Schwierigkeiten bei OCR

- Buchformate
- Buchzustand
- Sprachvielfalt
- Unterschiedliche Drucktechniken
- Unterschiedliche Schriftarten
- Unterschiedliche Druckintensität
- Knitterfalten (durch Feuchtigkeit)
- bei Scans
    - gewölbte Buchseiten
    - schräge/verzerrte Scans
    - verkrümmte Textlinien
    - störende Farbflecken
    -
    - Schwarzer Rand
    - Durchscheinen
- bei Layouterkennung
    - Platzierung Seiten-/Kapitelzahl
    - Handschriftliche Anmerkungen
    - komplexes Layout bei Zeitungen
    - logische Reihenfolge der Textblöcke
    - Leserichtung
    - Tabellen

→ Bildvorbearbeitung wichtig

Lexika

- wesentlicher Bestandteil von OCR-Erkennung
- laufen während der OCR-Erkennung mit
- Ohne Lexika: fehlerbehaftet, weil Engine (Abbyy/tesseract) keine Vorgabe hat, mit der abgeglichen werden kann u. sich daher falsche Begrifflichkeiten herauspickt (→ Konfidenz/Vertrauen)
- Abbyy bringt standardmäßig 80/90 Lexika mit, Auswahl versch. Sprachen mgl., aber im Großen u. Ganzen: black box
- Google: steckt derzeit viel Energie u. Geld in die open source Engine tesseract
- Hypothetische Lexika: werden aus historischen Varianten von Wörtern gebildet (z. B.: th→ t)
- Für das Einbinden eigener Lexika sind Schnittstellen nötig
    - Tesseract: Einbinden von eigenen Wortlisten mgl.
- Ideal/noch Zukunftsmusik: Lexika auf bestimmten Zeitraum abgestimmt anbieten

Strukturerkennung/Layout

- Druckbereich
- Überschriften
- Fußnoten
- Platzierung Kapitel-, Seitenzahl
- Anmerkungen
- Verzeichnisse

Was alles zur OCR-Erkennung gehört

- Zusätzliches
    - Lexikonerstellung
    - Eigennamenerkennung

    - Nachkorrektur ((halb-)automatisch, kollaborativ)
- Hinsichtlich der Segmentierung
    - Segmentierung der Vorlage in Zeichen, Wort, Zeilen
    - Layoutsegmentierung (bei Zeitungsartikeln relevant)
- Hinsichtlich der Scans
    - Randerkennung
    - Geometrische Korrektur (Gerade-Rücken)
    - Farbreduktion und Binarisierung (Umwandlung des Bildes in s/w-Bild/Binärbild; es geht hierbei um die Trennung der Objekte, die geprüft werden sollen, vom Untergrund)
- Strukturanalyse
- Evaluation

→ An erster Stelle steht immer die Aufgabe: Material klassifizieren und kategorisieren → je nach Eigenarten des Materials ergibt sich ein anderer Workflow

Verarbeitungskette in der OCR
- Bildvorbearbeitung
- Segmentierung
- OCR
- Nachkorrektur
- Evaluation
- Strukturanalyse

Interoperabilität und Ausgabeformate
- Abbyy xml
- METS/ALTO: Buchstabenpositionen können gerahmt werden; Polygone besser, da dadurch eine bessere Beschreibung mgl. ist
- hOCR: Format, welches von Google/Tesseract kreiert wird
- TEI
- Page xml: im Rahmen von IMPACT entwickeltes xml

Warum Bildvorverarbeitung
- Publikation von optimierten Scans im Netz
- Ausdruck von optimierten Scans (print on demand)
- Höhere Erkennungsgenauigkeit → verbesserte OCR

OCR-Praxis
- Bildvorbearbeitung, z. B. durch
    - Randentfernung
    - Geometrische Korrektur
    - Binarisierung
- ! Tiff-Header f. OCR wichtig (sind u. a. Infos über Auflösung drin)
- eine manuelle Korrektur des OCR-Ergebnisses ist immer nötig
- Wichtig: Vorarbeit nötig → man muss den eigenen Bestand sehr genau sichten u. kennen
- Klären des Anwendungszwecks wichtig (z. B. grafische Hervorhebung v. Suchbegriffen)
- Vgl. Engines / Engines
    - open source – Ocropus (für Layoutsegmentierung) und Tesseract (für Texterkennung)
    - Kommerziell – Abbyy (vereint Layoutsegmentierung und Texterkennung)