BIBLIOTHECA ARABICA

Deep-Learning-basierte Texterkennung arabographischer Handschriftenkataloge

Herausforderungen und Best Practices

Coffee Lecture, 25.06.2026

Daniel Kinitz (SAW Leipzig)

Sächsische Akademie der Wissenschaften zu Leipzig

Ablauf

- Bibliotheca Arabica: Kontext
- Herausforderungen

3 Arbeitsbereiche

Makroperspektive

u.a.
Textgenres,
Textpraktiken

Bsp.
Hadith
Commentaries,
Randkommen-
tare

Digitale BA

Mikroperspektive

u.a.
MS-Vermerke,
Bibliotheken

Bsp.
Shirwani

Leitung: Prof. Verena Klemm

3 Arbeitsbereiche

Digitale BA

https://khizana.saw-leipzig.de

Graph Database (Knowledge Graph)

- Data from > 100 manuscript catalogues (print & online)
- Manuscript notes on owners, readers, etc.
- …

Use Case OCR:

- 250 Bände Handschriftenkataloge (ar/fa/en + Umschrift) erkennen
- Challenge: viele Eigennamen (NLP-Postkorrektur schwierig)

Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Verarbeitungskette

scan

OCR

full text

parse

import

graph database

authority control

public research plattform

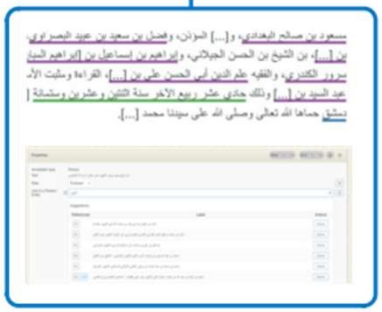
manuscript

transcribe

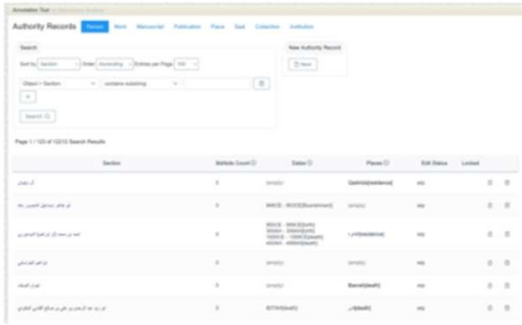
full text (manuscript notes)

annotate

authority control



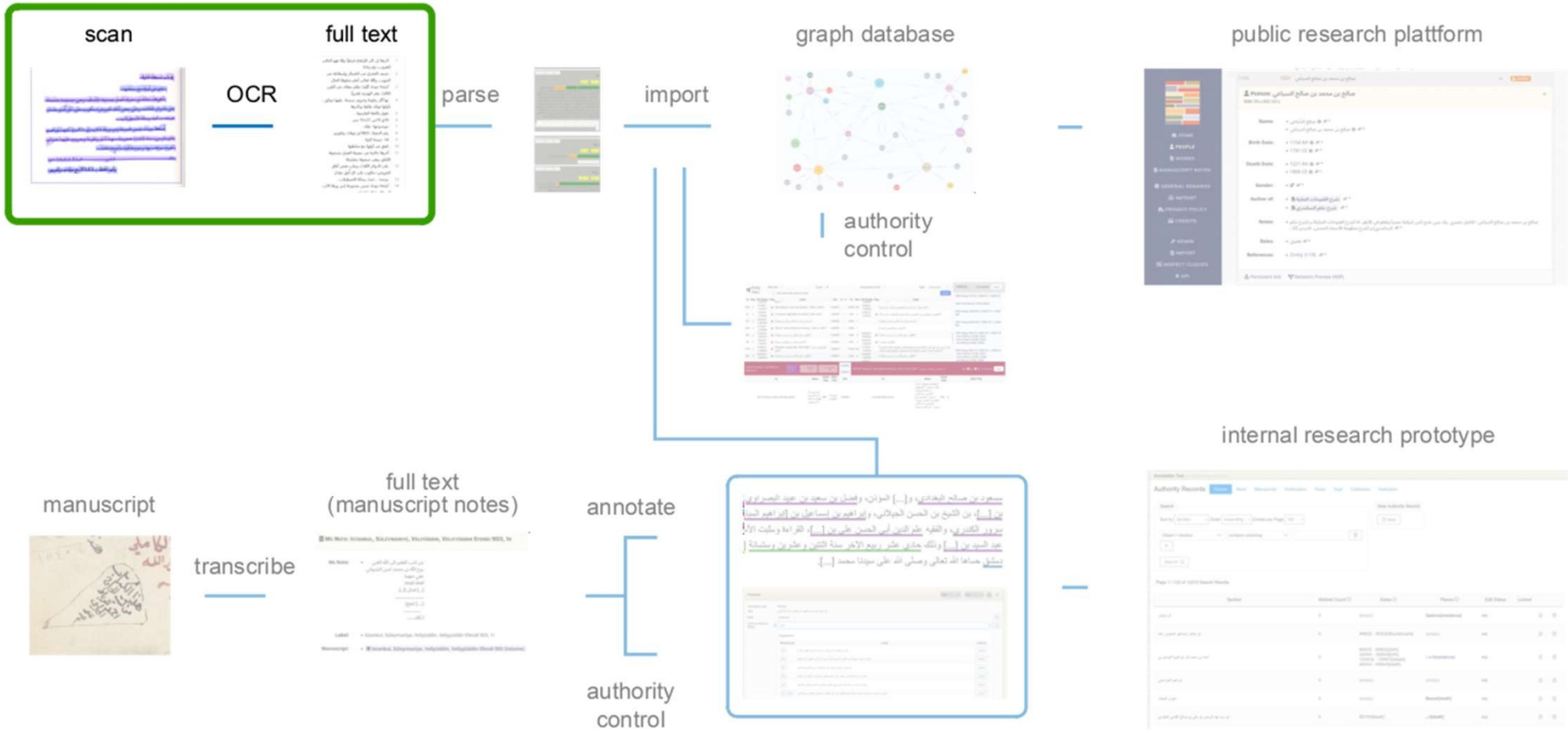
internal research prototype



Sächsische Akademie der Wissenschaften zu Leipzig

BIBLIOTHECA ARABICA

Verarbeitungskette



Historische Ausgangslage: arabographisches OCR

- Gedrucktes Arabisch, Persisch, Urdu usw. (RTL, Ligaturen) nicht mit herkömmlichen OCR-Anwendungen gut erkennbar (hoher Aufwand, relative schlechte Ergebnisse)

- Ab Mitte der 2010er Jahre erste frei verfügbare DL-basierte Anwendungen für OCR (kraken, Ben Kiessling), aber: zunächst ungeeignete/schlechte Transkriptionsinterfaces für RTL-Sprachen; → eScriptorium Repo seit 2018 Jahren auf gitlab.com

→ Character Accuracy: > 0.99 möglich, word accuracy: > 0.97 möglich

OCR: WORKFLOW

eScriptorium

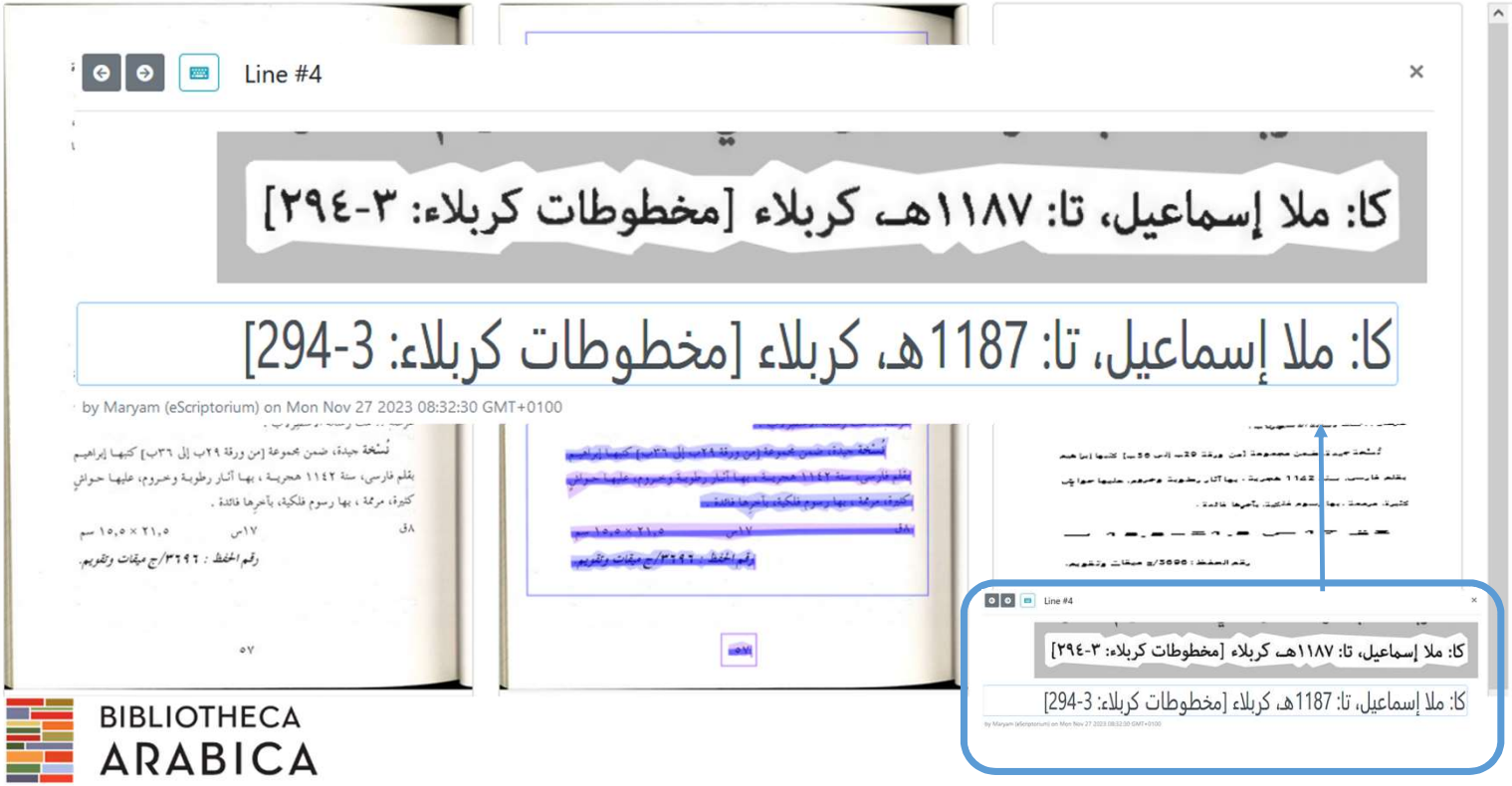
kraken

image

segmentation

training/correction

full text



1 آخرها: إن كان الإرتفاع شرقياً، وإلا فهو الباقى للغروب، مع زيادة
2 نصف التعديل فى الشمال وإسقاطه فى الجيوب، والله تعالى أعلم بحقيقة الحال.
3 نُسْخة جيدة، كُتِبَت بقلم معتاد، فى القرن الثالث عشر الهجرى تقديراً،
4 بها آثار رطوبة وخروم، مرممة، عليها حواشٍ ، بأولها فوائد فلكية وبآخرها
5 نقول باللغة الفارسية .
6 10ق 19س 21×15 سم
7 موضوعها : فلك .
8 رقم الحفظ : 3825/ج ميقات وتقويم.
9 14- نسخة ثانية
10 تتفق فى أولها مع سابقتها.
11 آخرها: خاتمة فى معرفة العمل بصحيفة الآفاق، وهى صحيفة مشتملة
12 على الدوائر الثلاث، وعلى بعض آفاق العروض؛ مكتوب على كل أُفق مقدار
13 عرضه . . تمت رسالة الاسطرلاب .
14 نُسْخة جيدة، ضمن مجموعة [من ورقة 29ب

BIBLIOTHECA ARABICA

Herausforderungen & Best Practices

Herausforderung: Kinderkrankheiten App

Rechte Spalten: Reihenfolge Zeilen vertauscht

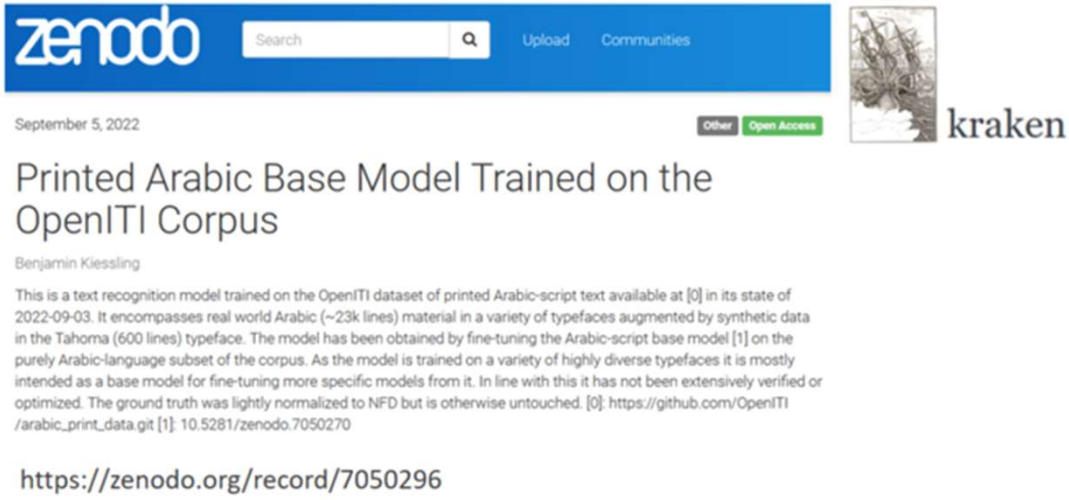
Best Practices:

- App selbst fixen (pull request) → teuer
- auf Update warten / Arbeitspakete verschieben → umständlich
- eigenes Script → pragmatisch
- Workarounds, inklus. manuelle Arbeit SHKs

Herausforderung: Trainingsdaten generieren → viel Arbeit

mit wenig Manpower …

- bis vor Kurzem Fonts/Typen einzeln trainieren, von Grund auf neu
- Sep. 2022: erstes Gesamtmodel für gedrucktes Arabisch → Finetuning



kraken

Herausforderung: Korrekturlesen Hunderter Seiten

mit wenig Manpower …

→ 2 Modelle trainieren und diffs vergleichen

Herausforderung: gemischte Schriften

→ gemischte Schriften (arabisch/persisch): Homographen und Verwechslungen pers. vs. arab. Ziffern

Bsp.: Verbundkatalog irak. Handschriften: Bd. 1-17 arab. Zeichensatz, Bd. 18 arab.+pers.

lat: 0123456789

ar: ٠١٢٣٤٥٦٧٨٩

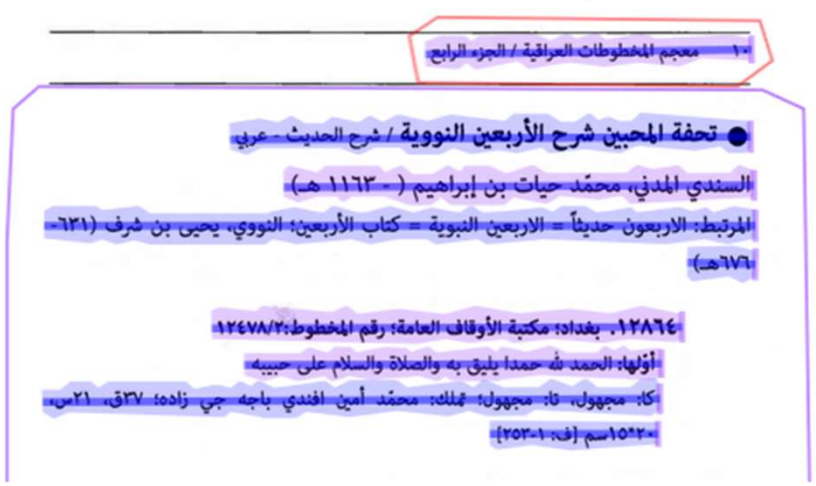
fa: ۰۱۲۳۴۵۶۷۸۹

→ pers.-arab. Homographen bzw. Verwechslungsmöglichkeiten
arabisches ʿAyn (ع) → erkannt als pers. 6 (۶) → neues Modell
Herausforderung: den genauen Zeichensatz kennen (großen Textkorpus)

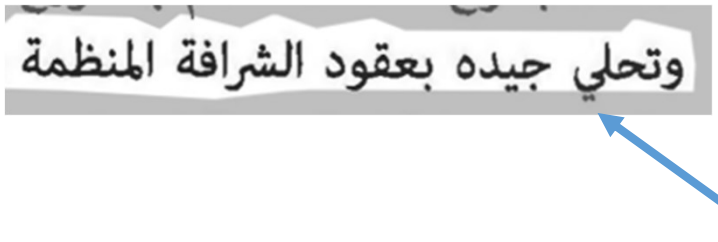
+ pers. Bearbeiter: arabische 1 (١ U+0661) vs. Persische 1 (۱ U+06F1) → Modell trainiert verschiedene, Zeichen, die m.E. nur aus politischen Gründen versch. code points erhalten haben

Herausforderung: Nichttrainierbarkeit der Masken

Trainierbar:
Regionen
Baselines



Bis dato nicht trainierbar:
Masken um Zeilen (Polygone)



Möglicher Workaround:

Statistik: seltener Endbuchstaben ى → Postkorrektur über Liste

Technischer Lösungsansatz: kraken → party

Herausforderung: Dependencies: py bidi

→ (fehlerhafte?) Python-Implementierung für Unicode bidirectional algorithm → Zero width no-joiner (Bindehemmer) gelöscht

می‌باید mit Bindehemmer (korrekt)

میباید ohne Bindehemmer (→ inkorrektes Persisch, ggf. bedeutungsunterscheidend)

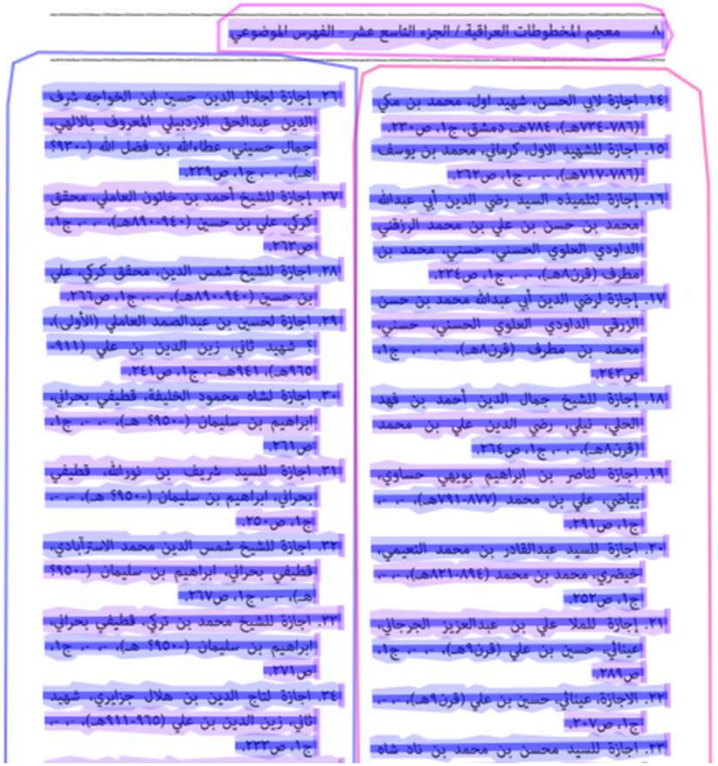
می@باید **Workaround**: durch Zeichen ersetzen, das nicht gelöscht wird

https://github.com/MeirKriheli/python-bidi/issues/4

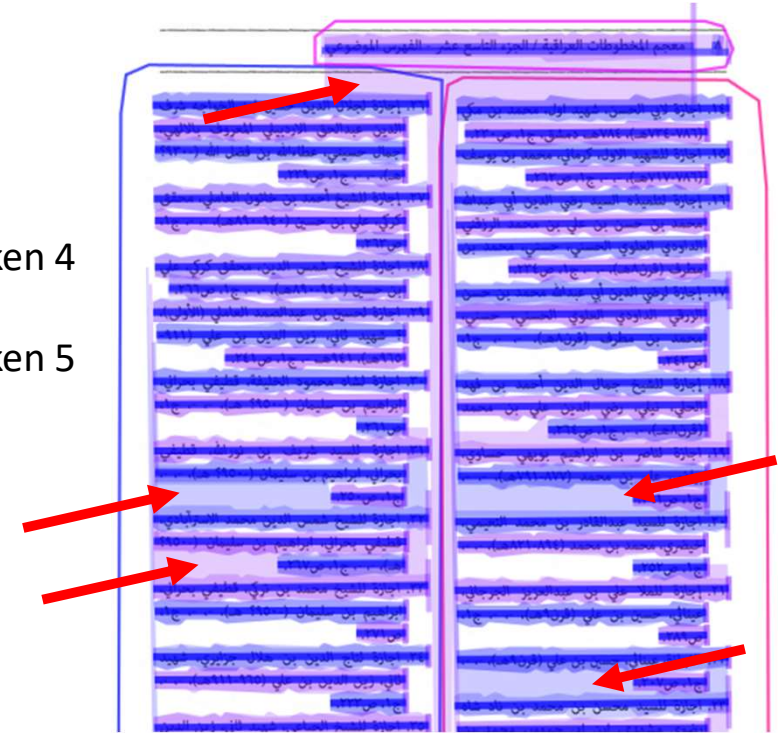
→ **Problem**: Abhängigkeit von Privatinitiativen / Einzelentwicklern

Herausforderung: Regression

Regression bei Masken (Polygone um Zeilen), mit kraken bis dato nicht trainierbar



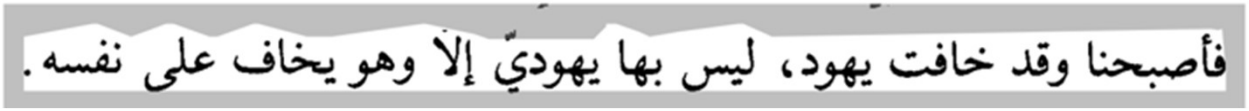
kraken 4
→
kraken 5



Workaround:
kraken 4: segmentation
kraken 5: recognition

Herausforderung: manuelle Fehler + Overfitting

Bild



Erkannt فأصبحنا وقد خافت يهود، لوقعتنا بعدو الله فليس بها يهودي إلا وهو يخاف على نفسه.

Kauderwelsch
(bestimmter Abschnitt)

?

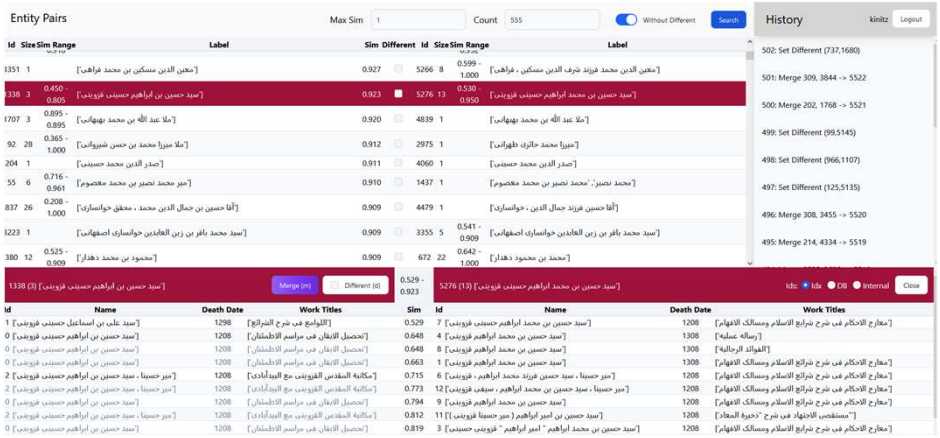
Ground Truth فأصبحنا وقد خافت يهود، لوقعتنا بعدو الله فليس بها يهودي إلا وهو يخاف على نفسه.

! Grund: Nachnutzung vorhandener Scan-Text-Paare (Volltexte Bücher)

Lösungsansatz: Kenne deine Trainingsdaten/GT!
Kenne die Schwächen deines Modells!
Implementiere ausreichend Checks!

Mögliche weitere Anwendungsfälle DL

- Handwritten Text Recognition (mehr Trainingsdaten)
- Normierung: identische Personen, Werke in Daten finden (s. Bild)
- Arabische Schrift ←→ latinisierte wiss. Umschrift (nicht eineindeutig)
- Retrieval Augmented Generation: Sprachmodell + eigene Quellen → Befrag den Chatbot über unsere Daten!



Vorschlagssystem für identische Personen
(M. Reckziegel/ D. Kinitz)

Fragen?
Anmerkungen?

Literatur

D. Kinitz (forthcoming): Deep Learning-Based OCR of Printed Arabographic Resources. Challenges and Best Practices in a Production Environment. In „Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology“, 12 2026. https://ebook.ek.szte.hu/index.php/btk-magyarnyelviirodalmi-intezet/catalog/series/wpcl

D. Kinitz / Th. Efer (2023): *Towards a Dynamic Knowledge Graph of a Non-Western Book Tradition.* In: Baillot et al. (eds.). Digital Humanities 2023: Book of Abstracts. Graz 2023, pp.216-217.