

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

Christian Reul

Zentrum für Philologie und Digitalität "Kallimachos" (ZPD)
Universität Würzburg





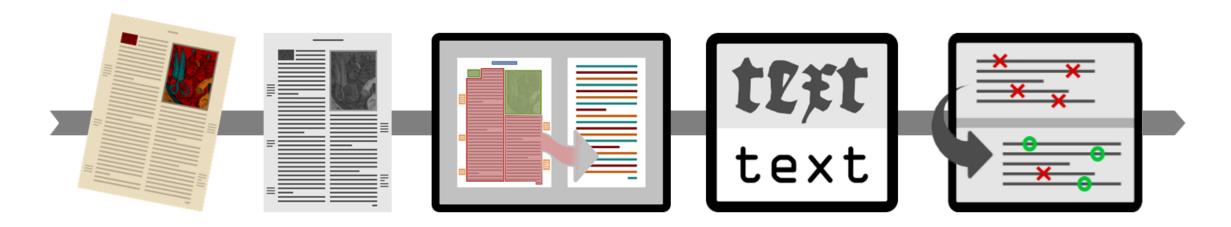
Gliederung

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

Er wird eifrig gesammelt.

Er wird eifrig gesammelt.



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-MOACPRVDENTIS-

fimo Senatui ciuitatis Noriberge, Dominis fuis colendifsimis, Sebaldus Heyden S. D.



Voties de nostrarum Scholarum disciplina cogito, Viri Clarissis mi, nonnihil admirari soleo, cur studij literarij loca olim Græcis ab ostio, Scholæ: Latinis ue?

re illos artium ingenuarum. ac Philosophiæ studia, ita uilia ac leuia habuisse, ut eano nisi per otium, & tanos res ludicras sectanda esse censerent. Presertim quum ob eaipsa studia, cæteras omnes gentes, barbaras pre se uocarent, ac tanos omnis humanitatis expertes despicerent. Cue ius elationis causas certe non otiosas,

CLARISSI-MOACPRVDENTIS-

simo Senatui ciuitatis Noriberge, Dominis suis colendissimis, Sebaldus Heyden S. D.



Voties de nostrarum Scholarum disciplina cogito, Viri Clarissis mi, nontihil admirari soleo, cur studij literarij loca olim Græcis ab os tio, Scholæ: Latinis ue?

ro Ludí dicia sint. Nam non ausim putar re illos artium ingenuarum. ac Philosophiæ studia, ita uilia ac leuia habuisse, ut ea no nisi per otium, & tanci res ludicras sectanda esse censerent. Presertim quum ob ea ipsa studia, cæteras omnes gentes, barbaras pre se uocarent, ac tanci omnis humanitatis expertes despicerent. Cue ius elationis causas certe non otiosas,

acis ab o=

Voties de nostrarum Scholarum disciplina cogito, Viri Clarissis mi, nonnihil admirari soleo, cur studij literarij loca olim Græcis ab os tio, Scholæ: Latinis ue?

ro Ludí dicta sint. Nam non ausim putal re illos artíum ingenuarum. ac Philosophiæ studia, ita uilia ac leuía habuisse, ut eano nisi per otium, & tanos res ludicras sectanda esse censerent. Presertim quum ob ea ipsa studia, cæteras omnes gentes, barbaras pre se uocarent, ac tanos omnis humanitatis expertes despicerent. Cue ius elationis causas certe non otiosas,

CLARISSI-MOAC PRV DENTISfimo Senatui ciuitatis Noriber-

ge, Dominis suis colendissimis,

Sebaldus Heyden

A 2 multo

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 quirin Vienen. Dawas Ver kepser gar fro 83 sein son en und materien erlengert und sch-Sonder beghinsel Volmarct in Breughden. Stultop zc. Om inquastultop ut 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.)

bat ihn wegen seines eigenen Glutes, sich nicht in alle Zweige deringemischten Mathematikeinie wins ragt Sigurds Schreckhorn hell hinaus Borfaal, und ihre fconen Tone zogen alle nur eine Kluft, die uns von dem gewünschten So traumte er; ein Schwarm aufflatternber Gulen in unsern Tagen von hundert Personen neun und und entbedte zulett einen ichmalen hölzernen Steg; lassen, daß das schwächliche Kind des reichen Harter baburch verrathen würde, daß er wirklich eine Neigung Geschäftig eilten die Diener herbei. Der Berzog und mich in Harnisch. Es handelt sich also nicht um einen find der Erste, dem ich von meinem Leben erzähle, damit Frakturschriften 19. Jh.

Cto procederet, omnium, quorum et

mem notare, uidetur initium illius iubere observari, in parte, mentio. Disserui autem ista illius sixæ temperatű quiddam valent ê serre posim gratiæ, pro summis er 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), $1\rightarrow 2$ 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 % |
|---------------|------------------|-------------|
| × | | 1,92 |
| | X | 1,95 |
| X | Х | 1,80 |
| G1 | 2,84 | |



Beispielergebnis mit ca. 1,6% CER

Bose keuschlichen leben und wil unserm herren Bosen veindt mit seinem gesist, wann wenn er ein behafftes mensch ansah so must deindt außfaren und La Ler keyser gestarb La ward Ballus keyser. Ler was der innekframen Con stancia gebomer freundt. Der het einen Bosen

bas ksuschlichen Icben vnd wil vnserm herzen 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 gebozner freundt. der het einen bößen



Gliederung

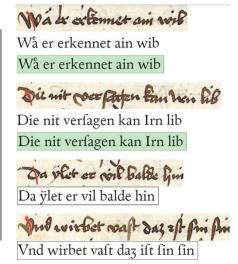
- 1. Motivation und Grundlagen
- 2. Methoden
- 3. Anwendungsbeispiel Drucke
- 4. Anwendungsbeispiel Handschriften
- 5. Diskussion und Ausblick



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

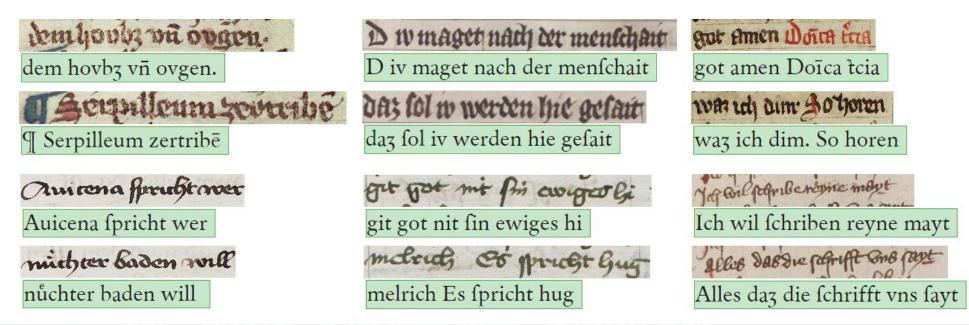






Trainingsdaten

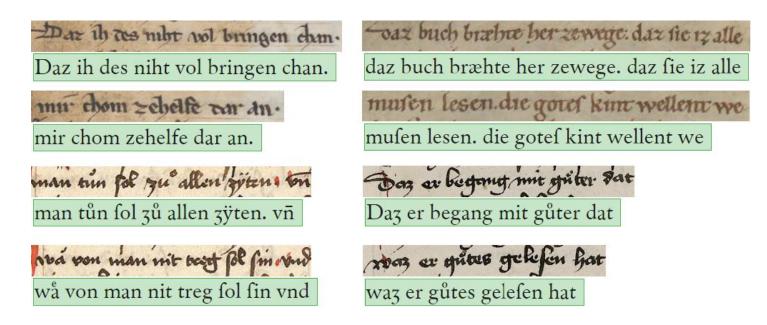
- Selbst erstellt oder aus frei verfügbaren Quellen zusammengesucht
- Frühes 13. bis spätes 15. Jh.: Kindheit Jesu, Marienleben, Parzival, ...
- 35 Werke, knapp 300 Seiten, ca. 12,5k Zeilen → zwei gemischte Modelle



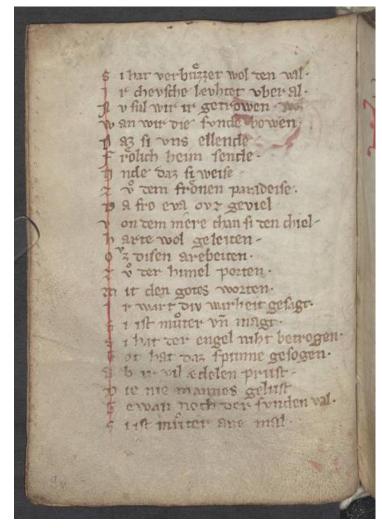


Evaluationsdaten

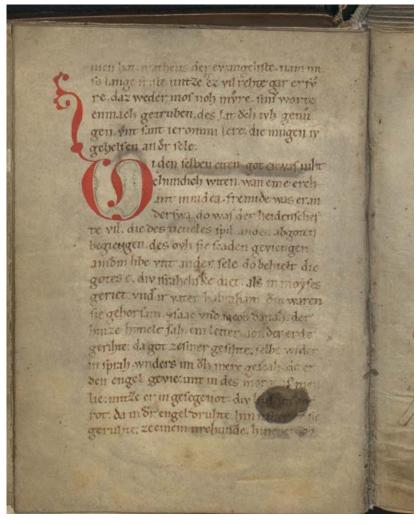
- Vier weitere Handschriften (2x Gotische Buchschrift, 2x Bastarda)
- Kindheit Jesu, Marienleben (x2), Der Welsche Gast (x2)
- Jede Hs. noch einmal unterteilt in Trainings- (32) und Evaluationsseiten (8-18)



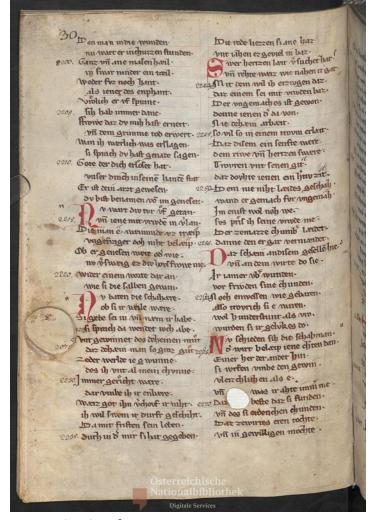




Wernher-Wien



Wernher-Krakau



Handschrift-B





Welscher-Gast-1

novil day done allela marme Titel mellegit New day man fret leffer winden utilin manut nit water fol I to found rate wire rely des . viny. Dutte Gabavar operome Soz alle die Die mit Bodffort BRimeet fix son walter waterman doct der troften Pute mani was Dut founded at or all good feeting fir Sax fine feet andered Ger Affet Jan ex fol the das wir der geon conductioning Bellen fin die vooget zu Geen Gaungebil (2nd by wir to der frientfor Gille mieme. The des mit woolben roolgen (2nd orde out des forten Bride an ourcome Africe for Privato andern Miles alten & Dinomic day gefffully Das ma duccif finde dream comen Bofon Gody Baur Paris wie hom mit rologen will ober with bodate iff Drive gale orgate Gilde Moone Gic Sarumbe mor peftferfon for Ond. Grentle Jam Day cont yot ame maiper Bat orgetie don Athelie voice alle zit. Toly maine de babit And wie day doe thit Ber In micge: factio (and foreign wie wife day Din der Ballt fine beedier on fine Breiffe pond Indes Builing grabes vient the det day durity opit And ortico dem griten Phortite dos da Too land Du nik man air gut ding hit roccince mant aleram Bifes wit porticity (Men wie liefter

overate mast farma over tit Wind speal am wenty coo den letzernbus in Darmile of Boil Das since dem and mit entrast war Das expern bit Valute fin whourff for Das ma wider wite Das ma wher more Insported dient refit vac gram ma titles doch congress is a mane up 8ic wirth & comme mostoco name aler vaen Na gaige In maning wift barumb hi co our totten And amouniste och den Sie leve worder spracefor . 20, 10 ic mans was die fineften vo den truftgen land. en Day hi Day folle tim day his where mer warn Dut page org was filling feiderich da werben pol of fingsof descriper xi Da Apriche ich wie at uf winer matery fumon Em am wemporon wil wider Darla Phi wacume die Gorafinor also Gaige Ou wie and to doc Good force on finis enden walken mit Angaine dans mit welden liften ma Die inhuget flietlen get An Dit name die tjort Fact your die egge vin den mide vintering il buy towns manufert Ic marthe reth aim . wenig wood ask Mus page wit min force Plagett de he su' ारित्रकारिया केल किया मिन्नियं के कि Antwicte vow Role Dann win Guit an - 17. Dut pourte das am Gere dem acu von dem Lowen zolegen BE to Blainen acu on ame tiowen Infine mult trangen weeker minmer north mere (Du ver 3 wiftin

I for for outh day boot effen me Eman bringe Die er feerentst mont er firt behitten wol Bedeutjathen (In dem munt Do er timbe and porcele miche Die Wite or I'm dom hund har whe anieman lieve fiel and den defetten egit dem berhel jo sie trinthen wolfent Det som har fie dar 30 gebunden aver trimbe of Dem berfer frest Den Bimpt Souchesfon Litten miste Sin man Von dem stefetten Fin poir affen den afe die leve mins Obe ine do right gelatte that To im petter er effen pl agan pot and, after atte first mir de four Die confeger of Sur Der de felle gil der renteen hames onen pot and Day ofery monday Day men mit off mit Buden handen man pot out frather ment. Day men mit den gemaffe refe in die philpet wife mit Der hanne

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 | Verb. |
|----------|------------|-------|
| | CER in % | 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 | Pretrained | Verb. PT/FS |
|----------|--------------|------------|-------------|
| | CER in % | CER in % | 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: <u>Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition</u> and Finetuning. HiP'21.
 - Reul, Tomasek, Langhanki, Springmann: <u>Open Source Handwritten Text Recognition on Medieval Manuscripts using Mixed Models and Document-Specific Finetuning</u>. submitted to DAS22.



Aktuelle Arbeiten I

Erquant te brit apres nome que ganuai

Et quant ce vint apres nonne que gauuāi Et quant ce vint apres nonne que gauuāi

fu auques lasses et que le bras lycommenca fu auques lasses et que le bras lycommenca

adouloir le morholt qui bien sen aparceuoit adouloir le morholt qui bien sen aparceuoit

Frz. Handschrift aus dem 15. Jh. Fehlerrate: etwa 0,6%

Transkription/Korrektur: Prof. Burrichter

holf sive the Heff huymais faid et bog

li dist Sire chlīr Jl iest lhuymais tard et vo' li dist Sire chlīr Jl est huymais tard et vo'

estes lasses et trauailles et Je aussi Sy aly

estes lasses et trauailles et Je aussi Sy aly

vngs tant essaie laut que bien nous deuons
vngs tant essaie laut que bien nous deuons

entrognoistre Je ne ledy pas ne pō vo9 louer entcognoistre Je ne ledy pas ne pō vo9 louer



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 hu<mark>ni</mark>us pr. Georgij, gravius affligat. Ego quidem huius

Joachim Camerarius der Ältere Schmierfink Fehlerrate: etwa 3,2%

Transkription/Korrektur: Dr. Schlegelmilch

Portectio & diamnuc poterit affiduintate precationis christi protectio & diamnuc poterit affiduitate precationis christi protectio & diamnuc poterit affiduitate precationis christi mos vyorolo : has gener formus corbera. His die anae exorari. hinc igitur speremus caetera. His die anae exorari. hinc igitur speremus caetera. His die

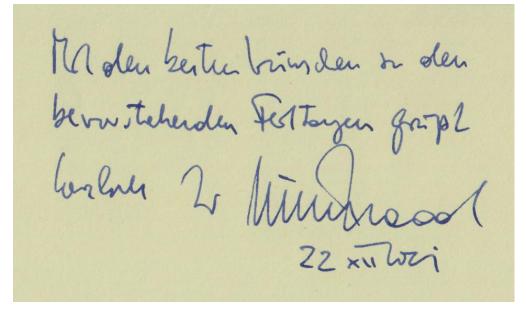
bus accepi lrās scriptas ad Illustr. pr. Georgiū. bus accepi lrās scriptas ad Illustr. pr. Georgiū.

Com aut mila spes mi facet me druj reperturu Coum aut nulla spes mī fiuereit me brevi reperturu cum aut nulla spes mī fuerit me brevi reperturu



Ausblick: Schlimmer geht immer

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



Würzburger Musikprofessor, der anonym bleiben möchte; Dezember 2021

