

# 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)

MEDIE ET INFIMÆ LATINITATIS  
Glossarium a CAROLO DU FRESNE  
DOMINO DU CANGE  
A MONACHIS ORDINIS S. BENEDICTI  
D. P. CARPENTERII  
G. A. L. HENSCHEL  
GLOSSARIUM GALLICUM, TURCICUM, INDIACUM ET ALERI, DISCUSSIONES  
EDITIONIS MOYA nata prolixa unde datur usque  
LUDOVICO PAVRE  
Habentes in Iudicio de Efficiente de Fratre et successore necesse habuit ut Amperius de Patis.



PDF (searchable)

ABACISTA Vide Abacus.#

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

Text

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

<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  
 $n_{GT} = c + s + d$ ,  $n_{OCR} = c + s + i$

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

(often  $n = n_{GT}$  or  $n = n_{OCR}$  - 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:  $N_{\text{pot}} = T_p + F_p = N_{\text{OCR}}$
  - number of good items:  $N_{\text{good}} = T_p + F_n = N_{\text{GT}}$
- **Precision: proportion of good items in retrieved set,  $p = T_p / N_{\text{pot}}$  (Reinheitsgrad)**
- **Recall: proportion of good items retrieved,  $r = T_p / N_{\text{good}}$  (Ausbeute)**

# Precision and recall in OCR

- we have:

- $T_p = c$
- $T_n = 0$  (we want to recognize all items, none are originally bad)
- $N_{GT} = c + s + d$
- $N_{OCR} = c + s + i$

- 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:
  - $F_p = s + i$
  - $F_n = s + d$  (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 differuit:

*Oratio Periclis funebrit.*

**M**ulto quidem corū qui ex hoc hacētū loco uerba fecerunt, hunc legibus institutum morem in concione dicendi ad exequias defunctorum in bello, ut pulchrum laudant. Mihi uero satis esse uisum est, ut uorum præstatiū faciū honores declarare, qualia circa bustum hoc publice infrastructa cōspicis: nec in uno uiro multorum uirtutes perilitari 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, & hominem diligit, aliquid

Kreütter

ner erscheinig / vnserer teütscher zaun oder hagwurzel / gar nicht / welche der mehretheil balbierer für rechte Aristokochiam rotum / dam einsamlend. Diſc. Diſer wurgel etwas mit wein myrrhen vnd pfeffer getrunknen / reiniget die weiber von überflügigem vn- / rath der müter / treide auf die an- / dgeburt vñ weiber menses. Ein / salb gemachte vonn diſer wurgen

**C**Mediolanū igit̄ ciuitas potissima tol-  
us Cisalpine gallicæ Metropolis e yrbiuſ  
ceterarū in: impante Assuero psarū rege  
anno mudi. 4340. e an xpi aduentū 359. a  
gallicis senonēsibus n̄ adua/ v̄m̄l̄ astucie  
voluit: sed aucta e instaurata fuit. Lā enī  
Iosie hebreoz iudicis tēpore a dignissi-  
mis auctoribus primo conditā fuisse me-  
morie proditū ē. Nec certe credidū ē v̄ta  
ferax: tāq̄ opulēta regio v̄q̄s ad Senonē-  
sum galloꝝ tēpora sine urbe extiterit: Lāz  
Bonifacius Utterbičia ep̄s: e Decius au-  
xoniū vir illustris i carbalago nobiliū cui  
ratuſ velint ihām ē Troianoz tēporibus  
clarissimā fuisse. Nā ē Sicābri Hermāic  
populi a Sicābria priami soror dicti: Tro-  
fa cuera Samsonis iudicis tēporibus: oc-  
cupatis Ungarie ac Banarie, ac Banarie, p

nerlen Sache nicht wohl bestehen könne. Da also der Stadt-Schreiber zu Bella in dem Processe zwischen Marco und Julio dem Marco eine Schrift und Deduction aufgesetzet, gleichwohl in eben dieser Sache so wohl vor als nachher registriert, und sich als Actuarium aufgeführt; So gewinnet es das Anheben, ob habe er allerdings ein crimen pravaricationis begangen und einige Straße verdient. Alsdieweil aber kein Gesetze vorhanden ist, welches einem Actario in eben der Sache, worinnen er registriert, Schriften zu versetzen ausdrücklich verbietet,

clockwise: printing year (author)

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

# The challenge (II): special glyphs

Pontanus: *Progymnasmata Latinitatis* (1589)

nis indicium. Quid sequebatur? *S.* De tonis  
seu accentibus nescio quid. *A.* Iam recordor.  
Nosse etiam quo tono, acuto, gravi, in-  
flexovbi vtendum. Adhaec de interpunctio-  
nibus, quæ videlicet nota hypodiastoles dis-  
tingienda, quæ contra per ὑφ' ἐμ' coniun-  
genda, quando demum syllaba porrecta su-  
per se pusilla linea insignienda, quando semi-  
lunula inferiore ad breuitatem indicandam,  
quando comma, quando punctus, quando bi-  
na pucta, seu colon, quando interrogationis  
signum, quando parētheseos nota adhiben-  
da. *S.* Dicebat in orthographia locum esse  
non deriuationibus duntaxat, notationibus,  
sive etymologiis, originibus, sed consuetu-  
dini etiam; videndumque quid solerent eru-  
diti: qui si discreparent, & alij hoc, alij alio  
modo verbum idem scriptitarent, plurium  
valere oportere iudicium. *A.* In reprehen-  
sionem denique vocabat eos, qui cum perui-

historical fonts

long s (ſ)

historical ligatures:  
Æ, æ, œ, ſt, ct

Polytonic Greek words

diacritics

abbreviations

historical spellings

# 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 sūmū bonū et beatitudinē abs-  
qz om̄i deliberatōne vel p̄electōne Vnde dicit au-  
gustinus in soliloquij. Deus quē amat omne qd̄  
amare potest: siue sciens: siue nesciens. Circa neu-  
trā istarū est meritū vel demeritū: quia nec volū-  
tas: virtus em & virtū voluntaria sunt. Volun-  
taria aut̄ diuidit in duas: scilicet amiciciā & con-  
cupiscentiā. Amiciciā diligim⁹ illud quod ppter  
se diligimus. Concupiscentiā vero diligimus illud  
cui bonū volum⁹: scz ad delectandū in eo. vtro-  
qz istorū modorū diligimus deū naturaliē: & ange-  
li etiā in primo statu. Sed diligebat angelus deū  
sup om̄ia amore occupiscentiē: scz in ipso delectan-  
do sup om̄ia. Nec tñ seq̄tur q̄ haberet caritatem  
quia nō diligebat deū ppter ipm deū sed ppter se :

velie nolit aspenc sumu bonu ce beaticuome al?sqzonn veliberaeone velpelec^oneVnoevicicau Aus^mus in soliloquisi^eus que amat omne qv amarc potest:s»uesciens.smenesciens Circa neu era ls^aru esk mencu vet vcmmenturquia ncc volu ras vireus em viciu voluntana (une V^!un-tana auc ouivic in ouas: scilicet amicia Le concupilcencia 5Vmiceria vilizim? illuo quov zpter sevili^imus Concupiscens vcro viliquimus illuo cui bonu volum?:1cz as velec^anvu in co Vtrot-qz is^or^ moyov oiliquimus veu naeuralitRLR an^e li eria in primis l^acu Leo viliquebat an^clus veu sup omia amorc ocupiscentie lcz in ipo vclec^an - vo sup omia Alec cn seueur q? kaberee caritacem quia no viligybavt veu ^fpfer ipm veu seo zx>c se :

An incunabulum printing has special abbreviation signs, e.g. p q p q Q q 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 artificially 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.)

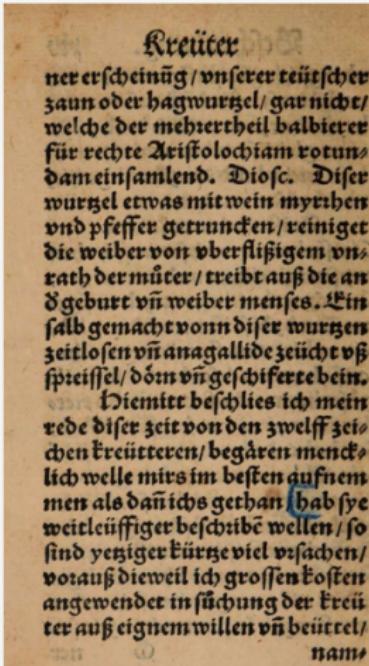
velit nolit appetit sumū bonū et beatitudinē abf :  
qz om̄i deliberatōne vel p̄electōne Vnde dicit au  
gustinus in soliloquijs · Deus quē amat omne qd  
amare potest:sue sciens:sue nefcienſ. Circa neu  
trā iftarū est meritū vel demeritū:quia nec volū  
tas. virtus em & vitiū voluntaria fuit. Volun  
taria aut̄ diuidit̄ in duas: scilicet amiciciā & con  
cupiscentiā. Amicicia diligim⁹ illud quod ppter  
se diligimus. Concupiscētia vero diligimus illud  
cui bonū volum⁹ : scz ad delectandū in eo . vtro  
qz istoꝝ modoꝝ diligimus deū naturalit̄: & ange  
li etiā in primo statu. Sed diligebat angelus deū  
sup om̄ia amore occupiscentie . scz in ip̄o delectan  
do sup om̄ia. Nec tñ seq̄tur q̄ haberet caritatem  
quia nō diligebat deū ppter ip̄m deū sed ppter se :

velit nolit appetit sumū bonū et beatitudinē abf ·  
qz om̄i deliberatōne vel p̄electōne Vnde dicit au  
gustinus in soliloquijs · Deus quē amat omne qd  
amare potest:sue sciens:sue nefcienſ. Circa neu  
trā iftarū est meritū vel demeritū : quia nec volū  
tas · virtus em & vitiū voluntaria fuit · Volune  
taria aut̄ diuidit̄ in duas: scilicet amiciciā & con  
cupiscentiā · Amicicia diligim⁹ illud quod ppter  
se diligimus · Concupiscētia vero diligimus illud  
cui bonū volum⁹ : scz ad delectandū in eo · vtro  
qz istoꝝ modoꝝ diligimus deū naturalit̄: & ange  
li etiā in primo statu · Sed diligebat angelus deū  
sup om̄ia amore occupiscentie · f3 in ip̄o delectan  
do sup om̄ia · Nec tñ seq̄tur q̄ haberet caritatem  
quia nō diligebat deū ppter ip̄m deū sed ppter fe :

- 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	Abbyy FR 11 (84.65%)	OCROpus (98.84%; raw uncorrected output)
 <p>Kreüter      ner erscheinung / vnserer teütscher      zaun oder hagwurzel / gar nicht/      welche der mehrtheil balbierer      für rechte Aristolochiam rotun-      dam einsamlend. Diosc. Diser      wurzel etwas mit wein myrrhen      vnd pfeffer getruncken/ reiniget      die weider von vberflzigem vn-      rath der müter/ treibt aufz die an      d geburt vn weider menses. Ein      salb gemacht vonn diser wurzen      zeitlosen vn anagallide zeücht vñ      spreissel/ dörn vñ geschiferte bein.      Niemitt beschlies ich mein      rede diser zeit von den zwelf zei-      chen kreütteren/ begären menck-      lich welle mirs im besten aufnem      men als dañ ichs gethan hab fye      weitleüfftgerdeschricle wellen / fo      sind yetziger kürze viel vrsachen/      vorauf dieweil ich groskenkosten      angewendet insüchung der kreü-      ter auf eignem willen vñ beütcel/      nam</p>	<p>Kreücev      ner erscheinng / vnserer teütscher      zäun oder hagwurzel/ garnicht/      welche der mehrertheil balbierer      für rechte Aristolochiam rotun-      dam einsamlend. Diosc. Diser      würget etwas mit wein Myrrhen      vnd Pfeffer getryncken/ reiniget      die weider von vberflzigem vn-      rath der müter/ treibt aufz die an      d geburt vn weider menses. Ein      falbgemachtvonn diser würgen      zeitlosenvnanagallidezeüchtvñ      spreissel/ dörnvngeschiferte dein.      iZiemitt deschlies ich mein      rede diser zeitvonden zwelfzei-      chenkreütteren/degären menck-      lich welle mirs im dessen aufnem      men als dañ ichs gethan / had fye      weitleüfftgerdeschricle wellen / fo      sind yetziger kürze viel vrsachen/      vorauf dieweil ich groskenkosten      angewendet insüchung der kreü-      ter auf eignem willen vñ beütcl/      nam</p>	<p>Kreüter      ner erscheinüg/ vnserer teütscher      zaun oder hagwurzel/ gar nicht/      welche der mehrertheil balbierer      für rechte Aristolochiam rotun-      dam einsamlend. Diofc. Difer      wurtzel etwas mit wein myrrhen      vnd pfeffer getruncken/ reiniget      die weider von vberflzigem vn-      rath der müter/ treibt aufz die an      d geburt vn weider menses. Ein      falbgemacht vonn difer wurtzen      zeitlosen vñ anagallide zeücht vñ      spreissel/ dörnvñ geschiferte bein.      Hiemitt beschlies ich mein      rede difer zeit von den zwelf zei-      chen kreütteren/ begären menck-      lich welle mirs im besten aufnem      men als dañ ichs gethan/ hab fye      weitleüfftgerbeschridē wellen/ fo      sind yetziger kürze viel vrsach_en/      vorauf dieweil ich groffen kosten      angewendet in füchung der kreü-      ter auf eignem willen vñ beüt_tel/      ncn:</p>

# Mixed typefaces: old and new methods

Augustinus Leyser (1735)

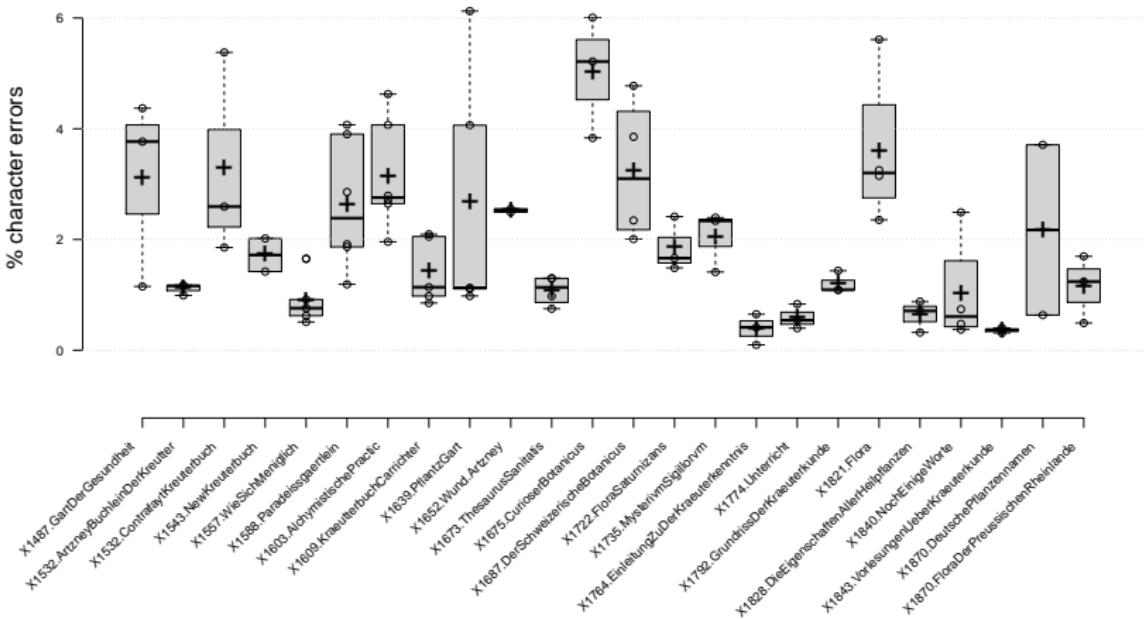
sumque committunt, arg. L. 24. C. de Procuratoribus. Et sic  
Jcti Helmstadienes menfe februario anni clo Is ccXXVIII.  
responderunt: Daferne die fämmtliche Meistere beyder  
Innungen in diese Denunciation oder Klage nicht gewil-  
liget, sondern ein Theil derselben die Klagende davon ab-  
gemahnet, und deshalb von diesen geschimpfet und ge-  
kräncket worden; so hätte den Klagenden nicht gebühret,  
den Namen der fämmtlichen Mettere 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 Procuratoribus. Et sic  
Jcti Helmstadienes menfe februario anni clo Is ccXXVIII.  
responderunt: Daferne die fämmtliche Meistere beyder  
Innungen in diese Denunciation oder Klage nicht gewil-  
liget, sondern ein Theil derselben die Klagende davon ab-  
gemahnet, und deshalb von diesen geschimpfet und ge-  
kräncket worden; so hätte den Klagenden nicht gebühret,  
den Namen der fämmtlichen Mettere unter ihre Klage  
zu setzen, 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](https://doi.org/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>.