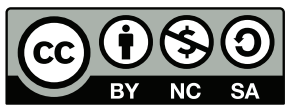
# Gemischte OCR-Modelle für die Erkennung gedruckter Texte seit Gutenberg

Uwe Springmann

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




<philtag n="14"/>, Universität Würzburg

2017-03-17

# Was wollen wir mit OCR erreichen?

- Motivation:
  - Erhöhung der Textverfügbarkeit (überall, jederzeit)
  - Mehrwert gegenüber Papier: auf elektronischem Text kann man suchen; indexieren; annotieren
- 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
- schneller und billiger als manuelle Transkription
- Zielkonflikt: *Masse* (automatisch, geringere Qualität) versus *Klasse* (manueller Eingriff, höchste erreichbare Qualität)

# Die Ausgangslage: 567 Jahre moderne Druckgeschichte

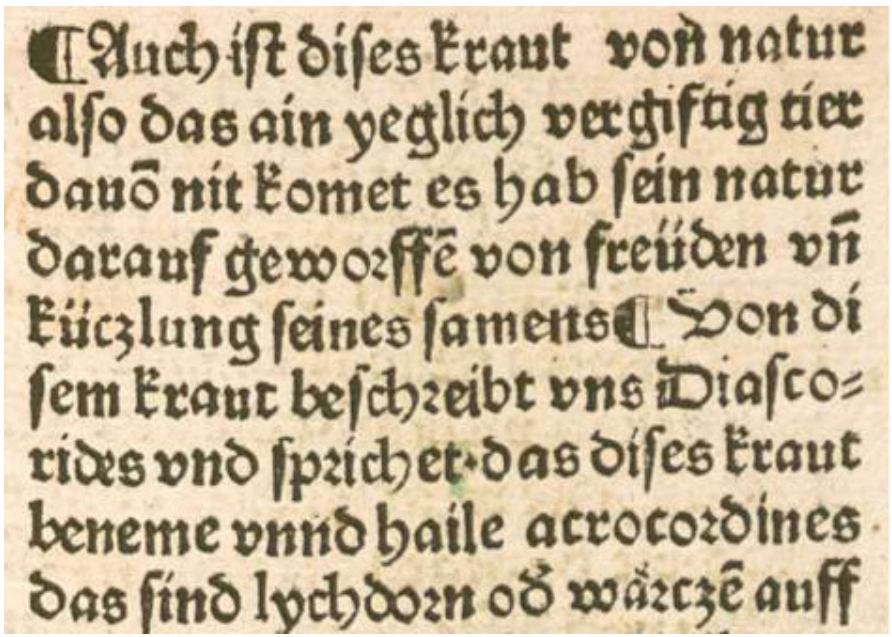
- moderne Druckgeschichte: seit Gutenberg (1450)
- Inkunabeln: 1450-1500, ca. 30.000 Titel, davon 70% auf Latein
- VD16-18: *Verzeichnisse der im deutschen Sprachraum erschienenen Drucke (16. bis 18. Jahrhundert*, etwa 50% der Titel auf Latein)

| 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 mit OCR zu erfassen
- heute schon 70 - 100 Millionen Seiten gescannt (OCR-ready bis auf Strukturerkennung)

# Individuell trainierte Modelle erreichen gute Qualität

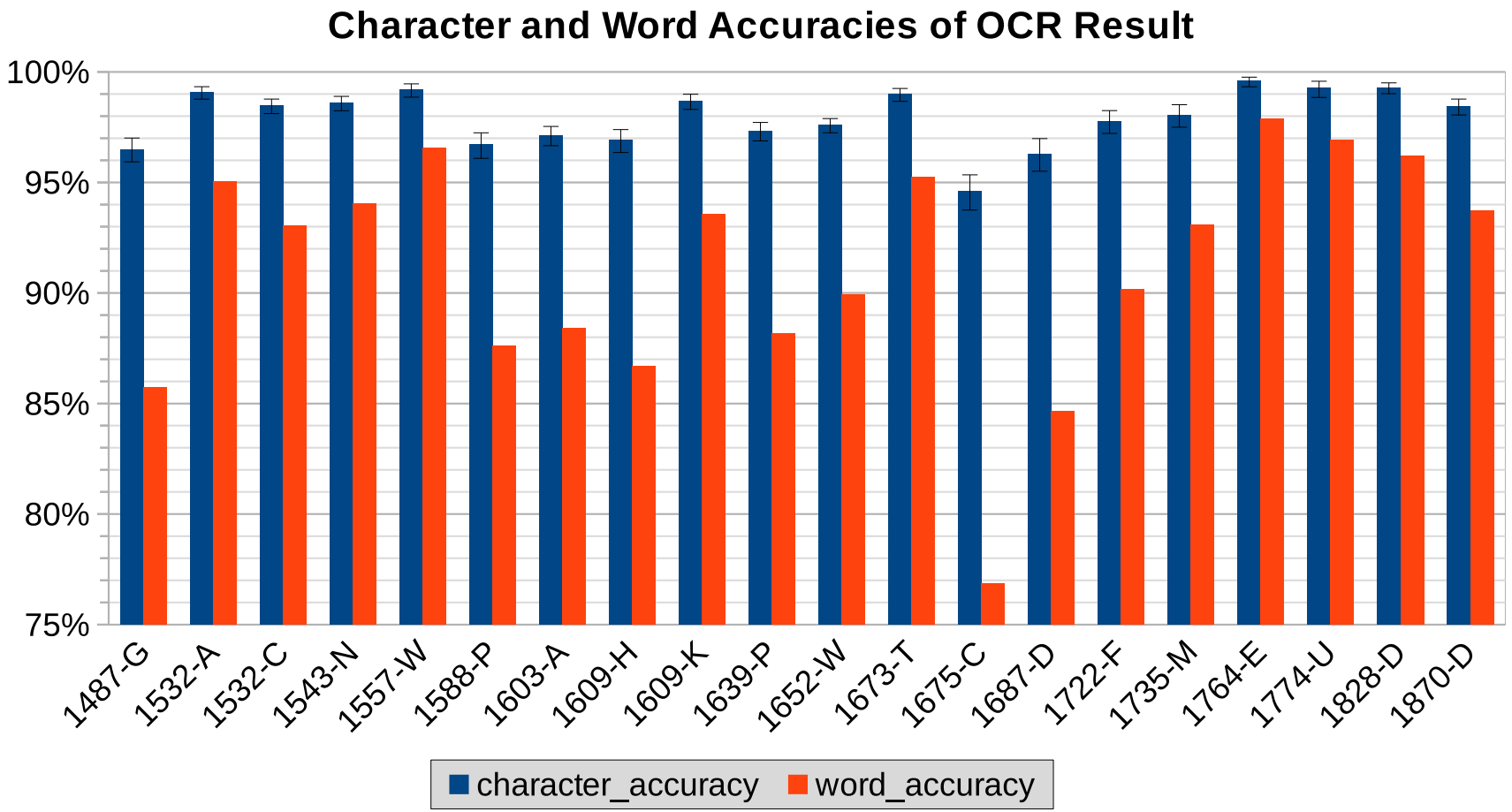
Gart der Gesundheit, 1497, OCRopus-Modell




¶ Auch iſt diſea kraut von natur
alſo das ain yeglich vergiftig tier
dauõ nit komet es hab ſein natur
darauf geworffẽ von freüden vñ
küczlung ſeines ſamens¶ Von di
ſem kraut beſchreibt vns Diaſco=
rides vnd ſprichet + das diſes kraut
beneme vnnd haile atrocordines
das ſind hychdorn od warczẽ auff

# Die OCR-Qualität hängt nicht vom Alter des Drucks ab

deutsche Frakturdrucke aus 4 Jahrhunderten (RIDGES-Korpus, HU Berlin)



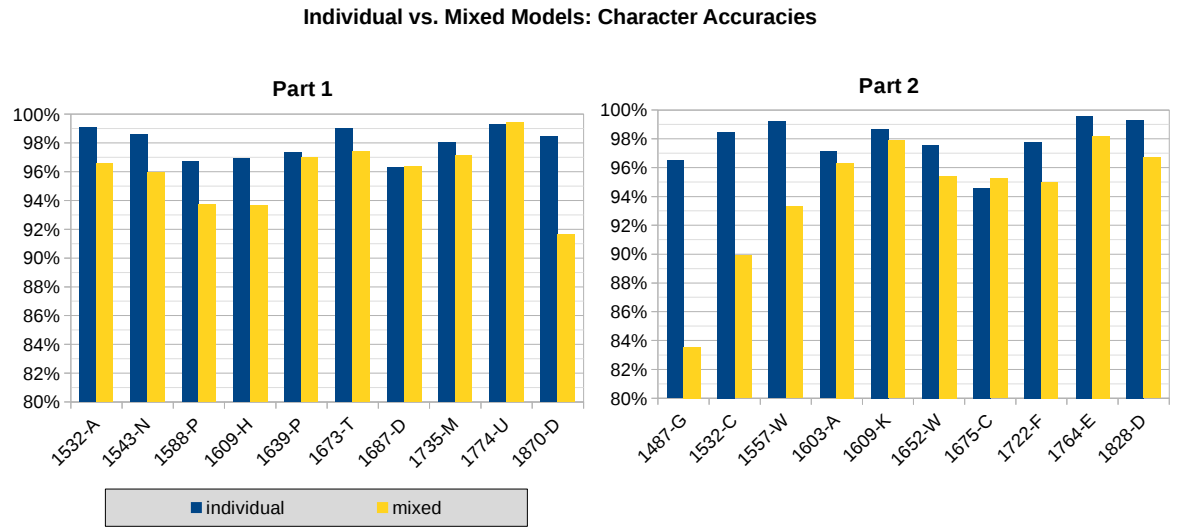
# Individuelle Modelle verallgemeinern schlecht

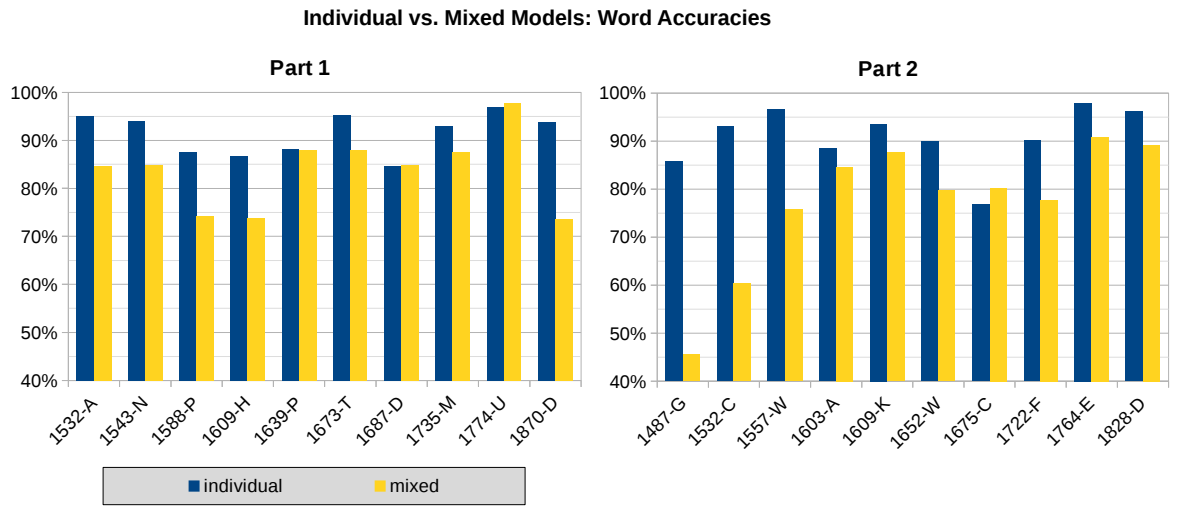
Zeichenerkennungsraten. (Zeilen: Drucke; Spalten: individuell trainierte Modelle)

| | 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** |

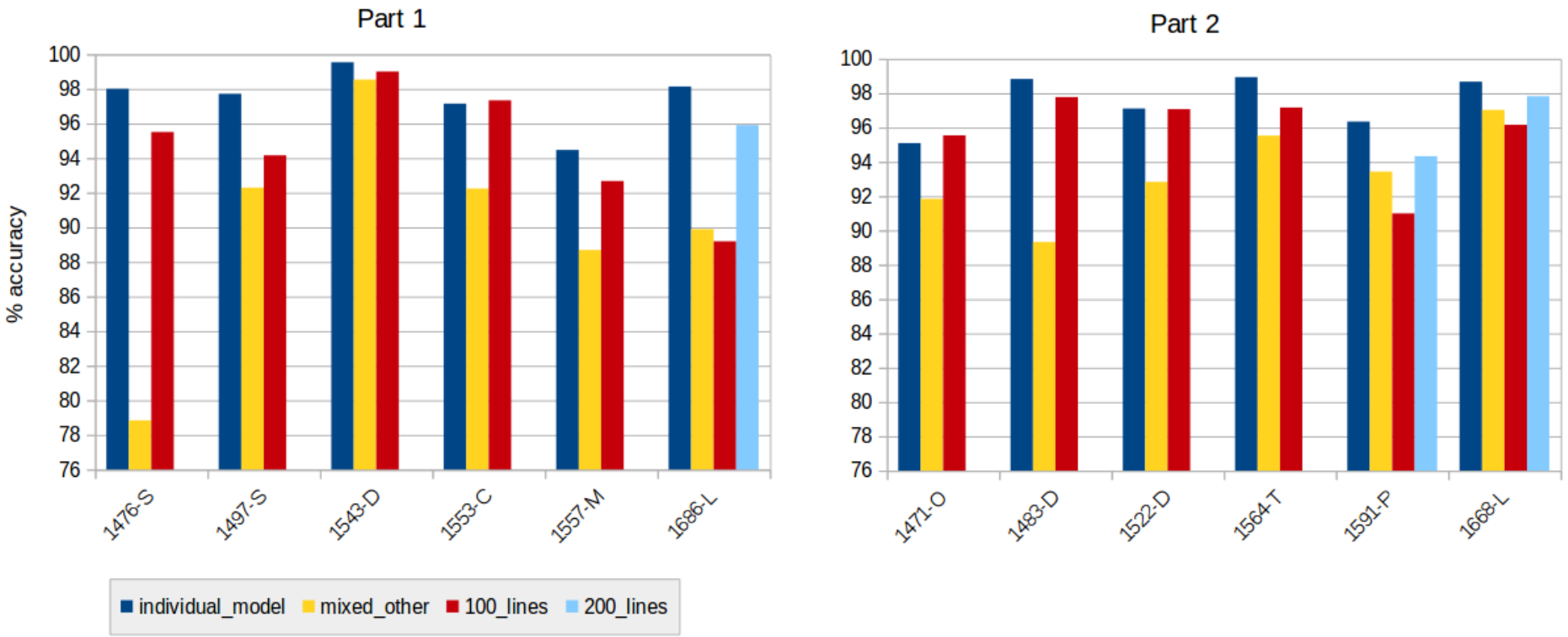
# *Gemischte Modelle* verallgemeinern besser

Vergleich der Erkennungsraten von individuellen und gemischten Modellen



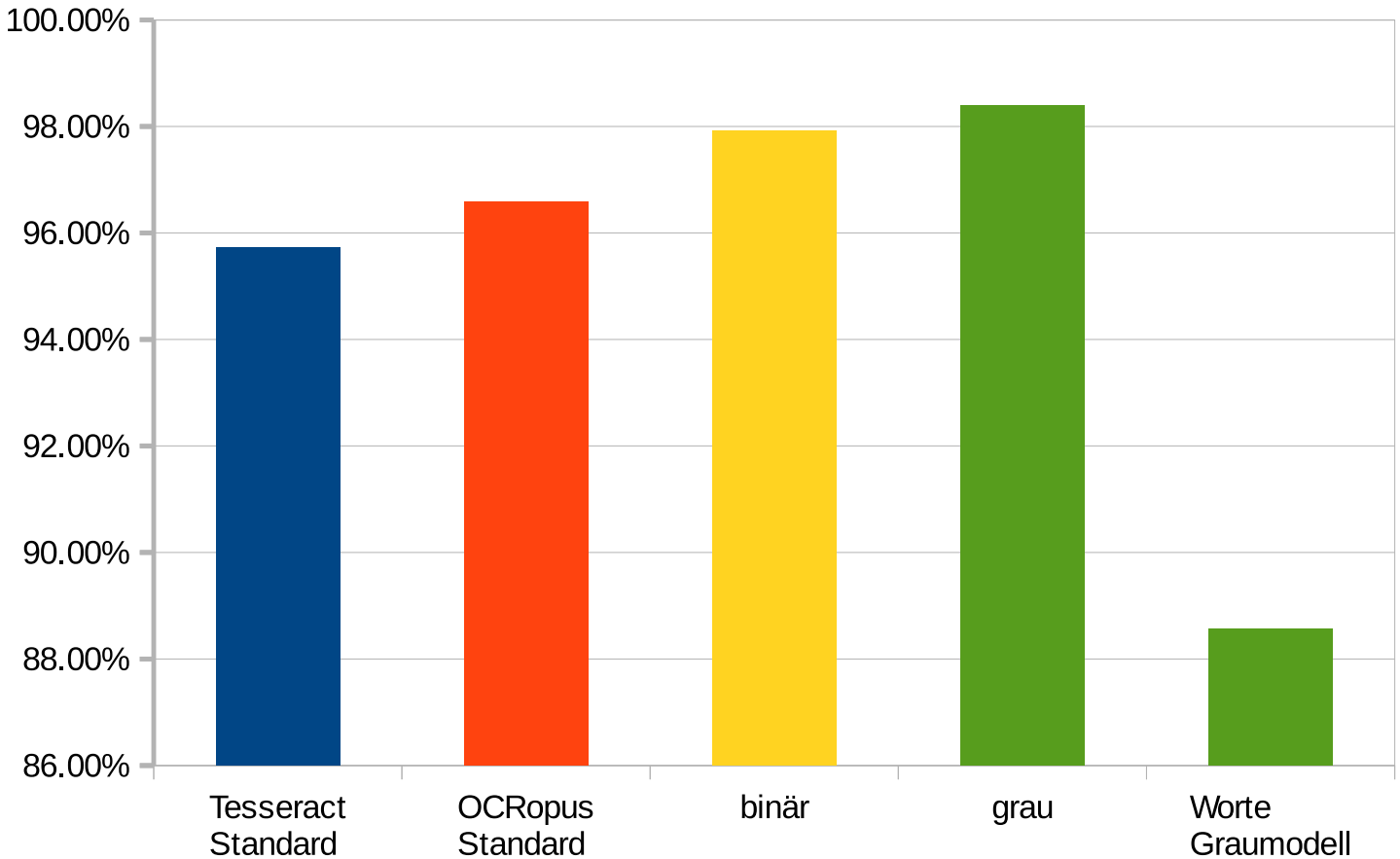



# Dasselbe Bild ergibt sich für Antiqua-Drucke (Latein)



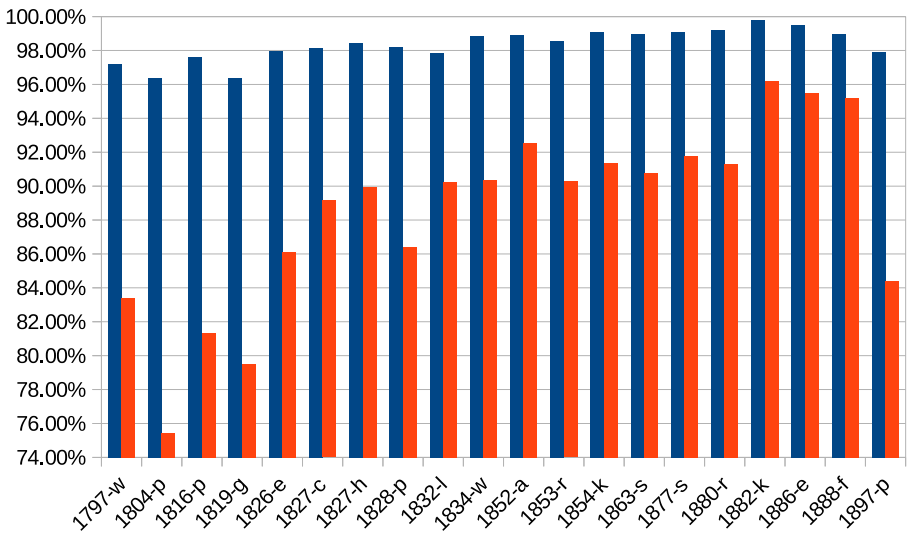
# Deutsche Frakturdrucke des 19. Jahrhunderts

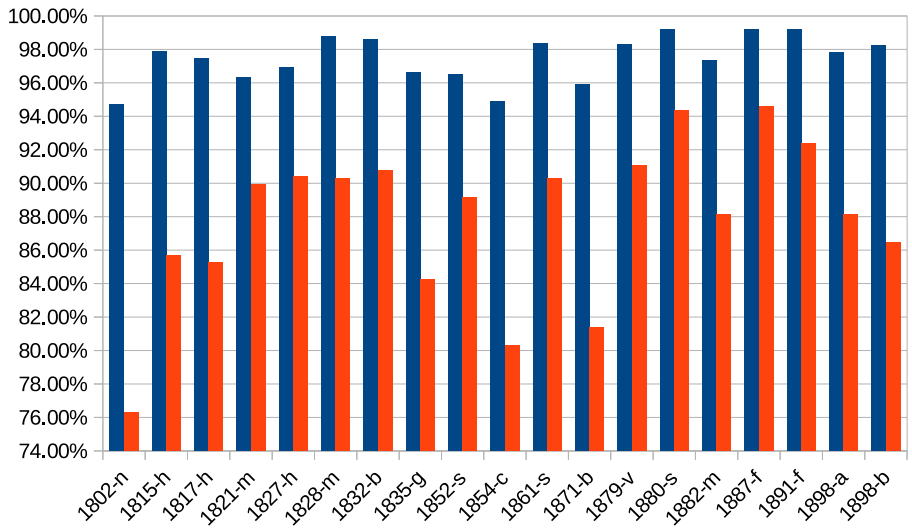
Vergleich von Standardmodellen (synthetisch trainiert) mit gemischten Modellen (39 Frakturdrucke, auf echten Drucken trainiert)



# Individuelle Erkennungsraten mit *gemischten Modellen*

blau: Zeichen; rot: Wörter

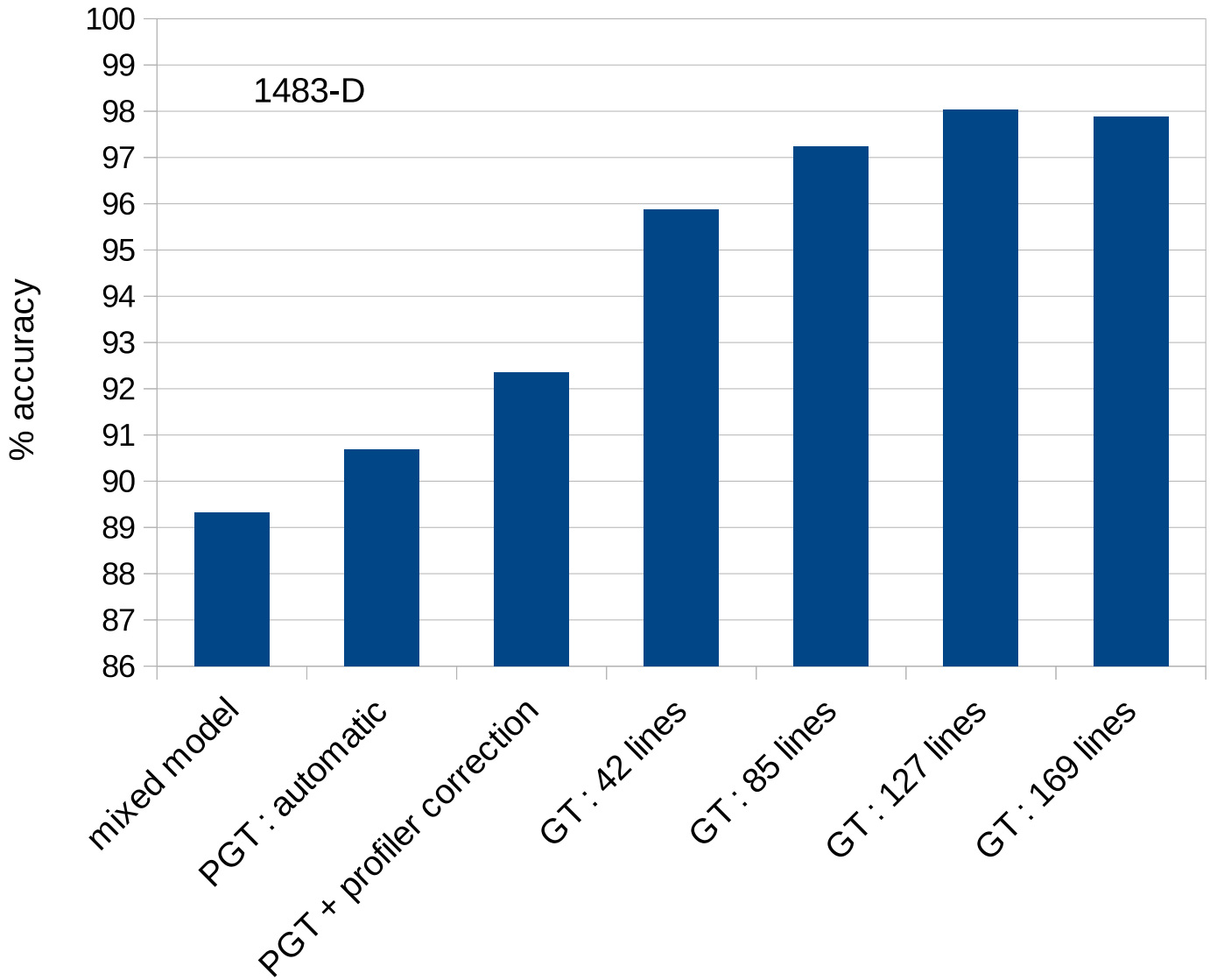


# Zielkonflikt Masse versus Klasse?

- 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*)



# 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)

# Anhang

# Änderungen bei Tesseract

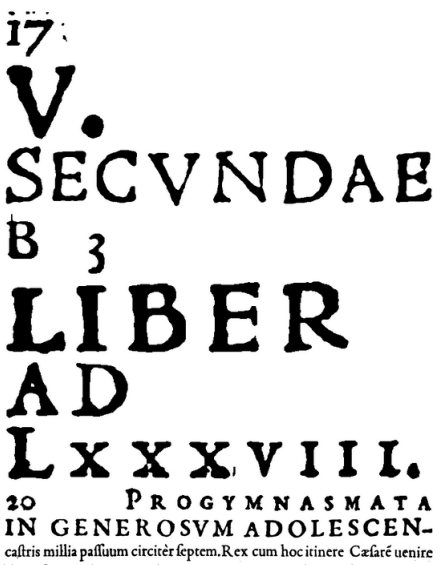
- erhebliche Verbesserungen (30%-40%) durch Erkenner auf Basis neuronaler Netze (LSTM)
- der alte Classifier-Code (Basis der Würzburger Modelle vom letzten Jahr) fliegt raus
- derzeit keine Frakturmodelle für den neuen Code
- bisher nur synthetisches Training: Erzeugung künstlicher Bilder von Druckseiten mit vorgegebenem Text, Trainingsmöglichkeit auf historischen Drucken unklar

# Neuer Classifier immer besser als alter?

theraysmith commented Feb 7, 2017

> *Please provide examples of where you get better results with the old engine.* ***Right now I'm trying to work on getting rid of redundant code,*** *rather than spending time fighting needless changes that generate a lot of work. [ … ]. AFAICT, apart from the equation detector,* ***the old classifier is now redundant.***

Beispiele aus der Praxis:


17
V.
SECVNDAE
B 3
LIBER
AD
LXXXVIII.
20 PROGYMNASMATA
IN GENEROSVM ADOLESCEN-
castris millia passuum circitèr septem. Rex cum hoc itinere Cæsaré uenire

# Test alt gegen neu

**Old method: tesseract -l lat –oem 0 –psm 7:**

```
17:
V.
SECVNDAE
B 3
LIBER
AD
Lxxxvxn.
zo PROGYMNASMATA
IN GENEROSVM ADOLESCEN-
```

**New method: tesseract -l lat –oem 1 –psm 7:**

```
177:
V,.
SECV NDAHE
B- 5
LI B E D.
A D
Lx x XV II IL.
209 P R o cy M N ^ s M A T 4
IN GE NE R O SVM A D O L E S CE N-
```

# Aber:

theraysmith commented Mar 8, 2017

> … *yes* ***I would still like to remove the old classifier*** *and take out a lot of code with it. I'm going to review the replies to my request for "old better than new", and thanks to those that provided them, with a view to making new better than old on those problems.*

# Vielen Dank für Ihre Aufmerksamkeit!

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