SE Historische Korpuslinguistik: Digitalisierung und Optical Character Recognition

Uwe Springmann

Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin

2017-04-26

Einführung

Begriffe: Digitalisierung - OCR - Volltextdigitalisierung

- Digitalisierung:
  - oft nur: Herstellung von Seitenbildern von Dokumenten durch Fotografie, Scanning
  - Text wird nicht als solcher erkannt, Bild ist nicht durchsuchbar
- Optical Character Recognition (OCR):
  - automatisches Verfahren zur Erkennung von Text in Seitenbildern
  - erkannter Text kann als unsichtbare Ebene hinter die Seitenbilder gelegt werden (Bild wird dann durchsuchbar)
- Digitalisierung + OCR wird manchmal *Volltextdigitalisierung* genannt
  - Achtung: auch dieser Begriff ist mehrdeutig; manchmal ist damit lediglich der Scan sämtlicher Seiten eines Druckwerks gemeint

Die Ausgangslage: 567 Jahre moderne Druckgeschichte

- moderne Druckgeschichte: seit Gutenberg (1450)
- Inkunabeln: 1450-1500, ca. 30.000 Titel, davon 70% auf Latein
- VD16-18: *Verzeichnisse der im deutschen Sprachraum erschienenen Drucke*

VD	Anzahl Titel	davon bild-digitalisiert
VD16	110.000	61.000
VD17	300.000	133.000
VD18	600.000	145.000

- 200 - 300 Millionen Seiten zu erfassen
- heute schon 70 - 100 Millionen Seiten gescannt

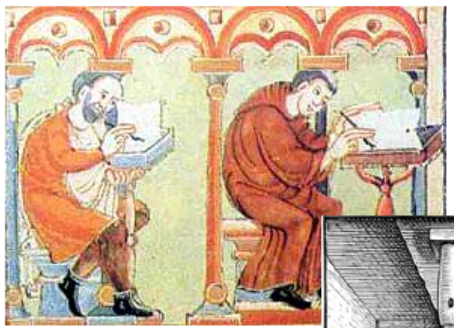
Der Blick aus 10km Höhe

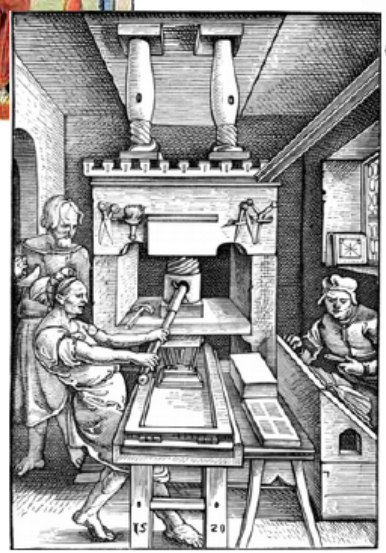
Die Forschungsfrage: Wie können wir die Druckproduktion seit Gutenberg in maschinenverarbeitbaren Text verwandeln?

- Motivation:
  - Erhöhung der Verfügbarkeit (überall, jederzeit)
  - Mehrwert gegenüber Papier: auf elektronischem Text kann man suchen (und finden)
- ungeheure Materialfülle nur mit maschinellen Hilfsmitteln zu bewältigen
- maschinenverarbeitbar (*machine-actionable*), nicht nur maschinenlesbar:
  - bezieht sich nicht nur auf Text, sondern auch auf Metadaten
  - z.B. Autor, Zeit, Ort, Genre, Strukturinformationen
  - maschinenverarbeitbare Form erlaubt Korrekturen, Textkritik, Annotationen

Textüberlieferung und Medienwechsel

Manuscript





Printing

PDF (Image)

PDF (searchable)

ABACISTA Vide Abacus.#

ABACIUM [gap: Greek word(s)] , Abacus, Fragm. Petronii: Abacia et cucumi omnia exposcit, etc.#

ABACOT pileus augustalis Regum Anglorum duabus coronis insignitus. Vide Chron. an. 1463. Edvv. IV. pag. 666. col. 2. lib. 27. Ita Spelman.#

Text

```
<pb id='s0004' n='4'/>
<p><term>ABACISTA</term>
<def>Vide <hi
rend='italic'><ref>Abacus<
/ref>.</hi>#</def></p>
<p><term>ABACIUM</term>
<def><gap desc='Greek
word(s)'
resp='sampling'/>, Abacus,
Fragm. Petronii: <hi
rend='italic'>Abacia et
cucumi omnia exposcit,
```

TEI

Kurze Geschichte der OCR

- 1913: Fournier d'Albe, Optophone (Lesehilfe für Blinde)
- 1929: Gustav Tauschek, Reading Machine
- 1931: Emanuel Goldberg, Statistical Machine (Retrieval von Metadaten auf Mikrofilmen)
- 1974: Ray Kurzweil, Reading Machine
- Hauptanwendungsgebiete heute: automatischer Dokumentenworkflow ("papierloses Büro")
- kommerzielle Anbieter: ABBYY (*Finereader*), Nuance (*Omnipage*), Canon (*ReadIris*)
- seit einigen Jahren gute Open-Source-Software verfügbar:
  - Tesseract (seit 2005; Ray Smith, früher HP Labs, jetzt Google)
  - OCRopus (seit 2009; Tom Breuel, früher DFKI Kaiserslautern, jetzt Nvidia)

OCR aktuell

OCR: Stand der Forschung

- OCR liegt im Schnittpunkt von Mustererkennung, künstlicher Intelligenz und Computer Vision (“hot topics”, denken Sie an selbstfahrende Autos)
- Forschung bisher hauptsächlich in kommerziellen Firmen (Hersteller von Scannern und Kopierern) und daher Firmengeheimnis
- allgemeine Ansicht: OCR ist ein gelöstes Problem! (gilt nur für Drucke ab dem 20. Jahrhundert: > 99% richtig erkannte Zeichen)
- OCR historischer Drucke:
  - Hindernisse aufgrund Typographie etc. (sh. oben)
  - vorhandene OCR Engines liefern schlechte Ergebnisse (unter 85% Zeichengenauigkeit
  - für Firmen uninteressant aufgrund geringen Ertragspotentials (30.000 Inkunabeltitel 1450-1500, weltweite Buchproduktion 2014: 1 Mio Titel)

Projekte und Forschungsgruppen zur OCR historischer Drucke

- EU IMPACT Projekt (2008-2012)
- CIS, LMU München: Klaus Schulz
  (Nachkorrektur von OCR-Resultaten, seit 2004)
- OCR-D Projekt der DFG: Koordinationsprojekt zur Weiterentwicklung von Verfahren zur OCR historischer Drucke, seit 2015)
- Early Modern OCR Project (EMOP): Laura Mandell
  (Texas A&M University, 2012-2015)
- DFKI Kaiserslautern, Andreas Dengel & Co.
  (Fortsetzung der Forschung von Tom Breuel)
- Kallimachos-Projekt: Hans-Günter Schmidt
  (Universität Würzburg, 2014-2017)
- Ocular: Dan Klein, Taylor Berg-Kirkpatrick
  (University of California, Berkeley, seit 2013)

OCR Grundlagen

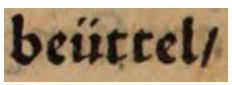
OCR Verarbeitungskette (“Workflow”)

die komplette OCR Verarbeitungskette besteht aus mehreren Schritten (Schritt 3 ist für historische Drucke entscheidend):

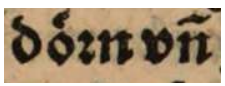
1. Bildbeschaffung
2. Vorverarbeitung
3. *Transkription* (“ground truth”) und Modelltraining
4. Zeichenerkennung (eigentliche OCR)
5. Evaluation
6. Nachbearbeitung: *Fehlerkorrektur*, *Annotation*, ...

Fehlerarten

- Elementarfehler (*elementary edit operations*):
  - Hinzufügungen (*insertions*)
  - Auslassungen (*deletions*)
  - Umwandlungen (*substitutions*)
- Beispiele:



(beüttel) → **beüt tel** (Hinzufügung eines Leerzeichens)



(d̊rn vñ) → **d̊rnvñ** (Auslassung eines Leerzeichens)

(erſcheinũg) → **erſcheinüng** (Substitution ũ → ü)

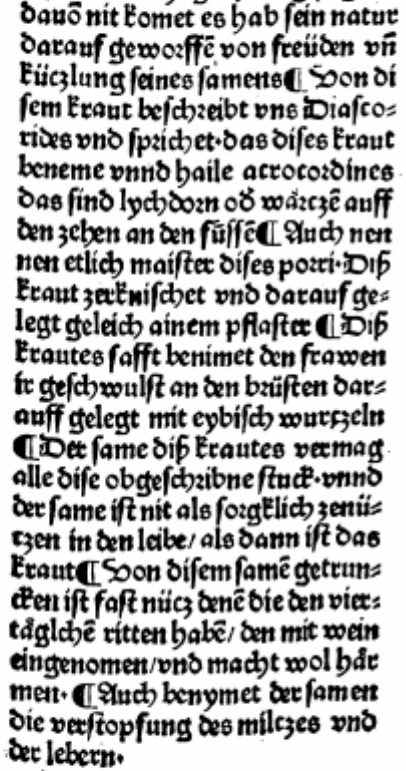
OCR-Kennzahlen: Fehlerrate und Genauigkeit

- OCR-Fehler: fehlerhaft erkannte Elemente (Zeichen, Wörter)
- *Fehlerrate*:
  Anzahl Fehler / Anzahl aller Elemente
  Anzahl Fehler = Summe der Elementarfehler
- *Genauigkeit (accuracy)*:
  Anzahl richtig erkannter Elemente / Anzahl aller Elemente
  = 1 - Fehlerrate
- Beispiel:
  - 2% Zeichenfehlerrate
  - entspricht 98% Zeichenerkennungsrate (*accuracy*)

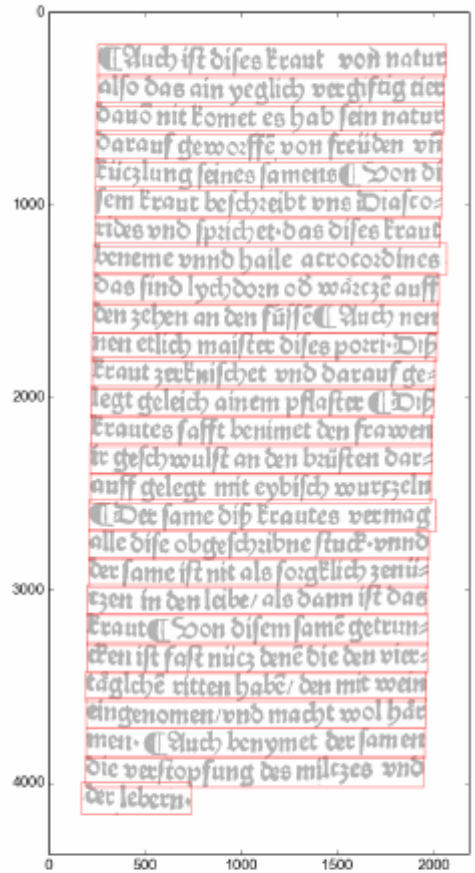
Beispiel Vorverarbeitung: Gart der Gesundheit

Johann Wonnecke von Kaub (J. von Cuba), Gart der Gesundheit (1485, hier 1487)

original image



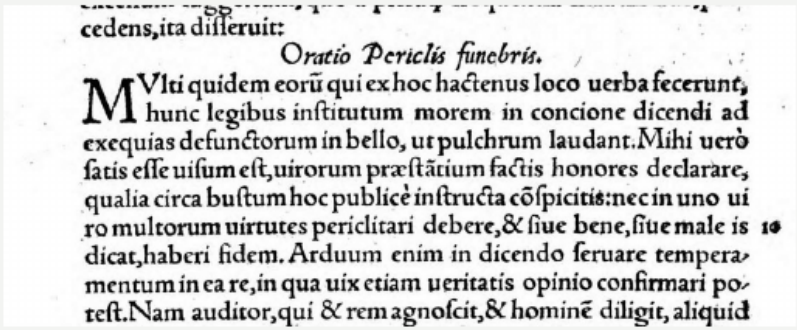
binarized text zone



line segmented

Hindernisse I: historische Schrifttypen

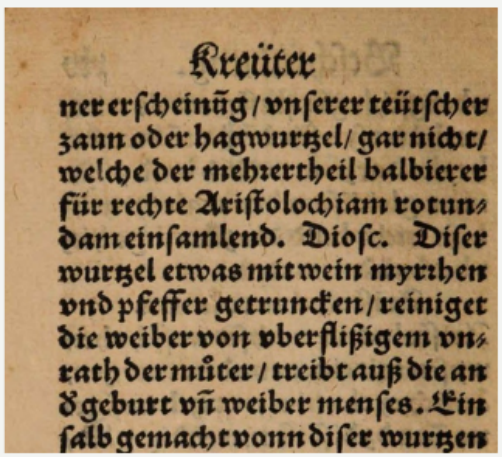
im Uhrzeigersinn: Valla (1564), Foresti (1487), Leyser (1735), Bodenstein (1557)



cedens,ita differuit:

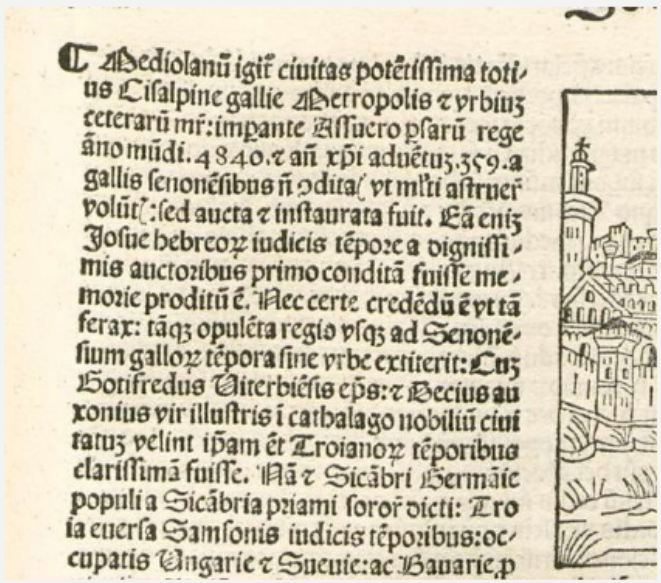
Oratio Periclis funebris.

MVlti quidem eorũ qui ex hoc hactenus loco uerba fecerunt, hunc legibus institutum morem in concione dicendi ad exequias defunctorum in bello, ut pulchrum laudant. Mihi uerò satis esse uisum est, uirorum præstãtium factis honores declarare, qualia circa bustum hoc publicè instructa cõspicitis: nec in uno uiro multorum uirtutes periclitari debere, & siue bene, siue male is dicat, haberi fidem. Arduum enim in dicendo seruare temperamentum in ea re, in qua uix etiam ueritatis opinio confirmari potest. Nam auditor, qui & rem agnoscit, & hominẽ diligit, aliquid

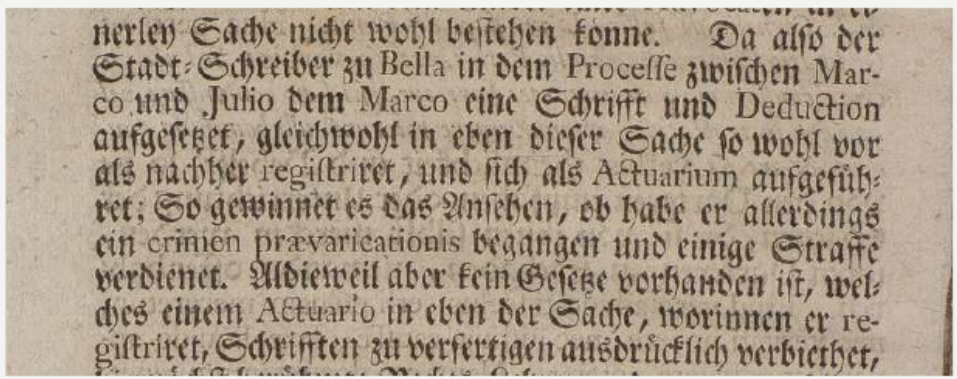


Kreüter

ner erscheinũg / vnserer teütscher zaun oder hagwurtzel / gar nicht / welche der mehrertheil balbierer für rechte Aristolochiam rotundam einsamlend. Dioſc. Diſer wurtzel etwas mit wein myrrhen vnd pfeffer getruncken / reiniget die weiber von vberflißigem vnrath der mũter / treibt auß die an ð geburt vñ weiber menſes. Ein ſalb gemacht vonn diſer wurtzen



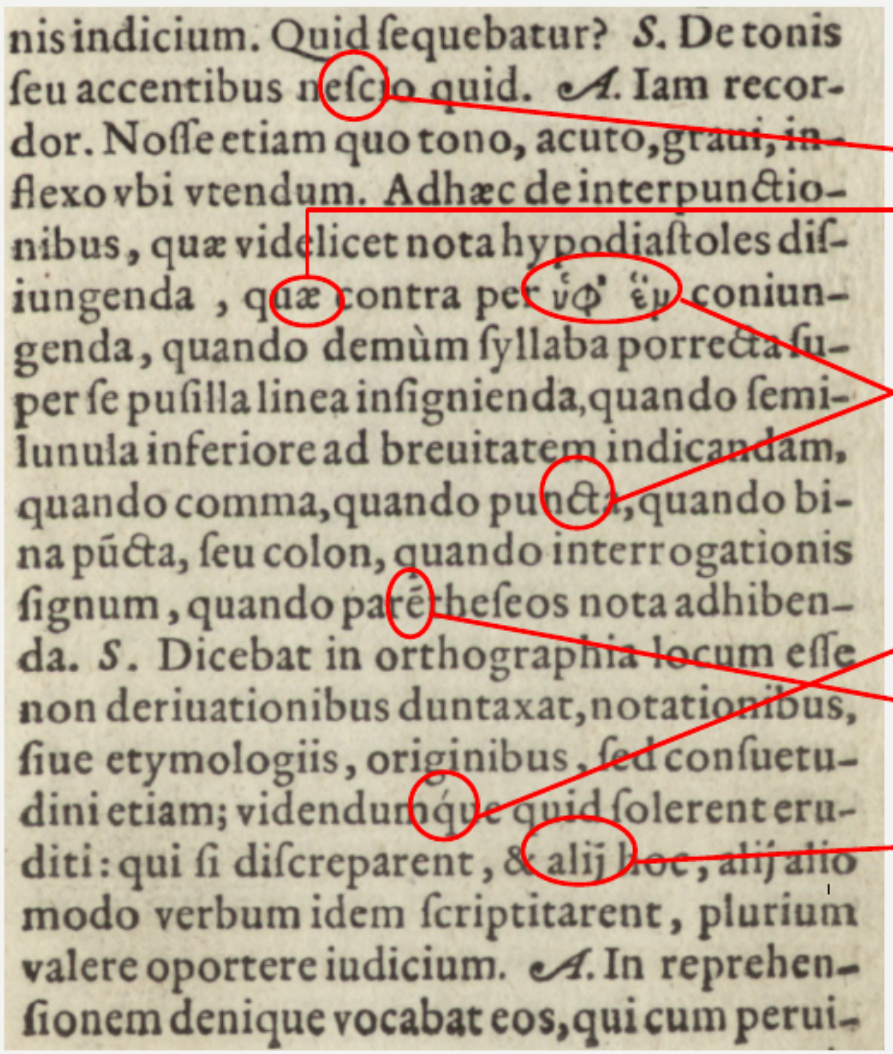
Mediolanũ igit̃ ciuitas potẽtiſſima totius Ciſalpine gallie Metropolis ꝛ vrbiũ ceterarũ mr̃: impante Aſſucro pſarũ rege ãno mũdi. 4840. ꝛ añ xp̃i aduẽtuz. 359. a gallis ſenonẽſibus ñ odita(vt mlti aſtruer volũt(: ſed aucta ꝛ inſtaurata fuit. Eã enĩz Joſue hebreoꝝ iudicis tẽpore a digniſſimis auctoribus primo conditã fuiſſe memorie proditũ ẽ. Nec certe credẽdũ ẽ vt tã ferax: tãq opulẽta regio vſq ad Senonẽſium gallorꝝ tẽpora ſine vrbe extiterit: Cuĩz Bonifredus Viterbiẽſis epc̃s: ꝛ Beciꝰ auxonius vir illuſtris i catbalago nobiliũ ciuitatuz velint ipam ẽt Troianoꝝ tẽporibus clariſſimã fuiſſe. Mã ꝛ Sicãbri Germãie populi a Sicãbria priami ſororꝰ dicti: Troia euerſa Samſonis iudicis tẽporibus: occupatis Vngarie ꝛ Sueuie: ac Bauarie, p



nerley Sache nicht wohl beſtehen könne. Da alſo der Stadt-Schreiber zu Bella in dem Proceſſe zwiſchen Marco und Julio dem Marco eine Schrifft und Deduction aufgeſetzet, gleichwohl in eben dieſer Sache ſo wohl vor als nachher regiſtriret, und ſich als Actuarium aufgeführet; So gewinnet es das Anſehen, ob habe er allerdings ein crimen prævaricationis begangen und einige Straffe verdienet. Alldieweil aber kein Geſetze vorhanden iſt, welches einem Actuario in eben der Sache, worinnen er regiſtriret, Schrifften zu verfertigen ausdrücklich verbiethet,

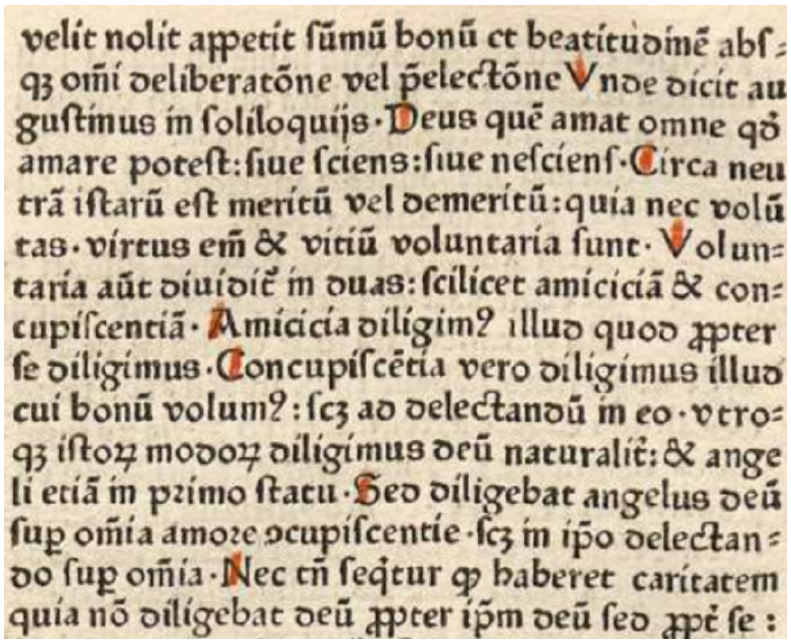
Hindernisse II: historische Schreibweisen, Zeichenrepertoire

Pontanus, Progymnasmata (1589)



OCR auf Inkunabeln: aussichtslos?

Beauvais: *Speculum naturale* (1476); ABBYY FR11 Fraktur 68% acc.



velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ=
qꝫ oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs·Deus quẽ amat omne qđ
amare poteſt:ſiue ſciens:ſiue neſcienſ·Circa neu
trã iſtarũ eſt meritũ vel demeritũ:quia nec volũ
tas·virtus eñ & vitiũ voluntaria ſunt·Volun=
taria aũt diuidit in duas:ſcilicet amicicią & con=
cupiſcentią·Amicicia diligim? illud quod ꝓpter
ſe diligimus·Concupiſcẽtia vero diligimus illud
cui bonũ volum?:ſcz ad delectandũ in eo·vtro=
qꝫ iſtoꝝ modoꝝ diligimus deũ naturaliť:& ange
li etiã in primo ſtatu·Sed diligebat angelus deũ
ſup oñia amore ocupiſcentie·ſcz in ipõ delectan=
do ſup oñia·Nec tñ ſeqtur qꝺ haberet caritatem
quia nõ diligebat deũ ꝓpter ipm deũ ſed ꝓpť ſe :

velie nolit aspenc sumu bonu ce beaticuome al?s-
qzonn veliberaeone velpelec^oneVnoevicicau
Aus^mus in soliloquijs'^eus que amat omne qv
amarc potest:s»uesciens.smenesciens Circa neu
era ls^aru esk mencu vet vcmenturquia ncc volu
ras vireus em viciu voluntana (une V^o!un-
tana auc oiuivic in ouas: scilicet amicicia Le con-
cupilcencia 5Vmicicra vilizim? illuo quov zpter
sevili^imus Concupiscena vcro viliqimus illuo
cui bonu volum?:1cz as velec^anvu in co Vtro-
qz is^or^ movoy oiliqimus veu naeuralitrLL an^e
li eria in prims l^acu Leo viliqebat an^clus veu
sup omia amorc ocupiscentie lcz in ipo vclec^an -
vo sup omia Alec cn seqeur q? kaberee caritacem
quia no viligvbat veu ^fpter ipm veu seo zx>c se :

Inkunabeln haben häufig besondere Abkürzungszeichen, z.B. ꝑ ꝓ p̈ qꝫ Qꝶ q̈ ſcz.

(Rydberg-Cox 2009) (unsere Hervorhebung): *"Because of the prevalence of these glyphs,* ***incunabula cannot be processed using OCR software.*** *Commercial OCR programs produce almost no recognizable character strings, let alone searchable text. …* ***Other methods*** *must be explored."*

OCR für historische Drucke

Andere (OCR-) Methoden: Rekurrente neuronale Netze

- Schlagwort:
  Rekurrente neuronale Netze mit langem Kurzzeitgedächtnis
  RNN mit LSTM, *Hochreiter and Schmidhuber (1997)*
- Methode hatte große Erfolge bei Mustererkennung
  (Verkehrszeichen, Gesichtserkennung, ...)
- auf OCR-Erkennung erstmals angewandt von *Breuel et al. (2013)*
- auf Erkennung von Frühdrucken adaptiert von *Springmann et al. (2014); Springmann, Lüdeling, and Schremmer (2015); Springmann (2015)*

Wie lernt das neuronale Netz?

Idee (Breuel):

- zerschneide Bild einer Textzeile in viele vertikale Streifen (500-1000)
- ordne den Streifen (Pixels) einer Zeile die diplomatische Transkription (Labels) der Zeile zu
- das Netzwerk gewichtet die Verbindungen seiner internen Speicherzellen (Gedächtnis) so, dass eine Verbindung von Inputdaten (Pixels) zu Outputdaten (Labels) entsteht
- Lernen geschieht selbsttätig (Klassifizieren von benachbarten Streifen zu kodierten Glyphen)
- nach einiger Zeit erkennt es vorher nicht gesehene Zeilen mit guter Genauigkeit

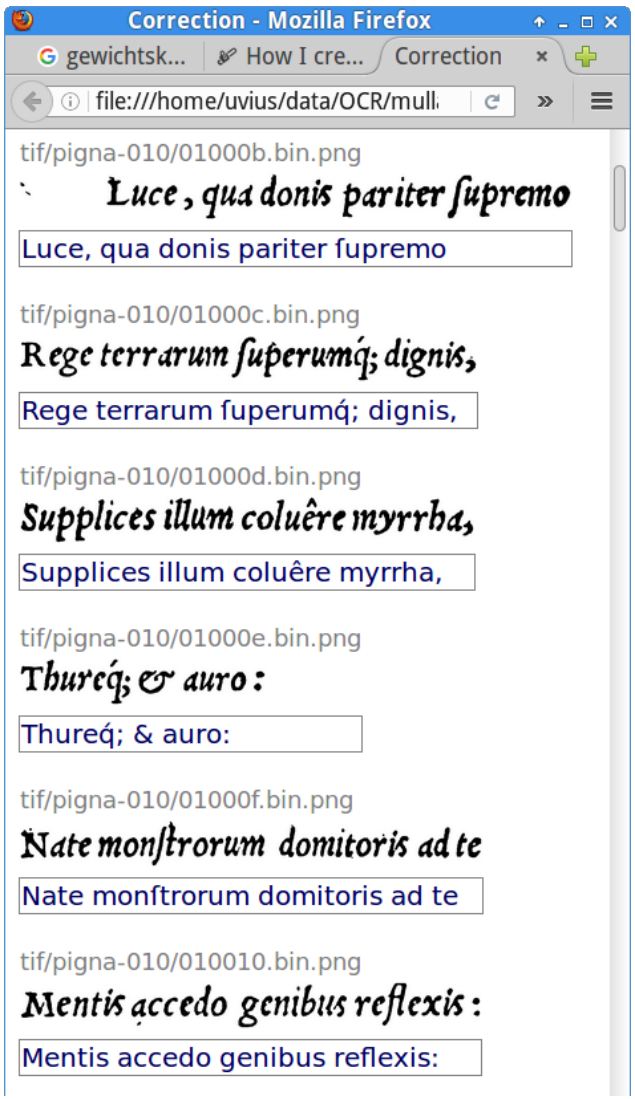
Die Zerlegung von Zeichen in einzelne Streifen als Grundeinheiten ist der Schlüssel für die bessere Erkennung gegenüber einer Mustererkennung auf Zeichenebene!

Trainieren eines OCR-Modells für ein Buch

Das Modelltraining gliedert sich in die folgenden Schritte:

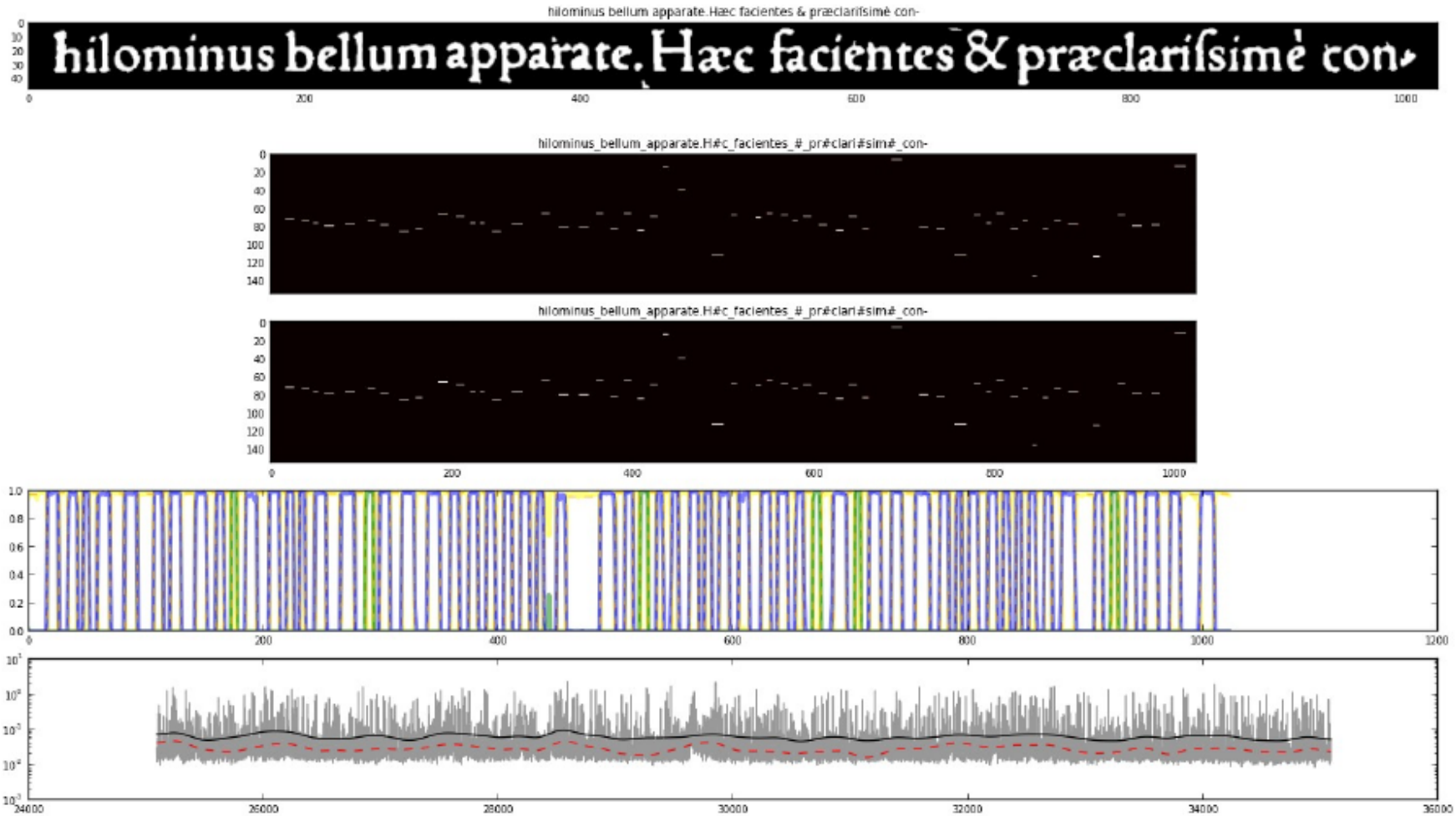
1. Beschaffen der Scans
2. Zerlegen der Seitenbilder in einzelne Zeilen
3. Herstellung einer diplomatischen Transkription (*ground truth*) dieser Zeilen
4. Aufteilen der Bild- und Textzeilen in eine Trainings- und eine Testmenge
5. Training auf der Trainingsmenge
6. Testen auf Testmenge:
   - Testergebnis ok: Erkennen des ganzen Dokumentes
   - Testergebnis zu schlecht:
     Korrekur der Erkennung einiger weiterer Seiten zu ground-truth-Qualität,
     Hinzufügen zur Trainingsmenge und Rücksprung auf Nr. 5

Transkription für OCR-Zwecke im Browser



- Eingabe über Zeilensynopse im Browser mit geeigneter Schriftart, z.B. Junicode
- Glyph-Repertoire bestimmen
- Paläographie-Kenntnisse notwendig
  - Ligaturen (z.B. ft, æ)
  - Suspensionen (z.B. domin⁹ = dominus)
  - Kontraktionen (z.B. oẽs = omnes, eps = episcopus)
- weitere Voraussetzungen:
  - historische Linguistik
  - Schreibvarianten
  - Latein (70% der Inkunabeldrucke)

Dem Netz beim Lernen zuschauen



Modelltraining

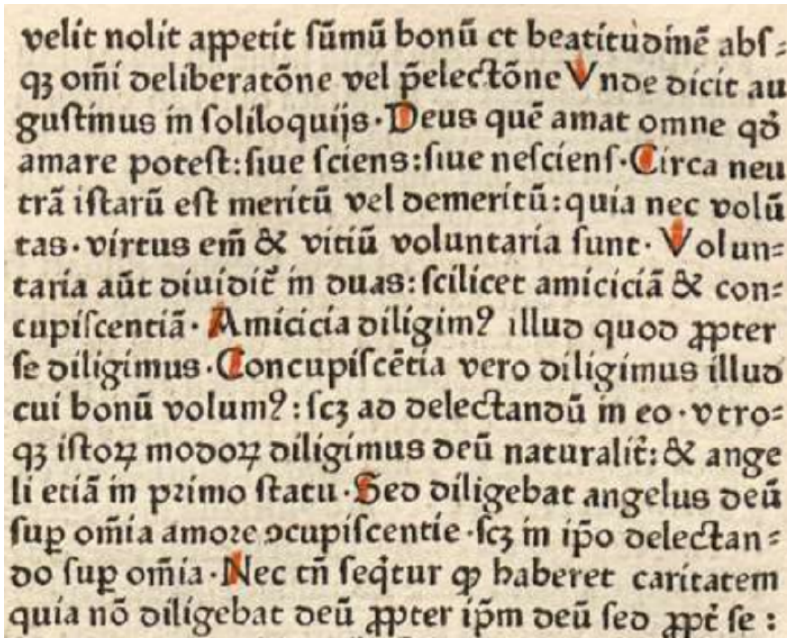
(Pigna: Carmina, 1553)

nach einer Weile (hier: nach 49.021 Lernschritten):

```
49021 15.09 (456, 48) train/pigna-010/010015.bin.png
   TRU: u'Spirtium \u017facro tibi \u017fempiternum'
   ALN: u'Spiritiium \u017facro tibi \u017fempiternum'
   OUT: u'Spiritum \u017facro tibi \u017fempiternum'
49022 9.94 (558, 48) train/pigna-010/01000b.bin.png
   TRU: u'Luce, qua donis pariter \u017fupremo'
   ALN: u'Luce, qua donis pariter \u017fupremo'
   OUT: u'Luce, qua donis pariter \u017fupremo'
49023 4.07 (267, 48) train/pigna-010/01000e.bin.png
   TRU: u'Thureq\u0301; & auro:'
   ALN: u'Thureq\u0301; & auro:'
   OUT: u'Thureq\u0301; & auro:'
```

Noch einmal Beauvais, *Speculum Naturale*

Trainiertes OCRopus-Modell (dieser Ausschnitt: 99% acc.)


velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ=
qȝ oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs · Deus quẽ amat omne qđ
amare poteſt: ſiue ſciens: ſiue neſcienſ · Circa neu
trã iſtarũ eſt meritũ vel demeritũ: quia nec volũ
tas · virtus eñ & vitiũ voluntaria ſunt · Volun=
taria aũt diuidit in duas: ſcilicet amicitiã & con=
cupiſcentiã · Amicicia diligim⁹ illud quod ꝓpter
ſe diligimus · Concupiſcẽtia vero diligimus illud
cui bonũ volum⁹ : ſcȝ ad delectandũ in eo · vtro=
qȝ iſto₄ modo₄ diligimus deũ naturaliť: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſuꝑ oñia amore ocupiſcentie · ſcȝ in iꝑo delectan=
do ſuꝑ oñia · Nec tñ ſeq̇tur q̱ haberet caritatem
quia nõ diligebat deũ ꝓpter iꝑm deũ ſed ꝓpť ſe :

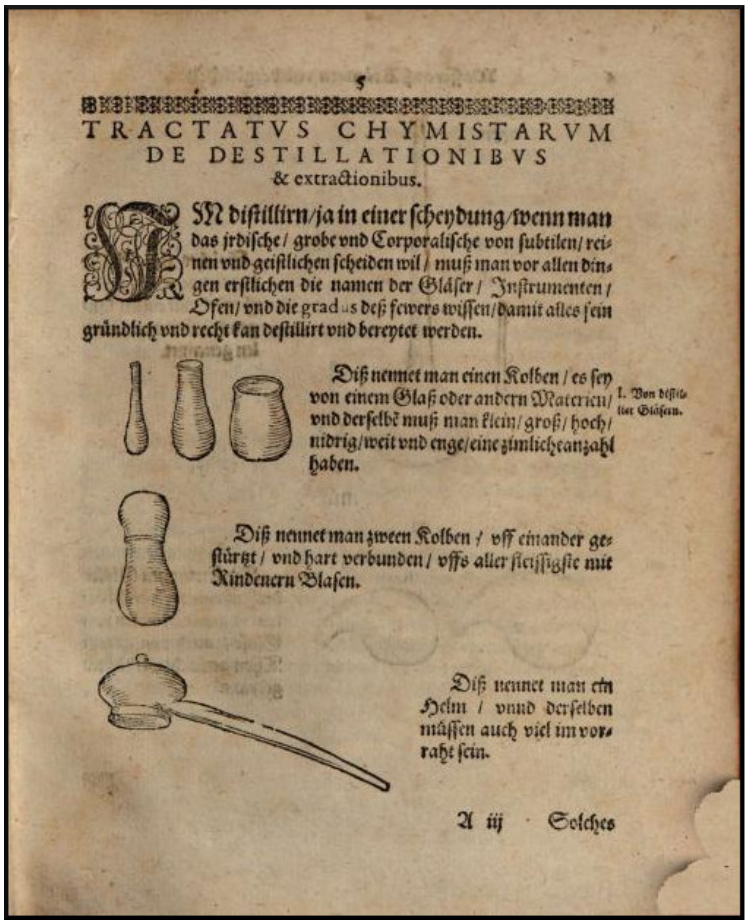
velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ ·
qȝ oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs · Deus quẽ amat omne qđ
amare poteſt:ſiue ſciens:ſiue neſcienſ· Circa neu
trã iſtarũ eſt meritũ vel demeritũ : quia nec volũ
tas · virtus eñ & vitiũ voluntaria ſunt · Volune
taria aũt diuidit in duas: ſcilicet amicitiã & con=
cupiſcentiã · Amicicia diligim⁹ illud quod ꝓpter
ſe diligimus · Concupiſcẽtia vero diligimus illud
cui bonũ volum⁹ : ſcȝ ad delectandũ in eo · vtro=
qȝ iſto₄ modo₄ diligimus deũ naturaliť: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſuꝑ oñia amore ocupiſcentie · ſcȝ in iꝑo delectan=
do ſuꝑ oñia · Nec tñ ſeq̇tur q̱ haberet caritatem
quia nõ diligebat deũ ꝓpter iꝑm deũ ſed ꝓpť ſe :

- nur noch 4 Fehler! (rechts: rot und blau markiert)
- trainiert auf 13 Seiten, getestet auf weiteren 4 Seiten
- 98% mittlere Zeichenerkennungsrate (rohes, unkorrigiertes OCR-Ergebnis)
- ohne Verwendung eines Sprachmodells

20 deutsche Frakturdrucke aus 4 Jahrhunderten (RIDGES)

RIDGES-Korpus: deutsche Kräutertexte vom 15. bis 20. Jahrhundert

links Libavius, Alchymistische Practic, 1603, rechts NN, Curioser Botanicus, 1675



5

TRACTATVS CHYMISTARVM
DE DESTILLATIONIBVS
& extractionibus.

M distillirn/ja in einer scheydung/wenn man das irdische/ grobe vnd Corporalische von subtilen/ reinen vnd geistlichen scheiden wil/ muß man vor allen dingen erstlichen die namen der Gläser/ Instrumenten/ Ofen/ vnd die gradus deß fewers wissen/ damit alles sein gründlich vnd recht kan destillirt vnd bereytet werden.

Diß nennet man einen Kolben/ es sey von einem Glaß oder andern Materien/ vnd derselbē muß man klein/ groß/ hoch/ nidrig/ weit vnd enge/ eine zimliche anzahl haben.

1. Von distillier Gläsern.

Diß nennet man zween Kolben/ vff einander gestürtzt/ vnd hart verbunden/ vffs aller fleissigste mit Rindenern Blasen.

Diß nennet man ein Helm/ vnnd derselben müssen auch viel im vorraht sein.

A iij Solches

[8]

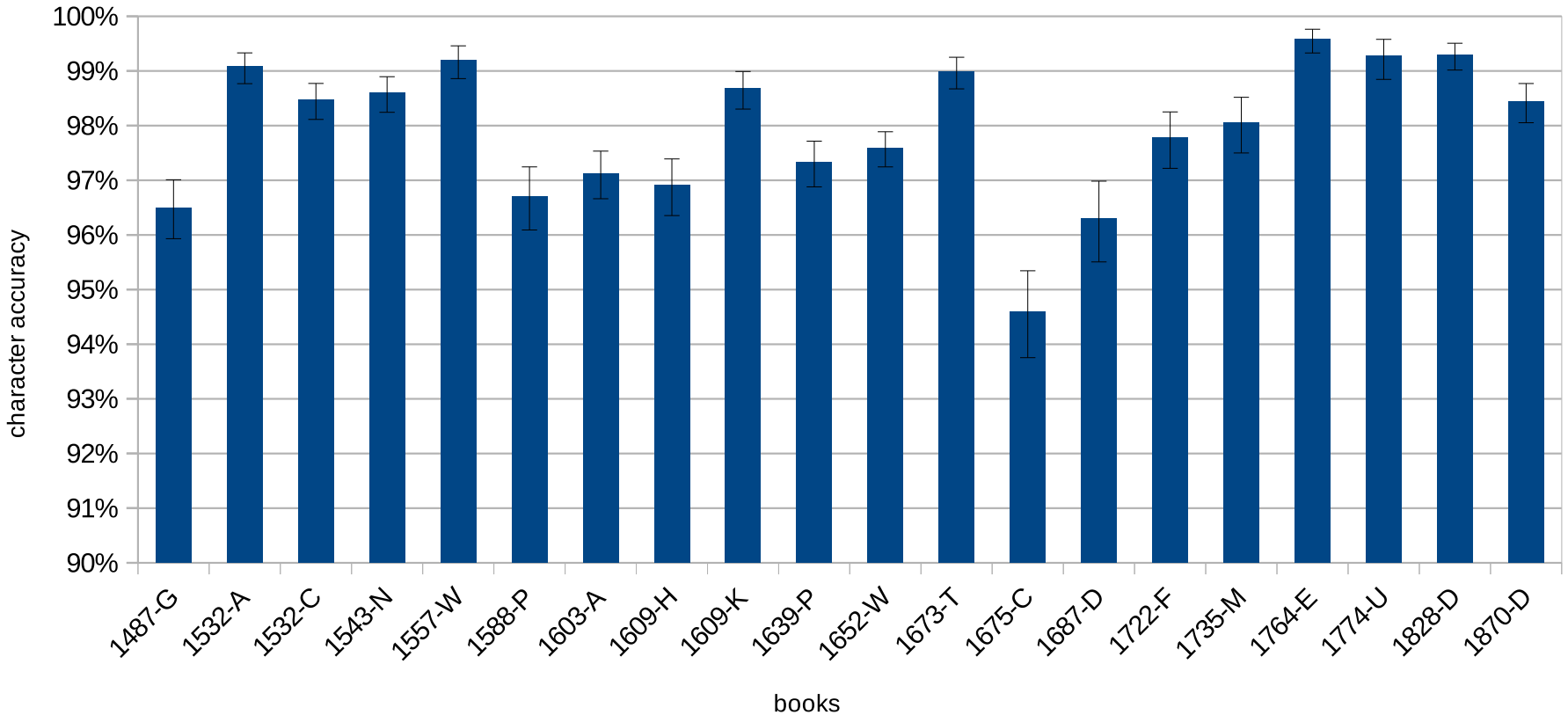
und hoher. Die Blätter gleichen dem Beyfuß/ sind gantz graufarbig/ zerspalten/ eines würtzhafften Geruchs und sehr bittern Geschmacks. Trägt statt der Blüte fast Trauben-weiß gelbe Knöpflein/ worinnen der Same enthalten. Die Wurtzel ist dicke/ hölzicht/ zäsericht/ doch nicht bitter/ sondern guten Geschmacks.

3. Ort. Wermuth wächst gerne an steinichten/ bergichten und ungebauten Orten/ dahero es hin und wieder/ öffters aber bey alten eingefallenen Baustellen gefunden wird.

4. Zeit. Wermuth blühet im Julio und Augusto. Im Junio aber ist er in seiner besten Krafft/ da es auch eingesamlet werden soll.

5. Vermehrung. Ob gleich der Wermuth in Gärten/ damit man sie täglich zur Hand haben möge/ kan gepflantzet werden/ wird sie doch hierdurch nicht verbessert/ weßwegen man denn diese/ so von sich sel-

Stand der Technik: Zeichenerkennungsraten



Erkennungsraten liegen zwischen 94,6% und 99,6%.

(Springmann and Lüdeling 2017)

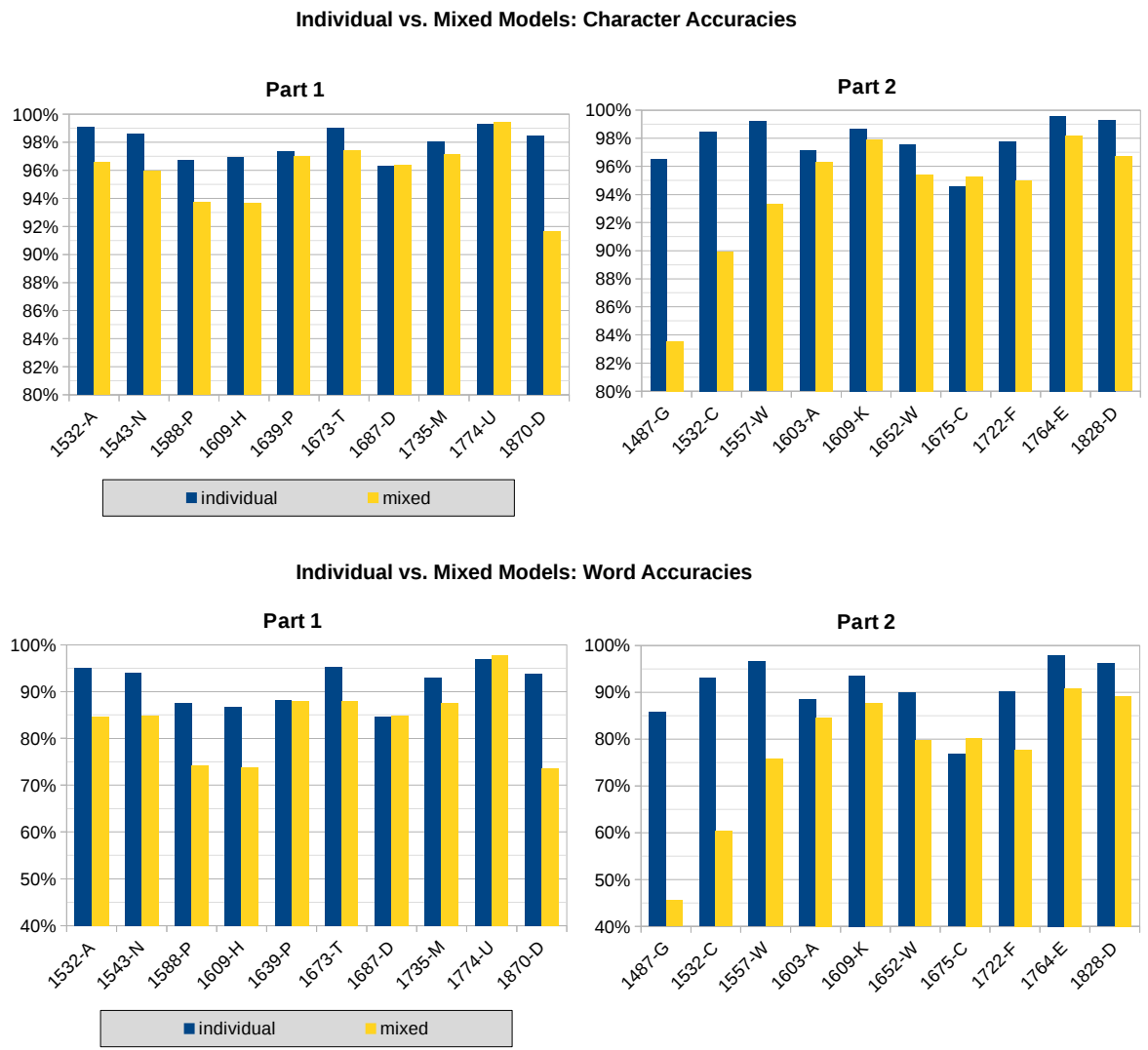
Muss man jede Type separat trainieren? Die Typographiebarriere

Zeilen: Drucke; Spalten: Modelle

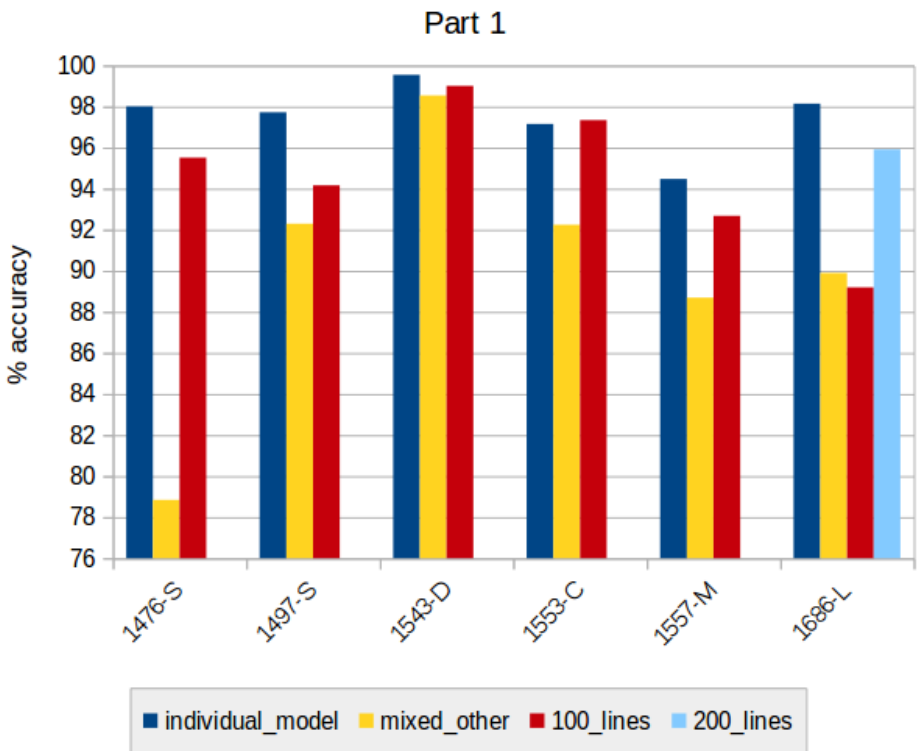
	1487-G	1532-A	1532-C	1543-N	1557-W	1588-P	1603-A	1609-H	1609-K	1639-P	1652-W	1673-T	1675-C	1687-D	1722-F	1735-M	1764-E	1774-U	1828-D	1870-D
1487-G	**96.5**	77.0	74.7	75.4	79.3	72.4	74.1	68.1	72.6	73.8	71.0	70.4	77.5	64.6	72.6	71.3	66.6	72.2	64.2	54.7
1532-A	84.3	**99.1**	85.4	85.7	90.5	84.8	90.5	87.2	89.4	83.5	87.2	81.9	90.6	79.7	74.1	84.2	67.9	75.8	70.3	59.2
1532-C	80.0	74.9	**98.5**	84.5	84.4	72.9	80.9	74.5	67.7	77.8	63.8	67.0	80.7	60.1	67.5	66.3	63.6	67.6	56.2	46.4
1543-N	89.9	88.6	91.0	**98.6**	91.6	85.2	86.9	85.0	86.1	85.5	80.5	81.5	88.9	74.9	80.0	84.3	73.8	75.7	71.1	60.8
1557-W	90.0	84.8	87.6	84.0	**99.2**	82.1	83.6	79.8	76.7	84.1	83.4	70.2	89.7	69.1	73.4	78.9	66.9	79.5	76.4	63.8
1588-P	69.2	71.2	66.9	66.8	72.2	**96.7**	86.6	86.2	85.1	88.4	90.3	84.8	88.6	82.1	76.4	79.1	72.3	74.9	70.0	62.3
1603-A	78.4	81.9	79.7	78.5	78.5	89.0	**97.1**	**95.7**	91.4	90.0	83.8	87.9	87.5	84.6	85.7	84.6	76.3	76.6	64.3	63.1
1609-H	67.7	72.8	72.4	69.3	68.8	86.4	93.6	**96.9**	87.8	84.3	80.0	81.5	82.9	78.2	76.5	76.9	65.3	66.9	59.6	58.9
1609-K	83.1	83.4	81.6	82.6	83.3	93.9	**97.0**	**96.2**	**98.7**	92.7	92.1	90.9	93.3	91.5	84.7	88.2	80.3	82.5	76.7	68.0
1639-P	79.7	80.1	77.7	79.3	82.0	91.8	92.6	91.7	91.0	**97.3**	94.5	89.3	93.6	86.7	86.2	86.9	81.1	86.5	75.8	70.1
1652-W	71.5	77.1	71.4	61.4	76.7	91.6	89.0	85.8	85.8	92.4	**97.6**	87.8	92.0	86.8	82.7	84.8	78.8	83.0	72.8	66.1
1673-T	73.0	79.1	70.3	69.0	77.3	88.8	91.8	88.7	90.6	90.3	91.1	**99.0**	93.5	90.6	87.8	88.2	86.3	83.9	78.3	70.3
1675-C	72.0	72.6	73.3	76.3	75.8	88.5	82.7	85.3	84.9	91.7	89.1	82.4	94.6	80.8	78.7	80.9	76.8	79.4	73.0	66.2
1687-D	74.2	76.7	63.7	64.0	68.1	82.2	89.3	87.0	88.7	87.6	89.5	90.3	94.2	**96.3**	86.6	84.7	84.5	83.7	77.5	69.2
1722-F	75.8	71.5	70.5	72.2	73.2	81.4	88.5	84.7	84.7	89.3	83.5	87.3	92.2	84.7	**97.8**	91.6	87.5	86.9	77.0	73.0
1735-M	79.0	80.1	77.8	81.0	82.5	85.1	90.8	86.1	87.6	91.6	87.3	90.1	92.0	86.8	94.7	**98.1**	90.8	91.5	86.9	85.1
1764-E	82.7	78.2	73.8	70.3	78.2	91.3	88.8	85.7	88.4	93.6	92.5	95.0	**97.2**	91.1	**95.6**	**95.3**	**99.6**	**96.2**	93.0	88.4
1774-U	81.6	80.6	79.9	76.3	84.6	92.7	92.6	90.5	90.3	**95.8**	**95.5**	93.0	**96.5**	91.2	94.4	**95.5**	**96.4**	**99.3**	94.3	87.2
1828-D	75.2	77.0	77.3	67.3	78.6	86.1	84.8	82.3	84.7	89.7	89.2	87.6	93.5	83.1	88.0	90.7	93.9	92.4	**99.3**	93.5
1870-D	71.3	71.6	69.2	65.6	69.9	81.3	80.4	80.1	79.8	84.9	82.3	84.5	87.4	81.3	86.1	84.2	86.6	84.5	88.2	**98.4**

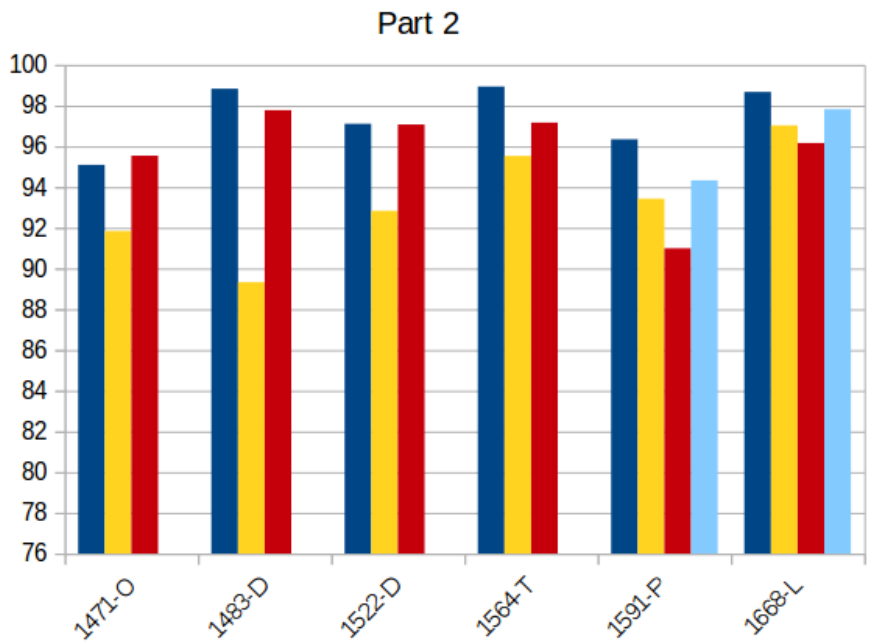
Gemischte Modelle verallgemeinern besser

Vergleich der Erkennungsraten von individuellen und gemischten Modellen



Dasselbe Bild ergibt sich für Antiqua-Drucke (Latein)



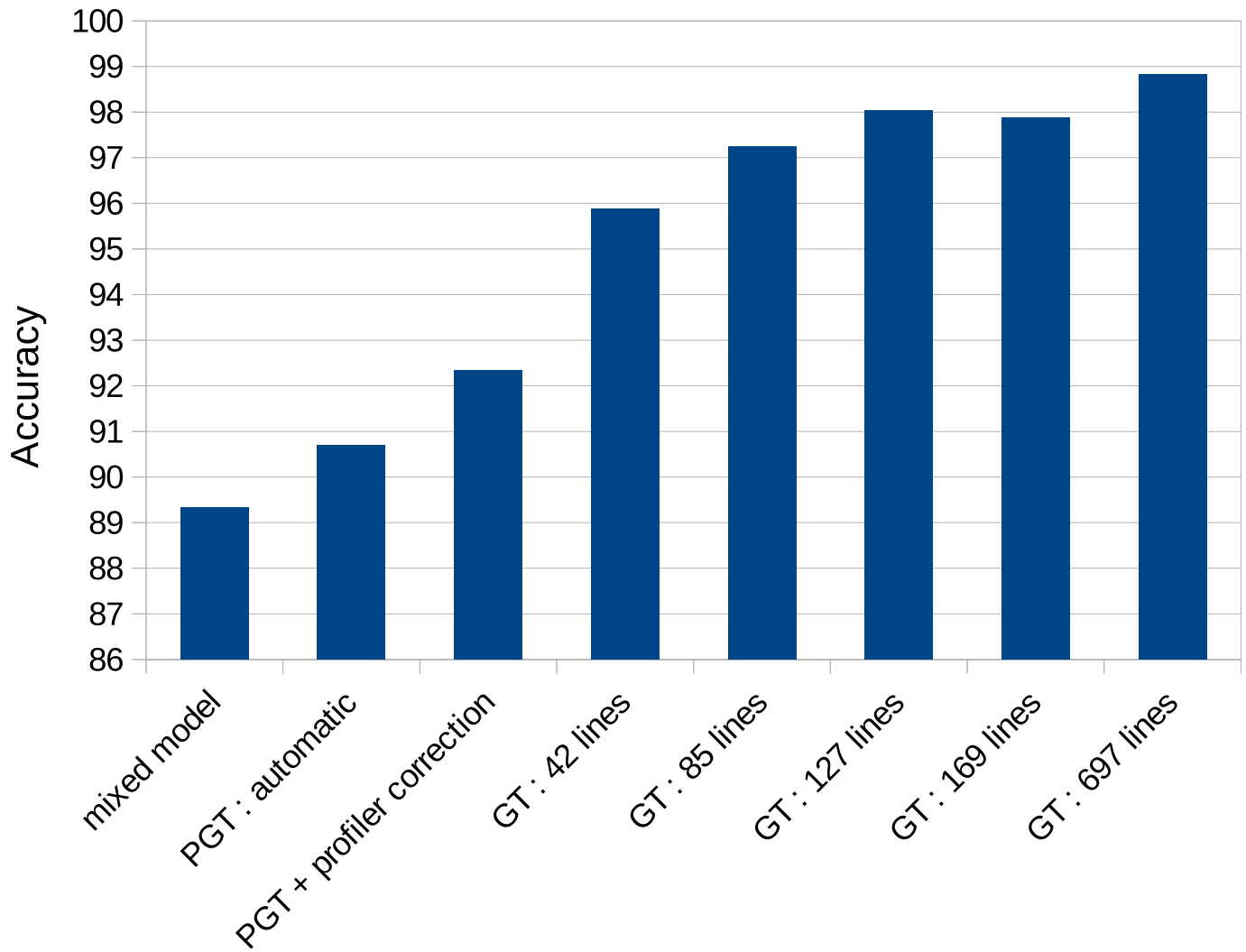


Zielkonflikt Masse versus Klasse?

- Hoffnung: Für viele Anwendungsfälle reicht Qualität einer Massen-OCR mit gemischten Modellen aus
- wo das nicht reicht, Anwendung der *Cowboy-Methode* (W.P. Klein):
  - Verwendung der Massen-OCR als Startnäherung an den gedruckten Text
  - *Draufsatteln von mehr Qualität* durch Nachkorrektur und Training eines individuellen Modells!
  - oft genügen schon wenige nachkorrigierte Zeilen
- nachkorrigierte Zeilen in Verbesserung gemischter Modelle einfließen lassen (*circulus virtuosus*)

Von Masse (*gemischtes Modell*) zu Klasse (*individuelles Modell*)

Beispiel Biondo, Decades (1483)



Fehlerstatistik der häufigsten Fehler

Part 2 mit gemischten Modell_1 (links) und mit individuellen Modellen (rechts)
Hauptfehlerquelle: Wortabstände
(vorher Löschen von Wortabständen: *merges*, nachher *splits* und *merges*)

```
errors        4510
total        64780
err          6.962 %

973    _     " "
183    c     e
143    r     _
100    _     i
71     a     _
47     t     r
37     _     t
32     __    ": "
31     i     _
30     i     ĩ
```

```
errors        1393
total        64780
err          2.150 %

46     " "     _
26     _       " "
24     _
22     i       _
21     c       e
19     _
17     _       i
```

Grenzen einer unicode-basierten OCR

- **Glyphen** (Graphe) sind Oberflächenformen von **Zeichen** (Graphemen):
  - Beispiel: a, *a*, **a** (Alloglyphen bzw. Allographe)
- das zugrundeliegende Zeichen ist jedesmal "LATIN SMALL LETTER A" (U+0061)

http://www.unicode.org/standard/principles.html:

> *The Unicode Standard does not define glyph images. The standard defines how characters are interpreted, not how glyphs are rendered. The software or hardware-rendering engine of a computer is responsible for the appearance of the characters on the screen. The Unicode Standard does not specify the size, shape, nor style of on-screen characters.*

- Keine Codierung und Erkennung von Allographen möglich (von Ausnahmen abgesehen)
- eine **glyphentreue** OCR gibt es nicht, sondern maximal eine **zeichentreue**

OCR für manche Fragestellungen nicht hinreichend

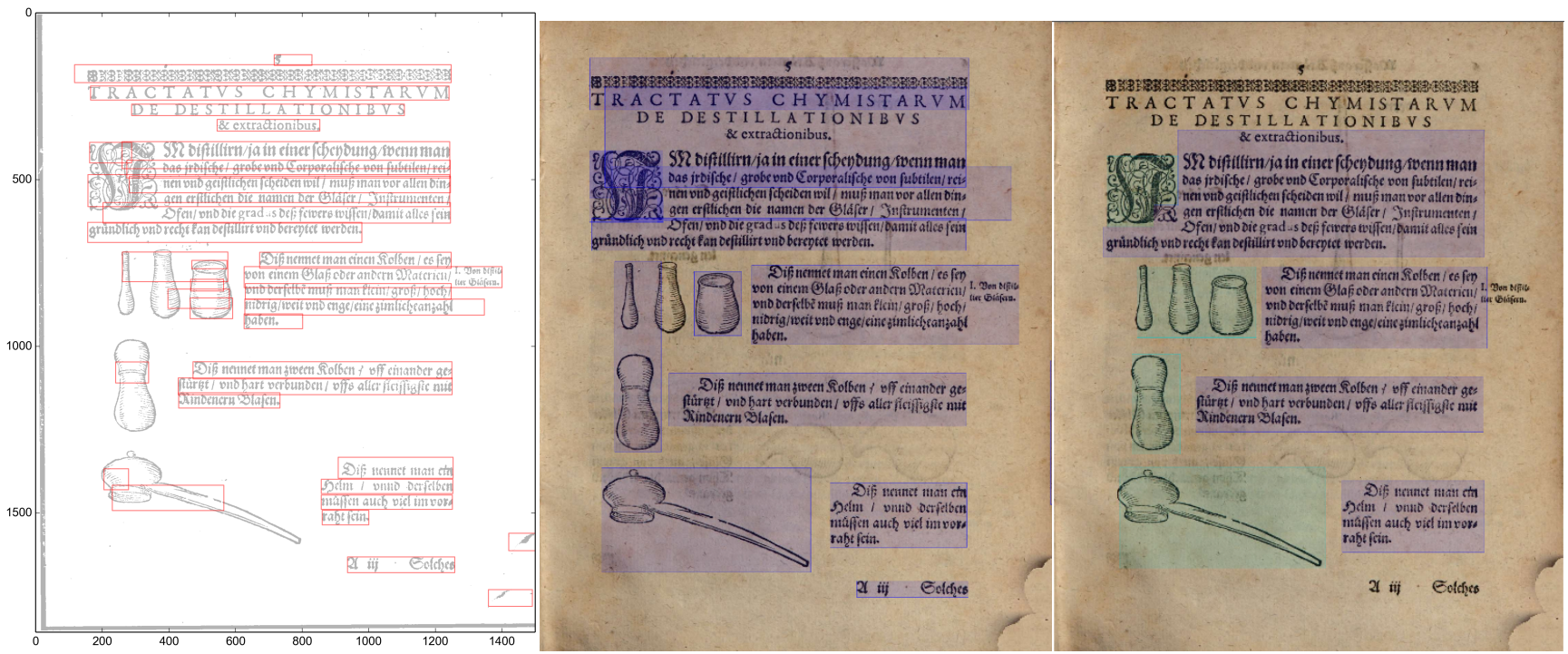
- kritisches Bewusstsein für Beschänkungen der Methode wichtig
- Oberflächenformen vermitteln kontextuelle Bedeutung:
  - Latein-Deutsch (Antiqua-Fraktur)
  - Hervorhebung (Sperrdruck, kursiv, fett)
  - etc.
- Konsequenz: weitere Annotation (z.B. in TEI) notwendig
- Weiterentwicklung automatischer Verfahren denkbar (Erkennung von Schriftarten)

Diesseits der OCR-Erkennung: Segmentierung

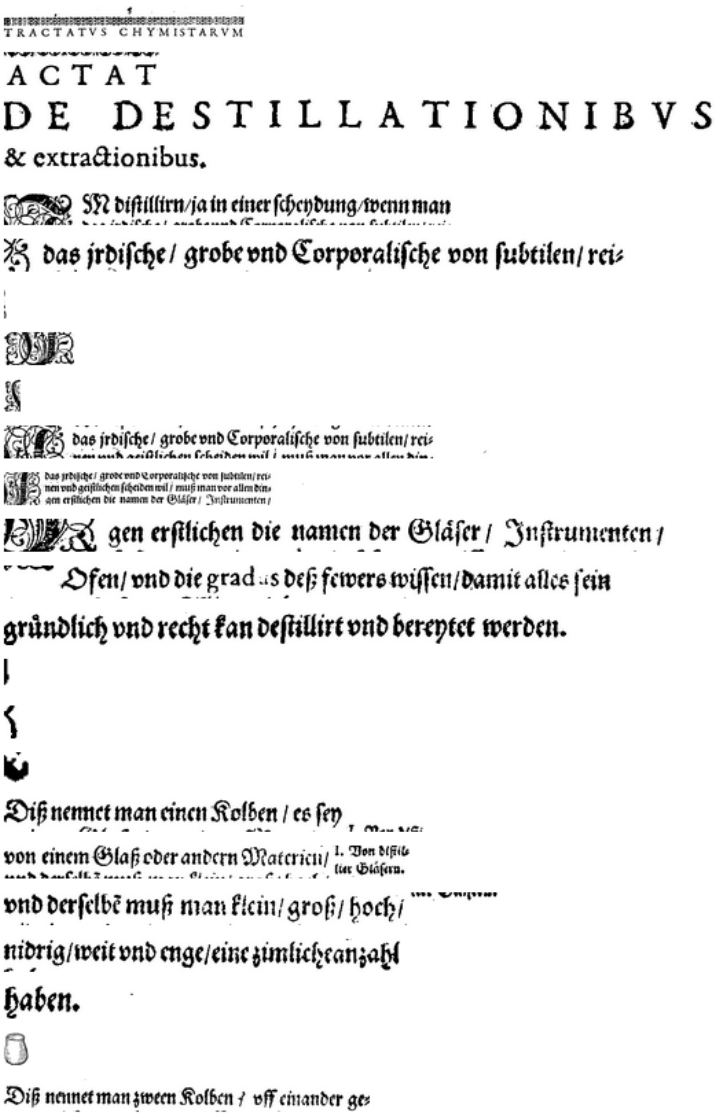
- was nicht richtig segmentiert wird, kann auch nicht erkannt werden
- es fehlt an einem Open-Source Werkzeug zur verläßlichen Segementierung von Druckseiten:
  - Text- und Nicht-Text-Bereiche
  - Segmentierung in Textzeilen
  - semantische Auszeichnung: Überschrift, Seitennummer, Kopf- und Fußzeilen, Marginalien, ...
- Probleme von OCRopus: Bilderkennung, unterschiedliche Schriftgrößen auf einer Seite
- auch Tesseract hat Probleme bei der Segmentierung

Beispiel: Segmentierung von Libavius

v.l.n.r.: OCRopus, Tesseract, ABBYY



Beispiel: Segmentierung von Libavius (Tesseract, OCRopus)



TRACTATVS CHYMISTARVM

ACTAT

DE DESTILLATIONIBVS

& extractionibus.

M diſtillirn/ja in einer ſcheydung/wenn man

das jrdiſche/ grobe vnd Corporaliſche von ſubtilen/ rei-

gen erſtlichen die namen der Gläſer/ Inſtrumenten/

Ofen/ vnd die grad ..s deß fewers wiſſen/damit alles ſein

gründlich vnd recht kan deſtillirt vnd bereytet werden.

Diß nennet man einen Kolben / es ſey

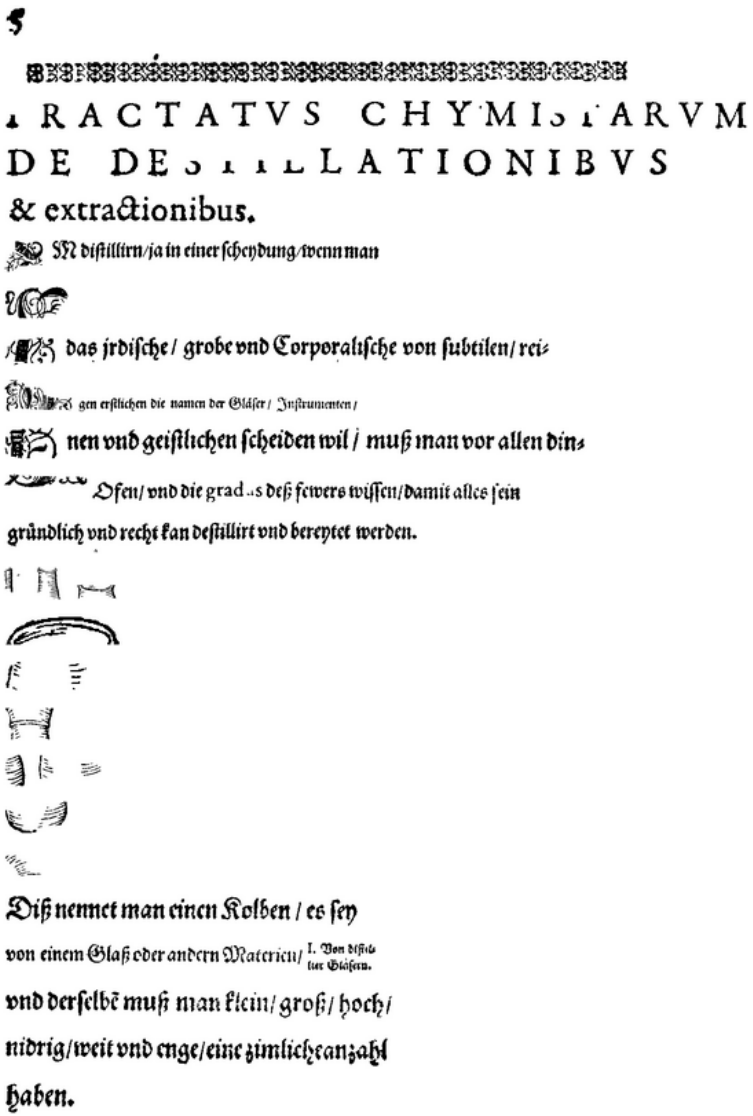
von einem Glaß oder andern Materien/ I. Von diſtil- lier Gläſern.

vnd derſelbē muß man klein/ groß/ hoch/

nidrig/weit vnd enge/eine zimliche anzahl

haben.

Diß nennet man zween Kolben / vff einander ge-



5

RACTATVS CHYMI..ARVM

DE DE..LLATIONIBVS

& extractionibus.

M diſtillirn/ja in einer ſcheydung/wenn man

das jrdiſche/ grobe vnd Corporaliſche von ſubtilen/ rei-

gen erſtlichen die namen der Gläſer/ Inſtrumenten/

nen vnd geiſtlichen ſcheiden wil/ muß man vor allen din-

Ofen/ vnd die grad ..s deß fewers wiſſen/damit alles ſein

gründlich vnd recht kan deſtillirt vnd bereytet werden.

Diß nennet man einen Kolben / es ſey

von einem Glaß oder andern Materien/ I. Von diſtil- lier Gläſern.

vnd derſelbē muß man klein/ groß/ hoch/

nidrig/weit vnd enge/eine zimliche anzahl

haben.

Jenseits der OCR-Erkennung: Normalisierung

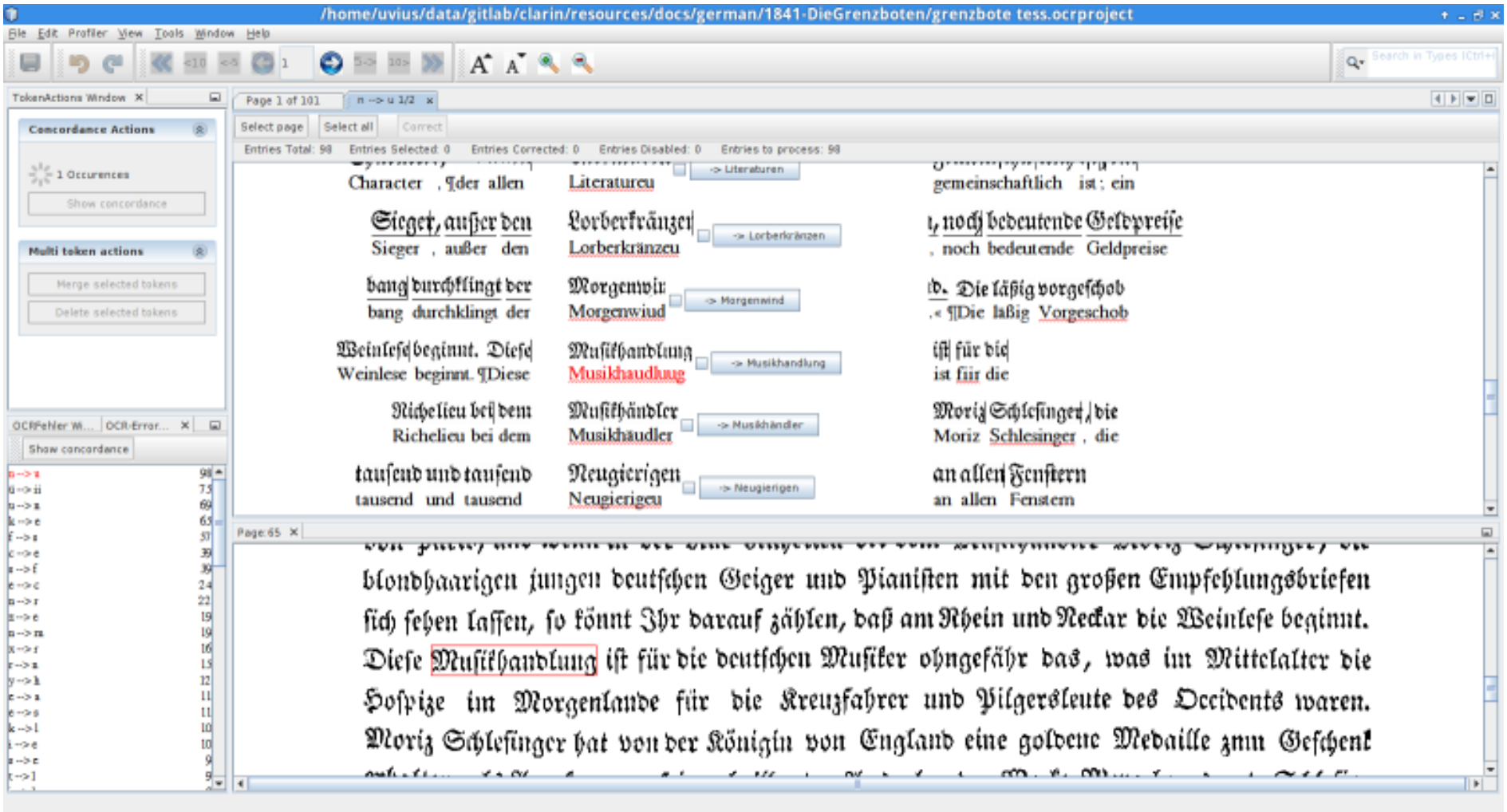
- Wie sucht man auf diesem Text?

  Concupiſcẽtia vero diligimus illud
  cui bonũ volum⁹ : ſcʒ ad delectandũ in eo · vtro=
  qʒ iſto₄ modo₄ diligimus deũ naturaliṫ: & ange
  li etiã in primo ſtatu · Sed diligebat angelus deũ
  ſup om̃ia amore ɔcupiſcentie · ſʒ in ip̃o delectan=
  do ſup om̃ia · Nec tn̄ ſeq̇tur q̱ haberet caritatem
  quia nõ diligebat deũ ꝑpter ip̃m deũ ſed ꝑpṫ ſe :

- Normalisierung auf heutige Schreibweise als Annotationsebene notwendig, siehe dazu z.B. Bollmann, Petran, and Dipper (2011), Jurish (2013)

Jenseits der OCR-Erkennung: Nachkorrektur

Nachkorrektur: z.B. mit dem interaktiven CIS-Tool PoCoTo



Resümee und Ausblick: Wo stehen wir?

Resümee und Ausblick

- Wir können heute bereits Drucke der gesamten modernen Druckgeschichte bis hinunter zu Gutenberg mit hoher Genauigkeit (> 95%) durch eine typentrainierte OCR erkennen.
- Es fehlt an verlässlicher *ground truth* und an einer Kultur des offenen Datenaustausches. Auch die Urheberrechtsfrage von Scans (Stichwort *copyfraud*, *Schutzrechtsberühmung* bei gemeinfreien Inhalten) und die damit einhergehende Nichtverfügbarkeit hochaufgelöster Bilddaten erschwert das Modelltraining.
- Eine koordinierte Initiative deutscher Institutionen könnte auf Basis der vorliegenen Scans eine OCR-Erfassung durchführen und zentral zur Nachkorrektur anbieten. Der dadurch entstehende *ground truth*-Vorrat könnte automatisiert zur Modellverbesserung genutzt und damit ein sich stets verbessernder Zirkel in Gang gesetzt werden, an dessen Ende das gesamte bildmäßig erfasste Material als hochgenauer elektronischer Text vorliegt.

Wenn Sie mehr erfahren möchten

- (Semi-) automatische Verbesserung von OCR-Resultaten ausgehend von gemischten Modellen (Springmann, Fink, and Schulz 2016)
- CIS OCR Workshop (Springmann and Fink 2016)
- PoCoTo, ein Nachkorrekturwerkzeug für historische OCR
- Ocrocis (Springmann and Kaumanns 2015):
  A project manager interface to OCRopus
- Ocrocis Tutorial (Springmann 2015)
  Ausführliche Anleitung zum Trainieren eigener Modelle
- ein allgemeines OCR Tutorial (Springmann 2014)

Vielen Dank für Ihre Aufmerksamkeit!

Dr. Uwe Springmann
☙ digital humanist ☙
vorname [A T] nachname.net

Literaturangaben I

Bollmann, Marcel, Florian Petran, and Stefanie Dipper. 2011. “Applying Rule-Based Normalization to Different Types of Historical Texts – an Evaluation.” In *Human language technology challenges for computer science and linguistics*, 166–77. Springer.

Breuel, Thomas M, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. 2013. “High-Performance OCR for Printed English and Fraktur Using LSTM Networks.” In *2th International Conference on Document Analysis and Recognition (Icdar), 2013*, 683–87. IEEE.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural computation* 9 (8). MIT Press: 1735–80.

Jurish, Bryan. 2013. “Canonicalizing the Deutsches Textarchiv.” In *Proceedings of Perspektiven einer corpusbasierten historischen Linguistik und Philologie (Berlin, 12th - 13th December 2011)*, edited by Ingelore Hafemann. Vol. 4. Thesaurus Linguae Aegyptiae. Berlin, Germany: Berlin-Brandenburgische Akademie der Wissenschaften. http://edoc.bbaw.de/frontdoor.php?source_opus=2443.

Literaturangaben II

Rydberg-Cox, Jeffrey A. 2009. "Digitizing Latin Incunabula: Challenges, Methods, and Possibilities." *Digital Humanities Quarterly* 3 (1). http://www.digitalhumanities.org/dhq/vol/3/1/000027/000027.html/#p7.

Springmann, Uwe. 2015. "Ocrocis: A high accuracy OCR method to convert early printings into digital text – A Tutorial." http://cistern.cis.lmu.de/ocrocis/tutorial.pdf.

Springmann, Uwe, and Florian Fink. 2016. "CIS OCR Workshop v1.0: OCR and postcorrection of early printings for digital humanities." doi:10.5281/zenodo.46571.

Springmann, Uwe, and David Kaumanns. 2015. "Ocrocis – a high accuracy OCR method to convert early printings into digital text." http://cistern.cis.lmu.de/ocrocis/.

Springmann, Uwe, and Anke Lüdeling. 2017. "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.

Literaturangaben III

Springmann, Uwe, Florian Fink, and Klaus U Schulz. 2016. "Automatic Quality Evaluation and (Semi-) Automatic Improvement of OCR Models for Historical Printings." *ArXiv E-Prints*. http://arxiv.org/abs/1606.05157.

Springmann, Uwe, Anke Lüdeling, and Felix Schremmer. 2015. "Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten." DHd-Tagung 2015, Graz. http://gams.uni-graz.at/o:dhd2015.p.34.

Springmann, Uwe, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, and Florian Fink. 2014. "OCR of historical printings of Latin texts: problems, prospects, progress." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 57–61. DATeCH '14. New York, NY, USA: ACM. doi:10.1145/2595188.2595197.