

# Text Technologies for Historical Disciplines

Tobias Hodel (Walter Benjamin Kolleg, Digital Humanities)

Bern, 04. & 05.05.2023

Bild: Stable Diffusion, Prompt:  
«A deep neural network  
in the style of a Dali painting»



# Program

## Introductions & Hands-On

Thu., 4th of May

- Morning: Text Recognition (Transkribus)
- Afternoon: Natural Language Process/Information Extraction (Flair)

Fri., 5th of May

- Morning: Scholarly Editions (TEI & TEI-Publisher)
- Afternoon: Linked Open Data (Geovistory)

# How we proceed

- Each segment will be introduced theoretically (only superficial)
- Followed by a short hands-on introduction
  - [then we will probably need a break]
- Now you test it out yourself (with your own material, if possible)
- Finally, let's present your approach and discuss possible next steps

# Philosophy

Get your hands «dirty»

- Create (throw-away) accounts
- Try out!
- Try again
- You can't break anything
- Ask anything
- Help each other out

# Slides and Downloads

We use GitHub as our «Cloud»  
But you don't have to learn git.

Link regular: [https://github.com/DominicWeber/CUSO\\_UNIL](https://github.com/DominicWeber/CUSO_UNIL)

Link short (may rot): <https://bit.ly/3nrSD9g> bit.ly/3nrSD9g

# Round of Introduction

- What's your background?
- What are you interested in (language-/topic-wise)?
- What would you want to achieve?

# Handwritten Text Recognition

*u*<sup>b</sup>

*b*  
UNIVERSITÄT  
BERN

Tobias Hodel (Walter Benjamin Kolleg, Digital Humanities)

Bern, 04.05.2023

Screenshot of a digital manuscript transcription interface:

The interface includes a top navigation bar with icons for search, zoom, and document management, followed by a status bar showing "In Progress".

The left sidebar contains links for Server, Ov (Transcription view docking state), Tools, Logout, and various document management options like Document, Find, Document Manager, User Manager, Versions, Jobs, Recent Documents, and User activity.

The main area displays a medieval manuscript page in Latin script. The text is partially transcribed in blue ink below the original red ink. A handwritten note "[309 a.] (1361 IX. 14.)" is visible on the right side of the page.

The transcription results are shown in a table:

ID	Title	Pages	Uploader	Upload
2533...	TRAINING_TESTSET_Ad fontes: KB...	1	tobias.hodel...	Sat No
2531...	TRAINING_TESTSET_Ad fontes: KB...	1	tobias.hodel...	Fri No
2200...	Kopialbuch_20	3	tobias.hodel...	Mon S
2200...	Kopialbuch_19	3	tobias.hodel...	Mon S
2200...	Kopialbuch_18	3	tobias.hodel...	Mon S
2200...	Kopialbuch_17	3	tobias.hodel...	Mon S
2200...	Kopialbuch_16	3	tobias.hodel...	Mon S
2200...	Kopialbuch_15	3	tobias.hodel...	Mon S
2200...	Kopialbuch_14	3	tobias.hodel...	Mon S
2200...	Kopialbuch_13	3	tobias.hodel...	Mon S
2200...	Kopialbuch_12	3	tobias.hodel...	Mon S
2200...	Kopialbuch_11	3	tobias.hodel...	Mon S
2200...	Kopialbuch_10	3	tobias.hodel...	Mon S

Bottom left corner: 2-4 verordnet.hät. ↪

# Goals

We want to...

- understand the basics of Text Recognition and be able to produce specialized models on our own
- develop scenarios for the use of Text Recognition for your research
- get to know export functionalities and ways to publicize transcriptions
- get to know challenges for project based on machine learning

# Before we start

Please:

- > Register at <https://readcoop.eu/>
- > Download the latest version of Transkribus:  
<https://readcoop.eu/transkribus/download/>

There is a lite (web-version) available: I recommend to learn the (full) downloadable software (if you want to use Transkribus for research).

Mac User: Make sure to copy Transkribus to your program folder.

Java-Problems: [https://download.oracle.com/java/17/latest/jdk-17\\_macos-x64\\_bin.dmg](https://download.oracle.com/java/17/latest/jdk-17_macos-x64_bin.dmg)

# A very short overview: From Optical Character Recognition to (Automatic/Handwritten) Text Recognition

1990ies: Commercial Optical Character Recognition

2000thes: Recognition of «Fraktur» with OCR (EU Projekt IMPACT, ABBYY XIX)

2010ns: Simple Language Model / Intorduction of neural networks

2020ies: Dominance of *deep learning*  
(high-dimensional language model)

Starting 2021: Transformer-based models for text recognition (trOCR)

# Projekt READ

## Recognition and Enrichment of Archival Documents

Making archival (esp. handwritten) documents more accessible

Research infrastructure – Transkribus

Funded until mid-2019 by the European Union (H2020)

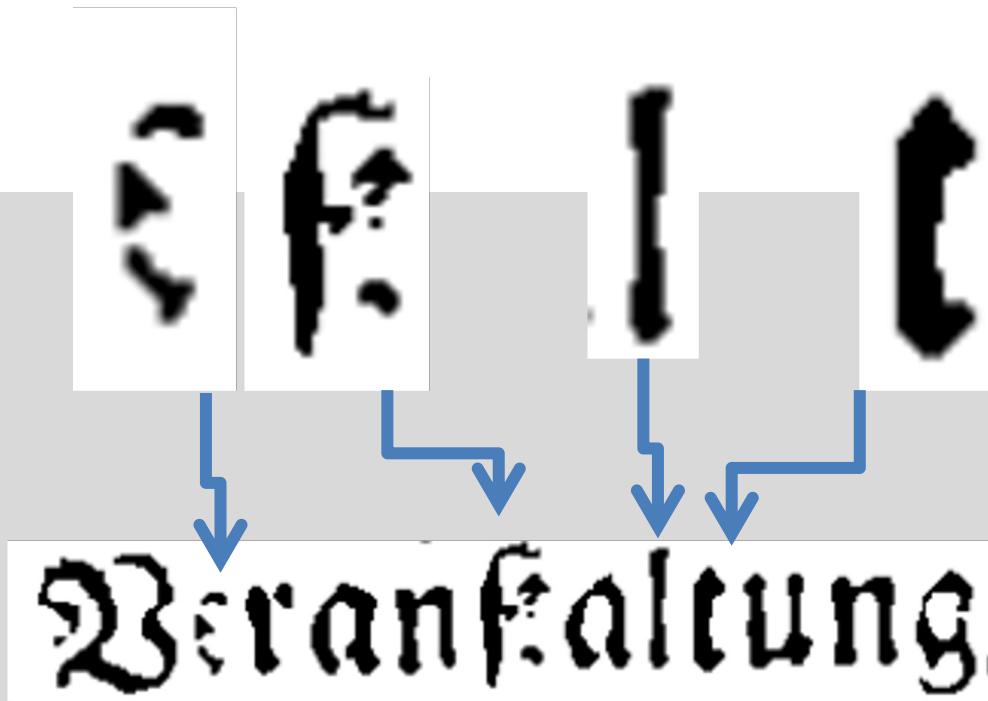
15 European partners

# Research perspectives of READ

- Recognition of layout and text structures
- Recognition of handwriting (Handwritten Text Recognition)
- Text recognition with dictionaries
- Writer identification --> proof-of-concept
- Best-practices for recognition of large amounts of documents
- Digital Humanities in archives and scholarly practices

**READ**

# Automated Text Recognition?



# Automated Text Recognition

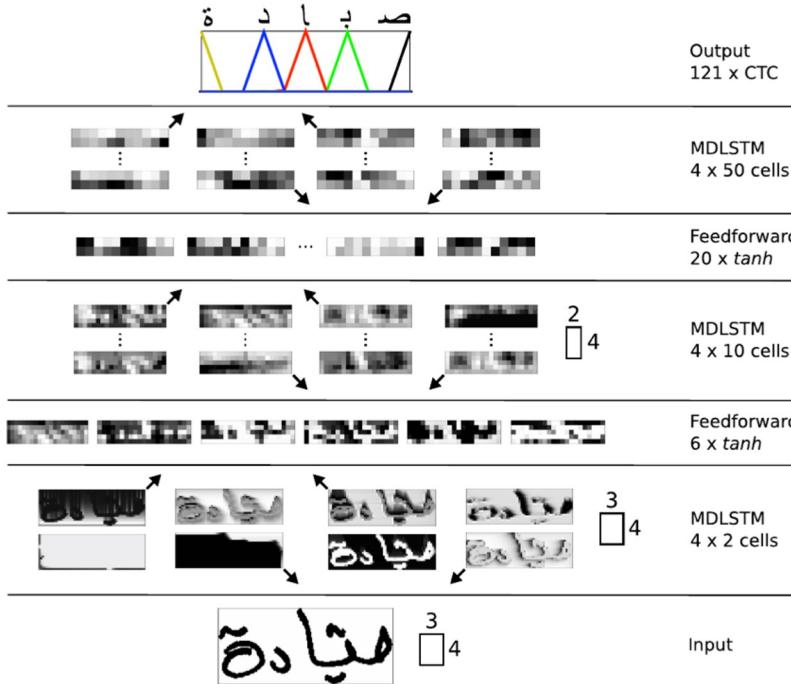
- Machine learning using neural networks
- Processes writing by line, rather than by character
- Needs to be trained by being shown document images and transcripts
- More training data → more accurate recognition
- Create a model to transcribe and search a collection of documents

# Recognising printed text

- . Neural networks can also process printed text – with less training data!
- . Transcribe documents or use OCR engine in Transkribus
- . Use these transcripts to train a model
- . Results with <1% CER are possible



## Von der Eingabe zur Netzausgabe



(Bild aus GRAVES, SCHMIDHUBER: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks)



# Carolingian Minuscules

5'000 words, 1'000 lines:

Validation pages: 7% CER

24

illecm: que p[ro]fec[ta]cc fo[rum] mitat  
timorem per quic uiuersc que  
prius non sine formidine obser-  
vabec absq[ue] ullo labore uel ut  
nec tunc eliter & consu&udine  
incipi & custodire. Non icem timo-  
re gehenn[us], sed amore xp[ist]i: & con-  
su&udine ipse bone & delecta-  
tione uirtutum: quae dñs icem  
in operchum suum mun dum  
auitatis & pec[un]ia[rum] sc̄o dignebit[ur]  
demonstrare; **De officiis diuinis** **III**  
**H**i emis tempore i[n] hoc t[ri]bus;  
idesat: a[et]er[na]l[er] dis nob[is] bem br[iti]s; in festo marie pentecoste  
usque in pascha. iux ta considera-  
tionem rectionis o[ct]ava horee noc-  
ti surgen dum est: ut mo di ce-  
am plius de media nocte pecus &

# Early Modern Letters

48'000 words, 6'450 lines:

**Validation pages: 2,5% CER**

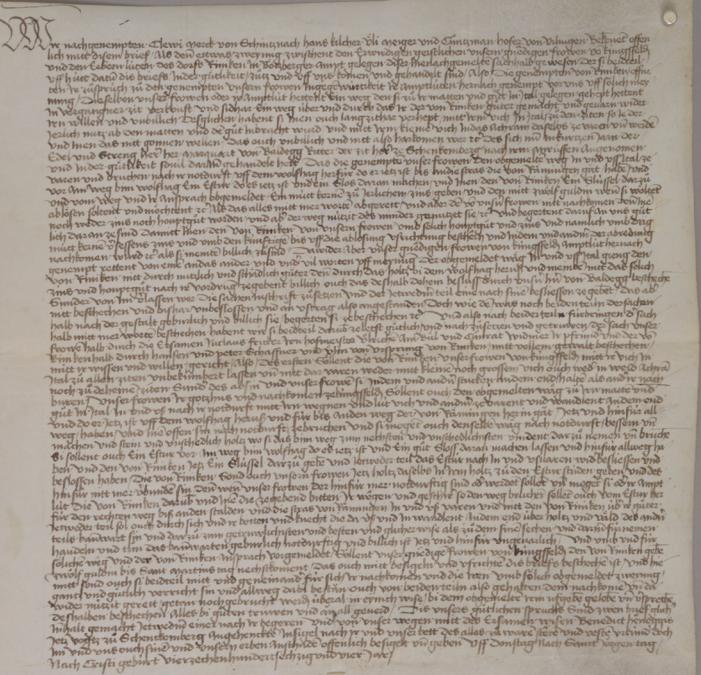
allor lind by lar St. venthar Dreyfus 27

mit was foriden und was gretig an habe sol  
gleichsich van dreyt enz. dies geschenk an  
vader gott seg lob und dank da stadt unter  
der sonden salt der sol gro kommen zet  
der wyls doy gow in den vryd wahr  
nun wond gow in dant, der geygafft  
zu vryl ghet will ich myt uouw zelten  
fall vnde den dor ouw si sol kommen  
dys ob minc inticht dorwan aben und reicht  
hiedam doy sol nicht anje balt favord  
si myt myt dor, wo den ic minc dor  
megevob dagos will zelten sin dor  
is dor souven so den day, si myt myt dor  
aldeinde regenten den, si geban enz en  
my dor want nglieh laugen gemit  
da fouldert und wird allerien in sic gelegten  
ob gan van vryt gemit si gher de forst  
dag dor myt myt dorwan myt dor  
maghe, isam vordern spilt nun, in miche  
out endom dor em vler da vryd er  
ghebrugem vnder agen, wie sollan dor

A.

# Charter 15th century

77'000 words, 3'500 lines:  
**Validation pages: 4,8% CER**



# Council Minutes 19th Century

148'000 words, 31'000 lines:

**2,5% CER**

14. Januar 1851.

Actum Samstags den 4. Januar 1851.  
In Gegenwart des gesammten Regierungsrathes.

Wiederholungspflichten von 30,00 aufgestellt werden.  
Unterschrift des Befürworters  
der in den nächsten Jahren der Gemeindeverwaltung, die  
gegenüber dem Bürger  
jeweilige Rechtsvorschriften bringt, zu legen  
31. Wurz möglicherfalls in Form eines Jahres-  
aufstellens nach Name, Name, Vorname und Zweck  
des Aufstellens sowie in einem Zeichen, wird die  
in der Stadtstaatsprüfung bestehende, am 31.  
März eines möglichen Jahres abgelegte Prüfung vorliegen,  
imponieren bestehende, bei welcher Anlaß aufgestellt  
befindenden Urkunden bestehend zu unterscheiden  
sollen, die auf den angegebenen Prüftag im Kreis-  
land befinden.

Es wird diejenige Wirkung der Variation bestimmt, welche übereinstimmend mit dem Befestigungsmaßstab die Verhältnisse der Anordnung der zu befestigenden Teile verlangt.

Opportunities for organized labor were 28% greater between 1960 and 1970 than during the previous decade, and opportunities for self-employed workers increased by 10%.

Dorfgründung von höchstens 400 m auf das Bezirksgericht Wittenburg und der Direktion des Finanzamtes zu beziehen.

gung überwiegt.

Überprüfung der Ergebnisse für die Erhaltung, Begehung und Abstellung in  
der Zeit vom 1. April bis zum 30. September 1912.

# Generalizing of Writing Types Current/19th century

HTR MODEL	HTR ENGINE	CER MEAN %	CER MEDIAN %	CER UPPER BOUND (WORST)
German Kurrent M2	HTR+	3.43	2.76	9.13
	PyLaia	18.77	13.30	51.05
Transkribus German Kurrent	HTR+	5.90	4.85	10.20
RRB	HTR+	9.15	8.13	16.28

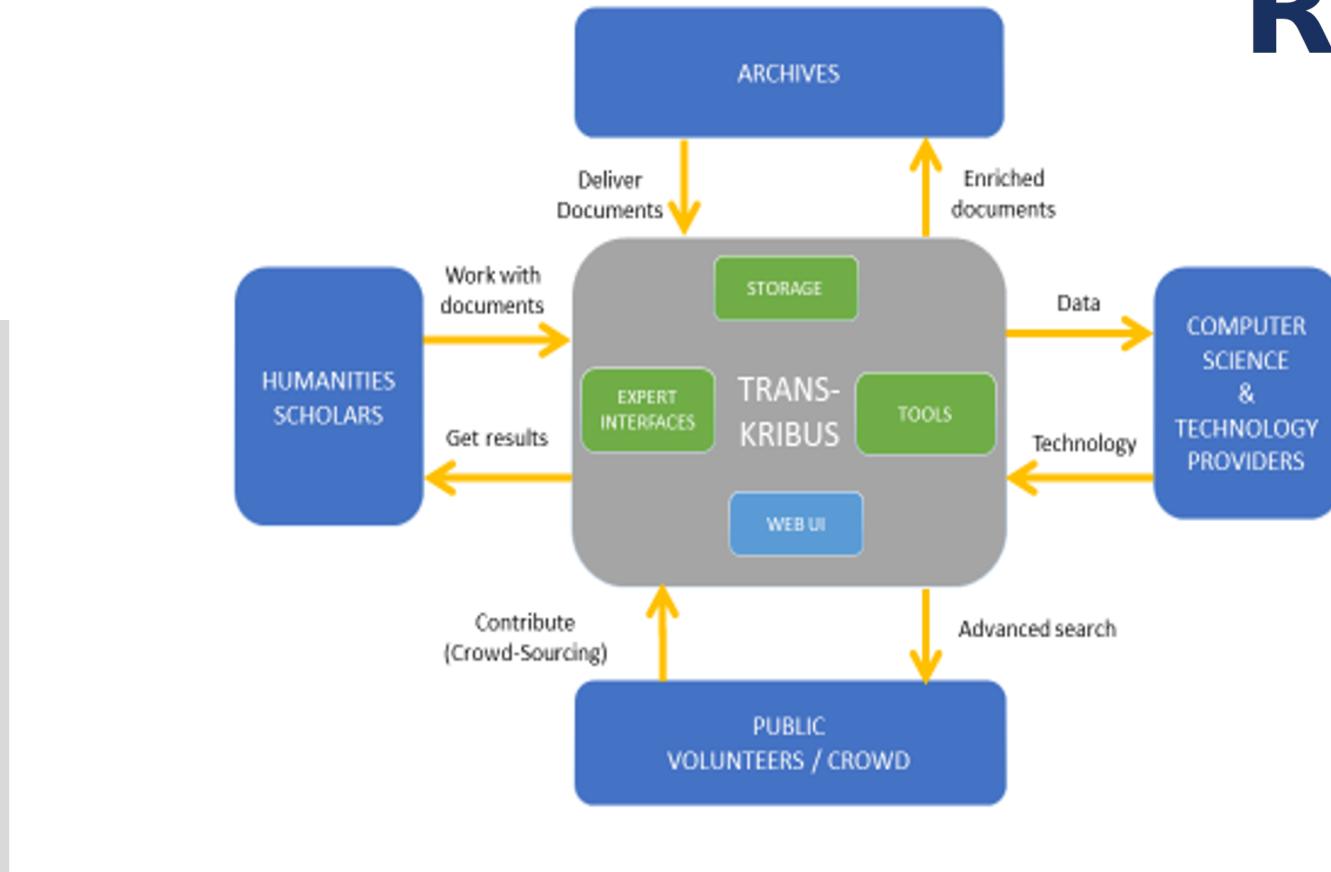
1. Januar 1851.

Actum Suntays den 4. Januar 1851.

In Gegenwart des gesammten Regierungsrathes.

Die bei Hoffnung auf Vorsichtserfolgen von 30,0% aufgeführten Differenzen sind auf die vorliegenden Schriften zurückzuführen, die im Kurrent-Schrifttypus durch die früheren Kontrollen ausgeschlossen worden waren. Diejenigen bei großartigen Versuchen, welche am 31. März vorgenommen wurden, haben ebenfalls nach Ramm, Wenzl, Goretzka und Grub das Ergebnis fast ausgängig zu bestätigen, und die

Zur Freiheit gegeben, dass, um Ablösungen in  
der Form der Schrifttypus  
aufzufinden, die bestehende  
Schrift, geschafft durch eine Reihe von  
verschiedenen Versuchen, um 18,0% das Ergebnis, auf die von dem Druckmeister



*u*<sup>b</sup>

Look at Transkribus

---

<sup>b</sup>  
UNIVERSITÄT  
BERN

**READ**

# READ COOP

## Business model

Cooperative

More than 60 members

(Institutes & private members)

Buy-in:

250€ / 1'000€

500 Credits included

Costs of Credits: Varying depending on amount.

<https://readcoop.eu/transkribus/credits/>

Ca. 0.11-0.24 € / page (text recognition)

# Visual Feature Recognition

page-number

marginalia paragraph

320.

Daudor fuhkend und den Rita  
Loutinum das grose Gabium u. Zalpa  
in der fl. Und zu Daudor Basel un  
König zu h. ringlang zu  
gehen, und gewiss und vor  
galgen, Gabium das nu  
gut doppelt - domisior  
W. Daudor h., soll das in da

page-number

marginalia paragraph

321 preamble

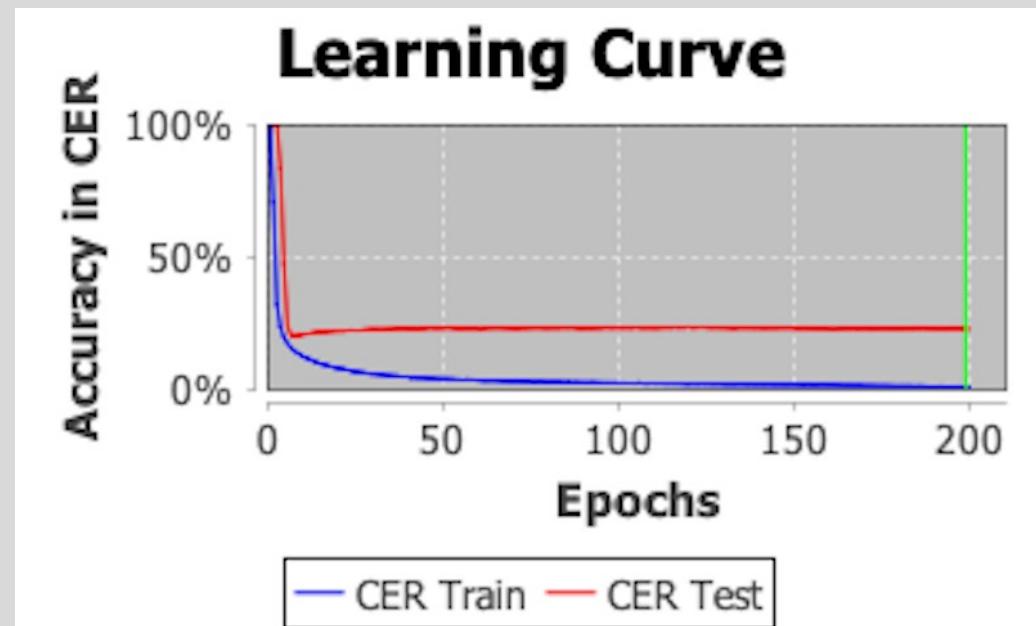
Von für mit Rauf Enföning wo hör ich  
Girolf und H. Narfingar mit Girolf  
lex s. off. Willam von Hessen aufgenommen  
Anföre, warhier unten anhören  
Herr P. Ich will nun siehe  
Anföllan zu Blattföllan be  
grungen sind, - haben Magdeburg  
zu H. Gabium in der hund

# Visual Feature Recognition

	1461.	Werner Hofer, Gerichtsschreiber der Stadt Zürich, erklärt, da Oldelhart Völlinger von dem Predigerkloster Zürich ein Häuschen an dem kleinen Predigergässchen an dem Brach zu seinem länglichen Niersbrach gekauft habe, mit der Bestimmung, dass aus ihrem Nachlass 20 Pfund Penninge an das Kloster bezahlt werden sollen, so habe er, nachdem ihm dieselbe all ihr Gut versprochen habe, dem Kloster Gewähr geleistet, dass, wenn der Nachlass für Bezahlung der 20 Pfund nicht hinreichen sollte, es das Mangolende ergangen werde.
date	Februar 9. (Mittag vor der pfaffen vasnacht.)	Hofer siegt.
call-number	St. A Z. Obmannamt str. 286. Original, Pergament. Siegel hängt.	

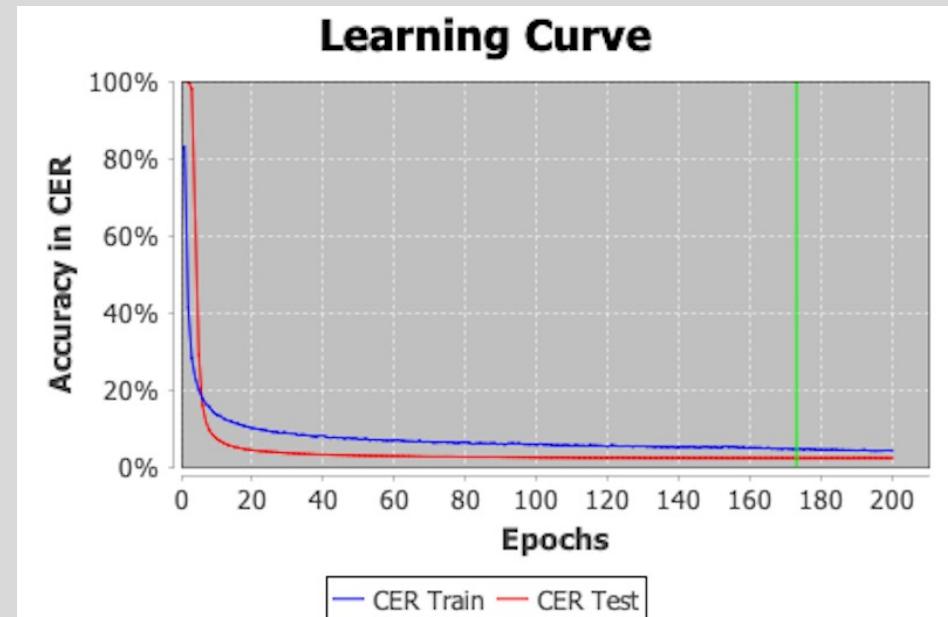
# Interpretation of Graphs

Diverging curves: overfitting



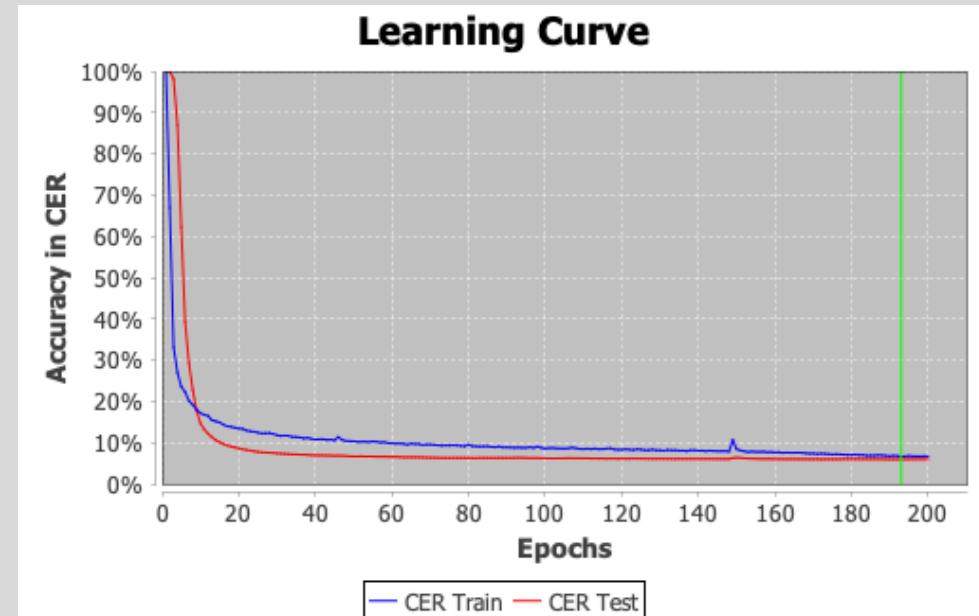
# Interpretation of Curves

Keep on training...



# Interpretation von Trainingskurven

## Austrainiertes/optimales Netz



## How-To's & Links

Read&Search (Hanserezesse)

<https://transkribus.eu/r/rezesse-niederdeutscher-staedtetage/#/>

Tutorials:

<https://readcoop.eu/transkribus/resources/how-to-guides/>