

Keyword Spotting, sharing HTR models and Transkribus web

Günter Mühlberger

University of Innsbruck,

Digitisation and Digital Preservation Group

The logo consists of the word "READ" in a bold, dark blue, sans-serif font. The letters are closely spaced and have a slight shadow effect.

Agenda

- Keyword Spotting
- Sharing models
- Transkribus Web

Keyword spotting

Transcription vs. searching

- Transcription
 - In order to get good transcriptions in an automated way you need a lot of training data
 - Especially for large collections this might be a challenge – thousands of pages will be necessary
- Searching
 - Is it necessary to have a transcription of the text to be able to perform good search results?
- Keyword spotting
 - Is a method to be able to search a collection without using the automated transcription of the text
- Magic?

Dogs – Cats - Birds



Labels



= DOG



= CAT



= BIRD

Input



= DOG



= CAT



= BIRD

Output



= DOG



= CAT



= BIRD

Input

TRAINING

Output



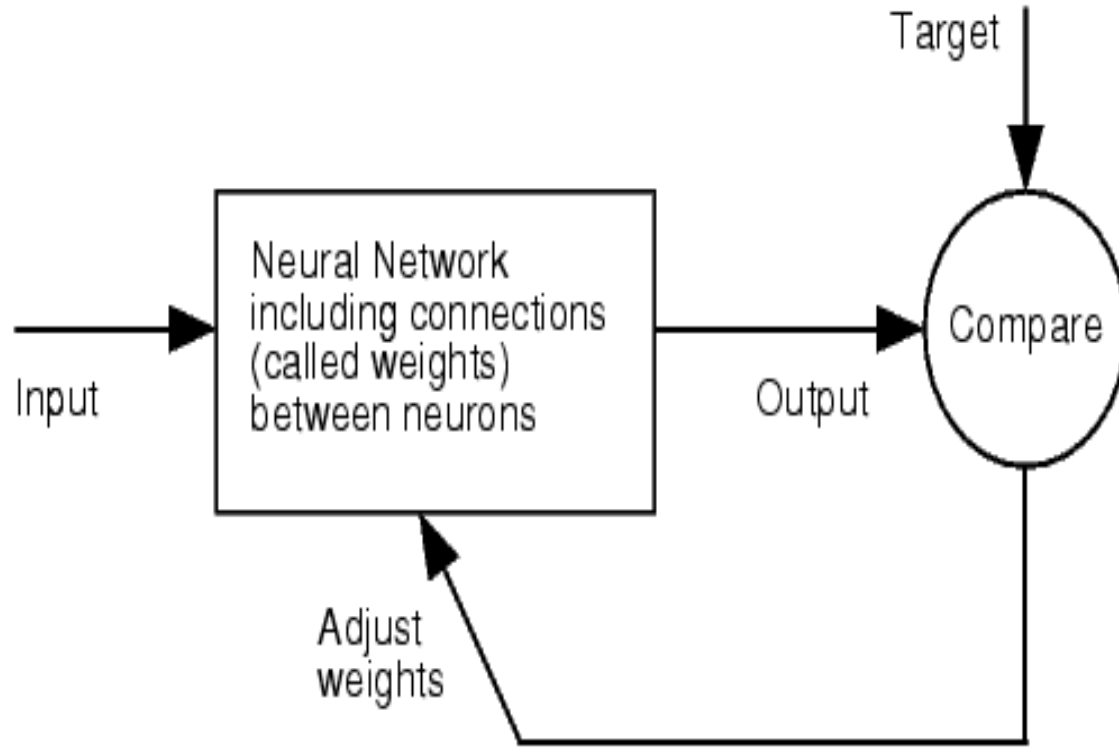
= DOG



= CAT



= BIRD



= DOG



= CAT

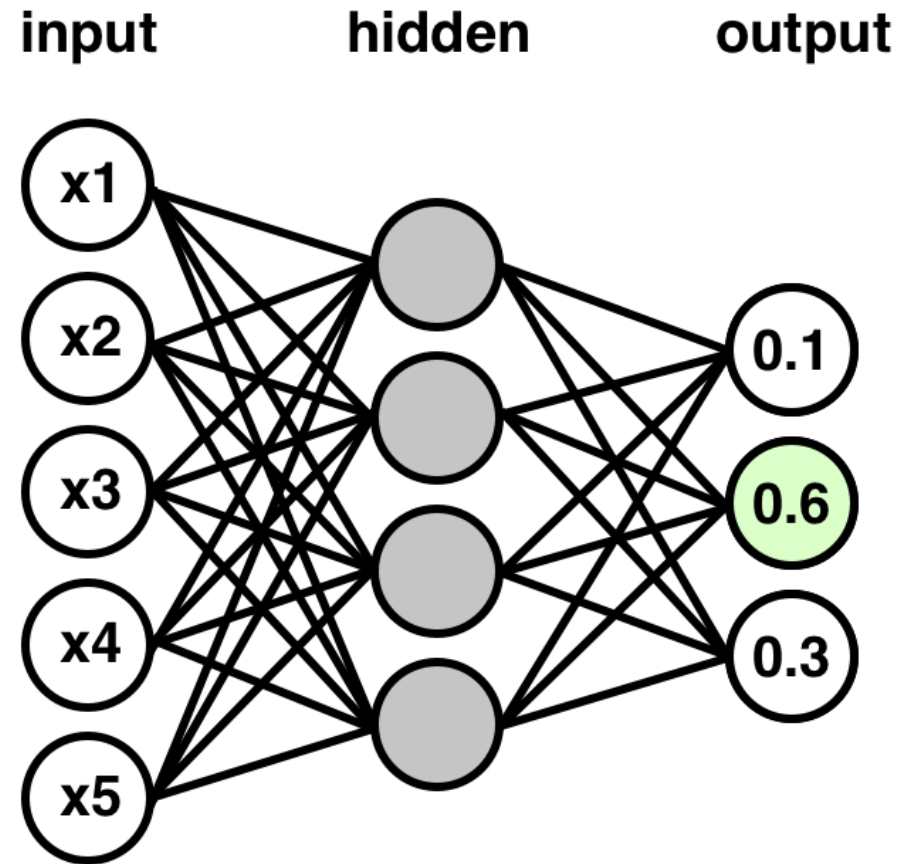


= BIRD

Application = Recognition



?



= BIRD

= CAT

= DOG

Recognition



?

0,1 = BIRD

0,7 = CAT

0,6 = DOG

Recognition



?

0,1 = BIRD

0,7 = CAT

0,6 = DOG

Recognition



?

0,1 = BIRD

0,7 = CAT

0,6 = DOG

“Transcription”



?

0,7 = CAT

“Keyword spotting”



?

0,1 = BIRD

0,7 = CAT

0,6 = DOG

This image appears in my result list since the confidence value is higher than 0,5 (*but the transcription would be wrong*)

Keyword Spotting

- Like for CAT, DOG, BIRD for every character in the alphabet a confidence is computed, with which the neural network states, that a character appears at a specific spot of the image
- Transcription
 - = the characters with the highest confidence
- Keyword Spotting
 - = the search string which is above a given confidence value

Implementation in Transkribus

- All documents recognized in Transkribus are stored with a confidence matrix
- Method 1 – University Rostock – CITlab Team
 - Search is performed directly on the confidence matrix
 - Is therefore available directly after the recognition process
 - All data can be exploited, e.g. search with regular expressions
 - Fast for small collections, but relatively slow compared for large collections – implemented as a job
 - Produktive version since late 2017
- Method 2 – Technical University Valencia – PRHLT Team
 - Based on the confidence matrix an index is generated. Based on n-grams of characters, not words
 - Index reduces the amount of possible options
 - This index can be fed into a standard full-text engine such as Lucine/Solr
 - Quick response time
 - Con: Some loss of information
 - Demoversion in Transkribus implemented

Search for...

Documents

Fulltext (Solr)

Tags

KWS

Keyword (Solr)

Search HTR text for single words

Search in:

1895

All Documents

Search for keyword:

prison

Confidence Threshold:

25

<

>

Search!

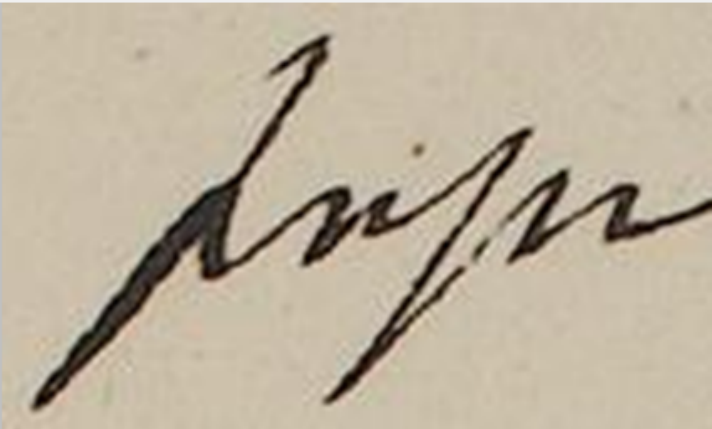
Previous page

Next page

Search results (43 hits, page 1 of 1):

Probability	Word	Document	Page	Line ID
82	prison	[box_001]	303	null
70	prison	[box_001]	221	null
65	prison	[box_001]	229	null
59	prison	[box_001]	548	null
55	prison	[box_001]	120	null
52	prison	[box_001]	29	null
51	prison	[box_001]	715	null
50	prison	[box_001]	627	null

Preview



Close

Search results – first page

Search

Time line



Tree View

Level 1+ (233)
Level 1+ (122)
Level 1+ (40)
Level 2+ (1700)
Level 2+ (1200)
Level 1+ (544)
...

Results: <1-10 > Date / Relevance



Image snippet – keyword highlighted
2-3 lines height
Fonds/series/subseries/document title



Image snippet – keyword highlighted
2-3 lines height
Fonds/series/subseries/document title



Image snippet – keyword highlighted
2-3 lines height
Fonds/series/subseries/document title



Image snippet – keyword highlighted
2-3 lines height

Archival units

Title 1 (40)
Title 2 (20)
...

Places

Helsingfors
(3567)
Tampere (2342)
...

Keywords

House selling
(1872)
Children care
(2000)

Review selection

Search

Selected results: <1-10 > Date / Relevance

Send to
Transkribus
PDF, Excel
Crowd-sourcing

☒ X

Image snippet – keyword highlighted
5-10 lines height

Comment / Tag

Go to page / correct snippet

☒ X

Image snippet – keyword highlighted
5-10 lines height

Comment / Tag

Go to page / correct snippet

DEMO

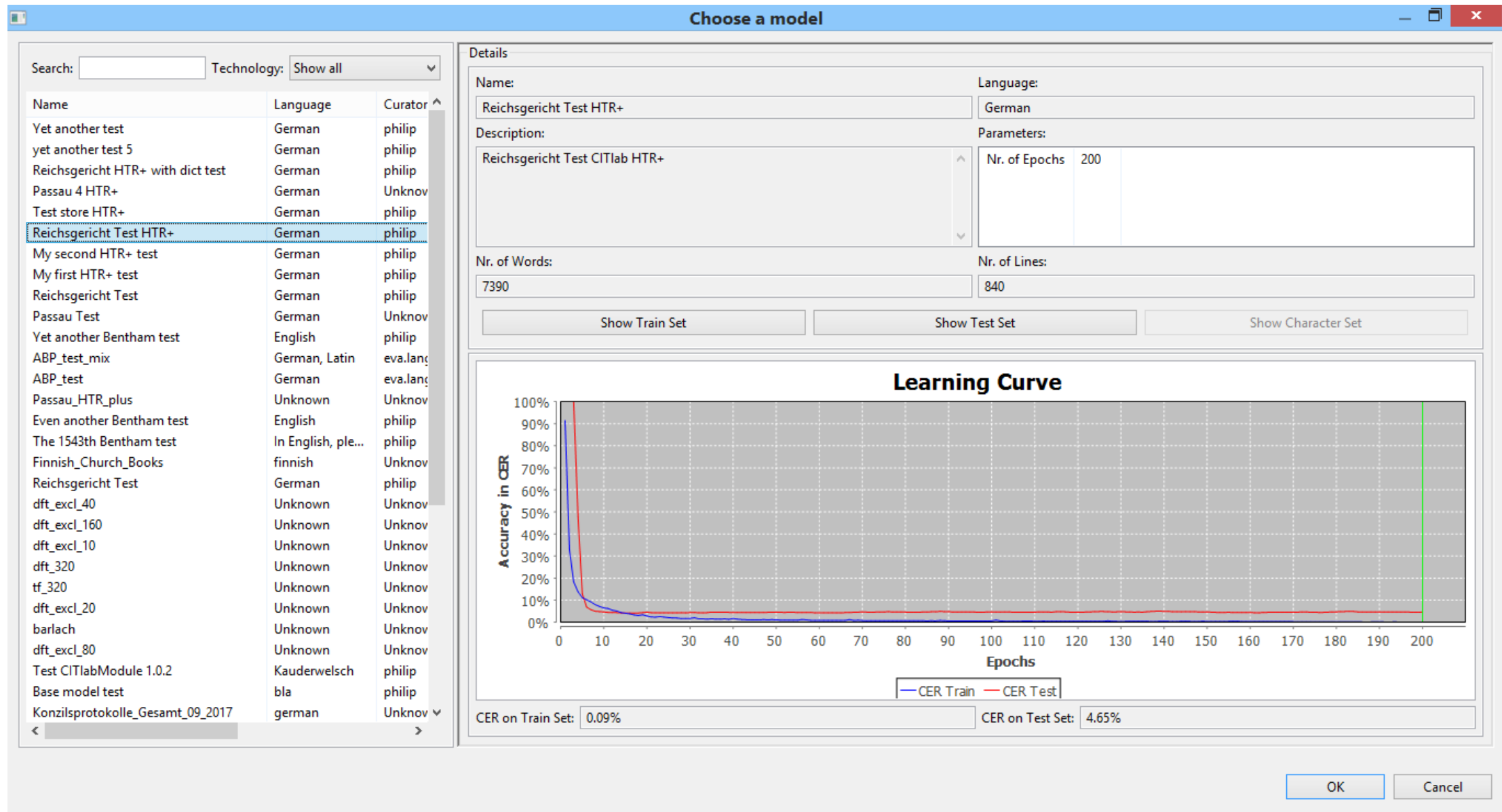
Sharing HTR models

Model training

- Currently available for users who have written us an email
- Extremely nice and easy feature
- Hundreds of meaningful models were trained so far
- At the heart of Transkribus
 - Share models (not necessarily the documents) with other users
- How to?

Considerations

- Some metadata needed
 - Language
 - Time of writing
 - Script (printed/handwritten, writing style,...)
 - Place where it was written
 - OBIG.de database of scripts
 - Description of documents used for training
- However
 - More important is it that other users who want to reuse a model have examples of the documents available
- Sharing
 - Either the complete document (but this might be work in progress, the document not belong to the user, etc...)
 - Or just a few lines
 - Credits



Random lines – approved by user who wants to share

Kaiser von Fürstenthümern gratis und franco zu liefern, wofür

Platz in den bestellten Plätzen der Luftkammer, welche die Füllung nicht ausfüllt.

25 23, 25 St. mit dem 1. September 1900 als Fälligkeitstag zu verfahren.

Under consideration

- User feeds some lines with ground truth to Transkribus
- User selects some metadata, such as language and period
- Transkribus runs these lines against all available models
- A chart is provided with the best models

Webinterface

Main idea is to be able to involve users into transcription projects without the need that they learn to work with the expert client

My Collections

Collections with which you are associated.

Public Projects

« 1 2 3 »

Collection	# Documents	
BNE_KWS_TEST	9	Open
Bohisto	32	Open
Brabrand-Aarslev 1928-1933	3	Open
Briefe	0	Open
Collegie van Landdrost en Heemraden	8	Open

Project Bohisto

[Back to overview](#)

Edit Special Characters

Document:

Status:

From 1 to 12 of 574



Seite 1

Done

Changed by
barbara.denic
olo

Plaintext Annotation



Seite 2

Done

Changed by
barbara.denic
olo

Plaintext Annotation



Seite 3

Done

Changed by
barbara.denic
olo

Plaintext Annotation



Seite 4

Done

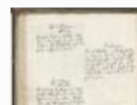
Changed by
barbara.denic
olo

Plaintext Annotation



Seite 5

Done



Seite 6

Done



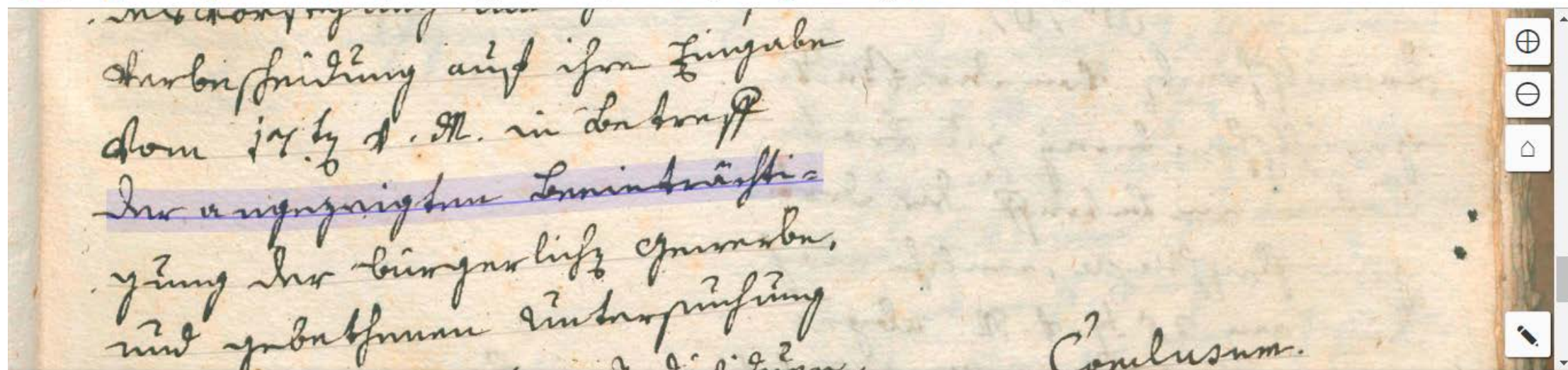
Seite 7

Done



Seite 8

Done



x² x₂ B / U abe Special Characters Annotate ... ?! Unclear

1	Anlangen der ... bürger: Han-	#
2	dels Vorsteherung um ehebäldeste	#
3	Verbescheidung auf ihre Eingabe	#
4	Vom 17: v. M. in Betreff	#
5	der angezeigten Beinträchti-	#
6	gung der bürgerlich Gewerbe,	#
7	und gebethenen Untersuchung	
8	der befingerzeigten Individuen.	

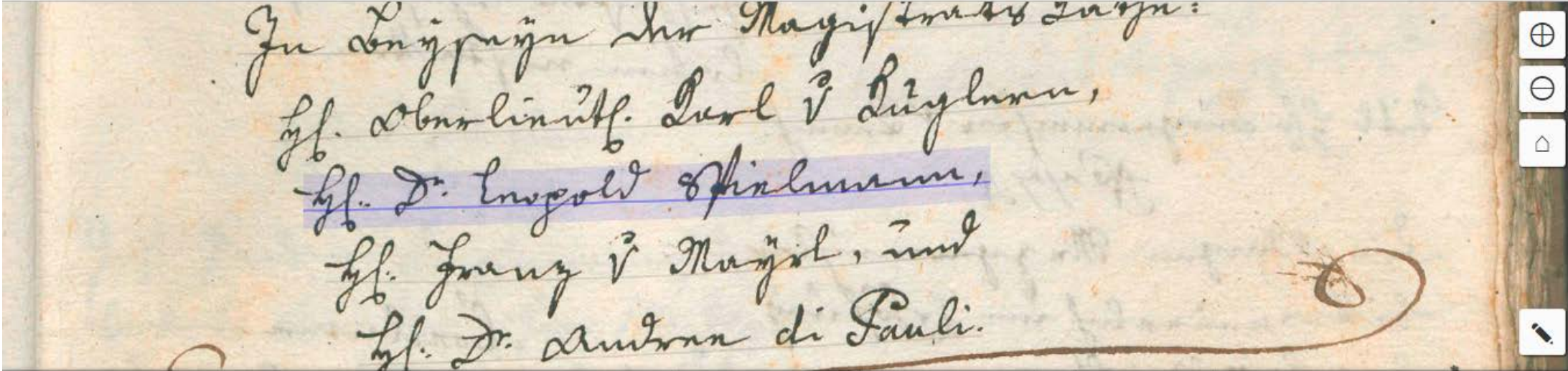
In Bearbeitung

Save Changes

News – READ ProjeTranskribus als WeProjektePhilologisProject - Bohistohttps://transkribushttps://transkribushttps://transkribus

← → ↺ https://transkribus.eu/r/read/sandbox/application/?mihyro=1&markers=1&collId=2082&docId=4381&pageId=4

Apps6 KalenderHome - Research PaREAD CallREAD WEBREAD WUIKWSKalenderPROJEKTE - TranskriSandbox



⊕ ⊖ 🏠 ✎

⌂

x² x₂ B / U abe Special Characters Annotate ... ?! Unclear

Notice
Notice

Occupation
Occupation

Firstname
Firstname

DateOfDeath
DateOfDeath

– + × Dr. Leo...

Text Region 3

1 Actum Bozen den 1^{ten} Julius 1793.

2 Vor Titl Hⁿ Bürgermeister Jos. v Remich p.

3 In Beyseyn der Magistrats Räte:

4 H. Oberlieut. Karl v Kuplern,

5 H: Dr. Leopold Spielmann,

6 H: Franz v Mayrl, und

7 H: Dr. Andree di Pauli.

In Bearbeitung

Save Changes

Thank you a lot for your attention!

More information

<https://read.transkribus.eu/>

<https://transkribus.eu/>

<https://scantent.cvl.tuwien.ac.at/en/>

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

