



# Text2Image matching

How to use existing images and transcripts to  
produce Automated Text Recognition

*Tobias Hodel  
Gundram Leifert*

State Archives of Zurich  
University of Rostock





Another perspective to train an ATR

Workflow

Text2Image on StAZH images

More demos and impressions





## Another perspective to train an ATR

Workflow

Text2Image on StAZH images

More demos and impressions





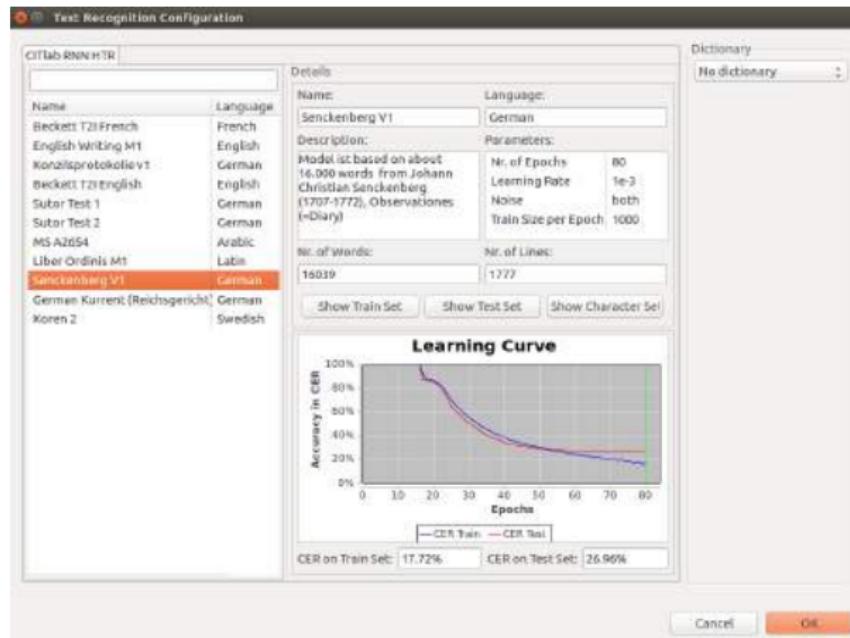
## Another perspective to train an ATR

The screenshot shows a software application window with a toolbar at the top and a sidebar on the left. The main area contains a document with handwritten German text. Below the text is a list of numbered annotations:

- 2 seines-unterm-1ten-Deember~~ell~~
- 3 nße.-in-betref-der-Organisa~~ell~~
- 4 tion-des-Sucnrrkontingents~~ell~~
- 5 gefaßten-Beschlußes-nach-an~~ell~~
- 6 gehörten-Gutachten-und-Frey→~~ell~~
- 7 ervorschlag-der-Militäirkom~~ell~~
- 8 mißion-vom-30ten-pus, zu-der



## Another perspective to train an ATR





## Another perspective to train an ATR

We want to have a good Automated Text Recognition (ATR)!





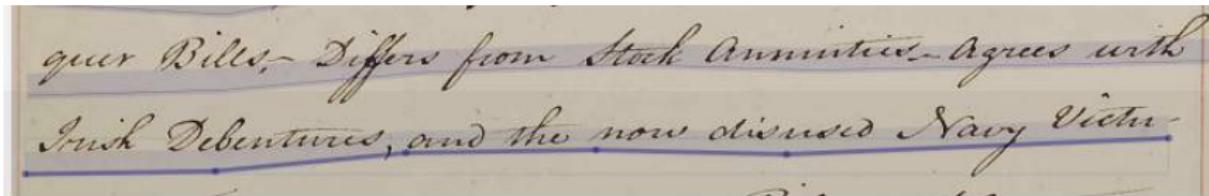
## Another perspective to train an ATR

We want to have a good Automated Text Recognition (ATR)!

We know:

- the more training data, the better the ATR
- classical training data production is expensive

ground truth (GT) – training sample



quer Bills, - Differs from Stock Annuities - Agrees with ↘  
Irish Debentures, and the now disused Navy Victu- ↘



## Another perspective to train an ATR

We want to have a good Automated Text Recognition (ATR)!

We know:

- the more training data, the better the ATR
- classical training data production is expensive

### Another option

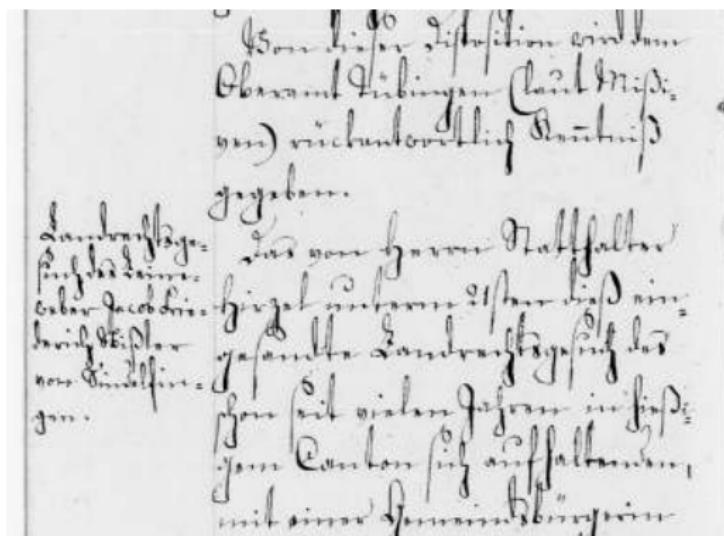
**we know:** for many documents the transcripts are already available

**but:** the transcripts are not aligned to the text lines of the images

**goal:** create training data from these images and transcripts for classical training



## Regierungsratsprotokolle StAZH (large scale demonstrator)



(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. [Landrechtsgesuch des Leineweber Jacob Friedrich Nißler von Sindelfingen.](#) Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(b) Transcripts



Another perspective to train an ATR

Workflow

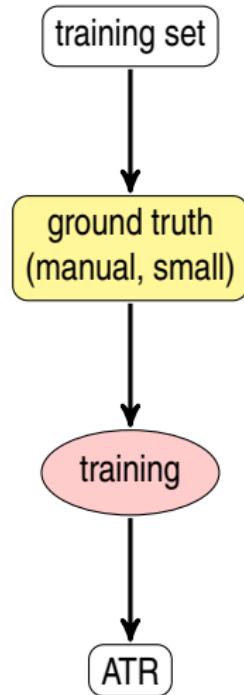
Text2Image on StAZH images

More demos and impressions



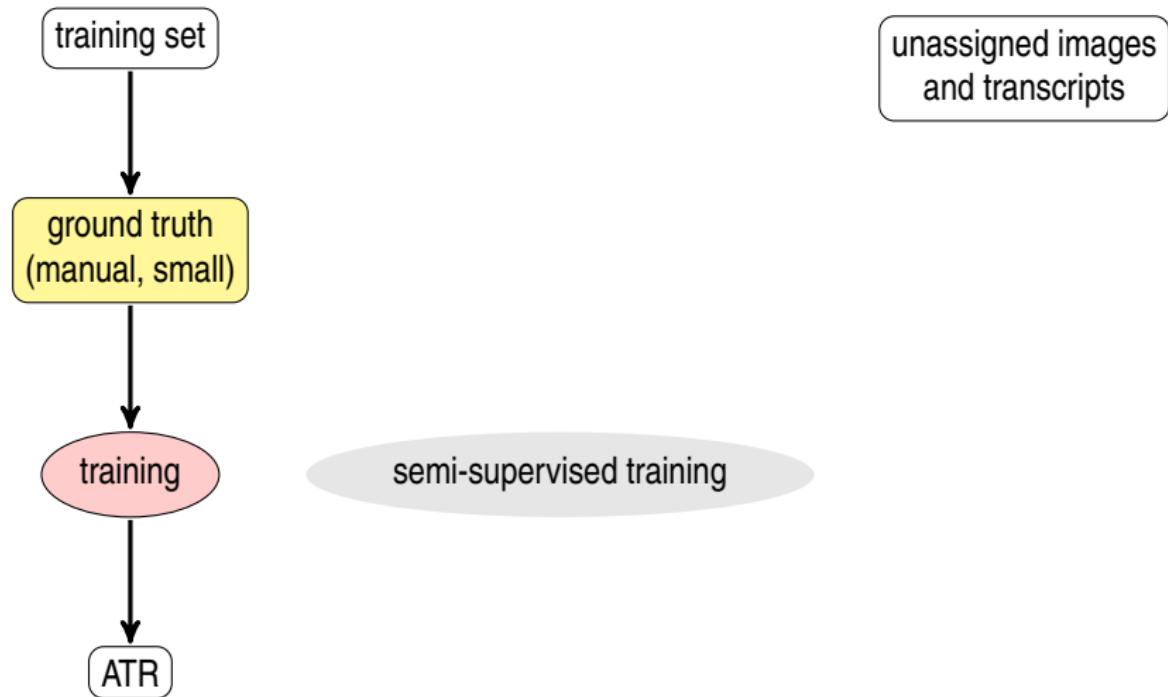


## classical training workflow

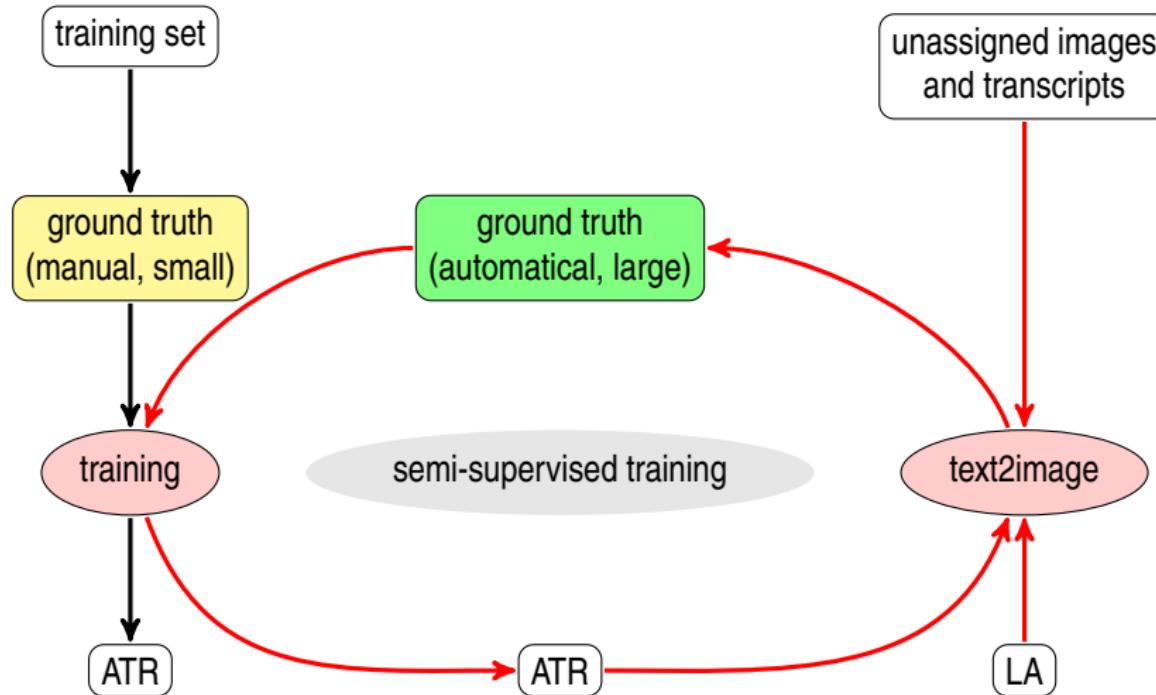




## semi-supervised training workflow



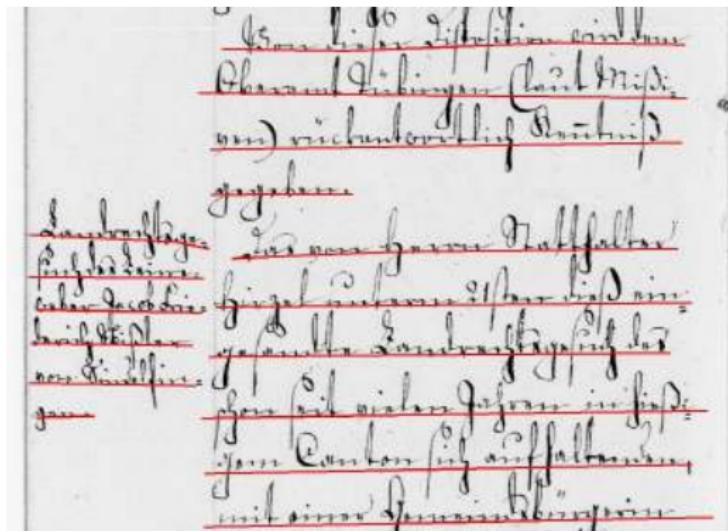
## semi-supervised training workflow





## workflow for Text2Image

transcript with corresponding image with baselines



(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. [Landrechtsgesuch des Leineweber Jacob Friedrich Nißler von Sindelfingen](#). Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeinsbürgerin

(b) Transcripts



Another perspective to train an ATR

Workflow

Text2Image on StAZH images

More demos and impressions



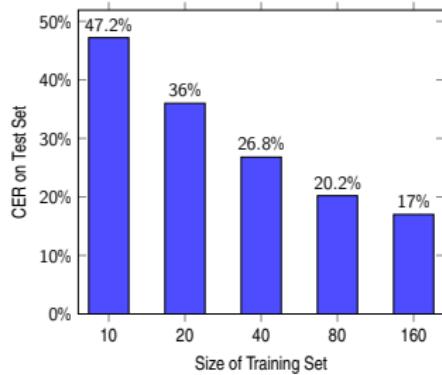


## Text2Image on StAZH images

### StAZH test collection (part)

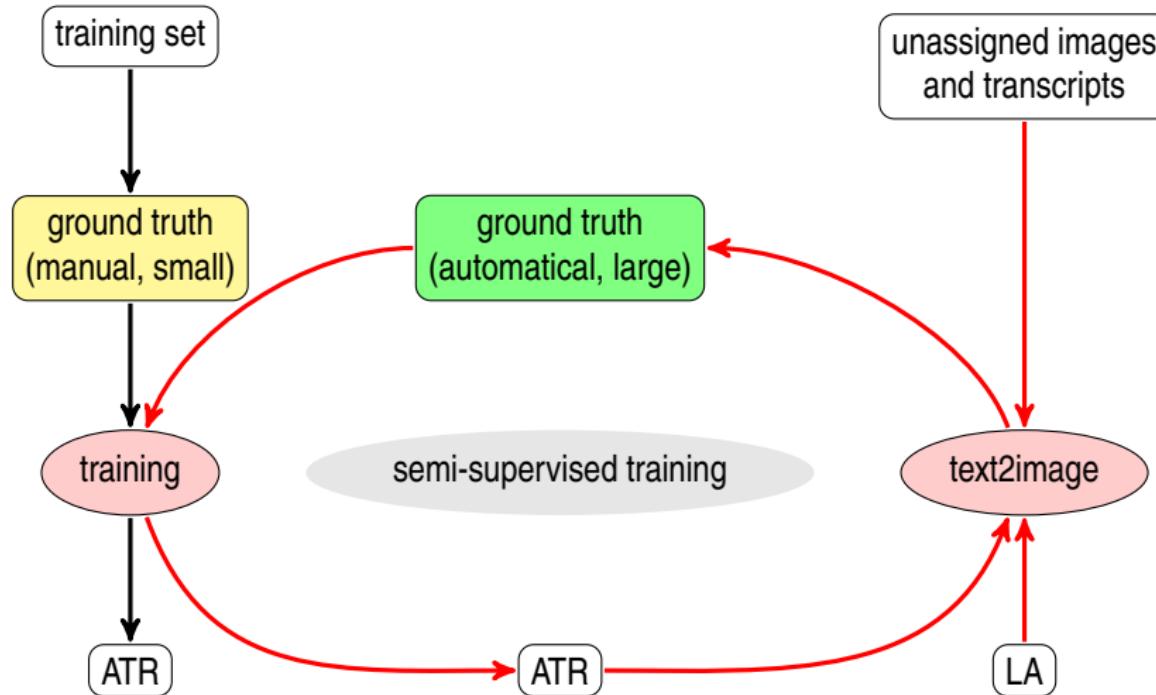
| Size of collection | pages |
|--------------------|-------|
| training set       | 160   |
| test set           | 40    |

### Character Error Rate on test set





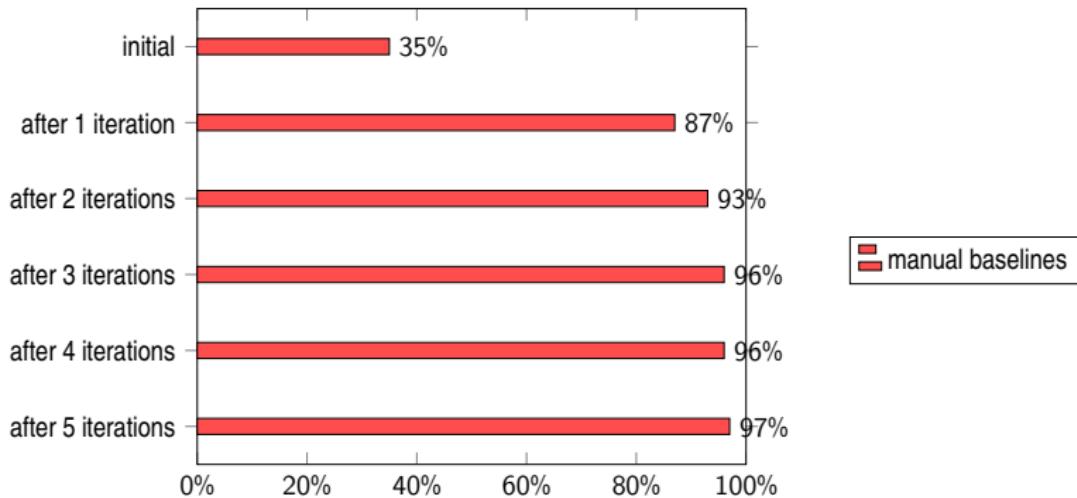
## semi-supervised training workflow





## Text2Image on StAZH images

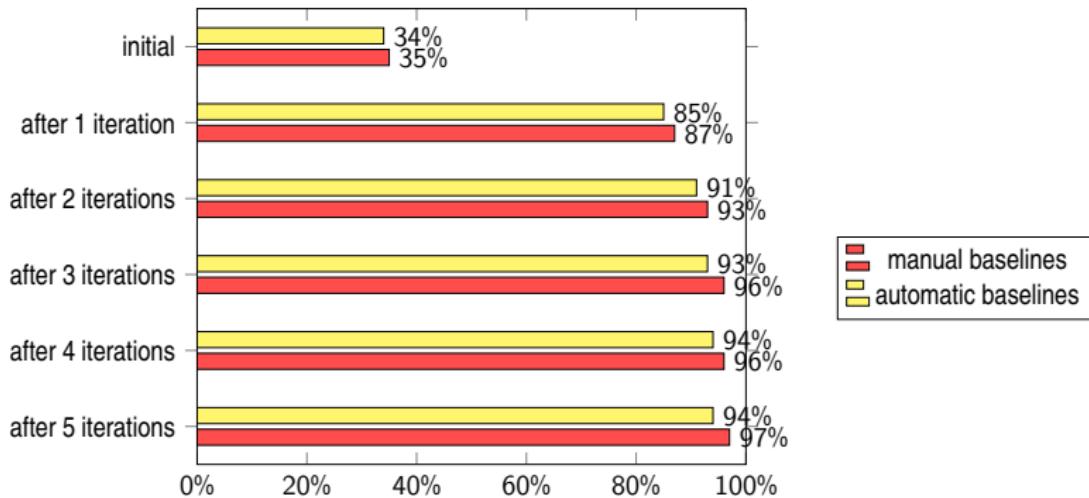
### Percentage of aligned Groundtruth





## Text2Image on StAZH images

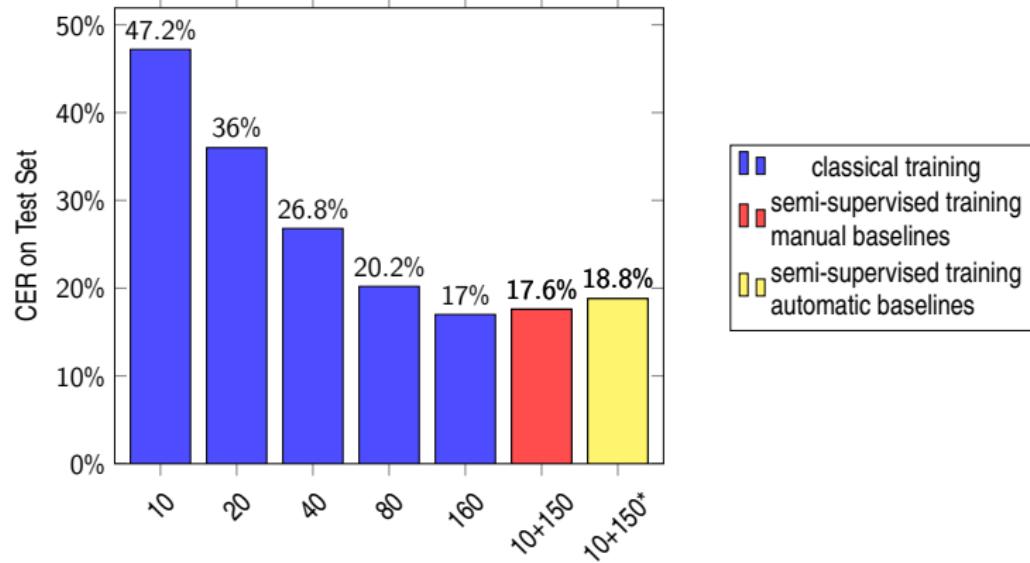
### Percentage of aligned Groundtruth





## Text2Image on StAZH images

### Character Error Rate on Testset





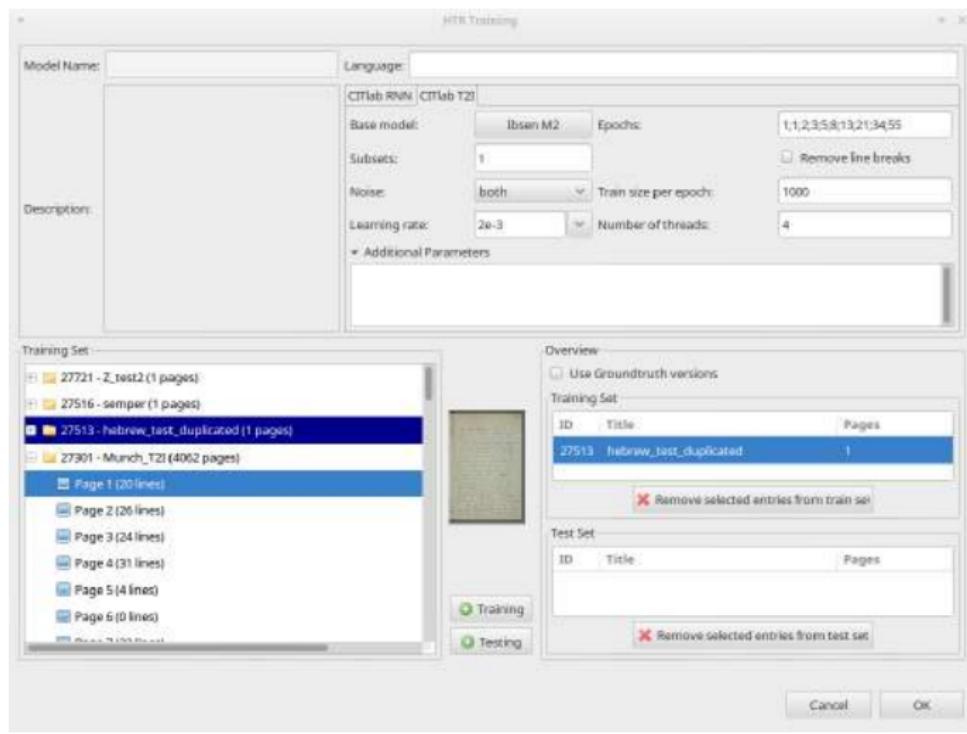
## What can we handle?

- unknown/new characters in transcripts
- missing line breaks in transcripts
- missing hyphenations in transcripts
- missing baselines and transcripts
- wrong baselines
- wrong order between baselines and transcripts
- right-to-left languages
- **hard:** abbreviations in transcripts





## Text2Image in Transkribus





Another perspective to train an ATR

Workflow

Text2Image on StAZH images

More demos and impressions





Kanton Zürich  
Direktion der Justiz und des Innern  
Staatsarchiv

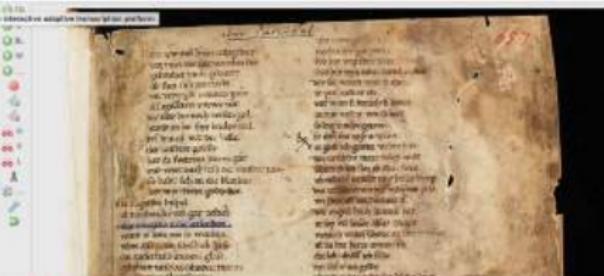
# Project READ

## Text2Image

### Demo



## Medieval Text

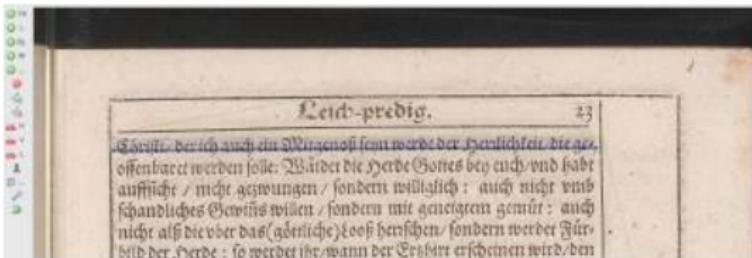


The screenshot shows a medieval manuscript page with two columns of handwritten text in a Gothic script. Below the page, a transcription of the text is provided, numbered from 28 to 38. The transcription includes some red annotations.

28 so habt sich an die blanken,  
29  
30 der mit steten gedanken  
31 Diz fligende-bispele.  
32  
33 ist tumben lüten gar zensné.  
34  
35 sine mugens nñht erdenken.  
36  
37  
38

- 2 Transcription
- 2 Transkribus t2i with „generic model“  
(about 20% CER, no GT)
- 2 Identification of ~50% of lines
- 2 230 pages
- 2 Trained HTR: 13% CER (on test set)

## Printed Text (16<sup>th</sup> century)

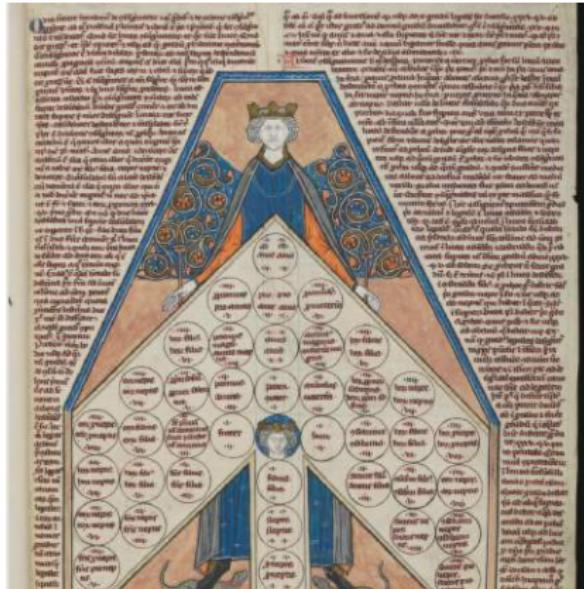


Christi / der ich auch ein-Mitgenoß seyn werde der Herrlichkeit / die ge-<sup>3</sup>  
offenbart werden sole: Würdet die Herde Gottes bey euch / vnd habt<sup>4</sup>  
aufsicht / nicht gezwungen / sondern williglich: auch nicht vmb<sup>5</sup>  
schändliches Gewissens willen / sondern mit geneigtem gemüt: auch<sup>6</sup>  
nicht aß die über das (göttliche) Looß herrschen / sondern werdet Für-<sup>7</sup>  
bild der Herde: so werdet Ihr / wann der Erzähler erscheinen wird / den<sup>8</sup>  
unvergleichlichen Krantz der Herrlichkeit darvon bringen. Aber es<sup>9</sup>  
folget auch die Rechnung<sup>10</sup>  
Hinden nach / des-vnützten Knechts.<sup>11</sup>  
Dem es / über seine schlimme Haltung vnd vergrabenes Talent /<sup>12</sup>  
wunder-vbel ergehet. Höret /<sup>13</sup>

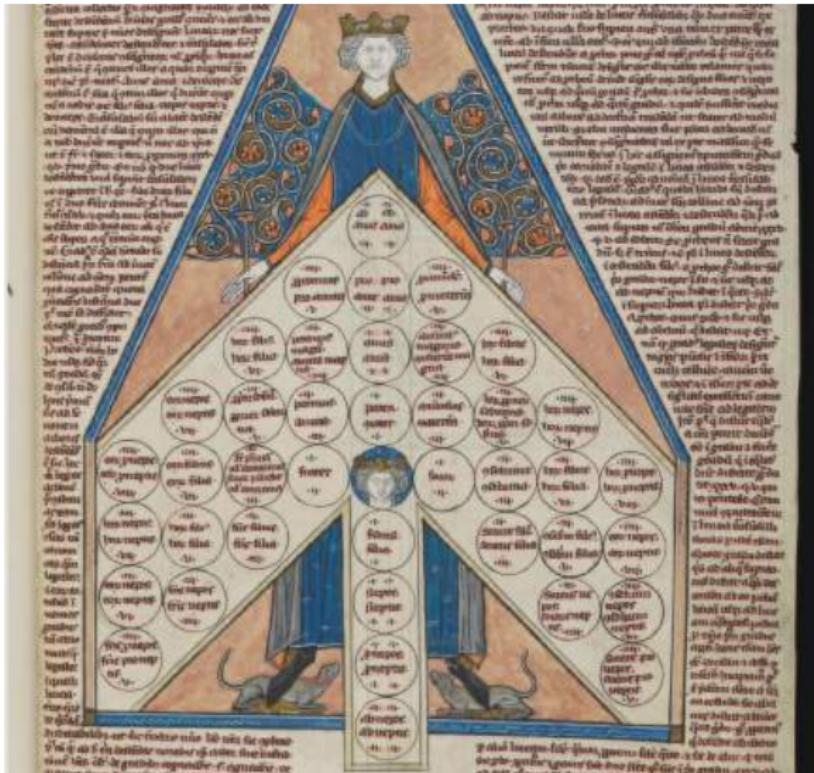
- 2 Transcription
- 2 Transkribus t2i with „fraktur model“  
(about 20% CER, no GT)
- 2 Identification of ~90% of lines
- 2 36 pages
- 2 Trained HTR: 5% CER

# Input

- 2 Images of pages
- 2 Text files (one txt file per pages; also TEI/Word-Doc)




 Direktion der Justiz  
 und des Innern  
**READ**





THANKS ...

CIT lab Group

**CITIAB**

Computational Intelligence Technology

MoU Partner SME



PLANET artificial intelligence GmbH

EU HORIZON2020 Grant



**READ** Recognition and Enrichment  
of Archival Documents

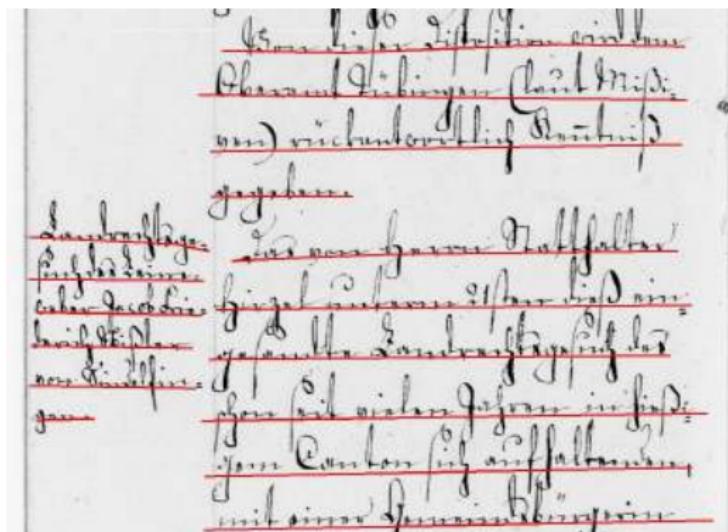


... for your kind attention!



## workflow for Text2Image

transcript with linebreaks



(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben.

Landrechtsge-  
such des Leine-  
weber Jacob Frie-  
derich Nißler  
von Sindelfin-  
gen.

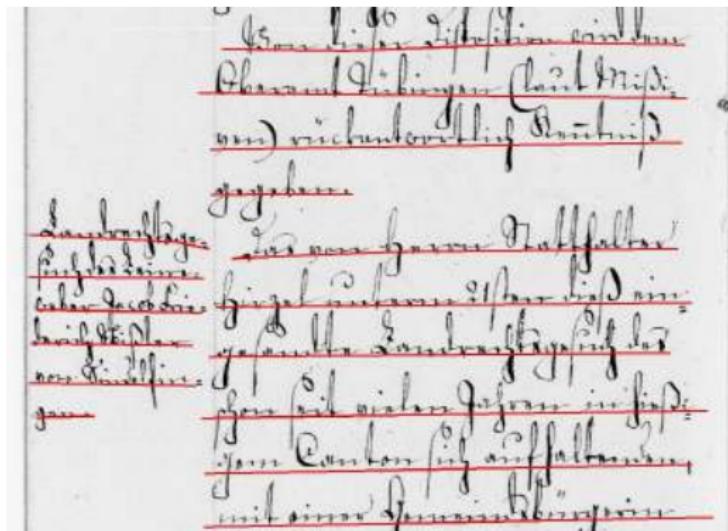
Das von Herrn Statthalter Hirzel unterm 21sten dieß ein- gesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeinsbürgerin

(b) Transcripts



## workflow for Text2Image

transcript without linebreaks



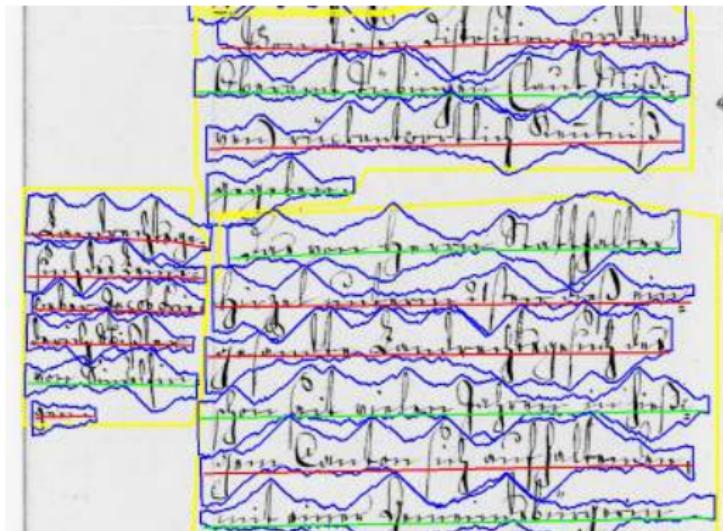
(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. [Landrechtsgesuch des Leineweber Jacob Friedrich Nißler von Sindelfingen](#). Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeinsbürgerin

(b) Transcripts



## Matching lines



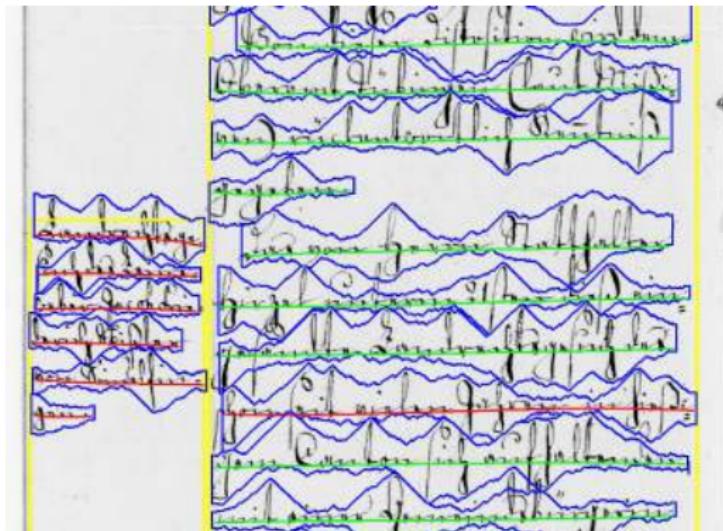
(a) Image

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. **Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen.** Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin

(b) Transcripts



## Matching lines



(a) Image

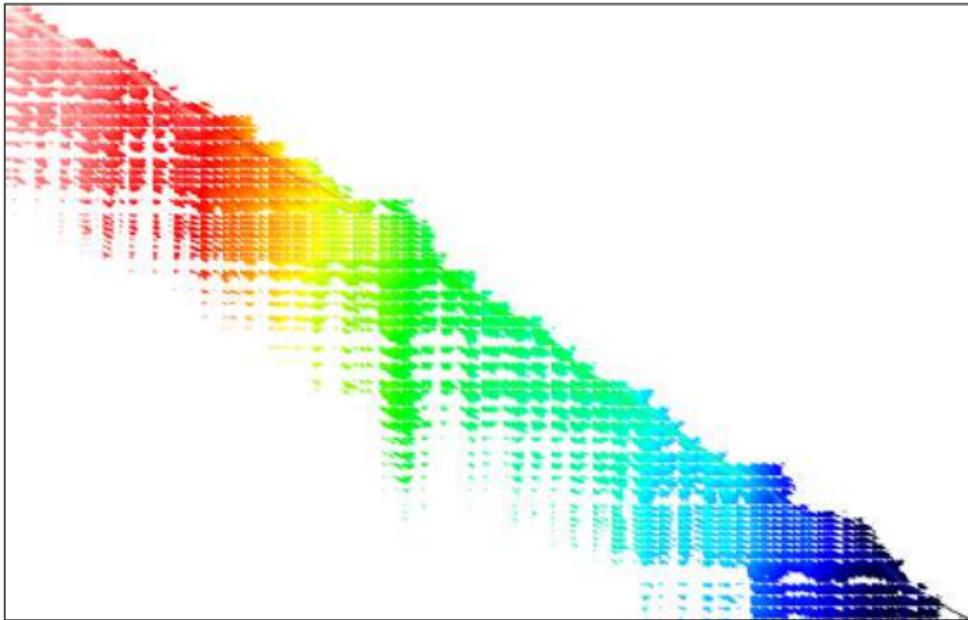
(b) Transcripts

Von dieser Disposition wird dem Oberamt Tübingen (laut Mißiven) rückantwortlich Kenntniß gegeben. [Landrechtsgesuch des Leineweber Jacob Friederich Nißler von Sindelfingen](#). Das von Herrn Statthalter Hirzel unterm 21sten dieß eingesandte Landrechtsgesuch des schon seit vielen Jahren in hießigem Canton sich aufhaltenden, mit einer Gemeindsbürgerin



## workflow for Text2Image

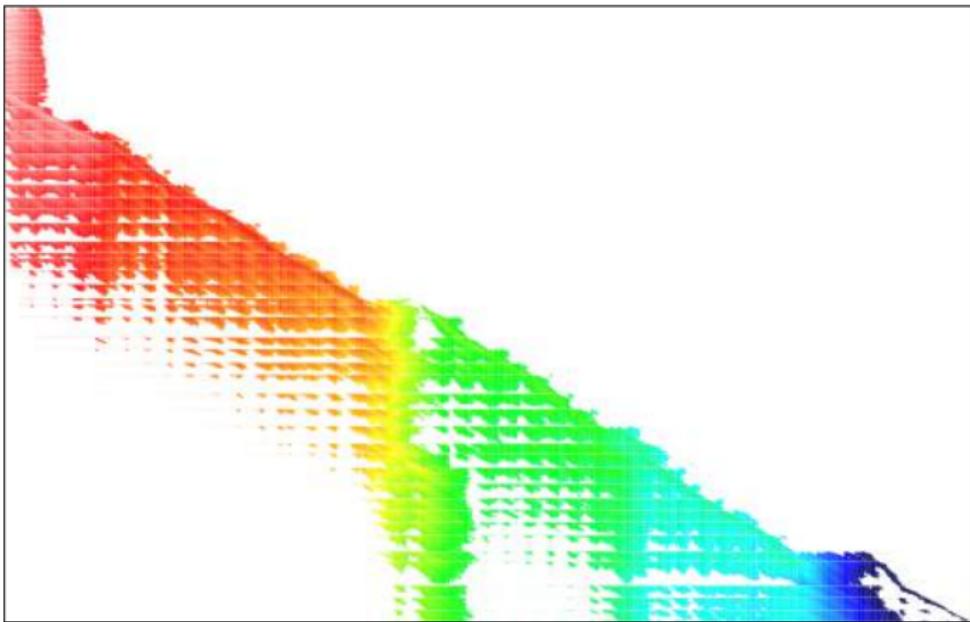
Some nice images from the assignment procedure...





## workflow for Text2Image

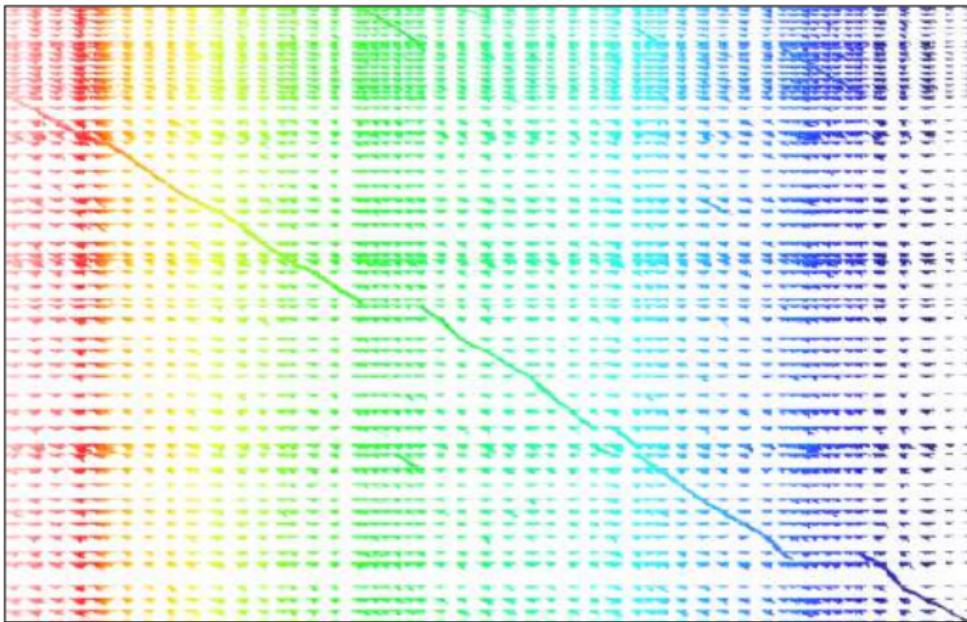
Some nice images from the assignment procedure...





## workflow for Text2Image

Some nice images from the assignment procedure...





## Text2Image matching

We want to have a good Automated Text Recognition (ATR)!

We know:

- the more training data, the better the ATR
- classical training data production is expensive

## Another option

- for many documents the transcripts are already available
- so far, the transcripts are not aligned to the text lines of the images
- we can create training data from these images and transcripts to train an ATR