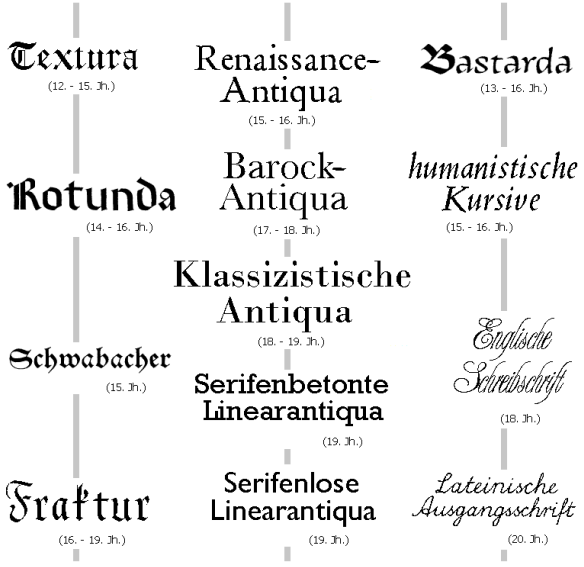
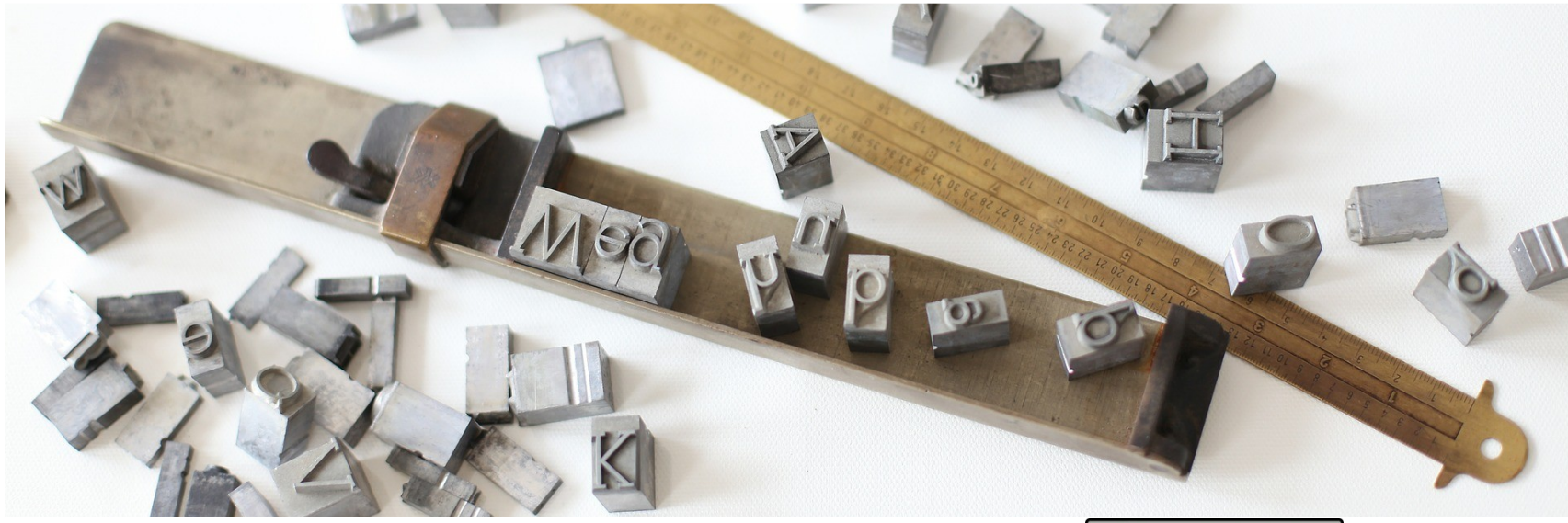
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UB
UNIVERSITÄTSBIBLIOTHEK

# Neue Modelle
## dank GT-Aufwertung und Anreicherung



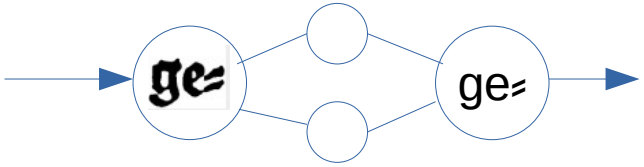
Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST



09.06.2021

# In Kürze: Was sind Tesseract-Modelle?

- Modelle sind trainierte künstliche neuronale Netze
- Künstliche neuronale Netze bilden Eingabewerte auf Ausgabewerte mit Hilfe von Knoten und gewichteten Knotenverbindungen ab
- Training: Optimierung der Gewichte, um die Eingabewerte möglichst gut auf die Ausgabewerte abzubilden
- Trainingsdaten: Zeilenbilder (Eingabewert) und Texte (Ausgabewert)




- Die Qualität und der Umfang der Trainingsdaten bestimmen die Leistungsfähigkeit des Netzes maßgeblich

- Tesseract verwendet Modelle, um in Digitalisaten bestimmte Sprachen/Schriftsysteme zu erkennen

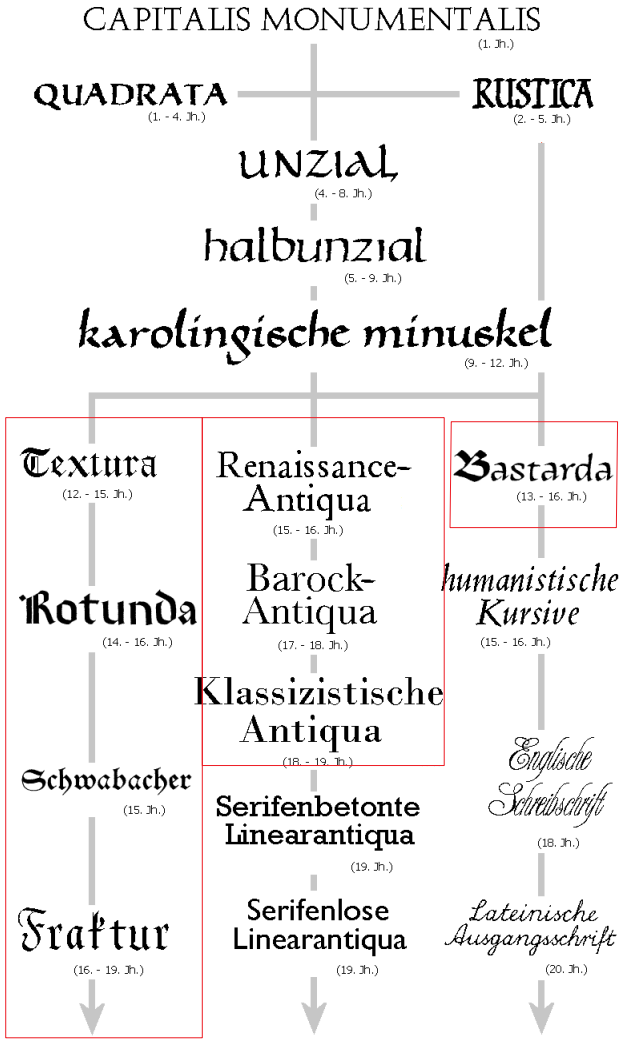
# In Kürze: Was ist Aufwertung und Korrektur von Ground Truth?

- Durch die Aufwertung können GT-Datensätze auf ein höheres Transkriptionslevel angehoben werden
  - Bsp.: **OCR-D Level 1** auf **OCR-D Level 2** durch:
    - Umwandlung des **sz** zu **ß**
    - Umwandlung des **runden s** zu **langen s**
    - …

**Gefahr**: Regelbasierte Umwandlung kann nur gewisse Fälle abdecken!

- Die Vereinheitlichung der Normalisierung
- Korrektur von Abschreibfehlern
- Entfernung von nicht konformen oder fehlerhaften Bild-Text Dateien

# Training neuer Fraktur-Modelle



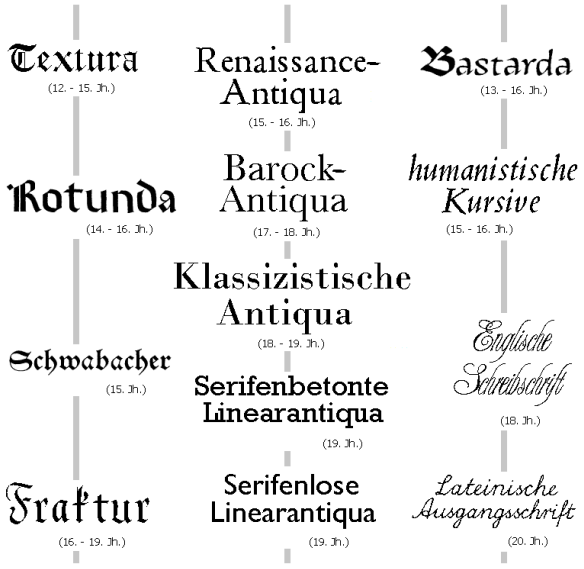

https://typo-info.de/entwicklung-der-schrift/

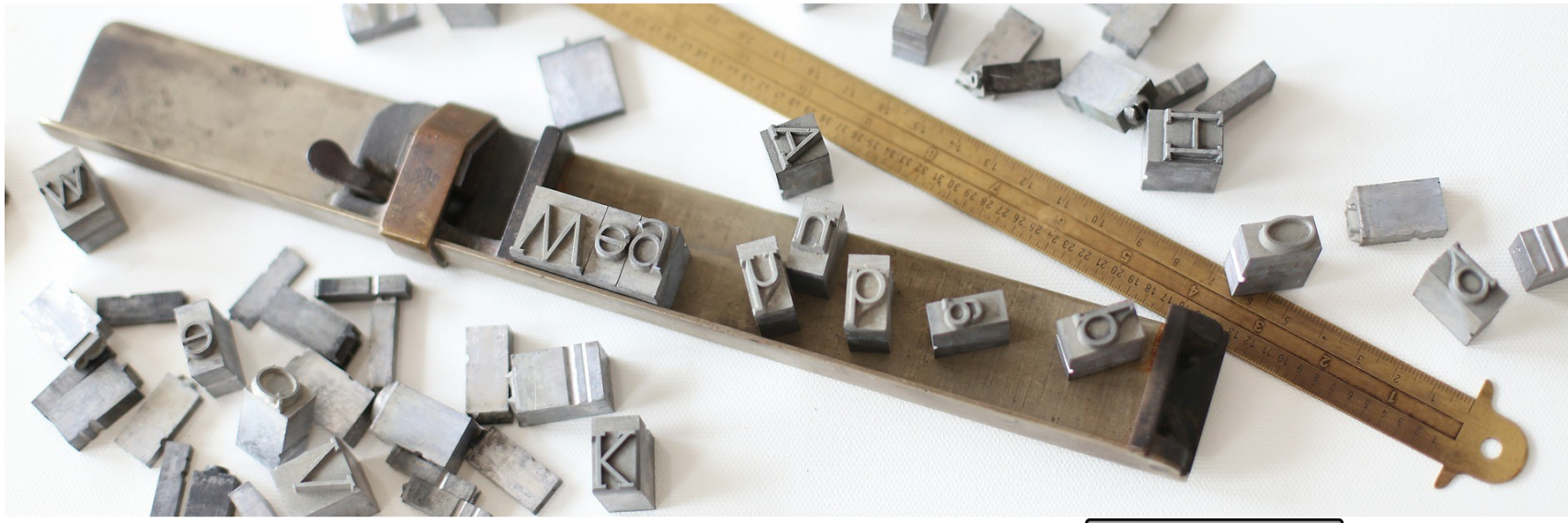
Neue generische Modelle
sind sprachenunabhängig und
decken auch viele Schriftarten ab,
ebenso wie **fett** und *kursiv* gedruckte Variationen.
Erleichtert Volltexterkennung für typische Druckwerke.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UB
UNIVERSITÄTSBIBLIOTHEK

# Neues Modell (2019): GT4HistOCR



Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST



09.06.2021

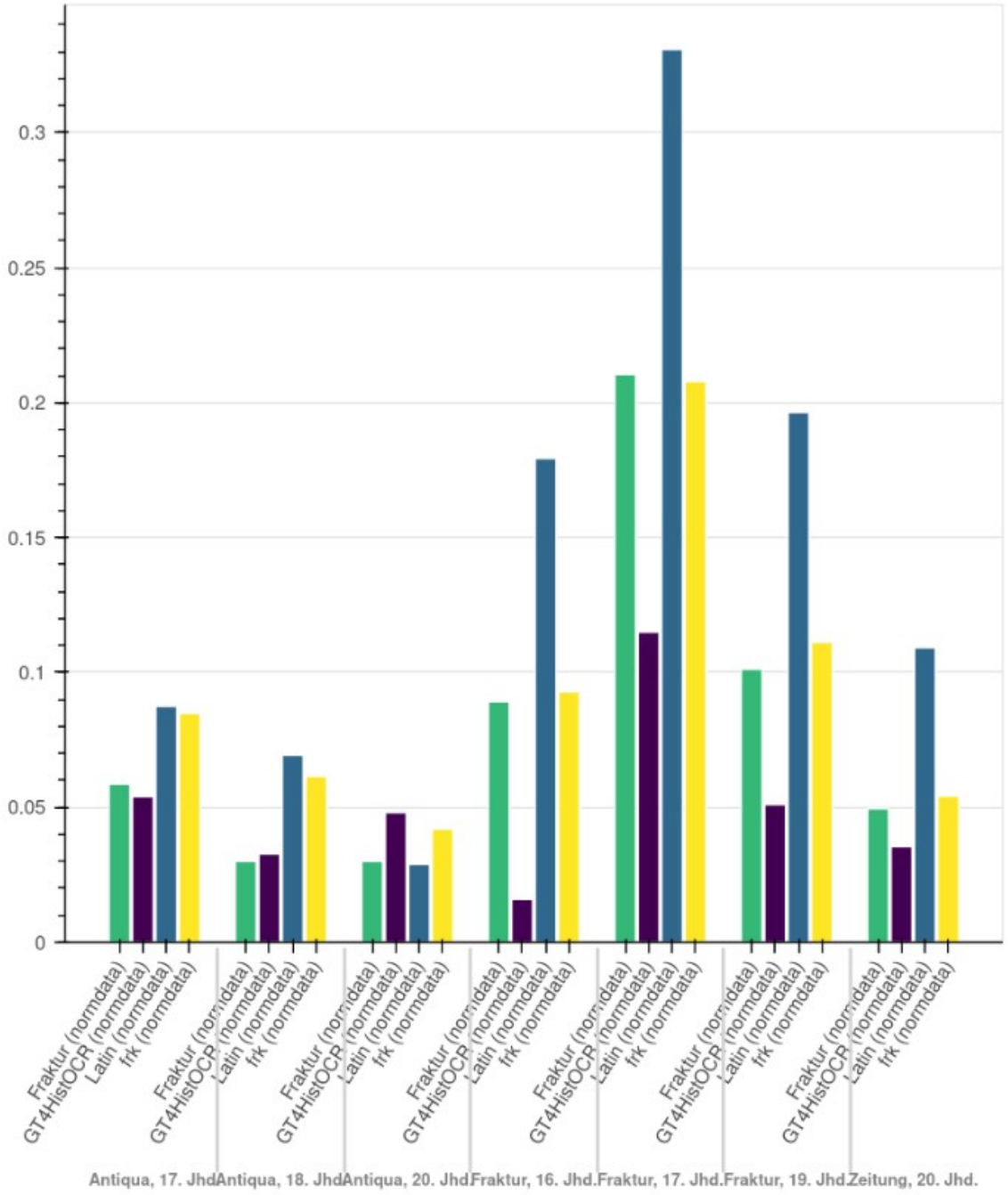
# Verwendete Datensätze

- GT4HistOCR (Originaldatensatz: https://zenodo.org/record/1344132#.YL90dDqxVH4)
  - Datensätze
    - EarlyModernLatin
    - Kallimachos
    - RIDGES-Fraktur
    - RefCorpus-ENHG-Incunabeln
    - dta19
  - Über 50 verschiedene Druckwerke aus dem 15.–19. Jahrhundert
  - Unterschiedliche Schriftarten und Sprachen
  - ca. 313.173 Zeilen

# CER

- Antiqua, 17. Jhd. (Französisch)
- Antiqua, 18. Jhd. (Französisch)
- Antiqua, Anfang 20. Jhd.
- Fraktur, 16. Jhd.
- Fraktur, 17. Jhd.
- Fraktur, 19. Jhd.
- Zeitung, 20. Jhd.

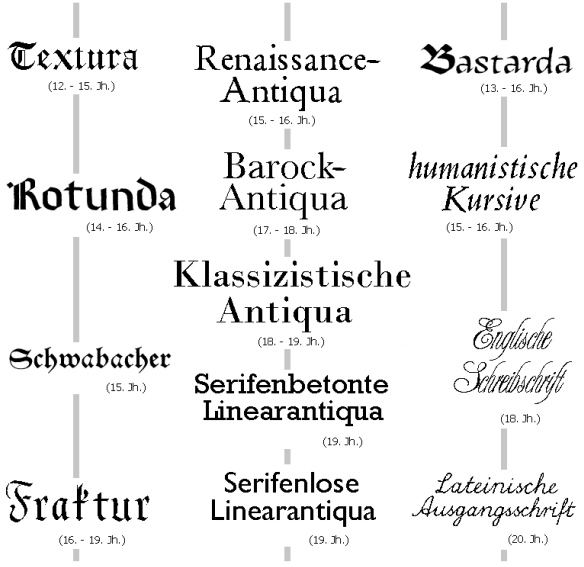
- Fraktur
- GT4HistOCR
- Latin
- frk

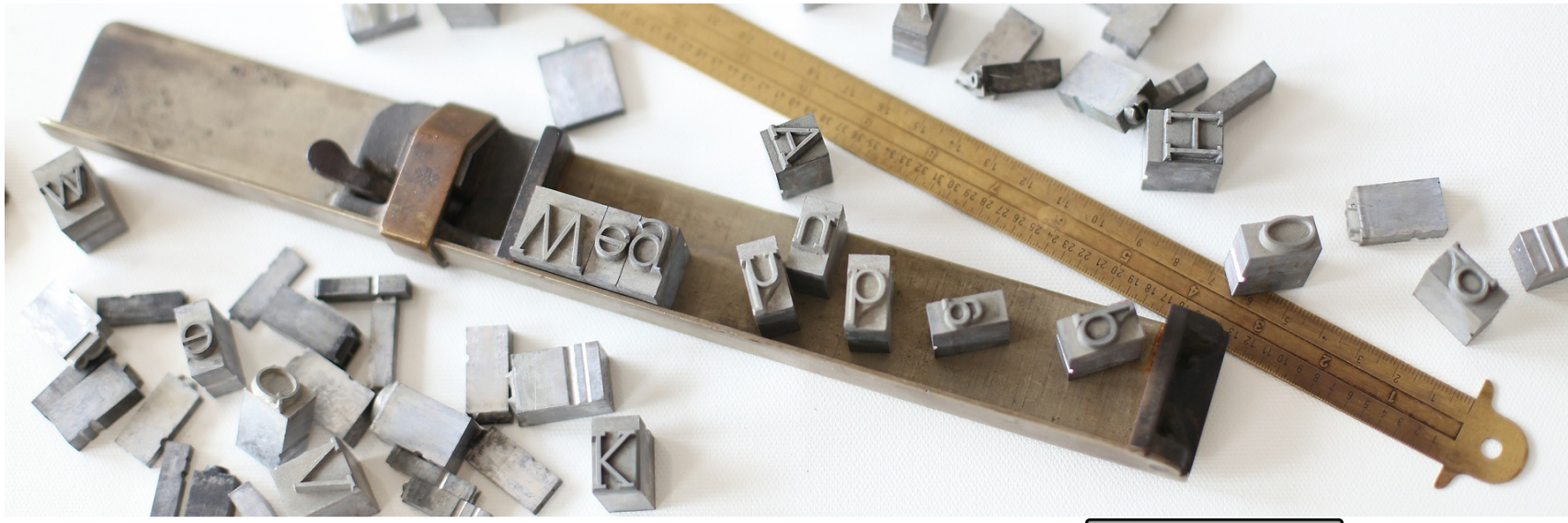


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UB
UNIVERSITÄTSBIBLIOTHEK

# Neues Modell (2021): frak2021



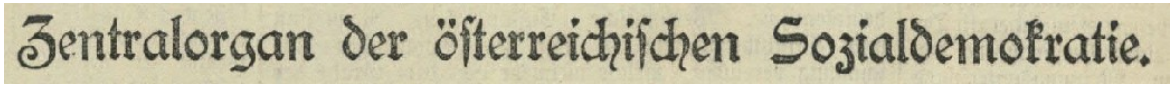
Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST



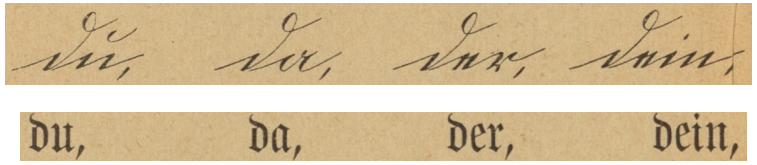
09.06.2021

# Anreicherung der Trainingsdaten

- GT4HistOCR (Originaldatensatz: https://zenodo.org/record/1344132#.YL90dDqxVH4)
- AustrianNewspapers (https://github.com/UB-Mannheim/AustrianNewspapers)
  - Transkription: Österreichische Nationalbibliothek (wurden für das Modell *ONB_Newseye_GT_M1+* für Transkribus verwendet)




  - Umfang: 57541 Zeilen
  - Inhalt: Deutschsprachige Zeitung aus dem Zeitraum 19. – Anfang 20. Jahrhundert
- Fibeln (https://github.com/UB-Mannheim/Fibeln)
  - Transkription: UB Mannheim und GEI Braunschweig nach OCR-D Level 2

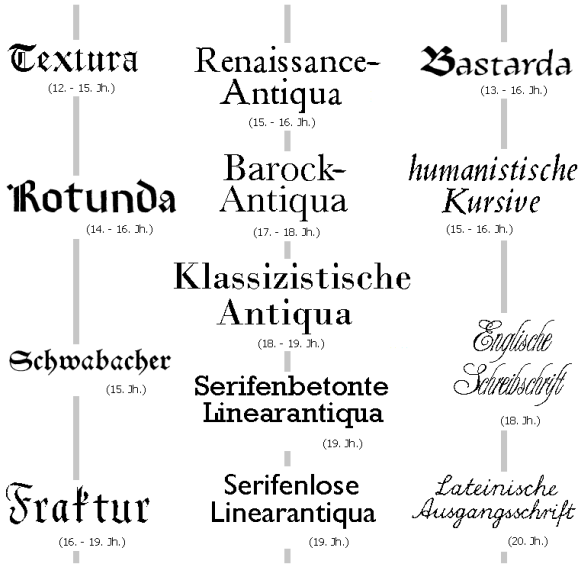


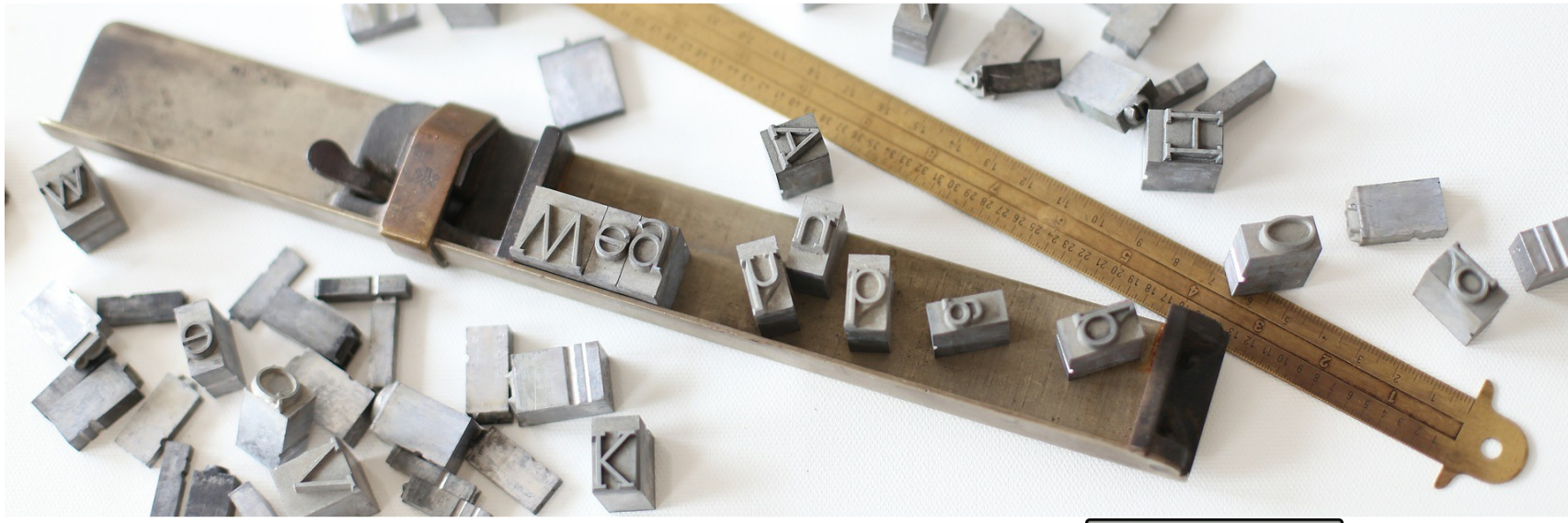

  - Umfang: 3871 Zeilen
  - Inhalt: Deutsche Lernfibeln aus dem Kaiserreich (19. Jahrhundert) mit Schreib- und Druckschrift

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UB
UNIVERSITÄTSBIBLIOTHEK

# Aufwertung und Korrektur
## auf OCR-D Level 2



Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST



09.06.2021

# Regelbasierte Aufwertung und Korrektur

- Entfernung von Textdateien mit mehrfach Zeilen
- Doppelte Leerzeichen zu einfachen Leerzeichen
- Satzzeichen bei Satzenden an den vorangegangen Buchstaben ziehen
- Bei Datumsangaben Minuszeichen durch Halbgeviertstrich ersetzen
- Ersetzen von Bruchdarstellungen 1/2 --> ½

# Komplexere Problemfelder

- ä --> a͛ 
- sz --> ß 
- s --> ſ 
- r --> ꝛ 
- Trennzeichen
  - --> ⸗ 
- Fehlerhafte Bild und Text Zuordnung

# GTReval

- Skriptbasierte, automatische Aufwertung und Korrektur von GT-Zeilenpaaren durch spezialisierte Modelle und nutzerspezifische Regeln
- Nutzerspezifische Regeln
  - Bspw. OCR-D Level Richtlinien 2
    - s --> ſ
- Grundvoraussetzung
  - das Modell muss den Text gut genug erkennen,
    um durch einen Vergleichsalgorithmus Zuordnung zu erhalten
  - Das Modell muss die zu erkennenden Glyphen sehr gut erfassen (spezialisiert)
- Analyse der GT auf die Häufigkeit der Glyphen und die Einhaltung von Transkriptionsregeln/-richtlinien (bspw. OCR-D Level 2)

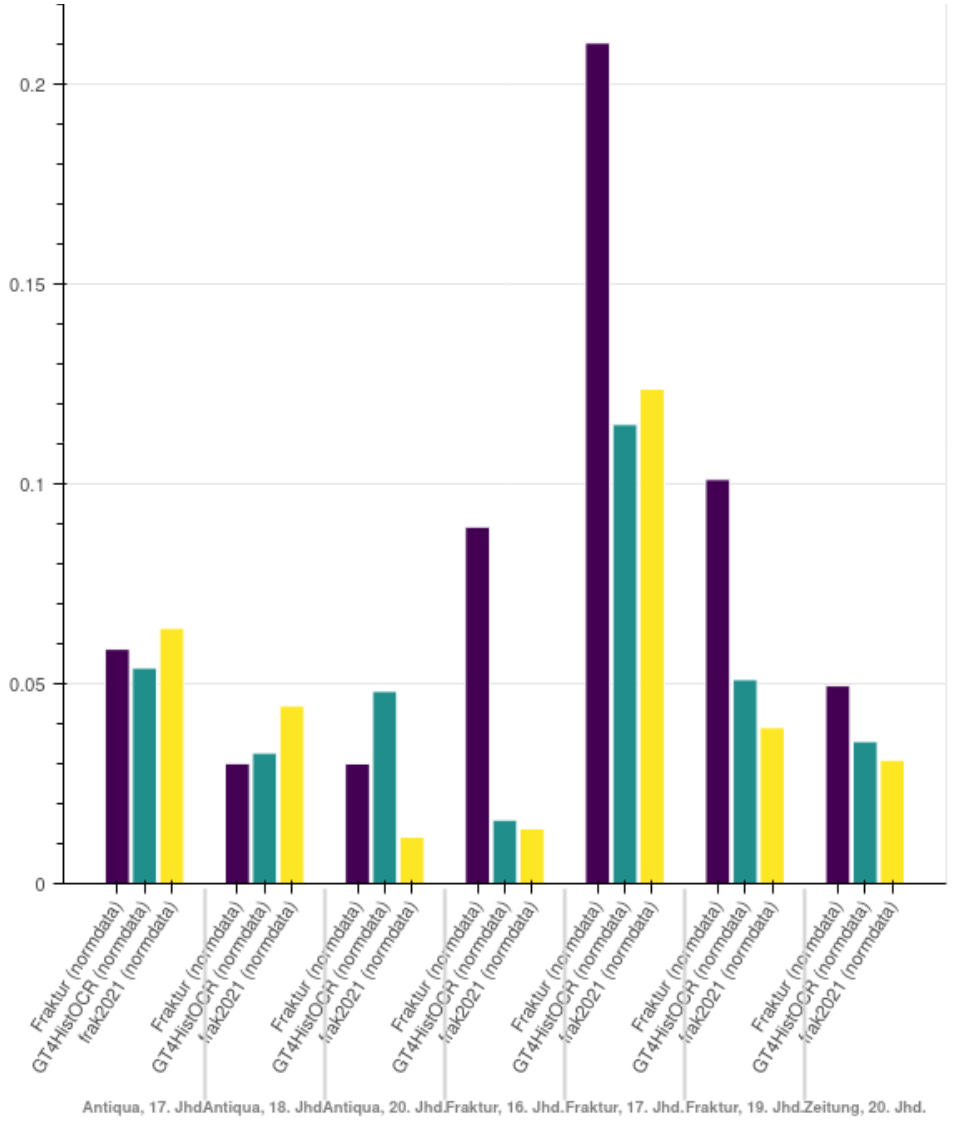
# GTCheck

- Programm zur manuellen Korrektur von Änderungen an git-verwalteter GT
- Web-GUI mit browserbasierter Rechtschreibprüfung
- Korrekturen werden via Git versioniert
- Features
  - Einblendung der vorangegangen und folgenden Zeile
  - Virtuelles Keyboard
  - Rückgängigmachung der letzten Korrektur

# frak2021 – CER

- Antiqua, 17. Jhd. (Französisch)
- Antiqua, 18. Jhd. (Französisch)
- Antiqua, Anfang 20. Jhd.
- Fraktur, 16. Jhd.
- Fraktur, 17. Jhd.
- Fraktur, 19. Jhd.
- Zeitung, 20. Jhd.

- Fraktur
- GT4HistOCR
- frak2021



# frak2021 – Stärken / Schwächen

- Stärken
  - Erkennungsgeschwindigkeit (sehr kleines Modell)
  - Fraktur
  - Historische Antiqua (ohne diakritische Zeichen)
  - Zeitungen Anfang 20. Jahrhundert
  - Separatoren besonders schräger Doppelstrich
- Schwächen
  - Französische Texte
  - Diakritische Zeichen
  - Verwechselungen:
    - N und R
    - t und k

# Fazit

- Die neuen Modellen schneiden bei den meisten Vorlagen deutlich besser ab als die Standard-Tesseract-Modelle für Fraktur
- Es gilt stets zunächst, die Modelle zu testen – derzeit gibt es keinen eindeutigen Gewinner
- Die reduzierte Netzwerktiefe sollte als Faktor untersucht werden
- Verbesserungspotential
  - Aufwertung und Korrektur
  - Erweiterung des Zeichensatzes
  - Ausgleich von unterrepräsentierten Zeichen
  - Aufbau des neuronalen Netzes

# Ausblick

- OCR-D (2021–2023)
  Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung

  Ziel dieses Projektes ist, dass Einrichtungen (zum Beispiel Bibliotheken) möglichst einfach die Module des OCR-D-Workflows nachtrainieren können,
  so dass bessere Erkennungsraten für spezifische Werke erreicht werden können.

  - Werksspezifisches Training scheint ein erfolgversprechendes Mittel, um besonders hohe Genauigkeit mit recht überschaubaren Ressourceneinsatz zu erreichen
  - Es sollen weiterhin die Datensätze aufgewertet und korrigiert werden
  - Neue Modelle basierend auf zusätzlichen GT-Daten sind geplant

# Literatur

- **OCR-D Richtlinien**: https://ocr-d.de/de/gt-guidelines/trans/