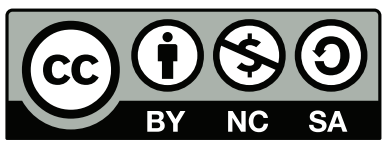
Module 1
Challenges & Methods

Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München (LMU)



2015-09-14

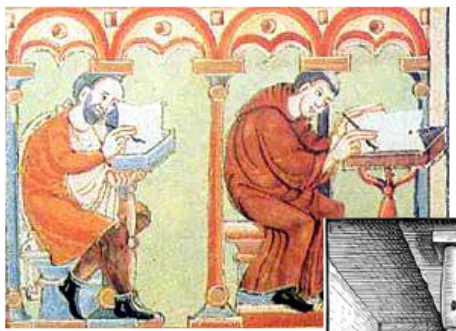
Goals

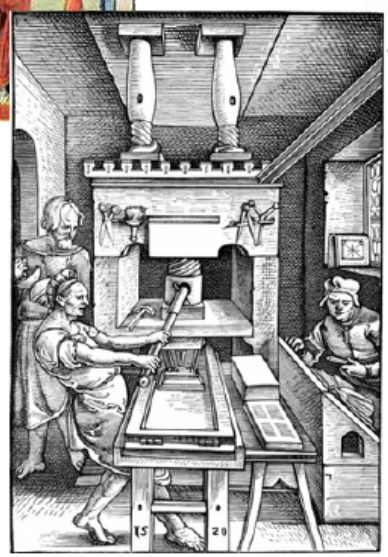
1. make electronic representations of (all) documents universally available
   - make scanned images of document pages accessible over the internet
2. make scanned images searchable
   - OCR (with errors)
3. make one representation as machine-actionable electronic text
   - annotation, postcorrection

- can be seen as large-scale program or as individual project focused on specific documents
- this workshop: mostly concerned with steps 2 and 3 above

Transmission of texts

Manuscript





Printing

PDF (Image)

GLOSSARIUM
MEDIÆ ET INFIMÆ LATINITATIS
CONDITUM A CAROLO DU FRESNE
DOMINO DU CANGE
AUCTUM
A MONACHIS ORDINIS S. BENEDICTI
CUM SUPPLEMENTIS INTEGRIS
D. P. CARPENTERII
ADELUNGII, ALIORUM, SUISQUE
DIGESSIT
G. A. L. HENSCHEL
SIMULQUE
GLOSSARIUM GALLICUM, TABULÆ, INDICES AUCTORUM ET RERUM, DISSERTATIONES
EDITIO NOVA aucta pluribus variis clarorum scriptorum
LÉOPOLD FAVRE

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

Introduction to OCR

OCR: definition & history

- **Optical Character Recognition (OCR): automated conversion of images of printed pages to machine-actionable text**
- early applications: reading device for blind people (Fournier d'Albe: Optophone, 1913; Kurzweil: Reading Machine, 1974)
- today important business: paperless office, automatic workflow
- leading proprietary products: *Finereader* (ABBYY), *Omnipage* (Nuance), *ReadIris* (Canon)
- good open source software available since 2005: *Tesseract* (Ray Smith, HP Labs, now Google), *OCRopus* (Tom Breuel, DFKI Kaiserslautern, now Google)

OCR workflow

- the complete OCR workflow consists of several steps (step 3 is optional):
  1. image acquisition
  2. preprocessing
  3. (ground truth production, model training)
  4. recognition
  5. evaluation
  6. postprocessing: annotation, error correction, tagging, ...

OCR research

- OCR belongs to pattern recognition, artificial intelligence, computer vision (hot topics)
- product related proprietary research mostly done in commercial companies (scanning hardware manufacturers, Google)
- general opinion: OCR is a solved problem! (for 20th century printings and beyond: >99% correctly recognized characters)
- not at all true for earlier printings: Gothic scripts, non-Latin alphabets, unusual glyphs, complex layout, book degradation from usage and ageing
- much academic research on postprocessing of commercial engine OCR output (spelling correction, annotation, search in noisy data)

Renewed interest in OCR

- massive digitization (=scanning!) of historical printings (newspapers, books): Google Books (scan 130 mill. books until 2020), libraries (Bavarian State Library has > 1 mill. books scanned, HathiTrust: > 10 mill. books)
- long term goal of funding institutions: make all scanned books available in text form (must be automatic process = OCR)
- EU IMPACT project (2008-2012)
- CIS: Prof. Schulz (postcorrection, since 2004)
- Open Greek and Latin project, Greg Crane (U Leipzig)
- Early Modern OCR Project (eMOP), Laura Mandell (Texas A&M University)
- Dan Klein, Taylor Berg-Kirkpatrick (University of California, Berkeley): *Ocular*

Digression: OCR errors, OCR quality measures

Important concepts to know

- we talk of OCR errors as misrecognized elements (characters or words)
- *error rate*: errors / all elements
- *accuracy*: correctly recognized elements / all elements = 1 - error rate
- the rest of this section is more mathematical and serves as background reading

OCR errors

OCR errors can be classified as elementary edit operations:

- misspelled characters: *substitutions, s*
- spurious symbols: *insertions, i*
- missing text: *deletions, d*

for OCR sometimes additional elementary operations: * symbol splits, e.g. m -> in * symbol merges, e.g. cl -> d

Example:

- exerciſed → exercifed (*substitution* of long s by f)
- in → m (*deletion* of i followed by substitution n → m)
- having → hav ing (*insertion* of blank, resulting in word split)

Levenshtein distance, error rate, accuracy

Levenshtein distance (LD): the minimum number of edit operations to transform an input string into an output string

Example: *ernest to nester*: LD = 4

- delete *er* at beginning and insert *er* at end
- not: substitute each letter separately (6 operations!)
- -> now we have an unambiguous definition of *s+i+d*
- the single errors *s,i,d* may not be unique (ab -> ba: s=2 or d=1,i=1)!

We have *errors (s,i,d)* and *correct output tokens (c)* (4 oberservables) with
$n_{\mathrm{GT}} = c + s + d, n_{\mathrm{OCR}} = c + s + i$

Error rate: ratio of errors to "all" tokens (n), $e = \frac{s+i+d}{n} = \frac{s+i+d}{c+s+i+d}$

(often $n = n_{\mathrm{GT}}$ or $n = n_{\mathrm{OCR}}$ - watch out for used definitions!)

error rate can be measured at character (CER) or word (WER) level

Accuracy: ratio of correct tokens to "all" tokens, $A = \frac{c}{c+s+i+d} = 1 - e$

Definition of precision and recall

think Cinderella, picking out lentils with the help of birds:

The good ones go into the pot,
The bad ones go into your crop

- four cases:
  - True positives, T_p: good ones picked out
  - False positives, F_p: bad ones falsely picked out or good ones damaged
  - True negatives, T_n: bad ones correctly eaten
  - False negatives, F_n: good ones missed, falsely eaten or damaged
- summing up:
  - number of items picked out: $N_\mathrm{pot} = T_\mathrm{p} + F_\mathrm{p} = N_\mathrm{OCR}$
  - number of good items: $N_\mathrm{good} = T_\mathrm{p} + F_\mathrm{n} = N_\mathrm{GT}$
- **Precision: proportion of good items in retrieved set, $p = T_\mathrm{p}/N_\mathrm{pot}$ (Reinheitsgrad)**
- **Recall: proportion of good items retrieved, $r = T_\mathrm{p}/N_\mathrm{good}$ (Ausbeute)**

Precision and recall in OCR

- we have:
  - $T_\mathrm{p} = c$
  - $T_\mathrm{n} = 0$ (we want to recognize all items, none are originally bad)
  - $N_\mathrm{GT} = c + s + d$
  - $N_\mathrm{OCR} = c + s + i$
- therefore:
  - $p = \frac{c}{c+s+i}$
  - $r = \frac{c}{c+s+d}$
- now we can identify F_p and F_n in terms of OCR errors:
  - $F_\mathrm{p} = s + i$
  - $F_\mathrm{n} = s + d$ (not missed items, but damaged and destroyed items)
- make one measure out of two:
  - F-measure, harmonic mean of p and r
  - $F = \frac{2pr}{p+r}$

Historical OCR

OCR for historical printings?

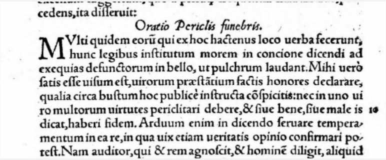
In historical documents we often find:

- lots of different printing types
- high variability in letter shapes
- special glyphs, script and alphabet mixtures
- high variability in spelling, morphology, and syntax → variable context
- right justification in manual typesetting leads to:
  - abbreviations (vnd, vñ)
  - insertions of consonants (von, vonn)
  - narrow inter-word spacing

Therefore:

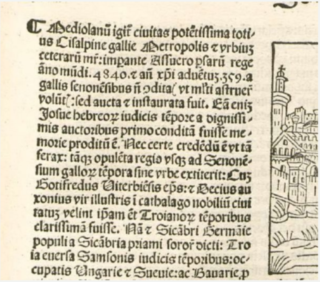
- results are often unsatisfactory for broken scripts (Gothic, Fraktur) and earlier texts (Piotrowski 2012; Strange et al. 2014)

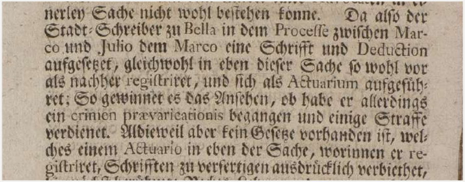
The challenge (I): historical typographies

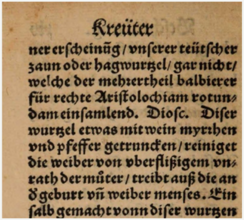

cedens, ita disseruit:

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


Mediolanũ igit ciuitas potẽtissima totius Cisalpine gallie Metropolis ꝛ vrbiũ ceterarũ mr: impante Assuero psarũ rege ãno mũdi. 4840. ꝛ añ xpi aduẽtũ 359. a gallis senonẽsibus ñ ɔdita (vt mlti astruũt volũt): sed aucta ꝛ instaurata fuit. Eã enĩ Josue hebreoꝝ iudicis tẽpoꝛe a dignissimis auctoꝛibus primo conditã fuisse memoꝛie proditũ ẽ. Nec certe credẽdũ ẽ vt tã ferax: tãq opulẽta regio vsq ad Senonẽsium galloꝝ tẽpora sine vrbe extiterit: Cũ Gotifredus Viterbiẽsis eps: ꝛ Decius auxonius vir illustris i catbalago nobiliũ ciuitatũ velint ipsam ẽt Troianoꝝ tẽporibus clarissimã fuisse. Nã ꝛ Sicãbri Germãie populi a Sicãbria pꝛiami soroꝛ dicti: Troia euersa Samsonis iudicis tẽpoꝛibus: occupatis Ungarie 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,


Kreüter

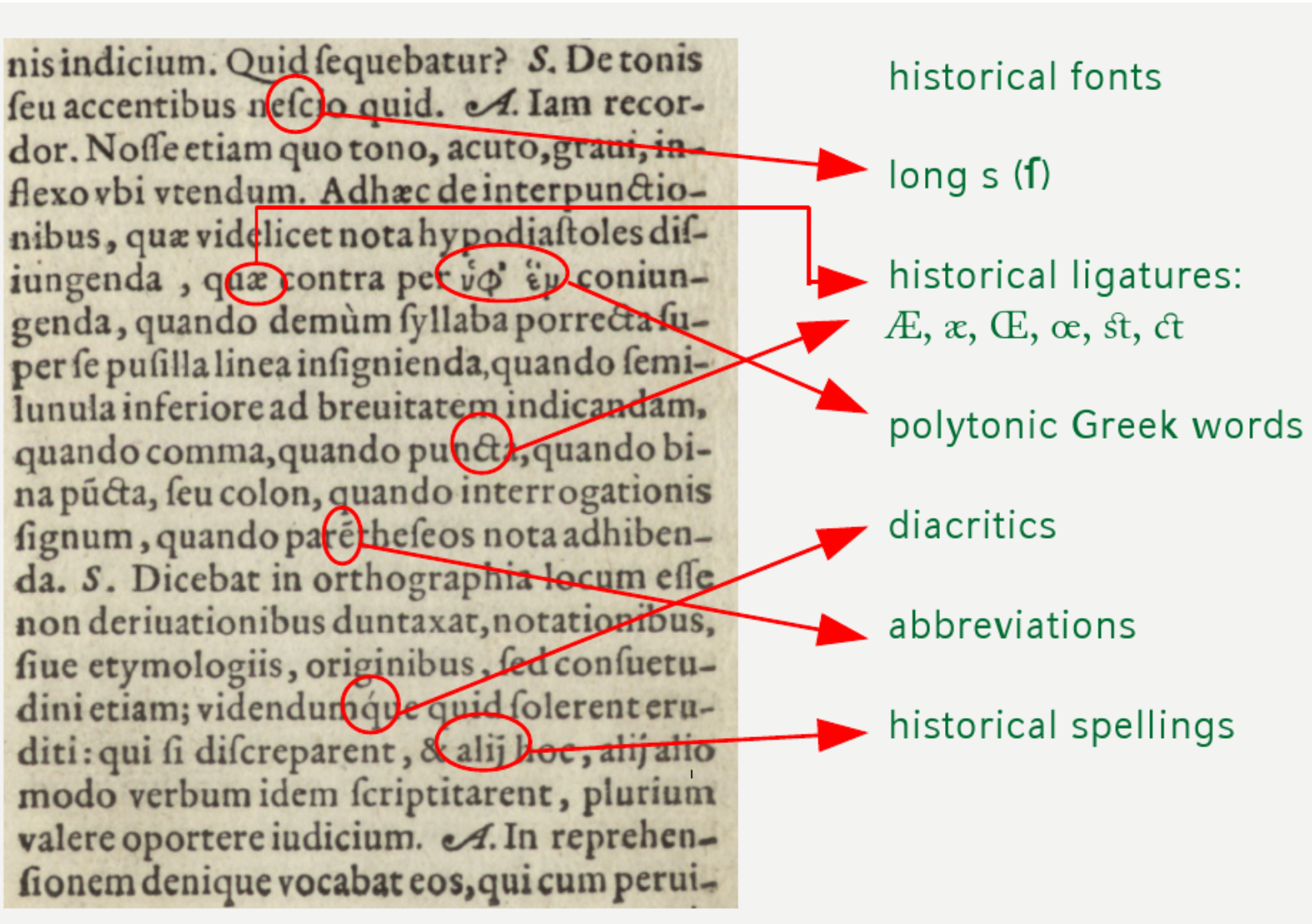
ner erscheinũg / vnserer teütscher zaun oder hagwurtzel / gar nicht / welche der mehrertheil balbierer für rechte Aristolochiam rotundam einsamlend. Diosc. Diser wurtzel etwas mit wein myrrhen vnd pfeffer getruncken / reiniget die weiber von vberflüßigem vnrath der müter / treibt auß die andgeburt vñ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

clockwise: printing year (author)

1564 (Valla), 1487 (Foresti), 1735 (Leyser), 1557 (Bodenstein)

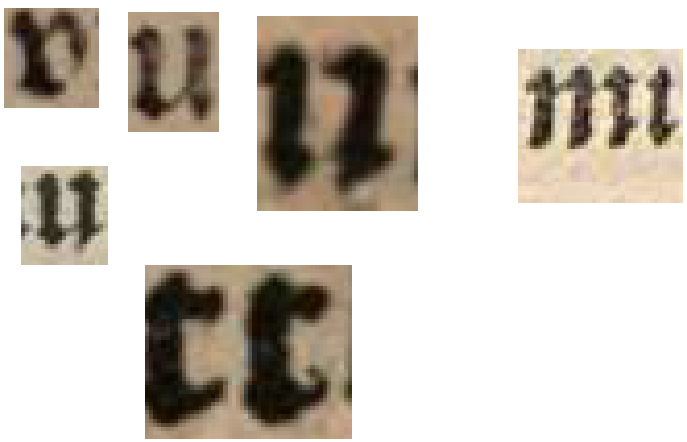
The challenge (II): special glyphs

Pontanus: Progymnasmata Latinitatis (1589)



The challenge (III): historical fonts, historical spellings

(Anke Lüdeling, HU Berlin)

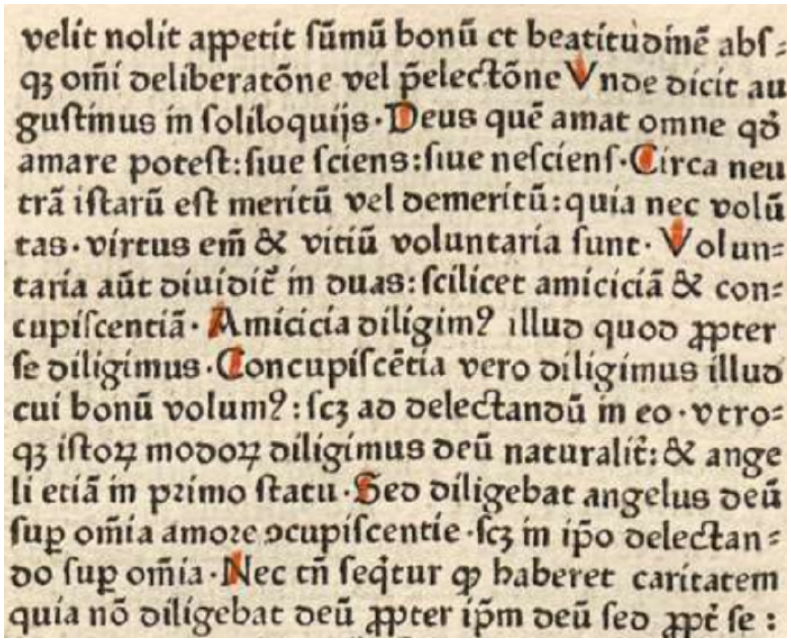


u? n? tt? un? v?

meüßörlin (modern: Mäusöhrlein)? brey (modern: Brei)? brust (brnst)?

The challenge (IV): incunabula

Beauvais: Speculum naturale (not after 1476); ABBYY FR11 Fraktur 68% acc.



velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ=
q3 oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs·Deus quẽ amat omne qꝺ
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 amicicia & con=
cupiſcentiã·Amicicia diligim? illuꝺ quoꝺ ꝑter
ſe diligimus·Concupiſcẽtia vero diligimus illuꝺ
cui bonũ volum?:ſcz aꝺ delectandũ in eo·vtro=
q3 iſtoꝝ modoꝝ diligimus deũ naturaliter:& 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 ꝑter ipm deũ ſeꝺ ꝑꝛ 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 :

An incunabulum printing has special abbreviation signs, e.g. ꝑ ꝓ p̈ q3 Q̾ q̾ ſcz.

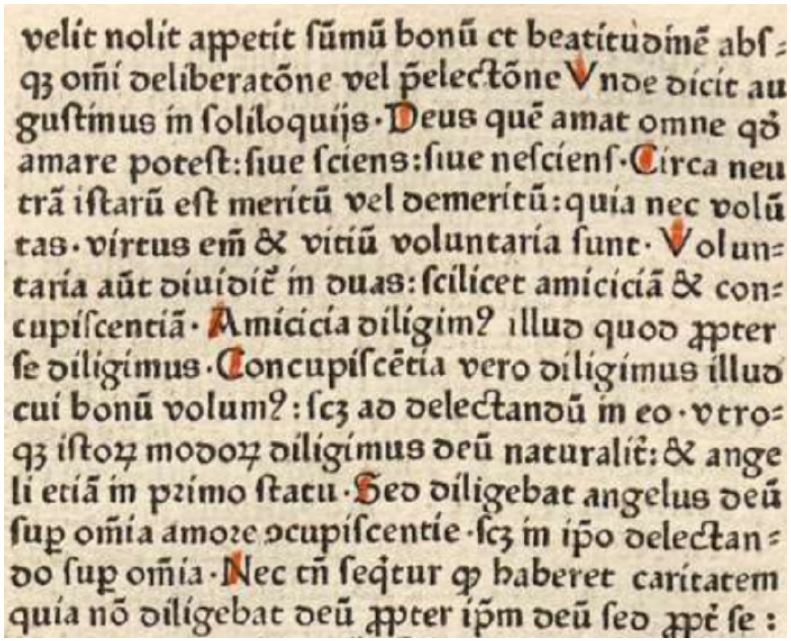
(Rydberg-Cox 2009) (our emphasis): *"Because of the prevalence of these glyphs,* ***incunabula cannot be processed using OCR software.*** *Commercial OCR programs produce almost no recognizable character strings, let alone searchable text. …* ***Other methods*** *must be explored."*

Other (OCR) methods: OCR with recurrent neural networks

- recurrent neural network (RNN) with long short-term memory (LSTM) as
- invented by (Hochreiter and Schmidhuber 1997), first applied to OCR by (Breuel et al. 2013)
- input layer: pixel values of vertically sliced text lines (500–1000 frames)
- memory layer: 100 hidden memory blocks
- output layer: character representations (glyphs)
- needs training (either on artifically generated images from text or ground truth corresponding to printed text)
- learns by adjusting weights between connections of layers
- does not need a language model
- can be trained on a lot of scripts and languages, even on mixed cases

Trained models for incunabula

Trained OCRopus model (this passage: 99% acc.)



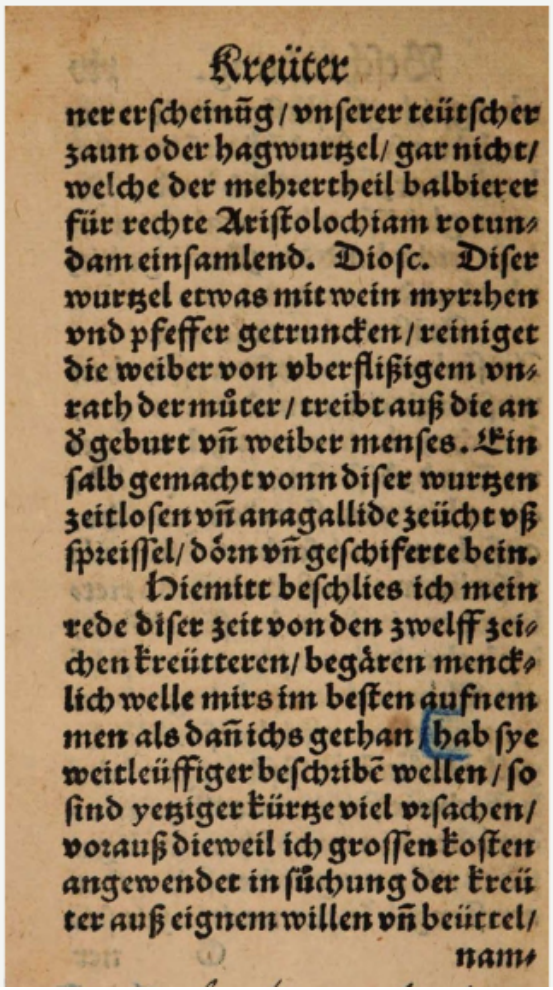
velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ ·
qȝ oñi deliberatõne vel p̃electõne Vnde dicit au
guſtinus in ſoliloquijs · Deus quẽ amat omne qd̓
amare poteſt:ſiue ſciens:ſiue neſcienſ· Circa neu
trã iſtarũ eſt meritũ vel demeritũ : quia nec volũ
tas · virtus eñ & vitiũ voluntaria ſunt · Vol une
taria aũt diuidi᷑ in duas: ſcilicet amicicia & 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 naturali᷑: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſup oñia amore ocupiſcentie · ſcȝ in iꝑo delectan
do ſup oñia · Nec tñ ſeq̓tur q haberet caritatem
quia nõ diligebat deũ ꝓpter iꝑm deũ ſed ꝓp̓ ſe :

- trained on 13 pages, tested on additional 4 pages
- 98% average character accuracy (raw, uncorrected output)
- no language model employed

Schwabacher font: old and new methods

Adam von Bodenstein (1557); ABBYY FR 11 Fraktur + hist. lexicon

original page



Kreüter
ner erſcheinūg / vnſerer teütſcher zaun oder hagwurtzel / gar nicht / welche der mehrertheil balbierer für rechte Ariſtolochiam rotun- dam einſamlend. Dioſc. Diſer wurtzel etwas mit wein myrrhen vnd pfeffer getruncken / reiniget die weiber von vberflißigem vn- rath der mūter / treibt auß die an ð geburt vñ weiber menſes. Ein ſalb gemacht vonn diſer wurtzen zeitloſen vñ anagallide zeücht vß ſpreiſſel / dörn vñ geſchiferte bein.
Hiemitt beſchlies ich mein rede diſer zeit von den zwelff zei- chen kreütteren / begären menck- lich welle mirs im beſten aufnem men als dañ ichs gethan / hab ſye weitleüffiger beſchribē wellen / ſo ſind yetziger kürtze viel vrſachen / vorauß dieweil ich groſſen koſten angewendet in ſūchung der kreü ter auß eignem willen vñ beüttel /
nam-

Abbyy FR 11 (84.65%)

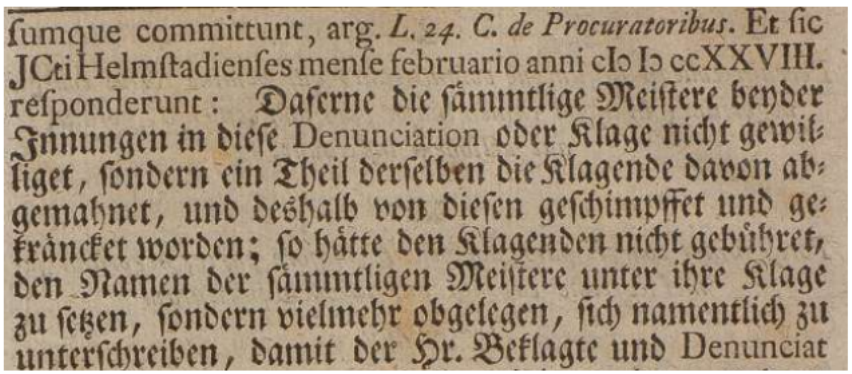
Kreücev
ner erscheinng / vnserer teutscher zäun oderhagwuryel/ garnicht/ welche der mehrertheil baldierer fiir rechte Aristolochiam rotun- dameinsamlend. Diosc. Diser würget etwas mit wein Myrrhen vnd Pfeffer getrvncken/ reiniget die weider von vderssissigem vm rath dermäter/ treibt auß die an Lgedurt vn weider menses. Ein faldgemachtvonndiser würgen zeitlosenvnanagallidezeüchtvß spreissel/ dörnvngeschiferte dein.
iZiemitt deschlies ich mein rede diser zeitvonden zwelffzei- chenkreütteren/degären menck- ltch welle mirs im dessen aufnem men als dan ichs gethan / had fye weitleüfftgerdeschridc wellen / f» sind yegigerkürgeviel vrsachen/ voraußdieweil ichgrosicnkosten angewendet insüchungder kreü ter auß eignem willen vndeücrcl/

OCRopus (98.84%; raw uncorrected output)

Kreüter
ner erſcheinüg/ vnſerer teütſcher zaun oder hagwurtzel/ gar micht/ welche der mehrertheil balbierer für rechte Ariſtolochiam rotun= dam einſamlend. Dioſc. Diſer wurtzel etwas mit wein myrrhen vnd pfeffer getruncken/ reiniget die weiber von vberflifzigem vn= rath der můter/ treibt aufz die an d geburt vñ weiber menſes. Ein falb gemacht vonn diſer wurtzen zeitloſen vñ anagallide zeücht vfz ſpreiffel/ dórnvñ geſchiferte bein.
Hiemitt beſchlies ich mein rede diſer zeit von den zwelff zei= chen kreütteren/ begären menck= lich welle mirs im beſten aufnem men als dañ ichs gethan/ hab ſye weitleüffiger beſchribē wellen/ ſo ſind yetziger kürtze viel vrſach_en/ voraufz dieweil ich groſſen koſten angewendet in ſůchung der kreü ter aufz eignem willen vñ beüt_tel/
ncn=

Mixed typefaces: old and new methods

Augustinus Leyser (1735)



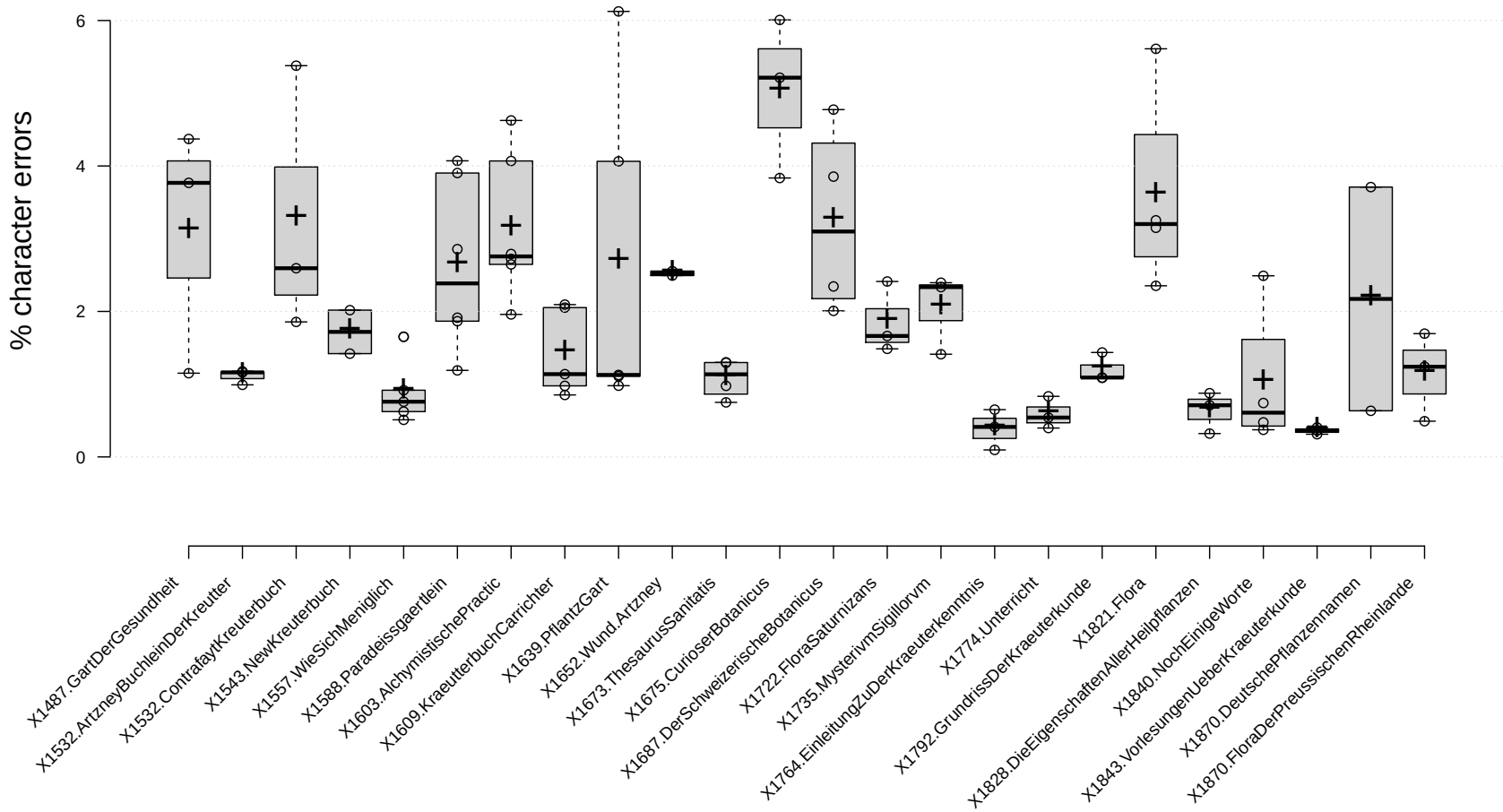
ſumque committunt, arg. *L. 24. C. de Procuratoribus.* Et ſic JCti Helmſtadienſes menſe februario anni cIↃ IↃ ccXXVIII. reſponderunt: Daferne die ſämmtlige Meiſtere beyder Innungen in dieſe Denunciation oder Klage nicht gewilliget, ſondern ein Theil derſelben die Klagende davon abgemahnet, und deshalb von dieſen geſchimpffet und gekräncket worden; ſo hätte den Klagenden nicht gebühret, den Namen der ſämmtligen Meiſtere unter ihre Klage zu ſetzen, ſondern vielmehr obgelegen, ſich namentlich zu unterſchreiben, damit der Hr. Beklagte und Denunciat

ſumque committunt, arg.L. 24. C. de Procuratcribus.Et ſic JCti Helmſtadienſes menſe februario anni cIↃIↃ ceXXVIII. reſponderunt: Daferne die ſämmtlige Meiſtere beyder Innungen in dieſe Denunciation oder Klage nicht gewilliget, ſondern ein Theil derſelben die Klagende davon abgemahnet, und deſhalb von dieſen geſchimpffet und gekräncket worden; ſo hätte den Klagenden nicht gebühret, den Namen der ſämmtligen Mettere unter ihre Klage zu ſeten, ſondern vielmehr obgelegen, ſich namentlich zu unterſchreiben, damit der Hr.Beklagte und Denunciat

- mixed typefaces: Fraktur for German, Antiqua for Latin.
- trained on 40 pages, tested on 8 pages.
- mean acc. 97% (ABBYY 77%, Tesseract 82%)

OCR over the centuries

residual error on 24 herbal texts from 1487 to 1870: individually trained models, RIDGES Corpus (Springmann, Lüdeling, and Schremmer 2015)



Conclusions

- for modern material (even including 19th century Fraktur) the pretrained models of ABBYY, Tesseract and OCRopus give very good results (above 98% character accuracy) (Breuel et al. 2013)
- for older material, missing language models (Latin) and the above-mentioned challenges severely limit the performance of pre-trained models to about 85% (incunables even less); even perfect lexica will raise accuracies to just about 90% (Springmann et al. 2014)
- trained OCRopus models will consistently give > 95% (up to 99%) accuracies depending only on the quality of the scans, not on printing date

References I

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.

Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers.

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.

References II

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.

Strange, Carolyn, Daniel McNamara, Josh Wodak, and Ian Wood. 2014. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers." *Digital Humanities Quarterly* 8 (1). http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.