Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Von „gemischt“ zu „werksspezifisch“
Modelltraining für die Texterkennung von historischen Drucken und Handschriften

Christian Reul

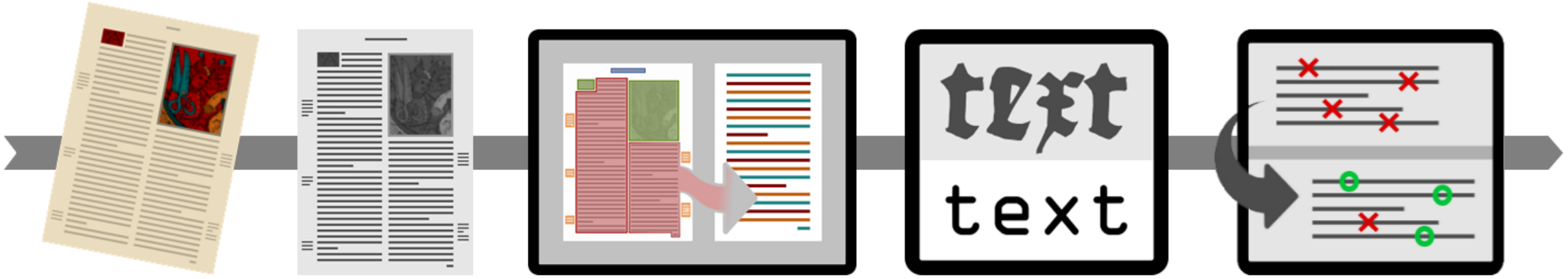
Zentrum für Philologie und Digitalität „Kallimachos“ (ZPD)
Universität Würzburg

03.02.2022

Gliederung

1. **Motivation und Grundlagen**
2. Methoden
3. Anwendungsbeispiel Drucke
4. Anwendungsbeispiel Handschriften
5. Diskussion und Ausblick

Grundlagen – Workflow und Modelle



- Hauptkomponenten: Vorverarbeitung, Segmentierung, OCR, Nachkorrektur
- Moderne OCR-Ansätze arbeiten auf Zeilen- und nicht mehr auf Zeichenbasis
- Sog. Modelle extrahieren Text aus Textzeilenbildern
  - Modelle müssen trainiert werden (Neuronale Netze)
  - Trainingsdaten bestehen aus Zeilenbild-Text-Paaren

Grundlagen – Gemischte und werksspezifische Modelle

- Gemischte Modelle:
  - Im Normalfall auf einer Vielzahl von Quellen trainiert
  - Vorteil: Out-of-the-Box Anwendung, kein werksspezifisches Training
  - Nachteil: ungenauer als werksspezifische Modelle
- Werksspezifische Modelle:
  - Im Normalfall exklusiv für die Erkennung genau einer Quelle trainiert
  - Vorteil: meist deutlich genauer als gemischte Modelle
  - Nachteil: benötigt Ground Truth (GT), also Zeilen und deren Transkriptionen
- Kombination: werkspezifisches Training ausgehend von gemischten Modellen („Finetuning“)

Motivation

- Unterschiedliche Strategien für verschiedene Anwendungsszenarien und Bedarfe
  - Massenvolltexterkennung: Out-of-the-Box Anwendung gemischter Modelle
  - Textproduktion für digitale Edition: Umfangreiches werksspezifisches Training, um bestmögliche Erkennung als Grundlage für die manuelle Nachkorrektur zu erzeugen
  - Fließender Übergang
- Idee/Ziel: Breit aufgestellte Modelle trainieren, die …
  - out-of-the-box auf eine Vielzahl (Alter, Sprache, Schriftart) von Materialien angewendet werden können
  - als Ausgangspunkt für weiteres Finetuning dienen können
- Pragmatischer Ansatz!

Gliederung

1. Motivation und Grundlagen
2. **Methoden**
3. Anwendungsbeispiel Drucke
4. Anwendungsbeispiel Handschriften
5. Diskussion und Ausblick

Ausbalancieren der Daten

Problem:

- Daten sehr unausgeglichen: Anzahl GT Zeilen pro Quelle schwankt zwischen <50 und 10k+
- Modelle optimieren sich in Richtung der im Korpus dominanten Daten

Lösung:

- Ausbalancieren durch Definition ausgewählter Seiten (etwa 50-150 Zeilen pro Quelle, je nach Aufkommen im Korpus)
- Zweistufiger Ansatz:
  1. Training auf allen verfügbaren Daten
  2. „Refinement“ des resultierenden Modells nur auf den ausgewählten Seiten

Nutzung unterschiedlicher Binarisierungen

Problem:

- Art der Binarisierung kann Performanz gemischter Modelle erheblich beeinflussen
- Keine fixe Universallösung, sondern Einsatz bei den Endnutzern abhängig …
  - vom Material
  - deren persönlicher Präferenz

Lösung:

- Training der gemischten Modelle auf unterschiedlichen Binarisierungen
  - Größere Robustheit gewährleistet breitere Anwendung
  - Künstlich aufgeblähte Trainingsmenge erhöht Performanz

CLARISSI-
MO AC PRVDENTIS-
simo Senatui ciuitatis Noriber-
gę, Dominis suis colendissimis,
Sebaldus Heyden
S. D,

Voties de nostrarum Scholarum disciplina cogito, Viri Clarissi-mi, nonnihil admirari soleo, cur studij literarū loca olim Græcis ab o-tio, Scholæ: Latinis uero Ludi dicta sint. Nam non ausim putare illos artium ingenuarum. ac Philoso-phiæ studia, ita uilia ac leuia habuisse, ut ea nō nisi per otium, & tanq̄ res ludicras sectanda esse censerent. Præsertim quum ob ea ipsa studia, cæteras omnes gentes, barbaras præ se uocarent, ac tanq̄ omnis humanitatis expertes despicerent. Cu-ius elationis causas certe non otiosas,

A 2 multo

Normalisiertes Graustufenbild

CLARISSI-
MO AC PRVDENTIS-
simo Senatui ciuitatis Noriber-
gę, Dominis suis colendissimis,
Sebaldus Heyden
S. D,

Voties de nostrarum Scholarum disciplina cogito, Viri Clarissi-mi, nonnihil admirari soleo, cur studij literarū loca olim Græcis ab o-tio, Scholæ: Latinis uero Ludi dicta sint. Nam non ausim putare illos artium ingenuarum. ac Philoso-phiæ studia, ita uilia ac leuia habuisse, ut ea nō nisi per otium, & tanq̄ res ludicras sectanda esse censerent. Præsertim quum ob ea ipsa studia, cæteras omnes gentes, barbaras præ se uocarent, ac tanq̄ omnis humanitatis expertes despicerent. Cu-ius elationis causas certe non otiosas,

A 2 multo

Binarisierung 1

CLARISSI-
MO AC PRVDENTIS-
simo Senatui ciuitatis Noriber-
gę, Dominis suis colendissimis,
Sebaldus Heyden
S. D,

Voties de nostrarum Scholarum disciplina cogito, Viri Clarissi-mi, nonnihil admirari soleo, cur studij literarū loca olim Græcis ab o-tio, Scholæ: Latinis uero Ludi dicta sint. Nam non ausim putare illos artium ingenuarum. ac Philoso-phiæ studia, ita uilia ac leuia habuisse, ut ea nō nisi per otium, & tanq̄ res ludicras sectanda esse censerent. Præsertim quum ob ea ipsa studia, cæteras omnes gentes, barbaras præ se uocarent, ac tanq̄ omnis humanitatis expertes despicerent. Cu-ius elationis causas certe non otiosas,

A 2 multo

Binarisierung 2

Gliederung

1. Motivation und Grundlagen
2. Methoden
3. **Anwendungsbeispiel Drucke**
4. Anwendungsbeispiel Handschriften
5. Diskussion und Ausblick

Daten

- Selbst erstellt oder aus frei verfügbaren Quellen zusammengesucht
- Training
  - 21.5k Seiten aus 642 Werken
  - ca. 450 Jahre Druckgeschichte (ca. 1450 - 1900) → große Variabilität an Drucktypen
  - Verschiedene Sprachen: Deutsch, Latein, Französisch, Niederländisch
- Evaluation
  - 29 Werke zwischen 1506 und 1849
  - Repräsentative Abdeckung hinsichtlich Alter, Druckqualität, Schriftarten, etc.
  - Auswahl, Segmentierung und Transkription von vier bis fünf Seiten pro Buch

Dem hoeghen marschalck sent qurijn
Dienen. Da was der keyser gar fro Dz sein syn
en vnd materien erlengert vnd sch-
Sonder beghinsel Volmaect in Vreughden.
Stultoꝝ ꝛc̄. Q m̄ inquā stultoꝝ vt
complexus sit: ut sub æquatore est
Und angehängter besondern Tabell

Kallimachos Subkorpus, größtenteils Narrenschiffe (15./16. Jh.)

des siècles. Ces couronnes que j'ai si
contraste dans les physionomies de ceux
monstrant l'integrité de mon ame, vous
savant transporté de joie, que nous
les genitoires tant feconds, & les Vto-
Va donc pour Lisette; je n'en serai pas

Französisch (17.-19. Jh.)

hat ihn wegen seines eigenen Glückes, sich nicht in
alle Zweige der gemischten Mathematik, wie
wins ragt Sigurds Schreckhorn hell hinaus
Vorsaal, und ihre schönen Töne zogen alle
nur eine Kluft, die uns von dem gewünschten
So träumte er; ein Schwarm aufflatternder Eulen
in unsern Tagen von hundert Personen neun und
und entdeckte zuletzt einen schmalen hölzernen Steg;
lassen, daß das schwächliche Kind des reichen Hart-
er dadurch verrathen würde, daß er wirklich eine Neigung
Geschäftig eilten die Diener herbei. Der Herzog und
mich in Harnisch. Es handelt sich also nicht um einen
sind der Erste, dem ich von meinem Leben erzähle, damit

Frakturschriften 19. Jh.

cto procederet, omnium, quorum et
nem notare, videtur initium illius iubere obseruari,
in parte, mentio. Disserui autem ista
illius fixæ temperatũ quiddam ualent ê
ferre posim gratiæ, pro summis & immortalibus

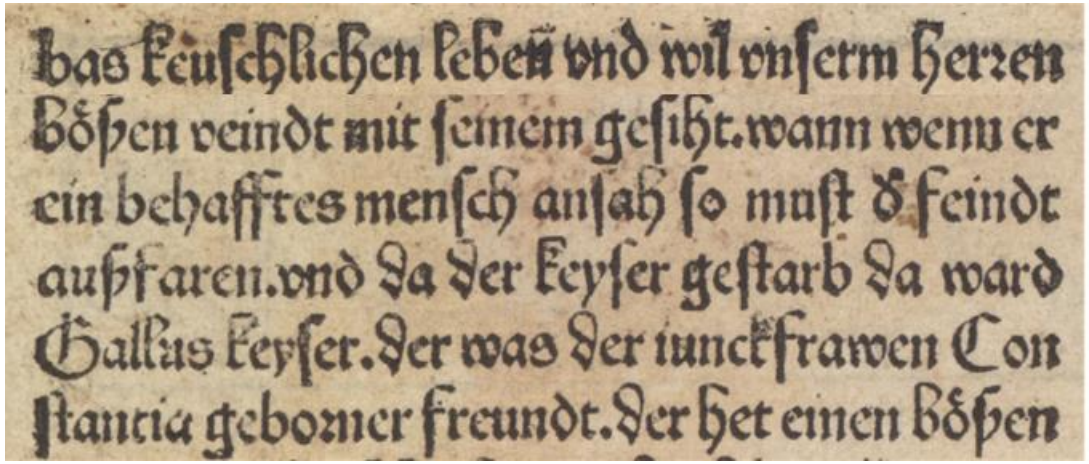
Camerarius (16. Jh.)

Evaluation

- Vergleich von Stufe 1 (alle Daten), Stufe 2 (nur ausgewählte Daten), deren Kombination sowie dem weit verbreitetem GT4HistOCR Standardmodell
- Bestimmung der Zeichenfehlerrate (Character Error Rate, CER)
- Stufe 1 und 2 vergleichbar (Balance > Masse), 1→2 am besten
- Deutlich (ca. 40%) besser als GT4HistOCR
- Mit Abstand größte verbleibende Fehlerquelle (ca. 50%): Einfügungen und Löschungen von Leerzeichen
- Durch technische Optimierungen (Datenaugmentierung, tiefere Neuronale Netze etc.) weitere Senkung auf ca. 1,6%

alle Daten	ausgew. Daten	CER in %
x		1,92
	x	1,95
x	x	1,80
	GT4HistOCR	2,84

Beispielergebnis mit ca. 1,6% CER


bas keuſchlichen leben vnd wil vnſerm herren
böſen veindt mit ſeinem geſiht. wann wenn er
ein behafftes menſch anſah ſo muſt ð feindt
außtaren. vnd da der keyſer geſtarb da ward
Gallus keyſer. der was der iunckfrawen Con
ſtantia geborner freundt. der het einen böſen

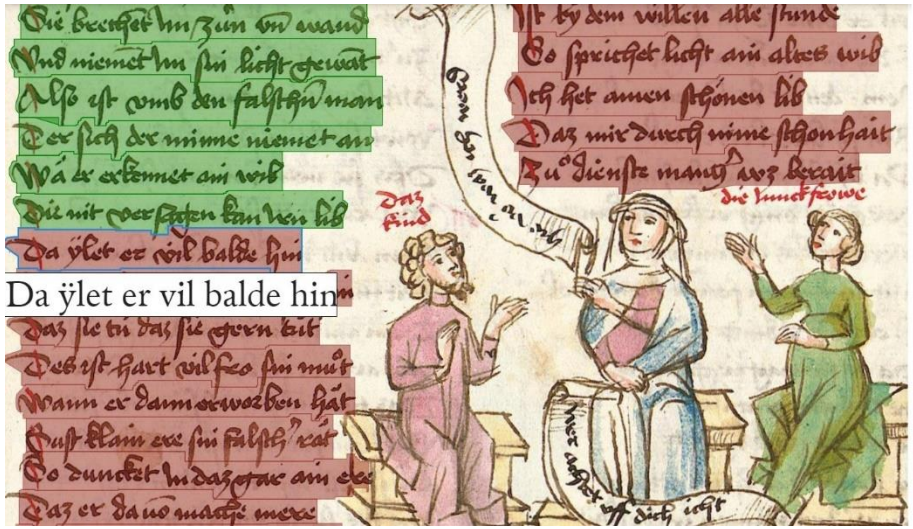
bas kſuſchlichen lcben vnd wil vnſerm herren
bößen veindt mit ſeinem geſiht. wann wenn er
ein behafftes menſch anſah ſo muſt ð feindt
außtaren. vnd da der keyſer geſtarb da ward
tGallus keyſer. der was der iunckfrawen Con
ſtantia geboꝛner freundt. der het einen bößen

Gliederung

1. Motivation und Grundlagen
2. Methoden
3. Anwendungsbeispiel Drucke
4. **Anwendungsbeispiel Handschriften**
5. Diskussion und Ausblick

Material: Mittelalterliche Handschriften

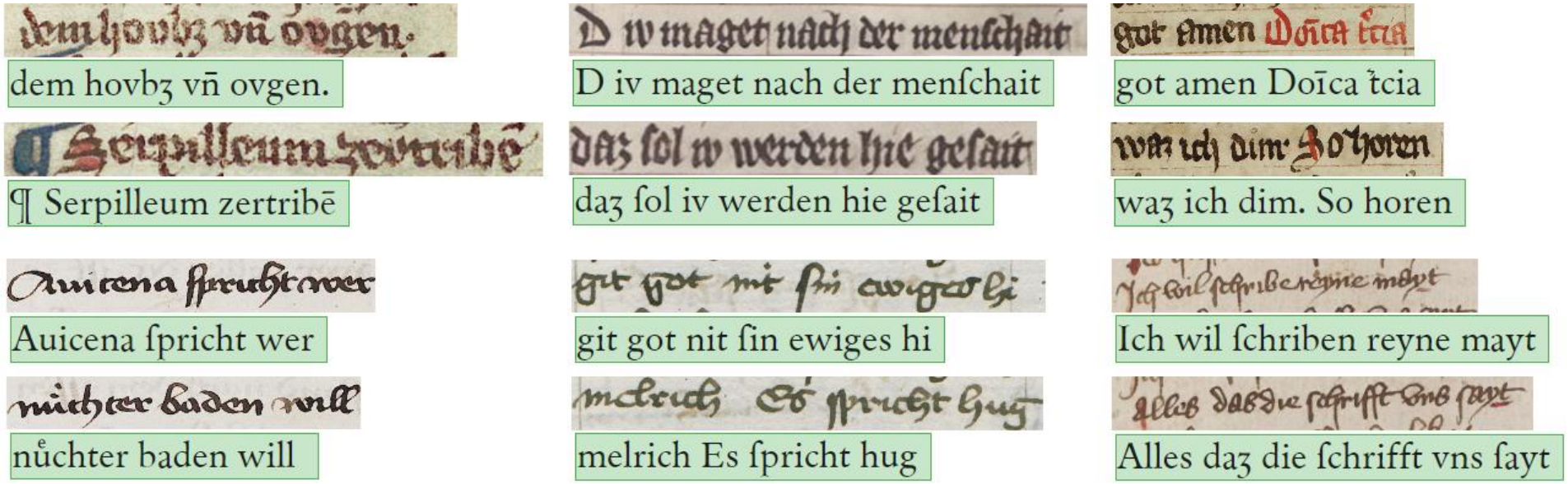
- Kooperation mit Dr. Stefan Tomasek (Lehrstuhl für deutsche Philologie, ältere Abteilung) anhand des Projekts „Konrad von Fußesbrunnen: Kindheit Jesu“
- Vielen Dank an die Projektgruppe und die fleißigen Hilfskräfte!
  - Dr. Stefan Tomasek, Florian Langhanki, Maximilian Wehner
  - Susanne Bremer, Lisa Gugl, Sebastian Hammer, Kiara Hart, Ursula Heß, Leonie Kampmann, Annika Müller, Anne Schmidt



Wå er erkennet ain wib
Wå er erkennet ain wib
Die nit verſagen kan Irn lib
Die nit verſagen kan Irn lib
Da ylet er vil balde hin
Da ylet er vil balde hin
Vnd wirbet vaſt daz iſt ſin ſin
Vnd wirbet vaſt daz iſt ſin ſin

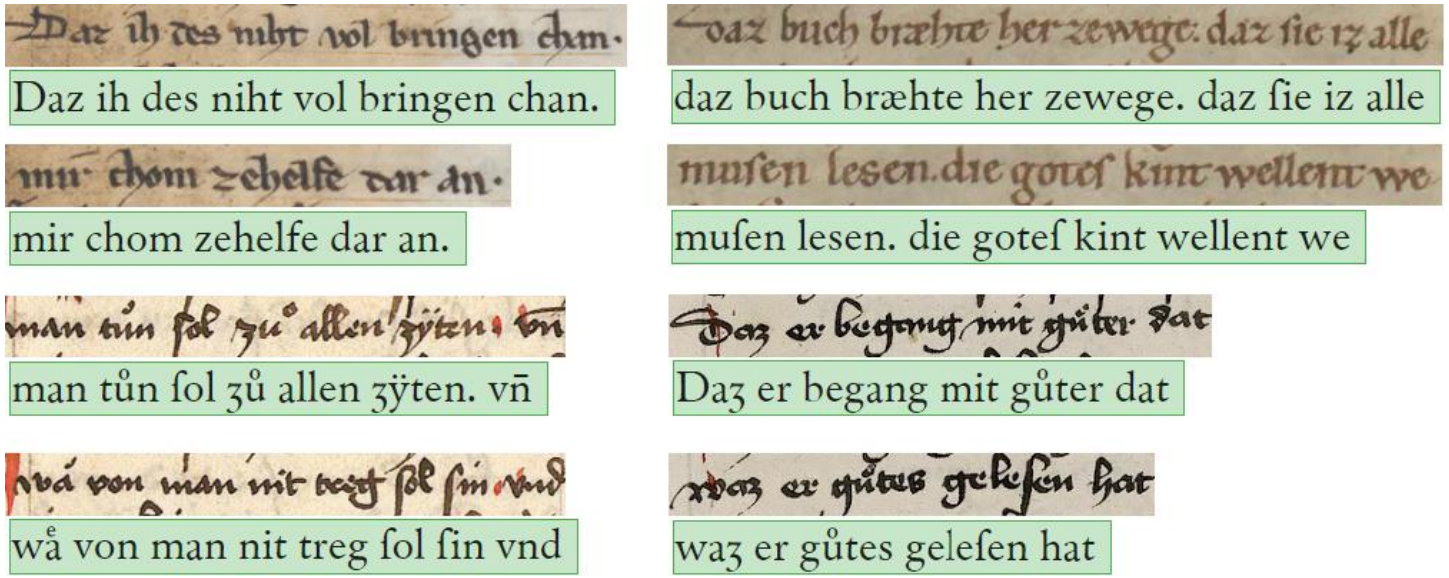
Trainingsdaten

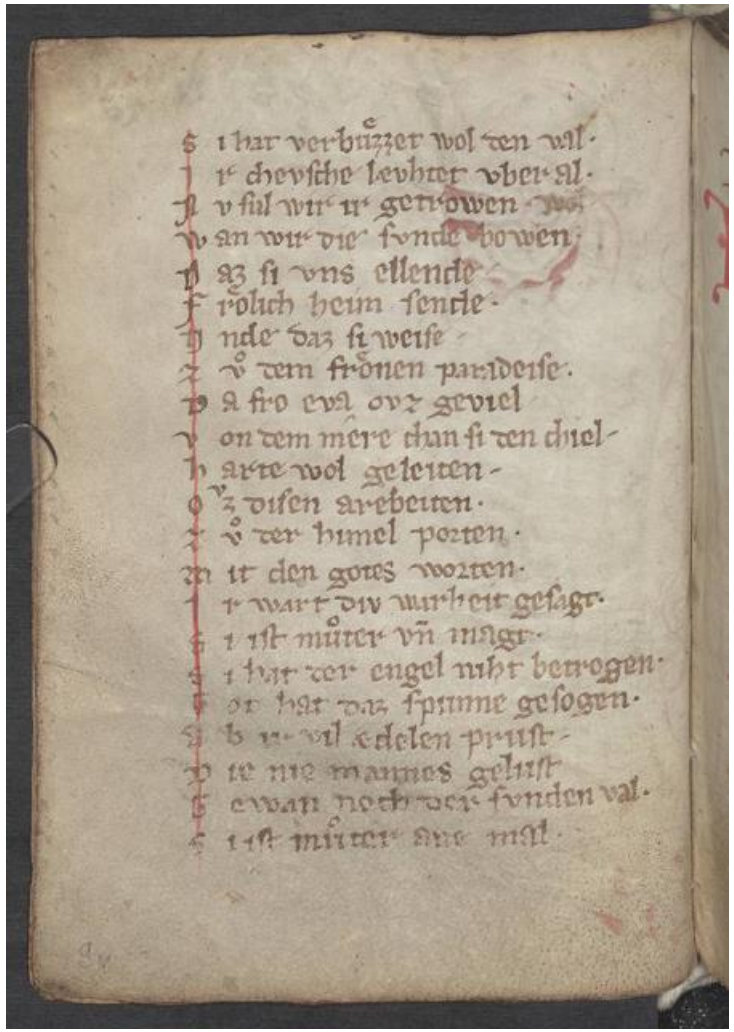
- Selbst erstellt oder aus frei verfügbaren Quellen zusammengesucht
- Frühes 13. bis spätes 15. Jh.: Kindheit Jesu, Marienleben, Parzival, …
- 35 Werke, knapp 300 Seiten, ca. 12,5k Zeilen → zwei gemischte Modelle



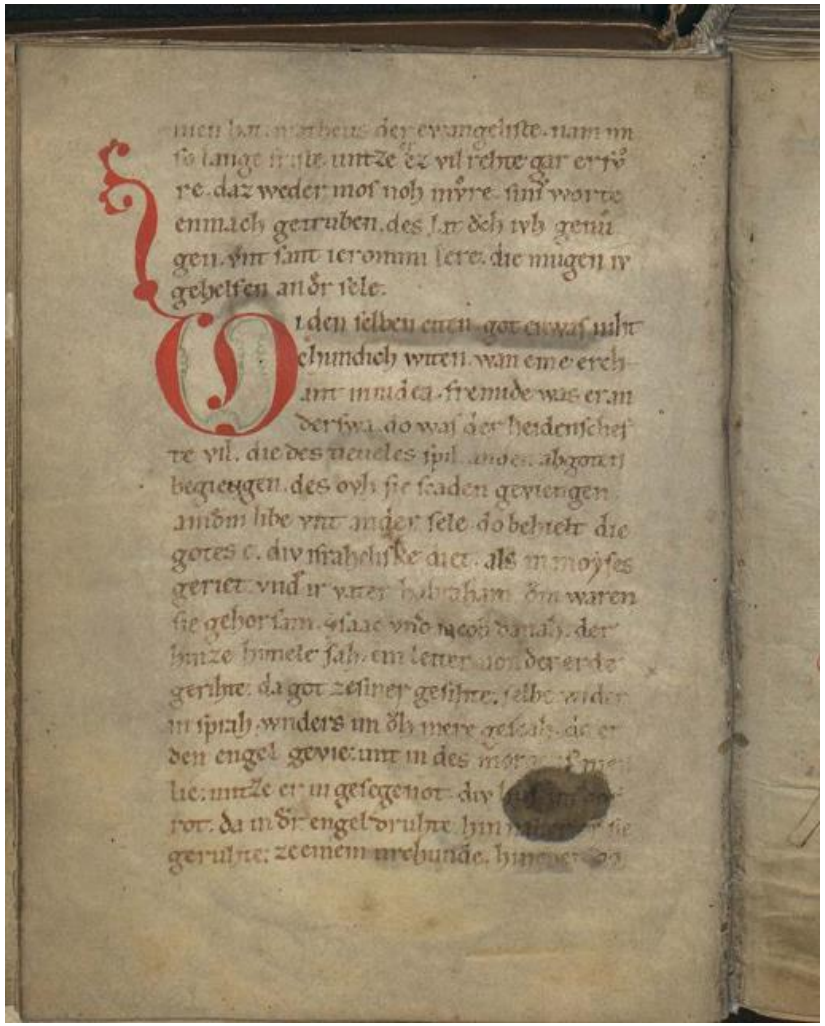
Evaluationsdaten

- Vier weitere Handschriften (2x Gotische Buchschrift, 2x Bastarda)
- Kindheit Jesu, Marienleben (x2), Der Welsche Gast (x2)
- Jede Hs. noch einmal unterteilt in Trainings- (32) und Evaluationsseiten (8-18)

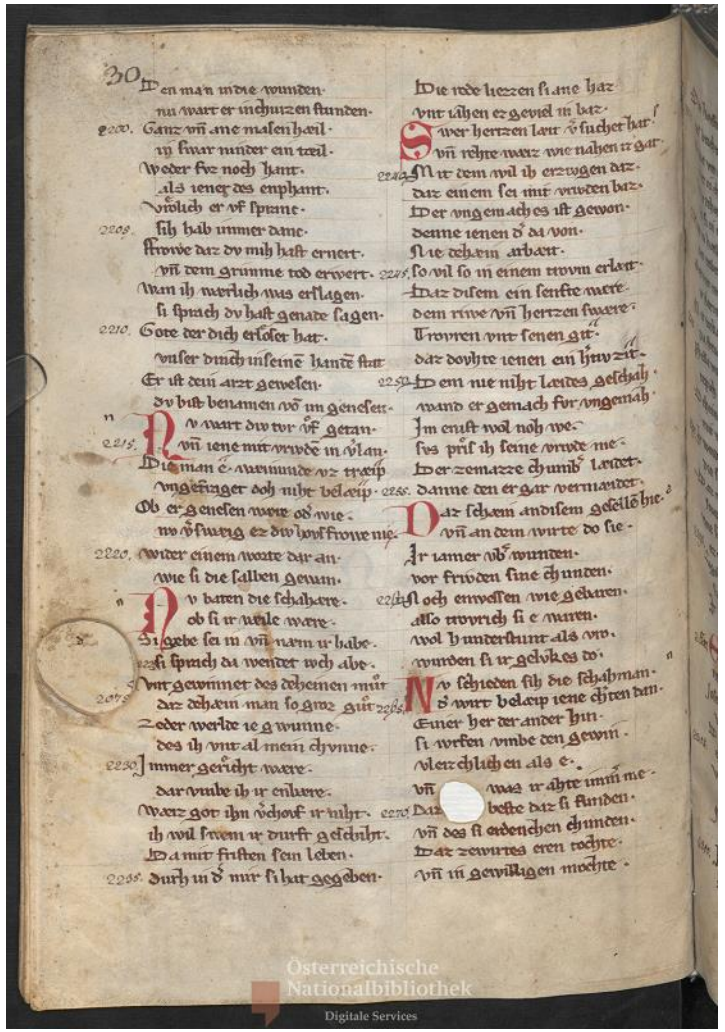




Wernher-Wien

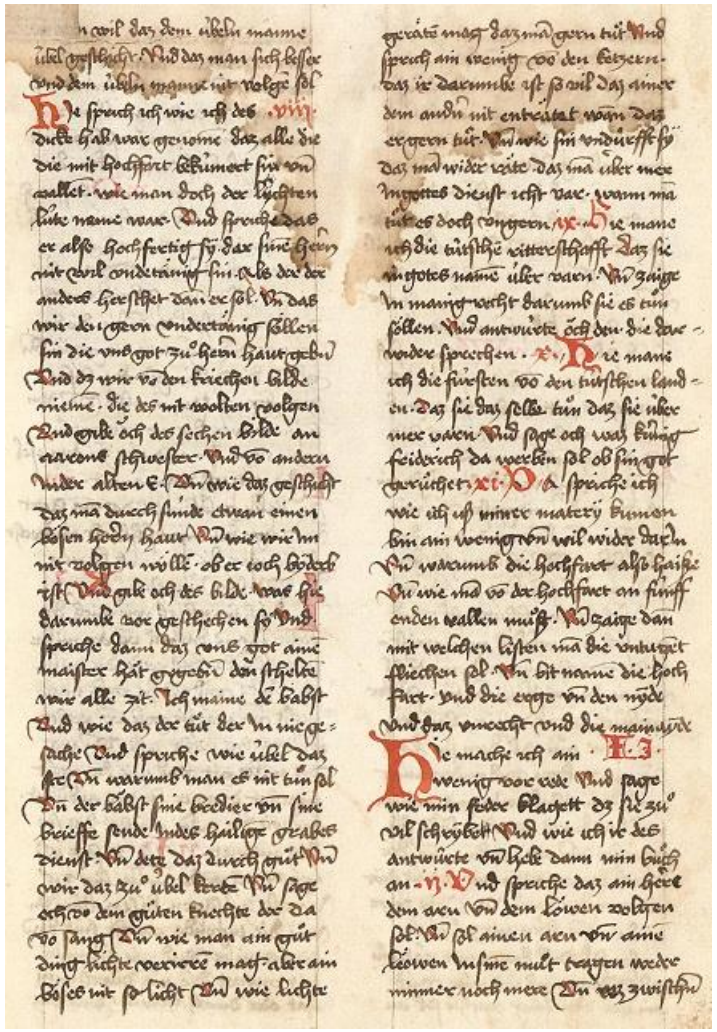


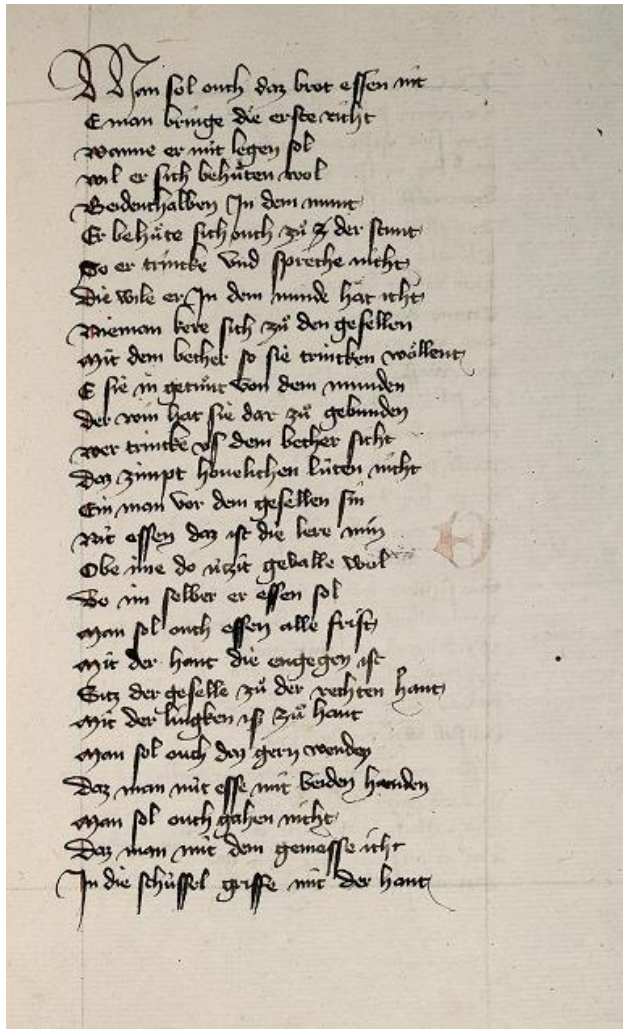
Wernher-Krakau



Handschrift-B

Welscher-Gast-1





Welscher-Gast-2

Simulation einer schrittweisen Transkription

- Iterative Verdopplung der Trainingsmenge
  - 2, 4, 8, 16, 32 Seiten GT
  - 0 Seiten: out-of-the-box Ergebnis des gemischten Modells
- Evaluation auf fixen Evaluationsset und Berechnung der CER (%); Durchschnitt über alle Werke
- Angesichts des Materials/Modells gute ootb CER
- Training auf lediglich zwei Seiten halbiert die CER
- Weitere deutliche Steigerungen (bis deutlich unter 2% CER), jedoch abnehmende Effizienz
- Verbleibende Fehler von Leerzeichen und Punkten dominiert

# Seiten	Pretrained CER in %	Verb. in %
0 (ootb)	6,22	-
2	3,27	48
4	2,58	21
8	2,17	16
16	1,94	11
32	1,65	15

Nutzen gemischter Modelle

- Vergleich der Ergebnisse …
  - beim Starten des Trainings From Scratch
  - bei Nutzung eines vortrainierten Modells als Ausgangspunkt (Pretrained)
- Verb. PT/FS: Verbesserungsfaktor Pretrained im Vergleich zu From Scratch
- Riesiger Unterschied, speziell bei wenig Trainingsmaterial (85% bei 2 Seiten!)
- Effekt nimmt mit steigender Seitenzahl stetig ab, ist jedoch selbst bei 32 Seiten noch klar erkennbar (knapp 40%)

# Seiten	From Scratch CER in %	Pretrained CER in %	Verb. PT/FS in %
2	21,12	3,27	85
4	10,74	2,58	76
8	5,82	2,17	63
16	3,78	1,94	49
32	2,72	1,65	39

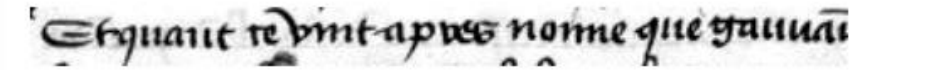
Gliederung

1. Motivation und Grundlagen
2. Methoden
3. Anwendungsbeispiel Drucke
4. Anwendungsbeispiel Handschriften
5. **Diskussion und Ausblick**

Zusammenfassung

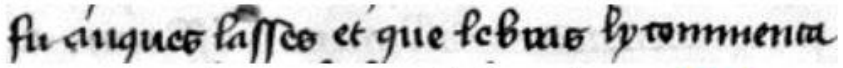
- Gemischte Modelle trainiert für …
  - Drucke in lateinischer Schrift zwischen 1450 und 1900
  - Mittelalterlichte Handschriften (Gotische Buchschriften und Bastarden)
- Hochperformant …
  - bei der out-of-the-box Anwendung
  - als Ausgangspunkt für Finetuning
- Balance > Masse → lieber viele Quellen mit jeweils wenig Zeilen, als umgekehrt
- Selbst mit wenigen Seiten große Verbesserungen durch Finetuning möglich
- Weitere Erkenntnisse zu Einfluss von Netzstrukturen und Binarisierungen sowie in-domain und out-of-domain Pretraining; für Details s.
  - Reul, Wick, Nöth, Büttner, Wehner, Springmann: Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning. HiP'21.
  - Reul, Tomasek, Langhanki, Springmann: Open Source Handwritten Text Recognition on Medieval Manuscripts using Mixed Models and Document-Specific Finetuning. submitted to DAS22.

Aktuelle Arbeiten I



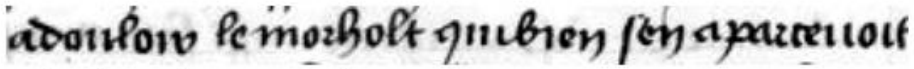
Et quant ce vint apres nonne que gauuāi

Et quant ce vint apres nonne que gauuāi



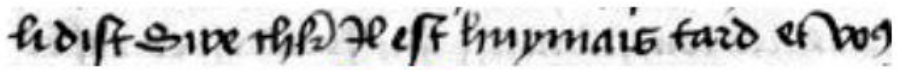
fu auques lasses et que le bras lycomnmenca

fu auques lasses et que le bras lycommenca



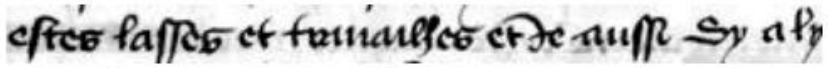
adouloir le morholt qui bien sen aparceuoit

adouloir le morholt qui bien sen aparceuoit



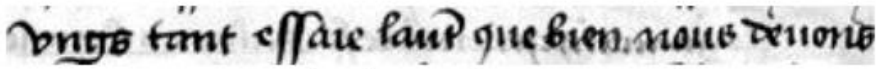
li dist Sire chlr̄ Jl iest lhuymais tard et vo⁹

li dist Sire chlr̄ Jl est huymais tard et vo⁹



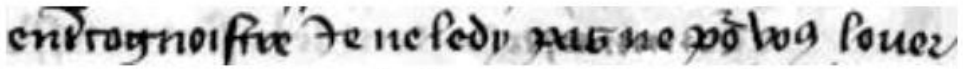
estes lasses et trauailles et Je aussi Sy aly

estes lasses et trauailles et Je aussi Sy aly



vngs tant essaie lauſt que bien nous deuons

vngs tant essaie lauſt que bien nous deuons

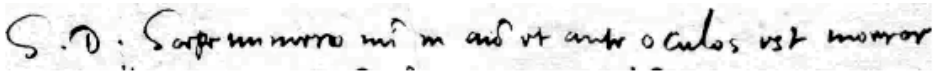


enſtcognoistre Je ne ledy pas ne pō vo9 louer

enſtcognoistre Je ne ledy pas ne pō vo9 louer

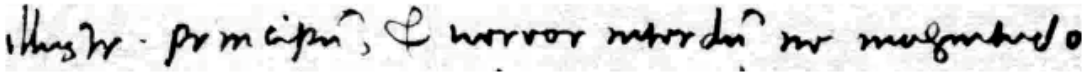
Frz. Handschrift aus dem 15. Jh.
Fehlerrate: etwa 0,6%
Transkription/Korrektur: Prof. Burrichter

Aktuelle Arbeiten II



S.D. Saepenumero mī in aniō et ante oculos est moeror

S.D. Saepenumero mī in aiō et ante oculos est moeror

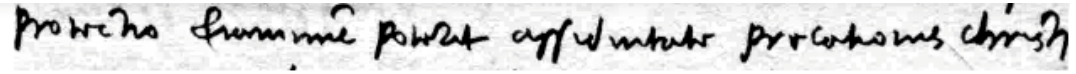


illustr. principū, & vereor interdū ne magnitudo

illustr. principū, & vereor interdū ne magnitudo

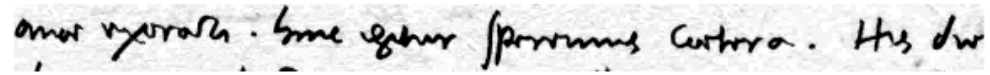
Ppr. Georgij, gravius affligat. Ego quidem hunius

pr. Georgij, gravius affligat. Ego quidem huius



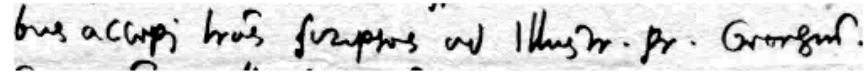
Pprotectio &iamnūc poterit affiduintate precationis christi

protectio &iamnūc poterit affiduitate precationis christi



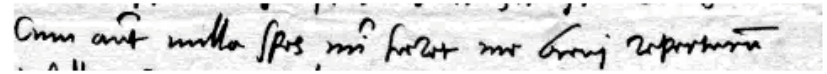
anae exorari. hinc igitur ſperemus caetera. His die

anae exorari. hinc igitur ſperemus caetera. His die



bus accepi lrās ſcriptas ad Illustr. pPr. Georgiū.

bus accepi lrās ſcriptas ad Illustr. Pr. Georgiū.

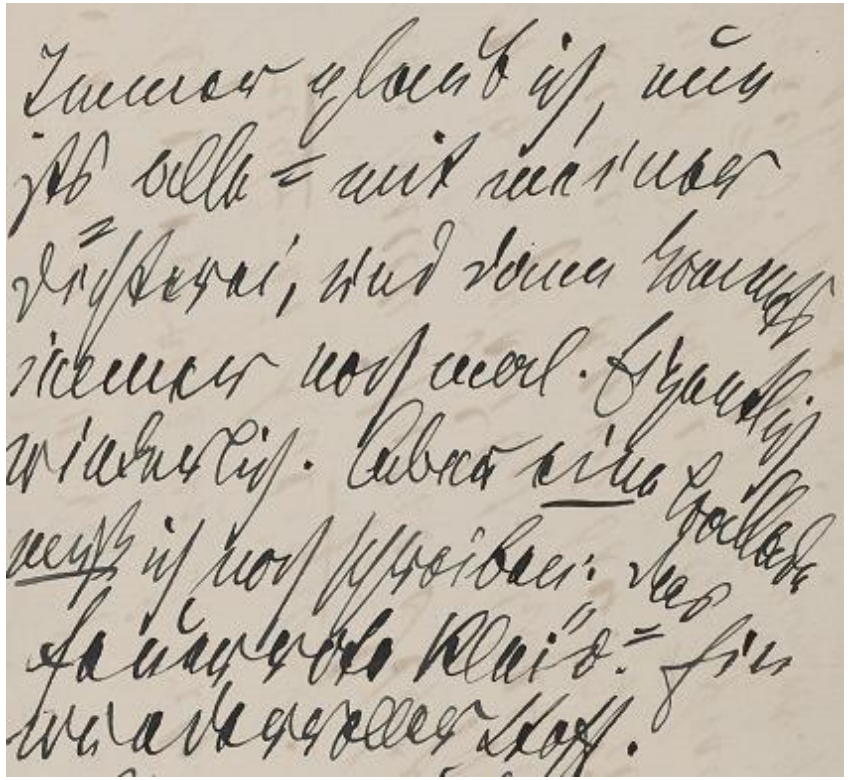


Ccum autť nulla ſpes mī fiuereit me brevi reperturū

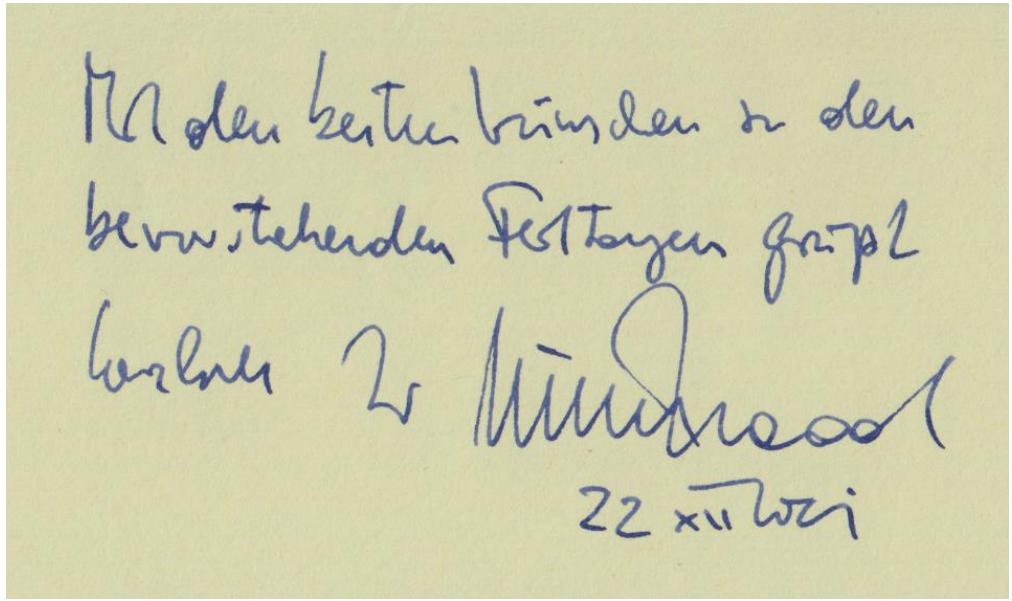
cum auť nulla ſpes mī fuerit me brevi reperturū

Joachim Camerarius der ~~Ältere~~ Schmierfink
Fehlerrate: etwa 3,2%
Transkription/Korrektur: Dr. Schlegelmilch

Ausblick: Schlimmer geht immer



Ein Herr Liliencron (?); Mitte 19. Jh. (?)
Antragskooperation mit Univ. Hamburg



Würzburger Musikprofessor, der anonym bleiben möchte; Dezember 2021