Collaborating on Machine Reading: Training Algorithms to Read Complex Collections

Carrie Pirmann, Brian King, Bhagawat Acharya ‘20, Katherine Faull
Bucknell University

https://tinyurl.com/BUDSC19MoravianLives
Moravian Lives

Tracing the Movements and History of Members of the Moravian Church (1750–2012)

Search among over 60 000 Memoirs written by members of the Moravian church from 1750 to the present day. The map visualizes place and date of birth and death of the authors.

The Moravian Memoirs project is a collaboration between the Center for Digital Humanities and the Centre for Critical Heritage Studies at University of Gothenburg, Bucknell University, USA, The Moravian Archive in Herrnhut, Germany and The Moravian Archive in Bethlehem, USA.
Moravian Memoirs

The Moravian archives in Herrnhut, Germany and Bethlehem, PA house over 65,000 memoirs or “ego-documents” of members of the church, dating from 1750 to the present day.

There are many other smaller archives around the world with more uncatalogued and undigitized documents.

The corpus is an important resource for historians of religion, gender, race, linguists, and genealogists.
Prof. Katie Faull, Moravian Lives, PI
Dr. Diane Jakacki, Encoding, Project Manager
Carrie Pirmann, Transkribus Project Manager
Carly Masonheimer, Student Researcher
Marita Gruner, PhD Student, Transcriber
Jess Hom, Student Researcher
Marleina Cohen, Student Researcher
Morgan McMullen, Student Researcher
Bhagawat Acharya, Student Researcher
Justin Schaumberger, Student Researcher
Mike McGuire, Programmer
Prof. Brian King, Deep Learning, Computer Science
Transcription Desk

Built with Scripto plugin in a WordPress frame

---

Brief notices concerning the life and happy departure from this world late evening of the late Miss Marianne Spence who fell asleep in Jesus, on Monday morning, June the 15th, 1866: at Fulneck.

Our late Sister was born on October the 10th 1844 at the Mission Station of Bethabara in the Island of Jamaica where her parents laboured in the Lord’s Vineyard as esteemed and useful servants of the Brethren Church. From her very birth and more especially on the day of her baptism in November the 24th she was by her parents dedicated unto the Shepherd and Bishop of our souls as His redeemed property, with the servant prayer that He himself would lend and keep her as one of His sheep:- This prayer was answered: for the Lord our Saviour was ever mindful of her and through the days of her childhood & youth and to the very end of her sojourn here below manifestly led her with his rod and staff— The blessed name of Jesus she learnt to know and to love from her earliest years. — Brought up as a little child in the nurture and admonition of the Lord under the parental roof she early learnt to love the house, and the word of the Lord, and to pray. — She was naturally not of a robust constitution, a weakness in her chest was perceptible even in her youngest years tho’ as far as I can hear, she never materially offered from this cause at that time. — In the
Challenges to Crowdsourcing Transcriptions

- Slow, painstaking, and meticulous work
- We need transcribers who have some facility with reading cursive; this is becoming more difficult to find among students who might want to work on the project
- Crowdsourcing produces results that vary in levels of accuracy
- Our rate of images being acquired is far exceeding the manual transcription rate (i.e., a weeklong trip to the Archives may result in digital images for 100 memoir documents, which equals hundreds of pages to be transcribed)
February 1772

Christian Bell was born in London Decr 7th
1742 and came as a little Boy into the Children's
Economy, he was a very lovely Child, beloved by all
about him. When he arrived to his Boys Age, he
was put apprentice to our dear late Br Charlesworth
in the Merchant Business, during which time he
behaves himself as an orderly & faithful Apprentice
and gave full Satisfaction to Br Charlesworth & others.

1-1 February 1772
1-2 Christian Bell was born in London Decr 7th
1-3 1742 and came as a little Boy into the Children's
1-4 Economy, he was a very lovely Child, beloved by all
1-5 about him. When he arrived to his Boys Age, he
1-6 was put apprentice to our dear late Br Charlesworth
1-7 in the Merchant Business, during which time he
1-8 behaved himself as an orderly & faithful Apprentices,
1-9 and gave full Satisfaction to Br Charlesworth & others
1-10 about him. I was evident that there was a Work
1-11 of Grace in his Heart, be those who were acquainted
What is Transkribus?

- Comprehensive platform for the automated recognition, transcription, and searching of historical documents
- Originally developed at University College, London; development team is now housed at University of Innsbruck (Austria)
- Uses predefined handwriting models to machine transcribe handwritten documents (handwritten text recognition, or HTR)
- Users can “train” their own models, specific to the handwriting style(s) of the documents in a corpus
- Accuracy rates of 95% and above are possible with adequate amounts of training data
Transkribus Workflows

- Establishing a training corpus of data from the Moravian memoirs
  - Batch upload digital images to Transkribus’ servers, and match existing transcriptions to digital images to ensure diplomatic transcriptions of the documents
- These documents have been used to train HTR models, which can be reconfigured and tweaked as necessary, with the goal of achieving a CER (character error rate) of between 5-10%.
- Apply HTR models to new, untranscribed documents; assess models for best results
- The resultant transcriptions generated by Transkribus still need to be checked by humans, a job done by Carrie, Katie, Jess, and Morgan.
Building an HTR Model

- Eyeballing for similar handwriting styles to build models - sometimes leads to good results; sometimes not
- Have created 30+ models, best CER is 5.42%
- Other models with CER of 11% or lower have performed well against documents
Refining Models

Samuel Benade memoir

First computer generated transcription
Refined Model
differing indeed both in arrangement & culture, but all under the care and management of one & the same Gardener; who appoints Laborers for each division of his garden & vineyard, instructing them in the proper mode of cultivating that part which is exclusively committed to them. How it is well known, that different plants require both a different treatment and a different soil, & would not equally prosper in a soil less congenial to their texture and nature. So it is with the different denominations in the Christian Church. All of them, that confess Jesus Christ to be their Lord and Master, and walk in the way of his precepts and commandments deriving from him wisdom, righteousness, sanctification and redemption, and keeping the word of his patience, belong to the Lord's vineyard, & occupy different compartments where they receive a different treatment under their peculiar form and constitution; there they find themselves at home and thrive best. Thus there are many excellent Children of God, whom our constitution, discipline and outward form would not suit, and who would not like to be confined within our fold. And thus on the
Room for Improvement

Our late sister was born on the 11th of Feb
8th at Hecford Cwm Tenbis in the county of
Yorkshire; the god brought her to the
principles and doctrines of the established
Church of England and by a pious Mother
had early looked in the paths of virtue, and
though she lost her Mother in early age yet
the impressions she had received from her
Mother's godly advice, to be puictco at thousete and li the reatirgy
of the ward of God wese nerer fryotterrn.
and oose a soure of glat blessing to hu he
Subsequently the Holy Dpont earhor to fe
hnsf os a pone cunce, and to fix to
(hist for mrey. _ Oer late Tiste opegain
Deep Learning: Making HTR a reality

What is deep learning?

Christian Bell was born in London Decr 7th
Alan Turing

1. The Invitation Game.

I propose to consider the question, 'Can machines think?'
This should begin with definitions of the meaning of the terms 'machine' and 'think'. These definitions might be framed so as not to reflect any possible kind of use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?', is to be sought in a statistical survey such as a college poll. But this is absurd. Instead of attempting such a definition, I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator sits in a booth apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to X and Y in any order. If X is A, then A must answer. It is A's turn answer.
Human Learning vs. Machine Learning

We study

We take an exam

We assess our performance

We (hopefully) makes adjustments if performance is low!

72% of 1000 images correct
Machine Learning

A ML algorithm:
“Learns” about your data

identifies key **features** in your data that
distinguish between different labels

Data... and LOTS of it!

More importantly, “labeled” data

Program that can automate labeling of data
And the list goes on...
A Standard Machine Learning Pipeline

https://www.datanami.com/2018/09/05/how-to-build-a-better-machine-learning-pipeline/
Learning curve

AI → Machine Learning → Deep Learning

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://developer.nvidia.com/deep-learning
The rise of “deep” learning

“Deep learning is getting lots of attention lately and for good reason. It’s achieving results that were not possible before.” - Mathworks.com

Facial recognition

Autonomous driving

Image restoration

Handwritten Text Recognition (HTR)
Automatic feature extraction

A deep learning model learns key “features” of the input that distinguish between labels.
“Deep” learning - The take away

- Traditional machine learning:
  - Great on smaller, well-defined classification / regression problems from **structured** data
  - Limited results on large, high-dimensional, unstructured data
    - Image, audio, video, time series, etc.
  - Often requires substantial time to prepare / reduce data so that methods could learn

- Deep learning
  - Works with complex, high-dimensional, large datasets
  - Works with **unstructured** data

- Deep learning - there is no free lunch
  - Computational costs can be huge
  - LOTS of data required to perform well
Transkribus developers leverage deep learning to create a series of HTR models
The Transkribus pipeline

- Training examples
  - One folder for printed lines, another for handwritten lines
  - As many text lines as possible
- Data augmentation
- Training
- Neural network
- Input text line
- Classification result

https://read.transkribus.eu/
Transkribus live demo
Challenges

- **Transkribus**
  - Supplies only one English handwriting model, which was created based on a corpus of materials with one single handwriting style
  - User is also allowed to train their own models
- **For Moravian Lives Project?**
  - Not all images are “clean”!
  - Inconsistent digitizing (using a smartphone camera!)
  - Artifacts from bleedthrough, poor image quality, paper decomposition from age, etc.
  - The default Transkribus English handwriting model is not a good match with the majority of our documents
  - Scribes can often exhibit substantially different script styles
Challenges

Poor image quality
Challenges

Unique handwriting style
Improving the HTR models

A model that is built on handwriting that is representative of samples from the Moravian Lives project will perform better than the default English model, which is built from only one handwriting PROVEN! But, can we do even better?

A model built from highly similar handwriting should perform even better than a single, broadly trained model

QUESTION: How can we identify similar handwriting?
Scribe identification?

- We do not know the identity of the scribe for every memoir in the Fulneck corpus
- Human intervention?

Our approach:
- We made an assumption on a handful of examples from the same memoir that they also used the same scribe
Auto identification of the scribe

Hypothesis: Developing a model to predict the scribe can prove that deep learning can distinguish between handwriting styles.

How? Get good, high quality handwriting examples from different scribes, and train a deep learning model!

1. Improve the selection of handwriting on each image to eliminate artifacts not representing handwriting
2. Create small, random “patches” representing snippets of script from different scribes
   ○ Why? We don’t care about the text, only the style of the handwriting.
   ○ These are our input into a model
   ○ Output? The scribe
Generate and train a dataset

- We have images from different memoirs.
  - Divide the images from each memoirs for training and testing
- For every image, generate random patches of same size
- Patches are filtered to maximize the likelihood of capturing ONLY patches that contain text.
  - Filtered using the standard deviation to determine if a patch is mostly the same color (probably no writing) or different colors (probably writing)
  - Currently working on better ways to generate patches
- Classify these patches and train the model
Example Patches
Example Patches (continued)
Model is trained from 300 patches per memoir using 6 memoirs

For a new image, select 15 random patches from the image

For any image, the scribe with the most predictions over all patches wins the prediction!
<table>
<thead>
<tr>
<th>Memoir</th>
<th>Prediction</th>
<th>Molther Philip Henry</th>
<th>Jungmann</th>
<th>Mack</th>
<th>Wittke</th>
<th>Ramftler</th>
<th>Benade Samuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benade Samuel</td>
<td>IMG_4320.jpg</td>
<td>0.0005</td>
<td>0.0000</td>
<td>0.0679</td>
<td>0.0000</td>
<td>0.0042</td>
<td>0.9274</td>
</tr>
<tr>
<td></td>
<td>IMG_4322.jpg</td>
<td>0.1316</td>
<td>0.0001</td>
<td>0.1157</td>
<td>0.0001</td>
<td>0.0013</td>
<td>0.7512</td>
</tr>
<tr>
<td></td>
<td>IMG_4329.jpg</td>
<td>0.0011</td>
<td>0.0001</td>
<td>0.0439</td>
<td>0.0002</td>
<td>0.0235</td>
<td>0.3012</td>
</tr>
<tr>
<td></td>
<td>IMG_4330.jpg</td>
<td>0.0106</td>
<td>0.0001</td>
<td>0.0090</td>
<td>0.0001</td>
<td>0.0027</td>
<td>0.9778</td>
</tr>
<tr>
<td></td>
<td>IMG_4332.jpg</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0043</td>
<td>0.9656</td>
</tr>
<tr>
<td></td>
<td>IMG_4343.jpg</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0003</td>
<td>0.9997</td>
</tr>
<tr>
<td></td>
<td>IMG_4561.jpg</td>
<td>0.3444</td>
<td>0.0116</td>
<td>0.1273</td>
<td>0.0014</td>
<td>0.0164</td>
<td>0.5089</td>
</tr>
<tr>
<td>Jungmann</td>
<td>Jungmann.jpg</td>
<td>0.1009</td>
<td>0.6639</td>
<td>0.0558</td>
<td>0.1645</td>
<td>0.0154</td>
<td>0.0493</td>
</tr>
<tr>
<td></td>
<td>Women_s memoirs - 126 of 192.jpg</td>
<td>0.0000</td>
<td>0.9999</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>Women_s memoirs - 129 of 192.jpg</td>
<td>0.0034</td>
<td>0.9381</td>
<td>0.0045</td>
<td>0.0534</td>
<td>0.0005</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>Women_s memoirs - 131 of 192.jpg</td>
<td>0.1265</td>
<td>0.7995</td>
<td>0.0006</td>
<td>0.0656</td>
<td>0.0018</td>
<td>0.0061</td>
</tr>
<tr>
<td></td>
<td>Women_s memoirs - 153 of 192.jpg</td>
<td>0.0269</td>
<td>0.8127</td>
<td>0.0001</td>
<td>0.0692</td>
<td>0.0855</td>
<td>0.0057</td>
</tr>
<tr>
<td>Mack</td>
<td>10.jpg</td>
<td>0.0011</td>
<td>0.0000</td>
<td>0.9272</td>
<td>0.0004</td>
<td>0.0008</td>
<td>0.0705</td>
</tr>
<tr>
<td></td>
<td>19.jpg</td>
<td>0.0011</td>
<td>0.0000</td>
<td>0.8872</td>
<td>0.0001</td>
<td>0.0670</td>
<td>0.0445</td>
</tr>
<tr>
<td></td>
<td>3.jpg</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.9064</td>
<td>0.0026</td>
<td>0.0001</td>
<td>0.0005</td>
</tr>
<tr>
<td>Molther Philip Henry</td>
<td>Ful_3_04_MoltherPhilipHenry-01.jpg</td>
<td>0.8713</td>
<td>0.0009</td>
<td>0.0630</td>
<td>0.0271</td>
<td>0.0037</td>
<td>0.0341</td>
</tr>
<tr>
<td></td>
<td>Ful_3_04_MoltherPhilipHenry-10.jpg</td>
<td>0.7829</td>
<td>0.0600</td>
<td>0.1200</td>
<td>0.0197</td>
<td>0.0112</td>
<td>0.0061</td>
</tr>
<tr>
<td></td>
<td>Ful_3_04_MoltherPhilipHenry-25.jpg</td>
<td>0.7522</td>
<td>0.1658</td>
<td>0.0006</td>
<td>0.0255</td>
<td>0.0517</td>
<td>0.0041</td>
</tr>
<tr>
<td>Ramftler</td>
<td>IMG_4718.jpg</td>
<td>0.0011</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0721</td>
<td>0.0284</td>
</tr>
<tr>
<td></td>
<td>IMG_4733.jpg</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9954</td>
</tr>
<tr>
<td></td>
<td>IMG_4747.jpg</td>
<td>0.0023</td>
<td>0.0593</td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0829</td>
<td>0.1150</td>
</tr>
<tr>
<td></td>
<td>IMG_4750.jpg</td>
<td>0.0620</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.8998</td>
</tr>
<tr>
<td></td>
<td>IMG_4757.jpg</td>
<td>0.0466</td>
<td>0.0182</td>
<td>0.0002</td>
<td>0.0016</td>
<td>0.0053</td>
<td>0.0267</td>
</tr>
<tr>
<td></td>
<td>IMG_4768.jpg</td>
<td>0.0645</td>
<td>0.0009</td>
<td>0.0561</td>
<td>0.0001</td>
<td>0.8562</td>
<td>0.0222</td>
</tr>
<tr>
<td>Wittke</td>
<td>Women_s memoirs - 112 of 192.jpg</td>
<td>0.1328</td>
<td>0.6283</td>
<td>0.0815</td>
<td>0.1751</td>
<td>0.0012</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>Women_s memoirs - 113 of 192.jpg</td>
<td>0.2785</td>
<td>0.3448</td>
<td>0.0803</td>
<td>0.1580</td>
<td>0.0581</td>
<td>0.0824</td>
</tr>
</tbody>
</table>
Prediction Visualized
Prediction Visualized

The bottom graph shows an incorrect prediction.
Prediction Results

Accuracy: 71.43%

Correct vs Incorrect Predictions
Challenges
Outputs & Next Steps
What happens after transcription?

- Batch export of TEI-compliant XML, raw TXT, and keyword searchable PDFs directly from Transkribus
- Documents housed in GitHub repository
- Semantic and text analysis
- Automated entity extraction (XSLT)
- Building personography and placeography
- Working toward digital editions
- … and more possibilities!
Linking Entities

Using Palladio to visualize person-person networks

Extracted persName entities from the Fulneck collection
What have we learned?

- The value of interdisciplinary, collaborative work
- The value of having undergraduate students work as research assistants - they each bring their own skills and interests and can help further develop parts of the project
- Progress on creating English-language models is informing our work developing German script models
- Benefits of multiple project management tools (Trello, GitHub, Google Apps, Slack)
Questions?

Moravian Lives: http://moravianlives.org/
Moravian Lives - Transcribed Memoirs: http://moravianlives.scholar.bucknell.edu/
Transkribus: https://transkribus.eu/Transkribus/
Moravian Lives GitHub: https://github.com/moravianlives/ML

Project Contacts:
- Katherine Faull (PI): faull@bucknell.edu (@KatherineFaull)
- Carrie Pirmann (Transkribus): carrie.pirmann@bucknell.edu (@librariancarrie)
- Brian King (Machine Learning): brian.king@bucknell.edu (@bkingcs)
- Bhagawat Acharya ‘20 (Algorithm Development): ba017@bucknell.edu
Datum der 18.9.

lässt und "ab ab" ab

→ Wenn was würden sie machen, wenn sie keine Haarbezüge hätten?

→ Wenn man so machen.

Wenn sie keinen Haarbezüge hätten,

Wenn sie kein Schnupfen hätten,

Wenn sie kein Käppchen hätten,

Wenn sie kein Perücken hätten,

Wenn sie sich das Mundgezecht verbrannt hätten

→ Wenn die Guppe hätten

→ Wenn sie keine Schlummer hätten.

Lösung: S. 08

Permanente 8

Sach. 128130

Praxis Bussig.