WIEVIEL SIND 85% WERT:
QUALITÄT VON OCR UND NER-VERFAHREN FÜR DIE FORSCHUNG

Elisa Herrmann

24.05.2018

Qualitätsansprüche der DFG

- Praxisregeln „Digitalisierung“ 2/13:
  - >99,5% wissenschaftlich zuverlässig
  - 80-90% schlecht
  - <80% Gesamtnutzen fragwürdig
  - „Dabei herrscht wenig Einigkeit darüber, welche Messkriterien und -verfahren angelegt werden.“

- Praxisregeln „Digitalisierung“ 12/16:
  - „Eine Genauigkeit unter 95 % sollte möglichst nicht vereinbart werden.“
  - „Unterhalb einer Genauigkeit von 99,5 % ist bei manueller Erfassung ein Ergebnis ungenügend.“

Messungen zur Textgenauigkeit

- Fehlertypen:

ſreugenSchnld
erläſſt vns' mehr ,

Substitution

falsches Einsetzen eines Zeichens

Insertion

Einfügen eines Zeichens, das in der Originalvorlage nicht auftritt

Deletion

Zeichen wird gelöscht

Rejection

Zeichen wird übergangen

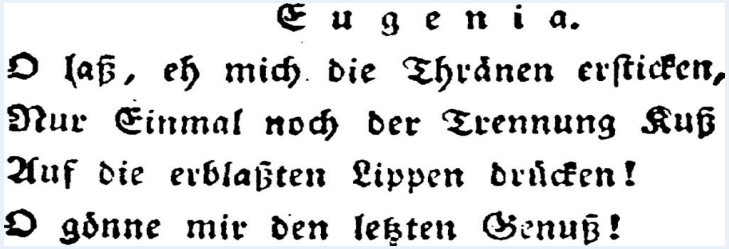
Messungen zur Textgenauigkeit

- Word Error Rate / Word Recognition Rate

$$WER = \frac{\#WERR}{\#WORD} \qquad WRR = \frac{\#CRW}{\#WORD}$$

- Character Error Rate / Character Recognition Rate

$$CER = \frac{\#CERR}{\#CHR} \qquad CRR = \frac{\#CRC}{\#CHR}$$



Eugenia.
O laß, eh mich die Thränen erſticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

Eugeaia.
O laß, H) mich. die Thtäaea erIHckm,
Nur Einqu noch der Trennung Kuß
Auf die erqußteu Lippen drückenx
O gönne mie den legten Genußx

WER: 0,28 (28%)

WRR: 0,72 (72%)

CER: 0,141 (14,17%)

CRR: 0,85 (85%)

Messungen zur Textgenauigkeit

- Precision

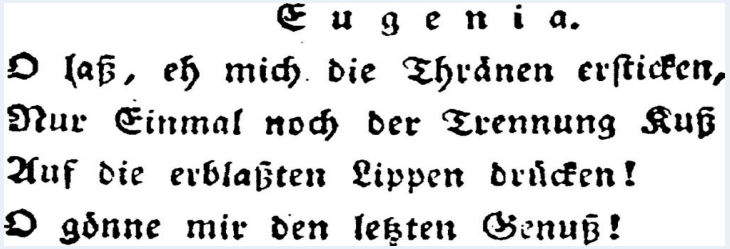
$$P = \frac{\#CCR}{\#COCR}$$

- Recall

$$R = \frac{\#CCR}{\#CHR}$$

- F-Maß

$$F - Maß = 2 * \frac{\frac{\#CCR}{\#CHR} * \frac{\#CCR}{\#COCR}}{\frac{\#CCR}{\#CHR} + \frac{\#CCR}{\#COCR}}$$


Eugenia.
O laß, eh mich die Thränen erſticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

Eugeaia.
O laß, H) mich. die Thtäaea erIHckm,
Nur Einqu noch der Trennung Kuß
Auf die erqußteu Lippen drückenx
O gönne mie den legten Genußx

Precision: 0,853 (85,34%)

Recall: 0,825 (82,5%)

F-Maß: 0,839 (83,9%)

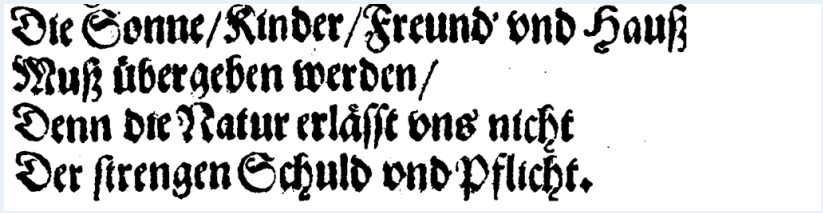
Qualitätsergebnisse

- Abhängig von:
  - Material & Vorverarbeitung
  - Engine
  - Training

- Allgemein:
  - 20. Jahrhundert problemlos 99%
  - Untrainiert, Segmentierungsansatz: <85%
  - Trainiert, Segmentierungsfrei: >95%

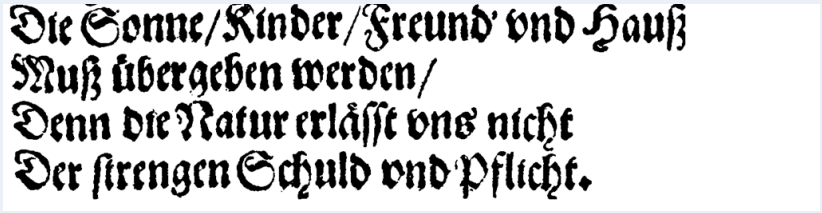
Abhängigkeiten - Preprozessing

Otsu (1979)

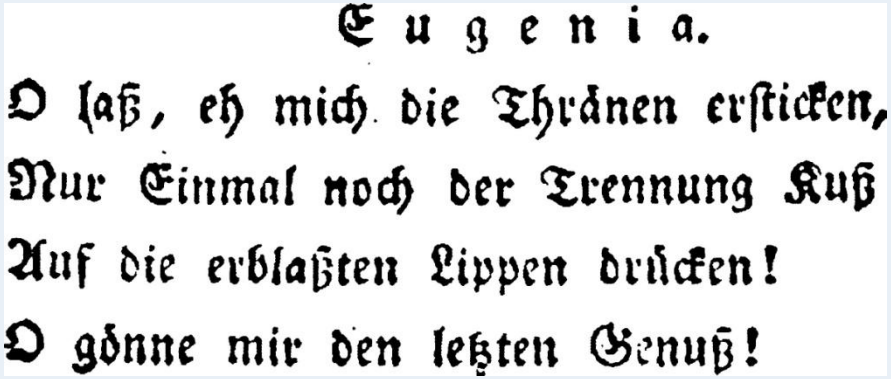

Die Sonne/Kinder/Freund' vnd Hauß
Muß übergeben werden/
Denn die Natur erläſſt vns nicht
Der ſtrengen Schuld vnd Pflicht.

Die Sonne/Kindrr/Frenud' vnd Hauß
Muß übergeben werden/ ' '
Denn dirNatnr erläſſt vns' mehr '
Der ſreugenSchnld ondPflichr.

Wolf und Doermann (2002)


Die Sonne/Kinder/Frennd' vnd Hauß
Muß übergeben werden/
Denn deeNainr erläſſt vns nicht
Der ſtrengen Schuld vndPflicht.

Abhängigkeiten - Verfahren



Eugenia.

O laß, eh mich die Thränen erſticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gönne mir den letzten Genuß!

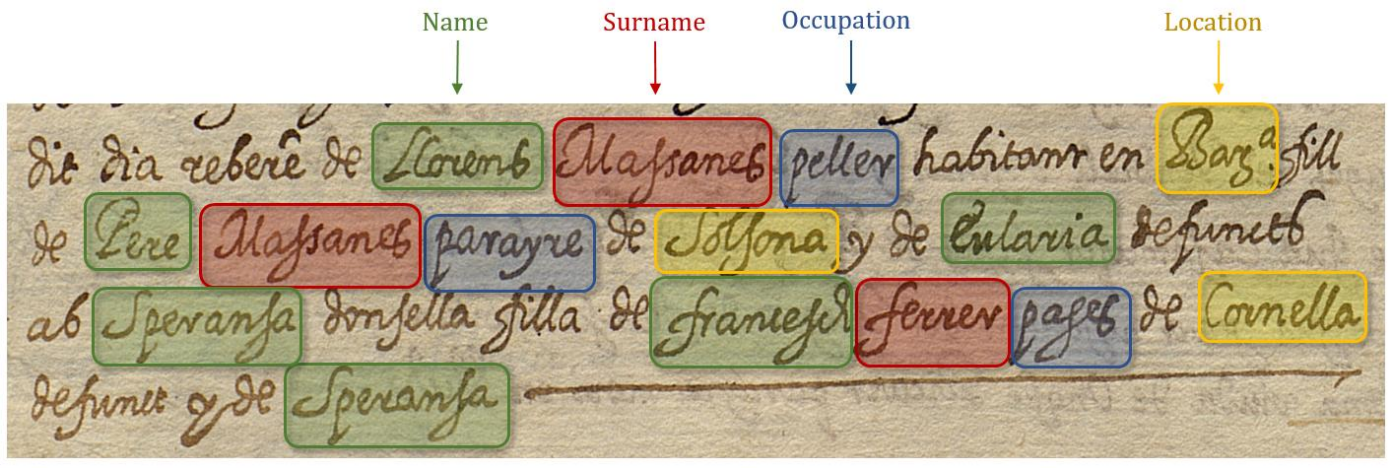
Tesseract (Smith 2007)

Eugeaia.
O laß, H) mich. die Thtäaea erIHckm,
Nur Einqu noch der Trennung Kuß
Auf die erqußteu Lippen drückenx
O gönne mie den legten Genußx

OCRopus (Breuel 2008)

E u g e n i a.
O haß, eh mich die Thränen ersticken,
Nur Einmal noch der Trennung Kuß
Auf die erblaßten Lippen drücken!
O gdnne mir den letzten Genuß!

Name Entity Recognition



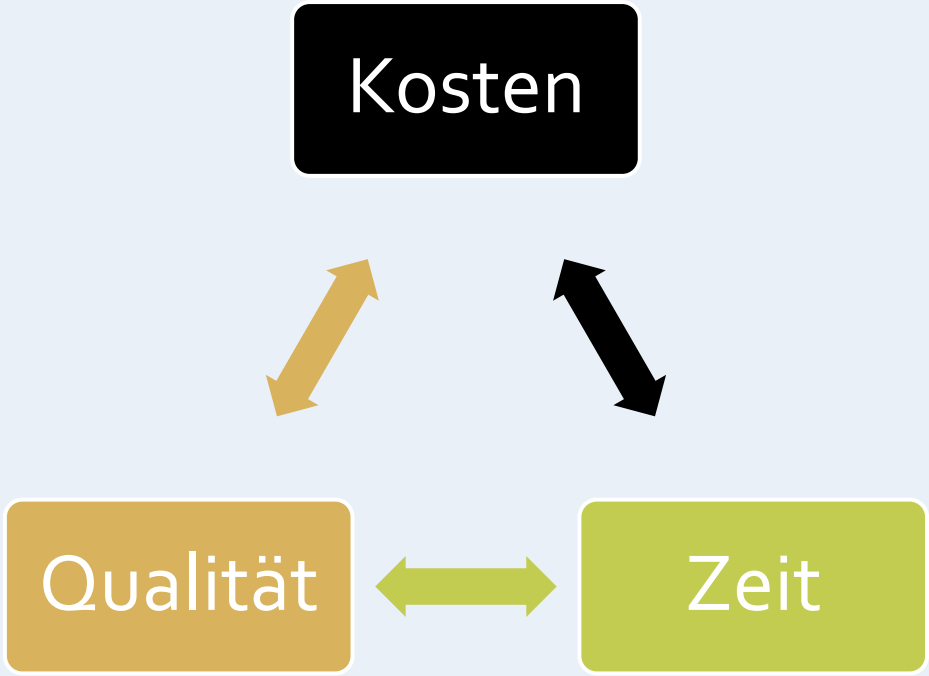
http://rrc.cvc.uab.es/files/MRecord_NE.png

- Ergebnisse der ICDAR2017 : Information Extraction in Historical Handwritten Records

http://rrc.cvc.uab.es/?ch=10&com=evaluation&task=1&f=2&e=1

Forschung vs. Praxis

- Magisches Dreieck:



Das Projekt OCR-D

- Projektpartner
  - Herzog August Bibliothek Wolfenbüttel
  - Berlin-Brandenburgische Akademie der Wissenschaften
  - Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
  - Steinbuch Centre for Computing (KIT)

- 2 Phasen:
  1. Analyse von Entwicklungsbedarfen, Aufbau der Koordinierungsstruktur und Konzeption der Projektphase
  2. Ausschreibung und konzeptionelle Begleitung der Modulprojekte

- Gefördert von der Deutschen Forschungsgemeinschaft

Projektziele

- **Konzeptionelle Vorbereitung der Transformation der VD-Drucke (16.-18. Jh.) und der Drucke des 19. Jh. in maschinenlesbare Form.**

- Technische Vorbereitung der massenhaften, maschinellen Vervolltextung von digitalisierten Drucken des 16.-19. Jh.

- Text- **und** Strukturerfassung zur Schaffung wissenschaftlich nutzbarer Forschungsdaten

- Antworten auf sich daraus ergebende konzeptionelle, informationswissenschaftliche und organisatorische Fragen entwickeln

- Lückenschluss zwischen Forschung und Praxis

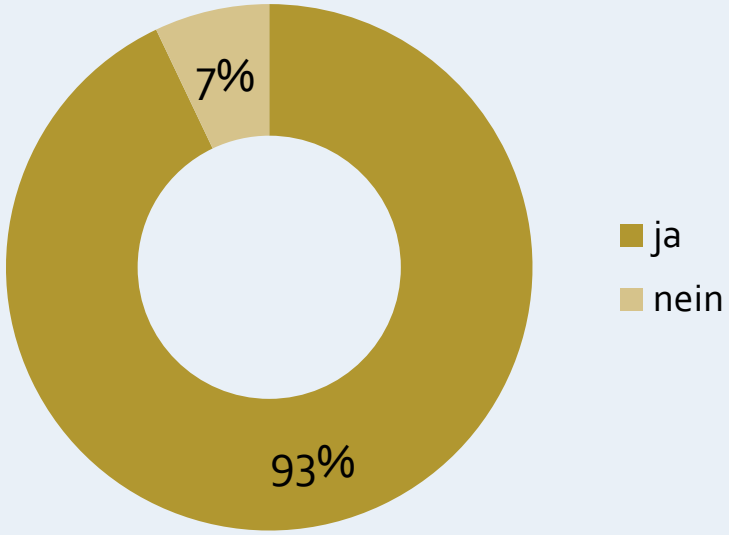
Qualität im OCR-D Prozess

- Drei Aspekte:
  - Qualitätskontrolle der Modulprojektergebnisse (mit Ground Truth)
  - Qualitätskontrolle ohne Ground-Truth als Forschungsprojekt
  - Formale Kontrolle der Softwarequalität ISO/IEC 25010:2011

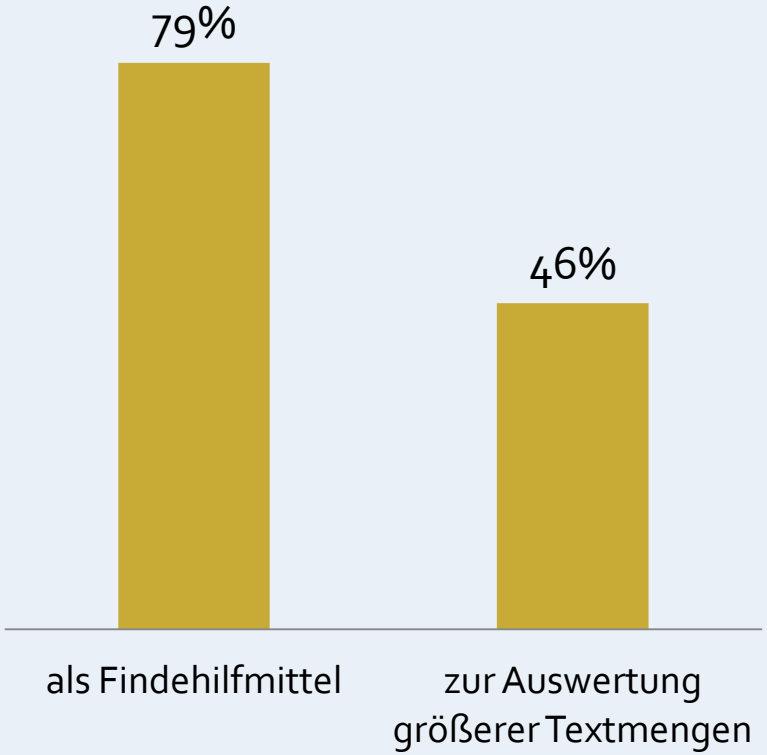
- Aktualisierung der Qualitätsbegriffs
  - Qualitätsbeschreibung nach Verwendungsmöglichkeiten, Gewichtung der Fehlerart

Qualitätsbegriff nach Nutzungsszenario

1. Beziehen Sie maschinell erkannte Texte (OCR-Texte) gedruckter Werke in Ihre Arbeiten/Forschung ein?

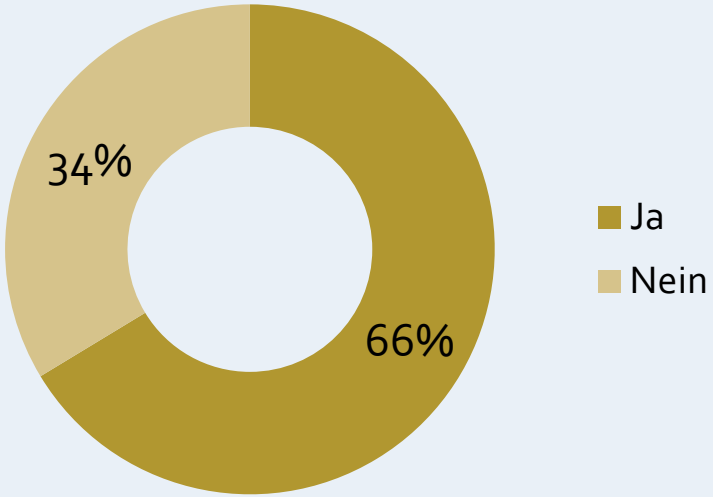


2. Wenn ja, benutzen Sie diese Texte



Qualitätsbegriff nach Nutzungsszenario

7. Sehen Sie auch eine (wissenschaftlich) nutzbringende Verwendung von OCR-Texten mit relativ hohem Fehlergrad?



Qualitätsbegriff nach Nutzungsszenario

EPIS J-OLJE

JOANNEM AKEPPLERUM

MATHEMATICUM C^SAREUM

SCRIPTi;
INSERTIS AD EASDEM

RESPONSIONIBUS KEPPLERIANIS,

QUOTQUOT HACTENUS REPERIRI

OPUS NOVUM, Qy6°RECONDITA KEPPLERIAN^E

DOCTRIN^E CAPITA DILUCIDE EXPLICANTUR,ET HISTORIA
LITERARIA IN UNIVERSUM MIRIHCE ILLUSTRATUR,

NUNC PRIMUM

CUM PR^FATIONE DE MERITIS GERMANORUM IN MATHESIN,

INTRODUCTIONE IN HISTORIAM LITERARIAM S./ECULORUM
XVI. ET XVII. ET JO. KEPPLERI VITA

Le texte affiché peut comporter un certain nombre d'erreurs. En effet, le mode texte de ce document a été généré de façon automatique par un programme de reconnaissance optique de caractères (OCR). Le taux de reconnaissance estimé pour ce document est de **86.43** %.
En savoir plus sur l'OCR

Wieviel sind 85% wert?

- Nutzungsszenario für 85%
  - OCR-Texte als Findehilfsmittel (bei Negativsuche)
    - Keyword search
    - Phrase search
    - Access via content structure
  - Überblick für die tiefergehende wiss. Arbeiten u. weitere Bearbeitungsschritte
- Definition der Qualität
  - Bestimmung der AQL
  - Definition der Messkriterien/ Metriken
- Bewertung in Abhängigkeit von Umfang und Vorarbeiten
  - Priorisierung des „Magischen Dreiecks“

FRAGEN

www.ocr-d.de

Elisa Herrmann, herrmann@hab.de