THE NATIONAL ARCHIVES
SERVICE OF FINLAND

Transkribus

Maria Kallio
Senior Research Officer
National Archives of Finland
maria.kallio@arkisto.fi

13.9.2016

READ

Transkribus is

- A downloadable platform for the automated recognition, transcription and searching of historical documents.
- Go to: http://transkribus.eu

Functions

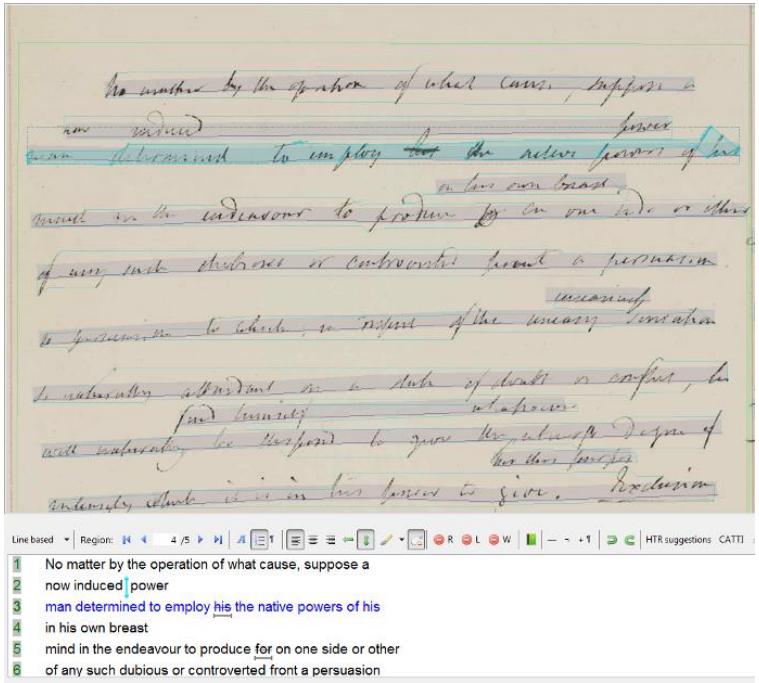
- HTR-engine
- OCR-editor
- Tagging index words
- Keyword spotting
- Writer identification
- Diverse search tools

THE NATIONAL ARCHIVES SERVICE OF FINLAND

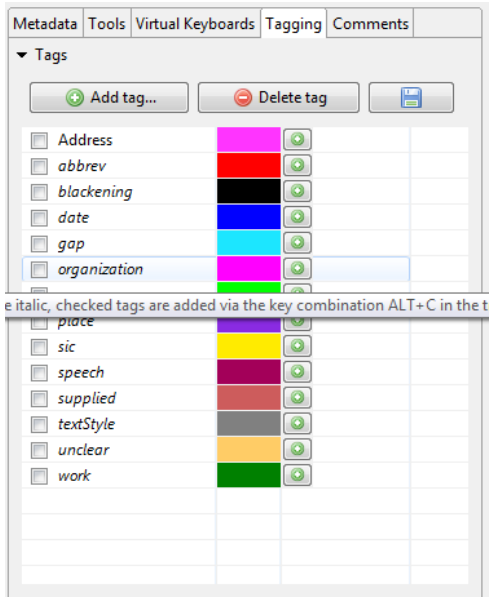
READ

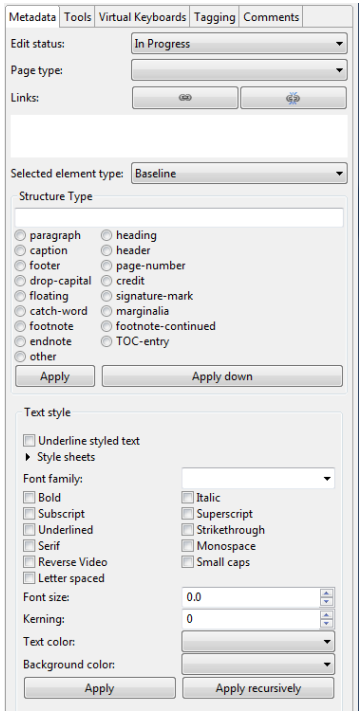
A transcription undertaken in Transkribus can be used for two main purposes:

1. **Simple transcription –** to train the HTR engine



2. **Advanced transcription** – to create a rich transcript (could be basis for digital edition)





THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

Workflow – preparing documents for HTR

Register for a Transkribus account

Download Transkribus

Upload images of your document collection to Transkribus

Create at least 30 pages of “ground truth” training material:
1. segmentation,
2. transcription

Contact Transkribus team to generate a HTR model for your collection

Use this model to generate transcripts of the rest of your collection

Check and correct the results of the HTR

Export your transcripts (PDF, XML, TEI, RTF etc.)

THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

Ground Truth

The HTR engine needs to be trained by being shown examples of transcribed documents

Representative selection from a manuscript collection, at least 30-50 pages

Each document needs to be:

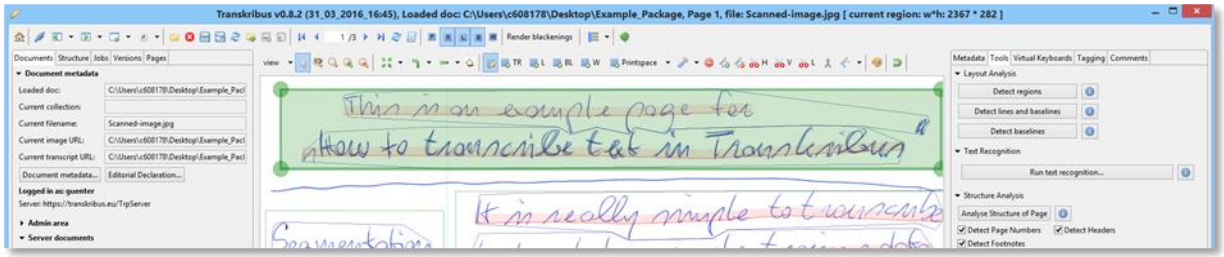
1. **Segmented**
2. **Transcribed**

THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

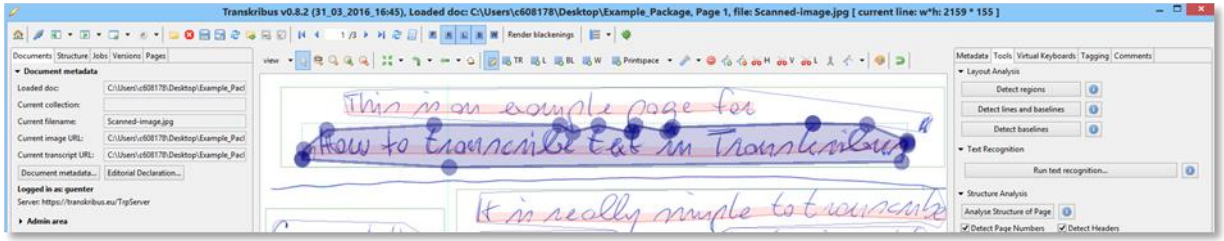
1. Segmentation

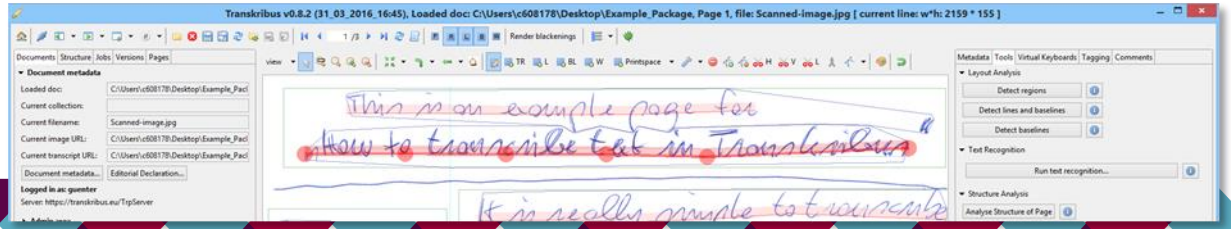
Lines in the manuscript image need to be **segmented** to connect them to the lines of the text transcript.



Text regions

Line regions





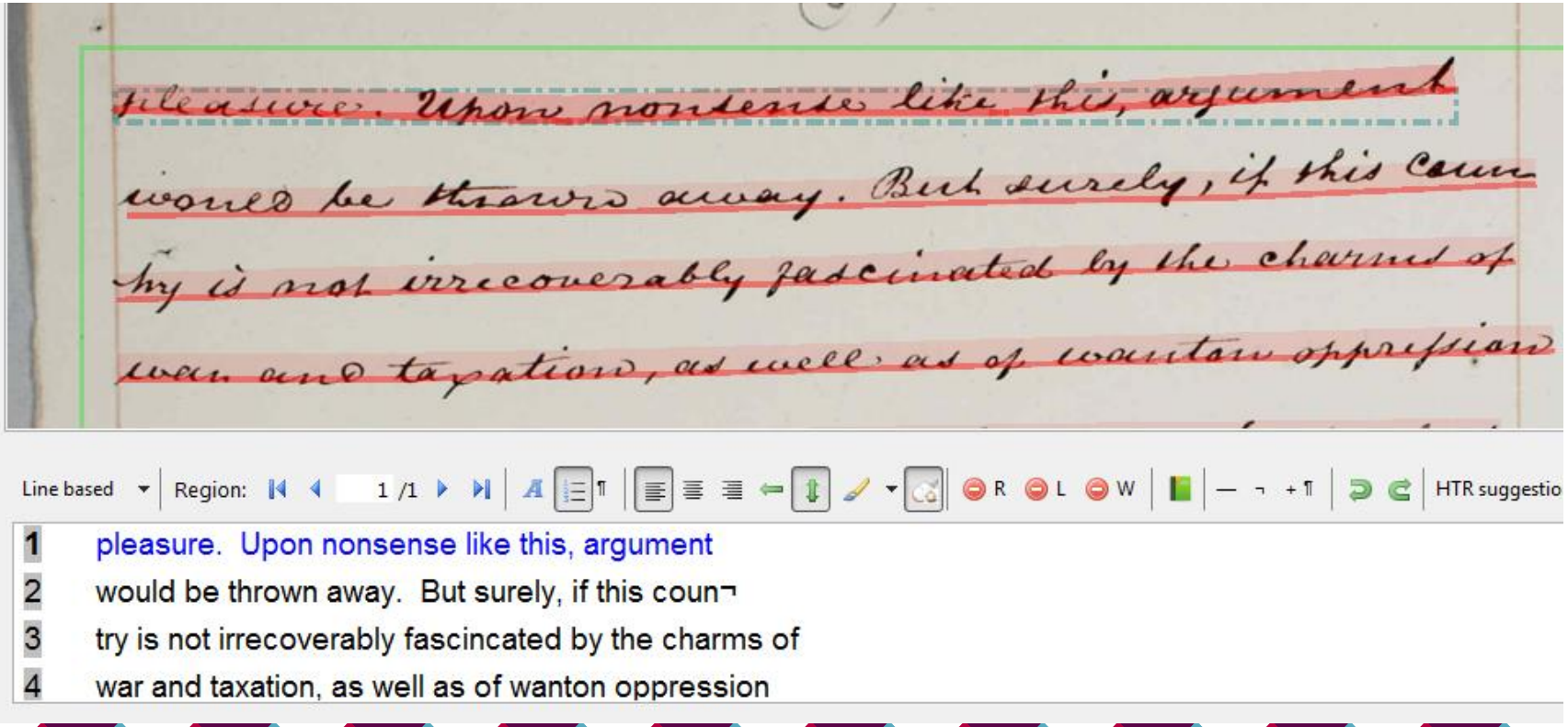
Baselines

THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

2. Transcription

Transcription should be diplomatic - words transcribed exactly as they appear on the page.



THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

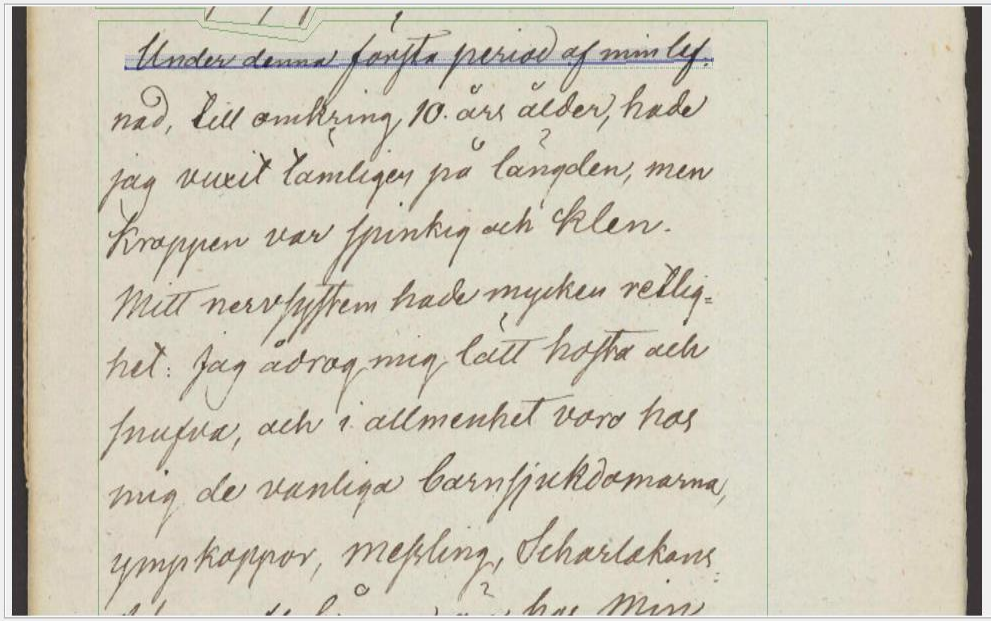
Using HTR technology

- Once ground truth has been created, the Transkribus team can generate a HTR model for your collection
- Use this model to automatically transcribe the rest of your collection!
- Measure accuracy of HTR using Character and Word Error rates – results of 10-20% are possible

The First HTR-model for NAF

- Memoirs of Albrecht Fredrik Richard de la Chapelle (1785–1859) *Underrättelser för Mina Barn*
- 800 images, of which 50 were transcribed and proofread

Transkribus v0.8.7 (10_08_2016_14:58), Loaded doc: Copy of NAF_504_Original_Scans_Training_Set_TEST, ID: 5060, Page 1, file: 1325022.jpg [Image Meta Info: (Resolution:72.0, w*h: 2908 * 3495)] [current line: w*h: 1759 * 53]



Under denna första period af min lef-
nad, till omkring 10 års ålder, hade
jag vuxit tamligen på längden, men
kroppen var spinkig och klen.
Mitt nervsystem hade mycken retlig-
het: jag ådrog mig lätt hosta och
snufva, och i allmenhet voro hos
mig de vanliga barnsjukdomarna,
ymp koppor, mäsling, Scharlakans

Line based | Region: 3 /3 | HTR suggestions CATTI

1 under·denna·första·person·af·min·lif
2 med·,·till·omkring·10·års·ålder·hade
3 jag·visat·tämligen·på·lång·den·men
4 kropp·en·var·synlig·och·klen
5 Matts·merna·hösten·hade·mycken·måttlig
6 het·.·jag·ådrag·mig·lätt·hosta·och
7 hufvud·och·i·allmenhet·voro·hos
8 mig·de·vanliga·barn·sjukdom·armen
9 yppig·köp·hur·,·Med·ting·Schack·lekamen
10 telser·nas·svära·re·an·hos·Min

Virtual archive

VENICE TIME MACHINE

Crowdsourcing

- Web interface in preparation
- The goal is to improve the accessibility and usability of extensive archival collections
- Transcribe Bentham - An example of a successful project

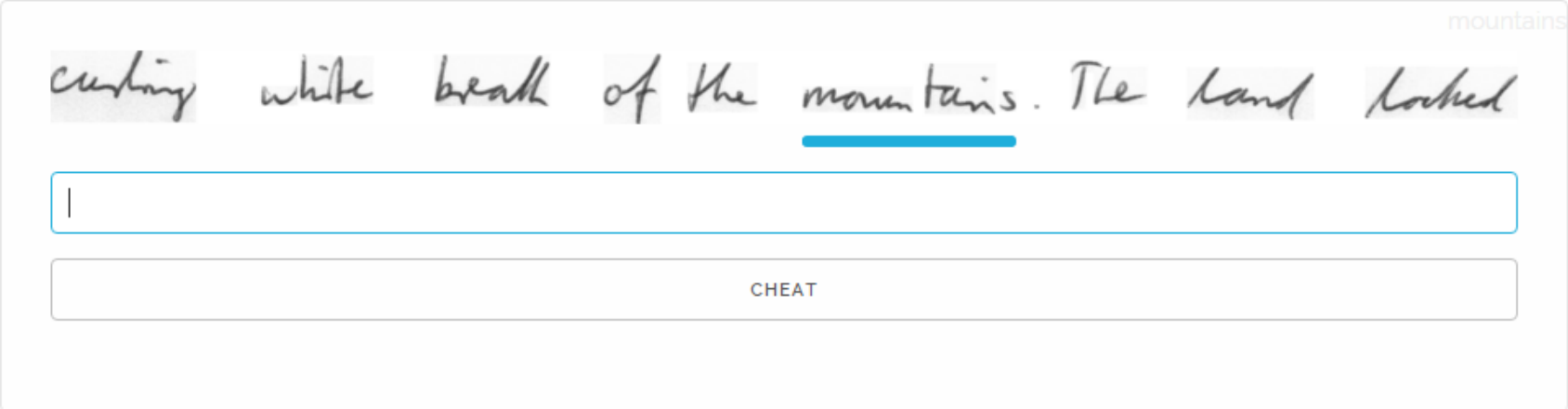
Digital editions

- Enables new forms for publications
- Nådendalsdiplomen (Diplomatarium Fennicum)
- Letters of Alfred Escher

THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

E-learning tutorial – practice reading a particular hand

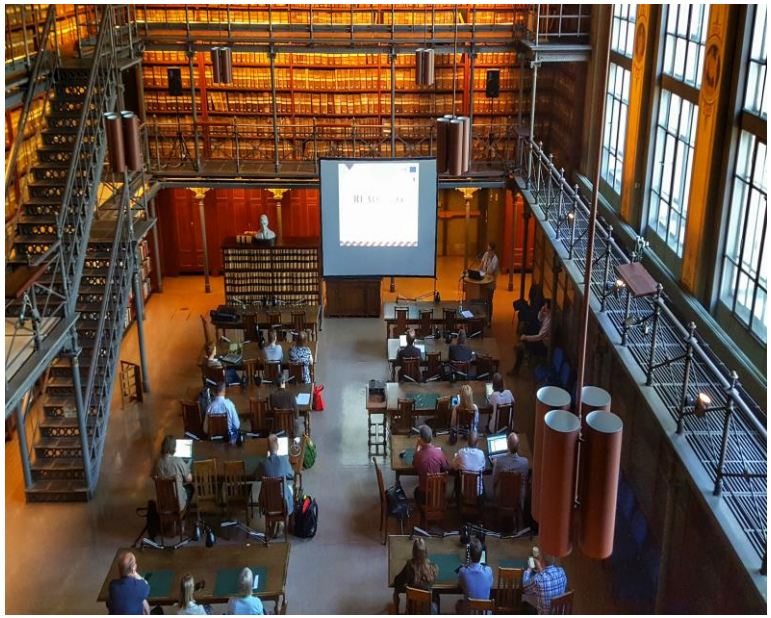


THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

Spreading the word…

- Transkribus services are free
- Open access, Open source https://github.com/transkribus/
- email@transkribus.eu for advice/help
- Possible to become part of READ with a **Memorandum of Understanding**



THE NATIONAL ARCHIVES SERVICE OF FINLAND

READ

Transkribus website: https://transkribus.eu/

Transkribus wiki: https://transkribus.eu/wiki/

READ website: https://read.transkribus.eu/

Transkribus github: https://github.com/transkribus/

Contact: email@transkribus.eu