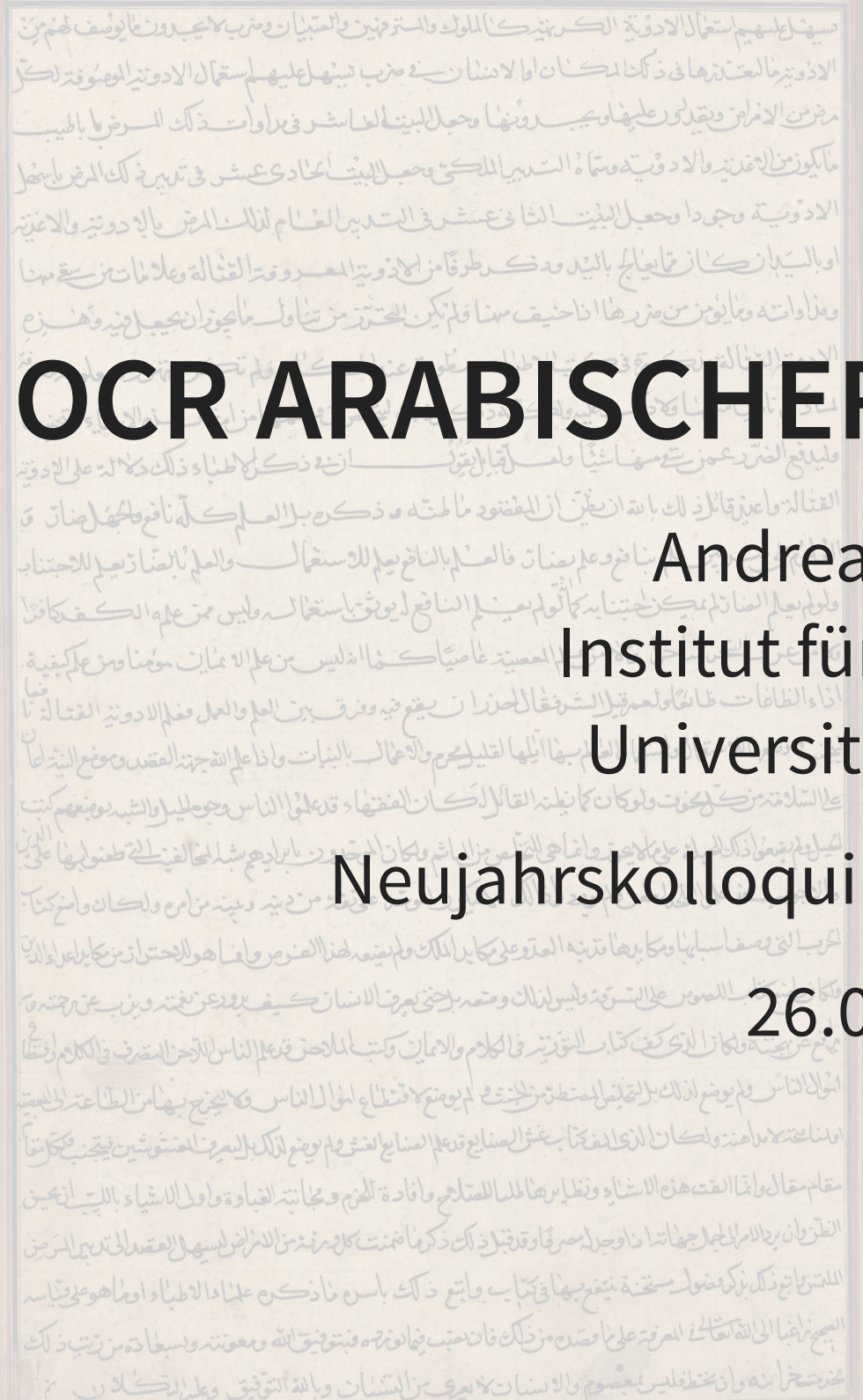
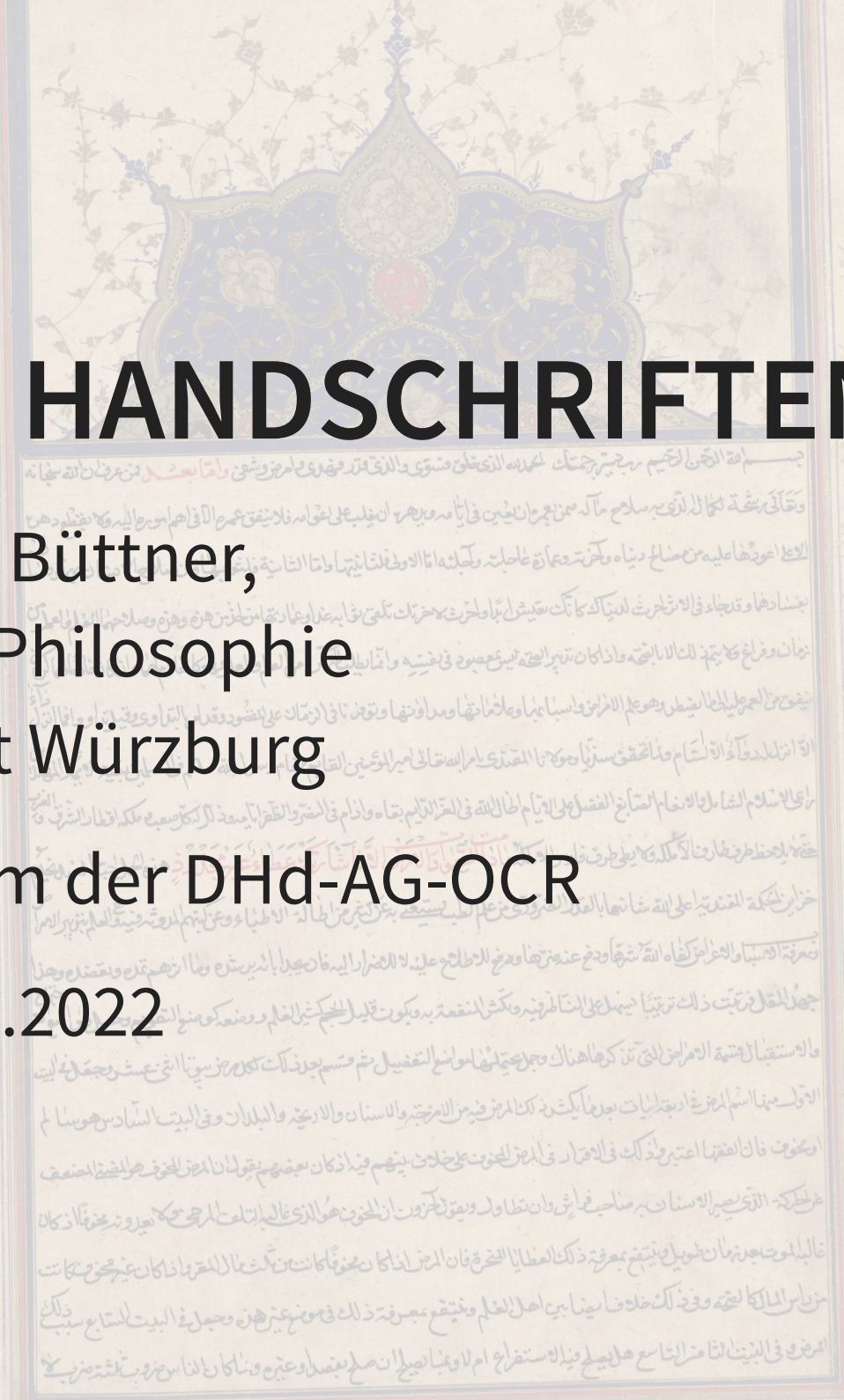


OCR ARABISCHER HANDSCHRIFTEN

Andreas Büttner,
Institut für Philosophie
Universität Würzburg
Neujahrskolloquium der DHd-AG-OCR

26.02.2022



ÜBERBLICK

1. OCR arabischer Schrift
2. Projekt: Leiden UB, Or. 680 für Ptolemaeus Arabus et Latinus
3. Segmentierung
4. GT und Modelle
5. Exkurs: Alignierung existierender GT
6. Modelltraining

OCR ARABISCHER SCHRIFT

- seit der Einführung zeilenbasierter trainierbarer OCR im Prinzip unproblematisch
- Unterstützung für Bidirektionalität erforderlich
- verfügbare GT deutlich weniger als für lateinische Schriften

PROJEKT: PTOLEMAEUS ARABUS ET LATINUS

- Edition der ar. und lat. Versionen der astronomischen und astrologischen Texte des Ptolemaeus
- <https://ptolemaeus.badw.de>

ALMAGEST-HANDSCHRIFT LEIDEN UB, OR. 680, FF. 2R-219R

- Almagest in der ar. Übersetzung al-Ḥajjāj b. Yūsuf b. Maṭar und Sirjis b. Hiliyyā al-Rūmī (214/828-9)
- Handschrift undatiert, ältester Besitznachweis 615/1219
- 172 von 435 Seiten bereits transkribiert

SEGMENTIERUNG

- Regionensegmentierung in LAREX
- Zeilensegmentierung Ocropus/Kraken (vgl. OCR4all)
- Vereinigung nebeneinanderliegender Zeilensegmente
- manuelle Nachkorrektur dringend nötig

الدور وهو على خط الاستواء من جهة الشمال والجنوب في كل وقت من السنة ويكون في
الاعمال اعني اوقات السنة التي تكون من بعد الاعتدالين ويكون من جهة الشمال والجنوب في كل وقت من السنة
ولكن يكون في بعض اوقات من جهة الجنوب والاعمال في بعض اوقات من جهة الشمال والجنوب في كل وقت من السنة
في الاصل الذي هو عند مركز قوس البروج وهو خلاف ذلك في بعض اوقات من جهة الشمال والجنوب في كل وقت من السنة
ويزداد في الشمال والجنوب من قوس البروج نحو الشمال والجنوب في كل وقت من السنة
اطول ايام البرهان الذي هو في الوسط والجنوب في كل وقت من السنة
من جهة الشمال والجنوب في كل وقت من السنة
في الشمس جهة من جهة الخارج وعلى ذلك اختلفت في جهة الاضراسي انما وجدنا هاهنا
فانك التذوق بولها من قبل انفسهم ما وجدنا من جهة الشمال والجنوب في كل وقت من السنة
سابع فباستنا فاننا لا نرى من جهة الشمال والجنوب في كل وقت من السنة
الخارج لا يكون له غير من جهة الشمال والجنوب في كل وقت من السنة
فانك التذوق بولها من قبل انفسهم ما وجدنا من جهة الشمال والجنوب في كل وقت من السنة

عسر ح قايو ولانه اذا نقص خطه
في خطه بك مصروبا في منله ويكون هو
وهو يعني ان يكون خطه كلب بل لا اله الا
علا ان بعد القدر مسيره الا وبعده
بعده في نفسه وبعث ان بعد القدر الحرف ثمانية
واحد او ثنا وعسرون في نفسه وبعده
بور على علامه ح فاذا وصلنا خط
بل فلان راويه يه كل يكون حرا واحدا
الردوانا الفاعله بلهانه وسن حرا
هانه وسن حرافيه يكون حرس
حرس في الدرس وحمس ح في نفسه
بكل الفاعله الراويه بلهانه وسن
فلان الذي به يكون فخرية مانه و

GT UND MODELLE

- arabische Drucke: OpenTI
- arabische Handschriften: RASM2018/RASM2019

OPENITI

- [github: OpenITI/OCR_GS_Data](#)
- arabische Drucke: ~7k Zeilen
- Zeilenbilder, binarisiert

RASM

- [doi: 10.23636/1135](https://doi.org/10.23636/1135)
- ICDAR: Recognition of Historical Arabic Scientific Manuscripts
- Seitenbilder, Farbe, PAGE XML
- Problem: keine Rotationswinkel
- ~2.6k Zeilen nutzbar ohne Rotationskorrektur

MODELLE

- [github: Calamari-OCR/calamari_models_experimental](#)
- def_arabic: auf Daten des Arabic Latin Corpus trainiert
- weitere Modelle für ar. Texte werden dort in nächster Zeit veröffentlicht

EXKURS: ALIGNIERUNG EXISTIERENDER GT

- 6k von 12k Zeilen bereits transkribiert
- Seiten- und Zeilenumbrüche in Transkription markiert
- Problem: Fehler in Zeilensegmentierung und hohe Fehlerrate bestehender Modelle → Gefahr von Fehlalignierungen und dadurch fehlerhafter GT

STRATEGIE

1. OCR mit bestehendem Modell
2. Alignierung mit optimaler Übereinstimmung
3. Übernahme von Zeilen mit einer unter einem Schwellwert liegenden CER als GT
4. Training eines neuen Modells
5. zurück zu 1.

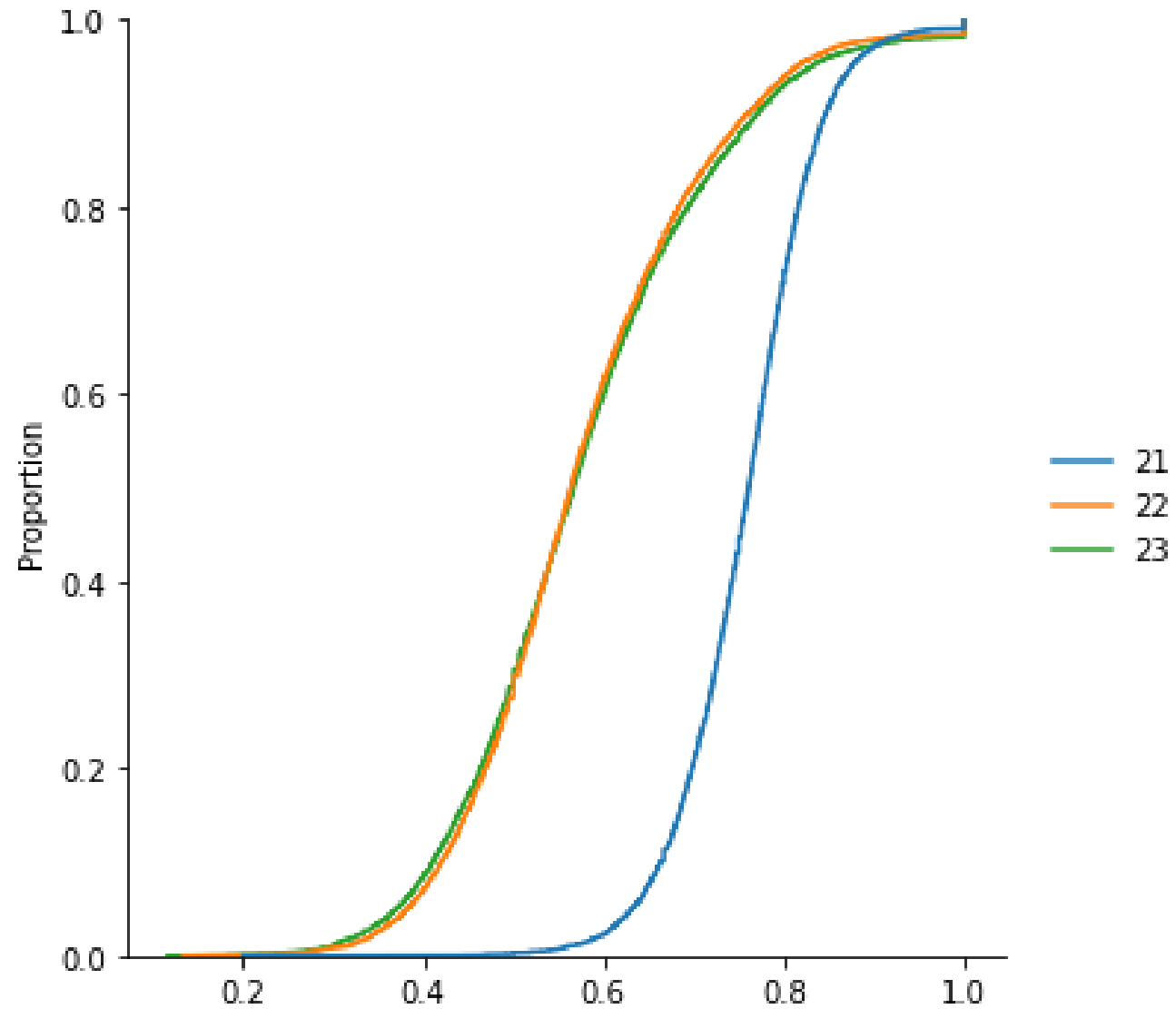
IMPLEMENTIERUNG DER ALIGNIERUNG

- einspaltiges Layout, einfache Sortierung der Zeilensegmente → Sequenzalignierung
- Alignierung Zeile-Zeile ähnlich Needleman–Wunsch, dabei Minimierung des Levenshtein-Abstands der Zeilen
- bei "Equal" oder "Replace" mit ausreichend hoher Ähnlichkeit: Transkription gilt als Ground Truth

MODELLTRAINING

- jeweils Set aus fünf Modellen CF-validiert
- fünffache Augmentierung
- $cnn=40:3 \times 3, pool=2 \times 2, cnn=60:3 \times 3, pool=2 \times 2,$
 $cnn=120:3 \times 3, lstm=200, lstm=200, lstm=200, dropout=0.5$
- drei Modelle (CER jeweils auf Validierungsdaten):
 - Drucke (OpenITI): avg. CER 1.66%
 - Handschriften (RASM): avg. CER 16.77%
 - Drucke, dann Handschriften: avg. CER 15.84%

ALIGNIERUNG 1: THRESHOLD?



21: Drucke, 22: MSS, 23 Drucke+MSS.

Anteil der GT-Zeilen in Abhängigkeit von der CER.

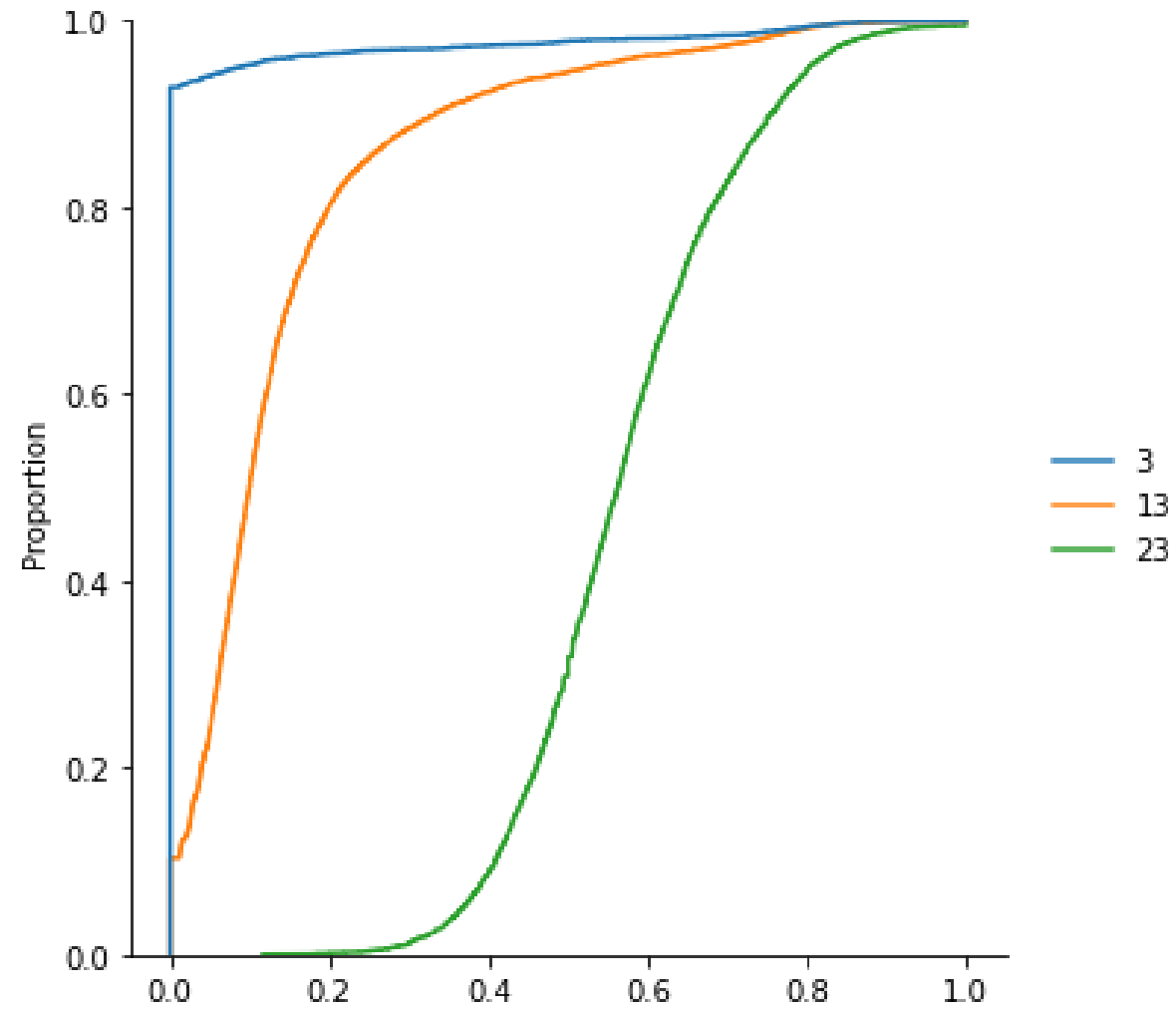
→ Modell 3, Threshold 0.4, 493/6087 Zeilen

FINETUNING 1, ALIGNIERUNG 2

- avg. val. CER 9.05%
- → Threshold 0.4, 5719/6087 Zeilen

FINETUNING 2, ALIGNIERUNG 3

avg. val. CER 3.52%



23: Drucke+MSS, 13: Finetuning 1, 3: Finetuning 2.

Anteil der GT-Zeilen in Abhängigkeit von der CER.

→ Threshold 0.1, 5877/6087 Zeilen

GT	PRED	COUNT	PERCENT
{}	{}	35	5.25%
{ب}	{}	20	3.00%
{ا}	{}	19	2.85%
{ي}	{}	17	2.55%
{ت}	{}	15	2.25%
{و}	{}	13	1.95%
{ل}	{}	13	1.95%
{ن}	{}	10	1.50%
{ت}	{ي}	10	1.50%
{م}	{}	10	1.50%
{ي}	{ت}	9	1.35%
{ }	{}	8	1.20%
{أ}	{}	8	1.20%
{ث}	{}	7	1.05%
{ر}	{}	7	1.05%
{ن}	{ي}	6	0.90%
{إ}	{أ}	6	0.90%
{ف}	{}	6	0.90%
{د}	{}	5	0.75%
{ء}	{}	5	0.75%

The remaining but hidden errors make up 65.67%

AUSBLICK

- Training und Evaluation von spezifischen Segmentierungsmodellen
- weitere ar. Handschriften, z.B. Tunis, Dār al-kutub al-waṭaniyya, 7116
- Veröffentlichung von vortrainierten OCR-Modellen