DIE VIELFÄLTIGEN HERAUSFORDERUNGEN FÜR AUTOMATISIERTE TEXTERKENNUNG AUF DOKUMENTEN DES HERDER-INSTITUTS

- Ziel: Automatisierte Erstellung von Volltexten zu Beständen des Herder-Instituts
- … als Grundlage für weitere NLP- und DH-Verfahren (NER, Topic Modelling etc.)
- Bestände sind…
  - sehr heterogen (zeitlich, Medientypen, Inhalte, …)
  - vergleichsweise kleinteilig -> Arbeitsaufwand GT-Erstellung etc. für ‚maßgeschneiderte' Modelle hoch in Relation zum Output
  - nur in Teilen und nicht systematisch digitalisiert
- Geringe Personalausstattung

WAS BISHER GESCHAH…

- Auswahl von zwei exemplarischen Dokumenten für HTR/OCR:
  - Briefwechsel Julie und Eduard von Oettingen, DSHI 190 Livland 33, 11c, ~3000 Seiten (Handschrift 2. Hälfte 19. Jhd., siehe Abb. 1)
  - Georg v. Krusenstjern: Einsatz. Tagebuchblätter, Briefe und Notizen aus dem zweiten Weltkriege 1941-1945, DSHI 190 Krusenstjern 2002, ~400 Seiten (Typoskript Mitte 20. Jhd., siehe Abb. 2).
- Erstellung von Evaluationskorpora (50seitige Samples, händisch transkribiert)
- Erprobung von off-the-shelf HTR/OCR-Modellen u.a. mit Transkribus, ocr4all, recognAIze, nopaque und Abby FineReader
- Training/Finetuning von Modellen zur Layouterkennung für Handschrift in Transkribus
- Training/Finetuning von OCR-Modellen für Typoskript in eScriptorium

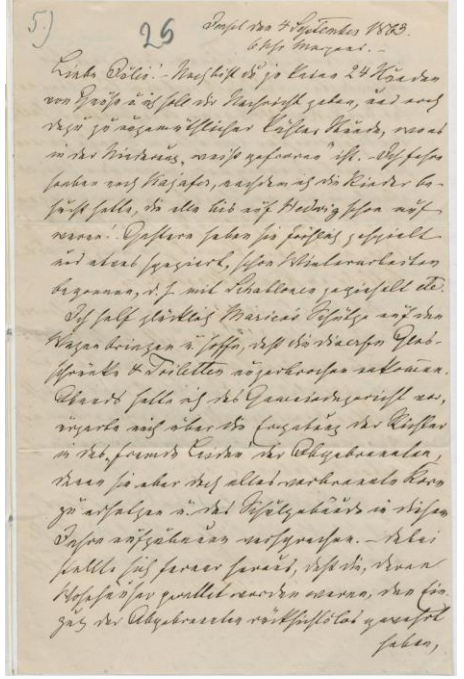
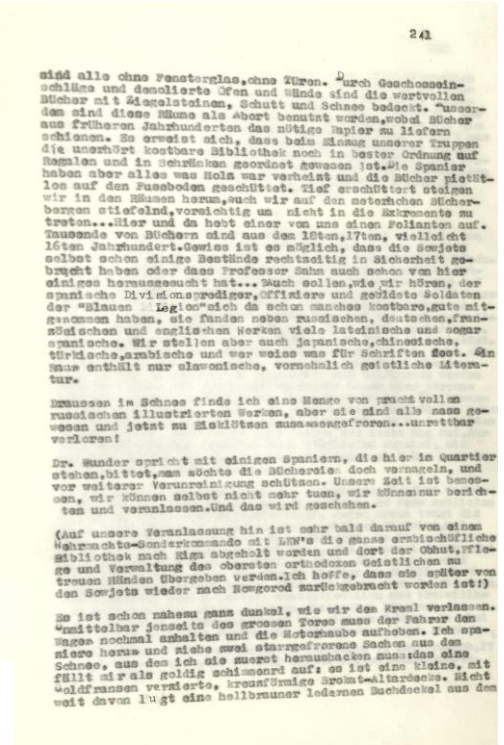


Abb. 1: Brief Eduard v. O. an Julie v. O., 4. September 1863



241

sind alle ohne Fensterglas,ohne Türen. Durch Geschosseinschläge und demolierte Öfen und Wände sind die wertvollen Bücher mit Ziegelsteinen, Schutt und Schnee bedeckt. Ausserdem sind diese Räume als Abort benutzt worden,wobei Bücher aus früheren Jahrhunderten das nötige Papier zu liefern schienen. So erweist sich, dass beim Rückzug unserer Truppen die unersetzt kostbare Bibliothek noch in bester Ordnung auf Regalen und in Schränken geordnet gewesen ist.Die Spanier haben aber alles was Holz war verheizt und die Bücher plötzlich auf den Fussboden geschüttet. Tief erschüttert stehen wir in den Räumen herum,auch wir auf den meterhohen Bücherbergen stiefelnd,vorsichtig um nicht in die Krippenste zu treten...Hier und da hebt einer von uns einen Folianten auf. Tausende von Büchern sind aus dem 16ten,17ten, vielleicht 16ten Jahrhundert.Gewiss ist es möglich, dass die Sowjets selbst schon einige Bestände rechtzeitig in Sicherheit gebracht haben oder dass Professor Sahs auch schon von hier einiges herausgeschafft hat... Doch sollen,wie wir hören, der spanische Divisionsprediger,Offiziere und geadelte Soldaten der "Blauen Legion"sich da schon manches kostbare,gute mitgenommen haben, sie fanden neben russischen, deutschen,französischen und englischen Werken viele lateinische und sogar spanische. Wir stellen aber auch japanische,chinesische, türkische,arabische und vor weiss was für Schriften fest. Ein Raum enthält nur slawonische, vornehmlich geistliche Literatur.

Draussen im Schnee finde ich eine Menge von prachtvollen russischen illustrierten Werken, aber sie sind alle nass gewesen und jetzt zu Eisklötzen zusammengefroren...unrettbar verloren!

Dr. Wunder spricht mit einigen Spaniern, die hier in Quartier stehen,bittet,man möchte die Bücherberge doch vernageln, und vor weiterer Verunreinigung schützen. Unsere Zeit ist bemessen, wir können selbst nicht mehr tun, wir können nur berichten und veranlassen.Und das wird geschehen.

(Auf unsere Veranlassung hin ist sehr bald darauf von einem Wehrmachts-Sonderkommando mit LKW's die ganze orthodoxe Bibliothek nach Riga abgeholt worden und dort der Obhut,Pflege und Verwaltung des obersten orthodoxen Geistlichen zu treuen Händen übergeben worden.Ich hoffe, dass sie später von den Sowjets wieder nach Nowgorod zurückgebracht worden ist!)

Es ist schon nahezu ganz dunkel, wie wir den Kreml verlassen. Unmittelbar jenseits des grossen Tores muss der Fahrer den Wagen nochmal anhalten und die Motorhaube aufheben. Ich spaziere herum und suche zwei eingefrorene Sachen aus den Schnee, aus den ich die zweite mühsam herausbacken muss:das eine fällt mir als goldig schimmernd auf: es ist eine kleine, mit Goldfransen verzierte, kreuzförmige Brokat-altardecke. Nicht weit davon liegt eine hellbraune ledernen Buchdeckel aus den

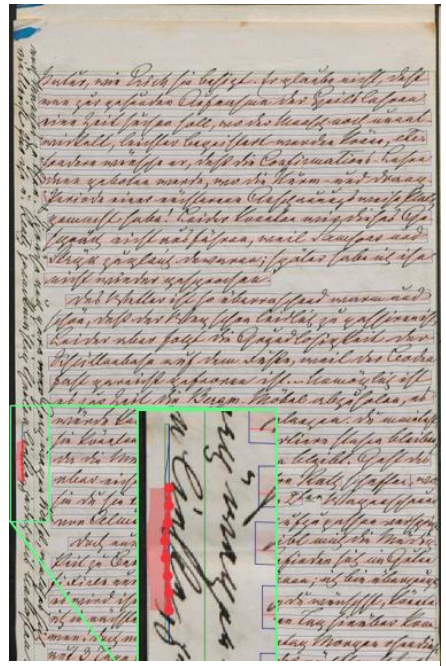
Abb. 2: Krusenstjern, Einsatz, S. 241

ERSTE ERGEBNISSE UND FRAGEN

- **Erste Ergebnisse**:
  - off-the-shelf Modelle mit sehr unterschiedlichen Ergebnissen

	Transkribus	eScriptorium	Abby	Tesseract	Nopaque
CER	**1.96%**	9.79%	17.91%	17.51%	23.07%
WER	**14.99%**	44.73%	49.54%	50.57%	59.83%

Ergebnisse Evaluation OCR auf 25 Seiten-Sample Bericht Krusenstjern, jeweils bestes Model



Screenshot Transkribus, Layout Analysis mit selbst trainiertem Model, aufgenommen 20.5.2024

  - Ergebnisse für Typoskript mit minimalen Trainingsdaten (10 Seiten) deutlich optimierbar (mittels eScriptorium: CER **2.58%**, WER: **14.14%**)
  - Herausforderungen: Layout-Erkennung, v. a. für Handschriften; Evaluierung von Layout Recognition
  - Für ‚kleine' Sprachen fehlen tw. Modelle v. a. für historische Sprach- und Schriftstufen (u. a. Polnisch)
  - Bildqualität vermutlich ein Faktor, aber schwer zu beziffern

Fragen

  - Layout-Erkennung ein Problem, v. a. bei Handschriften, aber auch bei schlechten Scans. Wie hiermit umgehen? Frage der Trainingsdatenmenge oder grundsätzlicheres Problem?
  - Wie Layout-Erkennung evaluieren?
  - Trainings-Daten-Erstellung: Standards für Transkriptionen (und Layouts) um Kompatibilität mit bestehenden Trainingsdaten zu gewährleisten?
  - Ressourcen für osteuropäische Sprachen (Modelle, Trainingsdaten, …)?