



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

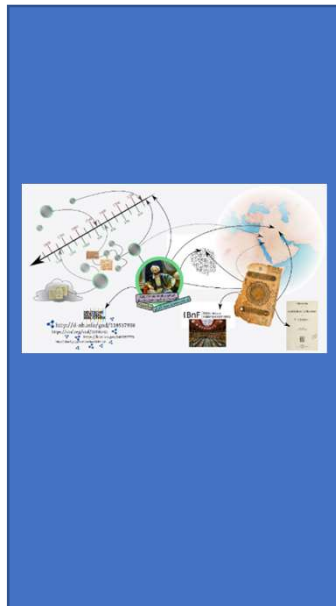
3 Arbeitsbereiche

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



KHIZANA BETA

BIBLIOTHECA ARABICA

BIBLIOTHECA ARABICA'S REFERENCE WORK ON THE ARABIC MANUSCRIPT TRADITION

KHIZANA aims to be a comprehensive bio-bibliographical reference work on agents and works related to Arabic manuscripts, focusing on the period between the 12th and the 19th centuries CE. As a reference work on Arabic literature, it integrates **sources** relevant to Bibliotheca Arabica. The most important source types are:

- data from manuscript catalogues (print & online)
- data from biographical and bibliographical works and
- manuscript notes on owners, readers, etc.

Using graph based technologies, special emphasis is laid on providing evidence (i.e., **provenance**) for every piece of information. Work on the KHIZANA will be expanded continuously in the next years. Currently, the following entries are available (including possible duplicates):

PERSONS 122,191	WORKS 120,881	MANUSCRIPTS 97,855	MSNOTES 71,287
---------------------------	-------------------------	------------------------------	--------------------------

[Sign In]

<https://khizana.saw-leipzig.de>

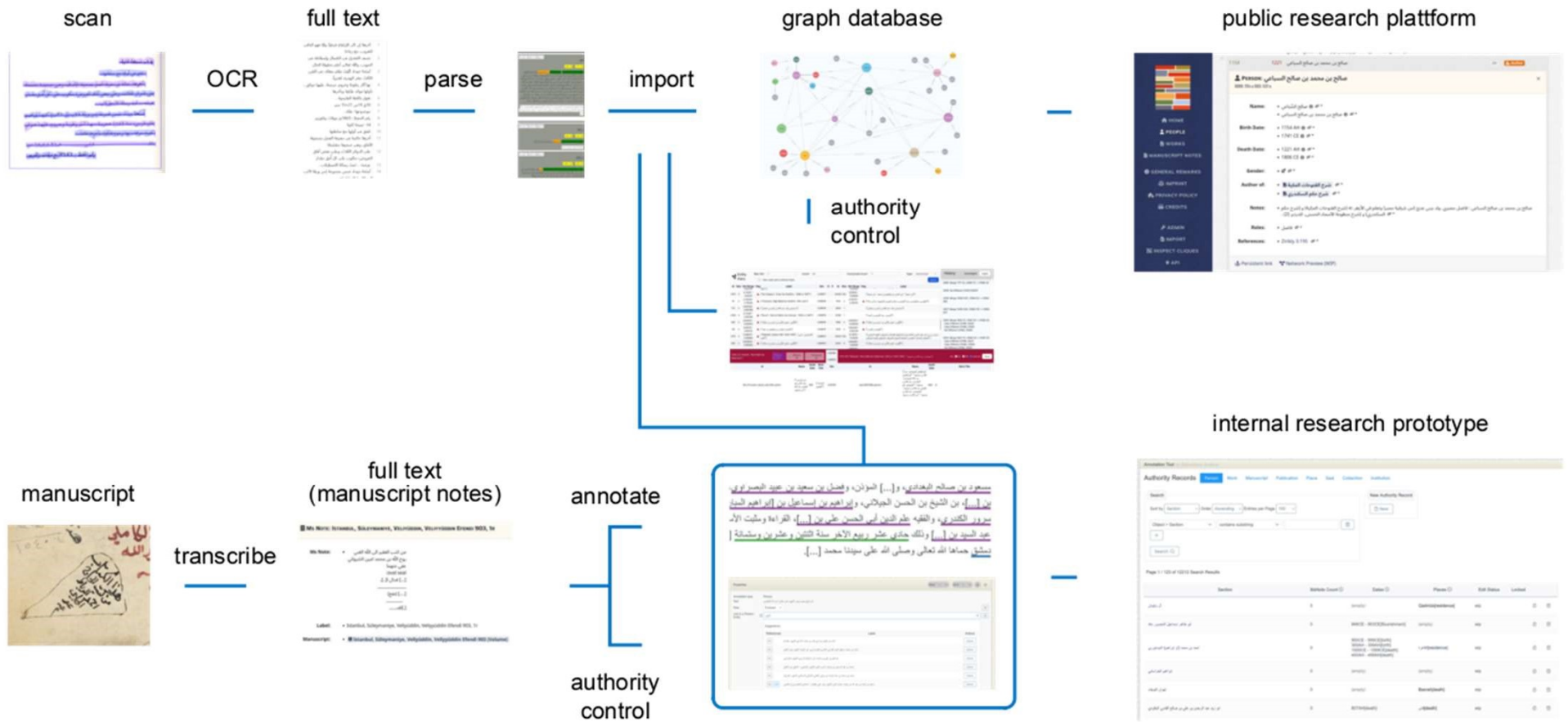
Graph Database (Knowledge Graph)

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

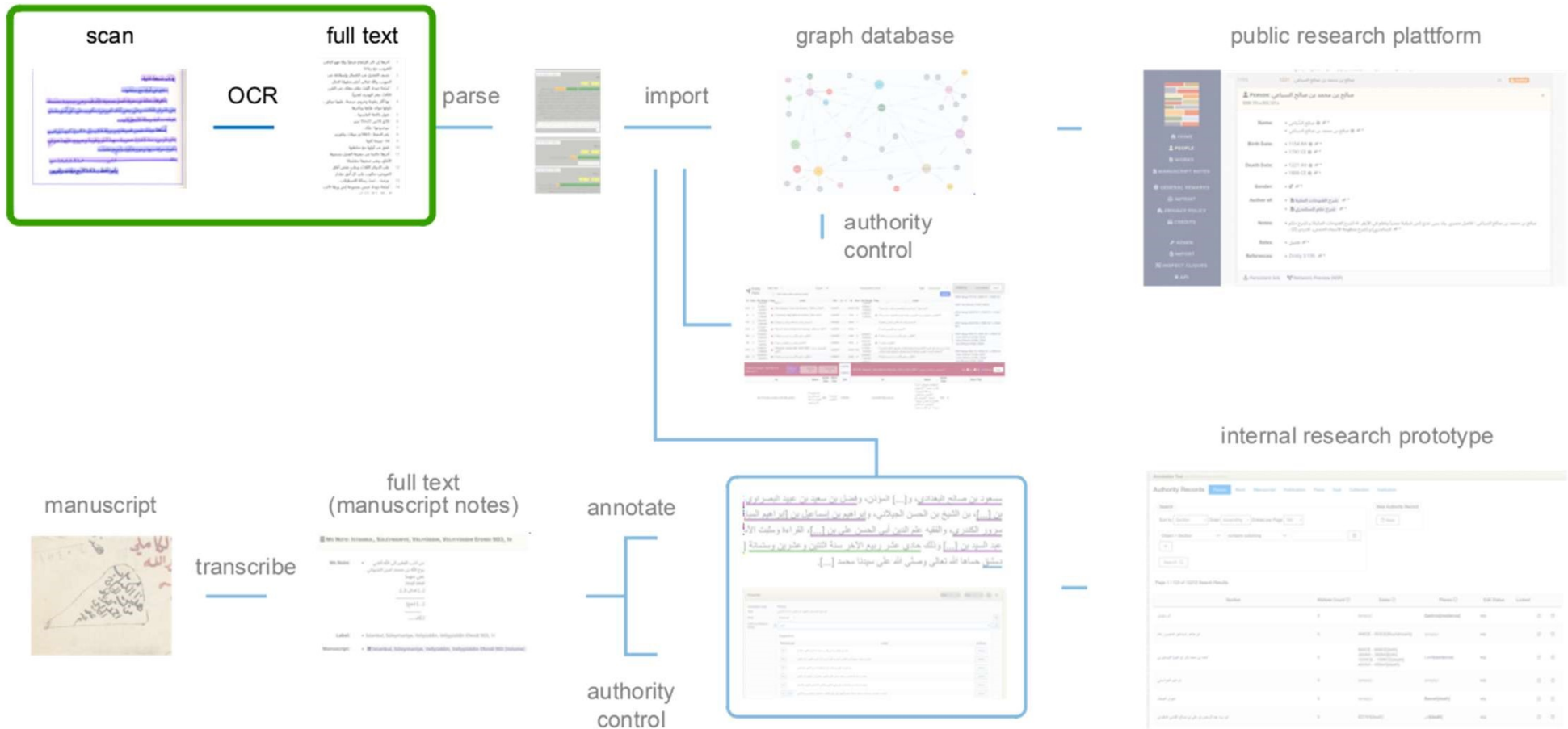
Use Case OCR:

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

Verarbeitungskette



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](https://gitlab.com/eScriptorium)
- Character Accuracy: > 0.99 möglich, word accuracy: > 0.97 möglich

OCR: WORKFLOW



eScriptorium

kraken

image

segmentation

training/correction

full text

Line #4

كا: ملا إسماعيل، تا: ١١٨٧هـ كربلاء [مخطوطات كربلاء: ٣-٢٩٤]

كا: ملا إسماعيل، تا: 1187هـ، كربلاء [مخطوطات كربلاء: 3-294]

by Maryam (eScriptorium) on Mon Nov 27 2023 08:32:30 GMT+0100

نسخة جيدة، ضمن مجموعة (من ورقة 29 إلى 36) كتبها إبراهيم بن محمد فارس، سنة 1142 هجرية، بها آثار رطوبة وخروم، عليها حواشي كثيرة، مرمة، بها رسوم فلكية، وآخرها فلكة. ١٧ س ١٥,٥ × ٢١,٥ سم رقم الحفظ: ٣٦٩٦/ج مقيات وتقوم.

نسخة جيدة، ضمن مجموعة (من ورقة 29 إلى 36) كتبها إبراهيم بن محمد فارس، سنة 1142 هجرية، بها آثار رطوبة وخروم، عليها حواشي كثيرة، مرمة، بها رسوم فلكية، وآخرها فلكة. ١٧ س ١٥,٥ × ٢١,٥ سم رقم الحفظ: ٣٦٩٦/ج مقيات وتقوم.

نسخة جيدة، ضمن مجموعة (من ورقة 29 إلى 36) كتبها إبراهيم بن محمد فارس، سنة 1142 هجرية، بها آثار رطوبة وخروم، عليها حواشي كثيرة، مرمة، بها رسوم فلكية، وآخرها فلكة. ١٧ س ١٥,٥ × ٢١,٥ سم رقم الحفظ: ٣٦٩٦/ج مقيات وتقوم.

Line #4

كا: ملا إسماعيل، تا: ١١٨٧هـ كربلاء [مخطوطات كربلاء: ٣-٢٩٤]

كا: ملا إسماعيل، تا: 1187هـ، كربلاء [مخطوطات كربلاء: 3-294]

by Maryam (eScriptorium) on Mon Nov 27 2023 08:32:30 GMT+0100

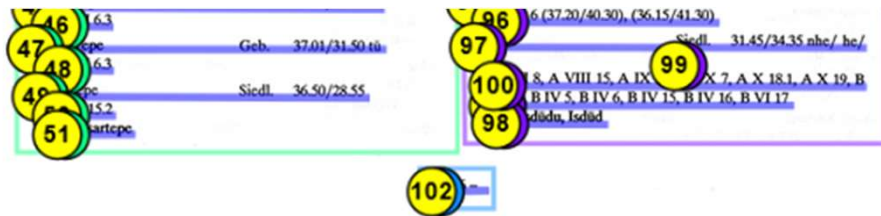
- 1 آخرها: إن كان الإرتفاع شرقياً، وإلا فهو الباقي للغروب، مع زيادة
- 2 نصف التعديل في الشمال وإسقاطه في الجنوب، والله تعالى أعلم بحقيقة الحال.
- 3 نُسخة جيدة، كُتبت بقلم معتاد، في القرن الثالث عشر الهجري تقديراً،
- 4 بها آثار رطوبة وخروم، مرمة، عليها حواشٍ، بأولها فوائد فلكية وبآخرها
- 5 نقول باللغة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

BIBLIOTHECA ARABICA



Herausforderungen & Best Practices

Herausforderung: Kinderkrankheiten App



Rechte Spalten: Reihenfolge Zeilen vertauscht

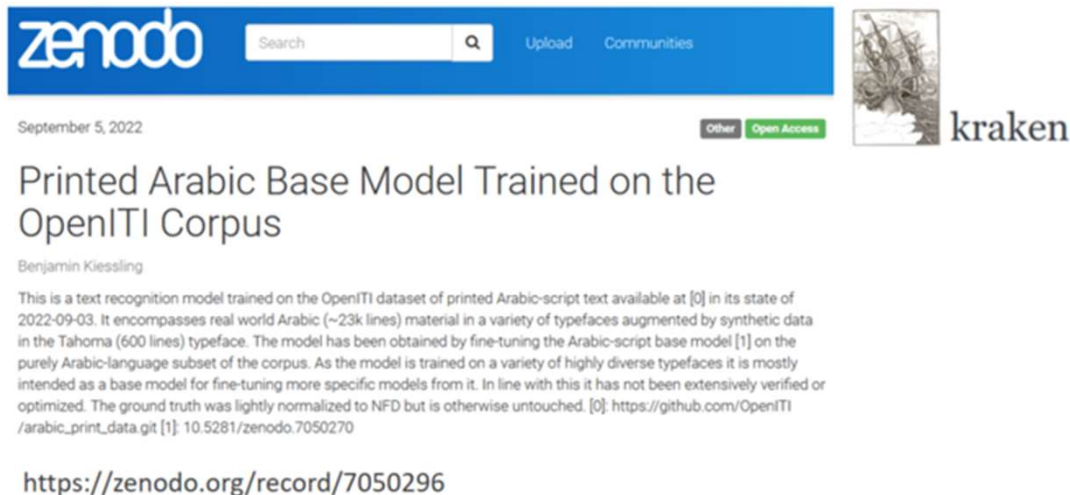
Best Practices:

- 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 Gesamtmodell für gedrucktes Arabisch → Finetuning



The screenshot shows the Zenodo interface for a record titled "Printed Arabic Base Model Trained on the OpenITI Corpus" by Benjamin Kiessling. The page includes a search bar, navigation links for "Upload" and "Communities", and a date of "September 5, 2022". A "kraken" logo is visible on the right. The main text describes the model's training on the OpenITI dataset, mentioning the use of synthetic data and the Tahoma typeface. A URL is provided at the bottom: <https://zenodo.org/record/7050296>.

zenodo Search Upload Communities

September 5, 2022 Other Open Access

kraken

Printed Arabic Base Model Trained on the OpenITI Corpus

Benjamin Kiessling

This is a text recognition model trained on the OpenITI dataset of printed Arabic-script text available at [0] in its state of 2022-09-03. It encompasses real world Arabic (~23k lines) material in a variety of typefaces augmented by synthetic data in the Tahoma (600 lines) typeface. The model has been obtained by fine-tuning the Arabic-script base model [1] on the purely Arabic-language subset of the corpus. As the model is trained on a variety of highly diverse typefaces it is mostly intended as a base model for fine-tuning more specific models from it. In line with this it has not been extensively verified or optimized. The ground truth was lightly normalized to NFD but is otherwise untouched. [0] https://github.com/OpenITI/arabic_print_data.git [1] 10.5281/zenodo.7050270

<https://zenodo.org/record/7050296>

Herausforderung: Korrekturlesen Hunderter Seiten

mit wenig Manpower ...

→ 2 Modelle trainieren und diffs vergleichen



وتحلي جیده بعقود الشرافة المنظمة من دولة ذوي الشرف والسيادة الشريفة ... إلى

وتحلى جیده بعقود الشرافة المنظمة من دولة ذوي الشرف والسيادة الشريفة ... إلى

by admin (import) on Mon Jun 03 2024 12:01:01 GMT+0100

Toggle transcription comparison-



iraq_indo_digits_vol_1_2_5_best (current)

iraq_mixed_digits_vol123_17_18_20pages

وتحلى جیده بعقود الشرافة المنظمة من دولة ذوي الشرف والسيادة الشريفة ... إلى
وتحلى جیده بعقود الشرافة المنظمة من دولة ذوي الشرف والسيادة الشريفة ... إلى

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: ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩

fa: ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩

→ pers.-arab. Homographen bzw. Verwechslungsmöglichkeiten
arabisches 'Ayn (ع) → erkannt als pers. 6 (٦) → neues Modell
Herausforderung: den genauen Zeichensatz kennen (großen Textkorpus)

+ pers. Bearbeiter: arabische 1 (١ U+0661) vs. Persische 1 (١ 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)

وتحلي جيده بعقود الشرافة المنظمة

Möglicher Workaround:

Statistik: seltener Endbuchstaben ع → Postkorrektur über Liste

Technischer Lösungsansatz: kraken → party

Herausforderung: Dependencies: py bidi

→ (fehlerhafte?) Python-Implementierung für Unicode bidirectional algorithm → Zero width non-joiner (Bindehemmer) gelöscht

می‌باید mit Bindehemmer (korrekt)

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

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

not working for zero width non-joiner #4

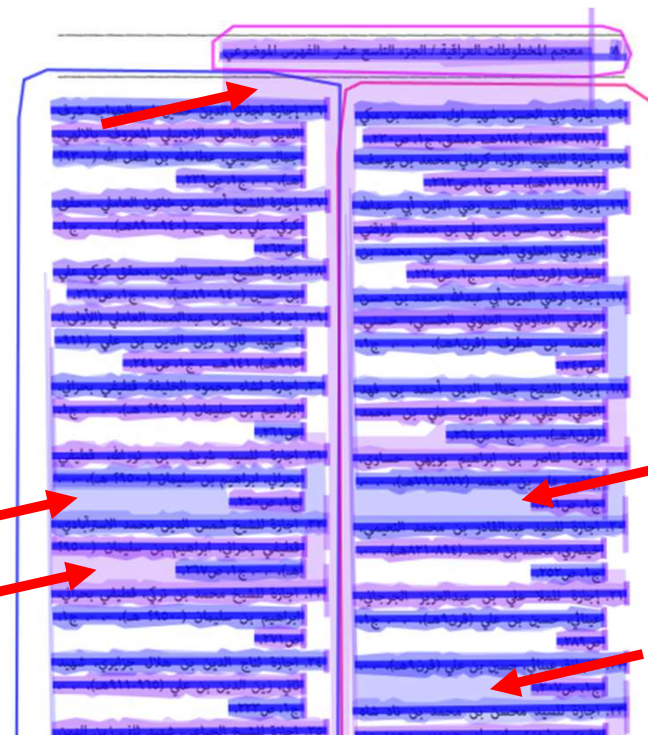
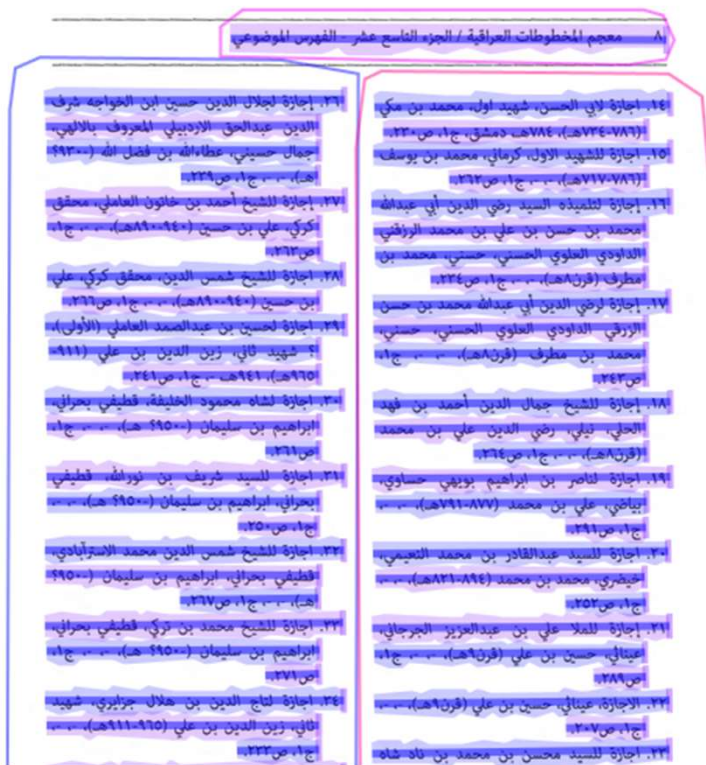
 Closed Mahdizade opened this issue on Feb 28, 2016 · 1 comment

  MeirKriheli closed this as completed on Jul 23 2024! <https://github.com/MeirKriheli/python-bidi/issues/4>

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

Herausforderung: Regression

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



kraken 4
→
kraken 5

Workaround:
kraken 4: segmentation
kraken 5: recognition

Herausforderung: manuelle Fehler + Overfitting

Bild

فأصبحنا وقد خافت يهود، ليس بها يهودي إلا وهو يخاف على نفسه.

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 \leftrightarrow latinisierte wiss. Umschrift (nicht eindeutig)
- Retrieval Augmented Generation: Sprachmodell + eigene Quellen \rightarrow Befrag den Chatbot über unsere Daten!

Entity Pairs

Max Sim: 1 Count: 555 Without Different Search History Kinitz Logout

id	Size	Sim	Range	Label	Sim	Different	id	Size	Sim	Range	Label
3351	1			[معلم الدين مسكين بن محمد فراهي]	0.927		5266	8	0.599 - 1.000		[معلم الدين محمد فرزند تريف الدين مسكين - فراهي]
338	3	0.450 - 0.927		[سيد حسين بن ابراهيم حسيني فونيني]	0.923		5276	13	0.530 - 0.950		[سيد حسين بن محمد ابراهيم حسيني فونيني]
1707	3	0.895 - 0.923		[ملا عبد الله بن محمد بهواني]	0.920		4839	1			[ملا عبد الله بن محمد بهواني]
92	28	0.365 - 1.000		[ملا ميرزا محمد بن حسن فيروكاني]	0.912		2975	1			[ميرزا محمد حائري طهراني]
204	1			[صدر الدين محمد حسيني]	0.911		4060	1			[صدر الدين محمد حسيني]
55	6	0.716 - 0.961		[مير محمد نصير بن محمد مقصوم]	0.910		1437	1			[محمد نصير - محمد نصير بن محمد مقصوم]
837	26	0.208 - 1.000		[آقا حسين فرزند جمال الدين - حواسري]	0.909		4479	1			[آقا حسين فرزند جمال الدين - حواسري]
1223	1			[سيد محمد باقر بن زين العابدين حواسري اصنافي]	0.909		3355	5	0.941 - 0.909		[سيد محمد باقر بن زين العابدين حواسري اصنافي]
380	12	0.525 - 0.909		[محمود بن محمد دهان]	0.909		672	22	0.642 - 1.000		[محمد بن محمود دهان]
1388	8			[سيد حسين بن ابراهيم حسيني فونيني]	0.529 - 0.923		5276	13			[سيد حسين بن محمد ابراهيم حسيني فونيني]

502: Set Different (737,1680)
501: Merge 309, 3844 -> 5522
500: Merge 202, 1768 -> 5521
499: Set Different (99,5145)
498: Set Different (966,1107)
497: Set Different (125,5135)
496: Merge 308, 3455 -> 5520
495: Merge 214, 4334 -> 5519

1388 (8) Merge (8) Different (8) 0.529 - 0.923 5276 (13) id: id: D8: Internal Close

id	Name	Death Date	Work Titles	Sim	id	Name	Death Date	Work Titles
1	[سيد علي بن اسماعيل حسيني فونيني]	1298	[الواجع في شرح الفرائض]	0.529	7	[سيد حسين بن محمد ابراهيم حسيني فونيني]	1208	[معارج الاكابر في شرح فرائع الاسلام ومسالك الانهار]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تمثيل الايمان في مراسم الاضحيان]	0.648	4	[سيد حسين بن محمد ابراهيم فونيني]	1208	[مناجاة عشاق]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تمثيل الايمان في مراسم الاضحيان]	0.648	8	[سيد حسين بن محمد ابراهيم فونيني]	1300	[الاقوال الربانية]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تمثيل الايمان في مراسم الاضحيان]	0.663	1	[سيد حسين بن محمد ابراهيم فونيني]	1300	[معارج الاكابر في شرح فرائع الاسلام ومسالك الانهار]
2	[مير حسينيا - سيد حسين بن ابراهيم حسيني فونيني]	1208	[مكتبة المغنين القويدي مع اليباداد]	0.715	6	[سيد حسين فرزند محمد ابراهيم - فونيني]	1208	[معارج الاكابر في شرح فرائع الاسلام ومسالك الانهار]
2	[مير حسينيا - سيد حسين بن ابراهيم حسيني فونيني]	1208	[مكتبة المغنين القويدي مع اليباداد]	0.773	12	[مير حسينيا - سيد حسين بن محمد ابراهيم - سنيقي فونيني]	1208	[معارج الاكابر في شرح فرائع الاسلام ومسالك الانهار]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تمثيل الايمان في مراسم الاضحيان]	0.794	9	[سيد حسين بن محمد ابراهيم فونيني]	1208	[معارج الاكابر في شرح فرائع الاسلام ومسالك الانهار]
2	[مير حسينيا - سيد حسين بن ابراهيم حسيني فونيني]	1208	[مكتبة المغنين القويدي مع اليباداد]	0.812	11	[سيد حسين بن مير ابراهيم - مير حسينيا فونيني]	1208	[مستكشف الانهار في شرح لاجرة البحار]
0	[سيد حسين بن ابراهيم حسيني فونيني]	1208	[تمثيل الايمان في مراسم الاضحيان]	0.819	3	[سيد حسين بن محمد ابراهيم - مير ابراهيم - فونيني حسيني]	1208	[معارج الاكابر في شرح فرائع الاسلام ومسالك الانهار]

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.