



OCR4all – Eine (semi)automatische Open Source Software für die OCR historischer Drucke

Christian Reul

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



05.05.2021



Gliederung

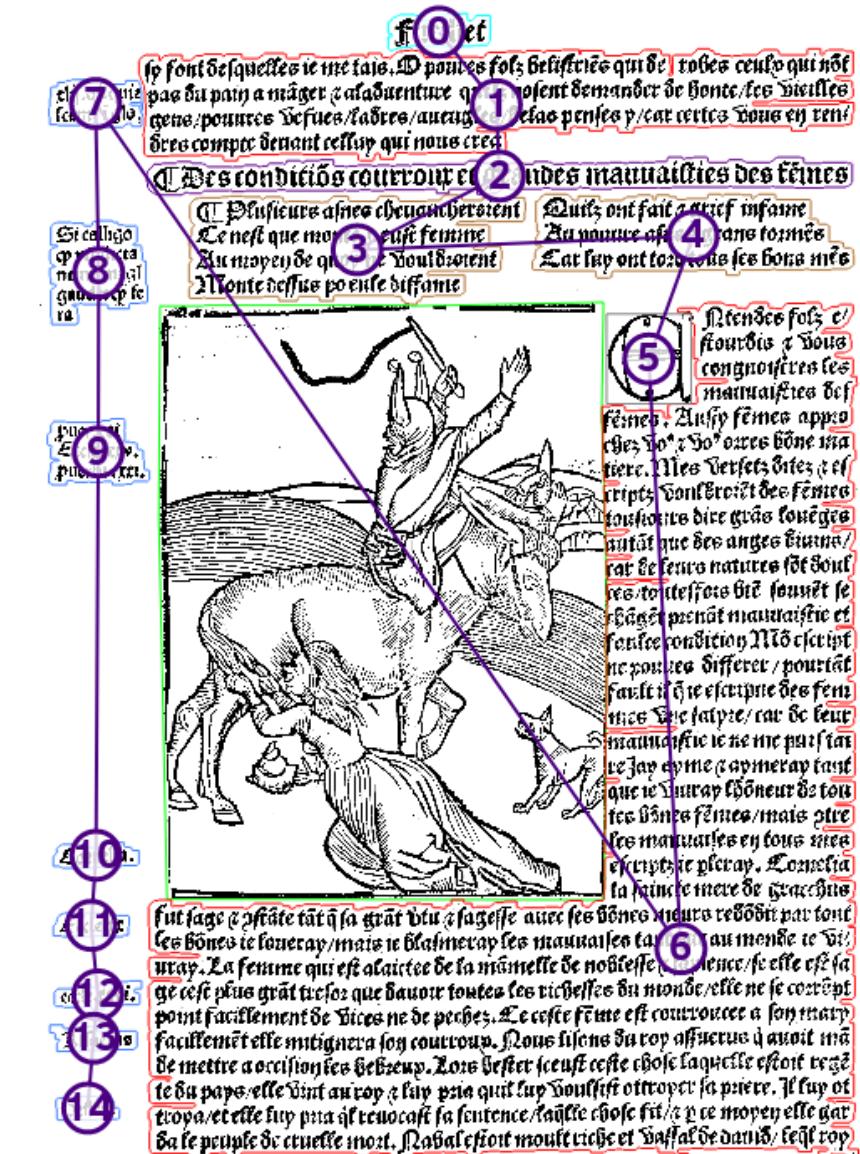
- 1. Einleitung**
2. Submodule
3. Workflow
4. Live Demo
5. Evaluation
6. Diskussion und Ausblick

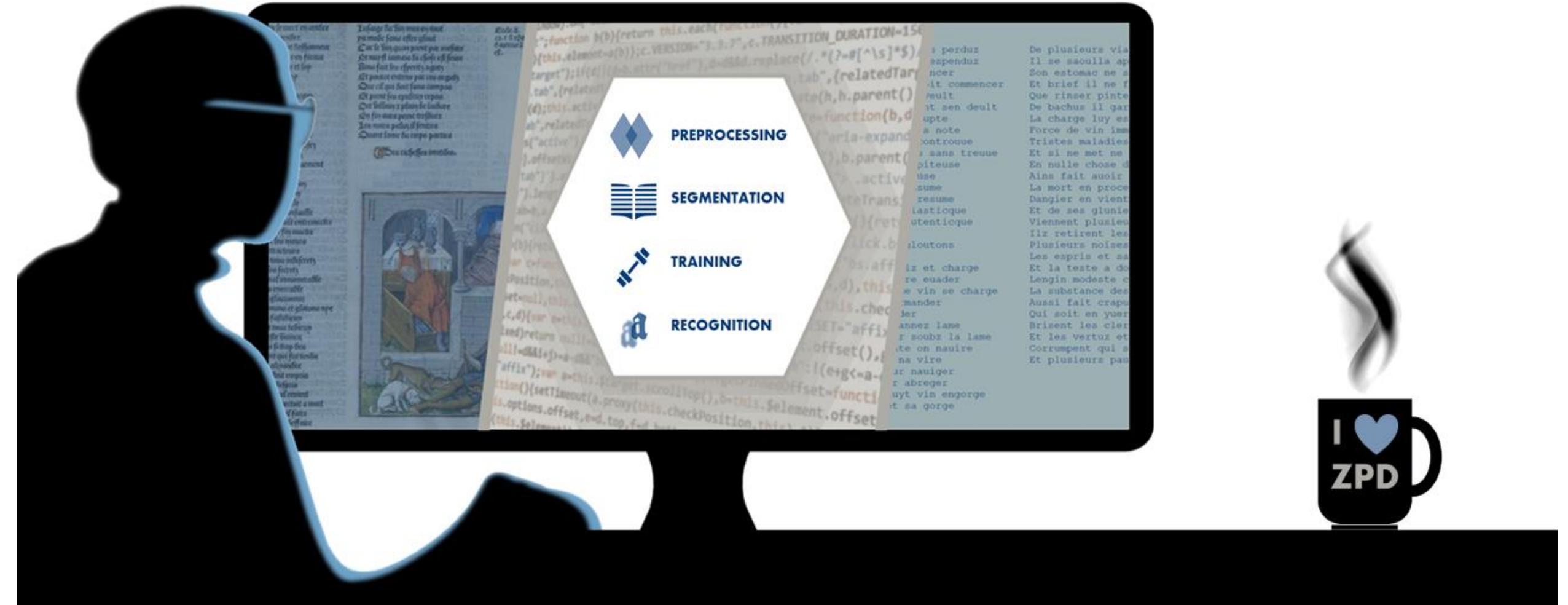
OCR4all – Motivation und Überblick

- Bestehende Open Source OCR Software ist mächtig, kann nicht-technische Nutzer jedoch schnell überfordern:
 - Installation nicht unbedingt trivial
 - Keine grafische Benutzeroberfläche, sondern ungewohnte Kommandozeile
 - ...
- Die Idee hinter OCR4all:
 - Verständlich und anwendbar auch für nicht-technische Nutzer
 - Nutzung von Docker/VirtualBox → Einfache Installation, Unabhängigkeit vom Betriebssystem
 - Basiert auf unterschiedlichen Open Source OCR Tools

OCR4all – Historie und Aktueller Stand

- Ursprünglich entwickelt für die OCR sehr (sehr!) alter Drucke ([Projekt Narragonien digital](#)):
 - Sehr komplizierte Layouttypisierung
 - Training werkspezifischer Modelle unerlässlich aufgrund der höchst varianten Typographie
 - Anspruch häufig: 100% Erkennungsgenauigkeit für Layout und OCR
→ Nutzer nehmen dafür einen gewissen manuellen Korrekturaufwand in Kauf
- Mittlerweile erfolgreich auf großer Bandbreite von Drucken (15. bis 21. Jh.) eingesetzt
- Hauptziel: Erhöhung des Automatisierungsgrads und der Robustheit
- Work in progress!





www.ocr4all.de

Gliederung

1. Einleitung

2. Submodule

3. Workflow

4. Live Demo

5. Evaluation

6. Diskussion und Ausblick

Aktuell integriert

- OCropus
 - Vorverarbeitung
 - Automatische (Zeilen)Segmentierung
- Calamari
 - Texterkennung
 - Modelltraining
 - Evaluation
- LAREX
 - Ursprünglich zur interaktiven Regionensegmentierung inklusive semantischer Auszeichnung entwickelt
 - Mittlerweile umfangreiches Korrekturtool, u. a. für
 - Regionen- und Zeilenkoordinaten
 - Semantische Typisierung und Lesereihenfolge
 - Text (Ground Truth Erstellung für Training!)

Zwischenfazit und -ausblick

- Bislang begrenzte Auswahl
 - Ursprünglich für konkreten Anwendungsfall ([Narragonien digital](#)) entwickelt
 - Damaliges Ziel: erstmal funktionierenden Workflow schaffen
 - Fokus auf ausgewählte „beste“ Lösungen
- Anbieten weiterer Lösungen unbedingt intendiert
 - Problemlose Anbindung, solange klar definierte Schnittstellen eingehalten werden
 - Workflow und Lösungen für einzelne Schritte frei kombinierbar
 - Baukastenprinzip!
- Klingt nach OCR-D?!
Später mehr...
„Nimm die Binarisierung von OCropus, die Segmentierung von Tesseract und die Texterkennung von Calamari“
(Konstantin Baierer, vor 20min)

Gliederung

1. Einleitung
2. Submodule

3. Workflow

4. Live Demo
5. Evaluation
6. Diskussion und Ausblick

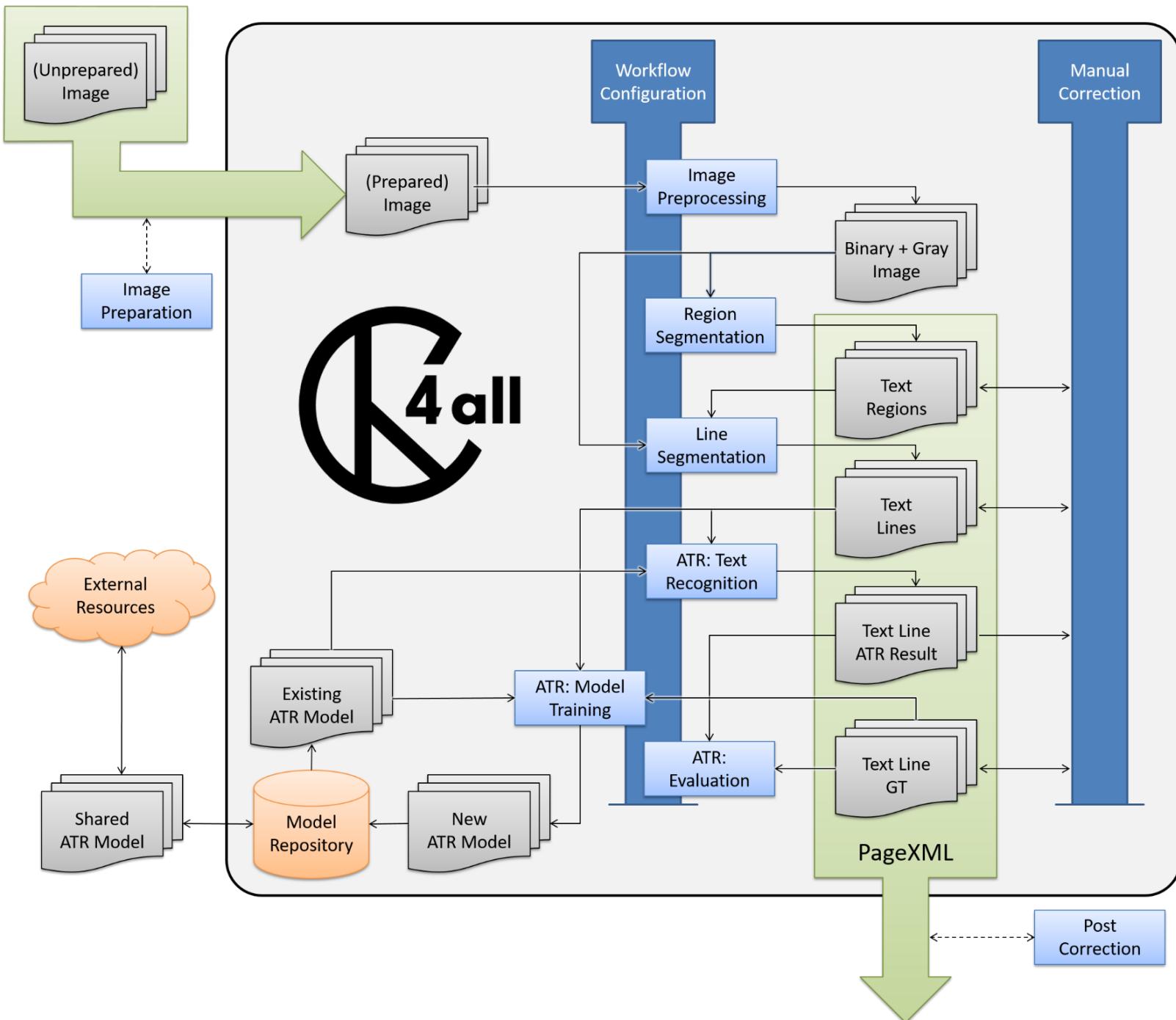


Image Preparation

- **Input:** unverarbeitete Bilder
Output: vorbereitete Bilder
- Notwendige Vorbereitungen variieren von Werk zu Werk:
 - Teilung von Doppelseiten
 - Rotation
 - Entfernung von störenden Bildrändern
 - ...
- Bisher nicht in OCR4all integriert
→ Open Source Tool [ScanTailor](#)

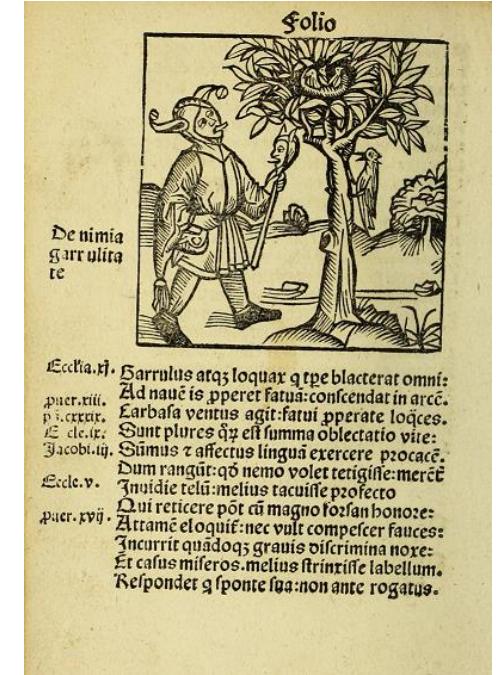
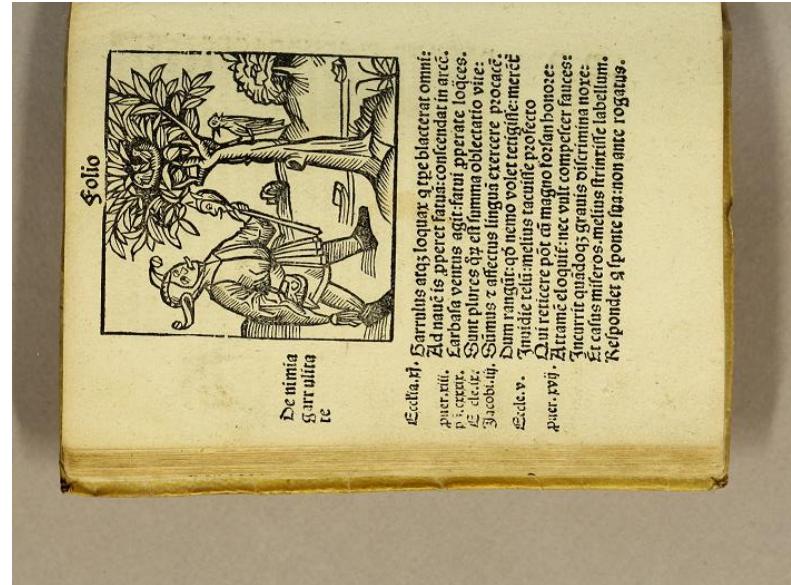
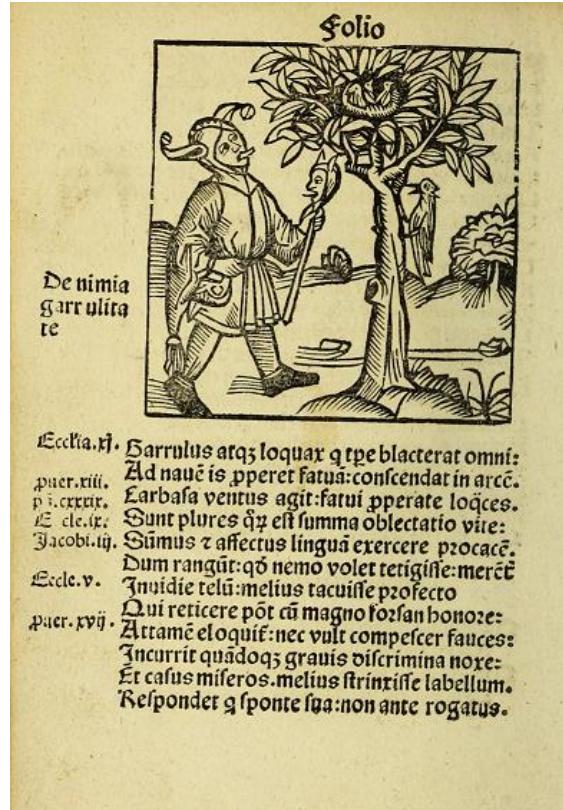


Image Preprocessing

- **Input:** Originalbilder (Farb- oder Graustufen- oder Binärbilder)
Output: geradegestellte Binärbilder
- Zwei Teilschritte, die die folgenden Arbeitsschritte erleichtern:
 - Binarisierung
 - Geradestellen
- Derzeit umgesetzt durch *ocropus-nlbin*



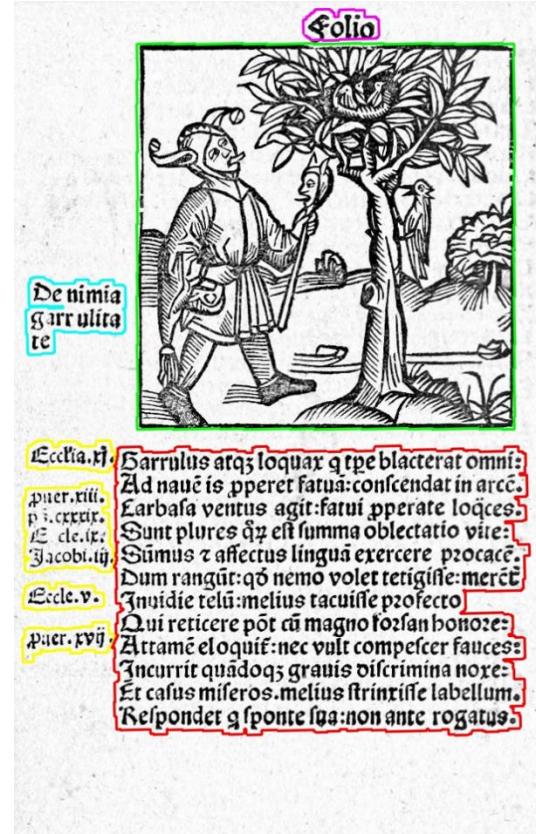
Ecclesia. xij. Harrulus atqz loquax q tpe blaterat omni:
puer. xiii. Ad nauē is pperet satua:conscendat in arcē.
p.i.cxxix. Larbasa ventus agit:fatuī pperate loqces.
E cle. ix. Sunt plures qz est summa oblectatio vite:
Jacobi. iij. Sūmus t affectus lingua exercere procacē.
Dum rangūt:qđ nemo volet terigisse:meret
Ecle. v. Invidie telū:melius tacuisse profecto
puer. xvij. Qui reticere pot cū magno forsan honore:
Attamē eloquit:nec vult compescer fauces:
Incurrit quādoqz grauis discrimina noxe:
Et casus miserōs.melius strinxisse labellum.
Respondet q sponte sga:non ante rogatus.



Ecclesia. xij. Harrulus atqz loquax q tpe blaterat omni:
puer. xiii. Ad nauē is pperet satua:conscendat in arcē.
p.i.cxxix. Larbasa ventus agit:fatuī pperate loqces.
E cle. ix. Sunt plures qz est summa oblectatio vite:
Jacobi. iij. Sūmus t affectus lingua exercere procacē.
Dum rangūt:qđ nemo volet terigisse:meret
Ecle. v. Invidie telū:melius tacuisse profecto
puer. xvij. Qui reticere pot cū magno forsan honore:
Attamē eloquit:nec vult compescer fauces:
Incurrit quādoqz grauis discrimina noxe:
Et casus miserōs.melius strinxisse labellum.
Respondet q sponte sga:non ante rogatus.

Region Segmentation

- **Input:** vorverarbeitete Bilder
- **Output:** Informationen über vorhandene Layoutelemente und deren Lesereihenfolge
- Tools/Methoden:
 - LAREX
 - Exakte Segmentierung inklusive semantischer Auszeichnung
 - Semiautomatisch, Intuitiv, adaptierbar und nachvollziehbar
 - Dummy Segmentation
 - Ganze Seite als ein Textsegment, Rest erledigt Zeilensegmentierung
 - Vollautomatisch und sehr schnell
 - Oft völlig ausreichend für moderate Layouts (z. B. typische Fraktur Romane aus dem 19. Jh.)

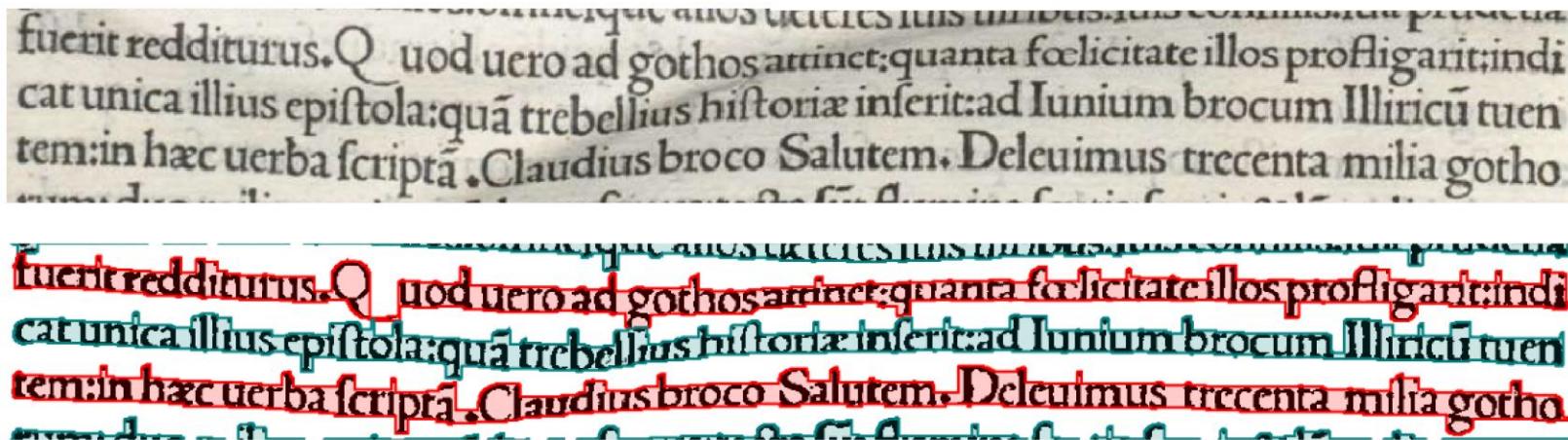


9

ten mit den Circassern und Arabern, die ich in russischen und französischen Diensten mitmachte. Endlich verflug mich das Schicksal nach Rom, wo mir meine letzten Mittel ausgingen. Ein Streifzug hatte mich unfähig gemacht, weitere große Märkte zu machen, obwohl ich noch fähig zum Garnisonsdienst war. Durch Vermittelung des französischen Kommandos erhielt ich einen Dienst in der päpstlichen Hofschiere, einen Dienst, der ruhig, gefährlos und dabei doch einträglich genug war, um ein höchst beschauliches und sorgenfreies Leben zu führen. Da standen wir in den bunten Drähten der alten Schweizer mit den Hellebarden in der Hand in den Kolonaden des Battians, zuweilen auch oben im Quirinal, und bewachten den Herrscher der Christenheit. Konnte das Schicksal mir wohl einen höhnischeren Streich spielen, als mich, den Glaubenslosen, den Spötter und modernen Atheisten zum Wächter der Päpste zu machen, zum Genossen von allerlei Abenteuern und Verlorenen, die der Wirbelwind des Schicksals aus aller Herzen Länder hier zusammengetrieben hatte? Gleichwohl wäre ich ganz glücklich gewesen, aber das Andenken an Wanda quälte mich Tage und Nächte. Unter den Marmorgestalten des Battians glaubte ich Wanda zu sehen, in den heiteren Tänzen, die das römische Volk an schönen Sonntagen in dem Giardino del Popolo am Coliseo aufführte, tauchte mir Wanda's Gesicht heraus, bei den pomphaften Kirchenfesten in Sanct Peter, unter den schat-

Line Segmentation

- **Input:** vorverarbeitete Bilder und Segmentierungsinformationen
Output: Zeilenpolygone
- Wichtiger Vorbereitungsschritt der eigentlichen Texterkennung
- Vorab Deskewing einzelner Regionen
- Zeilensegmentierung auch bei problematischer Ausgangslage möglich
- Derzeit umgesetzt durch angepasstes *ocropus-gpageseg*

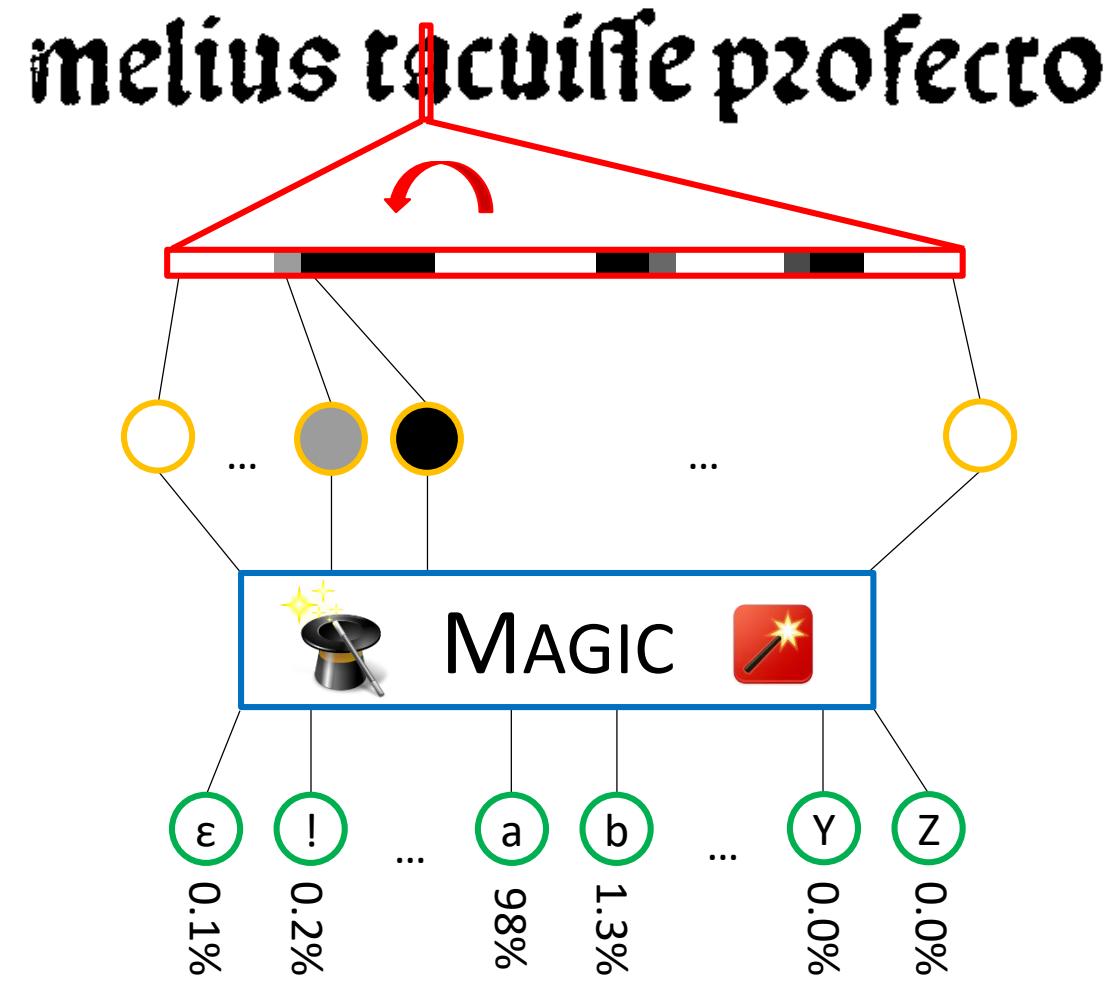


Character Recognition

- Input: Textzeilenbilder und OCR-Modelle
- Output: erkannte Textzeilen
- Derzeit umgesetzt durch *calamari-predict*

Garrulus atq; loquax q tpe blacterat omni:
Ad nauē is pperet fatuā:conscendar in arcē.
Earbasa ventus agit:fatui pperate loqces.
Sunt plures qu est summa oblectatio vite:
Sūmus z affectus lingua exercere procacē.
Dum rangūt:qd nemo volet tetigisse:merct
Jnuidie telū:melius tacuisse profecto

Earrulus atq; loquax q Epe blacterat omni:
Ad nauē is pperet fatuā:conscenda in arcē
Sunt plures qu est summa oblectatio vite:
Sūmus z affectus lingusta exercere procacē.
Dum rangūt:qd nemo volet tetigisse:merct
Jnuidie telū:melius tacuisse profecto



Post Correction – Ground Truth Production

- **Input:** Textzeilenbilder und Erkennungsergebnisse
Output: korrigierter Text (= Ground Truth, GT)
- GT wird für das Training von OCR-Modellen benötigt
- Anpassbares Virtuelles Keyboard ermöglicht die Verwendung spezifischer Sonderzeichen

Von dem Cirurgicus

fon dem iirurgicus

ix

zc

Von dem Cirurgicus

Von dem Cirurgicus

ix

IX

Post Correction – LAREX

Fueillet

sy font desqueselles ie me fais. O poures folz belistries qui de robes censy qui n'ot
pas du pain a mäger & aladventure quilz nosent demander de honte/les vieilles
gens/pourees defues/sabres/aveugles/helas pensez y/cat certes vous en ren-
dres compte devant celluy qui nous crea

Des conditiōs courronx et grandes mauuaisties des fēmes

Plusieurs asnes chevaucheroient
L'enest que monte y eust femme
Au moyeh de quoy ne vouldroient
Monte dessus po euse dissame

Quilz ont fait & grief infame
Au pouure asne & grans formēs
Car luy ont tozs tous ses bons mēs
Car luy ont tozs touo ses bons mēs

Ntendes folz e/
stourdis & vous
congnoistres les
mauaisties des
fēmes. Aussy fēmes appio
chez vo'z vo'z ortes hōne ma
tiere. Mes versetz ditez & es
criptz voulzroiet des fēmes

elij.di.quie
scam' i glos.

Si colligo
& vindicta
nemo magi
gaudet q̄ se
ra

puer.uti
Eccl. rev.
puerbi.uti.

Quilz ont fait & grief infame

Ouilz ont fait & grief infame

Au pouure asne & grans formēs

Au pouure asne & grans tozmēs

Car luy ont tozs tous ses bons mēs

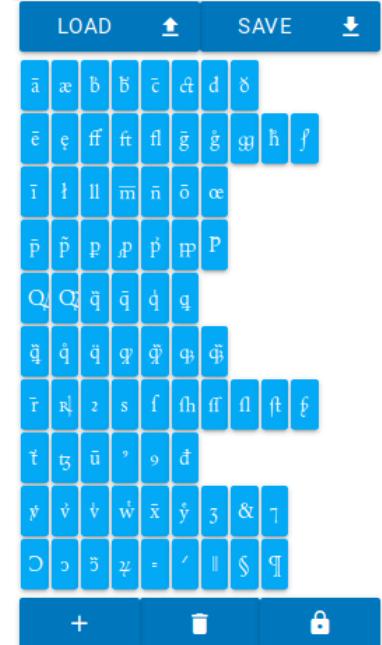
Car luy ont tozs touo ses bons mēs

H

C

Ntendes folz e/

Ntendes folz e=



Evaluation

- **Input:** OCR-Ergebnisse und Ground Truth für bestimmte Textzeilen
- **Output:** Zeichenfehlerrate (CER) und -statistiken
- Zeigt die Art und Häufigkeit der auftretenden Fehler an
- Auf Grundlage der Evaluation lässt sich erkennen, ob Fehler (noch) systematisch sind oder nicht
- Derzeit umgesetzt durch *calamari-eval*

GT	PRED	COUNT	PERCENT
{z}	{r}	33	17.65%
{}	{ }	11	5.88%
{i}	{i}	11	5.88%
{ff}	{f}	7	3.74%
{ff}	{ff}	5	5.35%
{t}	{i}	3	1.60%
{q}	{}	2	1.07%
{}	{}	2	1.07%
{ }	{}	2	1.07%
{x}	{}	2	1.07%

Training

Umfangreiche Unterstützung ...

- ... der zahlreichen Möglichkeiten des Calamari Trainingsprozesses
 - Ensemble-Training
 - Pretraining/Finetuning
 - Datenaugmentierung
 - ...
- ... des iterativen Trainingsansatzes
 - Ständiges Durchlaufen der Schritte Erkennung, Korrektur und Training zur Effizienzsteigerung
 - Modellverwaltung

Für Details siehe Veranstaltung zum Thema „Training“ und ggf. Live-Demo.

Live Demo!



Gliederung

1. Einleitung
2. Submodule
3. Workflow
4. Live Demo
- 5. Evaluation**
6. Diskussion und Ausblick

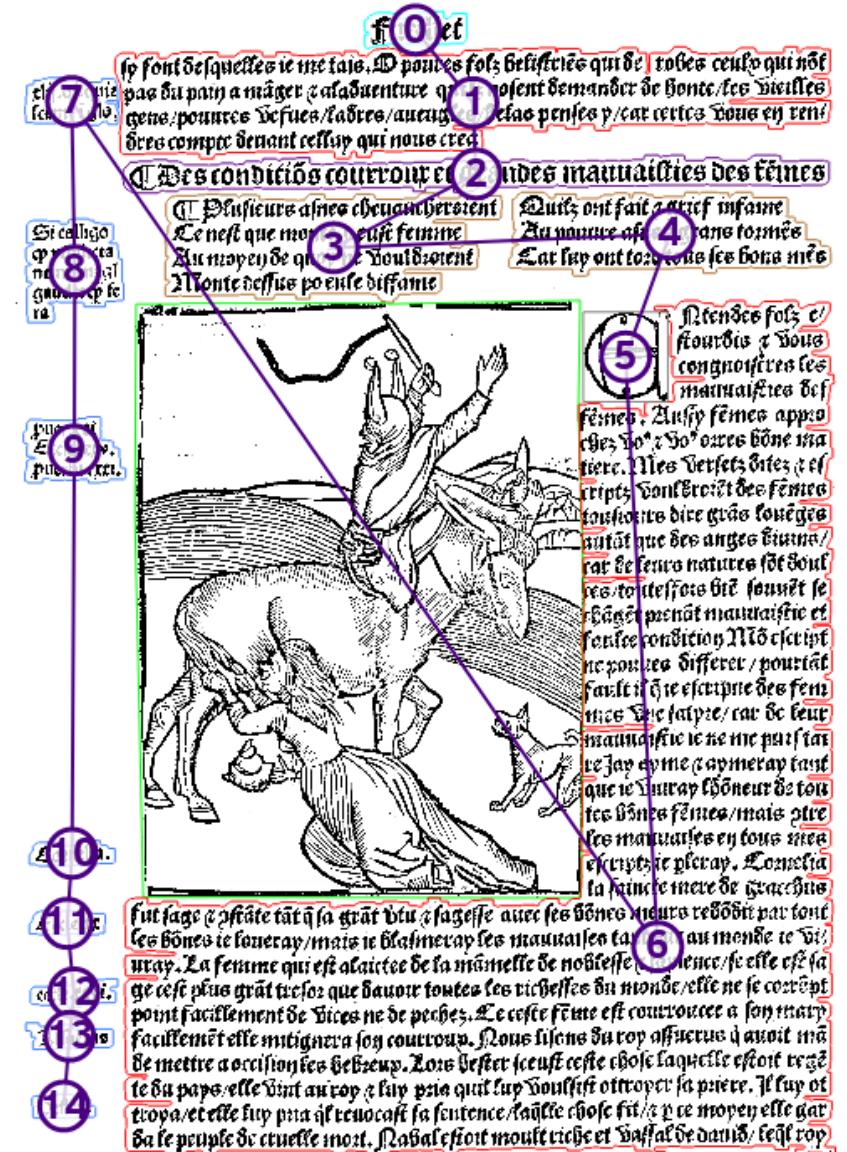
Evaluation – Frühdrucke bis 1600

■ Material

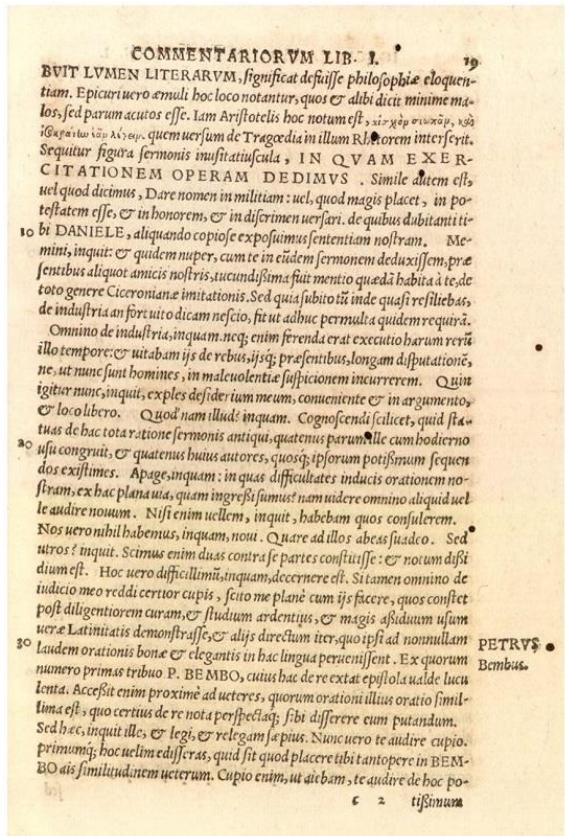
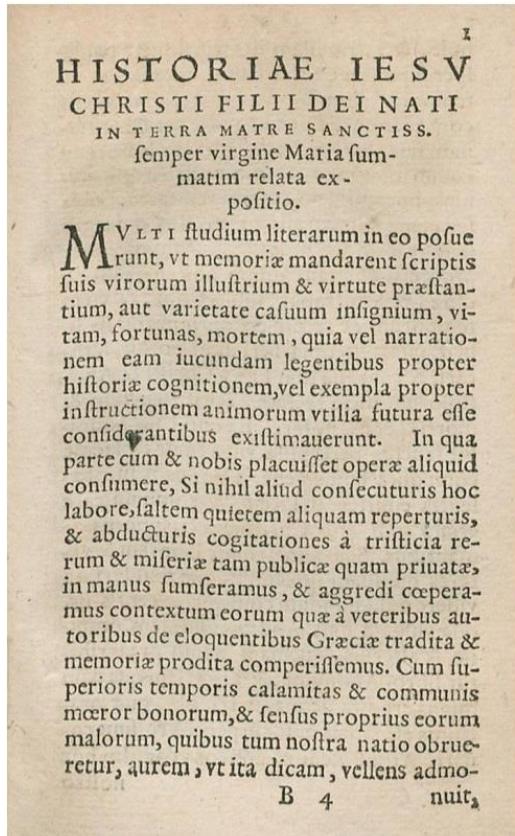
- 5 Ausgaben des „Narrenschiffs“
- 17 Werke des Universalgelehrten Joachim Camerarius
- 3 gemischte, früh-neuzeitliche Drucke (Praktikum)

■ Anforderungen / Ziele:

- Fehlerfreie Segmentierung und Reading Order sowie präzise semantische Auszeichnung
- Final: zitierfähiger Volltext (0% Fehler)
- Vorerst: 1% CER oder besser
- Bearbeitung in zwei Gruppen:
 - Unerfahrene, nicht-technische Nutzer
 - Erfahrene Nutzer



Frühdrucke bis 1600 – Beispiele I



109981/3

Vita salusq; homini est, is tota mente capeffat,
Cogiter, ediscat, meditetur, corde voluet
Includatq; pio, hac se consoletur, eadem
Pectus ad hostiles armans communiat ictus.
Qv o d patris e gremio celo descenderit alto
Filius eterni, simul ipse eternus, & intra
Mortalem celeste genus concluserit ortum,
Quen non ista caput totius machina mundi:
Factus homo in terris, Rerum non indigus ille
Nostrarum, sed cura fuit reparare salutem
Amissam & vitam nobis, ab origine prima
Quos peccatorum duro sub pondere pressos
Obruit ira Dei iusto commota furore,
Hec causa in terras celo hunc detraxit, vt effet
Saluator misericordum hominum, Deus uicus &
Rex,
Σατην, λυτεωτης, Σανατηφθορος ο επαν-
ροικης.

SEQVVNTVR ALII
quidam religiosi argumenti
versus compositi a Io-
achimo Came-
ratio.

K 4

PRE-

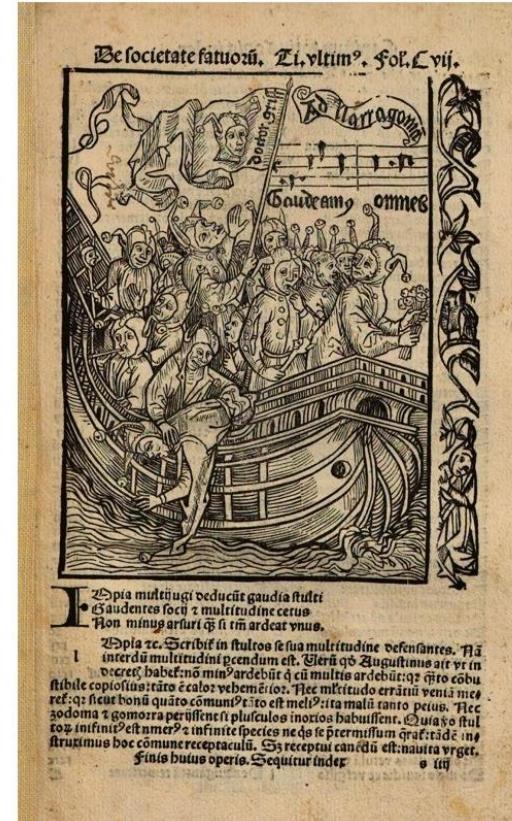
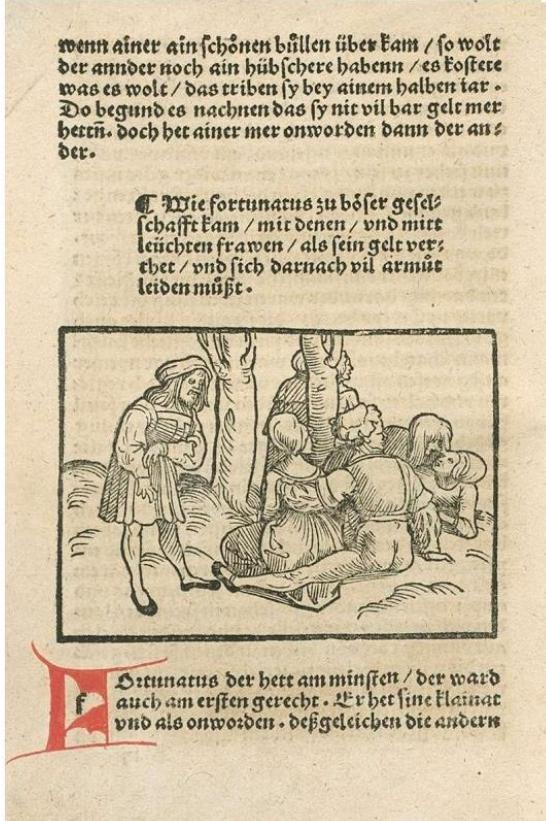
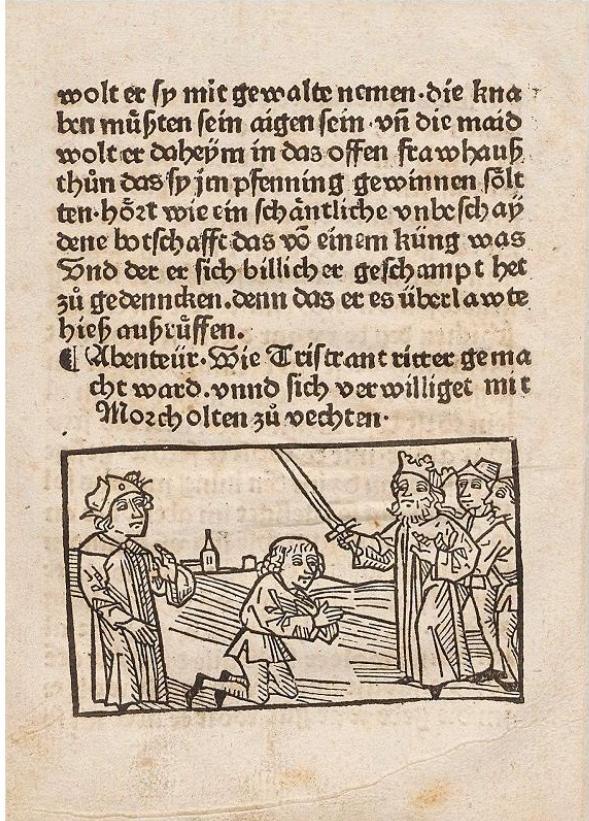
Vita salusq; homini est, is tota mente capeffat,
Cogiter, ediscat, meditetur, corde voluet
Includatq; pio, hac se consoletur, eadem
Pectus ad hostiles armans communiat ictus.
Qv o d patris e gremio celo descenderit alto
Filius eterni, simul ipse eternus, & intra
Mortalem celeste genus concluserit ortum,
Quen non ista caput totius machina mundi:
Factus homo in terris, Rerum non indigus ille
Nostrarum, sed cura fuit reparare salutem
Amissam & vitam nobis, ab origine prima
Quos peccatorum duro sub pondere pressos
Obruit ira Dei iusto commota furore,
Hec causa in terras celo hunc detraxit, vt effet
Saluator misericordum hominum, Deus uicus &
Rex,
Σατην, λυτεωτης, Σανατηφθορος ο επαν-
ροικης.

SEQVVNTVR ALII
quidam religiosi argumenti
versus compositi a Io-
achimo Came-
ratio.

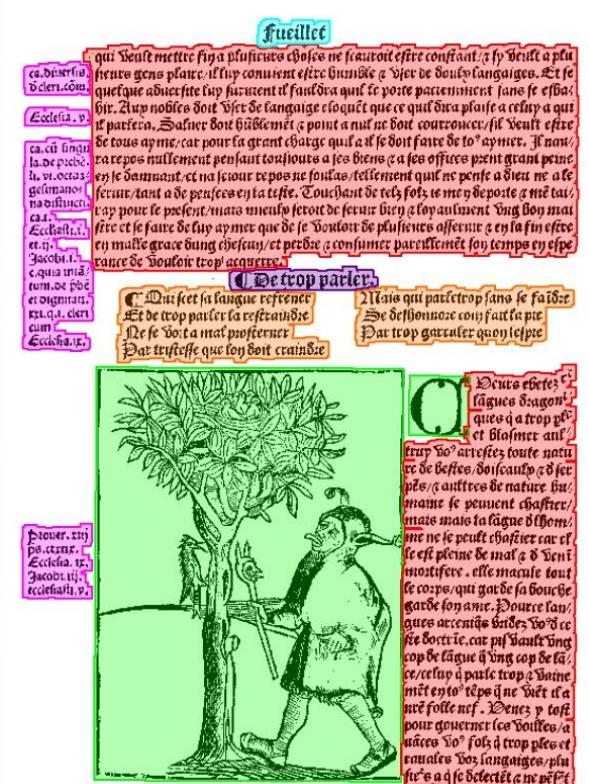
K 4

PRE-

Frühdrucke bis 1600 – Beispiele II



Frühdrucke bis 1600 – Beispiele III



Frühdrucke bis 1600 – Ergebnisse

	Unerfahrene Nutzer	Erfahrene Nutzer
Erreichte CER	$0,47\% \pm 0,22\%$	$0,49\% \pm 0,30\%$
Transkribiertes Trainingsmaterial	988 Zeilen	927 Zeilen
Korrekturzeit pro Zeile	$10s \pm 5,2s$	$5,5s \pm 2,4s$
Segmentierungszeit pro Seite	$1,1min \pm 0,5min$	$0,6min \pm 0,2min$

*Jeweils Durchschnittswerte, ggf. mit Standardabweichung

Beispielergebnis mit ca. 0,6% CER

bas keuschlichen leben vnd wil vnserm herren
dienen. Da was der keyser gar fro dʒ sein syn
verkert was. vnd seiner tochter nit wolt. vñ dy
andern zwu iunckfrawen blyben bei Constan
tia vnd dienten vnserm herren. Darnach wolt
Ballitanusnymmer Hertzog sein vnnd ward
demutig durch got vnnd gab sein gut den ar
men durch gotzwillen. vnd folget cristo nach.
vnd zwug den armen ir fuß. vñ goß in wasser
auff ir hend ee er in zu essen gab. vnd machet
vil siecher menschē gesunt. vñ vertrieb auch dy
bößen veindt mit seinem gesicht. wann wenn er
ein behafftes mensch anfah so muſt ð feindt
außfaren. vnd da der keyser gestarb da ward
Gallus keyser. der was der iunckfrawen Con
stantia geborner freundt. der het einen bößen

bas keuſchlichen Icben vnd wil vnſerm herzen
dienen. Da was der keyſer gar fro dʒ ſein ſyn
verkert was. vnd ſeiner tochter nit wolt. vñ dy
andern zwu iunckfrawen blyben bei Conſtan
tia vnd dienten vnſerm herzen. Darnach wolt
Ballitanusnymmer Hertzog ſein vnnd ward
demutig durch got vnnd gab ſein gut den ar
men durch gotzwillen. vnd folget crifo nach.
vnd zwug den armen ir fuß. vñ goß in waffer
auff ir hend ee er in zu essen gab. vnd machet
vil ſiecher menschē gesunt. vñ vertrieb auch dy
bößen veindt mit ſeinem gefiht. wann wenn er
ein behafftes mensch anfah ſo muſt ð feindt
außtaren. vnd da der keyſer geſtarb da ward
tGallus keyſer. der was der iunckfrawen Con
ſtantia geborner freundt. der het einen bößen

Gliederung

1. Einleitung
2. Submodule
3. Workflow
4. Live Demo
5. Evaluation
- 6. Diskussion und Ausblick**

Zusammenfassung

- OCR4all vielseitig einsetzbar
 - Erfolgreiche Verarbeitung von (historischen) Drucken des 15. bis 21. Jahrhunderts
 - Hauptanwendung: lokale Installation beim Nutzer
 - Nutzung als Serveranwendung prinzipiell bereits jetzt möglich
- Ergebnisse und dafür notwendiger Aufwand stark abhängig ...
 - ... vom Material
 - Automatische Segmentierung möglich?
 - Passendes gemischtes Modell vorhanden?
 - ...
 - ... von den Nutzeranforderungen
 - Grad der semantischen Auszeichnung?
 - Ansprüche hinsichtlich Genauigkeit?
 - ...

- Kooperationsvereinbarung Sommer 2020
 - Umsetzung von OCR-D Spezifikationen und Schnittstellen in OCR4all zum beiderseitigen Vorteil:
 - für OCR-D: bei Bedarf vereinfachter Zugang für die Nutzer, größere Reichweite
 - für OCR4all: erweiterte Auswahl an Werkzeugen, Flexibilität
 - Fortlaufender Austausch (Schnittstellen, Skalierbarkeit, kommende OCR Entwicklungen, GT, ...)
- Erfolgreicher Projektantrag *OCR4all-libraries* in dritter OCR-D Phase:
 - Kooperation zwischen GEI Braunschweig und Uni Würzburg (HCI, ZPD)
 - Volle Unterstützung der OCR-D Lösungen sowie deren Steuerung und Konfiguration über die GUI (Schnittstellen, Workflows, Settings, ...)
 - Ermöglichung einer Massenverarbeitung von Werken bzw. Werkclustern
 - Ausbau von LAREX als visuelle Erklärungskomponente (Fehleranalyse, Vergleich von Workflows, ...)
 - Optimierung der Usability
 - ...

Verarbeitung von Handschriften

- Erfassung von Drucken und Handschriften konzeptionell sehr ähnlich
- OCR4all bereits jetzt vielseitig im Handschrifteneinsatz
- Systematische Evaluation in Kooperation mit Dr. Stefan Tomasek (Lehrstuhl für Ältere Deutsche Literatur, Uni Würzburg)
 - Projekt Konrad von Fußesbrunnen: Kindheit Jesu
 - Material...
- Größtes Desiderat: robuste automatische Segmentierung
- Zeitnahe Anbindung vielversprechender, Baseline-basierter, trainierbarer (!) Lösungen:
 - [Kraken \(Kiessling: A Modular Region and Text Line Layout Analysis System\)](#)
 - Entwicklungen am Würzburger Lehrstuhl für Künstliche Intelligenz (Fischer, Hartelt, Gehrke, Puppe)
 - ...

Aktuelle Entwicklungen und Planungen

- Zeitnah: weitreichend überarbeitete Version mit zahlreichen Verbesserungen hinsichtlich Stabilität und Flexibilität
- Mittelfristig: größere Flexibilität hinsichtlich Material und Anwendungsszenarien
 - Handschriften
 - Massenvolltextdigitalisierung
- Weiterer DFG Antrag in Vorbereitung, u. a. Fokus auf
 - Optimierung des kollaborativen Arbeitens (Projekt-, Nutzer- und Taskverwaltung, Ressourcenmanagement, Backup und Versionierung, ...)
 - Bei Bedarf Steuerung einzelner Prozesse und Workflows aus LAREX heraus
 - Konfidenzbasierte Qualitätsanalyse und interaktive Nachkorrektur
 - Usability (Nutzerstudien, Überarbeitung des Schulungskonzepts und der Anleitungen, ...)

Interesse an OCR4all?

- Probieren Sie OCR4all auf Ihren eigenem Rechner mit Ihrem Material aus
 - www.ocr4all.de
 - https://github.com/OCR4all/getting_started
- Wir helfen gerne bei Fragen und Problemen: ocr4all@uni-wuerzburg.de
 - Installation und Nutzung
 - Bug Fixes und Feature Requests
 - Projekt-spezifisches Consulting
 - Server Support, falls die eigene Hardware nicht ausreicht
 - ...
- Wir sind auf Ihr (unverblümtes) Feedback angewiesen!
- ... aber bedenken Sie bitte immer: OCR4all ist umsonst und Work in Progress ☺

Vielen Dank für Ihre Aufmerksamkeit!

Acknowledgements:

- **OCR4all Web App:** Dr. Herbert Baier-Saip, Maximilian Nöth, Dennis Christ, Alexander Hartelt, Nico Balbach, Kevin Chadbourne
- **LAREX Web App:** Maximilian Nöth, Kevin Chadbourne, Nico Balbach
- **Calamari:** Dr. Christoph Wick, Andreas Büttner
- **Tests, Anleitungen, Nutzer Support und Artwork:** Maximilian Wehner, Raphaelle Jung, ...
- **Distribution via Docker und VirtualBox:** Björn Eyeslein, Yannik Herbst
- **Ideen und Feedback:** Dr. Uwe Springmann, Maximilian Wehner, Prof. Dr. Frank Puppe, Christine Grundig, Prof. Dr. Brigitte Burrichter, Prof. Dr. Joachim Hamm, ...
- **Förderung:** Lehrstuhl für Künstliche Intelligenz (Prof. Dr. Frank Puppe) und Zentrum für Philologie und Digitalität der Uni Würzburg, BMBF Projekt „Kallimachos“