Transkribus,
a research platform for the mass digitisation, transcription, recognition and searching of historical documents

Günter Mühlberger
University of Innsbruck,
Digitisation and Digital Preservation Group
Agenda

• Why this meeting?
• Transkribus - Future
• Some updates on technology
Why this meeting?
There is a great chance...

- Excellent prerequisites
- Digitisation by archives and libraries
- Digital Humanities projects
- Long term tradition in (digital) editing
- New technology, but already mature enough for broad usage
...for a Dutch Model

- A large dataset of transcripts available to everyone who wants to train machines to read historical Dutch documents
  - Archives, libraries
  - Scholars
  - Computer scientists
- One or more neural networks ("models") capable to read any kind of Dutch handwriting of the last 300-400 years with reasonable results
  - Reasonable results would be something around 10% Character Error Rate without further training or adaption
  - Based on Keyword Spotting this will already allow searching with high accuracy
- Data curation as the main task
  - Its not so much an IT task but the task to build and maintain a national dataset of transcriptions
Purpose of the meeting..

- Someone needs to take the lead...
- Spotting and collecting existing digital (and printed) editions
- Negotiating with data providers
- Transforming editions into machine readable data
  - Partly with automated Text2Image matching
  - Partly with support of service providers or crowd
- Transkribus is the logical place for this dataset, but the dataset can of course be used independently from Transkribus
Future of Transkribus
Transkribus future

• Projects ends on 30\textsuperscript{th} June 2019
• However, there is a strong demand for Transkribus services so that maintenance of the platform is already safeguarded until 2021
  • EU Project NewsEye (2018-2021)
  • German Science Funds project (2019-2020)
  • Project with National Archive Finland (2019)
  • Project with National Archive Netherlands (2019-2020)
  • National project with cadastre documents from Tyrol (2019-2020)
  • Project with Trinity College Dublin (2019-2021)
  • Project with State Archive Zurich (2019-2020)
  • More to come and under negotiation!
READ-COOP

• European Cooperative Society (SCE) as the legal framework
• Run and further develop the Transkribus platform
• Collaboration of independent entities
• Democratic constitution – members have the final say in the general meeting
• Customers become owners, owners become customers
• Direct benefit of members is the main goal
• No shareholder value
• Open to natural persons as well
Current state of affairs

• Statutes in pre-final version
  • Monistic system: Board of directors – general meeting

• Founding members
  • University Innsbruck, University Greifswald, Technical University Valencia, National Archive Finland, British Library, University Library Belgrade, Diocesan Archive Passau, University Rostock, ZAMG Vienna, Picturae, Geneanet France, etc..
  • Everyone invited to join!

• Membership shares
  • From 1000 (minimum) to 5000 EUR, 250 EUR for natural persons

• First board of directors will be formed in the next weeks
• Founding act shall take place before summer holidays
• Official start of business on 1\textsuperscript{st} of July (keep fingers crossed!)
Business

- Transkribus platform
  - Software as a service: We expect that users will work with the platform and connected services in a rather independent way (as it is already the case)
  - However, specific support and development tasks can also be offered for special projects

- Subscription fee for the platform itself
  - Around 3000-4000 EUR per year

- Page based prices for services
  - Between 10 and 16 Cent per page for HTR processing
  - Large quantities can be negotiated

- Discounts for coop members
Recent work and updates
Error rate tool – Advanced mode and sample mode
### Advanced Statistics

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>CER</th>
<th>Word Acc</th>
<th>Char Acc</th>
<th>Bag Tokens Prec</th>
<th>BT Recall</th>
<th>BT F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>20.79%</td>
<td>5.86%</td>
<td>79.21%</td>
<td>64.14%</td>
<td>0.835</td>
<td>0.822</td>
<td>0.835</td>
</tr>
<tr>
<td>Page 1</td>
<td>13.31%</td>
<td>4.22%</td>
<td>86.89%</td>
<td>95.76%</td>
<td>0.8096</td>
<td>0.8978</td>
<td>0.8936258059...</td>
</tr>
<tr>
<td>Page 2</td>
<td>17.97%</td>
<td>7.2%</td>
<td>82.03%</td>
<td>92.8%</td>
<td>0.8409</td>
<td>0.8672</td>
<td>0.85364615384...</td>
</tr>
<tr>
<td>Page 3</td>
<td>36.15%</td>
<td>8.84%</td>
<td>63.85%</td>
<td>91.16%</td>
<td>0.696800000000000...</td>
<td>0.8308</td>
<td>0.75769473684...</td>
</tr>
<tr>
<td>Page 4</td>
<td>8.03%</td>
<td>2.02%</td>
<td>91.07%</td>
<td>97.65%</td>
<td>0.9212</td>
<td>0.9244</td>
<td>0.92281300302...</td>
</tr>
<tr>
<td>Page 5</td>
<td>12.64%</td>
<td>2.76%</td>
<td>87.86%</td>
<td>97.24%</td>
<td>0.8917</td>
<td>0.8994</td>
<td>0.89555590681...</td>
</tr>
<tr>
<td>Page 6</td>
<td>28.23%</td>
<td>7.05%</td>
<td>79.75%</td>
<td>92.95%</td>
<td>0.8354</td>
<td>0.8354</td>
<td>0.835445035797...</td>
</tr>
<tr>
<td>Overall</td>
<td>19.91%</td>
<td>5.11%</td>
<td>79.21%</td>
<td>64.14%</td>
<td>0.835</td>
<td>0.822</td>
<td>0.835</td>
</tr>
</tbody>
</table>

---

**Error Rate Chart** | **Ref:** GT | **Hyp:** CITlab HTR: Amerikauswanderer_M1+

![Error Rate Chart Image](chart.png)

**Base folder:** C:\Users\c088178\Desktop

**File/Folder name:** Docid_142992

**Export path:** C:\Users\c088178\Desktop\Docid_142992.xls

Download XLS
Belebung zu Kriegsgefangenen gemacht, wirr.


Verlust der Engländer hat D. Gálvez nichts bes. dem richt. fähnen, weil sie ihn mit der ordentlichen Her
With the probability of 95% the CER for the entire document will be in the interval [3.879%, 11.019%] with the mean 7.440%.

By taking 4 times the number of lines the interval size can be cut in half. The CER for the sample pages is 2.93%.

Confidence Interval for CER

15:07:51.617  15:07:51.618
Sample

Upper and lower bound for CER
Browser based transcription interface (including tagging and baseline correction)
In Publiko = Politicus.

Aktum Boven den zweyten Jänner 1794
Vor Titl. Herrn Bürgermeister Joseph v. Remich
kaiserl. königl. Regierungsrah und T. L. M:
In beyseyn der Herrn Herrn Magistratsräthe ...
Oberlieutenant Karl v. Kuglern.
Technical updates

Simple search interface for New Zealand Alpine Journal
http://nzaj-archive.nz/
Volume 54, 2002, Page 84

... Mount Kenya, while I'm in Amsterdam, eating Dutch chocolate and ...

Volume 54, 2002, Page 88

... pickup point outside Amsterdam airport. I didn't bring you a latte ...
No summit on Mount Kenya

by Simon Carr

Have fun," Amy said as she walked through the departure gate at Nairobi airport. "I'll be thinking of you living it up on Mount Kenya, while I'm in Amsterdam, eating Dutch chocolate and sipping lattes. It will be tough."

That afternoon David and I took our gear to the campground in Nairobi where he had been staying. On entering the gate, it was as if I'd stepped back to my months in southeast Asia in 1985. I was surrounded by twenty-something Israelis, Brits, Germans... well-thumbed copies of Lonely Planet, Let's Go, Rough Guides on every table alongside the handrolled cigarettes. I instantly felt too old for this scene.

I'd met David a few years earlier in Boston, and we'd climbed together for a couple of seasons. He'd moved to Boulder, got good, became a refugee from the dotcom meltdown. His road trip had lasted nine months now, from Australia and New Zealand to Asia and on to Africa. So when Amy and I decided to do the safari thing in Africa, I emailed David suggesting we climb Mount Kenya. With internet connections everywhere, there is no longer the isolation I'd felt in 1985 when there were only post restante addresses in Kathmandu and Delhi for contact. However email wasn't quite foolproof...

"We've going light, I see," David said. "One 9mm rope, eh?"

I looked at him somewhat surprised. "Don't you have one?"

"No," he replied. "I thought you were bringing the gear."
Trainable layout analysis tool
1-1 7
2-1 Conclusum.
3-1 Conclusum.
New training interface for HTR models
ScanTent now delivered for testing to interested users all around the world
Yesterday...
• 19.258 images uploaded by single users
• 82 new users
• 322 unique logins to Transkribus expert client
• 1246 jobs processed
• 5 new HTR models trained by users
Thank you for your attention!

Further information

https://read.transkribus.eu/

https://transkribus.eu/

https://read.transkribus.eu/coop/

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 674943.