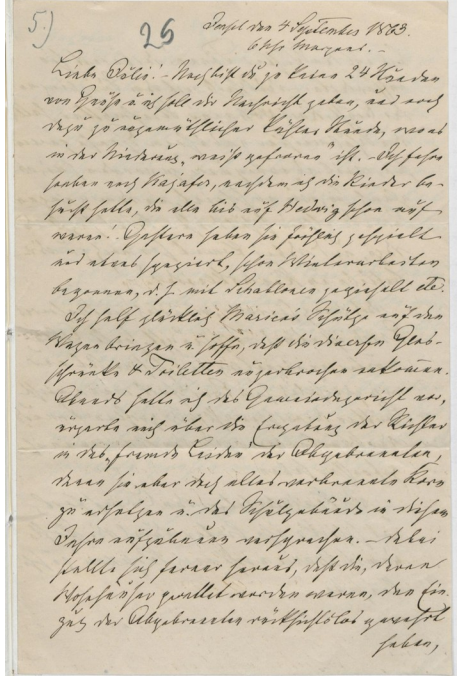
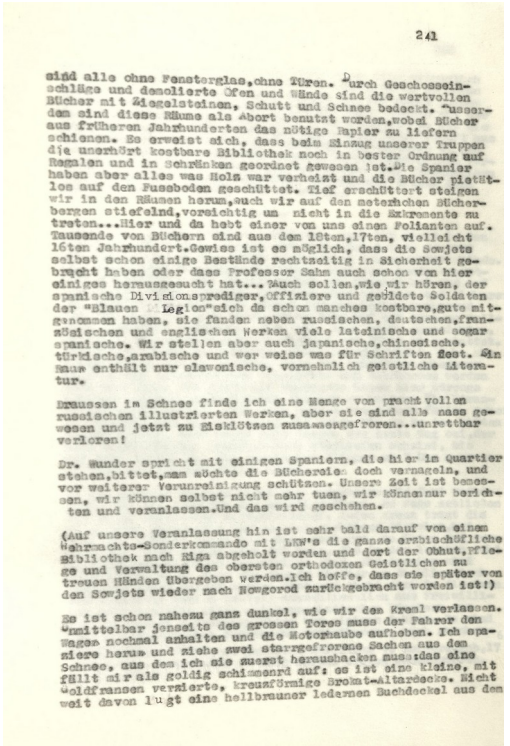
More Data – More Problems? Die vielfältigen Herausforderungen für Automatisierte Texterkennung auf Dokumenten (nicht nur) des Herder-Instituts

Ole Meiners
Arbeitsbereich Forschungsdatenmanagement
Abteilung Digitale Geschichte und Informationssysteme
Herder-Institut für historische Ostmitteleuropaforschung

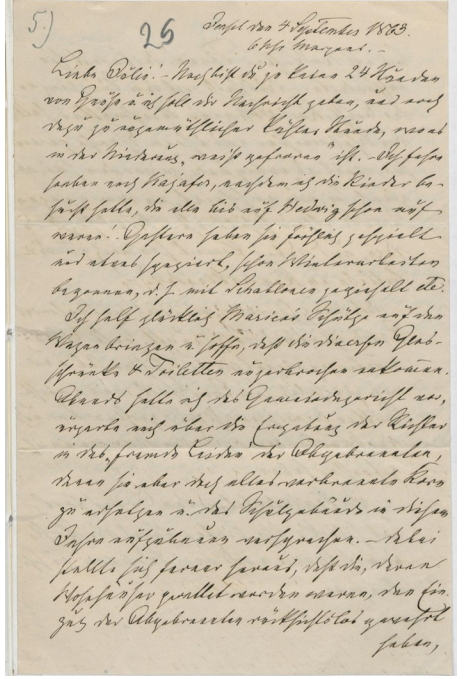
DIE VIELFÄLTIGEN HERAUSFORDERUNGEN FÜR AUTOMATISIERTE TEXTERKENNUNG AUF DOKUMENTEN DES HERDER-INSTITUTS

- 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 -> hoher Arbeitsaufwand GT-Erstellung etc. für ‚maßgeschneiderte' Modelle in Relation zum Output
  - nur in Teilen und nicht systematisch digitalisiert

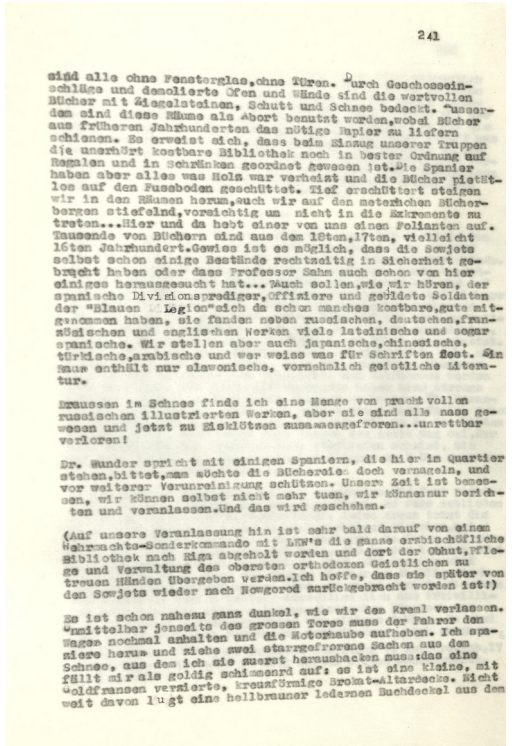




DIE VIELFÄLTIGEN HERAUSFORDERUNGEN FÜR AUTOMATISIERTE TEXTERKENNUNG AUF DOKUMENTEN (NICHT NUR) DES HERDER-INSTITUTS



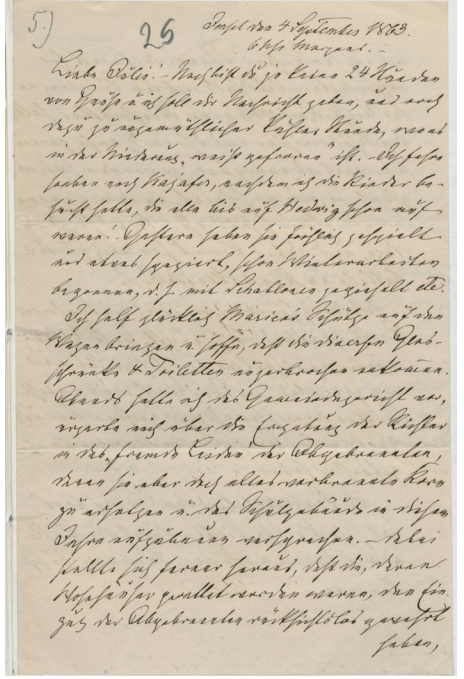
- Ziel: Automatisierte Erstellung von Volltexten zu Beständen des Herder-Instituts und anderen Textdigitalisaten
- … als Grundlage für weitere NLP- und DH-Verfahren (NER, Topic Modelling etc.)
- Bestände sind…
  - sehr heterogen (zeitlich, Medientypen, Inhalte, …)
  - vergleichsweise kleinteilig -> hoher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

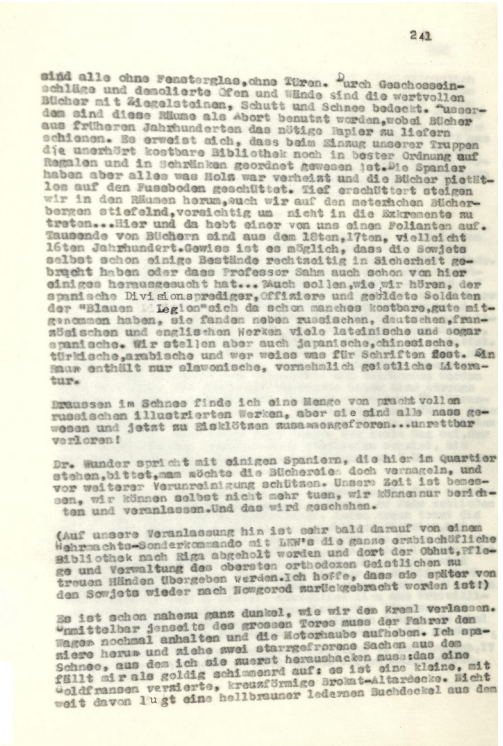


WAS BISHER GESCHAH…

- 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~~, eScriptorium, ocr4all, ~~recognAlze~~, ~~nopaque~~, PERO und Abbyy FineReader
- (Nach-)Training/Finetuning von Modellen zur Layouterkennung für Handschrift in Transkribus
- (Nach-)Training von OCR-Modellen für Typoskript über 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 unberührt 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 planlos auf den Fussboden geschüttet. Tief erschüttert steigen wir in den Räumen herum,auch wir auf den meterhohen Bücherbergen stiefelnd,vorsichtig um nicht in die Kaminente zu treten...Hier und da hebt einer von uns einen Folianten auf. Tausende von Büchern sind aus dem 18ten,17ten, vielleicht 16ten Jahrhundert.Gewiss ist es möglich, dass die Sowjets selbst schon einige Bestände rechtzeitig in Sicherheit gebracht haben oder dass Professor Sahn auch schon von hier einiges herausgesucht hat... Doch sollen,wie wir hören, der spanische Divisionsprediger,Offiziere und gebildete 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 wer 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 Bücherei 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 erzbischöfliche 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 das Kreml verlassen. Unmittelbar jenseits des grossen Tores muss der Fahrer den Wagen nochmal anhalten und die Motorhaube aufheben. Ich spaziere herum und ziehe zwei steifgefrorene Sachen aus den Schnee, aus dem ich sie 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 ein hellbrauner ledernen Buchdeckel aus dem

Abb. 2: Krusenstjern, Einsatz, S. 241

WAS SEITDEM GESCHAH…

- Texterkennung für Typoskript mit nachtrainiertem Model ‚einsatzfähig‘
  (CER: 1,82% WER: 8,87% auf Evaluationsdatensatz)
    - … Rechtslage leider unklar daher (noch) nicht publikationsfähig
- HTR auf Briefkorpus wegen Problemen bei Layout-Erkennung zunächst auf Eis gelegt (multi-direction layout recognition bislang nicht umsetzbar)

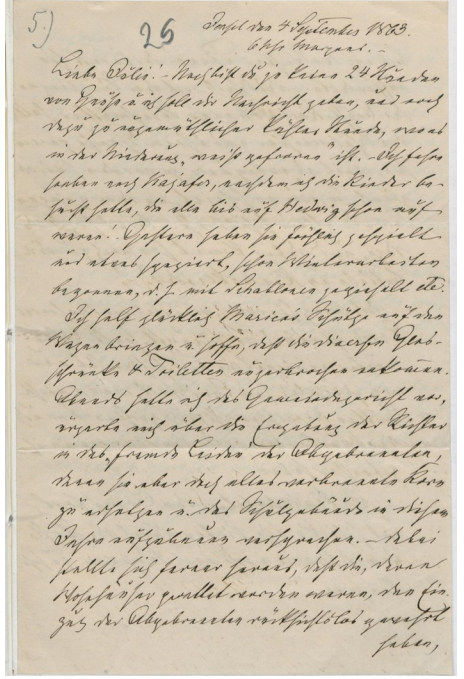
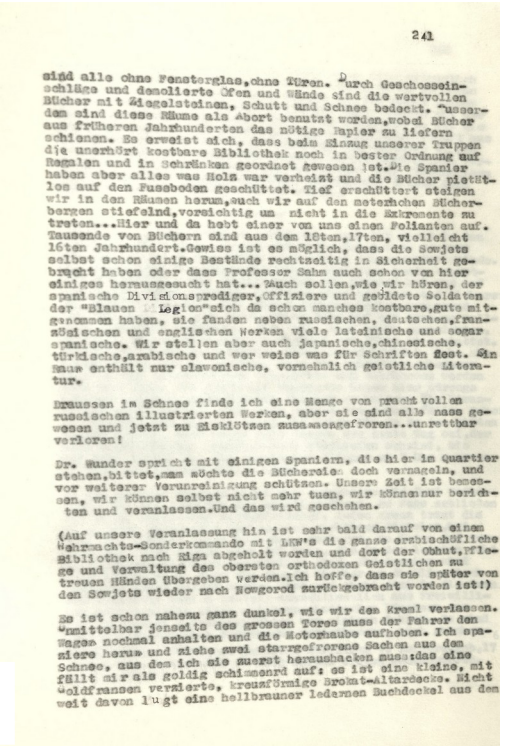


Abb. 1: Brief Eduard v. O. an Julie v. O., 4. September 1863



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 Einzug unserer Truppen die unerhört 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 steigen wir in den Räumen herum,auch wir auf den meterhohen Bücherbergen stiefelnd,vorsichtig um nicht in die Kaminecke 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 Sahn auch schon von hier einiges herausgesucht hat... Doch sollen,wie wir hören, der spanische Divisionsprediger,Offiziere und gebildete 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 wer weiss was für Schriften fest. Ein Haus 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 Bücherei 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 erzbischöfliche 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 das Kreml verlassen. Unmittelbar jenseits des grossen Tores muss der Fahrer den Wagen nochmal anhalten und die Motorhaube aufheben. Ich spaziere herum und ziehe zwei eingefrorene Sachen aus den Schnee, aus den ich sie mühsam herausbacken muss:das eine fällt mir als goldig schimmernd auf: es ist eine kleine, mit Goldfransen verzierte, kreuzförmige Brokat-altardecke. Nicht weit davon lugt eine hellbrauner ledernen Buchdeckel aus den

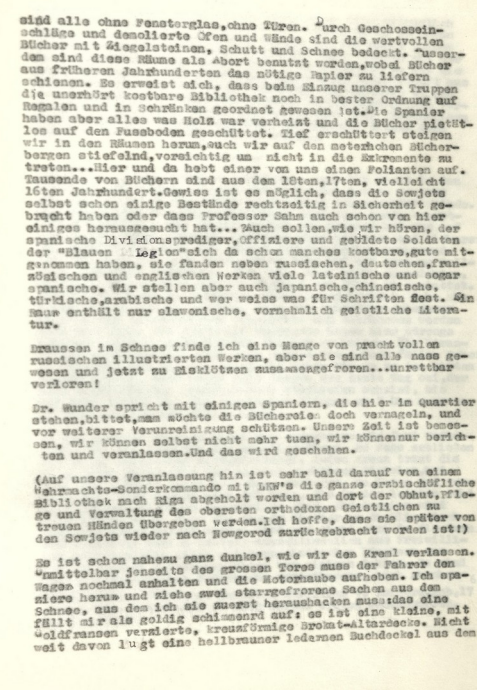
Abb. 2: Krusenstjern, Einsatz, S. 241

WAS SEITDEM GESCHAH…

- Texterkennung für Typoskript mit nachtrainiertem Model ‚einsatzfähig‘
  (CER: 1,82% WER: 8,87% auf Evaluationsdatensatz)
  - … Rechtslage leider unklar daher (noch) nicht publikationsfähig
- HTR auf Briefkorpus wegen Problemen bei Layout-Erkennung zunächst auf Eis gelegt (multi-direction layout recognition bislangnicht umsetzbar)
- Drittes Einsatzszenario: (Vertriebenen-)Zeitungen
  - u. a. „Ostpreußenblatt“ (1950-2003), „Preußische Allgemeine Zeitung“ (2003-2023)
  - „Sudetendeutsche Zeitung“ (derzeit: 1951-1955)
  - … und viele andere mehr (noch nicht digitalisiert)
  - Aktuell vorliegend: > 3850 Ausgaben, > 83.000 Seiten


Das Ostpreußenblatt
Organ der Landsmannschaft Ostpreußen
Gleiches Recht!
Die Stadtkirche von Wartenburg
Preußische Toleranz

Abb. 1: Titelseite „Das Ostpreußenblatt, Jg. 11, Folge 3, 16.01.1960


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 steigen wir in den Räumen herum,auch wir auf den meterhohen Bücherbergen stiefelnd,vorsichtig um nicht in die Reinecke 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 Sahn auch schon von hier sind,was heraussucht hat... Doch sollen,wie wir hören, der spanische Divisionsprediger,Offiziere und gewidete 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 wer weiss was für Schriften fest. Ein Raum enthält nur slavonische, 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 Bücherei 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 archivalische 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 das Kreml verlassen. Unmittelbar jenseits des grossen Tores aus der Fahrer den Wagen nochmal anhalten und die Motorhaube aufheben. Ich spaziere herum und ziehe zwei steifgefrorene Sachen aus den Schnee, aus den ich alle zuerst herausbacken muss:das eine fällt mir als goldig schimmernd auf: es ist eine kleine, mit Goldfransen verzierte, kreuzförmige Brokat-Altardecke. Nicht weit davon lugt eine hellbrauner ledernen Buchdeckel aus den

Abb. 2: Krusenstjern, Einsatz, S. 241

ZEITUNGEN

- Layout wie erwartet schlecht zu erkennen
  - … beste Ergebnisse bislang mit ABBYY FineReader und PERO OCR
  - … auch hier allerdings Reading Order mitunter stark fehlerhaft

Text- und Layout Erkennung Abbyy FineReader, Sudetendeutsche Zeitung, Jg. 1 Folge 01, 07.04.1951

Text- und Layout Erkennung PERO OCR, Sudetendeutsche Zeitung, Jg. 1 Folge 01, 07.04.1951

ZEITUNGEN

- Layout wie erwartet schlecht zu erkennen
  - … beste Ergebnisse bislang mit ABBYY FineReader und PERO OCR
  - … auch hier allerdings Reading Order mitunter stark fehlerhaft
- Herausforderungen:
  - Erstellung Ground Truth sehr arbeitsaufwändig
  - Layout-Erkennungsmodelle von ABBYY und PERO ermöglichen kein Finetuning/Adaption?
  - Gleichzeitig hohe Anforderungen an Erkennungsqualität (v.a. Text, mittelbar Layout) durch Forschungsprojekt

Text- und Layout Erkennung Abbyy FineReader, Sudetendeutsche Zeitung, Jg. 1 Folge 01, 07.04.1951

Text- und Layout Erkennung PERO OCR, Sudetendeutsche Zeitung, Jg. 1 Folge 01, 07.04.1951

FRAGEN

- Problem Layout-Erkennung; bei multi-direktionalen (Hand-)schriften, v. a. aber bei komplexen Zeitungs-Layouts.
- (Wie Layout-Erkennung evaluieren?)
- Tools für Layout-GT-Erstellung (eScriptorium zu umständlich für komplexe Layouts; LAREX? Weitere?)
- Publikation Evaluations-Datensätze? Verwendung als Trainingsmaterial nicht intendiert
- … und was ist überhaupt realistisch zu erreichen/wo lohnt der eigene Arbeitsaufwand? → v. a. für Layout-Erkennung

Vielen Dank für die Aufmerksamkeit!

Ole Meiners
Arbeitsbereich Forschungsdatenmanagement
Abteilung Digitale Geschichte und Informationssysteme
Herder-Institut für historische Ostmitteleuropaforschung