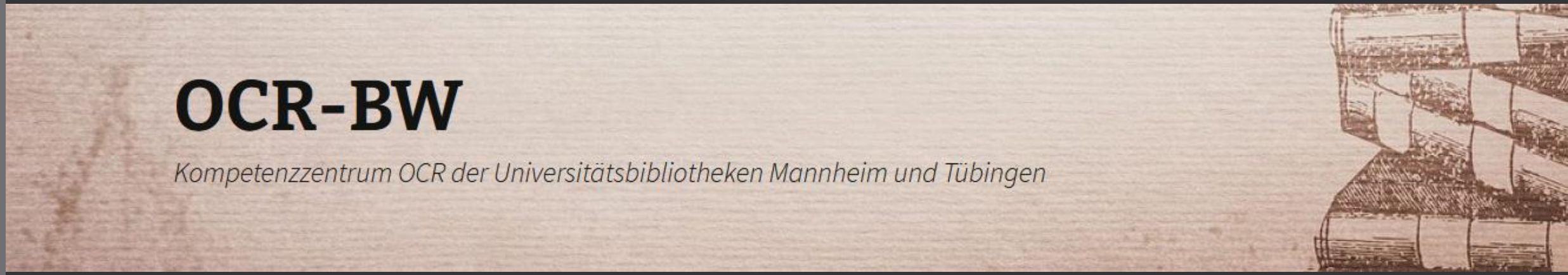
Werksspezifisches Training für ein historisches Werk am Beispiel der Weisthümer von Jacob Grimm



Präsentation im Rahmen des zweiten OCR-BW-Workshops am 9. Juni 2021
Fabian Voigtschild (Bayerische Staatsbibliothek)
Maria Nüchter (Universitätsbibliothek Frankfurt am Main)

Die Sammlung „Weisthümer“ von Jacob Grimm

- Rechtsquellen aus dem späten Mittelalter und der frühen Neuzeit
- Die Sammlung von Jacob Grimm umfasst insgesamt sechs Bände (etwa 5500 Seiten)
- gestaffelte Veröffentlichungen der Einzelbände von 1840 bis 1869
- unterschiedliche sprachliche Merkmale: verschiedene Varianten des Mittelhochdeutschen bis zum Lateinischen

Typographische Eigenschaften des Textes

- Schriftart historische Antiqua
- Umfang von etwa zwei Seiten je Kapitel
- einspaltiger Satz (einfaches Layout)
- Vorkommende Sonderzeichen je nach verwendeter Sprache:
  Römische Zahlen, Exponenten, Paragraphenzeichen (§), langes s (ſ), Zirkumflex (â), Caron (ǎ), Akzent (á), Ringdiakritik (å), diakritische Umlaute (a^{e}) und drei kursiv geschriebene griechische Buchstaben: Theta (ϑ), Beta (ϐ), Pi (Π).

Werksspezifisches Training

- Mithilfe von Tesseract (Version 4.0.0) können Modelle für bestimmte Sprachen oder Schrifttypen durch neuronale Netze erstellt werden
- Hiermit kann ein bereits vorhandenes Modell durch ein werksspezifisches Training auf typographische Besonderheiten und fehlende Zeichen hin verbessert werden
- Das Modell GT4HistOCR beruht auf einer Sammlung unterschiedlicher historischer Drucke im Umfang von 313.173 Textzeilen
- Während eines werksspezifischen Trainings werden im Prozess schrittweise neue Modelle generiert, die sich idealerweise kontinuierlich verbessern

Transkriptions-Richtlinien

- Transkription auf der Grundlage der von OCR-D bereitgestellten „Richtlinien zur Transkription der Volltexte für die Nutzung als Ground Truth“
- Level 2:
  - nur eigenständige Grapheme (vokalische Ligaturen) mit einem spezifischen Codepoint unter Nutzung von standardisierten Kodierungen (Unicode) abgebildet
  - Die Wiedergabe von Leerzeichen beschränkt sich darauf, dass diese ausschließlich Wörter voneinander trennen
  - Satzzeichen werden immer an das vorangegangene Wort herangezogen

Training Kennzahlen

- Verwendetes Modell für das werksspezifische Training: GT4HistOCR
- Transkription für das Training: 30 Seiten (ca. 1500 Zeilen)
- Transkription für die Validierung: 5 Seiten (ca. 250 Zeilen)
- Ergebnisse aus einem Trainingsdurchlauf:
  - 58 Modelle
  - 65.700 Iterationen

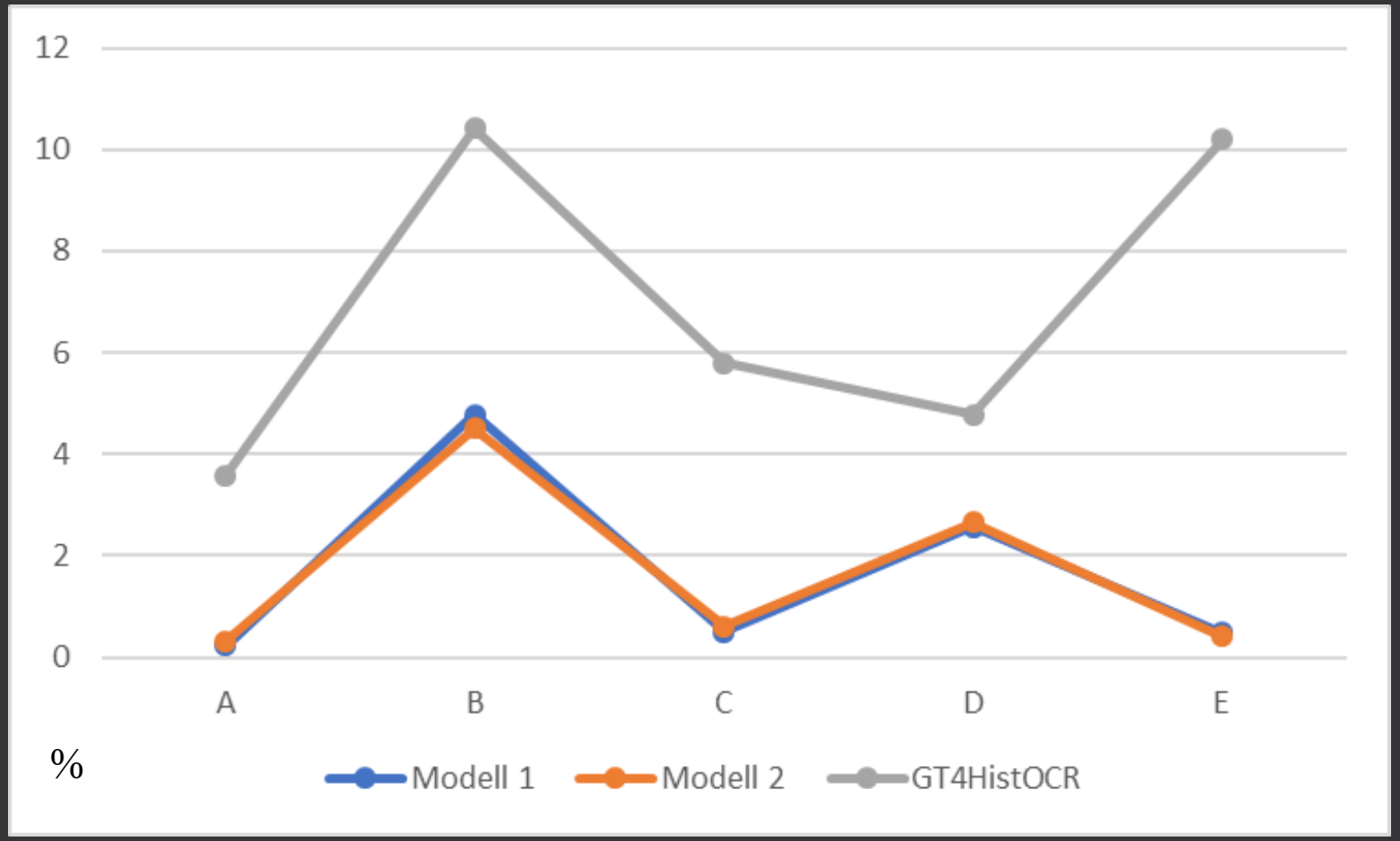
Validierung der OCR-Ergebnisse

- Vergleich der OCR-Ergebnisse mit der fehlerfreien Transkription
- Fehlerrate ergibt sich aus dem Quotienten der Anzahl an Fehlern und der jeweiligen Textlänge
- Character Error Rate (CER) oder alternativ Word Error Rate (WER)

Validierung der OCR-Fehlerrate (CER)

Seite	Modell 1	Modell 2	GT4HistOCR	Differenz
A	0,23%	0,32%	3,58%	91%
B	4,77%	4,51%	10,41%	57%
C	0,51%	0,61%	5,79%	89%
D	2,57%	2,67%	4,79%	44%
E	0,48%	0,41%	10,20%	96%

Validierung der OCR-Fehlerrate (CER)



Validierung der OCR-Ergebnisse

Transkription:	meier 3 ß ϑ ze besserung vervallen sin.
Modell 1:	meier 3 ß ß ze besserung vervallen sin.
Modell 2:	meier 3 ß ß ze besserung vervallen sin.
GT4HistOCR:	meier 3 4 & A, besserung vervallen sĩn.

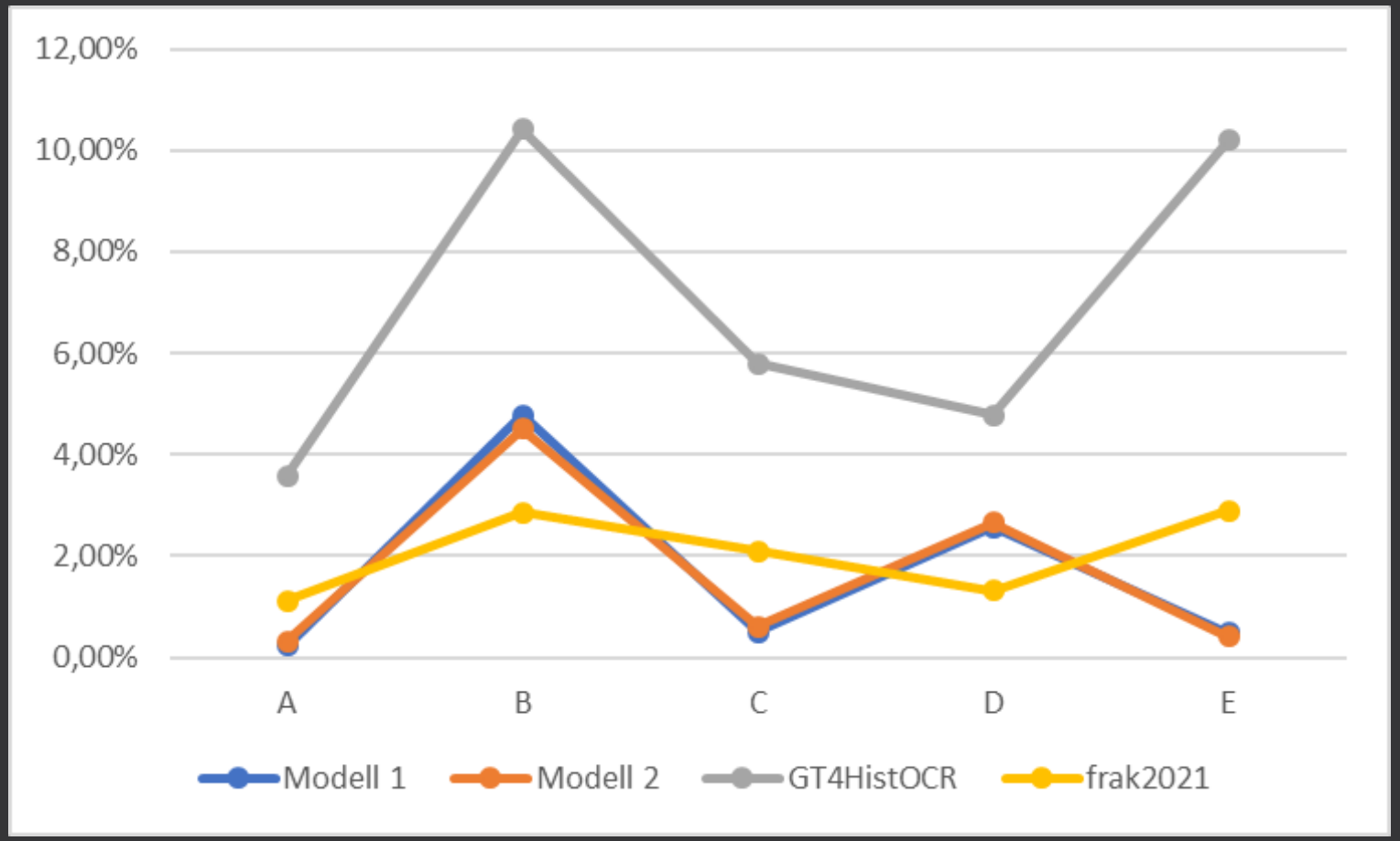
Validierung der OCR-Ergebnisse

Transkription:	die Esche vnd dannan gen Bůchs vnder dis
Modell 1:	die Esche vnd dannan gen Búchs vnder dis die
Modell 2:	die Esche vnd dannan gen Bůchs vnder dis
GT4HistOCR:	die Esche vnd dannan gen Biiehs vnder dis

Fazit

- Das werkspezifische Training erzielte in diesem Beispiel eine signifikant geringere Fehlerrate
- Das im Trainingsvorgang zuletzt ausgegebene Modell liefert nicht zwangsläufig die besten Ergebnisse
- Modell GT4HistOCR eignet sich als Ausgangspunkt für ein werkspezifisches Training

Ausblick


12,00%
10,00%
8,00%
6,00%
4,00%
2,00%
0,00%
A
B
C
D
E
Modell 1
Modell 2
GT4HistOCR
frak2021

Quellen

- Aletheia. https://www.primaresearch.org/tools/Aletheia
- GT4HistOCR. https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR
- OCR-D. Richtlinien zur Transkription für Ground Truth. https://ocr-d.de/de/gt-guidelines/trans/transkription.html
- Springmann, Uwe, Reul, Christian, Dipper, Stefanie, & Baiter, Johannes. (2018). GT4HistOCR: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1344132
- Tesseract. https://github.com/tesseract-ocr/tesseract
- Tesstrain. https://github.com/tesseract-ocr/tesstrain
- Weisthümer. https://github.com/UB-Mannheim/Weisthuemer