



Perspektiven der automatischen Texterfassung als Grundlage wissenschaftlicher Editionen

am Beispiel der Brief- und Schriftenausgabe
der Bernd Alois Zimmermann-Gesamtausgabe

Matthias Boenig, Hemma Jäger, Matthias Pasdzierny, Kay-Michael Würzner
{boenig|hemma.jaeger|pasdzierny|wuerzner}@bbaw.de

Geisteswissenschaftliche Forschungsdaten.
Methoden zur digitalen Erfassung, Aufbereitung und Präsentation
19. Oktober 2017



Übersicht

- Einleitung
 - ▶ die BERND ALOIS ZIMMERMANN-Gesamtausgabe
 - ▶ automatische Texterfassung
 - ▶ Motivation
- Workflowbeschreibung
 - ▶ Bildvorverarbeitung
 - ▶ Layoutanalyse
 - ▶ Zeichenerkennung
 - ▶ Textbearbeitung
- Perspektiven
 - ▶ Volltextverbesserung
 - ▶ OCR-D
 - ▶ Editionsunterstützung



Die Bernd Alois Zimmermann-Gesamtausgabe



Matthias



Die Bernd Alois Zimmermann-Gesamtausgabe



Matthias





Einleitung



Automatische Texterfassung

- Huge amount of text sources not available in *digital* form
 - ▶ Historical books, newspapers, and papers from the printing era
 - ▶ Hand-written manuscripts and letters
- Conservation by scanning or photographing
 - ▶ Images but **no text data**
- ⇒ Text retrieval or quantitative analyses impossible
- Automatic capturing of **text** within images
- Automatic capturing of **layout** within images (aka. Optical Layout Recognition)





Workflowbeschreibung



Übersicht



- Prozesse zur bestmöglichen Vorbereitung der Digitalisate für OLR und OCR
 - ▶ **Cropping:** Beschneidung des Digitalisats auf den Druckbereich
 - ▶ **Deskewing:** Rotation des Digitalisats zur Begradigung von Schrägstellungen
 - ▶ **Binarization:** Binäre Kodierung der Pixel (bedruckte Bereiche schwarz, nicht-bedruckte Bereiche weiß)
 - ▶ **Despeckling:** Entfernung von Bildartefakten (Verschmutzungen, sichtbare Papiermaserung etc.)
 - ▶ **Dewarping:** Begradigung von Wellen auf Zeilenebene
- starker Einfluss auf die Erkennungsqualität
- besondere Relevanz für historische Vorlagen



Bildvorverarbeitung: ScanTailor



- Prozesse zur Erkennung der Struktur auf Seiten- und Dokumentebene
 - ▶ **Page Segmentation:** Lokalisierung von zusammenhängenden Text- und Nichttextbereichen
 - ▶ **Region Classification:** Typisierung von Textbereichen
 - ▶ **Line/Character Splitting:** Lokalisierung der einzelnen Zeilen/Zeichen
 - ▶ **Document Analysis:** Konstruktion der logischen Dokumentstruktur (METS!)
- entscheidend für die korrekte **Rekonstruktion des Textflusses** und damit für maschinelle Auswertungen



Layoutanalyse: LAREX



- Kernkomponente der OCR
- Genauigkeit beeinflusst vom Typ des zugrundeliegenden **Algorithmus** und vom eingesetzten **Modell**
- aktuell Paradigmenwechsel: **zeichenorientiert** → **zeilenorientiert**
 - ▶ **Deep learning:** Tiefe (i.e. vielschichtige) neuronale Netzwerke zur Sequenzklassifizierung *(Hochreiter und Schmidhuber 1997)*
 - ▶ wesentlich weniger anfällig für **Zeichenvarianz**
 - ▶ eingebautes **Sprachmodell**
- auch schwierige historische Vorlagen in „OCR-Reichweite“ *(Springmann 2016)*

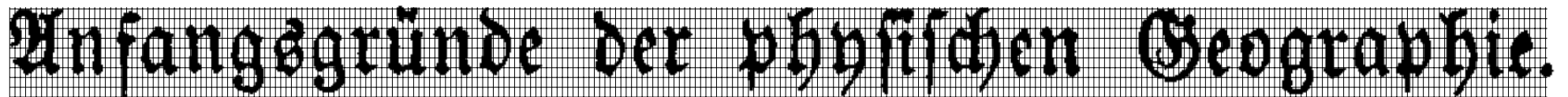


Sequence-focused recognition

- Targets one *line* of glyphs

Scaling: Uniform height for all lines

Feature extraction: Fixed number of horizontal rows, variable number of vertical columns: lines as sequences of binary-valued fixed-length vectors



Anfangsgründe der physischen Geographie.

- Context dependent (i.e. *transition probabilities*) recognition (requires larger amounts of training material → DTA)
- Segmentation into lines as pre-processing step
- Usually more robust to variance than character-focused approaches
- Open-source software OCRopus
 - ▶ Uses neural networks for sequence classification



Zeichenerkennung: Textvereinigung



Textbearbeitung



Matthias



Textbearbeitung: Oxygen



Matthias





Perspektiven



Volltextverbesserung







Matthias





Danke für Ihre Aufmerksamkeit!

