR technology using thousands of manuscript pages with varying dates, styles, languages and layouts. Testing the technology on a large and diverse data set will make it possible for computers to automatically transcribe and search any kind of handwritten document, from the Middle Ages to the present day, from old Greek to modern English. This research has huge implications for the accessibility of the written records of human history. The READ project is making this technology available through the Transkribus platform but is also developing other tools designed to make it easier for archivists, researchers and the public to work with historical documents. The workshop leaders will present prototypes of some of these tools. These include a system of automatic writer identification, an e-learning app to enable users to train themselves to read a particular style of writing, a mobile app to allow users to digitise and process documents in the archives and a crowdsourcing platform where volunteers can transcribe with the assistance of HTR technology. These tools will be open source and are designed to be used and adapted by other institutions and projects.

Introduction to Transkribus

HTR technology is made available through the Transkribus platform, which is programmed with JAVA and SWT (Mühlberger et al.) A transcription of a handwritten document can be undertaken in Transkribus for two main purposes. The first is a simple transcription – this allows users to train the HTR engine to automatically read historical papers. The second is an advanced transcription – this allows users to create a transcription of a document which may

serve as the basis of a digital edition. This presentation will explain both uses of Transkribus.

HTR engines are based on algorithms of machine learning. The technology needs to be trained by being shown examples of at least 30 pages of transcribed material. This helps it to understand the patterns which make up words and characters. This training material is known as 'ground truth' (Zagoris et al., 2012, Gatos et al., 2014). The workshop leaders will demonstrate how 'ground truth' training data can be prepared using Transkribus. Participants can work with images of their own documents, or experiment with test documents already on the system.

Transkribus can also be used simply for transcription. This presentation will explain how to create a rich transcription of a document in the platform, using structural mark-up, tagging, document metadata and an editorial declaration.

Working independently with Transkribus

In the last part of the workshop, the participants will be able to try out the functions of Transkribus on their own laptops. They will be supported by the workshop leaders, who will explain the different elements of the platform and then give participants the chance to practice each function for themselves. The workshop leaders will circulate around the room to answer any questions.

The workshop leaders will demonstrate the following tasks. After each demonstration, participants will be given 10-15 minutes to practice what they have learned.

- Document management – how to upload, view, save, move and export documents in standard formats (PDF, TEI, docx, PAGE XML)
- User management – how to allow specific users to view and edit documents
- Layout analysis – how to segment your documents to create training data for the HTR engines
- Transcription – how to create a rich transcript with tags and mark-up
- HTR – how to apply HTR models to automatically generate transcripts, how to conduct a keyword search of your documents, how to assess the accuracy of automatically generated transcripts

Question and Answer

The workshop will close with a Question and Answer session where participants can clarify anything they are unsure about. They will also have the opportunity to provide feedback on the Transkribus tool via our user survey.

Organizers

Louise Seaward

Dr. Seaward received her PhD in History from the University of Leeds (United Kingdom) in 2013. She is currently a research associate at University College London where she coordinates 'Transcribe Bentham', the scholarly crowdsourcing initiative which asks members of the public to transcribe manuscripts written by the British philosopher Jeremy Bentham (1748-1832). Outside of digital humanities and Bentham, her research interests relate to the history of censorship and the Enlightenment.

Maria Kallio

Maria Kallio works as a Senior Research Officer at the National Archives of Finland where she is responsible for collections, crowd-sourcing and dissemination within the READ project. Currently she is also finishing her PhD in History at the University of Turku (Finland). In addition to digital humanities, her research interests include medieval literacy and written culture in all its diversity.

Proposed audience

Humanities and digital humanities scholars, archivists, librarians, computer scientists

Guidelines for Participants

Participants should register to attend the workshop by sending an email to Louise Seaward. Participants will need to bring their own laptops and install Transkribus before attending the workshop. If participants are interested in working with their own documents, they should bring a selection of digital images to the workshop. Otherwise, it will be possible to work with test documents already on the platform.

Bibliography

Leifert, G., Strauß, T., Grüning, T., and Labahn, R. (2016). 'Cells in Multidimensional Recurrent Neural Networks', https://arXiv.org/abs/1412.2620v02

Mühlberger, G., Colutto, S., Kahle, P., (forthcoming) 'Handwritten Text Recognition (HTR) of Historical Docu-

ments as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP)' (pre-print)

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J.A., Toselli, A.H., and Vidal, E. (2014). 'Ground-Truth Production in the tranScriptorium Project', Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on Document Analysis Systems, 237-244

Stamatopoulos, N., and Gatos, B. (2015). 'Goal-oriented performance evaluation methodology for page segmentation techniques', 13th International Conference on Document Analysis and Recognition (ICDAR), 281-285.

Konstantinos, Z., Pratikakis, I., Antonacopoulos, A., Gatos, B., and Papamarkos, N. (2012). 'Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm", in: Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference, Bari, 103-108. DOI: 10.1109/ICFHR.2012.207.