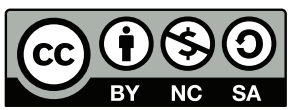
OCR von Inkunabeln: Herausforderungen und Herangehensweisen

Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München
und
Institut für deutsche Sprache und Literatur
Humboldt-Universität zu Berlin



<philtag n="13"/>, Universität Würzburg

2016-02-25

Einleitung

Warum interessieren uns Inkunabeln (Wiegendrucke)?

Gutenberg-Bibel B42, 1454

als die Druckkunst in der Wiege lag
(lat. incunabula, orum n.: Wiege)

- älteste Dokumente der modernen Druckgeschichte (1450-1500)
- Medienrevolution fällt in Umbruchzeit:
  - Vorabend der Reformation (Luther 1517)
  - Entdeckung Amerikas (Kolumbus 1492)
  - Wiederentdeckung der Antike (Renaissance)

Inkunabeln sind spannende Quellen einer Schlüsselperiode Europas!

Herausforderungen

Gestaltung und Typographie

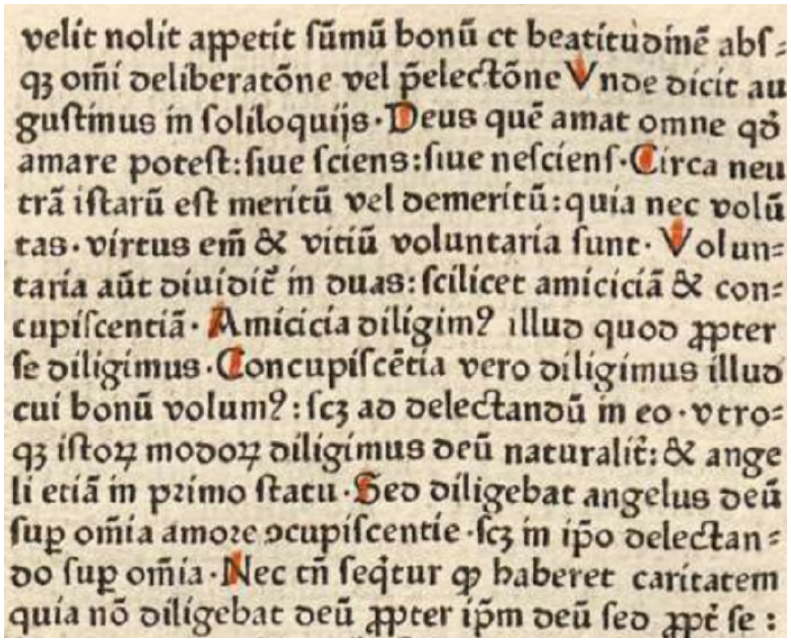
Rothschild-Bibel, 13. Jh., Paris

- absatzmarktorientierter Anschluss an Gestaltung von Manuskripten (illuminiertes Manuskript kostete soviel wie ein Haus)
- Farbe, Zierinitialen, Schmuck, Ligaturen, Sonderzeichen
- Papierbewirtschaftung: Spaltendruck, zahlreiche Abkürzungen
- noch keine Arbeitsteilung und Standardisierung: jeder Frühdrucker war auch Typenschneider und Setzer (Typenvielfalt!)

Hürden für automatische Segmentierung und Transkription!

OCR auf Inkunabeln: status quæstionis (bis 2014))

Beauvais: *Speculum naturale* (nicht nach 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 ꝯ Qꝰ q́ 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."*

Herangehensweisen

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)*

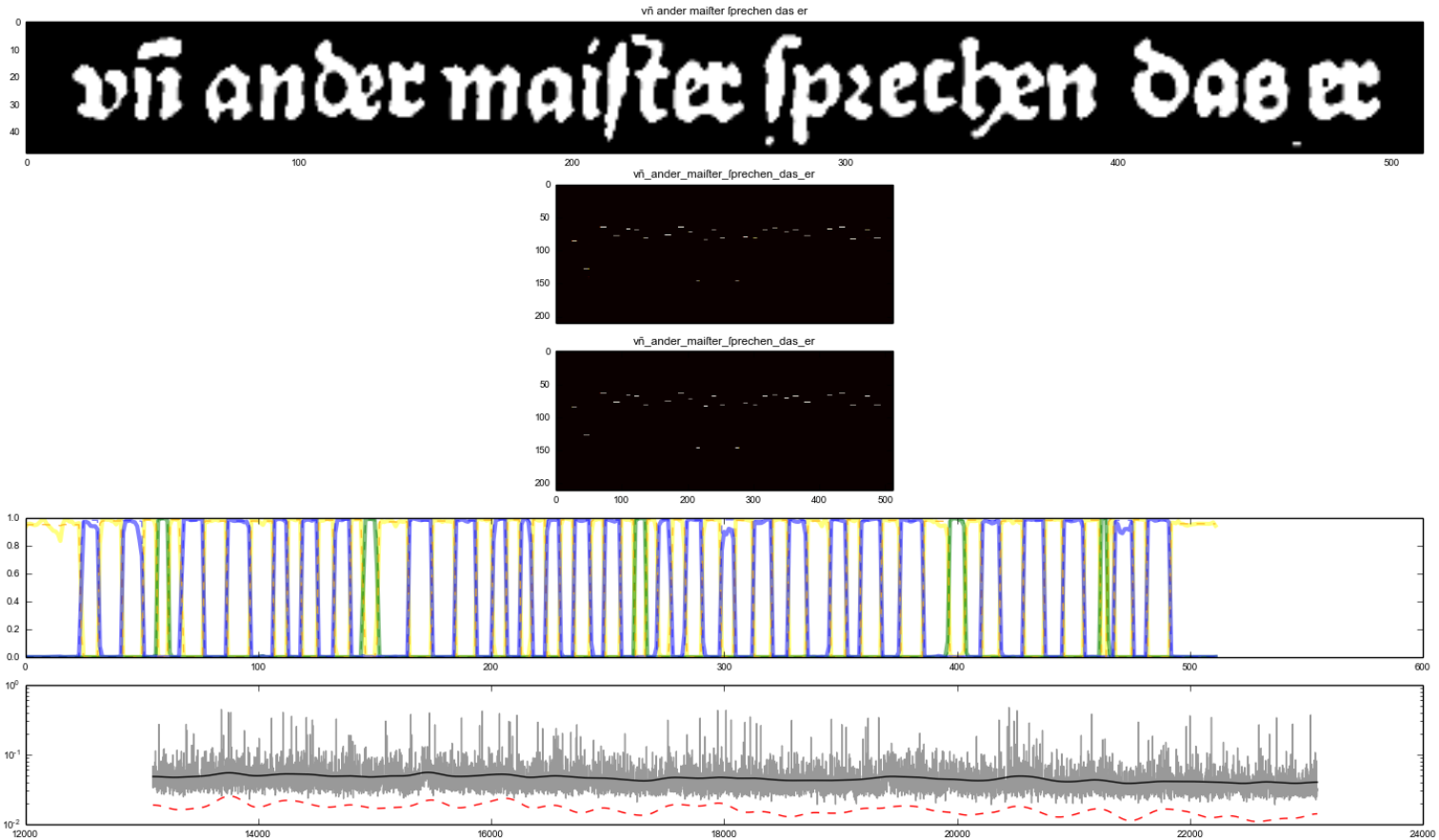
Wie lernt das neuronale Netz?

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!

Dem Netz beim Lernen zuschauen

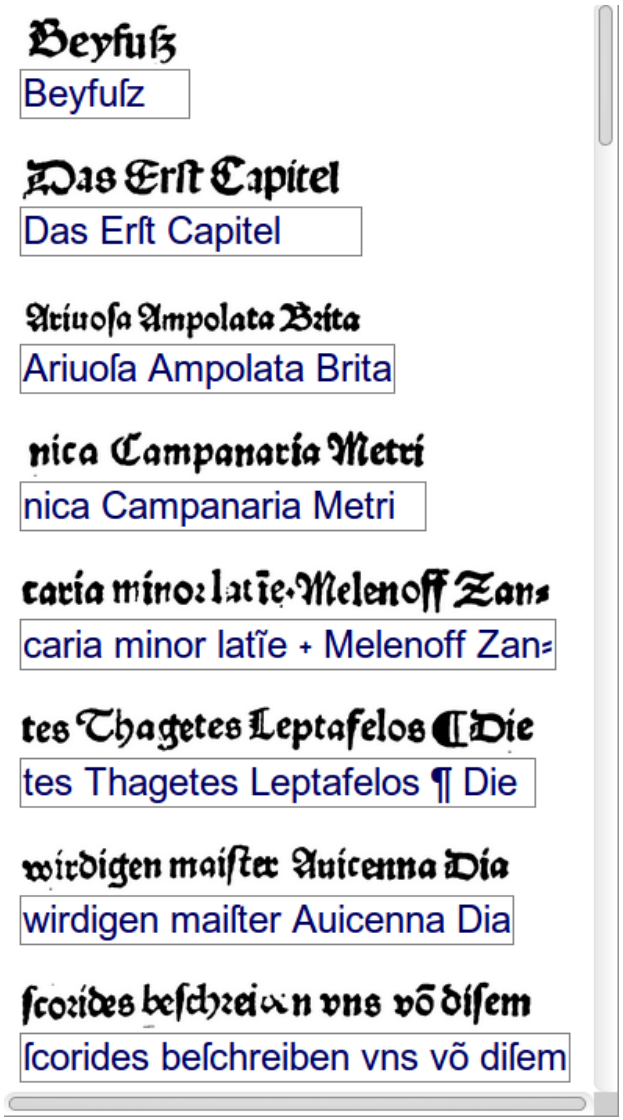


Trainieren eines OCR-Modells auf einer Schriftart

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

Diplomatische Transkription: *ground truth*-Erstellung



- Eingabe über Zeilensynopse im Browser mit geeigneter Schriftart, z.B. Junicode
- Glyph-Repertoire bestimmen (Häberle)
- Paläographie-Kenntnisse notwendig
  - Ligaturen
  - Suspensionen
  - Kontraktionen
- weitere Voraussetzungen:
  - historische Linguistik
  - Schreibvarianten
  - Frühneuhochdeutsch
  - Latein (70% der Inkunabeldrucke)

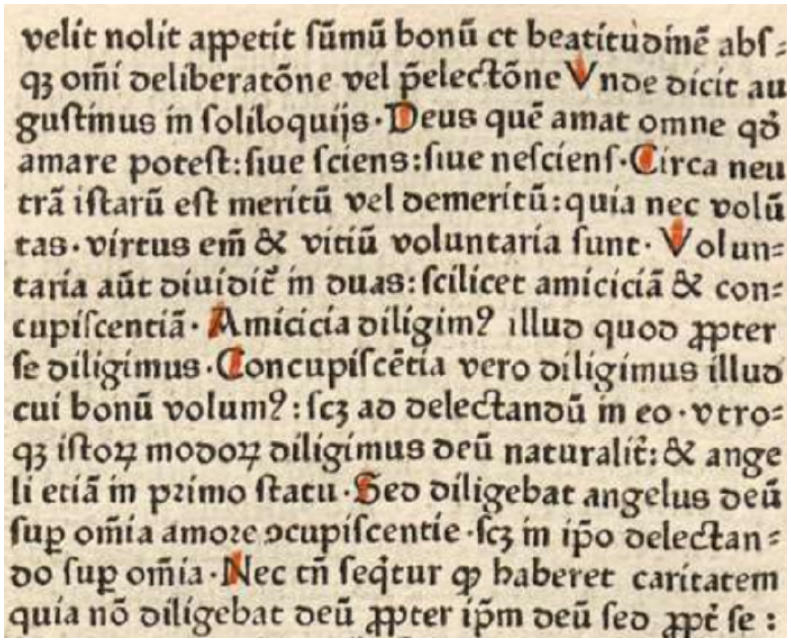
Modelltraining

nach einer Weile (hier: nach 23.055 Lernschritten):

```
23055 1.99 (497, 48) train/0004/01001f.bin.png
   TRU: u'vertreibt die \u017fchlangen die in den'
   ALN: u'vertreibt die \u017fchlangen die in den'
   OUT: u'vertreibt die \u017fchlan gen die in den'
23056 1.42 (508, 48) train/0002/010046.bin.png
   TRU: u'laxieren i\u017ft vnd purgieren / das dz'
   ALN: u'laxieren i\u017ft vnd purgieren / das dz'
   OUT: u'laxieren i\u017ft vnd purgieren / das dz'
23057 2.02 (514, 48) train/0001/01002e.bin.png
   TRU: u'che fraw wee mit ainem kind gat'
   ALN: u'che fraw wee mit ainem kind gat'
   OUT: u'che fra w wee mit ainem kind gat'
```

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 qd
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 diuidit in duas: ſcilicet amicicią & con
cupiſcentią · 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 naturalit: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſuꝑ oñia amore ocupiſcentie · ſcȝ in iꝑo delectan
do ſuꝑ oñia · Nec tñ ſeq̇tur q haberet caritatem
quia nõ diligebat deũ ꝓpter iꝑm deũ ſed ꝓp̄ 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

Offene Fragen

Muss man jede Type separat trainieren?

	1487-Gar	1532-Art	1532-Cor	1543-Nev	1557-Wie	1588-Par	1603-Alc	1609-Kra	1639-Pfla	1652-Wu	1673-The	1675-Cur	1687-Der	1735-My	1764-Einl	1774-Unt	1828-Die	1870-Deu
1487-GartDe	3,522	27,513	29,79	24,506	18,514	27,384	26,632	28,78	23,024	28,351	28,501	24,871	32,71	28,565	33,57	28,479	37,264	48,282
1532-Artzney	17,594	1,109	16,464	13,096	10,335	11,444	9,372	13,452	11,925	12,05	15,879	10,921	17,678	19,937	24,916	24,665	31,318	38,368
1532-Contraf	26,709	27,32	3,671	18,388	18,056	21,727	19,175	27,338	23,72	35,763	35,011	23,422	46,967	29,628	33,298	34,382	46,128	53,994
1543-NewKre	12,091	12,484	13,644	1,853	10,313	14,542	12,952	15,815	12,596	17,705	19,053	16,732	20,251	17,219	25,08	26,745	33,502	37,226
1557-WieSich	13,247	20,786	16,618	14,68	0,903	14,362	17,813	23,334	14,176	17,069	24,821	24,821	27,582	21,184	24,078	19,166	27,874	42,952
1588-Paradei	32,801	35,505	32,698	36,071	29,016	2,652	14,779	17,173	10,736	10,273	12,796	10,891	16,89	20,881	26,571	23,532	29,763	37,204
1603-Alchym	25,595	22,794	20,53	26,551	24,925	10,013	2,986	8,017	7,43	13,418	12,965	11,338	18,719	14,24	25,948	21,419	38,007	33,814
1609-Kraeutt	19,321	22,935	17,503	20,504	16,583	5,476	3,023	1,402	6,221	7,514	7,426	6,55	9,398	12,574	18,905	18,335	26,287	27,952
1639-Pflantz	21,093	23,525	23,426	24,808	19,01	8,181	7,348	10,113	2,416	4,898	9,563	6,548	13,512	13,112	18,077	13,545	24,442	28,607
1652-Wund-A	30,476	26,692	32,772	44,374	24,483	8,573	13,756	16,173	5,631	2,515	10,202	8,398	12,761	16,271	21,432	17,321	26,9	32,881
1673-Thesau	30,476	25,833	29,868	32,096	27,118	8,711	10,739	10,458	8,41	7,367	1,044	6,042	9,494	11,642	14,151	17,623	23,906	26,716
1675-Curiose	30,476	29,105	28,148	25,741	23,21	10,988	16,265	17,037	7,623	10,278	14,63	5	21,605	20	22,377	19,29	26,204	32,006
1687-DerSch	30,476	30,188	29,662	43,571	34,436	16,617	14,135	13,459	12,068	10,15	7,481	5,414	3,233	15,977	15,188	17,82	24,248	29,135
1735-Mysteri	30,476	24,794	21,765	23,214	23,872	12,315	12,94	15,608	7,935	10,998	9,121	8,133	13,006	2,14	9,055	8,331	15,838	15,904
1764-Einleitu	30,476	27,075	24,127	29,428	23,869	7,422	12,982	13,24	6,956	7,448	5,017	2,974	8,379	5,094	0,362	3,698	6,206	10,654
1774-Unterri	30,476	24,019	17,813	24,145	15,829	6,754	7,767	10,426	5,572	6,036	9,329	3,926	9,118	4,052	4,137	0,549	6,079	14,141
1828-DieEige	30,476	28,773	22,458	31,84	23,063	13,943	18,805	21,57	11,481	11,461	13,963	6,174	15,133	8,717	6,174	7,425	0,605	8,535
1870-Deutscl	30,476	35,064	29,749	41,771	33,245	18,299	22,429	21,407	15,907	16,152	14,803	11,429	17,134	15,457	12,86	13,883	13,126	1,738

Frakturmodelle angewendet auf Frakturdrucke des RIDGES-Korpus. Die Zahlen geben den Fehler der OCR-Erkennung auf Zeichenebene an (Prozentsatz falsch erkannter Zeichen).

- Spalten: Modelle
- Zeilen: Drucke
- beste Ergebnisse auf Diagonale (Modell passt zum Druck)
- Nichtdiagonalelemente: bei Frühdrucken starke Variation, später weniger
- keine systematische Untersuchung, durch verfügbares Material begrenzt
- Derzeit mangels Modellvielfalt / Trainingsdatenknappheit noch keine klare Aussage möglich!

Jenseits der OCR-Erkennung: Normalisierung

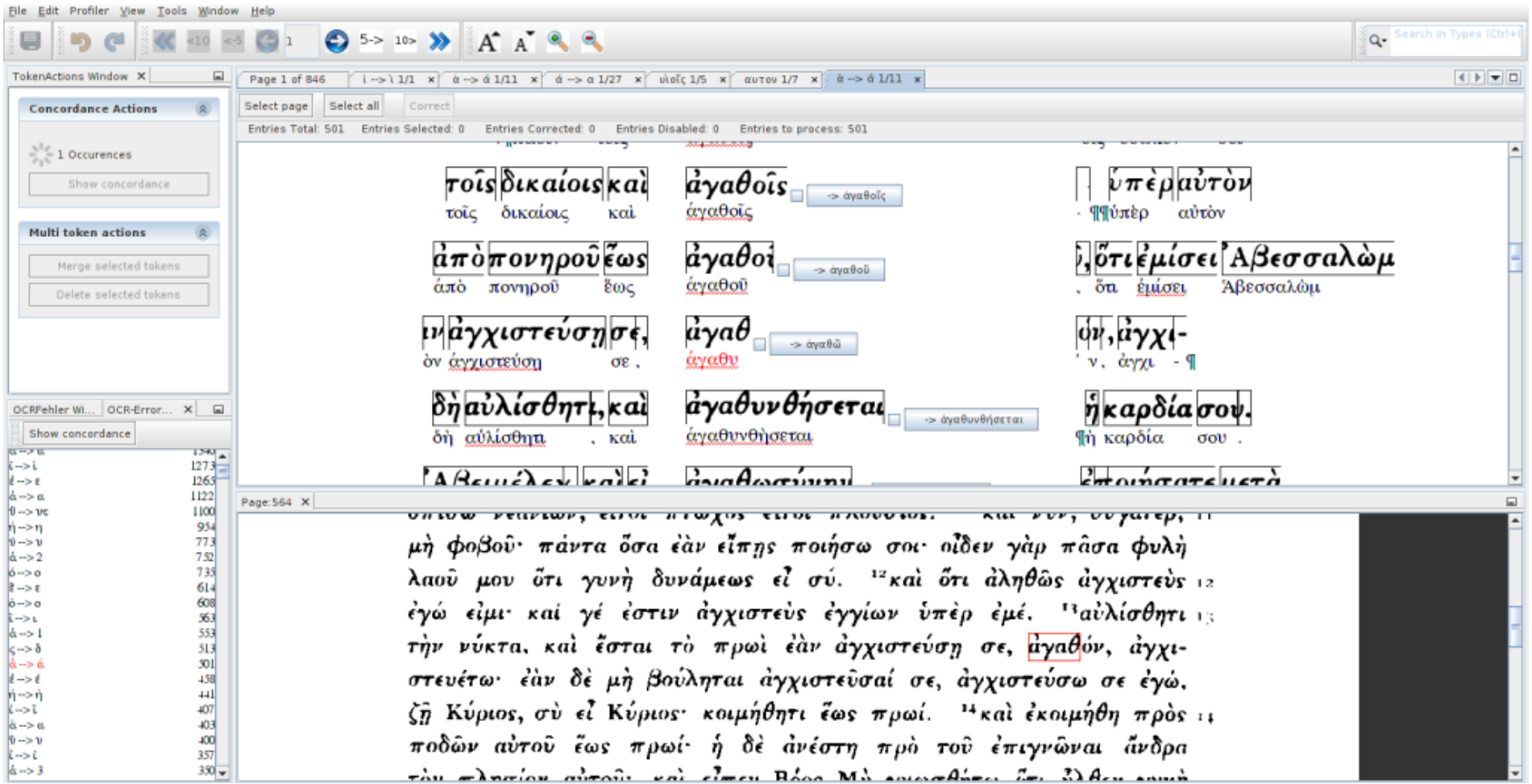
- Wie sucht man auf diesem Text?

  Von di
  ſem kraut beſchreibt vns Diaſco-
  rides vnd ſprichet + das diſes kraut
  beneme vnnd haile acrocordines
  das ſind lychdorn od wårczẽ auff

- Normalisierung auf heutige Schreibweise als Annotationsbeene 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



Vorschläge für ein koordiniertes Vorgehen

- Verfügbarmachung der vorhandenen Scans von Buchseiten in hochaufgelöster Form (nicht binarisierte tiffs anstelle binarisierter jpgs) für die Wissenschaft
- Akquisition eines Budgets zur Aufwertung der vorhanden Scanbestände durch OCR (für einen Bruchteil des für Scans ausgegebenen Budgets gewinnt man ein Mehrfaches an Nachnutzungspotential)
- Einrichtung eines gemeinsamen Daten-Repositories für OCR-Daten, die mit anderen bestandshaltenden Institutionen geteilt und zentral über eine Webschnittstelle zur Nachkorrekur angeboten werden (auch für *crowd-sourcing* geeignet)
- Zentrales Modelltraining in einem leistungsfähigen Rechenzentrum sowohl auf Einzeltypographien als auch für Typenmischungen, um die Einsetzbarkeit der Modelle zu verbessern
- Koordiniertes Vorgehen anstatt jeder für sich dasselbe machen (wie bei Scans)
- Ein solches Projekt würde eine weltweit einmalige Forschungsgrundlage auf Basis der Drucke aus deutschen Gebieten darstellen (einschließlich der in deutschen Gebieten gedruckten lateinischen etc. Werke)

Fazit

- 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

- CIS OCR Workshop (Springmann and Fink 2016)
- 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

Literaturangaben II

Aegyptiae. Berlin, Germany: Berlin-Brandenburgische Akademie der Wissenschaften. http://edoc.bbaw.de/frontdoor.php?source_opus=2443.

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/.

Literaturangaben III

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.