GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

GOEDOC – Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen

2016

Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich "Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930

Christoph Stollwerk

DARIAH-DE Working Papers Nr. 16

Stollwerk, Christoph: Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich "Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2016
(DARIAH-DE working papers 16)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2016-2
URN: http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2016-2-8

Dieser Beitrag erscheint unter der Lizenz Creative-Commons Attribution 4.0 (BY)

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

Abstract: Forschungsvorhaben sind oft an spezifische technische Anforderungen gekoppelt. Diese Studie liefert einen Überblick zu relevanten Anwendungsbereichen der OCR, dem Bereich der praxisbezogenen Massen-OCR für Bibliotheken sowie vergleichbaren Einrichtungen und dem anwendungsnahen Teil der OCR-Forschung. Für die systematische und vollständige Digitalisierung von Druckwerken wird in dieser Studie eine Übersicht zur Größe der Bestände bereitgestellt. Zudem wird überprüft, ob mit derzeit marktüblicher Software oder mit derzeit vorhandenen noch nicht marktüblichen Produkten, die Funktionalität anhand der Anforderungen älterer Drucke angepasst werden kann, ohne grundlegende Neuentwicklung dabei vorzusehen. Die Identifizierung von Forschungsfeldern im Bereich der angewandten Informatik, von denen die Ausdehnung der derzeit möglichen OCR-Verfahren auf bisher nicht behandelbare Korpora zu erwarten ist, wird ebenfalls geleistet. Darüber hinaus werden Forschungsfelder benannt, die eine Ausdehnung auf heute nicht routinemäßig mit OCR-Verfahren behandelbare Schriften (wie z.B. Handschriften) abdecken. Im Rahmen von vier Fallstudien sind für diese Zwecke Korpora definiert worden, um im Hinblick auf die Optimierungsmöglichkeiten von OCR-Verfahren die Ergebnisse in Fallstudien zu diskutieren, und um Empfehlungen und Ansätze zu konkretisieren.

Keywords: Optical Character Recognition, OCR, Digitalisierung, Alte Drucke

Optical Character Recognition, OCR, digitisation, historic prints

Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich „Alter Drucke“ zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930

Christoph Stollwerk

DARIAH-DE
Working Papers

Christoph Stollwerk: „Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich 'Alter Drucke' zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930“. *DARIAH-DE Working Papers* Nr. 16. Göttingen: DARIAH-DE, 2016. URN: urn:nbn:de:gbv:7-dariah-2016-2-8.

Dieser Beitrag erscheint unter der
Lizenz Creative-Commons Attribution 4.0 (CC-BY).

Die *DARIAH-DE Working Papers* werden von Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch herausgegeben.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Abstract

Forschungsvorhaben sind oft an spezifische technische Anforderungen gekoppelt. Diese Studie liefert einen Überblick zu relevanten Anwendungsbereichen der OCR, dem Bereich der praxisbezogenen Massen-OCR für Bibliotheken sowie vergleichbaren Einrichtungen und dem anwendungsnahen Teil der OCR-Forschung.

Für die systematische und vollständige Digitalisierung von Druckwerken wird in dieser Studie eine Übersicht zur Größe der Bestände bereitgestellt. Zudem wird überprüft, ob mit derzeit marktüblicher Software oder mit derzeit vorhandenen noch nicht marktüblichen Produkten, die Funktionalität anhand der Anforderungen älterer Drucke angepasst werden kann, ohne grundlegende Neuentwicklung dabei vorzusehen.

Die Identifizierung von Forschungsfeldern im Bereich der angewandten Informatik, von denen die Ausdehnung der derzeit möglichen OCR-Verfahren auf bisher nicht behandelbare Korpora zu erwarten ist, wird ebenfalls geleistet. Darüber hinaus werden Forschungsfelder benannt, die eine Ausdehnung auf heute nicht routinemäßig mit OCR-Verfahren behandelbare Schriften (wie z.B. Handschriften) abdecken.

Im Rahmen von vier Fallstudien sind für diese Zwecke Korpora definiert worden, um im Hinblick auf die Optimierungsmöglichkeiten von OCR-Verfahren die Ergebnisse in Fallstudien zu diskutieren, und um Empfehlungen und Ansätze zu konkretisieren.

Inhaltsverzeichnis

Abstract....2

1. Einleitung....5

2. Report zu bestehenden Forschungsfeldern....7

2.1 Forschungsstand....7

2.2 Zielsetzung....10

2.3 Fragestellungen....10

2.4 Hypothesen....12

2.5 Methoden....12

2.6 Testverfahren....14

2.7 Evaluation....17

2.8 Resultate....18

2.9 Diskussion....19

2.10 Praktische Relevanz....20

3. Die Fallstudien....22

3.1 Fallstudie 1 – Korpus 1....23
3.1.1 Resultate aus den Untersuchungen um Korpus 1....25
3.1.2 Korpus 1– Diskussion....26

3.2 Fallstudie 2 – Korpus 2....28
3.2.1 Überlegungen und Anmerkungen aus den Korpora 1 und 2....32

3.3 Fallstudie 3 – Korpus 3....34
3.3.1 Resultate und Ansätze aus den Erkenntnissen um das dritte Korpus....37

3.4 Fallstudie 4 – Resultate in Bezug auf Qualitätsunterschiede....40
3.4.1 Fallstudie 4 – Vergleiche auf Wortebene....42
3.4.2 Fallstudie 4 – Vergleiche auf Zeilenebene....45
3.4.3 Fallstudie 4 – Vergleiche auf Seitenebene....48
3.4.4 Fallstudie 4 – Diskussion....57
3.4.5 Ergebnisse aus Fallstudie 4....60

4. Abkürzungsverzeichnis....83

5. Abbildungsverzeichnis....84

6. Literaturverzeichnis....85

7. Webreferenzen....91

8. Anhang....92

1. Einleitung

Das Projekt mit dem Arbeitstitel „Machbarkeitsstudie zur vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930" beinhaltet die Erstellung zweier Berichte: Im ersten Teil wird eine Aufwandsabschätzung erarbeitet, in der marktübliche und nicht-marktübliche Software an größeren Datenbeständen auf ihre Funktionalität hin überprüft wird. Dabei wird besonders der Aspekt der Erweiterbarkeit von offener Software überprüft und es werden Aussagen über den Aufwand von Weiterentwicklungen getroffen, die ohne grundlegende Neuentwicklungen auskommen. Der zweite Teil umfasst die perspektivische Eröffnung von informationstechnischen Forschungsrichtungen, in denen sich Verfahren zur Optical-Character-Recognition (OCR) als bisher nicht ausreichend präzise erwiesen haben.

Noch immer sind OCR-Anwendungen nicht dazu in der Lage strukturierte Daten aus eingescannten Dokumenten korrekt abzuleiten. Diese und ähnliche Schwierigkeiten werden einen Teil der Untersuchung zur Identifikation von Forschungsbereichen, die in bisher nicht-behandelbaren Korpora zu erwarten sind, darstellen (Holley 2009). Museen, Bibliotheken, Archive und Kirchen stellen historische Dokumente in sehr großem Umfang bereit. Allerdings ist es meist nicht oder nur sehr schwierig möglich Zugriff auf diese Dokumente zu erhalten. Häufig sind die Dokumente aufgrund von Konservierung oder aus sicherheitsrelevanten Gründen nicht frei zugänglich. Digitalisierung ist in diesen Fällen oftmals die einzige Lösung diese historischen Bestände einem breiten Personenkreis zu eröffnen.

In der Vergangenheit haben sich eine Vielzahl von Projekten der Digitalisierung und Archivierung historischer Dokumente gewidmet.[1] Um die digitalisierten Daten für möglichst viele Personen effizient nutzbar zu machen, reicht es nicht aus sich auf langfristige Ablage oder Repräsentation des Digitalisierten zu fokussieren. Die in den letzten drei Jahrzehnten stattgefundene Digitalisierung steht nun vor Datenmengen, die ohne neue und dazu passende Werkzeuge nicht mehr auskommt.

Neben der reinen Bewältigung der Datenmenge, die bei OCR-Projekten schnell auf über 30 MB pro Seite steigen kann (Coyle 2006), hat man begonnen Problematiken im Bereich der Speicherung[2], Langzeitarchivierung und Erfassung von Dokumenten in Ansätzen zu lösen. In dieser Studie wird neben der Entwicklung von drei unterschiedlich gelagerten Korpora, auch auf Strukturen in Bezug auf Abläufe und Erweiterbarkeit von OCR-Software eingegangen.

Um alle Zielsetzungen des Vorhabens umzusetzen, wird eine formative Evaluation durchgeführt. Dabei werden unterschiedliche Schwachstellen (Kapitel 3.3) der Korpora in Verbindung mit den zu untersuchenden Software-Lösungen gesetzt, um zusätzliche Möglichkeiten der Optimierung aufzuzeigen.

Die Abschätzung des Aufwandes für eine Gesamtdigitalisierung gemeinfreier deutscher Drucke, oder richtiger: des Aufwandes für die systematische Anwendung von OCR Verfahren auf diese Drucke, leidet darunter, dass die Ergebnisse der Anwendung ein und desselben OCR Verfahrens auf zunächst ähnlich wirkende Texte, sehr divergierende Ergebnisse erbringen können. In der

1 http://monasterium.net; http://numeric.ws; http://nestor.sub.uni-goettingen.de; http://jstor.org
2 http://www.goethe.de/wis/bib/dib/de3616481.htm

vorliegenden Studie wurde daher ein mit überschaubarem Aufwand handhabbares Verfahren zur Etablierung von belastbaren Zahlen für einzelne Textkategorien aufgezeigt: Das Verfahren und deren Ergebnisse in Form der Etablierung eines Korpus für eine beschränkte, aber klar definierte Groundtruth geben Aufschluss über das Spektrum der Problemquellen. Der Begriff Groundtruth bezieht sich hier auf die Klassifikation einer Trainingsmenge oder hier eines Korpus im Kontext von Verfahren zur Überprüfung der entwickelten OCR-Verfahren und deren Relevanz für die Bewertung von OCR Software, sowie die Bewertung derzeit gängiger Systeme auf der Basis dieser Groundtruth, bilden das Ergebnis der Arbeit. Eine Hochrechnung, in Form der Untersuchung, wie groß der Anteil der hier untersuchten „Typen" von Drucken im Korpus insgesamt ist, konnte in dieser Studie nicht mehr umgesetzt werden. Unter einem „Typus" wird hier die Kombination spezifischer, für die Qualität von OCR Ergebnissen relevanter Features, wie Mischung von Fonts, Störungen durch mangelnde Druckqualität verstanden.

2. Report zu bestehenden Forschungsfeldern

Zur Identifikation von Forschungsfeldern, die im Bereich der angewandten Informatik angesiedelt sind, werden Verfahren zu aktuell verfügbarer OCR-Software in Bezug auf bisher nicht behandelbare Korpora untersucht und im Folgenden dargestellt. Korpora im historischen Kontext sind zahlreichen Problematiken unterworfen. Es handelt sich bei den hier angewandten OCR-Vorgehensweisen und Korpora um Sonderfälle, die einzeln betrachtet nicht repräsentativ für die OCR-Forschung gewertet werden können. Gleichwohl aber viele Punkte summieren, die als Engpässe in heutiger OCR-Software zu erwarten sind. Über die in Kapitel 4.3 erhobenen Features können Arbeitsbereiche den Forschungsdisziplinen explizit zugeordnet werden.

Um Verbesserungen von OCR-Verfahren über eine technische Perspektive einzuleiten, lassen sich aus der Informatik drei zentrale Felder ableiten: Die digitale Bildverarbeitung, Computer Vision und die Softwareentwicklung (z.B. Entwicklung von Algorithmen). Parallel zur Entstehung der "Digital Humanities" ist das Aufkommen digitalen Materials massiv gestiegen, so dass eine einzelne Technik nicht mehr abdecken kann, was an komplexen Anforderungen währenddessen entstanden ist. Es ist demnach wichtig, die einzelnen Techniken, die hier exemplarisch für unterschiedliche Arbeitsabläufe und Vorgehensweisen stehen, technologisch aufeinander abzustimmen, sodass auch komplexe Aufgaben, die durch historische Quellen vorgegeben sind, lösbar werden. Denn auch über die OCR haben sich im aktuellen Forschungsumfeld Unterdisziplinen gebildet, mit denen sich durch die reine Anwendung von Informationstechnik in keiner Form nutzbare Resultate in z.B. lateinischer Schrift ableiten lassen. Im Fall der Erkennung von Handschriften ist man davon abgewichen die Handschrift direkt in maschinenlesbaren Text umzuwandeln, sondern versucht ein Digitalisat in Hinblick auf seine Machbarkeit und bekannten Parametern hin zu klassifizieren.

Ein weiterer Aspekt ist die Untersuchung von Größe und Ausrichtung eines OCR-Projekts. Neben Projekten bei denen die massenhafte Erstellung und Verarbeitung (z.B. Google@BSB oder dem IMPACT-Projekt) von Digitalisaten im Vordergrund steht, sind kleine und mittlere Vorhaben auch zu berücksichtigen. Denn wie Blanke et. al anführen, sind die Parameter und Anforderungen eines Projekts ein wichtiger Erfolgsfaktor für den Ausgang eines OCR-Projekts. (Blanke, Bryant, and Hedges 2012a)

2.1 Forschungsstand

Ziel des vorliegenden Kapitels ist die Benennung von Forschungsfeldern der angewandten Informatik, bei der die Ausweitung auf bisher nicht-behandelbare Korpora zu erwarten ist und in der Gesamtperspektive der Prozessierung von erheblichem Zuwachs im Punkt Optimierung und Entscheidung zu erwarten ist. Die Verzeichnisse um den Gesamtbestand deutscher Drucke werden abgedeckt durch VD16[3] VD17[4] und VD18[5], sowie die Sammlung deutscher Inkunabeln

3 http://www.bsb-muenchen.de/1681.0.html
4 http://www.vd17.de
5 http://vd18.de

und Wiegendrucke[6]. Im Rahmen dieser groß angelegten Projekte sind bereits neben den Praxisregeln der Deutschen Forschungsgemeinschaft (DFG) weitere Standards entstanden, die im deutschen Raum für Konsens in Bezug auf die technischen Rahmenbedingungen der Digitalisierung, der Referenzierbarkeit und des Transfers über technische Barrieren hinweg gesorgt haben. Das Zentrale Verzeichnis Digitalisierter Drucke (ZVDD) bietet mit Blick auf die Aspekte der Organisation von Verzeichnissen, die aktuell einzige Quelle aller erschlossenen Drucke. Das heißt aber wiederum nicht, dass hier auch die Volltexte zu finden sind. Eine Infrastruktur, die dazu dient Quellen und Materialien zur Verfügung zu stellen, besteht demnach. Auch die dezentrale Architektur des Open Archives Initiative (OAI)-gestützten Ansatzes des ZVDD hat den Vorteil, dass hiermit eine übergeordnete Instanz zur Organisation aller deutschen Digitalisierungsprojekte zur Verfügung steht und Metadaten über Autoren, Titelstichwörter und Erscheinungsjahre über eine Internetseite für den Benutzer zur Verfügung stehen (Stäcker 2006).

Durch eine Homogenisierung der Metadaten entsteht ein entscheidender Vorteil: Wenn alle Zulieferer die Nutzung, Weiterentwicklung und Integration von Metadaten ermöglichen, werden dadurch nicht nur intrinsische Daten über ein Digitalisat zugänglich, sondern auch die Erweiterung und Ableitung neuer Informationen erst denkbar. Je nach gewählter Domäne, können in der Datenmodellierung eines Objekts auch extrinsische Daten zur Beschreibung von Artefakten genutzt werden. Dies dient zwar nicht unmittelbar eines Mehrwertes der geisteswissenschaftlichen Forschungsperspektive, führt allerdings dazu, dass die Zielsetzung der Volltext-Digitalisierung enorm transparenter dokumentiert wird und auf Dauer nachvollzogen werden kann.

Wissenschaftler sind nicht nur an Metadaten interessiert, die auf inhaltlicher Ebene einen Mehrwert darstellen, sondern werden zukünftig zunehmend an materialspezifischen Gegebenheiten (Weymann and Insitut 2010) interessiert sein. Daher sind Wertschöpfungsketten (Dr. Siegfried Dais, Prof. Henning Kagerman), die aufgrund bisher sehr fehlerhafter Ergebnisse nötig geworden sind, außerordentlich relevant. Die horizontale als auch die vertikale Integration aller Ergebnisse von OCR-Studien zu gewährleisten, versprechen viele Vorteile. Beispielweise wird über die Erweiterung von Metadaten erst ermöglicht Artefakte, die aus dem ersten Prozessierungsschritt im letzten wiederzuerkennen (Vorbereitung zu ‚Provenance').

Es ist unstrittig, dass die manuelle Korrektur oder Verfahren wie Double-Keying hohe finanzielle Hürden bergen, so ist die Inanspruchnahme der öffentlich freien Partizipation, um an valide Daten zu gelangen, ein vielversprechender Ansatz. Auf dieser Basis könnten sukzessiv korrekte Groundtruth-Korpora entstehen und man müsste nicht zwingend über einen Ansatz gehen, wie er im Deutschen-Textarchiv[7] angewandt wird. Über Crowdsourcing-Verfahren (Moyle, Tonra, and Wallace 2011), die ähnlich trivial gehalten sind wie Googles reCaptcha-Projekt[8] werden gerade zentrale Aspekte, die durch Trivialität und hoch-frequentierte Nutzer-System-Interaktion rasch wachsenden Gewinn von Informationen versprechen, in Aussicht gestellt.

6 http://www.zvdd.de
7 http://www.deutschestextarchiv.de
8 http://googleblog.blogspot.de/2009/09/teaching-computers-to-read-google.html

Eine Einordnung in Nutzergruppen und die damit verbundene Sorge um geringe Partizipation, wie es beim Projekt der Deutschen-Digitalen-Bibliothek stattgefunden hat (Virtual Identity AG – Deutsche Nationalbibliothek), würde hierbei nicht logisch sein und demnach auch nicht auftreten. Jeder Nutzer könnte einfach bei der Inanspruchnahme eines Services (zum Beispiel Login, Registrierung, Ausleihe, etc.) eine Kleinstmenge an Partizipation leisten.

Zwar entsteht auf diesem Weg nicht zwangsläufig eine große Anzahl von Groundtruth-Texten, aber eine stete Bereicherung, fernab von Projektrahmen und Budgetgrenzen, die unterschiedliche Vorhaben wesentlich einfacher realisierbar machen würden als bisher. Denn spricht man von textuellen Informationen, die zu Forschungszwecken genutzt werden sollen, ist unumgänglich, dass die Daten in einer sehr hohen Qualität vorliegen müssen (Holley 2009). Anders lassen sich quantitative Textanalysen, die zum Verständnis und der Untersuchung von Literaturgeschichte notwendig sind, nicht nutzen (Jannidis und Schöch 2014).

Neben der Erstellung weiterer Grundlagen, die zum einen das Ziel haben den Informationsaustausch über digitalisierte Werke zu ermöglichen und zum anderen sämtliches vorhandenes Material für Wissenschaftler, die nicht-textuelle, voll-automatisierte oder sammlungsspezifische Zwecke verfolgen, zu eröffnen, sollte auch die Hinwendung zu stärker technisch-orientierten Forschungsrichtungen berücksichtigt werden. Als zentrales Problem haben sich die so genannten „Multi"- oder „Mix"-Texte hervorgetan, dabei geht es vorrangig um zwei Phänomene:

1. Die Sprache wechselt auf Zeichen-, Wort- oder Absatzebene und 2. es befinden sich mehrere Zeichensätze im Dokument. Auf beide Schwierigkeiten wird in späteren Teilen dieses Berichts präziser eingegangen. Es sei darauf hingewiesen, dass die „Verwirrung" von Stil und Sprache als gelöst betrachtet werden kann (Biemann and Teresniak 2005). Denn in den hierzu zu untersuchenden Zeiträumen früher Drucke, befindet sich ein durchaus häufig wechselnder Sprachgebrauch aus Frühneuhochdeutschem, Neuhochdeutsch, Latein und Griechisch, aber auch Italienisch oder Französisch treten auf. Derartige Texte können entweder durch spezifische angepasste Modelle in der genutzten OCR-Software behandelt werden oder durch eigens dafür initiierte Neuentwicklungen, wie es etwa durch Satz-basierte Co-Occurrence-Graphen möglich wäre.

Neben der Ausdehnung von Forschungsrichtungen im Hinblick auf die optimierte Zugänglichkeit und Partizipation, bieten sich auch generische Möglichkeiten an, um die Qualität der Dokumente zu steigern. Neben Dokument-basierter Optimierungen, sind auch von Farb- und Segmentierungsverbesserungen, die profunde Teile der Potenziale ausmachen, zukünftig Optimierungen zu erwarten.

Da nicht alle Ansätze hier im Detail besprochen werden können, wird im Folgenden kategorisiert. Die aufgeführten Richtungen deuten wesentliche Verbesserungspotenziale im Hinblick auf die Qualität und Anwendbarkeit auf unbehandelbare Korpora an:

- Binarisierung in Dokumenten mit Störungen (Blanke, Bryant, and Speck 2012; Pratikakis, Gatos, and Ntirogiannis 2013; Konya et al. 2009; Sezgin 2004; Still Image Working Group 2010)
- Entzerrung von Zeichen- und Wörter-Orientierungen (Skew/Slant Correction) (Ávila and Lins 2005; Breuel, Beusekom, and Shafait 2009; Sadri and Cheriet 2009)

- Segmentierung und Seitenanalyse (Shafait, Keysers, and Breuel 2008)
    - Verbessertes Dokumentenverständnis von Regionen (Lidke, Thurau, and Bauckhage 2010)
    - Dokumentenverständnis von Bildern und Fußnoten (T. Breuel; Breuel 2003; T. Breuel 2003)
- Logische Layout-Analyse zur automatisierten Identifikation von Artefakten wie Titelblättern oder anderen strukturintensiven Objekten (Déjean and Meunier 2009)
- Qualitätssicherung von Daten mit Groundtruth-Ambition (Liu et al. 2012)
- Vorgehensweisen zur semantischen Erschließung von Dokumenten (Ratinov et al. 2011)
- Evaluation von textueller Informationen auf Basis von Ansätzen zur massenhaften Verarbeitung (Konya, Eickeler, and Seibert 2010)

2.2 Zielsetzung

Der vorliegende Bericht weist eine Vielzahl von Referenzen zu Forschungsliteratur auf und kann daher auch als Sammlung verstanden werden, in denen state-of-the-art Lösungen entweder genannt oder eruiert werden. Aus diesem Grund wird in der Untersuchung von den vier meist genutzten OCR-Implementationen aus dem proprietären sowie offen zugänglichen Bereichen als Grundlage genutzt, um daran drei parallel genutzte Korpora zu testen und an den OCR-Implementationen zu evaluieren.

Die Evaluation von OCR-Vorgehen hat sich als zentraler Aspekt der Schnittstelle zwischen den aus OCR-Implementationen resultierenden Texten, der Leistungsfähigkeit der OCR-Implementationen und bestehenden Ausgangsmaterialien herausgestellt. Daher ist erklärtes Ziel dieser Arbeit Workflows und Methoden zur Verfügung zu stellen, die dazu dienen Texte in einer Differenzanalyse über Distanzwerte vergleichbar zu machen. Die Differenzanalyse dient als evaluierende Instanz zur Bestimmung der Leistungsfähigkeit der OCR-Implementationen und wird zum Ende des Berichtes genutzt, um die entwickelten Korpora an den relativen Distanzmaßen zu vergleichen.

Um die OCR-Implementationen validen Testdaten gegenüberstellen zu können, werden Korpora entwickelt, die aus Blickrichtung der Datenmodellierung variierende Ziele verfolgen. Dabei sollen nicht nur Informationen in die Korpora fließen, die stellvertretend so groß sind, dass sie für einen "Job" einer massenhaften Verarbeitung exemplarisch stehen, es werden auch Daten genutzt, die das sehr heterogene Feld der hier vorliegenden Zeiträume abdecken und dabei bekannte Probleme der OCR-Entwicklung nicht vernachlässigen

2.3 Fragestellungen

Die Komplexität innerhalb von OCR-Projekten entsteht fast immer aus den miteinander zusammenwirkenden Arbeitsebenen, die in Form von Dokument-nahen, topographischen, strukturellen, logischen und analytischen oder funktionalen Abschnitten auftreten. Dies kann zusätzlich mit vielfältigen Überschneidungen unterschiedlicher Ebenen einhergehen. Eine Ebene kann aus Informationen zur Aufnahmesituation bestehen, während eine andere die

Binarisierung oder das gesamte Preprocessing dokumentiert. Mit dem Begriff Preprocessing ist hier eine Sammlung von Methoden gemeint, der man ein Digitalisat zuführt, bevor es mit der Zeichenerkennung prozessiert wird. Zum Preprocessing gehört z.B. die Binarisierung, Segmentierung oder Entzerrung eines Digitalisates. Wieder eine andere Ebene erfasst das Validieren von Ergebnissen. Zuletzt könnten topographische Informationen von Schreibräumen dargestellt und festgehalten werden.

In den meisten kollaborativen Umgebungen entsteht ein weiteres Problem: Die Auszeichnung der Werke stellt hier nur eine mögliche Interpretation dar. Eine neue Fragestellung erfordert wiederum das Einfügen einer erweiterten Beschreibungsebene. Ein Vorteil von Digitalen-Editionen besteht aber gerade in der Möglichkeit eine Edition als fortlaufenden Prozess zu betrachten und die transkribierten Ergebnisse immer wieder mit neuen Informationen anzureichern. Im Laufe späterer Analysephasen treten oft weitere Ebenen auf und führen zu immer komplexer ausgestalteten Meta-Informationen, die unterschiedlich kodiert sind und in unterschiedlicher Qualität vorliegen. Die Menge der theoretisch vorhandenen Ebenen und die Limitierung der zugrunde liegenden Strukturen sind demnach zu klären. Manche Ebenen konkurrieren oder korrelieren und die Modellierung in singulären Strukturen wie Meta-Dokumentationen fällt schnell als nicht ausreichend auf. Moderne, kollaborative Digitalisierungsprojekte mit Methodendokumentation werfen immer wieder die Frage auf: Welche Art von organisatorischer Dokumentation der Komplexität des OCR-Projektes wird erwartet?

Eine weitere Frage, die aus der Beobachtung bereits abgeschlossener und laufender Projekte hervorgeht, ist die der Nachnutzung von Resultaten[9]. Denn nicht nur Abfragen und Suchalgorithmen lohnen der wissenschaftlichen Überprüfung, auch Methoden wie Algorithmen zur Vermessung, Textmining und Rankingfunktionen, die die Identifizierung mittelalterlicher Entitäten erleichtern, sind in projektübergreifenden Rahmen mit Werkzeugen adäquat miteinander zu verbinden. Die Identifikation von Entitäten, seien es topologische, philologische oder rein technischer Natur geschuldete mittelalterliche Quellen, stellt Wissenschaftler in der Regel vor erhebliche Schwierigkeiten. Dazu zählen beispielsweise regional bedingte Varianten einzelnder Drucke, Übersetzungs- oder Transkriptionsfehler, genutzte aber unbekannte Sprachformen oder der Gebrauch von Namen. Nur mit intensiver Recherche bei gleichzeitig geringen Ressourcen finden derartige Identifizierungen innerhalb der meisten Erschließungsprojekte statt. Es gilt also zu prüfen, inwiefern die Kombination der Verfahren von einer Verbindung profitieren. Allerdings fehlt es an Ansätzen den Nutzer dazu zu bewegen, Werkzeuge und Quellen in einem dem Aufwand angemessenem Maß anzuwenden. Bestünde die Möglichkeit aus einer sehr großen Menge an historischen Daten eine möglichst begrenzte Menge potenziell identischer Entitäten zu extrahieren, könnte dem damit Abhilfe geschaffen werden. Doch auch hier entsteht die Fragestellung nach der Konstellation von Varianten. So sind zum Beispiel Algorithmen aus dem Bereich „Soft-Pattern-Matching" nicht zwangsläufig dazu geeignet, um phonetische Algorithmen umzusetzen. Über den Vergleich von Zeichenketten identifizierten Entitäten, bieten sich darüber hinaus auch Methoden aus dem Bereich des Text-

9 http://impact-project.eu; http://succeed-project.eu; http://emop.tamu.edu; http://deutsches-textarchiv.de

Mining an. Hier könnten statistische Methoden, wie zum Beispiel innerhalb von "word-space"-Modellen, zentral werden.

2.4 Hypothesen

Es existiert kein Korpus, das zusammengestellt wurde, um Schwierigkeiten bei der OCR zu analysieren. Das wird in dieser Studie nachgeholt und stellt potentiell signifikante Verbesserungen von OCR-Qualitäten und deren automatisierter Prozessierung in Aussicht.

Um die einzelnen Software-Implementationen zu evaluieren oder problemorientierte Korpora zu entwickeln, sind weitere grundlegende Basisarbeiten notwendig. Auch bei den OCR-Projekten, bei denen stärker kollaborativ gearbeitet wird, sind stärkere Maßnahmen zur Interoperabilität eine Ergänzung, die über semantische Anreicherung der Daten zu Optimierungen in der Qualität der Resultate führt. Aus diesem Grund wird in dieser Studie überprüft werden, inwiefern eine erweiterte Ebene von Metadaten stichhaltige Verbesserungen ohne grundlegende Neuentwicklungen fördert.

Nicht nur Position und Layoutinformationen (wie sie bspw. unter ALTO 1.0 genutzt werden[10]) von Entitäten können der datentechnischen Optimierung dienen, auch technische Hintergründe zur Prozessierung oder die paragraphen-orientierte Angabe einer Sprache machen Sinn (worauf auch die letzten Änderungen in ALTO 2.1 hindeuten, die am 20.02.2014 veröffentlicht wurden). Sobald versucht wird, auf Probleme einzugehen, die in der Prozessierung ihren Ursprung haben, zählen auch Informationen zu Ausgangsqualitäten der Ressourcen oder angewandte Algorithmen (zum Beispiel Binarisierungsalgorithmen oder Korrektur-Algorithmen) dazu.

Die meisten bestehenden OCR-Projekte halten sich an die Vorgaben der DFG (z.B. 300 DPI-Scans in Digitalisierungsprojekten zu verwenden – DFG 2013) und skalieren entweder Digitalisate bei zu geringer Auflösung hoch oder weisen zu gering klassifizierte Digitalisate sogar ab. Auch in Bereichen wie der Langzeitarchivierung, Datenhaltung, digitalen Bildverarbeitung, Computer Vision usw. würden damit Anknüpfungspunkte entstehen. Aus diesem Grund wird auch die Untersuchung moderner OCR-Software mit Hinblick auf die Leistungsfähigkeit bei geringeren Qualitäten des Ausgangsmaterials angestrebt. Denn nicht nur die allgemeine Bereitstellung zu Forschungszwecken stellt Rechenzentren vor kostenintensive Herausforderungen, auch den oben angesprochenen Forschungsrichtungen wäre in Bezug auf OCR Optimierungen zugearbeitet.

2.5 Methoden

Bestehende OCR-Lösungen orientieren sich noch immer vorwiegend an Buchstaben. Daher wird hier, in den getesteten OCR-Implementationen, auch auf Basis einzelner Buchstaben evaluiert. Im Bereich von Frakturschrift wird mit Umbrüchen in Formen (Buchstaben) gearbeitet (beispielsweise in ABBYY). Im Kontext von Handwriting Recognition ist wissenschaftlicher Konsens nicht mehr einzelne Bestandteile in den Fokus einer algorithmischen Untersuchung zu stellen (Sayre 1973). Es wird von Extraktionsmethoden zur Objekterkennung auf Basis von grafischen Schablonen gesprochen, die analytisch überprüft werden können, aber denen noch keine Semantik zuteil wird.

10 http://www.loc.gov/standards/alto

In verschiedenen Schritten werden die Bilddaten prozessiert und die Ergebnisse dieser Analyse werden als Merkmale in einem Vektor gehalten. Alternativ oder ergänzend dazu können mathematisch-codierte Beschreibungen von bereits segmentierten Entitäten stattdessen zusätzlich genutzt werden. Sowohl im Bereich der mathematischen Beschreibung, genauso wie in der Lösung mit Schablonen und der Segmentierung, stehen noch zu erwartende Implementationen aus, die von den sehr hohen Anforderungen im Bereich wissenschaftlicher Nutzung, gebraucht werden.

OCR-Verfahren lassen sich durch eine vermeintlich klare Struktur in Prozessschritte aufteilen: Die Vorverarbeitung – oder auch Preprocessing – ist ein zentraler Teil. Vor allem die Phase der Segmentierung eines Digitalisates, bei der auch das Entfernen von Störungen durchgeführt oder qualitative Tests gemacht werden, bietet Aufschluss über notwendige zusätzliche vorverarbeitende Schritte. Beim Preprocsessing wird alles vorbereitet, um in der nächsten Phase, dem Processing (hier: Feature-Extraction), optimale Voraussetzungen vorzufinden. Allerdings werden in diesem Schritt auch eine Menge der Daten aus dem Bild abgeleitet, die zum Verstehen und Regenerieren des Bildes notwendig sein könnten. Nachdem diese beiden Phasen abgeschlossen sind, folgt das Postprocessing (Nachverarbeiten), wobei meist domänenspezifische Informationen erhoben werden.

Bei der eigentlichen Prozessierung wird – wie bereits angedeutet – bei Repräsentation von Objekten zwischen Mustervergleichen (z.B. mit bestehenden Wissen über wiederkehrende Formen/Schablonen) und mathematisch kodierten Beschreibungen der Objekte (z.B. Vektoren, Histogramm-Verfahren), die Ausschnitte eines Digitalisates extrahieren und Inspektoren beschreiben, unterschieden. Man spricht in diesem Kontext daher auch fast immer von Erkennung oder Recognition. Diese Phase wird in manchen Lösungen nicht scharf getrennt von Verfahren zum Postprocsessing und steht in einem ständigen Wechselverhältnis. Zum einen werden Ergebnisse aus der Erkennung in Dokumenten mit Wörterbüchern und Lexika verglichen und auf der anderen Seite werden die erstellten Daten zu weiteren Zwecken, wie zum Beispiel Visualisierung, weiterverarbeitet. Dazu zählt auch bereits oft das schlichte Löschen von leeren Zeilen zur Reinigung oder Aufbereitung der Ergebnisse.

Ein klassischer Fall, in dem die Korrektur der Daten explizit zu Erkennungsmethode gehört, ist der Vorschlag von El Abed und Märgner (El Abed and Märgner 2007). Dabei wird vorgeschlagen, dass die automatisierte OCR-Korrektur bewältigt werden kann, indem man ein Digitalisat mehreren OCR-Erkennungsalgorithmen zuführt und die Ergebnisse über unterschiedliche Kombinationen (Mehrheitsprinzip, Ranglistenprinzip oder neuronales Netz (El Abed and Märgner 2007) prüft und darüber optimierte Ergebnisse erzeugt (Volk, Marek, and Sennrich 2010).

Eine weitere Möglichkeit ist mehrere Konfigurationen einer OCR-Engine zu kombinieren, anstatt mehrere OCR-Engines zu kombinieren. Diese wird auch „Multi-Stage Combination" genannt (Lin 2001). Dabei können – wie im Folgenden Bild illustriert – Abläufe generiert werden, die in ihrer Struktur parallel-kaskadierend oder umgekehrt kaskadierend-parallel verlaufen:

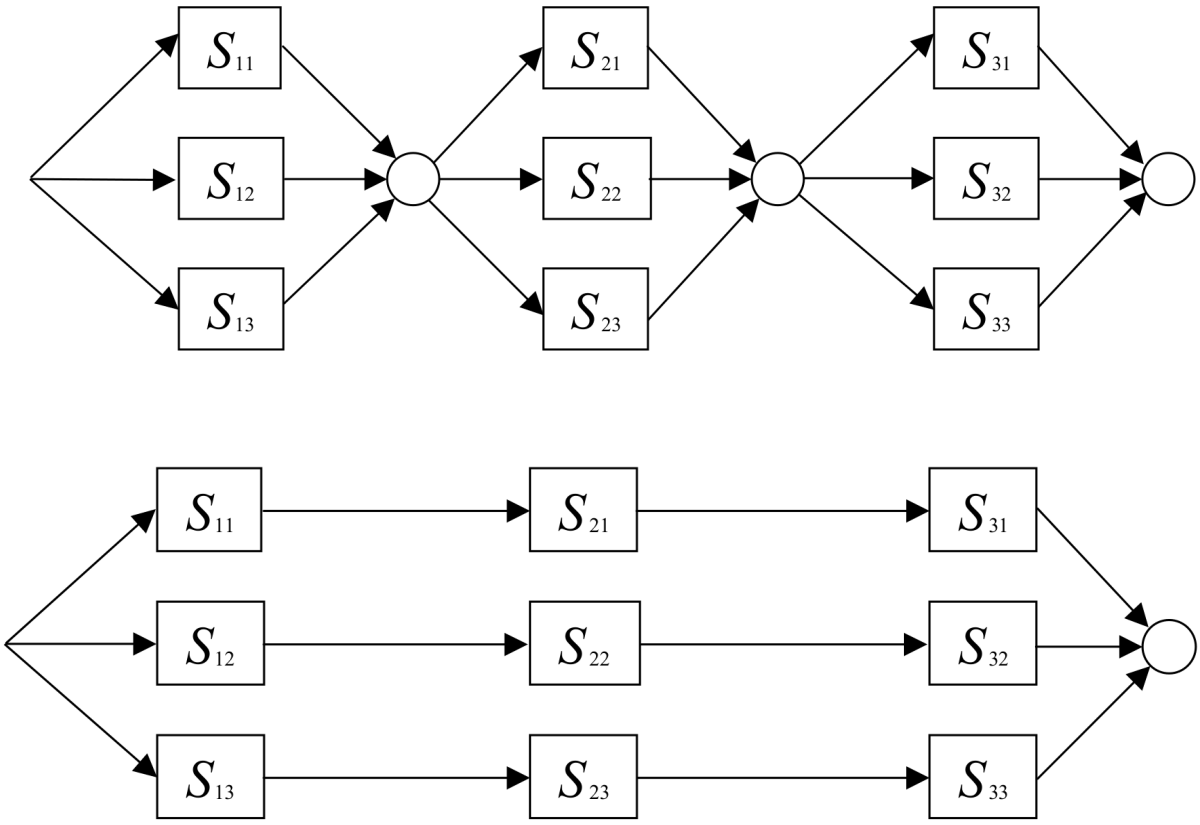


Abbildung 1: Mögliche Konfigurationen eines Multi-Stage Kombinations-Systems.

Abläufe, Prozessschritte oder Workflows zu kombinieren wird damit in eine Ordnung gefügt und hilft zweifelsfrei bei der Organisation der vielschichtigen Prozess-Architekturen, wie sie im Ocropodium-Projekt bereits angewendet wurden (Blanke, Bryant, and Hedges 2012b). Trotzdem bleiben in bestehenden OCR-Lösungen die Evaluation, das Fehler-Management und die Korrekturmöglichkeiten unterdeterminiert. Obwohl gerade das Fehlermanagement in größeren Projekten einen eklatanten Faktor, in Bezug auf den Erfolg im Einsatz von heterogenen Systemen, darstellt (Friske 2013; Pfleger 2014).

2.6 Testverfahren

Der Softwarepool, der zur Untersuchung von technischen Lösungen herangezogen wurde, ist aus einer Recherche abgeleitet worden. Die Kandidaten mit dem größten Potenzial auf Machbarkeit, sind sowohl in kommerzielle als auch frei verfügbare Lösungen zu unterteilen:

- Recognition Server 3.5 von ABBYY (vormals Finereader XIX)[11]
- BIT-ALPHA 6.5.8 von B.I.T. Bureau Ingéneur Tomasi[12]
- Tesseract 3.03 von HP & Google (mit Leptonica 1.6.9)[13]
- Ocropus 0.7 vom DFKI & Google[14]

11 http://frakturschrift.de/de:start
12 http://bit.dyndns.biz
13 http://code.google.com/p/tesseract-ocr
14 http://code.google.com/p/ocropus

Die beiden zuletzt angeführten Software-Projekte sind Quelltext-offen und können selbstständig erweitert werden. Die ersten beiden Produkte sind nur sehr begrenzt anpassbar und es bedarf an vielen Punkten noch der Rücksprache mit den Herstellern.

Ziel der hier angestrebten Testverfahren ist es Verbesserungs-Maßnahmen zu entwickeln. Diese Ziele stehen in direktem Zusammenhang mit den Ergebnissen, die bei der Durchführung der Evaluation entstehen. Die Potenziale einer Software in Bezug auf Brauchbarkeit bei Volltext-Digitalisierung des genannten Korpus zu untersuchen und dabei die Möglichkeiten zur Überprüfung der Ergebnisse plausibel nachvollziehbar darzulegen, ist demnach Ziel eines validen Workflows.

Tesseract ist eine Software, die bereits sehr lange entwickelt wird (1995 war sie eine Sieger-Implementation zur Zeichenerkennung). Sie wurde ursprünglich als Zeichenerkennungssystem konzipiert. Die Hinzunahme von Leptonica hat es ermöglicht die Software weiteren Nutzungsmöglichkeiten, als der reinen Zeichenerkennung, zuzuführen. Durch den Zugewinn von Leptoinca wird Seitensegmentierung als integraler Bestandteil der OCR-Lösung offeriert. Tesseract bestand hauptsächlich in seiner Funktion aus einer Bildanalyse-Software (Smith 2009). Sie steht zur freien Verfügung und wird nach wie vor stetig weiterentwickelt. Unzählige Projekte, die das Kommandozeilenprogramm mit grafischen Oberflächen ausstatten oder die Trainingsmöglichkeiten optimieren sind verfügbar und haben einen großen Anwenderkreis. Da Tesseract eine Linux-Kommandozeilenanwendung ist, ist die im Rahmen von Drittanwendungen oft geforderte Integration gut möglich. Allerdings konnte nur ein geringer Anteil in dieser Studie als funktionell einsetzbar getestet werden. Zudem hat sich viel am Entwicklungsstand der Software zwischen den Versionen 2 und 3 verändert. Die Dokumentation von Tesseract ist zwar recht ausführlich, aber aufgrund der langen Entwicklungstradition gleichzeitig auch sehr umfangreich, intransparent und diffus.

Gleichwohl muss angeführt werden, dass Tesseract zu der meist genutzten Referenz in OCR-Untersuchungen zählt. Denn kaum eine Software wird derartig oft in den Pool von OCR-Studien aufgenommen (Heliński, Kmieciak, and Parkoła 2012; Does and Depuydt 2013; Lopresti 2008; Clausner, Pletschacher, and Antonacopoulos 2013). Natürlich aber auch, weil es im IMPACT-Projekt in großen Umfang zur Anwendung kam.

Die zweite Open-Source-Software im Pool der hier genutzten Software Implementationen ist Ocropus. Sie war zu Beginn ihrer Entstehung als Frontend für Tesseract zu verstehen, da profunde Leistungen, in den allerersten Ansätzen noch auf Tesseract beruhten (Breuel 2008). In der Zwischenzeit konnte sich die Software allerdings von seinem prominenten Konkurrenten emanzipieren und setzt wesentlich stärker auf state-of-the-art Komponenten wie Multi-layer-perceptrons (MLP/mehrlagiges Perzeptron), Komponenten-Modellen (Breuel 2009a) und modernen Algorithmen zur Mustererkennung.[15] Ein wichtiger Aspekt von Ocropus ist die Bereitstellung von elementaren Funktionen wie Binarisierung, Segmentierung und der Generierung von visualisierten Ergebnissen in Form des eigens entwickelten Standards hOCR (Breuel). Ein hervorzuhebendes Entwicklungsergebnis ist die modulare Architektur der Open-Source-Software. Neben dem Erkennen von Zeichen beinhaltet Ocropus wichtige

15 http://arstechnica.com/information-technology/2007/10/hands-on-with-googles-ocropus-open-source-scanning-software

Arbeitsschritte, die für eine OCR-Software notwendig sind, um mit kommerziellen OCR-Produkten (wie z.B. B.I.T.-Alpha) vergleichbar zu bleiben. Hauptentwickler der Software ist Thomas Breuel von der Universität Kaiserslautern und dem Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI). Der große Erfahrungsreichtum aus den Bereichen Machine-Learning, Mustererkennung und Bildverständnis fließt in die Software ein. Neben den eigenen Entwicklungen zur Zei-chenerkennung (Tesseract), wird auch das Ocropus-Projekt über den gesamten Zeitraum bis heute von Firmen wie z.B. Google oder IBM gefördert. Die Alpha-Version aus dem Jahr 2007, hat konstant inkrementelle Verbesserungen erfahren, die von Version zu Version die Software einsatzfähiger machte und schließlich alle Bereiche im OCR-Workflow abgedeckt hat (Breuel 2008). Dementsprechend wird in dieser Studie von Version 0.7 ausgegangen, die im März 2013 veröffentlicht wurde. Seitdem wird von der state-of-the-art Software Ocropus als einer OCR-Software experimenteller Natur gesprochen.

Im Ocropus-Projekt wurde bereits früh das ambitionierte Ziel formuliert, die Version 1.0 zu veröffentlichen (Breuel 2009b). Oft werden in OCR-Projekten ähnlich ambitionierte Ziele formuliert, um im Anschluss wieder relativiert zu werden. Ein aktuelles Beispiel ist das eMOP-Projekt der Universität Texas, in dem ursprünglich 10 Seiten pro Sekunde zu prozessieren geplant waren[16] und nach der Hälfte des Projektzeitraums wurde dieses Vorhaben, das zwar mit reiner Brute-Force geplant war und extrem hohe Performanz demnach vorwies, wieder relativiert. Es kommt also nicht nur auf die angewandte Software im Projekt an, auch die Infrastruktur, auf der ein Vorhaben aufsetzt, ist z.B. im Hinblick auf In-/Output-Zeiten ein wesentlicher Faktor.

Das Softwareprodukt B.I.T.-Alpha stellt auf Basis der Berliner Untersuchung von Funeralschriften (Federbusch, Polzin, and Stäcker 2013) eine von zwei kommerziellen Software-Implementation dar, die in der vorliegenden Studie berücksichtigt werden. Sie wird entwickelt vom Ingenieurbüro „Ingénieur Tomasi" und ist im Projektzeitraum dieser Studie von diesem mehrfach aktualisiert worden. Herauszustellendes Merkmal ist der gewaltige Konfigurationsumfang der Software, der bereits in "Volltext via OCR" hinreichend beschrieben wurde. B.I.T.-Alpha hat bereits in diversen Digitalisierungsprojekten Anwendung gefunden und wird in großem Umfang über Dienstleistungsverhältnisse mit vereinbarten Qualitäten vom Entwickler genutzt (Federbusch, Polzin, and Stäcker 2013).

Modelle, die in der Arbeit mit BIT-Alpha entstehen, lassen sich in der Software abspeichern und stehen somit evaluierenden Maßnahmen grundsätzlich offen. Die Entwicklungsaktivitäten des Autors sind – wie bereits angedeutet – sehr hoch, allerdings liegt keinerlei Dokumentation des Produktes vor. Die Software verfügt zudem in keiner Form über Schnittstellen. Dies ist insofern problematisch, da in mittleren bis großen Digitalisierungsprojekten Architekturen vorliegen, bei denen davon auszugehen ist, dass die Komponenten zur Retrodigitalisierung in größere Kontexte eingebunden werden müssen. Dies wird an Projekten, wie dem Deutschen Textarchiv[17] der verschiedenen Projekte der Bayerischen Staatsbibliothek oder dem IMPACT-Projekt deutlich.

16 http://idhmcmain.tamu.edu/projects/Mellon/Gantt.pdf
17 http://deutschestextarchiv.de

Im Fall der BSB-Projekte fallen recht heterogene Konglomerate von Projekten zusammen und daher ist oft nicht transparent, welche OCR-Vorhaben mit welchen Softwarestrukturen umgesetzt wurden.

Die soeben angeführten Projektträger oder Projekte, sind Anwender der Software-Implementation von ABBYY. Dabei kommen entweder der Finereader in der Version 9 zum Einsatz, der Recognition Server in Version 3.5 oder auch das Software Development Kit (SDK) von ABBYY. Letzten Endes wird die ABBYY-OCR-Engine in der Version 7/8/9 oder 10 genutzt, was eine stetige Entwicklungsaktivität nachweist. ABBYY bietet für kleine und große Projekte Lösungen und Lizenzmodelle an, die bei Bedarf erweitert werden können. Hier zu nennen ist der Recognition-Server für Fraktur, bei dem eine Vielzahl von Arbeitsstationen eingerichtet werden können oder die SDK, die isolierte Funktionen in dokumentierter Form anbietet, um die funktionalen Teile der ABBYY-SDK in höherwertige Kontexte stellen zu können. Dies ist auch der Tatsache zu verdanken, dass der Anbieter von ABBYY bereits seit langen Jahren über die oben genannten Projekte (bspw. Bayerische Staatsbibliothek, (BSB)) an der Software weiterentwickelt.

2.7 Evaluation

Zur Evaluation der Qualität der Texte, die aus einer Software kommen, wurden verschiedene Korpora und Methoden entwickelt. Im Folgenden wird daher auf die Korpora und Methoden eingegangen.

Im ersten Korpus (Kapitel 4.1) wird die Trennung von repräsentativen und komplexen Daten zentral. Zum Anfertigen von Trainingsmodellen und zum Testen mit bereits bestehenden Modellen, die von den Autoren der Software kommen, ist dieses Korpus vorgesehen. Im zweiten Korpus (Kapitel 4.2) werden die massenhafte Verarbeitung und die Überprüfung der bisher entwickelten Workflows als Ziel formuliert. Das dritte Korpus weicht vom ersten in seiner Zielsetzung wiederum stark ab: Zum einen werden nun Texte in hoher Qualität genutzt und zum anderen werden die Daten nicht mehr durch die unten beschriebenen algorithmischen Lösungen evaluiert. Sie müssen manuell evaluiert werden, da dieses Korpus auf der Annotation von Problematiken beruht, die in OCR-Software als etabliert angesehen werden.[18] Auf die postulierten Korpora wird in Abschnitt 4 dieses Berichtes eingegangen.

In vier grundlegenden Schritten wird von der Datenvorbereitung bis zum Vergleich der Daten gegenüber Groundtruth-Daten der komplette Ablauf dargelegt. Damit wird eine gleichbleibende Struktur zugrunde gelegt, die es ermöglicht, die Ergebnisse miteinander zu vergleichen. Der Fokus der Evaluation der ersten beiden Korpora legt daher den Vergleich von Distanzwerten zwischen Groundtruth und einem Ergebnis einer OCR-Implementation zu Grunde. Die Validität der Ergebnisse bleibt in Bezug auf die Qualität der Groundtruth-ähnlichen Daten allerdings ungeklärt. Um die Qualität der Groundtruth-ähnlichen Daten zu überprüfen, ist es notwendig die Güte einer behaupteten Wahrscheinlichkeit einer Software von der Güte eines Textes zu trennen (Federbusch, Polzin, and Stäcker 2013). Dieser Rückgriff ist insofern notwendig, da man zwar nach einer Evaluation über Distanzwerte zu Ergebnissen einer spezifischen Software kommt, aber die eigentliche Qualität dennoch unbestimmt bleibt. Obwohl man davon ausgehen könnte, dass laut postulierter Skala (Holley 2009), sich alle online verfügbaren Texte

18 http://impactocr.wordpress.com/2010/05/07/an-overview-of-technical-solutions-in-impact

durch eine sehr hohe Präzision auszeichnen, ist dies nicht der Fall. Eine Überprüfung der Groundtruth-ähnlichen Daten wäre daher zusätzlich notwendig.

Nachdem die Qualität der Groundtruth-Daten bestimmt wäre, könnten erst mögliche Rückschlüsse in Bezug auf die Ergebnisse und die Güte der Erkennungsqualität einer Software geliefert werden.

In der vorliegenden Studie wurden zwei Methoden zur Prüfung der Qualität des Textes entwickelt und angewandt. Bei der ersten Methode sind Distanzen berechnet worden, die auf der Basis von Zeichenkettenvergleichen arbeiten. Bei der zweiten Methode haben automatisierte Verfahren keinen Sinn mehr ergeben, denn hier war es notwendig die einzelnen Problemfeatures manuell zu betrachten. Dazu wurden größere Übersichten erstellt, auf die an späterer Stelle noch näher eingegangen wird.[19]

Statistische Methoden, die erst bei hochqualitativen Ergebnissen vorliegen, bilden eine zentrale Rolle bei der Untersuchung und Eruierung von OCR-Resultaten. "Word-space"-Modelle basieren im Kern auf einer Verteilungshypothese, die besagt, dass Worte mit ähnlicher Bedeutung dazu tendieren in einem ähnlichen Kontext aufzutreten. Es ist daher möglich, dass eine Person in variierenden Quellen häufig in einem ähnlichen Kontext auftritt. Dieser Kontext kann sehr vielfältig sein: Chronologisch, orthographisch, topologisch, sozialer Art, berufsgebunden, nach Bildungsgrad, per Laufbahn oder über Personenkreise. Anhand dieser unterschiedlichen semantischen Dimensionen lassen sich wiederum Suchräume eingrenzen und als Filter für wissenschaftliche Zwecke nutzen.

2.8 Resultate

Eines der größten Probleme ist die unbekannte Qualität, die in allen existierenden Korpora vorherrscht. Ohne verifizierte Volltexte gestaltet sich die Formulierung valider Aussagen als schwierig oder wenig gehaltreich.

Eine weitere Schwierigkeit ergibt sich aus der Software und deren strukturellen Vorgaben. Dabei ist nicht der Format-Transfer, der lediglich einen Aufwandsfaktor bedeutet, gemeint, sondern vielmehr die Art und Weise der Prozessierung und deren Repräsentation. So ist es beispielsweise in Ocropus eine Vorgabe, dass alle Zeilen einer Seite einzeln der Schrifterkennung zugeführt werden müssen. Durch den modularen Aufbau sind Fehlerkorrekturen anders gelagert als bei den anderen OCR-Lösungen. Ob eine Fußnote oder Referenz als Zeile über dem Text korrekt positioniert ist, ist im Fall von Ocropus nicht automatisiert überprüfbar und müsste manuell durchgeführt werden (wie im hier vorliegenden dritten Korpus durchgeführt).

Andere ergonomische Engpässe lassen sich gut am Beispiel von BIT-Alpha erklären. Effekte von Änderungen der Konfiguration lassen sich nur überaus schwierig nachvollziehen, da die Software starke Schwächen in ihrer Oberfläche und der UI-Implementation aufweist. Zusätzlich werden nicht-nachvollziehbare interne Abläufe als konfigurierbar angeboten. Der Zweck einer Konfiguration ist nicht nachvollziehbar, denn Informationen zur Funktionsweise von Algorithmen

19 zu finden unter http://dariah.hki.uni-koeln.de

unterliegen laut Hersteller dem Betriebsgeheimnis (im Fall von BIT-Alpha ein selbst entwickelter Binarisierungsalgorithmus, der konfigurierbar aber nicht dokumentiert ist).

Auch den Konfigurationsmöglichkeiten des Trainings von Modellen und den Verarbeitungsschritten ist nur sehr schwierig oder teilweise gar nicht nachzukommen. Sehr rudimentär vorliegende Dokumentation erschwert den Umgang gegenüber Open-Source-Varianten zusätzlich. In diesen könnte man noch im Quelltext nachschauen.

Eine weitere Schwierigkeit ist durch die stetige Weiterentwicklung jeder angeführten Software gegeben. Gerade die Open-Source-Varianten unterliegen stetigen Iterationen. Ocropus liegt in Version 0.7 vor, Tesseract in Version 3.02, der Recognition Server von ABBYY in der Version 3.5 (früher Finereader 7 bzw. XIX) und BIT-Alpha in der Version 6.5.8. Die Referenztexte, die hier als Groundtruth-ähnliches Material vorkommen, der bei den Helmstedter Drucken hingegen sind mit der Version 5.4 der BIT-Alpha OCR-Implementation entstanden. Die vorliegende Untersuchung hebt nicht darauf ab, eine bestimmte Software zu empfehlen. Vielmehr werden Differenzen zu aktuell hoch-qualitativen online-verfügbaren Ergebnissen verdeutlicht. Aufgrund dessen macht nur eine Längsschnitt-Evaluation Sinn. Die Einbindung von Software Schnittstellen, in Form von SDK´s oder Programmierschnittstellen/application programming interfaces (API), ist je nach Projektumfang grundsätzlich abzuwägen. Eine aussagekräftige Struktur, in der die Implementationen wirklich auf ihr volles Leistungspotenzial untersucht werden, würde eine valide Bestimmung der Güte voraussetzen, die bis dato fehlt. Annäherungen an die Bestimmung von Distanzen zu Groundtruth-ähnlichem wird zunächst angestrebt.

Als vorläufige Resultate lassen sich festhalten, dass der von Staecker angeführte Pareto-Effekt (Federbusch, Polzin, and Stäcker 2013) im Bereich von Volltextdigitalisierung nicht nur durch die Arbeit an weiterführenden Korpora zu kompensieren ist. Auch die Abgrenzung von Bereichen, die im Zusammenhang mit Ort, Zeit oder Drucker stehen, sind sachdienliche Annäherungen an die Frage, ob man beispielsweise vom Drucker als Merkmal zur Klassifikation ausgehen kann. Eine weitere, aus den Resultaten dieser Fallstudien, ergangene Erkenntnis ist, dass es ohne technische Neuentwicklungen im Bereich der Programmierung, es weiterhin zu geringen Qualitäten der Ergebnisse kommen dürfte.

2.9 Diskussion

Es ist dringend anzuraten, dass Wissenschaftler selbstständig in die Lage versetzt werden müssen weitere Forschungsergebnisse zu erzeugen, die auf öffentlich-zugänglichen Architekturen liegen und demnach weniger organisatorischen Aufwand auf Seiten der Projektträger bergen. Aus diesem Grund ist in der vorliegenden Studie ein Teil der Arbeit in die Untersuchung von Möglichkeiten zur Bereitstellung investiert worden. Zwar wäre die Bereitstellung von Daten im „Tagged Image File Format" (TIFF) technisch ohne größere Einschränkungen zu bewerkstelligen, doch würden die Datenmengen Forscher in ungeahnte Abgründe führen. Auf der einen Seite geht aus der Fallstudie um JPEG/TIFF-Unterschieden hervor, dass die höchste Qualität nicht unbedingt zielführend ist und zusätzlich in Zusammenhang mit den Graustufen und Farbintensitäten stehen kann (Dhiman and Singh 2013). Im Rahmen von Big-Data-Studien könnten Wissenschaftler so eigenständig Resultate produzieren, die über Projektanträge und Bewilligungsebenen so nur schwierig oder in langwierigen Prozessen zu realisieren wären.

Da das Schlagwort Big-Data[20] nicht nur genutzt wird, um die zur Untersuchung von enorm großen Datenmengen beschriebenen Effekte zu benennen, wird auch extreme Komplexität untersuchbar, die sich durch eine Infrastruktur und den Zugriff auf Volltexte ideal untersuchen lässt. Es ist notwendig semantische Informationen nutzen zu können. Neben der Ergänzungen durch die Metainformation eines jeden Prozessschrittes, macht es auch Sinn intrinsische Informationen zur Erforschung historischer Inhalte bereitzustellen. Dementsprechend machen Lexika (z.B. Thesaurus Professionum (THEPRO) oder Thesaurus Locorum (THELO)) zwar einen enormen Sinn und sind zielführend, aber um sie der breiten Masse von Forschern zur Verfügung stellen zu können, sind Wissensmodelle in Form von Ontologien notwendig, die Datenbestände miteinander vernetzen. Das gilt genauso für den Wissenstransfer von extrinsischen Merkmalen.

Es bleibt demnach im Sinne des zuvor angeführten Pareto-Effekts der Hinweis auf die Nutzung von Bibliotheken und die Nutzung der vorliegenden Forschungsinfrastrukturen bestehen. OCR-Ergebnisse in transparenteren Kontexten zu sichern und deren Fortschritt nicht nur zu gewährleisten, sondern standardisiert und persistent zu dokumentieren ist ein Anliegen, auf das in dem vorliegenden Bericht immer wieder hingewiesen wird. Dieser ergeht auch an der schier unübersehbaren Heterogenität des Quellenmaterials und dessen Behandlung. Die nächste Stufe von OCR-Vorhaben wird sich auf die Erkennung von Handschriften beziehen, da sich in diesem Bereich bereits technisch ähnliche Lösungen, wie sie beispielsweise bei Ocropus genutzt werden, etabliert haben und die Erforschung um Sprachmodelle ein essenzieller Bestandteil dabei ist (Dhiman and Singh 2013).

2.10 Praktische Relevanz

Neben der Relevanz expliziter Auszeichnungsmöglichkeiten durch konsensuelle Serialisierungsformate ist die Interoperabilität damit grundsätzlich gewährleistet und ein progressiver Verlauf für Forschungsprojekte oder die Wissenschaft allgemein nicht mehr erschwert.

Konkret heißt das: Kann eine wissenschaftliche Arbeit, die sich nur mit der Annotation von Störungen intrinsischer Natur beschäftigt als Basis zur Untersuchung und Weiterentwicklung von Algorithmen zur Binarisierung genutzt werden? Würden daraus resultierende Fortschritte oder Erkenntnisse neben der schriftlichen Überlieferung auch einer datentechnischen Präservierung zugeführt, würde die Problematik nicht mehr entstehen. Vor dem Hintergrund der hohen Heterogenität des Quellenmaterials im Bereich von Retrodigitalisierung wäre dies eine minimale Anforderung. Wüsste man welcher Drucker welches Papier mit welchen Drucktypen in was für einer Druckpresse und der genutzten Farbe auf einem analogen Medium vorliegt, wären die Abläufe wesentlich transparenter nachzuvollziehen und den Aufwand für digitale Forensik könnte man einsparen (de Castro Polastroa and Almeidac).

Die Verschaltung der Arbeitsebenen und aller Prozessschritte ist daher ein wichtiger Schritt zur Erschließung der gesamtdeutschen Bestände der frühen Neuzeit und darüber hinaus. Dabei ist nicht unbedingt wichtig, dass die Daten für Forscher in exzellenter Qualität vorliegen, sondern vielmehr ist wichtig zu wissen, wie die Daten zustande gekommen sind und auf welchen Techniken sie beruhen. Im Umkehrschluss besteht dann die Möglichkeit, für diese Daten

20 http://mike2.openmethodology.org/wiki/Big_Data_Definition

Techniken zu entwickeln, die ein wiederkehrendes Szenario lösen und mit haltbaren Ergebnissen belegen.

Angenommen das oben genannte Szenario soll in einem initiierenden Testlauf umgesetzt werden. Alle digitalisierten Drucke des VD16, VD17, VD18 sollen Volltext-digitalisiert persistent abgelegt werden. Betrachtet man die Bestandszahlen der vorliegenden Drucke in den einzelnen Verzeichnissen ergibt sich folgende Tabelle:

Verzeichnis	Anzahl Drucke	Anzahl Digitalisate	Anzahl Seiten
VD 16	~ 110.000	~ 33.000	n.a.
VD 17	~ 282.000	~ 68.000	> 750.000
VD 18	~ 600.000	~ 110.000	17.828.856
VD 16-18	~ 992.000	~ 211.000	n.a.

Abbildung 2: Mögliche Konfigurationen eines Multi-Stage Kombinations-Systems.

Geht man davon aus, dass jede Seite ca. 30-90 MB unkomprimierte Bilddaten im TIFF-Format ausnutzt, bekommt man eine Vorstellung von der Größe der Datenmengen. Zum VD16 und VD17 konnten keine Angaben zum Seitenanfang gefunden werden. Im VD17 sind allerdings die Schlüsselseiten, wie etwa das Deckblatt oder Inhaltsverzeichnis, zu finden. Auch für die Anzahl der Drucke des VD18 steht noch in keinem korrekten Zusammenhang mit der Anzahl aller Digitalisate. Denn hier wird aktuell noch immer digitalisiert und die Zahl steigt stetig. Gleichwohl soll mit dieser Rechnung eine perspektivische Datenmenge bestimmt werden, die zunächst aus den Daten des VD18 einen interpolierten Durchschnitt für die Gesamtanzahl nutzt.

Wenn also 110.000 Digitalisate etwa ~18 Millionen Seiten enthalten, dann haben 210.000 verfügbare Digitalisate einen Seitenumfang von ~33 Millionen Seiten. Über diese interpolierte Hochrechnung von ~30-90 Megabyte (MB) pro Seite entstehen dann ~990-2970 Millionen MB, was umgerechnet ~966.796-2.900.388 Gigabyte (GB) oder ~144-2832 Terrabyte (TB) entspricht. Diese Zahlen sind allerdings bestenfalls als Annäherung zu verstehen und dienen einer Argumentation zum Vorgehen und zur Veranschaulichung der Datenmenge. Da diese weiter steigt (es wird stetig weiter digitalisiert), sollten Netzwerke großer Datenspeicher vereint werden. Die Öffnung bestehender Systeme für die Geisteswissenschaft steht allerdings noch aus.

3. Die Fallstudien

Im Rahmen des vorliegenden Berichts wurden verschiedene Fallstudien durchgeführt. In einem ersten Korpus wurde der Workflow (s.u.) entwickelt, mehrfach überprüft und iteriert. Inwiefern es möglich ist Trainingsmodelle für die einzelnen Software-Implementationen wurde im Rahmen dieser Studie geklärt. An diesem Korpus wurden alle Methoden zur Evaluation entwickelt. In einem zweiten Korpus, indem sehr heterogenes Material vorliegt, wurde dann überprüft, inwiefern eine Software-Implementation für massenhafte Verarbeitung geeignet ist. Die aus dem ersten Korpus resultierenden methodologischen Ergebnisse wurden hier erneut überprüft, verifiziert und gegebenenfalls angepasst. Ein drittes Korpus leitet in Abgrenzung zu den Korpora 1 und 2 eine qualitative Untersuchung von Fehlerquellen her.

Alle drei entwickelten Korpora sind als Liste von Signaturen im Anhang dieser Studie enthalten. Die Korpora sind im Rahmen dieser Studie im JPEG-Format und TIFF-Format vorhanden. Dabei wurde der Ansatz verfolgt, die Frage nach der geeigneten Bereitstellung von Datenmaterial, vor dem Hintergrund von begrenzten Transferzeiten und Speicherkapazitäten, zu klären.

Die Daten der Korpora entstammen aus den Beständen der Herzog-August-Bibliothek Wolfenbüttel und den damit verbundenen Helmstedter Drucken, den Beständen der Niedersächsischen Staats- und Universitätsbibliothek Göttingen und den Beständen der Bayerischen Staatsbibliothek. TIFF-Daten wurden auf Anfrage bereitgestellt, wohingegen die Akquise der JPEG-Daten durch Screenscraping-Verfahren über die Internetseiten der jeweiligen Bibliotheken geschehen ist. Auf die gleiche Weise sind auch die Groundtruth-ähnlichen Texte automatisiert heruntergeladen worden. Der Fokus lag dabei auf den Texten und nicht auf den Layoutinformationen. Aus diesem Grund sind auch alle Groundtruth-ähnlichen Texte von Zeilenumbrüchen befreit worden, so dass zuletzt der gesamte Text eine einzelne Zeile umfasst. Das hat zunächst den Vorteil, dass Layouterkennung nicht Teil der Evaluation ist und das Hauptaugenmerk auf die korrekte Erkennung der Texte gelegt wurde. Um die Präzision der Texterkennung zu bemessen, wurde sie in eine lange Zeichenkette übertragen, um sie Vergleichsanalysen zuzuführen. Dies geschah vor dem Hintergrund der zum Zeitpunkt dieser Studie noch gering entwickelten Segmentierungsalgorithmen und dem Fokus auf der Güte der Zeichenerkennung.

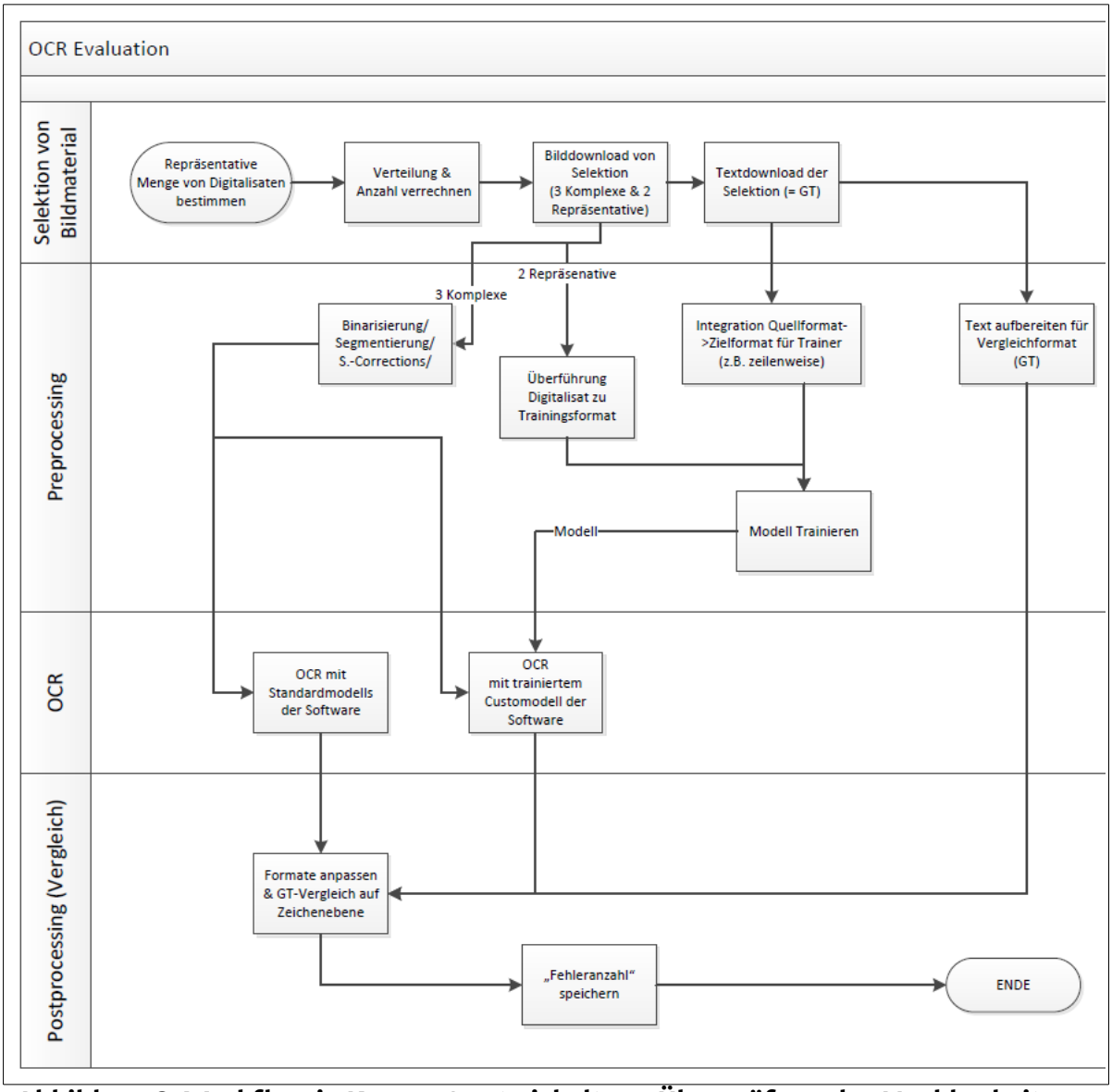


Abbildung 3: Workflow in Korpus 1 entwickelt zur Überprüfung der Machbarkeit von Trainingsmodellen.

3.1 Fallstudie 1 – Korpus 1

Aufgrund der Verfügbarkeit bestehender Digitalisate wird in einem ersten Lauf ein Korpus entwickelt, das einer chronologischen Ordnung unterliegt. Es wurden dabei zu Beginn des Projekts ausschließlich die Helmstedter Drucke der Herzog August Bibliothek (HAB) in Wolfenbüttel verwendet.

Die erste Version des Testsets besteht aus einer Menge von 100 Signaturen, die repräsentativ für problematische OCR-Bestände stehen. Eine Signatur steht dabei nicht zwingend für ein Buch. Die Referenzierung der genutzten Werke wird über die Angabe der Signatur und Seitennummer auch in diesem Bericht gemacht.

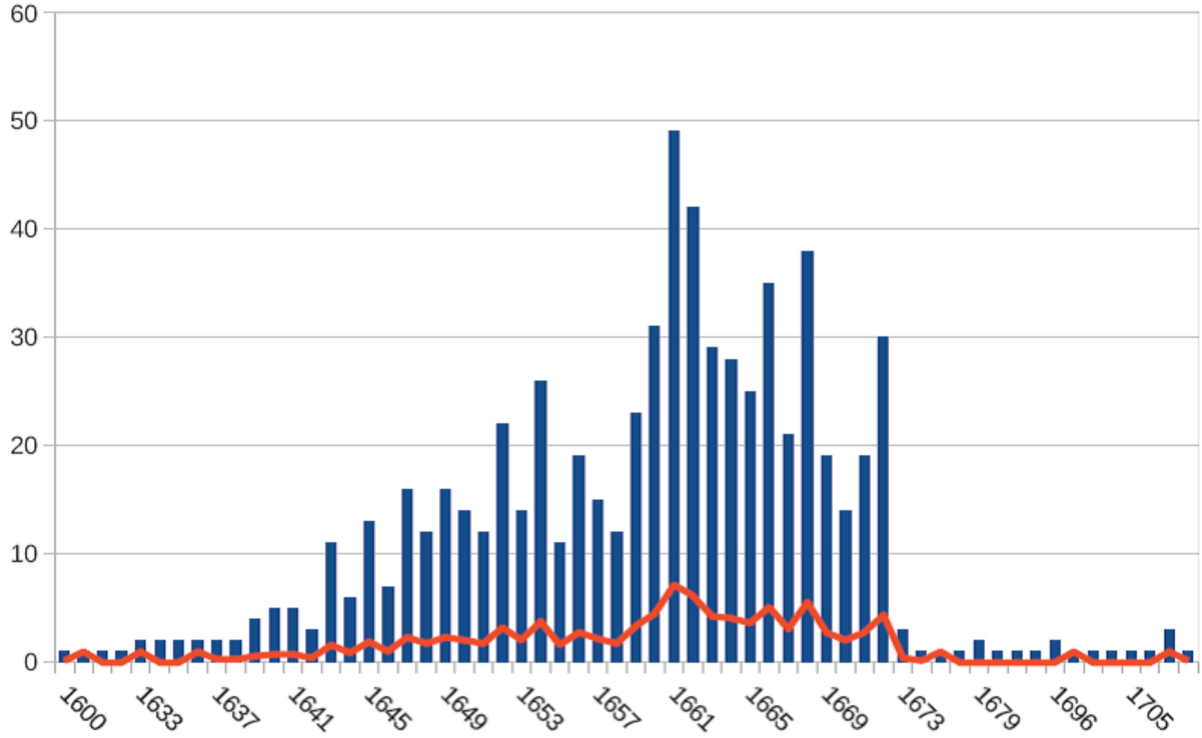


Abbildung 4: Übersicht des Zeitraums verfügbarer Drucke (blau) der HAB und Darstellung der Selektion der 100 Testkandidaten (rot).

Es wurden aus 694 online verfügbaren Volldigitalisaten (Volldigitalisate sind der Verbund von Bild und Text als Ergebnis einer OCR-Prozessierung) zufällig 100 Stück selektiert, die quantitativ und nach einer repräsentativen Häufigkeit über den Zeitraum aufgeteilt sind. Die tatsächliche Anzahl der verfügbaren Seiten eines Werkes ist dabei kein Kriterium gewesen. Das folgende Diagramm verdeutlicht das Volumen der verfügbaren Signaturen und erläutert die Selektion der 100 Testkandidaten daraus:

Der Großteil, der in Wolfenbüttel entstandenen Volltext-Digitalisate (Stand: 01.09.2013), liegt im Zeitraum von 1643-1672. Die bekannten Schwierigkeiten dieser Daten können in Effekte extrinsischer und intrinsischer Natur differenziert werden (Federbusch, Polzin, and Stäcker 2013). Man spricht von extrinsischen Effekten, wenn zum Beispiel schwankende Kontrastverhältnisse innerhalb eines Digitalisat entstanden sind. Also immer dann, wenn der OCR-Prozess an strukturelle Vorgaben des Quellenmaterials gebunden ist. In den meisten Fällen stehen Materialgegebenheiten oder Vorgehensweisen der Digitalisierung mit extrinsischen Effekten in Verbindung. Intrinsische Merkmale sind beispielsweise der Wechsel von Sprache oder verwendetem Zeichensatz (Federbusch, Polzin, and Stäcker 2013). Beide Problemaufrisse werden im späteren Verlauf dazu verwendet, um solide Kriterien zu entwickeln, an denen anschließend ein spezifisches drittes Korpus aufgebaut wird (ein vorläufiges Verzeichnis von 'Problem'-Features befindet sich in Kapitel 3.3).

Die 100 selektierten Signaturen sind dann manuell in einer zweiten Stufe in komplexe und repräsentative Seiten klassifiziert worden. Dazu wurden zuerst per Zufallsalgorithmus (PHP-Methode rand()) 15 Seiten aus dem gesamten Digitalisat automatisch ausgewählt, um danach 5 Seiten pro Signatur manuell zu klassifizieren. Die komplex-klassifizierten Seiten bilden einen Umfang von jeweils drei Seiten, um möglichst viele der oben genannten Probleme abzubilden. Die als repräsentativ klassifizierten Seiten entsprechen dem Großteil der im Digitalisat/Werk vorhandenen Seiten. Bei dieser Auswahl sind die ersten drei und letzten drei Seiten, sowie Einbände und Titelblätter vorweg herausgefiltert worden.

Insgesamt sind so 500 Testseiten in den Korpus 1 eingefügt worden. 300 zur Erstellung von Trainingsmodellen, die auf die 200 Repräsentativen angewandt werden sollten. Nach der Klassifikation ist eine erste Erweiterung des Korpus durchgeführt worden, bei dem keine Klassifikation aufgrund des Umfangs durchgeführt wurde. Das Testfeld und dessen Erweiterung werden in den folgenden Korpora immer impliziert.

3.1.1 Resultate aus den Untersuchungen um Korpus 1

Im Anschluss an die Verteilung der 500 Seiten auf eine Software-spezifische Ordnerstruktur, wurden sie jeder OCR-Implementation zur Prozessierung zugeführt. Die klassifizierten Seiten sind als Modelle für ein Training in der Software Tesseract und Ocropus genutzt worden. Aufgrund des fehlenden Sprachmodells in Tesseract für die Sprache Latein (roman-latin), sind die Ergebnisse durch hohe Distanzen in der Wiederverwendung gekennzeichnet. Auch das Modell von Ocropus hat sich als nicht-wiederverwendbar herausgestellt. Da aktuell nur ein geringer Anteil der Zeichen aus dem Unicode Transformation Format (hier: UTF-8) Zeichensatz unterstützt wurde. Alle unbekannten Zeichen, die dem Prozess als Zielvorgabe vorgegeben wurden, sind im späteren Volltext als Unicode-Zeichen aufgetreten. Das hängt zum einen mit dem begrenzten Zeichensatz zusammen und zum anderen in der Inkonsistenz der Software-internen Abbildung von Zeichen. Das Training von Modellen stellt sich im Hinblick auf geringe Wiederverwendbarkeit und hohen Ressourcenaufwand als kaum machbar heraus und es wird daher empfohlen trainierte Modelle an einzelnen Werken zu orientieren. Im Folgenden wird dargestellt, wie die gesamten 500 Seiten mehrfach durch die OCR-Prozesskette geschickt werden. Daraufhin werden Distanzen zwischen den Volltexten berechnet und visualisiert.

In den Prozessabläufen hat der ABBYY-Recognition Server 3.5 den am höchsten iterierten Entwicklungsstand aufgewiesen. Die Verteilung über mehrere Arbeitsstationen, Fraktur-Unterstützung, viele Sprachmodelle, hohe Geschwindigkeit und verteilte Arbeitsinfrastrukturen (Scan, Korrektur, Indizierung) sind wichtige Faktoren dieser Bewertung. Die schnelle Erarbeitung eines Konfigurationssets sind essenzielle Ansätze, die eine Anwender-Erfahrung sehr positiv beeinflussen. Im Gegensatz zur zweiten kommerziellen Lösung BIT-Alpha steht hier das Ergebnis weniger im Zusammenhang mit der Komplexität des Konfigurationssets. In beiden Produkten ist die Anwendung dieser Sets elementar und Trainingsmöglichkeiten sind stark beschränkt oder nicht zugänglich oder nachvollziehbar. In den beiden freien Implementationen (Tesseract und Ocropus) sind die Möglichkeiten der Konfiguration anders gelagert. Reine Textdateien, parametrisierte Methodenaufrufe oder die direkte Integration einer Schnittstelle in übergeordnete Konzepte, eröffnen potenziell ein universelles Konfigurationsmodell.

Diese Untersuchung geht davon aus, dass möglichst universelle Modelle, die bereits bei der Entwicklung genutzt wurden, eine hohe Leistungsfähigkeit besitzen und nutzt daher die im Softwarepaket hinterlegten bereits trainierten Modelle.

Die erzeugten Volltexte der 500 Testseiten werden als Zeichenkette verschiedenen Verfahren zur Differenzanalyse zugeführt. Das erste Verfahren ist eine Untersuchung der Längen der Volltexte. Erzeugt eine Software viele Störungen, werden teilweise reihenweise Punkte gesetzt, die sich auf eine Untersuchung der Längen negativ auswirken. Starke Filterung von Störungen, wie sie im Fall vom ABBYY Recognition Server 3.5 gemacht werden, sind daher für ein aufgeräumtes Schriftbild zwar sinnvoll, aber machen den Vergleich sehr unscharf. Allerdings sind die Referenztexte, gegen die hier verglichen wird (bisher Groundtruth-ähnliche Texte genannt), größtenteils mit BIT-Alpha innerhalb dieses Korpus prozessiert. Zudem sind die Referenztexte in einem Dienstleistungsverhältnis mit dem Hersteller der Software entstanden. Trotzdem könnte es Sinn machen die Groundtruth-ähnlichen Texte auf Störungen zu untersuchen. Syntaktische Notationen, die aus einer Reihe oder nur einem Punkt, Komma, Zirkumflex und dergleichen bestehen, würden das Ergebnis dann nicht mehr beeinflussen können.

3.1.2 Korpus 1 – Diskussion

Wie die vorangegangenen Untersuchungen gezeigt haben, besteht im Testset weiteres Potenzial auf Filterung. 100 Kandidaten, die näher untersucht werden, sind daher aus dem Angebot der 500 selektiert worden. Aus 25 Signaturen (100 Seiten) wurden Font-Wechsel in möglichst hoher Anzahl identifiziert, um dieses Problem-Feature zur Leistungsfähigkeit einer Software in Korrelation zu setzen. Ziel dieses Abschnitts ist es daher, eine Möglichkeit zu finden, um Font-Wechsel und die vermeintlich inhärenten Probleme anhand von Spitzen in später angeführten Diagrammen zu visualisieren.

An dieser Stelle wird exemplarisch ein vermeintlich offensichtlicher Fall besprochen: Auf Seite 23 der Signatur ‚223-2-quod-10s' befinden sich vier Drucktypen und zwei Sprachen (Deutsch und Latein). Bei der Untersuchung wird eine Abhängigkeit von Text und Struktur offensichtlich. So wird die Nutzung von Font-Wechsel und Sprachmix auf unterschiedlichen Zeilen von ABBYY, BIT-Alpha fast sehr gut erkannt. Auffällig dabei ist allerdings, dass ABBYY den Font-Wechsel und Sprachmix auf unterschiedlichen Zeilen besser beherrscht als den reinen Font-Wechsel und diese somit im Text fehlerhaft ausgibt. Bei BIT-Alpha ist dies anders, denn hier werden die Sprach- und Font-Wechsel sowohl auf Zeichen- als auch auf Zeilen-Ebene gut erkannt. Zudem wird deutlich, dass eine wesentlich stärkere Streuung von Störungen bei BIT-Alpha vorhanden ist, wie die folgende Darstellung verdeutlicht:

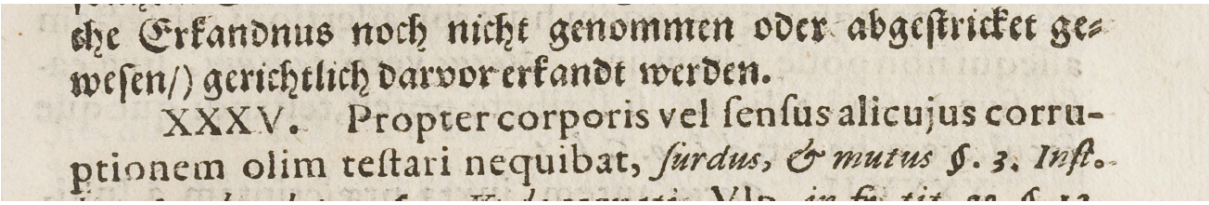


Abbildung 5: Ausschnitt des Original-TIFF aus Signatur '223-2-quod-10s' (S.23) der Helmstedter Drucke aus der HAB in Wolfenbüttel.

Ergebnis BIT-Alpha	ele Erkandnus noch mcht genommen oder abgesuicket ge - wesen /) gerichtnch darorerkandt werden .) c XXV . Proptercorporis vel sensus alicu , us corru - ptionem olun testau nequibat , surdue , & mutus § . 3 . Irst .
Ergebnis ABBYY	che Erkandnus noch nicht genommen oder abgestricket ge- wesen/) gerichtlich darvorerkandt werden. XXXV. Propter corporis vel fenfus alicujus corru- ptionem olim testari nequibat, & mutus $ .3. Infl·.
Ergebnis Ocropus	cye eCrfanbnue nocf) nidt genomnttn obcr abgejirifct gca tocjen/)gericytlif barvor trfanbt merben. XXXV. Propter corporis vel fenfus alicujus corru- ptionem olim teltari nequibat,/eraas, exarw g.1, Ief..
Ergebnis Tesseract	clgc @rf'anbnus nod; m'dgt genommm oDw-abgcfira'cfet ge; wefenn gerisbtlicb bawor-crfanbt werbm. XXX V .. P'roptfir corporis vel fenfus-aiicuj us corru- gtinnsm olim teltari nequibat,fi2rdw, émuzw p'. 3, Inf},

Veranschaulichen kann man sich die oben angeführten Ergebnisse, deren Reihenfolge auch gleichzeitig für seine Qualität steht, indem man sich den Absatz 'werden' (Ende von Zeile 2) anschaut und zudem in Zeile 4 'surdus & mutus'. Der Wechsel von gebrochener Schrift auf Serifen-Schrift (hier Antiqua) macht in jeder Software Probleme. Die Unterscheidung von Zeilen- und Zeichenebene ermöglicht es Schwierigkeiten, die Software-intern gelöst werden müssen, zu verdeutlichen.

Trotzdem sind die Problematiken keinesfalls auf eine einzelne Ursache zurückzuführen. Denn fehlende Wörterbücher, fehlende Dokumentation der Software und nicht nachvollziehbare Effekte von Konfigurationsmöglichkeiten finden sich in jeder Software wieder. Natürlich ist in den beiden Open-Source Implementationen ein wesentlich tieferer Einblick möglich. Der aber hier auch als höchster Aufwand einzuschätzen ist, da sie gleichzeitig die geringste Güte von Texten produziert haben.

Es ist anzumerken, dass vor der Zielsetzung von universellen Lösungsansätzen erst solide Methoden zur Evaluation vorhanden sein müssen, um eine Möglichkeit zur Überprüfung von Ergebnissen herbeizuführen. Ausführliche Erkenntnisse werden daher deutlicher, wenn die gesamten 25 Signaturen in ihrer Gesamtheit von Problemquellen überprüft und klassifiziert werden.

Deutlich ist auch, dass Strukturelemente von jeder Software unterschiedlich gehandhabt werden. Dies zeigt der Fall der Signatur 'p529-4f-helmst-3s' (S.114 im Werk) der Helmstedter Drucke. Hier werden verschiedene Strukturelemente in verschachtelten Absätzen und Fußnoten verwendet. Sofern kein Standard zur Verarbeitung von Strukturelementen genutzt wird, bieten kommerzielle Produkte nur teilweise Formate wie TEI, ALTO oder METS an. Dadurch entsteht eine Fehlerrate,

die auf bestimmten Seiten enorm hoch ist. Die rein textuellen Repräsentationen lassen sich demnach auch nicht wieder rekonstruieren im Sinne digitaler Äquivalenz. Sowohl Marginalien als auch Fußnoten können daher nur teilweise behandelt werden und die Handhabe dieser bleibt offen. Gerade Ocropus und Tesseract setzen Marginalien von ihrer Position her aus direkt in den Fließtext, sodass nicht mehr nachvollzogen werden kann, ob es sich um einen Textteil handelt. Partielle Unterstützung von strukturrelevanten Informationen, wie sie durch ALTO bereitgestellt werden, ist in diesem Sinne noch nicht ausreichend, bildet aber bis dato ein Novum, indem Strukturelemente und Inhalt akkumuliert vorkommen.

3.2 Fallstudie 2 – Korpus 2

Nachdem das erste Testset – bestehend aus den Helmstedter Drucken – einer ersten Analyse unterzogen wurde, ist es durch mehr als 2.500 Seiten aus dem Bestand der Bayerischen Staatsbibliothek ergänzt und auf das 15. Jahrhundert ausgeweitet worden. Für den Zeitraum von 1502 bis 1588 bezieht sich das Testset 1 quantitativ und chronologisch auf diese Erweiterung.

Zu allen bisher aufgeführten Digitalisaten liegen mit einer unbekannten OCR Software erzeugte Volltexte vor, die über die BSB-Webseite der Öffentlichkeit frei zugänglich sind. In der vorliegenden Untersuchung wird daher zunächst davon ausgegangen, dass hier höchstens von "Groundtruth-ähnlichem" die Rede sein kann.

Doch auch die bestehenden Korpora haben bereits, wie zu Beginn des Projektes festgestellt, spezifische Probleme. Die Helmstedter Drucke haben beispielsweise zum Teil die Besonderheit eines variierenden Kontrastverhältnisses über alle Seiten bestimmter Werke. Dies könnte – so eine Spekulation – auf eine fehlende Nachjustierung der Lichtsensorik beim Digitalisierungsvorgang zurückzuführen sein. Gerade deshalb ist es sinnvoll Metadaten zum historischen Entstehungs- und Digitalisierungshintergrund in die Prozessierung einzubeziehen. Die Informationen im Entstehungsvorgang der Volltextdigitalisierung sollten unbedingt berücksichtigt werden. Die folgende Tabelle zeigt daher die wichtigsten öffentlich zugänglichen Metadaten im Hinblick auf eine extrinsische Klassifikation oder Gruppierung:

Signatur BSB	Umfang (Seiten)	Erscheinungs-jahr	Verlag	Verlagsort	Autor/hrs
bsb10165677	152	1502	Aldus Manutius	Venedig	
bsb10139422	398	1531	Petri	Basel	Diodorus, Siculus; Boccaccio, Giovanni; Johannes Zonaras
bsb10168344	216	1531	Quentell	Coloniae	Nausea, Friedrich
bsb10203117	44	1552		Viennae Austriae	Corsi, Pietro; Aldus, David

bsb10178305	2785	1557	Oporinus	Basileae	Herold, Johannes
bsb10160279	192	1558		Francoforti	Andreä, Jacob
bsb10175361	144	1566	Plantin	Antverpiae	Sturm, Johannes
bsb10171607	712	1574	Rihelius	Argentorati	Cicero, Marcus Tullius
bsb10157993	25	1577	Gruppenbachius	Tubingae	Andreä, Jacob

Sowohl die Druckorte, als auch die Streuung der Jahreszahlen ist diffuser als das bislang in der Studie verwendete Korpus. Die Prozessierung in allen Implementationen hat sich daher hauptsächlich auf die Qualität der Groundtruth-ähnlichen Texte bezogen. Wohingegen hier auch die Zugänglichkeit zum Quellenmaterial und die Performance der Software als wichtigste Punkte der Untersuchung dargestellt werden.

Die mit Abstand schnellste Software, im Hinblick auf Performance und Usability war ABBYYs Recognition Server. Jede andere Software lag bei diesem Umfang mit Ergebnissen im Tagesbereich weit darüber. So hat BIT-Alpha im kommerziellen Vergleich recht lange für die gesamte Prozessierung benötigt. Aus einer Minute zu Beginn eines Batch-Prozesses sind mehrere Minuten zum Ende eines Batch-Prozesses (bei gleichbleibender Seitenkomplexität) geworden. Nutzt man Ocropus, entstehen bei größeren Mengen auch zusätzliche Verarbeitungszeiten, dadurch dass die stark segmentierten Arbeitsabläufe bei 2.500 Seiten auf mehreren Modellen mehrere Prozessierungen erfordern und somit In-/Output-Vorgänge starke Redundanzen erzeugen. Die Reduktion von Redundanzen stellt hierbei einen nicht zu unterschätzenden Zeitfaktor dar.

Trotz der vorläufigen Irrelevanz im Hinblick auf die eigentlichen Ergebnisse, sind die Digitalisate den gleichen Evaluationsmethoden zugeführt worden wie das Korpus 1. Dabei hat sich herausgestellt, dass die reine Alignierung auf Zeichenebene (Volltext als Zeile) nur teilweise Aussagekraft hat in Bezug auf die Güte eines Textes. Denn auch Leerzeichen und Störungen, die häufig in Form von Punkten, Kommas oder Sonderzeichen auftreten, werden von der Software erkannt. Daher werden auch Copyright-Zusätze, wie im Folgenden beispielhaft dargestellt, prozessiert, als Zeichen erkannt und dem Ergebnis-Text hinzugefügt.

Abbildung 6: Copyright-Hinweis der HAB in Wolfenbüttel. Verfügbar auf allen online verfügbaren Digitalisaten.

Dies ist insofern problematisch in Bezug auf Differenz-Untersuchungen oder die Messung einer Güte, denn es müssen entweder alle Zusätze manuell oder automatisiert vom Digitalisat wieder gelöst werden.

Auch die Markierung von Maßstäben taucht immer wieder in Digitalisaten auf, sodass nicht nur die Evaluation korrekter Zeichenketten berücksichtigt wird, auch die Evaluation von Wörtern als 'Token' sollte mehr berücksichtigt werden.

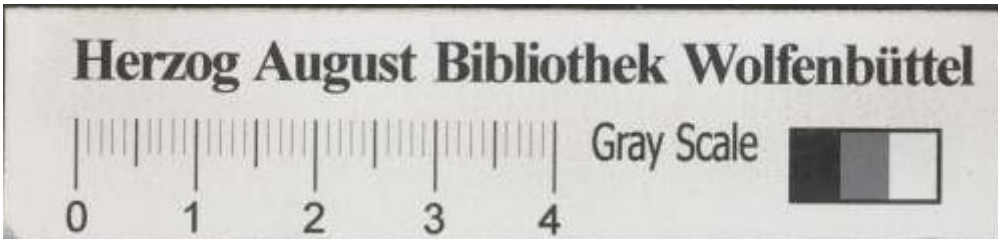


Abbildung 7: Markierung der Maßstäbe, die von Segmentierungs-Algorithmen über den Kontext herausgefiltert werden muss.

Aus diesem Grund kann der Schluss gezogen werden, dass sowohl die Resultate zur Zeichenebene, als auch die Resultate zur Wortebene grundsätzlich in Differenzanalysen miteinbezogen werden sollten. Die folgenden Diagramme stellen die Ergebnisse auf Basis der TIFF-Untersuchung auf Zeichenebene dar:

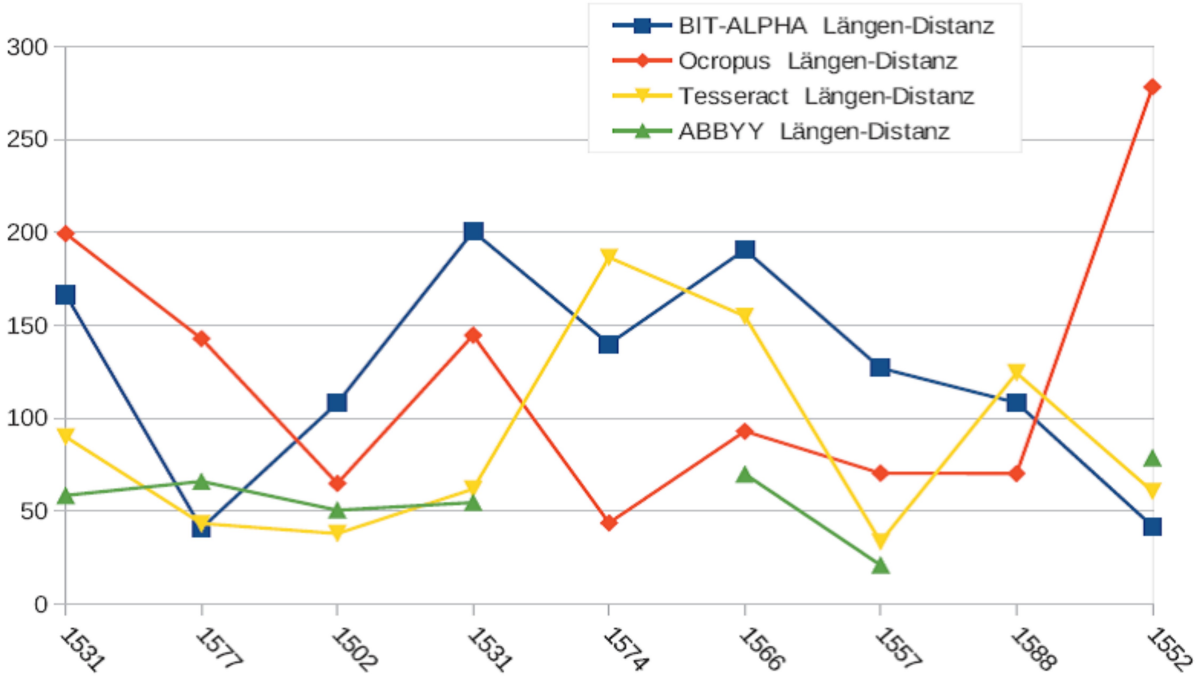


Abbildung 8: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation über die Länge der Texte.

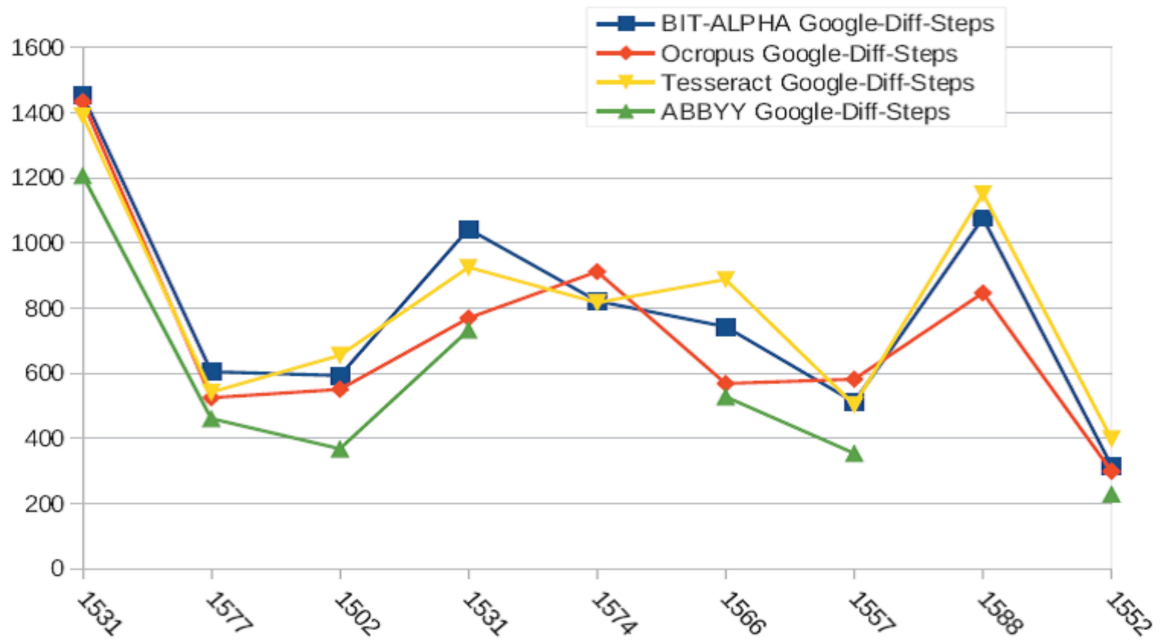


Abbildung 9: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation per Google-Diff-Steps-Algorithmus.

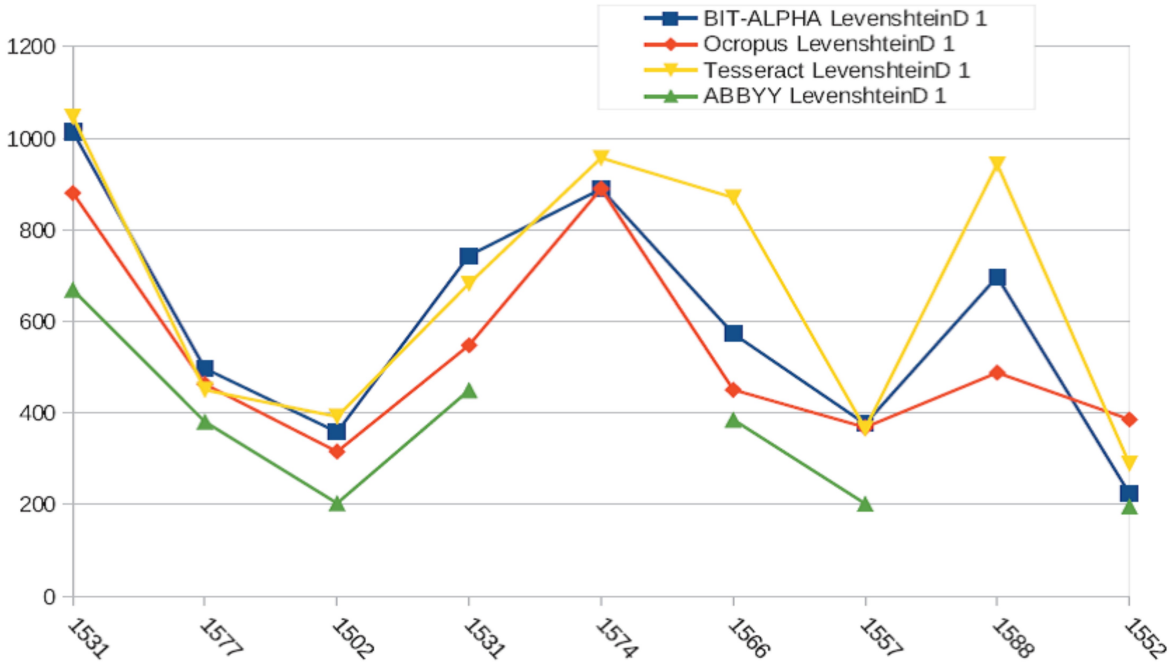


Abbildung 10: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation per Levenshtein-Distanz.

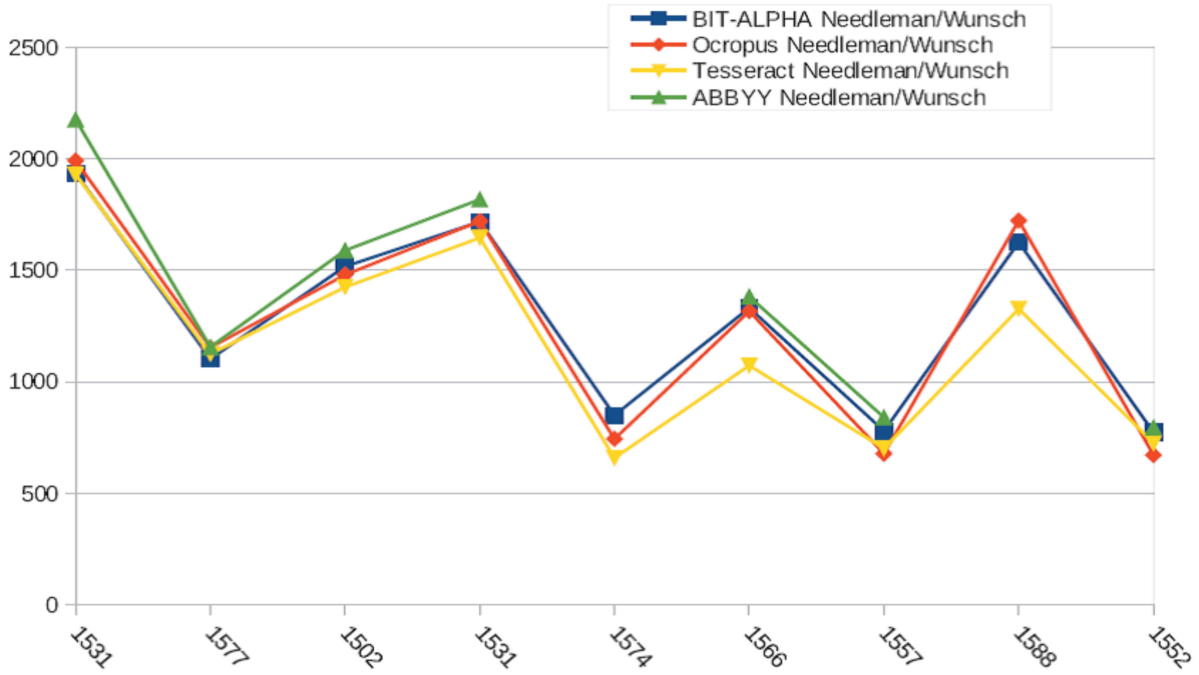


Abbildung 11: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation per Needleman/Wunsch-Algorithmus.

3.2.1 Überlegungen und Anmerkungen aus den Korpora 1 und 2

Die bisher erhobenen Daten und besprochenen Ergebnisse unterliegen zwei grundsätzlichen Schwierigkeiten. Sie sind als relative Bewertung einzuordnen, da die tatsächlichen Groundtruth-Daten für den deutschen Sprachraum fehlen. Eine Transformation der Ergebnisse (zum Beispiel Volltext als Zeile) haben Auswirkungen auf die Aussagekraft der Evaluation.

Aus diesem Grund sollten die angeführten OCR-Lösungen vornehmlich weiterentwickelt werden, sodass sie mehrheitlich zum Einsatz von interoperablen Vorhaben führen. Dazu könnte beispielsweise ein bestehendes Skript zur Zeichenprüfung so weiterentwickelt werden, dass die Qualität nicht nur in Dienstleistungsverhältnissen, sondern auch in Online-Projekten geprüft werden kann. Auf der Basis, der von Staecker postulierten Berechnung von Wahrscheinlichkeiten zum Rückschluss auf die Erkennungsqualität (Federbusch, Polzin, and Stäcker 2013), lässt sich damit effizient statistisch überprüfen welche Qualität vorliegt. Mit einem N-Gramm Ansatz wird dem Nutzer das Ergebnis gleichzeitig mit dem Bildausschnitt angezeigt und dieser entscheidet dann mit einem Tastendruck, ob der erkannte Volltext stimmt.

Der Bezug zwischen Bildausschnitt eines Wortes und erkanntem Volltext lassen sich aus den Angaben entsprechender Strukturelemente effizient ablesen. In ALTO werden die genauen Koordinaten und ein umgebendes Rechteck im erkannten Wort vorgesehen und könnten genutzt werden. Das setzt allerdings voraus, dass die Daten der Strukturelemente korrekt sind. Im Fall von BIT-Alpha waren die Koordinaten der erkannten Zeichen nicht korrekt. Dies steht dem zuvor angeführten Ansatz zur Überprüfung der Qualität entgegen.

Bekommt der Nutzer falsche Bilddaten zu einer Zeichenkette angezeigt, ist jegliche Prüfung der Qualität eines Textes zweifelsfrei hinfällig:

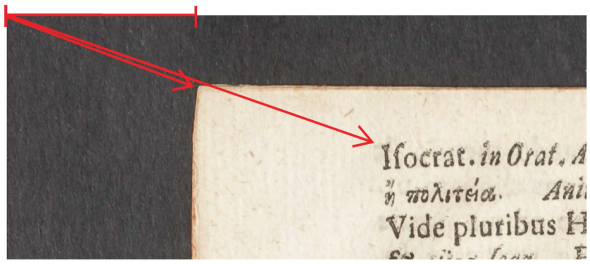


Abbildung 12: Ausschnitt des Digitalisats „li-sbd-115-2s". Das Wort 'lsocrat' befindet sich an der Position x=1089 & y=346 und hat die Abmessungen Breite=308 & Höhe=105.

```
<a:Page HEIGHT="4805" WIDTH="3602" PHYSICAL_IMG_NR="0" QUALITY="OK" PROCESSING="//">
  <a:PrintSpace HEIGHT="3751.9" WIDTH="3286.07" HPOS="105.31" VPOS="210.62">
     <a:TextBlock HEIGHT="2002" WIDTH="2202" HPOS="1113" VPOS="333">
        <a:TextLine HEIGHT="108.0" WIDTH="2162.0" HPOS="0.0" VPOS="7.0">
           <a:String HEIGHT="87.0" WIDTH="242.0" VPOS="28.0" HPOS="0.0" CONTENT="lsocrat" />
```

Abbildung 13: Ausschnitt des von B.I.T.-Alpha generierten ALTO-XML (Positionen)

Diese und weitere Schwächen in der Ausgabe von ALTO-XML sind zwar während der Prozessierung auffällig geworden, sie sind aber technisch schnell zu beheben. Es wird darauf hingewiesen, dass die Ausgabe der Meta-Daten in den Ergebnissen auch zu überprüfen ist, da auch die Angaben des Dateinamens in den ALTO-Dokumenten nicht korrekt sind:

```
<a:fileName>[redacted]\li-sbd-115-2s\00001.tif</a:fileName>
```

Abbildung 14: Ausschnitt des von B.I.T.-Alpha generierten ALTO-XML (Dateiname)

Bei dem oben genannten Digitalisat müsste es sich um die Datei 'li-sbd-115-2s\00012.tif' handeln. Ob jeder Ordner demnach im Batch-Prozess von BIT-Alpha intern gebündelt verarbeitet wird und das erste Bild – geordnet nach Dateiname – für alle weiteren ALTO-XML-Dokumente des Ordners genutzt wird, bleibt eine Vermutung, da verständlicherweise im Projektzeitraum keine Einsicht in die Quelltexte der Software gewährt wurde. Die Problematik, dass sehr oft in einer Stapelverarbeitung auf Daten zurückgegriffen wird, die zu Beginn des Stapels als Schablone abgeleitet wurden, stellt einen beliebten allerdings zu überprüfenden Ansatz in OCR-Projekten dar. Denn oft ist dies aufgrund von Werk-intrinsischen Merkmalen ausgeschlossen. Ein bemerkenswerter Zusatz, den BIT-Alpha leistet, ist die Sicherung von Konfigurationsdaten im Kopf des ALTO-Dokumentes. Ocropus und Tesseract erzeugen kein ALTO und ABBYY gibt zwar korrekte Daten aus, allerdings sind diese ohne Angabe von Konfigurationsdaten der OCR-Prozessierung.

Ein weiterer Ansatz, der sich als durchaus vielversprechend anbietet, ist die Erweiterung der Algorithmen zur Differenzanalyse. Bisher wurde ein erzeugter Volltext in eine lange Zeichenkette transformiert, um danach eine Metrik zur Anzahl von Operationen anzugeben. Die ist notwendig, um eine Zeichenkette in eine andere zu überführen. Eine zweistufige Analyse der Daten würde präzisere Rückschlüsse zulassen. Da dann nicht mehr ausschließlich auf Zeichenebene geprüft würde, was den meisten aktuell bestehenden Ansätzen im Forschungsfeld um OCR und Handwriting recognition (HWR) wiederum entspräche. Die Erweiterung des Algorithmus zur Differenzanalyse kann weiterhin auf Basis der Levenshtein-Distanz stattfinden. Allerdings sollte neben der Zeichenebene die Wortebene miteinbezogen werden. Würde man nur auf der Wortebene bei einem Missmatch (ungleiches Wort) die einzelnen Zeichen betrachten, wäre zwar mit höherer Latenz zu rechnen, aber die Aussagekraft der Textvergleiche wäre profunder. Auch die Anzahl der erkannten Worte, ohne die Anzahl der erkannten Zeichen, ist als wichtige Ergänzung zu betrachten. Es wird hier auf die Prägnanz im Bereich der automatischen Qualitätssicherung hingewiesen. Weil bereits bei der Wahl von Fenstergrößen von Vergleichsalgorithmen, im Umfeld von Wort-Token noch wichtige Erkenntnisse ausstehen.

Ein weiterer Lösungsansatz, der die Identifikation von Wörtern im Text leistet, ist daher die Konfiguration eines entsprechenden Tokenizers, der Wörter klassifiziert und korrekt selektiert. Punkte, Kommas, oder andere Sonderzeichen in Zeichenketten auszuschließen ist nicht trivial im Hinblick auf Zierdruckzeichen in den vorliegenden Korpora. Eine Näherung dazu kommt oft in Form von minimalen Wortlängen daher, was allerdings nicht ausreichend ist.

Ein weiterer Ansatz beläuft sich auf das angepasste Training von Modellen, die dem Sprachgebrauch der Texte im Wesentlichen entgegenkommen. Fehlende Wörterbücher machen dieses Unterfangen allerdings zu einer Vorgehensweise, die im vorliegenden Projektrahmen nicht geleistet werden konnte. Gleichwohl werden solide Modelle zur Erkennung von unterschiedlichen Zeichensätzen, hoher Wiederverwendbarkeit und unterschiedlichen Sprachen als sehr viel versprechend eingestuft. Die Umsetzbarkeit ist durch nicht existierende Grundlagen aber massiv erschwert. Die Komplexität derartiger Modelle würde keine der vorliegenden Software-Implementation auf heute verfügbarer Hardware nutzen können.

Die Untersuchung mit Wörterbüchern einer bestimmten Sprache in einem Digitalisat lässt auf fehlerhafte Verwendung von Trainingsmodellen zurückschließen. Besonders Vorkommen von 'the' oder 'each' in den erzeugten Volltexten mit lateinischer Herkunft lassen darauf schließen. Hier ist besonders der Fall von Tesseract zu bemerken. Es sollten daher zukünftige Ergebnisse auch im Rahmen von Plausibilität oder Unschärfe überprüft werden.

3.3 Fallstudie 3 – Korpus 3

Das dritte Korpus ist eine alternative Herangehensweise an die Frage: Wie kann man die Leistungsfähigkeit/Güte/Präzision einer OCR-Software überprüfen? Aus diesem Grund wurden aus dem Bestand der Helmstedter Drucke ein Korpus gebildet, das auf der Basis bekannter Fehler aufsetzt. Die Ansätze zu den Fehlern wurden dabei zum einen aus der Literatur aktueller Forschungsbeiträge gesammelt und zum anderen aus Erfahrungen von Mitarbeitern aus Digitalisierungsprojekten abgeleitet.

Die Fehlerquellen aus Digitalisierungsprojekten können sehr unterschiedliche Ursachen haben. Zunächst wird die Unterscheidung gemacht, ob eine Fehlerquelle aufgrund der intrinsischen Merkmale entstanden ist oder aufgrund von Ursachen, die durch Bearbeitung, Lagerung oder anderer nachträglicher Effekte ins Werk gelangt ist. Es wird daher vorgeschlagen, dass Fehlerquellen (im Folgenden Problem-Features genannt) in intrinsischen und extrinsischen Problem-Features unterschieden werden.

Im dritten Korpus sind daher manuell Seiten identifiziert worden, die mindestens ein Problem Feature beinhalten. Die folgende Liste zeigt die am häufigsten auftauchenden Problemfeatures.

Intrinsisch:

- Sprachmix
- Fontmix
- Fehlende Zeichensätze
- Ligaturen
- Gedrängtes Druckbild
- Schriftgrößenmix
- Abweichende Schriftregeln (u/v; I/J; zn/zu; nnd/und; uu/un/nn; nm/um/üm; cn/en/eu; ö/s;)
- Umlaute/Sonderzeichen (langes s etc.)
- Annotationen
- Satzspiegel

Extrinsisch:

- Geringe Druckqualität
- Geringe Qualität des Digitalisats (z.B. geringe Kontrastverhältnisse)
- Papierverfärbung (z.B. Flecken, Farbverläufe)
- Verschmutzung
- Schwankende Kontrastverhältnisse (Buchanfang → Buchmitte)
- Rechte oder linke Buchseiten
- Ränder
- Farbskalen/Copyright Hinweise
- Fehler aus Verarbeitungsorganisation

Extrinsische Probleme, die aufgrund von Software entstehen:

- Binarisierungsprobleme
- Verzerrungskorrekturen
- Segmentierungsschwierigkeiten
- GUI/Nachvollziehbarkeit/Usability
- Modularisierung des gesamten Prozesses/Vollautomatik
- Erkennung von Texten als Bild und umgekehrt
- Nachträgliche Fehlerkorrektur (z.B. per Lexikon oder linguistischen Methoden)
- Fehlende Ausgabeformate
- Fehlende Zeichensätze
- Inkompatible Trainingsmodelle
- Zonengrenzen falsch gesetzt
- Undifferenzierte Textblöcke
- Fehlende Wörterbücher

Diese Liste lässt sich erweitern, aber für die Untersuchung, der am häufigsten vorkommenden Problem-Features, wird hier eine Auswahl aus der oben genannten Liste entnommen und mit dem dritten Korpus in den Kontext der Machbarkeit von OCR-Möglichkeiten gestellt. Weiterführende Literatur findet sich in dem "Volltext via OCR"-Text (Federbusch, Polzin, and Stäcker 2013) aus Anwendersicht und zusätzlich in einer Dissertation (Strohmaier 2005), geschrieben an der Universität München.

Die Liste der Auswahl von angewandten Problemfeatures wird im Folgenden dargestellt. Um die Problemfeatures in den jeweiligen Ergebnissen zu überprüfen, sind sehr große HTML-Übersichten entstanden. In denen man mit einem Schlüssel auf die genannten Probleme zurückschließen kann. Die Übersichten sind so angelegt, dass neben jeder Seite die Auswahl getroffen werden kann, ob das OCR-Ergebnis das vorliegende Problem-Feature positiv erkannt hat (Test bestanden).

Problemfeature	Schlüssel
Fontmix (Absatzebene)	0001
Fontmix (Wortebene)	0002
Fontmix (Zeichenebene)	0003
Schriftgrößenmix	0004
Sprachmix	0005
Langes s/Zeichen	0006

komplexes Layout	0007
Abbildung	0008
Fußnoten/Marginalien	0009
Tintendurchdruck	0010
Sonderzeichen	0011
Verzerrung	0012
Farbsprung	0013
Ligaturen	0014
Zierbuchstaben	0015
Fettschrift/Normschrift-Wechsel	0016
Flecken	0017
Variierende Kontrastverhältnisse	0018
Knicke/Falten	0019
Gespiegelte Buchseite	0020
Zahlen	0021

Um die oben genannten Problemfeatures effizient zu visualisieren und evaluieren, wurde für die Aufbereitung im Rahmen dieser Studie eine Segmentierung der Präsentationsebene erstellt. Die Groundtruth-ähnlichen Referenztexte, die Bilddaten der dazugehörigen Seiten und die OCR Ergebnisse sind auf einer Webseite zusammengefasst dargestellt.[21]

3.3.1 Resultate und Ansätze aus den Erkenntnissen um das dritte Korpus

Bei dem Überblick auf alle bisher erarbeiteten Ergebnisse, könnte man versuchen die beste Software zu finden oder das beste Dateiformat zu bestimmen. Im Hinblick auf Datenmenge, Erkennungsqualität, Präzision oder Effizienz sind weitere Studien zu diesen Themen als Erkenntniszugewinn zu erwarten. Der Fokus liegt hier – wie zu Beginn erwähnt – auf der Untersuchung bereits bestehender Volltexte in öffentlich zugänglichen Korpora. Weiterhin ist es ein Ziel dabei, die Schwierigkeiten von Software und die Problematiken von Datenquellen zu beleuchten. Dabei sind bereits Einbrüche in der chronologischen Abfolge aufgezeigt worden, sowie die Distribution von Drucken ebenfalls untersucht wurde, was erst mit einer erneuten Erweiterung des Testsets möglich geworden ist. Diese Erweiterung beinhaltete weitere Druckorte und mehr Erscheinungsjahre.

Zudem ist es unbedingt notwendig aus der relativen Bewertung herauszukommen und daher wird dringend vorgeschlagen einen Service zur Überprüfung von Groundtruth-ähnlichen Texten öffentlich zugänglich zu machen, um Texte, die in vermeintlich sehr hoher Qualität vorliegen, zu verifizieren.

21 http://dariah.hki.uni-koeln.de

Unabhängig davon ist mit dieser Fallstudie auch die Perspektive auf die Arbeit mit Korpora gelegt worden und es wurden Kandidaten ermittelt, die mit den postulierten Problem-Features umgehen. Als weitere Tendenz soll aufgezeigt werden, ob verschiedene Trainingsmodelle im Kontext von Problemfeatures besser arbeiten als andere. Die folgende Grafik zeigt deutlich, dass der ABBYY-Recognition Server mit Abstand die besten Ergebnisse erzielt:

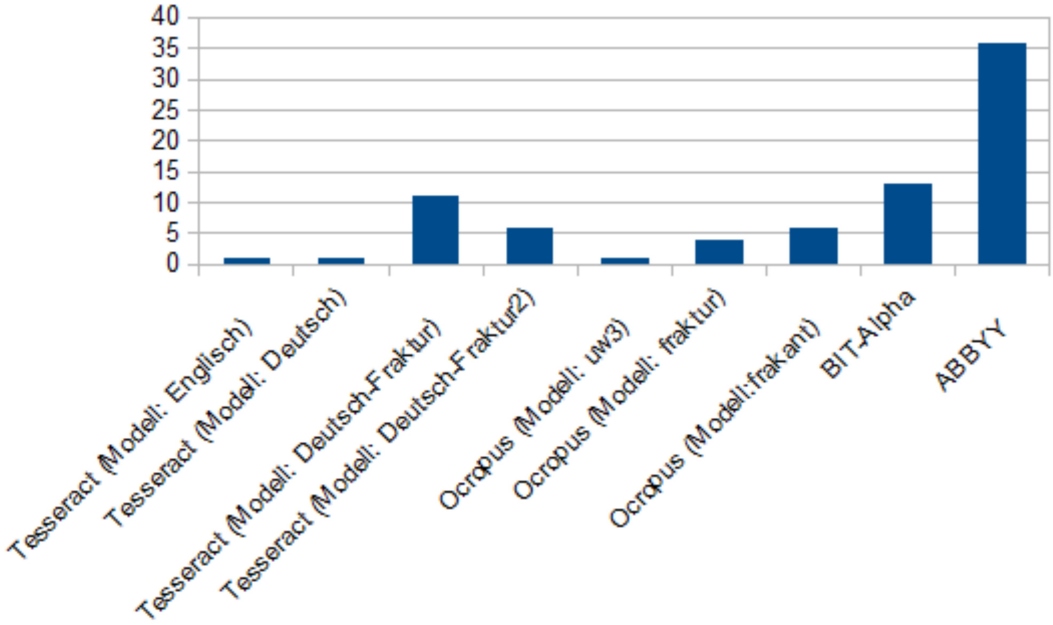


Abbildung 15: Gesamtübersicht der Erkennungsergebnisse des dritten Korpus.

In der Evaluation ist dabei der folgenden Logik gefolgt worden: Erkennt eine Software-Implementation im Ergebnis die Mehrzahl der Tests aus den Problemfeatures, wird dieses als positiv bewertet.

ABBYY führt an dieser Stelle die Punkteskala klar an. Zusätzlich wird deutlich, dass alle weiteren OCR-Implementationen keine großen Differenzen aufweisen. BIT-Alpha überragt Tesseract und Ocropus zwar, steht aber in großer Nähe zum Ergebnis von Tesseract mit einem Fraktur-Modell für die deutsche Sprache. Daher kann festgehalten werden, dass die Fraktur-Modelle in Open-Source Implementationen überaus empfehlenswert zu nutzen sind. Auch im ABBYY Recognition-Server ist das Modul 'Gothic' zur Erkennung von Frakturschrift in dieser Studie angewandt worden. Es lassen sich in ABBYY mehrere Lexika in einem Konfigurationsset verbinden, was in keiner anderen Software ohne weiteres möglich ist.

Unterscheidet man die Ergebnisse in intrinsische und extrinsische Problem-Features, werden die Graphen entropisch schwächer. Deutlich wird aber trotzdem, dass die OCR-Implementation ABBYY in vielen Belangen weiterentwickelt ist (wie zum Beispiel in der Verarbeitung von Problemfeatures zuvor deutlich wurde).

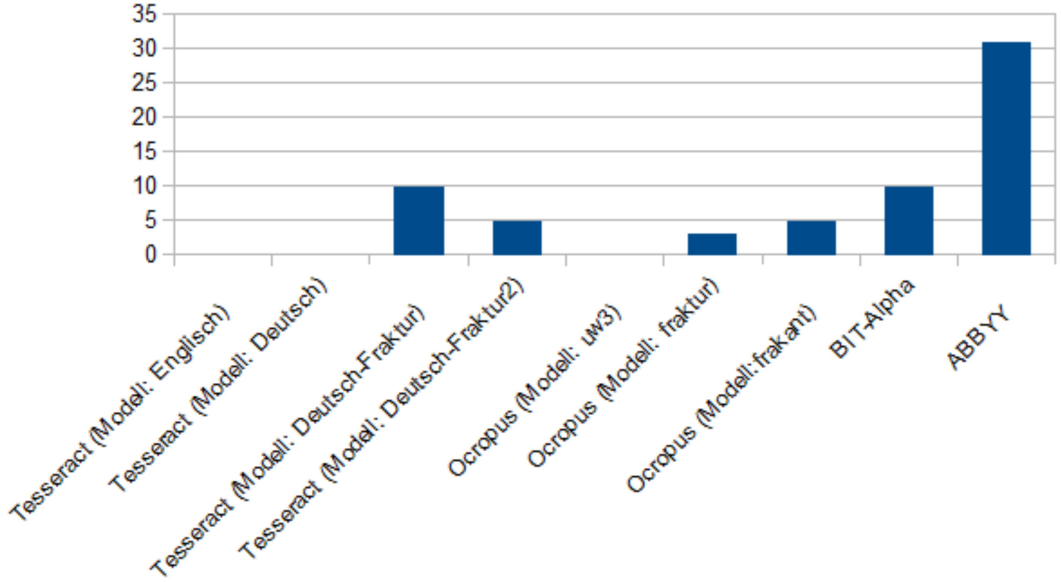


Abbildung 16: Intrinsische Problem-Features in der Ergebnis-Übersicht.

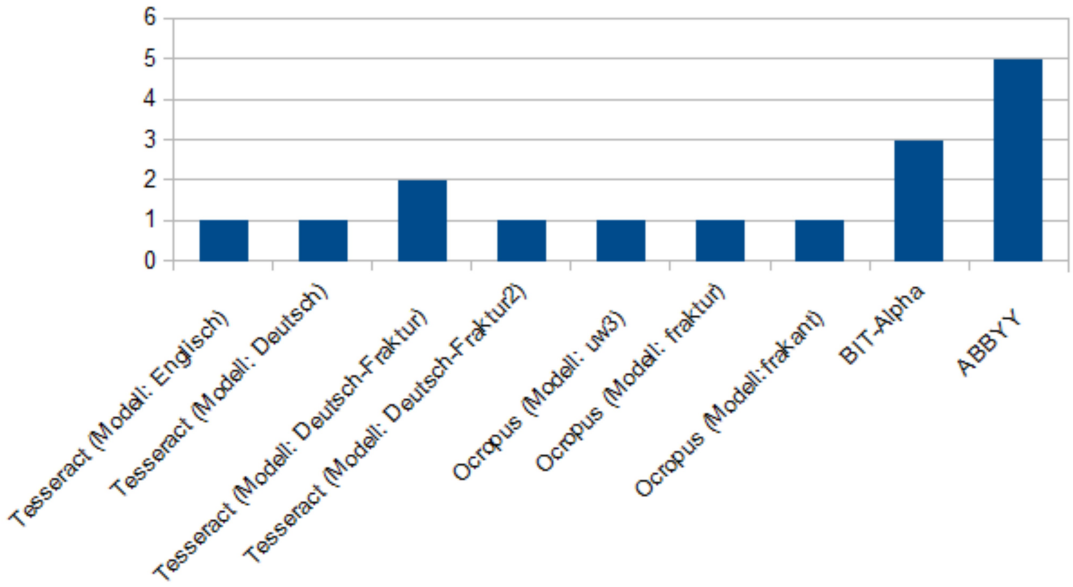


Abbildung 17: Extrinsische Problem-Features in der Ergebnis-Übersicht.

3.4 Fallstudie 4 – Resultate in Bezug auf Qualitätsunterschiede

Zur Überprüfung der Leistungsfähigkeit einer OCR-Software bei niedrig aufgelösten Bilddaten, wurde das dritte und erste Korpus in jeweils zwei Qualitäten erneut in allen vier OCR-Implementationen überprüft. Dabei war nicht entscheidend wie schnell oder welche Datenmengen verarbeitet wurden. Im Fokus steht die Analyse bzw. die Kontextualisierung, inwiefern eine Software mit geringerer Qualität als der vom DFG vorgegebenen 300 dpi zurechtkommt (DFG 2013). Zugleich wird die Frage beantwortet, ob es überhaupt Sinn macht neben der geringen online verfügbaren JPEG-Qualität auch bessere Qualitäten zu wissenschaftlichen Nutzung in Forschungsinfrastrukturen bereitzustellen oder ob bei geringeren Qualitäten mit ähnlichen Ergebnissen zu rechnen ist. Die folgenden Diagramme verdeutlichen dabei die Korrelation zwischen JPEG/TIFF-Qualitäten in allen vier OCR-Implementationen.

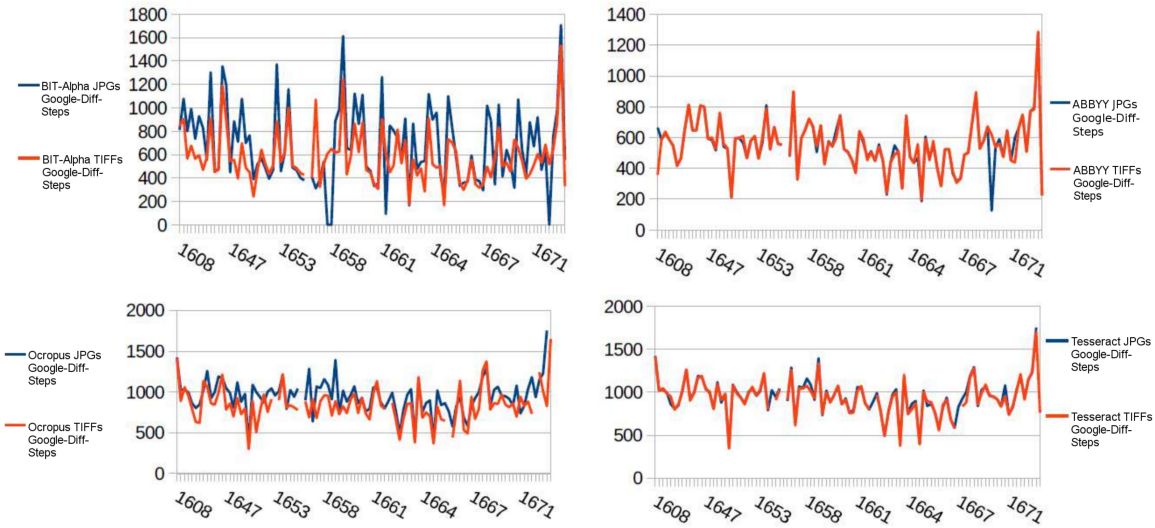


Abbildung 18: Gesamtübersicht der OCR-Ergebnisse von JPEG- und TIFF-Vergleichen.

Die daraus ableitbaren Erkenntnisse sind recht deutlich, betrachtet man die Ergebnisse von Tesseract und ABBYY's Recognition Server. Eine geringe Abweichung zwischen den Ergebnissen von TIFF-Daten zu JPEG-Daten lässt darauf schließen, dass sowohl Tesseract als auch ABBYY fast gleich gut mit TIFF oder JPEG zurechtkommen. Die Distanzen unterscheiden sich kaum in Bezug auf die Groundtruth-ähnlichen Daten. Bei stark optimierter Erkennungsqualität sind die Unterschiede zwischen verlustfreien TIFF-Dateien und JPEGs ähnlich gering. Die einzige Software, die an dieser Stelle deutlich zeigt, dass sie Bilddaten in TIFF benötigt, ist BIT-Alpha, denn auch Ocropus scheint kaum Probleme mit verminderter Qualität zu haben. Eine Prüfung der Güte eines Textes auf statistischen Näherungswerten ist daher notwendig und plausibel. Schaut man sich die folgende Gegenüberstellung an, wird klar, dass eine OCR-Implementation mit

Problem-Features in hohen Qualitäten scheinbar schlechter zurecht kommt als bei geringeren JPEG-Qualitäten. Es drängt sich der Verdacht auf, dass Schwierigkeiten, wie z.B. Flecke oder Tintenfraß in perfekter Qualität ein ähnliches Phänomen aufwerfen wie beim sogenannten „Übertrainieren" von Modellen.

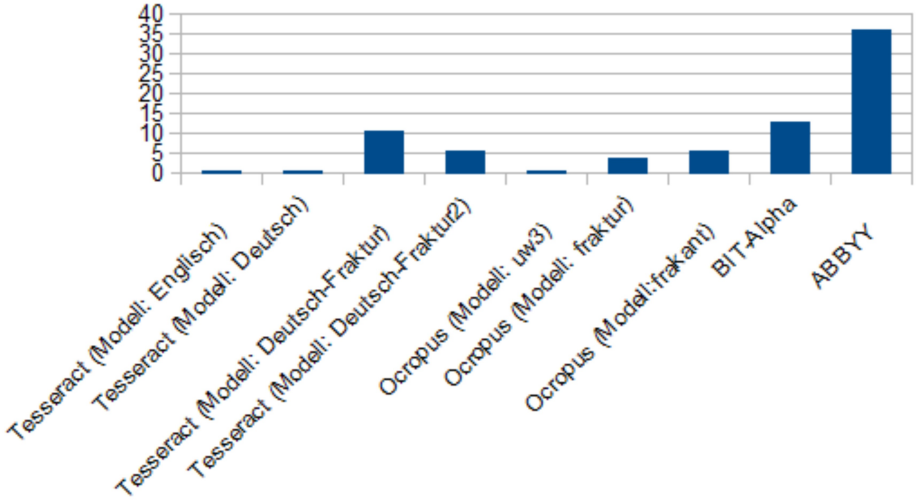


Abbildung 19: JPEG-Qualitäten und die intrinsischen/extrinsischen Ergebnisse der OCR-Verfahren.

Betrachtet man nur die Y-Achsen und deren Wertebereiche, wird deutlich, dass in der JPEG-Evaluation jede einzelne OCR-Implementation besser abgeschnitten hat, als die dazugehörige TIFF-Evaluation. Es wird die Tendenz deutlich, dass es ausreichend ist, JPEG-Daten für die Öffentlichkeit bereitzustellen. Als übliche Qualität werden meist TIFF-Daten und selten JPEG-Daten mit einer Qualität von 60% online zur Verfügung gestellt. Auf Gründe für derartige Ergebnisse und differenzierbare Falluntersuchungen bezogen, wird daher spezifischer im nächsten Abschnitt eingegangen.

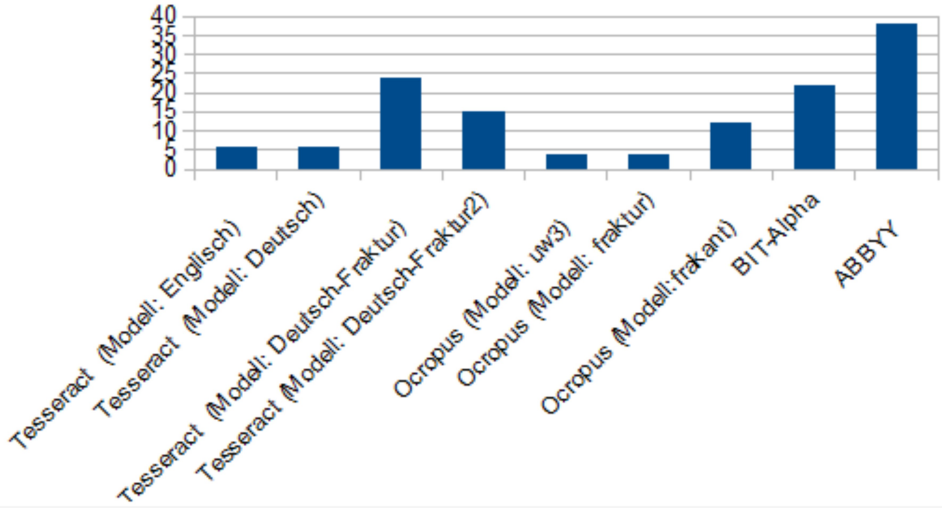


Abbildung 20: TIFF-Qualitäten und die intrinsischen/extrinsischen Ergebnisse der OCR-Verfahren.

Jede OCR-Software stellt ohne Ausnahme eine Reihe von algorithmischen Methoden bereit, die mit Bilddaten auf der Ebene von Pixeln arbeitet. Besonders häufig werden algorithmische Lösungen von der wissenschaftlichen Disziplin der Computergrafik bereitgestellt, die im Fall von Binarisierung das ganze Bild verarbeiten oder auch nur Teile filtern, wenn es um die Segmentierung oder die eigentliche Volltexterkennung geht. Daher wird im Folgenden in verschiedenen Konfektionierungen unterschieden und zugleich überprüft, wie sich Algorithmen zur Binarisierung in Bezug auf die Unterscheidung von JPEG-Formaten und TIFF-Formaten verhalten.

Es wird ein ähnlicher Ansatz wie von Schlarb verfolgt, bei denen die Präzision der Volltexterkennung dazu genutzt wird, Ausreißer in Volltexten zu identifizieren (Schlarb and Neudecker 2012; Buckley 2013).

Im Anschluss an die Prozessierung mit den OCR-Implementationen wurden Distanzmaße gebildet und verglichen. Der grundlegende Unterschied bezieht sich auf die Abwesenheit von Groundtruth, denn in dieser Studie wurden bereits bestehende Ergebnisse als Groundtruth-ähnlich miteinbezogen und ein Zielformat für öffentliche Einrichtungen soll als machbar erörtert werden. Das JPEG-Format bietet hier weitaus mehr Potential im Hinblick auf die Kostenersparnis. Aus diesem Grund wird als sinnvoll erachtet, dass verschiedene Erkennungsergebnisse und deren binarisierte Entsprechungen betrachtet werden. Einerseits wird der Bezug zwischen Ergebnissen der Binarisierung von JPEG-/TIFF-Quellen und den davon erzeugten Volltexten dargestellt, andererseits ist der Bezug von JPEG-/TIFF-Quellen und der Ergebnisse der algorithmischen Zeichenerkennung selbst sichtbar.

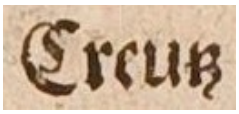
Die zur Binarisierung in einer OCR-Implementation genutzten Binarisierungs-Algorithmen hängen stark von den Parametern ab, die eingesetzt werden. Im Folgenden sind daher die Ergebnisse aus der Nutzung der Software mit Standardeinstellungen nach Installation dargestellt, so wie die entsprechende OCR-Implementation die verarbeiteten Bilder standardisiert oder initial prozessieren würde.

3.4.1 Fallstudie 4 – Vergleiche auf Wortebene

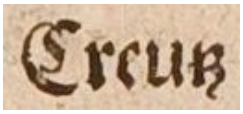
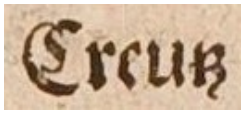
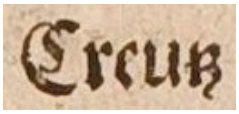
Es ist Konsens, dass Binarisierung der erste durchzuführende Schritt bei der Nutzung von Bild-Algorithmen ist. Alle weiteren Algorithmen nutzen dann die entstandenen Monochrombilder. Die Korrektur von brüchigen, fransigen Kanten (Ausbluten von Tinte) oder Verzerrungen (Skew-/Slantcorrection) findet üblicherweise auf Basis derartiger Bilder statt. Flecken, Farbschwellen und weitere Artefakte, die unter dem Sammelbegriff ‚Störungen' summiert werden, verwenden Algorithmen, die wesentlich komplexer und zugleich anfälliger für geringfügige Störungen sind, als wenn keine Binarisierung durchgeführt worden wäre. Der hohe Grad an Stilisierung der Bilddaten, der geschieht bevor die eigentliche Zeichenerkennung stattfindet, steht demnach im Zusammenhang mit den genutzten Algorithmen und wie bis dato angenommen in engem Zusammenhang mit der Qualität/Kompression eines Bildes. Folgt man der Tendenz von Zeichenerkennung im Bereich historischer Dokumente, geht die Entwicklung vor allem hin zu skalierungs- und rotationsinvarianten Algorithmen. Diese sind wesentlich robuster gegenüber den genannten Störungen.

In der folgenden Tabelle werden daher einzeln die Ergebnisse der OCR-Implementationen und deren Binarisierungsergebnisse im JPEG-/TIFF-Vergleich gegenüber gestellt.

ABBYY-Recognition-Server 3.5:

Originalbild TIFF	Binary TIFF	Original JPEG	Binary JPEG
 GT: „Creuz“	 „CreUtz“		 „Crcuy“

Ocropus 0.7:

Originalbild TIFF	Binary TIFF	Original JPEG	Binary JPEG
 GT: „Creuz“	 Non-linear (nlbin): „Treuy“ (fraktur) „Ereuz“ (frakant) „vErans“ (uw3)		 Non-linear (nlbin): „Trcug“ (fraktur) „Ereuz“ (frakant) „CCr0“ (uw3)
 GT: „Creuz“	 Sauvola: „Treus“ (fraktur) „Ereus“ (frakant) „vraie“ (uw3)		 Sauvola: „Creus“ (fraktur) „Creus“ (frakant) „Cre“ (uw3)

In dieser Gegenüberstellung wird deutlich, dass die Algorithmen zum Binarisieren weder mit JPEG, noch mit TIFF Bildern größere Schwierigkeiten haben. Meist gehen Algorithmen pixelweise mit einer Fensterfunktion und mehreren Threshold-Parametern über die gesamte Bildfläche und entscheiden dann, ob ein Pixel schwarz oder weiß im resultierenden Monochrom-Bild wird. Das funktioniert bei sehr großen TIFF-Formaten genauso gut wie bei JPEG-Bildern, die einen Bruchteil der Speichermenge benötigen. Anhand der wesentlich höheren Anzahl von schwarzen Pixeln in einem hochauflösenden TIFF ist auch die ‚dichtere/fette' Strichstärke zu erklären.

Inwiefern alle darauf aufbauenden Algorithmen nun auf die Ergebnisse der Binarisierung Bezug nehmen, ist an den Ergebnissen der Zeichenerkennung zu betrachten. Da bei der Untersuchung auf ein einziges Wort, kaum noch Segmentierungsalgorithmen zum Einsatz kommen.

Die Volltexte der einzelnen Kompressionsstufen einer OCR-Implementation sind ein deutliches Indiz für die Äquivalenz der angedeuteten qualitativen Ergebnisse. Das abgebildete Wort „Creuz" stammt aus ABBYY und ist mit einer Levenshtein Distanz von 2 im JPEG-Format erkannt worden und mit derselben Distanz von 2 im TIFF Derivat. Deutlicher wird es bei der Betrachtung der OCR-Implementation Ocropus, die als ‚state-of-the-art' Software in dieser Studie miteinbezogen wurde. Hier sind die Ergebnisse von TIFF-Daten auf den angelegten Modellen ‚fraktur' und ‚frakant' mit nicht-linearer Binarisierung (nlbin) geringfügig besser als im Fall von JPEG komprimierten Quellenmaterials. Im Fall der wesentlich häufiger angewandten und effizienteren Methode des Sauvola-Algorithmus sind die TIFF-Ergebnisse nur minimal schlechter oder fast äquivalent im Vergleich zu den genutzten JPEG-Daten.

Es wird also deutlich, dass sobald keine Segmentierungsalgorithmen zur Trennung von Wörtern angewandt werden, die Ergebnisse nur minimal schlechter auf Seiten der JPEG-Qualitäten sind. Auffällig ist auch, dass die experimentelle Implementation Ocropus unter den passenden Voraussetzungen in diesem Fall gleichwertige Ergebnisse wie kommerzielle Produkte abliefert.

Eine weiterverfolgte Hypothese lautet demnach: Moderne Methoden und Ansätze, wie sie in OCR-Software zur Anwendung kommen, liefern bei gleicher Aufgabestellungen äquivalente Ergebnisse, sofern die gewählten Zeichenmodelle konform sind.

Der ABBY Recognition Server 3.5 wurde anhand des ABBYY Finereader (XIX) entwickelt und befindet sich demnach seit 1993[22] in der Entwicklung, die OCR-Implementation Ocropus hingegen befindet sich aktuell in der Version 0.7 und wird seit 2007[23] entwickelt. Der Ansatz von Ocropus weist im Gegensatz zu dem von ABBYY deutliche architektonische Unterschiede auf. Einige wichtige – vor kurzem erreichte – Meilensteine in der Entwicklung um OCR-Software sind Lösungen, die robuste Algorithmen zum Dokumentenverständnis bereitstellen. Ein Überblick wird dazu in der Doktorarbeit von Konya gegeben (Konya 2013). Mit dem Fokus auf die Segmentierung von Wörtern und Zeilen, die essentieller Bestandteil von solider OCR-Software sind, wird auf das Thema grundlegend eingegangen.

Ein anderes weitreichendes Problem wird in allen OCR-Implementationen deutlich: Immer dann, wenn Marginalien teilweise zum Fließtext hinzugefügt werden oder ans Textende gesetzt werden. Die zuvor angeführte Dissertation setzt somit für diese Studien eine Marke in der Entwicklung um Segmentierungsalgorithmen. Allerdings wird kein Vorgehen vorgelegt mit Segmenten in der Retrodigitalisierung umzugehen, um sie sinnvoll aufzubereiten und weiterzuverwenden. Es bleibt daher offen inwiefern mit derartigen Unregelmäßigkeiten umgegangen werden kann.

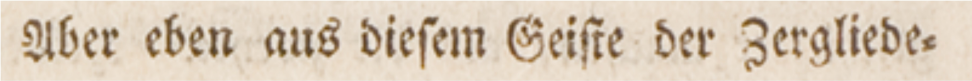
22 http://www.abbyy.de/unternehmen/geschichte

23 https://web.archive.org/web/20070414130618/; http://googlecode.blogspot.de/2007/04/announcing-ocropus-open-source-ocr.html

3.4.2 Fallstudie 4 – Vergleiche auf Zeilenebene

Neben der vorangegangenen Überprüfung auf Wortebene, wird nun die Binarisierung als Grundlage von Segmentierungsalgorithmen angeführt. Dabei ist zum einen das Ergebnis der Binarisierung (der Volltext) interessant und zum anderen ist auch zu prüfen ob korrelierende Ergebnisse zwischen dem monochromem Bild bzw. der Präzision des Volltexts und der Qualität des Bildes vorliegen.

ABBYY-Recognition-Server 3.5:

GT	Aber eben aus diesem Geiste der Zergliede-	
TIFF		
Binary TIFF		
Volltext	Aber eben aus diesem Geiste der Zergliede-	lvd1
JPEG		
Binary JPEG		
Volltext	Aber eben ans diesem Geiste der Zergliede-	lvd2

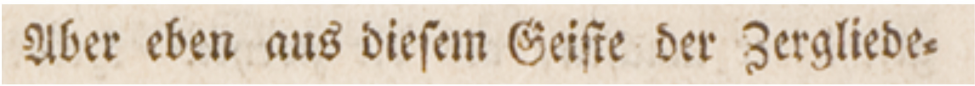
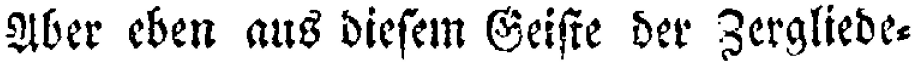
Der ABBYY-Recognition Server berechnet aus dem JPEG sowie aus der TIFF-Version des Eingabematerials fast identische Ergebnisse, obwohl beim JPEG-Format die üblichen Kompressions-Effekte (z.B. Quantisierung) bereits sehr deutlich auftreten. Am Ergebnis der Volltexte lässt sich ablesen, dass die Präzision der Segmentierungs- und Erkennungsalgorithmen sehr gut in beiden Qualitäten funktioniert hat. Nur in einem einzigen Buchstaben unterscheiden sich die Ergebnisse, wobei hier eine 158 Kilobyte (KB) TIFF-Datei gegenüber einer 23 KB JPEG-Datei steht. Es wird also deutlich, dass die zuvor formulierte These, bereits unter schlichten Faktoren, wie der Hinzunahme von Dateigrößen, gestützt wird. Zusätzlich wird deutlich, dass die JPEG-komprimierten Daten (mehr als sieben Mal kleiner) mit einer sehr guten Präzision in diesem Test aufwarten können.

In der OCR-Implementation Ocropus wird diese Tendenz weiter verfolgt und zusätzlich in den Kontext der genutzten Modelle gesetzt. Beide verfügbaren Binarisierungs-Algorithmen von Ocropus sind geprüft worden.

Nicht-lineare Binarisierung ist allerdings nicht mehr zeitgemäß und zudem sehr kostenintensiv. Effizientere Algorithmen, wie der von Sauvola entworfene, bilden dazu eine Alternative. Auch

der Niblack-Algorithmus liefert weitere Anschlussmöglichkeiten bei der Auswahl der passenden Binarisierung.

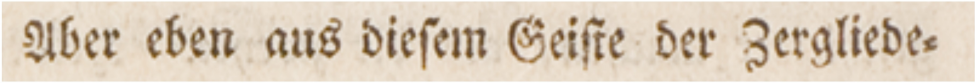
Ocropus 0.7 mit nicht-linearer Binarsierung:

GT	Aber eben aus diesem Geiste der Zergliede-
TIFF	
Binary TIFF	
Fraktur: Frakant: Uw3:	Aber eben aus diesem Geiste der Fergliede- Aber eben aus diesem Geiske der Zergliede- Slber cben aud bieian (ei(fe ber Reralibe
JPEG	
Binary JPEG	
Fraktur: Frakant: Uw3:	Aber eben aus diesem Geiste der Zergliede- Aber eben aus diesem Geike der Zergliede- 9lDer cben and bie(m Qei(e ber Reralibe

Mit nicht-linearer Binarisierung sind die Ergebnisse der Binary JPEG-Bilder und TIFF-Bilder erstellt worden. Im Vergleich sind sie zu den Ergebnissen von ABBYY fast identisch. Überprüft man die Verteilung von schwarzen und weißen Pixeln mit Histogramm-Verfahren, ergibt sich eine deutliche Korrelation. Auch die Ergebnisse der erkannten Volltexte sind mit nicht-linearer Binarisierung im Fall von JPEG-Kompression mit dem Fraktur-Modell von Ocropus nachzuprüfen. Die dazugehörige TIFF Variante enthält nur einen Fehler. Die beiden Frakant-Modelle sind nicht geeignet und werden aufgrund des vorliegenden Defizits nicht weiter einbezogen. Auch die Ergebnisse des uw3-Models, das mit Abstand die schlechtesten Ergebnisse geliefert hat, werden vernachlässigt. Das uw3-Modell ist im vorliegenden Kontext nicht geeignet.

Die Ergebnisse mit nicht-linearer Binarisierung lassen sich als robust und konkurrenzfähig zum ABBYY-Recognition Server bezeichnen. Gleichwohl wird die Vorverarbeitung (Preprocessing) mit einem effizienteren Algorithmus ‚Sauvola' in der folgenden Tabelle dargestellt.

Ocropus 0.7 mit Sauvola-Binarsierung:

GT	Aber eben aus diesem Geiste der Zergliede-
TIFF	
Binary TIFF	
Fraktur: Frakant: Uw3:	Aber eben ,aus diesem Geiste der Zergliede- Aber eben aus diesem Geiske der Zergliede- 8lfer ben au6 bie(em 8eiffe ber Rerglicbe
JPEG	
Binary JPEG	
Fraktur: Frakant: Uw3:	Aber eben aus diesem Geiste der Zergliede- Aber eben aus diesem Geiske der Zergliede- 8ifer cben. aud bie(m 8eiffe ber erglicbe

Man sieht deutlich die Unterschiede zu den vorangegangenen Ergebnissen. Kleinere Störungen, wie zwischen den Wörtern „eben" und „aus", sind deutlich erkennbar. Hervorzuheben dabei ist, dass die Störungen im TIFF-Bild leicht überwiegen, was beim Erzeugen des Volltexts über das Frakturmodell zu einem erkannten Komma geführt hat. Auch hier ist die Erkennung des Volltextes in der Kombination von JPEG-Kompression über Fraktur-Modell fehlerfrei verlaufen.

Vor dem Hintergrund, dass keine Konfiguration der Software auf das Material angepasst wurde (Ocropus Standardeinstellung für Sauvola ist sigma = 150 und k = 0.3), wird deutlich, dass im öffentlichen Zugang befindliches Material in JPEG-Kompression bisher die besten Ergebnisse geliefert hat. Weiter wird untersucht, ob die Entwicklung neuer robuster Algorithmen auf Basis von JPEG-Daten genauso effizient ist, wie mit TIFF-Daten. Darüber hinaus kann an dieser Stelle bereits davon ausgegangen werden, dass die üblichen Schwierigkeiten, die JPEG-Kompression mit sich bringt, kaum auftreten bei soliden Binarisierungsalgorithmen und identischer Auflösung. Die üblichen Artefakte, die bei Quantisierungsmechanismen von JPEG entstehen, werden gerade durch Binarisierung wieder kompensiert.

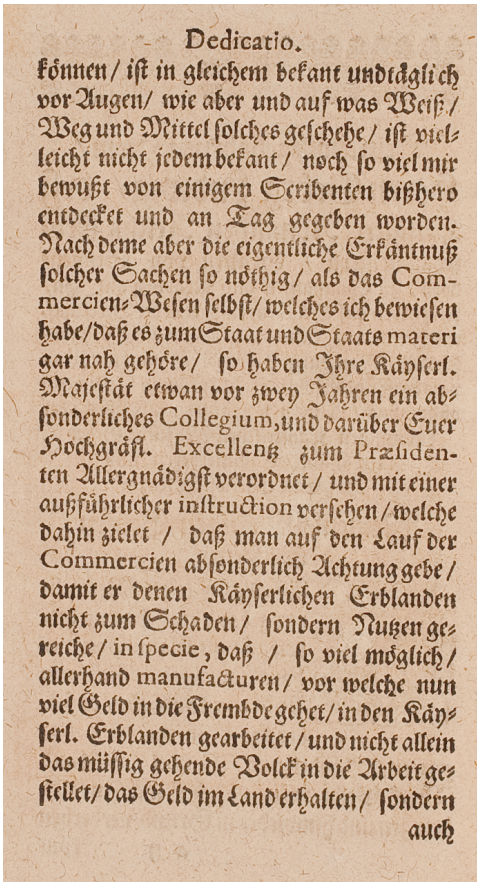
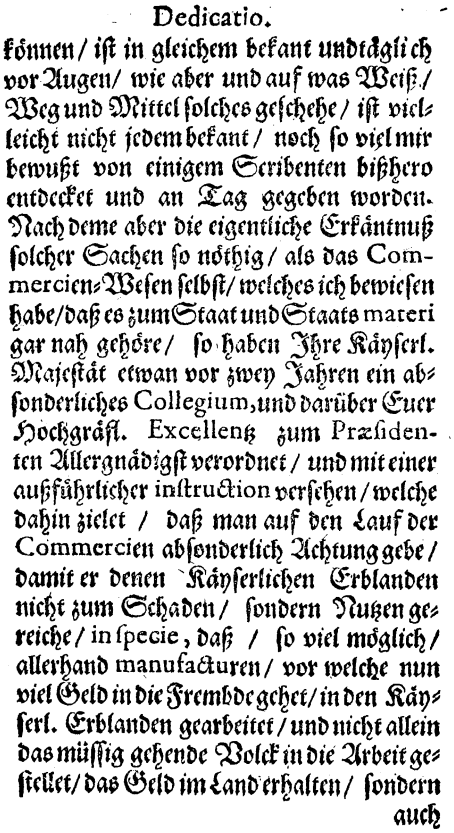
Die im vorigen Kapitel aufgestellte Trennung zwischen extrinsischen und intrinsischen Fehlerquellen, dient in der Diskussion um Binarisierung der Weiterentwicklung neuerer Algorithmen, die nicht mehr zur Kategorie Binarisierungsalgorithmen zählen, sondern zu reinen Segmentierungsalgorithmen.

Es ist wichtiger, Artefakte in einem Digitalisat zu erkennen als sie mit Binarisierungsalgorithmen in einem Schritt zu korrigieren. Insofern kann man von Machbarkeit im Sinne des Titels dieser Studie sprechen.

3.4.3 Fallstudie 4 – Vergleiche auf Seitenebene

Zuletzt wird eine Stufe weiter gegangen und neben der Wortsegmentierung in Zeilen auch die Segmentierung von Zeilen auf einer Seite in den OCR-Implementationen geprüft und in Kontext zu den Ergebnissen der Volltexterkennung gestellt. Als Maß wird erneut die Levenshtein-Distanz genutzt.

ABBYY-Recognition-Server 3.5:

Original TIFF:	Binarisierung TIFF:
	

Dedicatio.
können / ist in gleichem bekant und täglich
vor Augen / wie aber und auf was Weiß /
Weg und Mittel solches geschehe / ist viel-
leicht nicht jedem bekant / noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig / als das Com-
mercien-Wesen selbst / welches ich bewiesen
habe / daß es zum Staat und Staats materi
gar nah gehöre / so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium, und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet / und mit einer
außführlicher instruction versehen / welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe /
damit er denen Käyserlichen Erblanden
nicht zum Schaden / sondern Nutzen ge-
reiche / in specie, daß / so viel möglich /
allerhand manufacturen / vor welche nun
viel Geld in die Frembde gehet / in den Käy-
serl. Erblanden gearbeitet / und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet / das Geld im Land erhalten / sondern
auch

Groundtruth[24]:	Volltext-Ergebnis (TIFF):
Dedicatio.	Dedicatio*
kȯnnen/ iſt in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß/ Weg und Mittel ſolches geſchehe/ iſt vielleicht nicht jedem bekant/ noch ſo viel mir bewußt von einigem Seribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß ſolcher Sachen ſo nȯthig/ als das Commercien-Weſen ſelbſt/ welches ich bewieſen habe/ daß es zum Staat und Staats materi gar nah gehȯre/ ſo haben Jhre Käyſerl. Majeſtät etwan vor zwey Jahren ein abſonderliches Collegium, und darȯber Euer Hochgräfl. Excellentz zum Præſidenten Allergnädigſt verordnet/ und mit einer außfȯhrlicher inſtruction verſehen/ welche dahin zielet/ daß man auf den Lauf der Commercien abſonderlich Achtung gebe/ damit er denen Käyſerlichen Erblanden nicht zum Schaden/ ſondern Nutzen gereiche/ in ſpecie, daß/ ſo viel mȯglich/ allerhand manufacturen/ vor welche nun viel Geld in die Frembde gehet/ in den Käyſerl. Erblanden gearbeitet/ und nicht allein das müſſig gehende Volck in die Arbeit geſtellet/ das Geld im Land erhalten/ ſondern auch	können/ ist in gleichem bekattt Uttdkäglr'ch vor Augen/ wie aber und auf was Weiß/ Weg und Mittel solches geschehe / ist viel leicht nicht jedem bekam/ noch sovi/lmir bewußt von einigem Scribenkm bißhero entdecket und an Tag gegeben worden. Nachdeme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Comrnercien-Wesen selbst/ welches ich bewiesen habe/daß es ZumSkaak undStaaks materi gar nah gehöre/ so haben Ihre Kayscrl. Majestät ekwan vor zwey Jahren ein absonderliches Collegiummnb darüber Euer Hochgräsi. Excellens Zum Präsiden -im Allergnädigst verordnet / und mit einer ausführlicher initruction versehen/welche dahin zielet / daß man auf dm Lauf der Commcrcim absonderlich Achtung gebe/ damit er denen Kayserlichen Erblanden nicht zum Schaden/ sondern Nutzen gemche/infpecie,baß / so viel möglich/ allerhand manufa&umt/ vor welche nun viel Geld in die Frembdegchet/ in den Kayfcrl. Erblanden gearbeitet/und nicht allein dasmüssig gehende Volckindie Arbeit gestcllet/ das Geld im Landerhalten / sondern auch

Es soll aber zunächst unabhängig von der Betrachtung der Volltext-Ergebnisse auf Unterschiede zwischen dem Originalbild und dem Bild aus den Vorverarbeitungsschritten hingewiesen werden. Vor allem die Verzerrungs-Korrektur (Skew/Slant) wird bei ABBYY direkt in der Vorverarbeitung miteinbezogen. Es ist bekannt, dass die Vorverarbeitungs-Pipeline üblicherweise wie folgt aussieht:

Binarisierung → Skew/Slant-Korrektur → (Weichzeichnen) → Segmentierung → [Erkennung]

Da für moderne Binarisierungsalgorithmen, die auf der Grundlage von Fensterfunktionen arbeiten, die Auflösung ausschlaggebend ist, wird darauf hingewiesen, dass Interoperabilität von

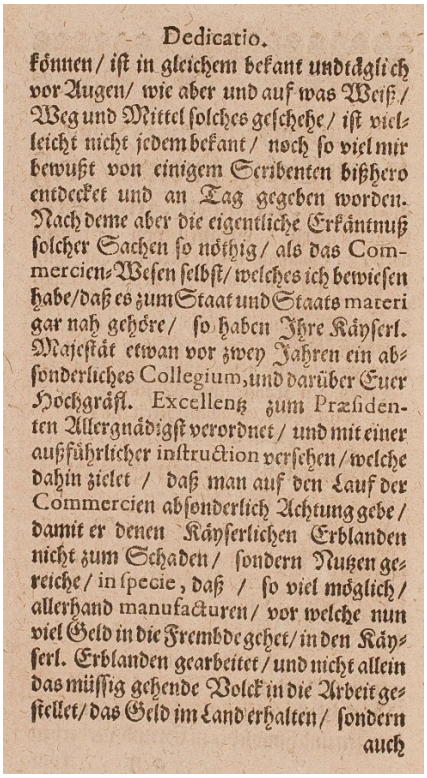
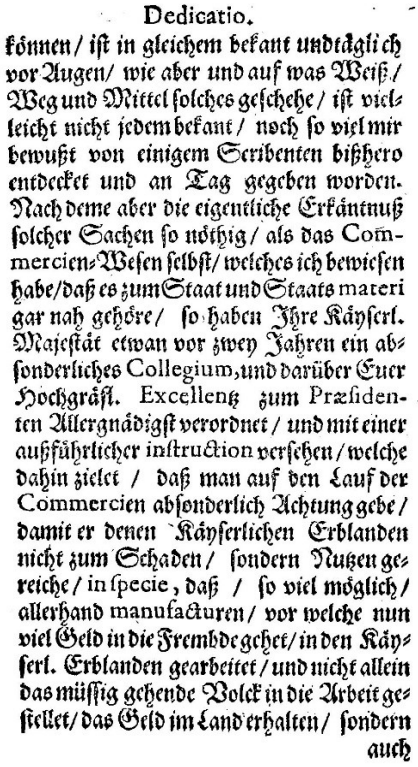
24 http://www.deutsches-textarchiv.de/book/view/becher_discurs_1668?p=8

entscheidender Bedeutung ist. Erst wenn die Effekte und Schwächen eines Algorithmus bekannt sind, lässt sich in anschließenden Schritten darauf eingehen. Eine Umgebung zur freien Verfügung interdisziplinär angesiedelter Wissenschaftsbereiche verspricht in dieser Hinsicht progressive Verbesserungen. Workflows und Konfigurationsmöglichkeiten persistent abzulegen sind ein Teil der Anforderungen einer solchen Umgebung. Die Arbeit mit Versionen, Editionen und Sammlungen der Volltexte ist ein wichtiger Optimierungszweig in der Arbeit mit OCR-Daten.

Unter Betrachtung des Volltextergebnis von ABBYY wird deutlich, dass der Großteil der entstandenen Fehler aufgrund von nicht korrekter Segmentierung oder fehlendem Wörterbuch Eintrag entstanden sind (Zeile: 10 [‚zum Staat und Staats']; 26 [‚das müssen']). Außerdem wird in diesem Vergleich erneut das Kompressionsverhältnis zwischen JPEG und TIFF-Daten deutlich. In der zweiten Zeile ist das letzte Wort ‚undtæglich' sehr prägnant. Ein kleiner schwarzer Fleck (Artefakt) führt dazu, dass aus dem Buchstaben ‚n' im TIFF Format ein ‚tt' wird und im JPEG-Format der korrekte Buchstabe ‚n'.

Die Gewichtung von Artefakten stellt demnach, wie in den vorigen Fällen gezeigt, ein Problem bei TIFF-Daten dar. Schnell wird ein Artefakt als Zeichen gedeutet, weil es bei der Binarisierung nicht entfernt wurde. Die Entwicklung neuer Algorithmen und Prozessketten wäre mit dieser Erkenntnis aber beschleunigt. Denn die Optimierung im Hinblick auf die Datenmenge stellt jeden Anbieter vor kleinere Latenz-Zeiten. Bei Cloud-/Crowd-gestützten Entwicklungsansätzen ist mit einen enormen Zuwachs an Iterationen zu rechnen. Kann der Algorithmen-Entwickler statt einmal mit TIFF-Material drei- bis siebenmal ein JPEG-Bild durch seinen Algorithmus prozessieren lassen, ist die Effizienz gesteigert.

ABBYY-Recognition-Server 3.5:

Original (JPEG):	Binarisierung (JPEG):
 Dedicatio. können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist vielleicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Commercien-Wesen selbst/ welches ich bewiesen habe/daß es zum Staat und Staats materi gar nah gehöre/ so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein absonderliches Collegium,und darüber Euer Hochgräfl. Excellentz zum Præsidenten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen gereiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käyserl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit gestellet/ das Geld im Land erhalten / sondern auch	 Dedicatio. können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist vielleicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Commercien-Wesen selbst/ welches ich bewiesen habe/daß es zum Staat und Staats materi gar nah gehöre/ so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein absonderliches Collegium,und darüber Euer Hochgräfl. Excellentz zum Præsidenten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen gereiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käyserl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit gestellet/ das Geld im Land erhalten / sondern auch
Groundtruth[25] Dedicatio. können/ iſt in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß/ Weg und Mittel ſolches geſchehe/ iſt vielleicht nicht jedem bekant/ noch ſo viel mir bewußt von einigem Seribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß ſolcher Sachen ſo nöthig/ als das Commercien-Weſen ſelbſt/ welches ich bewieſen habe/ daß es zum Staat und Staats materi gar nah gehöre/ ſo haben Jhre Käyſerl. Majeſtät etwan vor zwey Jahren ein abſonderliches Collegium, und darüber Euer	Volltext (JPEG): Dedicatio. fdtwen / ist in gleichem bekank undtäglich vor Augen/ wie aber und auf was Weiß/ Weg und Mittel solches geschehe/ ist viel leicht nicht sedem bekank/ noch so viel mir bewußt von einigem Scribcnten bißhero entdecket und an Tag gegeben worden. Nach dcme aber die eigentliche Erkantnuß solcher Sachen so nöthig / als das Commerden-Wefen selbst/ weiches ich bewiesen habe/daß es zumSkaak undStaaks materi gar nah gehöre/ so haben Ihre Kayserl. Majestät etwa» vor zwey Jahren ein ab sonderliches Collegium,mit) darüber Euer

25 http://www.deutsches-textarchiv.de/book/view/becher_discurs_1668?p=8

Hochgrafl. Excellentz zum Præſiden- ten Allergnadigſt verordnet/ und mit einer außführlicher inſtruction verſehen/ welche dahin zielet/ daß man auf den Lauf der Commercien abſonderlich Achtung gebe/ damit er denen Kayſerlichen Erblanden nicht zum Schaden/ ſondern Nutzen ge- reiche/ in ſpecie, daß/ ſo viel möglich/ allerhand manufacturen/ vor welche nun viel Geld in die Frembde gehet/ in den Kay- ſerl. Erblanden gearbeitet/ und nicht allein das müſſig gehende Volck in die Arbeit ge- ſtellet/ das Geld im Land erhalten/ ſondern auch	Hochgrafl. Excelleny zum Presiden ten Allergnadigst verordnet / und mit einer außführlicher inftruüion versehen/welche dahin zielet / daß man auf den Lauf der Commerden absonderlich Achtunggebe / damit er denen Kayserlichen Erblanden nicht zum Schaden / sondern Nutzen ge reiche /inspecie, daß / so viel möglich/ allerhand manufaäumt/ vor welche nun viel Geld in die Frembde gehet/ in den Käy- serl. Erblanden gearbeitet /und nicht allein daömüssig gehende Volckindie Arbeit ge stellet/ das Geld im Land erhalten / sondern auch

Nachdem bisher verdeutlicht wurde, dass Artefakte in Bilddaten, die im TIFF-Format oft detaillierter vorliegen, starke Auswirkungen auf die Erkennung von Zeichen haben, wird nun stärker auf die Präzision und den Bezug zu Groundtruth-ähnlichen Daten eingegangen. Einerseits wird der erkannte Text mit Groundtruth-Daten verglichen, um über Distanzwerte auf einen Wert für die Präzision der OCR-Ergebnisse zu kommen und andererseits wird die Darstellung der Segmentierung anhand der zeilenweisen Aufbereitung beurteilt.

Als Groundtruth-Daten sind erneut die Volltexte aus dem Deutschen-Textarchiv gewählt worden. Es sind – wie in der gesamten Studie – die Algorithmen der Google-Diff-Match Implementation genutzt worden, um eine Levenshtein-Distanz zu erhalten, die die Anzahl der Schritte zum Einfügen, Löschen oder Austauchen von Zeichen angibt, die benötigt werden, um hier ein OCR-Ergebnis in seine Groundtruth-Referenz zu überführen.

In der folgenden Tabelle sind die Ergebnisse zweier OCR-Implementationen in unterschiedlichen Konfigurationen und unterschiedlichem Eingabemedien untersucht worden. Die untersuchte Seite ist aus dem Bestand der HAB und trägt die Signatur ‚sc-09'. Die Seite 8 von Bechers 1668 verfassten Werk „Politischer Diskurs" wurde sowohl im JPEG als auch im TIFF Format in ABBYY und Ocropus prozessiert. In Ocropus wurden zudem mehrere Binarisierungsmöglichkeiten und Sprach-Modelle getestet. Daher liegen hier wesentlich mehr Ergebnisse vor als beispielsweise mit ABBYY prozessiert wurden. Die vollständige Untersuchung mit allen genutzten Modellen befindet sich im Anhang. In der vorliegenden Tabelle sind ausschließlich die Vergleiche/Distanzverhältnisse mit den geringsten Distanzen (der vermeintlich höchsten Präzision) dargestellt.

Nr	Quelle	Software	Binarisierung	JPEG/TIFF	Modell	Distanz[26]
1	Groundtruth	ABBYY	n/a (Fraktur)	TIFF	deu,lat grie,heb	112
2	Groundtruth	ABBYY	n/a (Fraktur)	JPEG	deu,lat grie,heb	133
3	Groundtruth	OCROPUS	non-linear	TIFF	Fraktur-Modell	179
4	Groundtruth	OCROPUS	non-linear	JPEG	Fraktur-Modell	171
5	Groundtruth	OCROPUS	Sauvola	TIFF	Fraktur-Modell	262
6	Groundtruth	OCROPUS	Sauvola	JPEG	Fraktur-Modell	200
7	Groundtruth	OCROPUS	Sauvola	JPEG	Fraktur-Modell	162

Deutlich wird, dass die Implementation von ABBYY bei TIFF-Daten die geringste Distanz zum Groundtruth hat und dies obwohl die JPEG-Ergebnisse nur 20 Schritte mehr brauchen. Bezieht man hingegen die gleiche Fragestellung auf die Ergebnisse der Software Ocropus, so sind die Verhältnismäßigkeiten genau umgekehrt: Bei nicht-linearer Binarisierung, sowie bei Sauvola-Binarisierung, sind die JPEG-Ergebnisse beide Male besser als die dazugehörigen TIFF-Ergebnisse. Ökonomisch betrachtet sind die JPEG-Daten damit effizienter, da die Dateigrößen bei TIFF 5.05 MB und bei JPEG 1.01 MB sind. Die TIFF-Ergebnisse haben lediglich eine Abweichung von etwa 10 Schritten, was sie unweigerlich durch Konfiguration und Kalibrieren optimieren ließe.

Der OCR-Software ABBYY TIFF-Daten einzugeben würde somit eine Vorteil von ~19% in der Präzision bewirken. Bei Ocropus ist bei nicht-linearer Binarisierung mit JPEG-Daten hingegen ein Vorteil von ~9% entstanden. Betrachtet man das Datenvolumen im Verhältnis dazu, sind die TIFF Daten ~500% größer als die JPEG-Datenmenge oder umgekehrt: Die JPEG-Daten bilden ~20% des Gesamtvolumens. Bei einem Verhältnis von „1:5“ beim Mehraufwand des Datenvolumens sind Vorteile bereits beim Datentransfer gesichert, bedenkt man beispielsweise, dass so Prozesse mehrfach angestoßen werden können. Versucht man die Bild-Daten der großen Masse bereitzustellen, sind hier ebenfalls Vorteile zu garantieren: Das ganze Buch ist mit JPEG-Daten ~240 MB groß und bei TIFF sind es ~1200 MB. Möchte man Verbesserungsmaßnahmen zur Optimierung von OCR-Verfahren einleiten, also in Hinblick auf die Entwicklung von Binarisierungs-, Segmentierungs- und Entzerrungsalgorithmen, werden damit neue Ansätze ermöglicht. Es macht Sinn den hier beschriebenen Anteil an historischen Dokumenten an eine möglichst große Gruppe von Wissenschaftlern freizugeben, damit diese effizient daran arbeiten können.

26 https://code.google.com/p/google-diff-match-patch

Die Betrachtung der Ergebnisse der Sauvola binarisierten Daten kann an dieser Stelle nur recht isoliert geleistet werden. Der Grund dafür ist eine fehlerhafte Implementation in der Software selbst, sodass die Skew-Korrektur für Sauvola-Binarisierung entfallen musste. Die Ergebnisse der gesamten Untersuchung befinden sich strukturiert im Anhang dieser Studie. Man kann deutlich erkennen, dass die Segmentierung stark von der Binarisierung abhängt. In den Fällen mit Sauvola-Binarisierung der Ocropus-Implementation ist die Verzerrungskorrektur (Skew/Slant) fehlerhaft implementiert, daher sind verhältnismäßig viele weiße Flächen um die nicht ausgerichteten Zeilen entstanden. Dadurch ist wiederum eine gesteigerte Anzahl an Artefakten in den segmentierten Daten vorhanden, was zur Vermehrung von potentiellen Fehlern in der Erkennung (vor allem von Satzzeichen) führt.

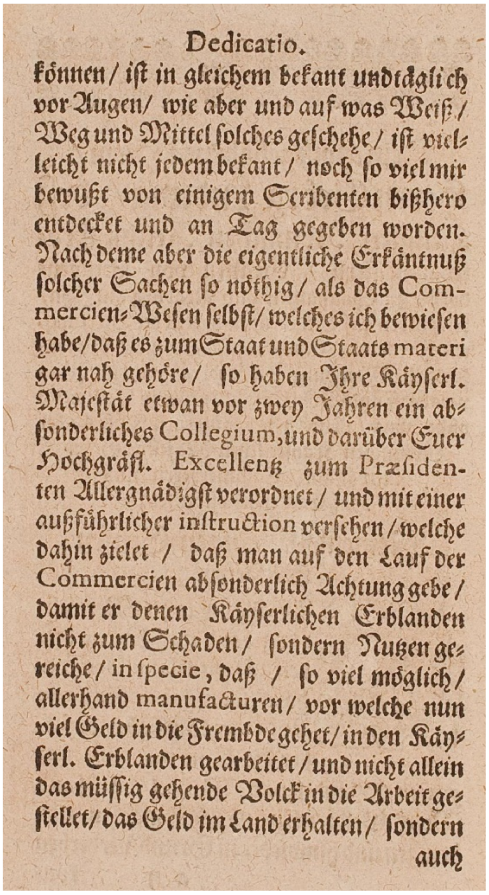
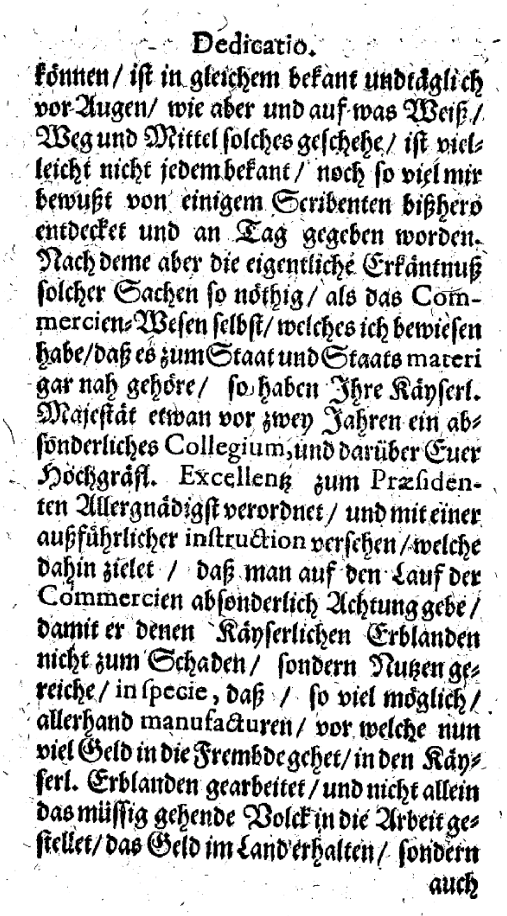
Auch wird in der folgenden Darstellung sehr gut deutlich, wieso die Algorithmen aufeinander aufbauen und inwiefern Zusammenhänge zwischen den Prozessschritten bestehen. Durch fehlende Rotation (Skew/Slant) ist nicht nur die Segmentierung der Zeilen betroffen, auch die Segmentierung der Wörter ist stark von unkorrigierten Verzerrungen betroffen. Die Trennung von Wörtern hat im folgenden Beispiel (Ocropus-JPEG-Sauvola-Frakant) nicht funktioniert.

Vornehmlich ist an diesem Beispiel in allen Belangen Optimierungspotenzial aus der Ableitung von Volltext über Bilddaten zu ersehen. Die Binarisierung mit Sauvola ist zwar wesentlich effizienter in Bezug auf Rechenressourcen, benötigt aber die Einstellung der Parameter Sigma und K, was zum einen Abläufe verlangsamt und zusätzlich Fachwissen erfordert. Auf diese Weise wären zahlreiche Artefakte in Form von kleineren Flecken und Punkten vermeidbar und eine initiale Konfiguration ist bei gegebenem Korpus-Umfang schnell sinnvoll. Aufgrund der fehlerhaften Implementation des ocropus-sauvola-scripts, wurde im Quellcode die Skew-Korrektur für diese Fallstudie deaktiviert. Die vorprozessierten Bilder sind daher nicht entsprechend in ihrer Drehung korrigiert. Daraus resultiert sowohl ein Fehlerpotential in Bezug auf das Ausschneiden der Zeilen (es resultieren viel zu große umgebende Rechtecke und damit zugleich darin viele Störungen)als auch eine Problematik, die auf die Segmentierung der Wörter zurückzuführen ist.

Trotz des also vermeintlich schlechten Ergebnisses, wird der Fokus jetzt auf die Testreihe Nr. 7 in der obigen Tabelle gesetzt. Obwohl die bisher dargestellten Umstände nicht ideal sind, sind mit dem Frakant-Modell und Standardeinstellungen aus den JPEG-Daten Ergebnisse mit einer geringeren Distanz als bei allen anderen Testläufen entstanden. Zusätzlich wurden ungetrennte Fontmixe auf dieser Seite (‚Collegium' [Z.14],‚Excellenz' [Z.15],'Commercien'[Z.19]) nur von der OCR-Implementation Ocropus korrekt erkannt.

Eine größere Schwierigkeit liegt demnach nicht mehr nur in der nutzbaren Technik, sondern vielmehr in deren Anwendung und Konfiguration. Sobald Testläufe massenweise durchgeführt werden können, macht es erst Sinn an der Grundlagen-Arbeit (OCR-Software und Digitalisierung) weiteren Aufwand zu investieren. Die Präzision zu messen, ist nur eine zu erwartende Komponente, die zukünftig wichtiger wird.

Ocropus 0.7 mit Sauvola-Binarsierung:

<table>
<tr>
<td>Original (JPEG):
 
Dedicatio.
können/ ist in gleichem bekant undtäglich
vor Augen/ wie aber und auf was Weiß/
Weg und Mittel solches geschehe/ ist viel-
leicht nicht jedem bekant/ noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig/ als das Com-
mercien-Wesen selbst/ welches ich bewiesen
habe/daß es zum Staat und Staats materi
gar nah gehöre/ so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium, und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet/ und mit einer
außführlicher instruction versehen/ welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe/
damit er denen Käyserlichen Erblanden
nicht zum Schaden/ sondern Nutzen ge-
reiche/ in specie, daß / so viel möglich/
allerhand manufacturen/ vor welche nun
viel Geld in die Frembde gehet/ in den Käy-
serl. Erblanden gearbeitet/ und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet/ das Geld im Land erhalten/ sondern
auch</td>
<td>Sauvola Binarisierung (JPEG):
 
Dedicatio.
können/ ist in gleichem bekant undtäglich
vor Augen/ wie aber und auf was Weiß/
Weg und Mittel solches geschehe/ ist viel-
leicht nicht jedem bekant/ noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig/ als das Com-
mercien-Wesen selbst/ welches ich bewiesen
habe/daß es zum Staat und Staats materi
gar nah gehöre/ so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium, und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet/ und mit einer
außführlicher instruction versehen/ welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe/
damit er denen Käyserlichen Erblanden
nicht zum Schaden/ sondern Nutzen ge-
reiche/ in specie, daß / so viel möglich/
allerhand manufacturen/ vor welche nun
viel Geld in die Frembde gehet/ in den Käy-
serl. Erblanden gearbeitet/ und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet/ das Geld im Land erhalten/ sondern
auch</td>
</tr>
<tr>
<td>Groundtruth[27]

Dedicatio.

können/ ist in gleichem bekant undtäglich

vor Augen/ wie aber und auf was Weiß/</td>
<td>Zeilensegmentierung/Volltext (JPEG/Frakant):

Dedicatio.
.- DöHicarlo .

können/ ist in gleichem bekant undtdglich.

wor Augenu wie aber und aufF was Weß/</td>
</tr>
</table>

27 http://www.deutsches-textarchiv.de/book/view/becher_discurs_1668?p=8

Weg und Mittel ſolches geſchehe/ iſt viel-	Weg und Mittel ſolches geſchehe/ iſt viel-
leicht nicht jedem bekant/ noch ſo viel mir	leicht nicht jedem bekant/ noch ſo viel mir
bewußt von einigem Seribenten bißhero	bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.	entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkåntnuß	Nach deme aber die eigentliche Erkäntnuß
ſolcher Sachen ſo nöthig/ als das Com-	ſolcher Sachen ſo nöthig/ als das Com-
mercien-Weſen ſelbſt/ welches ich bewieſen	mercien-Weſen ſelbſt/ welches ich bewieſen
habe/ daß es zum Staat und Staats materi	habe/daß es zum Staat und Staats materi
gar nah gehöre/ ſo haben Jhre Kåyſerl.	gar nah gehöre/ ſo haben Jhre Käyſerl.
Majeſtåt etwan vor zwey Jahren ein ab-	Majeſtät etwan vor zwey Jahren ein ab-
ſonderliches Collegium, und darüber Euer	ſonderliches Collegium, und darüber Euer
Hochgråfl. Excellentz zum Præſiden-	Höchgräfl. Excellentz zum Præſiden-
ten Allergnådigſt verordnet/ und mit einer	ten Allergnädigſt verordnet/ und mit einer
außführlicher inſtruction verſehen/ welche	außführlicher inſtruction verſehen/ welche
dahin zielet/ daß man auf den Lauf der	dahin zielet/ daß man auf den Lauf der

Commercien abſonderlich Achtung gebe/	Commercien abſonderlich Achtung gebe/
	. Commercien absonderlich Achtunggebe/
damit er denen Kåyſerlichen Erblanden	damit er denen Käyſerlichen Erblanden
	damit er denen Kiyserlichen Erblanden
nicht zum Schaden/ ſondern Nutzen ge-	nicht zum Schaden/ ſondern Nutzen ge=
	nicht zum Sondern Nusenges
reiche/ in ſpecie, daß/ ſo viel mögliche/	reiche/ in ſpecie, daß / ſo viel möglich/
	reiche/ inspcoia, daß / so viel möglich
allerhand manufacturen/ vor welche nun	allerhand manufacturen/ vor welche nun
	allerhand manufacturen/ vor welche nun
viel Geld in die Frembde gehet/ in den Kåy-	viel Geld in die Frembde gehet/ in den Käy=
	viel GeldindieFrembdegchetinden Käy-.
ſerl. Erblanden gearbeitet/ und nicht allein	ſerl. Erblanden gearbeitet/ und nicht allein
	serl. Erblanden gearbeitetundnichtallein
das müſſig gehende Volck in die Arbeit ge-	das müſſig gehende Volck in die Arbeit ge=
	dasmiissg gehende Volckindie Arbeitgee'
ſtellet/ das Geld im Land erhalten/ ſondern	ſtellet/ das Geld im Land erhalten/ ſondern
	sielletdas Geldimlanderhaltenn/ sondern
auch	auch
	auch

3.4.4 Fallstudie 4 – Diskussion

Der bisher oft angedeutete und immer wieder angesprochene Term ‚Vorverarbeitung' ist eklatant wichtig in Bezug auf sehr viele Fragestellungen, die sich mit dem Thema ‚Verbesserung von OCR-Ergebnissen' befasst. Ein Ergebnis dieser Studie ist die nachgewiesene Korrelation zwischen den Bildkompressionsverfahren und deren zu erwartenden Ergebnissen. Ein Algorithmus, der mit schlechterer Qualität gleiche oder effizientere Arbeit leistet, ist zweifelsfrei der bessere Algorithmus. In Bezug auf OCR heißt das nun, dass skalierungsinvariante Algorithmen zum Erkennen von Zeichen nicht isoliert betrachtet werden dürften, gleichwohl eine geringere Entropiedichte (Bildkompressionsverfahren) nicht zwangsläufig zu einem schlechteren Ergebnis führen muss.

Betrachtet man die historische Entwicklung von Segmentierungsalgorithmen, sind die Zusammenhänge zwischen Ergebnissen und algorithmischen Vorgehen schnell verdeutlicht. Ein varianzunabhängiger Segmentierungsalgorithmus stilisiert eine Zeile mit Wörtern soweit mit einem Weichzeichner, dass zum Schluss nur noch Pixelhaufen übrig bleiben, die zur Identifikation der Wortgrenzen solide und vor allem varianzunabhängige Arbeit leisten. Man spricht in der Fachliteratur daher auch von verbundenen Binary Large Objects (Blob) (Angadi and Kodabagi 2014) oder ‚connected components' (Smith 2007), die wiederum verbundene Elemente repräsentieren.

Abbildung 21: Layoutanalyse mit dem Algorithmus von Manmatha und Rothfeder.

Eine Layoutanalyse, wie die obige von Manmatha and Rothfeder, gibt zunächst Aufschluss über Bereiche, die Zeilen von Fließtext enthalten (Manmatha and Rothfeder 2005). Ihr Algorithmus hat sich laut Fachliteratur als überaus solide und robust herausgestellt. Selbst in Bildern mit viel Störung erzielt er gute Ergebnisse ohne kostenintensive Thresholding-Algorithmen bemühen zu müssen. Auch viele weitere Störungsquellen stehen nicht in Abhängigkeit mit diesem Algorithmus. Die Maße einer Zeile werden über verbundene 'Blobs' bestimmt, die durch einen anisotropischen Laplacian-Filter von einem Gauss-Kernel berechnet werden. Diese Berechnung zeichnet die Bilddaten mit Parametern weich, sodass Pixelhaufen übrig bleiben, die verbundene Elemente repräsentieren. Der gesamte Algorithmus bewirkt, dass das Bild invertiert, gesäubert und wie zuvor dargestellt weichgezeichnet wird. Danach wird es in Zeilen aufgeteilt und in Wörter zerteilt. Zuletzt werden umgebende Rechtecke erzeugt, mit denen die Wörter in Feature-Vektoren extrahiert werden.

Auch wenn die Layoutanalyse kaum noch als Forschungsvorhaben in der Volltexterfassung von historischen Dokumenten auftaucht, ist sie ein sehr stark bearbeitetes Feld im Kontext von OCR. Aus diesem Grund wäre der Zusammenschluss der Forschungsstände in Umgebungen mit

weitreichendem Zulauf äußerst ratsam. Denn auch wenn Projekte wie eMOP[28] oder auch das bereits genannte Projekt OCRopodium ihre anfänglich erwarteten Ziele relativiert haben[29], Retrodigitalisierung auf verteilten IT-Architekturen zur massenweisen Verarbeitung und Bearbeitung ist der zurzeit einzig sinnvolle Weg, um das Thema OCR weiter verhandeln zu können. Die vorliegenden Leistungsstände sind bereits weit fortgeschritten, aber dennoch werden bessere Ergebnisse notwendig werden.

Die vorliegenden Ergebnisse sind daher nicht den wissenschaftlichen Zielvorgaben entsprechend. Mit dieser Studie ist ein Diskurs eröffnet worden, der sowohl vor als auch hinter den etablierten Prozessketten der üblichen Software-Implementation ansetzt. Zum einen ist großer Wert auf die Qualität des Eingabematerials gelegt worden, um die Frage zu beantworten: Was sich an Verbesserungsmaßnahmen in dieser Hinsicht für Bibliotheken/Datenhaltern eröffnen lässt. Zum anderen sind zur Identifikation von Forschungsfeldern, in denen noch potentiell Bedarf besteht, Arbeiten dokumentiert worden, die sowohl Nutzer von OCR-Software, als auch Entwickler von OCR-Software, entsprechende Tendenzen aufweisen.

Das Hauptaugenmerk lag auf der Beantwortung von Fragen zur Machbarkeit und tendenziellen Arbeitsfeldern der Forschung, die zukünftig zu erwarten sind. Es wurde argumentiert und dargestellt, dass JPEG-Formate durchaus nutzbar sind, um etwa varianzunabhängige Algorithmen zu entwickeln oder Konfigurations-Sets zu finden, die in einer Software in Kombination mit einem historischen Druckerzeugnis beste Ergebnisse erzielen. Auch der Austausch von Daten und damit verbundenes kollaboratives Arbeiten ist mit JPEG-Daten sehr gut möglich.

Allerdings hängt die Wahl der Materialen und Werkzeuge immer vom Zweck des Vorhabens ab. Wird großer Wert auf Nutzbarkeit, Lesbarkeit und Exploration der Bilddaten gelegt, ist ggf. ein GeoTiff-Format (oder Pyramid-Images) sicher besser geeignet, als jedes kompressionsfreie TIFF oder quantisierte JPEG-Format.

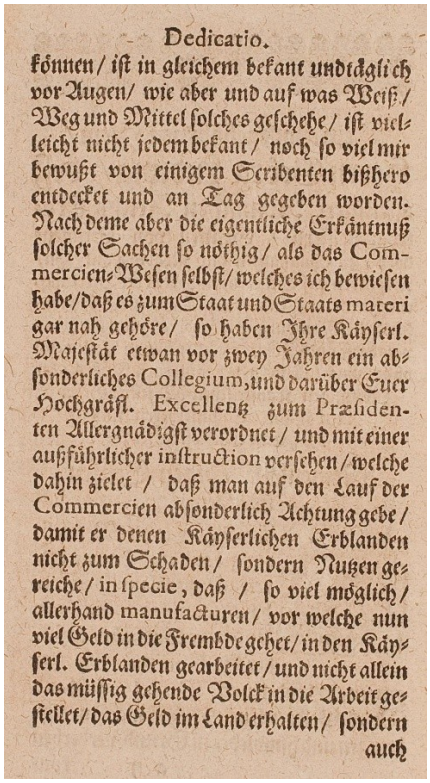
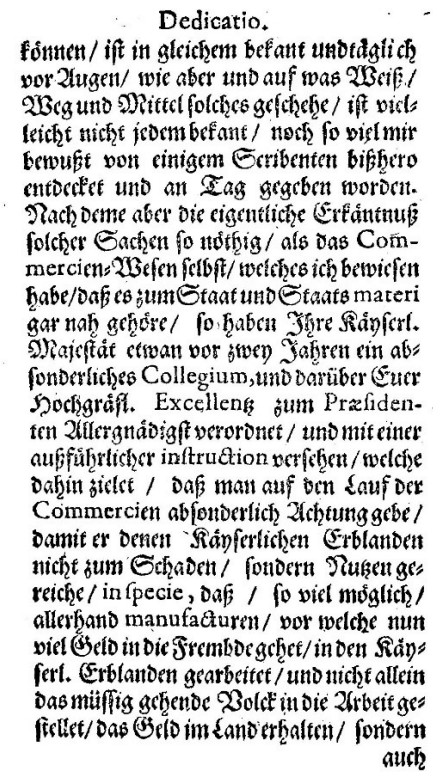
Unter dem Stichwort ‚BigData' befinden sich schon seit einiger Zeit viele Themen im Umlauf. Auch für den Kontext dieser Studie wurden Methoden entwickelt, die bisher im Kontext von OCR keine Anwendung gefunden haben Unter dem Term ‚OpenData' werden Zugänglichkeit und Nutzbarkeit im öffentlichen Sektor verhandelt und dieser appelliert zugleich an die Datenhalter: Ein essentielles und in dieser Studie inhärent befindliches Anliegen ist daher die Freigabe und Strukturierung aller Daten aus OCR-Projekten.

28 http://emop.tamu.edu

29 http://www.kcl.ac.uk/innovation/groups/cerch/research/projects/completed/ocropodium.aspx

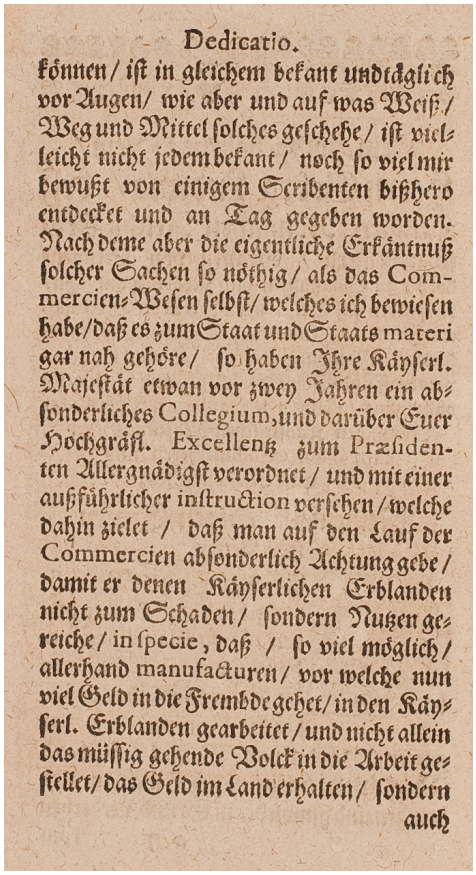
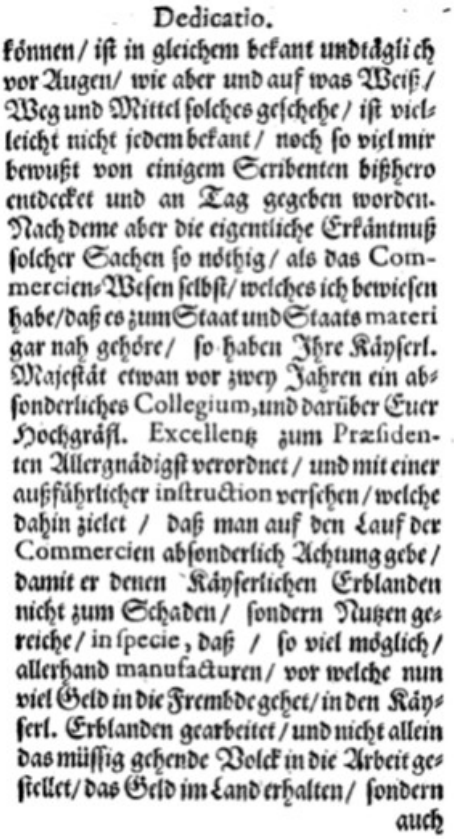
3.4.5 Ergebnisse aus Fallstudie 4

ABBYY-Recognition-Server 3.5:

Original (JPEG):	Binarisierung (JPEG):
 Dedicatio. können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist viel- leicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Com- mercien-Wesen selbst/ welches ich bewiesen habe/daß es zumStaat undStaats materi gar nah gehöre / so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein ab- sonderliches Collegium,und darüber Euer Hochgräfl. Excellentz zum Præsiden- ten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen ge- reiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käy- serl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit ge- stellet / das Geld im Land erhalten / sondern auch	 Dedicatio. können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist viel- leicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Com- mercien-Wesen selbst/ welches ich bewiesen habe/daß es zumStaat undStaats materi gar nah gehöre / so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein ab- sonderliches Collegium,und darüber Euer Hochgräfl. Excellentz zum Præsiden- ten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen ge- reiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käy- serl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit ge- stellet / das Geld im Land erhalten / sondern auch
Groundtruth: Dedicatio. kõnnen/ iſt in gleichem bekant undtãglich vor Augen/ wie aber und auf was Weiß/ Weg und Mittel ſolches geſchehe/ iſt viel- leicht nicht jedem bekant/ noch ſo viel mir bewußt von einigem Seribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkãntnuß ſolcher Sachen ſo nõthig/ als das Com- mercien-Weſen ſelbſt/ welches ich bewieſen habe/ daß es zum Staat und Staats materi gar nah gehõre/ ſo haben Jhre Kãyſerl. Majeſtãt etwan vor zwey Jahren ein ab- ſonderliches Collegium, und darũber Euer Hochgrãfl. Excellentz zum Præſiden- ten Allergnãdigſt verordnet/ und mit einer	Volltext (JPEG): Dedicatio. fdtwen / ist in gleichem bekank undtäglich vor Augen/ wie aber und auf was Weiß/ Weg und Mittel solches geschehe/ ist viel leicht nicht sedem bekank/ noch so viel mir bewußt von einigem Scribcnten bißhero entdecket und an Tag gegeben worden. Nach dcme aber die eigentliche Erkantnuß solcher Sachen so nöthig / als das Com- merden-Wefen selbst/ weiches ich bewiesen habe/daß es zumSkaak undStaaks materi gar nah gehöre/ so haben Ihre Kayserl. Majestät etwa» vor zwey Jahren ein ab sonderliches Collegium,mit) darüber Euer Hochgrafl. Excelleny zum Presiden ten Allergnadigst verordnet / und mit einer

außfůhrlicher inſtruction verſehen/ welche dahin zielet/ daß man auf den Lauf der Commercien abſonderlich Achtung gebe/ damit er denen Kåyſerlichen Erblanden nicht zum Schaden/ ſondern Nutzen gereiche/ in ſpecie, daß/ ſo viel mögliche/ allerhand manufacturen/ vor welche nun viel Geld in die Frembde gehet/ in den Kåyſerl. Erblanden gearbeitet/ und nicht allein das müſſig gehende Volck in die Arbeit geſtellet/ das Geld im Land erhalten/ ſondern auch	außführlicher inftruüion versehen/welche dahin zielet / daß man auf den Lauf der Commerden absonderlich Achtunggebe / damit er denen Kayserlichen Erblanden nicht zum Schaden / sondern Nutzen ge reiche /inspecie, daß / so viel möglich/ allerhand manufaäumt/ vor welche nun viel Geld in die Frembde gehet/ in den Käyserl. Erblanden gearbeitet /und nicht allein daömüssig gehende Volckindie Arbeit ge stellet/ das Geld im Land erhalten / sondern auch

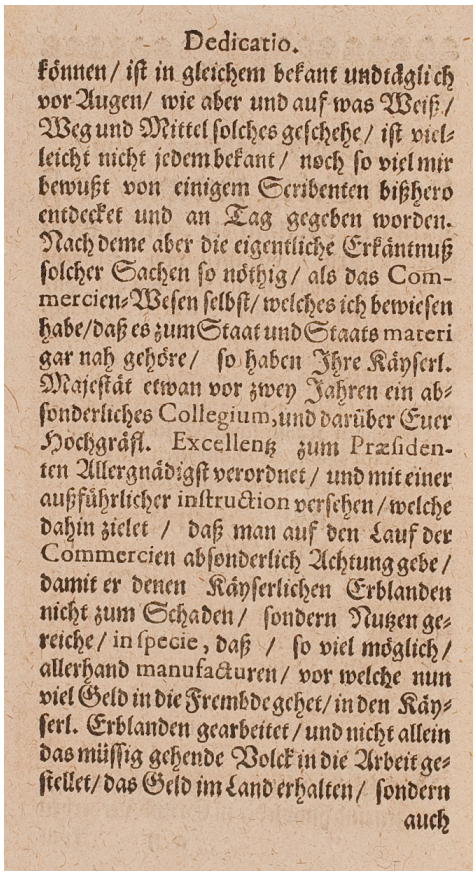
ABBYY-Recognition-Server 3.5:

Original TIFF:  Dedicatio. können/ iſt in gleichem bekant undtägli ch vor Augen/ wie aber und auf was Weiß/ Weg und Mittel ſolches geſchehe/ iſt vielleicht nicht jedem bekant/ noch ſo viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß ſolcher Sachen ſo nöthig/ als das Commercien-Weſen ſelbſt/ welches ich bewieſen habe/daß es zum Staat und Staats materi gar nah gehöre/ ſo haben Ihre Käyſerl. Majeſtät etwan vor zwey Jahren ein abſonderliches Collegium,und darüber Euer Hochgräfl. Excellentz zum Præſidenten Allergnädigſt verordnet/ und mit einer außführlicher inſtruction verſehen/welche dahin zielet / daß man auf den Lauf der Commercien abſonderlich Achtung gebe/ damit er denen Käyſerlichen Erblanden nicht zum Schaden/ ſondern Nutzen gereiche/ in ſpecie, daß / ſo viel möglich/ allerhand manufacturen/ vor welche nun viel Geld in die Frembde gehet/ in den Käyſerl. Erblanden gearbeitet/ und nicht allein das müſſig gehende Volck in die Arbeit geſtellet/ das Geld im Land erhalten/ ſondern auch	Binarisierung TIFF:  Dedicatio. können/ iſt in gleichem bekant undtägli ch vor Augen/ wie aber und auf was Weiß/ Weg und Mittel ſolches geſchehe/ iſt vielleicht nicht jedem bekant/ noch ſo viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß ſolcher Sachen ſo nöthig/ als das Commercien-Weſen ſelbſt/ welches ich bewieſen habe/daß es zum Staat und Staats materi gar nah gehöre/ ſo haben Ihre Käyſerl. Majeſtät etwan vor zwey Jahren ein abſonderliches Collegium,und darüber Euer Hochgräfl. Excellentz zum Præſidenten Allergnädigſt verordnet/ und mit einer außführlicher inſtruction verſehen/welche dahin zielet / daß man auf den Lauf der Commercien abſonderlich Achtung gebe/ damit er denen Käyſerlichen Erblanden nicht zum Schaden/ ſondern Nutzen gereiche/ in ſpecie, daß / ſo viel möglich/ allerhand manufacturen/ vor welche nun viel Geld in die Frembde gehet/ in den Käyſerl. Erblanden gearbeitet/ und nicht allein das müſſig gehende Volck in die Arbeit geſtellet/ das Geld im Land erhalten/ ſondern auch

Groundtruth:	Volltext-Ergebnis (TIFF):
Dedicatio.	Dedicatio*
kͤonnen/ iſt in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß/ Weg und Mittel ſolches geſchehe/ iſt viel- leicht nicht jedem bekant/ noch ſo viel mir bewußt von einigem Seribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß ſolcher Sachen ſo nöthig/ als das Com- mercien-Weſen ſelbſt/ welches ich bewieſen habe/ daß es zum Staat und Staats materi gar nah gehöre/ ſo haben Jhre Käyſerl. Majeſtät etwan vor zwey Jahren ein ab- ſonderliches Collegium, und darüber Euer Hochgräfl. Excellentz zum Præſiden- ten Allergnädigſt verordnet/ und mit einer außführlicher inſtruction verſehen/ welche dahin zielet/ daß man auf den Lauf der Commercien abſonderlich Achtung gebe/ damit er denen Käyſerlichen Erblanden nicht zum Schaden/ ſondern Nutzen ge- reiche/ in ſpecie, daß/ ſo viel möglich/ allerhand manufacturen/ vor welche nun viel Geld in die Frembde gehet/ in den Käy- ſerl. Erblanden gearbeitet/ und nicht allein das müſſig gehende Volck in die Arbeit ge- ſtellet/ das Geld im Land erhalten/ ſondern auch	können/ ist in gleichem bekattt Uttdkäglr'ch vor Augen/ wie aber und auf was Weiß/ Weg und Mittel solches geschehe / ist viel leicht nicht jedem bekam/ noch sovi/lmir bewußt von einigem Scribenkm bißhero entdecket und an Tag gegeben worden. Nachdeme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Com- rnercien-Wesen selbst/ welches ich bewiesen habe/daß es ZumSkaak undStaaks materi gar nah gehöre/ so haben Ihre Kayscrl. Majestät ekwan vor zwey Jahren ein ab- sonderliches Collegiummnb darüber Euer Hochgräsi. Excellens Zum Präsiden -im Allergnädigst verordnet / und mit einer ausführlicher initruction versehen/welche dahin zielet / daß man auf dm Lauf der Commcrcim absonderlich Achtung gebe/ damit er denen Kayserlichen Erblanden nicht zum Schaden/ sondern Nutzen ge- mche/infpecie,baß / so viel möglich/ allerhand manufa&umt/ vor welche nun viel Geld in die Frembdegchet/ in den Kay- fcrl. Erblanden gearbeitet/und nicht allein dasmüssig gehende Volckindie Arbeit ge- stcllet/ das Geld im Landerhalten / sondern auch

Ocropus 0.7 mit nicht-linearer Binarsierung:

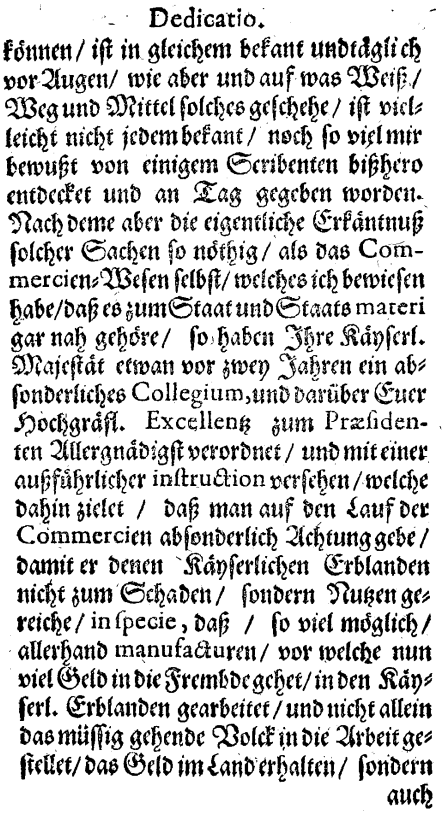
Original TIFF:



Dedicatio.

können / ist in gleichem bekant undtägli ch
vor Augen/ wie aber und auf was Weiß /
Weg und Mittel solches geschehe / ist viel-
leicht nicht jedem bekant / noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig / als das Com-
mercien-Wesen selbst/ welches ich bewiesen
habe/daß es zum Staat und Staats materi
gar nah gehöre / so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium,und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet / und mit einer
außführlicher instruction versehen / welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe /
damit er denen Käyserlichen Erblanden
nicht zum Schaden / sondern Nutzen ge-
reiche / in specie, daß / so viel möglich /
allerhand manufacturen / vor welche nun
viel Geld in die Frembde gehet/ in den Käy-
serl. Erblanden gearbeitet / und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet/ das Geld im Land erhalten / sondern
auch

Nicht-lineare Binarisierung (TIFF):



Dedicatio.

können / ist in gleichem bekant undtägli ch
vor Augen/ wie aber und auf was Weiß /
Weg und Mittel solches geschehe / ist viel-
leicht nicht jedem bekant / noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig / als das Com-
mercien-Wesen selbst/ welches ich bewiesen
habe/daß es zum Staat und Staats materi
gar nah gehöre / so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium,und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet / und mit einer
außführlicher instruction versehen / welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe /
damit er denen Käyserlichen Erblanden
nicht zum Schaden / sondern Nutzen ge-
reiche / in specie, daß / so viel möglich /
allerhand manufacturen / vor welche nun
viel Geld in die Frembde gehet/ in den Käy-
serl. Erblanden gearbeitet / und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet/ das Geld im Land erhalten / sondern
auch

Zeilensegmentierung/Volltext (TIFF/Fraktur):

Dedicatio.

Deiaatio.

können / ist in gleichem bekant undtägli ch

ksnnenz ist in gleschem bekant undtäkglich

vor Augen/ wie aber und auf was Weiß /

vorAugen? wic aber und auf was Weif

Weg und Mittel solches geschehe / ist viel-

Wegund Mittclsolches geschehe ist viel-

leicht nicht jedem bekant / noch so viel mir

leicht nichht jedembckant,? noch so vßelmir

bewußt von einigem Scribenten bißhero

bewußt von eingem Seribenten bißhero

entdecket und an Tag gegeben worden.

entdecket und an Tag gegeben worden-

Nach deme aber die eigentliche Erkäntnuß

Nachdeme aber die egentlsche Erkäntnuß

solcher Sachen so nöthig / als das Com-

solcher Sachen so nöthig? als das Conn-

mercien-Wesen selbst/ welches ich bewiesen

weroien»Wesen selbst- welchesichbewiesen

habe/daß es zum Staat und Staats materi

habesdaß eszumStaatundStaats mareri

gar nah gehöre / so haben Ihre Käyserl.

gar nah gehöre so- habcn Ihre Käyscrl.

Majestät etwan vor zwey Jahren ein ab-

Mdajestät etwan vor zwey Jahren ein abe

sonderliches Collegium, und darüber Euer

sonderliches Collegiuw.und darüber Euer

Hochgräfl. Excellentz zum Præsiden-

Hochgräfl. Hcallenz zum lralicen-

ten Allergnädigst verordnet / und mit einer

ten Allergnädsgstverordnctz und miteiner

außführlicher instruction versehen / welche

außführlicher iltruäionverschenz welche

dahin zielet / daß man auf den Lauf der

dahin ziclet - daß man auf den Lauf der

Commercien absonderlich Achtung gebe /

Correraien absonderlich Achtunggebe?

damit er denen Käyserlichen Erblanden

damit er denen Käyserlichen Erblanden

nicht zum Schaden / sondern Nutzen ge=

nicht zum Schadenz sondern Nuzenge-

reiche / in ſpecie, daß / ſo viel möglich /

reiche? in igeoio, daß - so viel möglich

allerhand manufacturen / vor welche nun

allerhand manukaäturen vor welche nun

viel Geld in die Frembde gehet / in den Käy=

vielGeld indieFrembdegchet?inden Käy-

ſerl. Erblanden gearbeitet / und nicht allein

serl. Erblanden gearbeitetkund uicht allein

das müſſig gehende Volck in die Arbeit ge=

dasmüissßz gchende Polck indie Arbeitges

ſtellet / das Geld im Land erhalten / ſondern

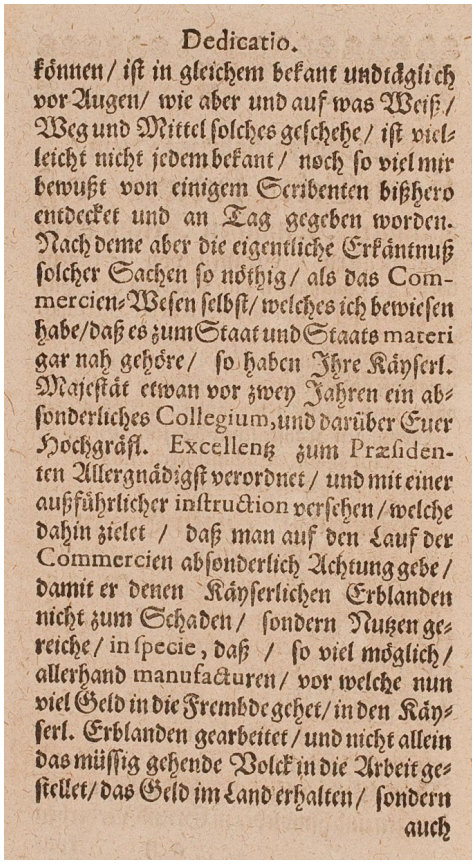
jicllet-das Geld im Tanderhaltenz sondern

auch

auch

Ocropus 0.7 mit nicht-linearer Binarsierung:

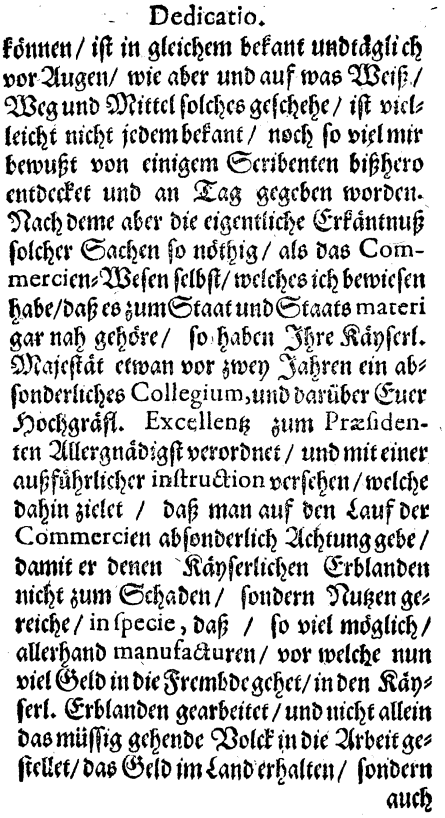
Original (JPEG):



Dedicatio.

können / ist in gleichem bekant undtägli ch
vor Augen/ wie aber und auf was Weiß /
Weg und Mittel solches geschehe / ist viel-
leicht nicht jedem bekant / noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig / als das Com-
mercien-Wesen selbst/ welches ich bewiesen
habe/daß es zum Staat und Staats materi
gar nah gehöre / so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium, und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet / und mit einer
außführlicher instruction versehen / welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe /
damit er denen Käyserlichen Erblanden
nicht zum Schaden / sondern Nutzen ge-
reiche / in specie, daß / so viel möglich /
allerhand manufacturen / vor welche nun
viel Geld in die Frembde gehet / in den Käy-
serl. Erblanden gearbeitet / und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet / das Geld im Land erhalten / sondern
auch

Nicht-lineare Binarisierung (TIFF):



Dedicatio.

können / ist in gleichem bekant undtägli ch
vor Augen/ wie aber und auf was Weiß /
Weg und Mittel solches geschehe / ist viel-
leicht nicht jedem bekant / noch so viel mir
bewußt von einigem Scribenten bißhero
entdecket und an Tag gegeben worden.
Nach deme aber die eigentliche Erkäntnuß
solcher Sachen so nöthig / als das Com-
mercien-Wesen selbst/ welches ich bewiesen
habe/daß es zum Staat und Staats materi
gar nah gehöre / so haben Ihre Käyserl.
Majestät etwan vor zwey Jahren ein ab-
sonderliches Collegium, und darüber Euer
Hochgräfl. Excellentz zum Præsiden-
ten Allergnädigst verordnet / und mit einer
außführlicher instruction versehen / welche
dahin zielet / daß man auf den Lauf der
Commercien absonderlich Achtung gebe /
damit er denen Käyserlichen Erblanden
nicht zum Schaden / sondern Nutzen ge-
reiche / in specie, daß / so viel möglich /
allerhand manufacturen / vor welche nun
viel Geld in die Frembde gehet / in den Käy-
serl. Erblanden gearbeitet / und nicht allein
das müssig gehende Volck in die Arbeit ge-
stellet / das Geld im Land erhalten / sondern
auch

Zeilensegmentierung & Volltext (JPEG/Fraktur):

Dedicatio.

Deeiaatio.

können / ist in gleichem bekant undtägli ch

ksunenz ist in gleschem bekant undtäkglsch

vor Augen/ wie aber und auf was Weiß /

vorAugen? wic aber und auf was Weifz

Weg und Mittel solches geschehe / ist viel-

Wegund Msttclsolchesgeschehe? ist viel-

leicht nicht jedem bekant / noch so viel mir

leicht nicht jedem bckant, noch so vielmir

bewußt von einigem Scribenten bißhero

bewußt von eingem Scribenten bißhero

entdecket und an Tag gegeben worden.

entdecket und an Tag gegeben worden.

Nach deme aber die eigentliche Erkäntnuß

Nach deme aber die esgentliche Erkäntnuß

solcher Sachen so nöthig / als das Com-

solcher Sachen so nöthig? als das Conm-

mercien-Wesen selbst / welches ich bewiesen

reroien-Wesen selbst-welches ich bewiesen

habe / daß es zum Staat und Staats materi

habesdaß eszumStaatundStaats mareri

gar nah gehöre / so haben Ihre Käyserl.

gar nah gehöre- so- haben Ihre Käyscrl.

Majestät etwan vor zwey Jahren ein ab-

Majeskät ctwan vor zwey Jahren ein abe

sonderliches Collegium, und darüber Euer

sonderlches Collegiuw.und barüber Euer

Hochgräfl. Excellentz zum Præsiden-

Hochgräfl. kcellenz zum lraücen-

ten Allergnädigst verordnet / und mit einer

ten Allergnädigstverordnctz und miteiner

außführlicher instruction versehen / welche

außführlicher iwltruäionverschen- welche

dahin zielet / daß man auf den Lauf der

dahin ziclet - daß man auf den Lauf der

Commercien absonderlich Achtung gebe /

Cotreraien absonderlich Achtunggebe?

damit er denen Käyserlichen Erblanden

damit er denen Käyserlichen Erblanden

nicht zum Schaden / ſondern Nutzen ge=

nicht zum Schadenz sondern Nuzenge-

reiche / in ſpecie, daß / ſo viel möglich /

reichef in sgeaio, daß so vscl möglich

allerhand manufacturen / vor welche nun

allerhand manutaäturenz vor welche nun

viel Geld in die Frembde gehet / in den Käy=

vielGeld indieFrembdegchet?inden Käy-

ſerl. Erblanden gearbeitet / und nicht allein

serl. Erblanden gearbeitetundnicht allein

das müſſig gehende Volck in die Arbeit ge=

dasmüüssßg gchende Polckindie Arbeitgee

ſtellet / das Geld im Land erhalten / ſondern

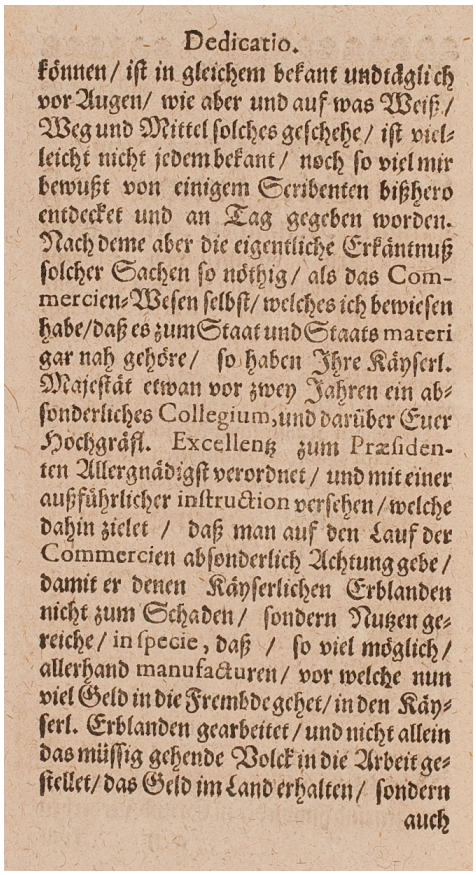
stcllet-das Geld sm Tanderhaltenz sondern

auch

auch

Ocropus 0.7 mit Sauvola-Binarsierung:

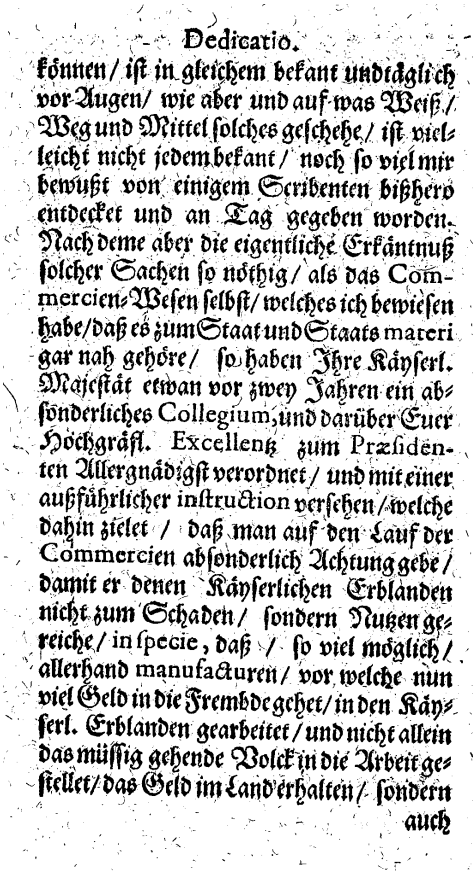
Original (TIFF):



Dedicatio.

können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist vielleicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Commercien-Wesen selbst/ welches ich bewiesen habe/daß es zum Staat und Staats materi gar nah gehöre / so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein absonderliches Collegium, und darüber Euer Hochgräfl. Excellentz zum Præsidenten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen gereiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käyserl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit gestellet / das Geld im Land erhalten / sondern auch

Sauvola Binarisierung (TIFF):



Dedicatio.

können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist vielleicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Commercien-Wesen selbst/ welches ich bewiesen habe/daß es zum Staat und Staats materi gar nah gehöre / so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein absonderliches Collegium, und darüber Euer Hochgräfl. Excellentz zum Præsidenten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen gereiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käyserl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit gestellet / das Geld im Land erhalten / sondern auch

Zeilensegmentierung/Volltext (TIFF/Fraktur):

Dedicatio.

- - Deehaarlo. -

können / ist in gleichem bekant undtäglich

ksnnenf ist in.gleschem bekant undtäglsch

vor Augen/ wie aber und auf was Weiß /

vorAugen? wse aber und aufwas WeifF

Weg und Mittel solches geschehe / ist viel-

Wegund Msttelsolchesgeschehe? isk niel-

leicht nicht jedem bekant / noch so viel mir

lescht nicht jedembekant noch so vfelmir

bewußt von einigem Scribenten bißhero

bewißt von' eingem Scrbenten bßßers

entdecket und an Tag gegeben worden.

entdegket und an Tag gegeben worden,

Nachdeme aber die eigentliche Erkäntnuß

Nachdeme aber die eßgentliche Erkäntnuß

solcher Sachen so nöthig / als das Com-

solcher Sachen so nöthsg? als das Cöth-

mercien-Wesen selbst / welches ich bewiesen

wraien»esen selbsk-welchesich bewiesen

habe / daß es zum Staat und Staats materi

habesdaß eszumStaatundStaats wateri

gar nah gehöre / so haben Ihre Käyserl.

gar nah gehöre sohabcn Ihre Käyserk.

Majestät etwan vor zwey Jahren ein ab-

Majeskät eewan vor zwey Jahren ein abs

sonderliches Collegium, und darüber Euer

sonderliches CollegiunJund darüberEuer

ten Allergnädigst verordnet / und mit einer

, Höchgräfl. Hcellent zum kraücen-

außführlicher instruction versehen / welche

außfüßrlscher iolkruäionverschenzwvelche

dahin zielet / daß man auf den Lauf der

dahin zielet - daß man auf'den Lauf der -

Commercien absonderlich Achtung gebe /

Cowroraien abssnderlsch Achtunggebe?

damit er denen Käyserlichen Erblanden

damit er denen Käyserlichen Erblanden

nicht zum Schaden / sondern Nutzen ge=

nicht zum Schadenz sondern Nuzenge-

reiche / in specie, daß / so viel möglich /

reiche? in sgeaio, daß - so vel möglsch? -

allerhand manufacturen / vor welche nun

allerhand manutaäturenz vor welche nun

viel Geld in die Frembde gehet / in den Käy=

vielGeld indieFrembdegchet?inden Kän-

serl. Erblanden gearbeitet / und nicht allein

serl. Erblanden gearbeitet-und nschtallein

das müssig gehende Volck in die Arbeit ge=

dasmüissgz gehende Volckfndie Arbeicges

stellet / das Geld im Land erhalten / sondern

- sellezdas GeldsmTanderhalten sondern

auch

auch

Zeilensegmentierung/Volltext (TIFF/Frakant):

Dedicatio.

Dötcarlo. .

können / ist in gleichem bekant und tägli ch

können/ ist ingleichem bekant undtdgltch.

vor Augen / wie aber und auf was Weiß /

wor Augenu wie aber undauf was Wenß/

Weg und Mittel solches geschehe / ist viel=

Wegund Mittelsolchesgeschehez st viel-

leicht nicht jedem bekant / noch so viel mir

leicht nicht jedembekant noch so vielmir

bewußt von einigem Scribenten bißherö

bewußt vor einigem Scribenten bißherö

entdecket und an Tag gegeben worden.

entdecket und an Tag gegeben worden,

Nachdeme aber die eigentliche Erkäntnuß

Nachdene äber die eigentlich Erkäntnuß

solcher Sachen so nöthig / als das Com-

solcher Sachen so nöthig/ als das Com-

mercien-Wesen selbst/ welches ich bewiesen

mercien-Wesenselbst/ welchesich bewiesen-

habe/daß es zum Staat und Staats materi

habedaßes zumStaatundStaats mareri

gar nah gehöre / so haben Ihre Käyserl.

gar nah gehöre/ souhaben Ihre Kiyserl.

Majestät etwan vor zwey Jahren ein ab-

Mtjcstit etwan vor zwey Jahren ein äb-

sonderliches Collegium, und darüber Euer

sönderliches CollegiumGunddarüberEuer.

Höchgräfl. Excellentz zum Præsiden-

Höchgräf. Excellens zum Preüiden- .

ten Allergnädigst verordnet / und mit einer

ten Allergnädigstverorwdnet/ und miteiner

außführlicher instruction versehen / welche

außführlicher ivltruKtion versehenwelche

dahin zielet / daß man auf den Lauf der

dahin ziclet daß man auf den auf der

Commercien absonderlich Achtung gebe /

. Commercien absonderlich Achtunggebe/

damit er denen Käyserlichen Erblanden

damit er denen Kiyserlichen Erblanden

nicht zum Schaden / sondern Nutzen ge=

nicht zum Schadenz sondern Nusenges

reiche / in specie, daß / so viel möglich /

reiche/ inspcdie, daß so viel möglich/

allerhand manufacturen / vor welche nun

' allerhand manufacturen/ vor welche nun

viel Geld in die Frembde gehet / in den Käy=

viel GeldimdieFrembdcgchetimden Käy.

serl. Erblanden gearbeitet / und nicht allein

serl. Erblanden gearbeitetundnichtallein

das müssig gehende Volck in die Arbeit ge=

dasmiisßßg gehende Volckdie Arbeiegee'

stellet / das Geld im Land erhalten / sondern

sielletzdas Geld imlaidtrhaltenz söibern

auch

' auch

Zeilensegmentierung/Volltext (TIFF/Uw3):

Dedicatio.

De4ioatlo

können / ist in gleichem bekant und täglich

B6nnen/ iff ingfticien Fefant unbiiglicy

vor Augen/ wie aber und auf was Weiß/

vor2ngen/ nie aber unbauf mas TIBeif/

Weg und Mittel solches geschehe/ ist viel-

%3unb iittcipolieegdidepe/ ijt nel

leicht nicht jedem bekant/ noch so viel mir

i(cyt nict icbebefant/ nocy fo rieimir

bewußt von einigem Scribenten bißhero

beruft oon' tinigen Ccribente bifpero

entdecket und an Tag gegeben worden.

eurbedfet unb an SEag gegcbcn torbc,

Nach deme aber die eigentliche Erkäntnuß

hacyrene afe bie egejrliye erfintnuf

solcher Sachen so nöthig/ als das Com-

jolicr Ciacyen fo ndtpig/ afs ras Com-

mercien-Wesen selbst/ welches ich bewiesen

mwrcialBeen iitff/ioclcycsicybeniejen

habe/daß es zum Staat und Staats materi

hafebafiesginCitaatunbetaats materi

gar nah gehöre/ so haben Ihre Käyserl.

gar nap gcpdre/ fobabc 3preSipjei.

Majestät etwan vor zwey Jahren ein ab-

EDtajcjiit etian vor jep ;yapre en abe

sonderliches Collegium, und darüber Euer

fonberniciee Collegiuov,unb bariibcreuer

Höchgräfl. Excellentz zum Præsiden-

Decygrifi. Excellos aun Praiiden-

ten Allergnädigst verordnet/ und mit einer

ro 2Ilergndbigjfverorbnet/ unbiiteiiier

außführlicher instruction versehen / welche

aufifiifrlicyer inihcuctionoerjcpenArciciye

dahin zielet / daß man auf den Lauf der

bapin jielet 1 baf.man auf ben aauf be

Commercien absonderlich Achtung gebe /

Commercien abjoiiberlic 2cytnggebe/

damit er denen Käyserlichen Erblanden

bamit er benen Sdpjerlirfenn Crblanben

nicht zum Schaden / sondern Nutzen ge=

nictgun e7cabei/ fonber Rligeg

reiche / in specie, daß / so viel möglich /

rtice/ infpcsie, baf / fo niei mdglicy/

allerhand manufacturen / vor welche nun

allerpanb manufadturen/ noi teicpe nun

viel Geld in die Frembde gehet / in den Käy=

xieieeibiibiejframbbcgcyet/inben Sdps

serl. Erblanden gearbeitet / und nicht allein

baemiifjig geenbe 9Bolenibie 2rbeitge

das müssig gehende Volck in die Arbeit ge=

fiilttasieib inianberpaitr/ fonber

stellet / das Geld im Land erhalten / sondern

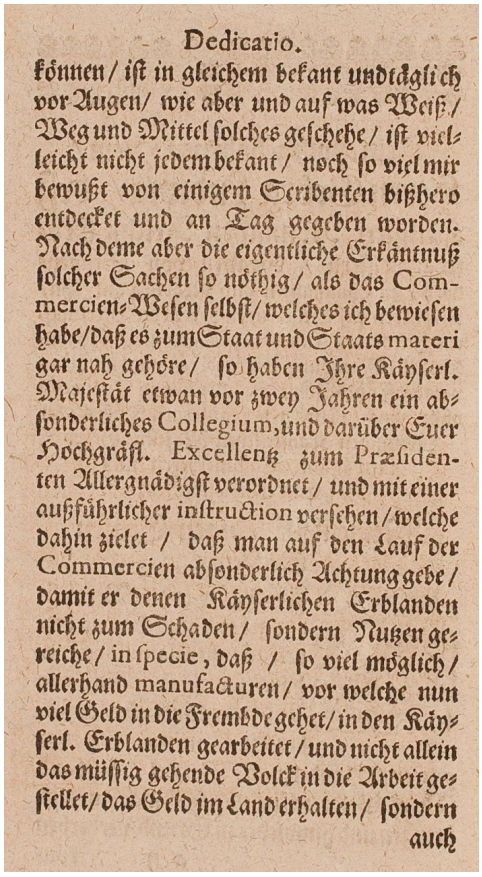
feri. Eriianben gearbeitet/unbnicyt afiein

auch

Auch

Ocropus 0.7 mit Sauvola-Binarsierung:

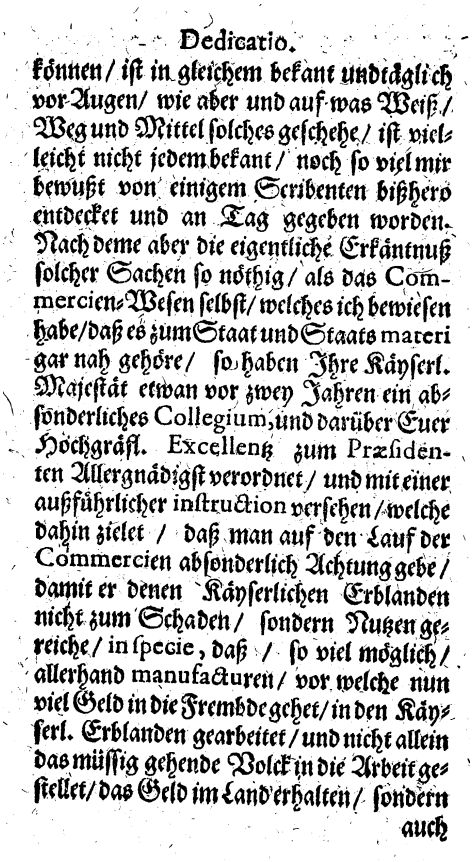
Original (JPEG):



Dedicatio.

können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist vielleicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Commercien-Wesen selbst/ welches ich bewiesen habe/daß es zum Staat und Staats materi gar nah gehöre / so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein absonderliches Collegium, und darüber Euer Hochgräfl. Excellentz zum Præsidenten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen gereiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käyserl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit gestellet / das Geld im Land erhalten / sondern auch

Sauvola Binarisierung (JPEG):



Dedicatio.

können / ist in gleichem bekant undtäglich vor Augen/ wie aber und auf was Weiß / Weg und Mittel solches geschehe / ist vielleicht nicht jedem bekant / noch so viel mir bewußt von einigem Scribenten bißhero entdecket und an Tag gegeben worden. Nach deme aber die eigentliche Erkäntnuß solcher Sachen so nöthig / als das Commercien-Wesen selbst/ welches ich bewiesen habe/daß es zum Staat und Staats materi gar nah gehöre / so haben Ihre Käyserl. Majestät etwan vor zwey Jahren ein absonderliches Collegium, und darüber Euer Hochgräfl. Excellentz zum Præsidenten Allergnädigst verordnet / und mit einer außführlicher instruction versehen / welche dahin zielet / daß man auf den Lauf der Commercien absonderlich Achtung gebe / damit er denen Käyserlichen Erblanden nicht zum Schaden / sondern Nutzen gereiche / in specie, daß / so viel möglich / allerhand manufacturen / vor welche nun viel Geld in die Frembde gehet / in den Käyserl. Erblanden gearbeitet / und nicht allein das müssig gehende Volck in die Arbeit gestellet / das Geld im Land erhalten / sondern auch

Zeilensegmentierung/Volltext (JPEG/Fraktur):

Dedicatio.

« - Deehaarlo.

können / ist in gleichem bekant undtäglich

könnenf ist in. gleßchem bekant undtäglich

vor Augen/ wie aber und auf was Weiß /

borAugen? wie aber und aufwas WeifF

Weg und Mittel solches geschehe / ist viel-

Wegund Msetelsolchesgeschehe? isk piel-

leicht nicht jedem bekant / noch so viel mir

leicht nscht jedembckant- noch so velmir

bewußt von einigem Scribenten bißhero

bewußt von' eingem Scrbenten bsßßerö

entdecket und an Tag gegeben worden.

entdecket und an Tag gegeben worden.

Nach deme aber die eigentliche Erkäntnuß

Nachdeme aber die eßgentliche Erkäntnuß

solcher Sachen so nöthig / als das Com-

solcher Sachen so nöthig? als das Cotn-

mercien-Wesen selbst / welches ich bewiesen

weroienWesenselbst? welches ich bewiesen

habe / daß es zum Staat und Staats materi

habe?daß eszumStaatundStaats mareri

gar nah gehöre / so haben Ihre Käyserl.

gar nah gehöre- so-ßaben Ißre Käyserl.

Majestät etwan vor zwey Jahren ein ab-

Majeskät etwan vor zwey Jahren ein abe

sonderliches Collegium, und darüber Euer

sonderliches Collegiuw.und darüber Euer

Höchgräfl. Excellentz zum Præsiden-

Höchgräfl. Hccelleny zum kraüen-

ten Allergnädigst verordnet / und mit einer

ten Allergnädßgstverordnctz und mitesner

außführlicher instruction versehen / welche

außfüßrlicher iwltruäionverschenfwelche

dahin zielet / daß man auf den Lauf der

dahsn zielet - daß man auf den Lauf der

Commercien absonderlich Achtung gebe /

Comroraien absonderlsch Achtunggebe?

damit er denen Käyſerlichen Erblanden

damit er denen Käpserlichen Erblanden

nicht zum Schaden / ſondern Nutzen ge-

nicht zum Schadenz sondern Nuzenge-

reiche / in ſpecie, daß / ſo viel möglich /

reichek in lgeaio, daß - so viel möglsch?

allerhand manufacturen / vor welche nun

allerhand manutaäturenz vor welche nun

viel Geld in die Frembde gehet / in den Käy-

vielGeld indieFrembdegchet-inden Kän-

ſerl. Erblanden gearbeitet / und nicht allein

erl. Erblanden gearbeitetzundnicht allein

das müſſig gehende Volck in die Arbeit ge-

dasmüüfßßg gehende Volckindie Arbeitges

ſtellet / das Geld im Land erhalten / ſondern

j.ellet-das Geld smTanderhaltenf sondern

auch

auch

Zeilensegmentierung/Volltext (JPEG/Frakant):

Dedicatio.

.- DöHicarlo .

können / iſt in gleichem bekant undtäglich

können/ ist in gleichem bekant undtdglich.

vor Augen / wie aber und auf was Weiß /

wor Augenu wie aber und aufF was Weß/

Weg und Mittel solches geschehe/ ist viel-

Wegund Mittelsolches geschche sk nmel-

leicht nicht jedem bekant/ noch so viel mir

leicht nicht jedembekant noch so vielmir

bewußt von einigem Scribenten bißhero

bewußt von einigem Scribenten bißhers

entdecket und an Tag gegeben worden.

entdecket und an Tag gegeben worden-

Nach deme aber die eigentliche Erkäntnuß

Nachdeme aber die eigentlich Erkäntnuß

solcher Sachen so nöthig/ als das Com-

solcher Sachen so nöthig/ als das Com-

mercien-Wesen selbst/ welches ich bewiesen

mercien-Wesenselbst/ welchesich bewiesen

habe/daß es zum Staat und Staats materi

habedaßes zumStaatund Staatsmareri

gar nah gehöre/ so haben Ihre Käyserl.

gar nah gehöre/ souhaben Ihre Kiyserl.

Majestät etwan vor zwey Jahren ein ab-

Majestit etwan vor zwey Jahren ein ab-

sonderliches Collegium, und darüber Euer

sönderliches CollegiumGunddarüber Euer

Höchgräfl. Excellentz zum Præsiden-

. Höchgräf. Excllens zum Preliden- .

ten Allergnädigst verordnet/ und mit einer

ten Allergnädigstverordnet/ undmiteiner

außführlicher instruction versehen/welche

außführlicher ivltruction versehen/welche

dahin zielet / daß man auf den Lauf der

dahin zielet daß man auf den auf der

Commercien absonderlich Achtung gebe /

. Commercien absonderlich Achtunggebe/

damit er denen Käyserlichen Erblanden

damit er denen Kiyserlichen Erblanden

nicht zum Schaden / sondern Nutzen ge=

nicht zum Schaden/ sondern Nusenges

reiche / in specie, daß / so viel möglich /

reiche/ inspcoia, daß / so viel möglich

allerhand manufacturen / vor welche nun

allerhand manufacturen/ vor welche nun

viel Geld in die Frembde gehet / in den Käy=

viel GeldindieFrembdegchetinden Käy-.

serl. Erblanden gearbeitet / und nicht allein

serl. Erblanden gearbeitetundnichtallein

das müssig gehende Volck in die Arbeit ge=

dasmiissg gehende Volckindie Arbeitgee'

stellet / das Geld im Land erhalten / sondern

sielletdas Geldimlanderhaltenn/ sondern

auch

auch

Zeilensegmentierung/Volltext (JPEG/Uw3):

Dedicatio.

DöHicarlo .

können / ist in gleichem bekant und täglich

Bdnne/ iff in giicpen iefant unbiiglic

vor Augen/ wie aber und auf was Weiß /

vor2ugen/ tie aber unb auf mas ilBeij/

Weg und Mittel solches geschehe / ist viel-

%Beunb ihiti(ofieegdidepe/ ijt nel

leicht nicht jedem bekant / noch so viel mir

l(ityt nicbt jcbembefant/ nocy fo rieimir

bewußt von einigem Scribenten bißhero

benuft oon' inigen Cicritenten bifpero

entdecket und an Tag gegeben worden.

etbecfet unb an SCag gegcbcn torten,

Nach deme aber die eigentliche Erkäntnuß

5hscybee ate bie egentlicye Crfdintnuf

solcher Sachen so nöthig / als das Com-

fofcer Cacycn fo ndtpig/ als bas Com-

mercien-Wesen selbst/ welches ich bewiesen

merciclBejen jlffi/toclyesicy beniejen

habe/daß es zum Staat und Staats materi

merciclBejen jlffi/toclyesicy beniejen

gar nah gehöre / so haben Ihre Käyserl.

gar nap gepdre/ fobafc 3oreSdpjei.

Majestät etwan vor zwey Jahren ein ab-

EDtajcjfit etian vor jnep japre en abe

sonderliches Collegium, und darüber Euer

fonberiitfes Collegiuo.unb bariibcr eurr

Hochgräfl. Excellentz zum Præsiden-

Decygrifi. Excellens gum Praiiden-

ten Allergnädigst verordnet / und mit einer

ton 2JIlcrgndbigfverorbnct/ unbniteiner

außführlicher instruction versehen / welche

aufififrlici)er inftcuctionocrjcpenAmcicoye

dahin zielet / daß man auf den Lauf der

bapin jielet 1 baff man auf ben aaufber'

Commercien absonderlich Achtung gebe /

Commcrcie abionberiicf 2cytnggebe/

damit er denen Käyserlichen Erblanden

bamit e benen Sipjerlirben dErblnnben

nicht zum Schaden / sondern Nutzen ge=

nicpt n Cicyabei/ fonbernn Thnengr

reiche / in specie, daß / so viel möglich /

rticye/ in fpcsic, baf 1 fo oiei mdglicy/

allerhand manufacturen / vor welche nun

allerfanb manufadturen/ roi teicye nuln

viel Geld in die Frembde gehet / in den Käy=

xieieeibinbicfjrambbcgcpet/inben Sips.

serl. Erblanden gearbeitet / und nicht allein

feri. Crifanben gearbeitet/unbnicf)taliein

das müssig gehende Volck in die Arbeit ge=

baemiiffig gcebe 9BofF(bie 2rbeitge4

stellet / das Geld im Land erhalten / sondern

ffciIct/baseeib inianberpaltrt/ fonbern

auch

auc

4. Abkürzungsverzeichnis

- ALTO – Analyzed Layout and Text Object, XML-Schema
- API – Programmierschnittstellen, Application Programming Interface
- BLOB – Binary Large Object
- DFG – Deutsche Forschungsgemeinschaft
- DFKI – Deutsche Forschungszentrum für Künstliche Intelligenz
- JPEG/JPG – Joint Photographic Experts Group
- GB – Gigabyte
- HAB – Herzog August Bibliothek Wolfenbüttel
- hOCR – Datenformat zur Repräsentation von Texterkennungsergebnissen
- HWR – Handwriting recognition
- HTML – Hypertext Markup Language
- IO – In-/Output; Ein-Ausgabe
- KB – Kilobyte
- MLP – Multi-layer-perceptrons, mehrlagiges Perzeptron
- METS – Metadata Encoding & Transmission Standard
- MB – Megabyte
- OAI – Open Archives Initiative
- OCR – Optical-Character-Recognition
- SDK – Software Development Kit
- TIFF – Tagged Image File Format
- TB – Terrabyte
- TEI – Text Encoding Initiative, hier: XML-Schema
- THELO – Thesaurus Locorum
- THEPRO – Thesaurus Professionum
- UTF – Unicode Transformation Format
- XML – Extensible Markup Language
- ZVDD – Zentrales Verzeichnis Digitalisierter Drucke

5. Abbildungsverzeichnis

Abbildung 1: Mögliche Konfigurationen eines Multi-Stage Kombinations-Systems.....14

Abbildung 2: Mögliche Konfigurationen eines Multi-Stage Kombinations-Systems.....21

Abbildung 3: Workflow in Korpus 1 entwickelt zur Überprüfung der Machbarkeit von Trainingsmodellen.....23

Abbildung 4: Übersicht des Zeitraums verfügbarer Drucke (blau) der HAB und Darstellung der Selektion der 100 Testkandidaten (rot).....24

Abbildung 5: Ausschnitt des Original-TIFF aus Signatur '223-2-quod-10s' (S.23) der Helmstedter Drucke aus der HAB in Wolfenbüttel.....26

Abbildung 6: Copyright-Hinweis der HAB in Wolfenbüttel. Verfügbar auf allen online verfügbaren Digitalisaten.....29

Abbildung 7: Markierung der Maßstäbe, die von Segmentierungs-Algorithmen über den Kontext herausgefiltert werden muss.....30

Abbildung 8: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation über die Länge der Texte.....30

Abbildung 9: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation per Google-Diff-Steps-Algorithmus.....31

Abbildung 10: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation per Levenshtein-Distanz.....31

Abbildung 11: Darstellung der OCR-Ergebnisse und deren Distanzen zu den Groundtruth-ähnlichen Texten im Rahmen der Evaluation per Needleman/Wunsch-Algorithmus.....32

Abbildung 12: Ausschnitt des Digitalisats „li-sbd-115-2s”. Das Wort ‘lsocrat‘ befindet sich an der Position x=1089 & y=346 und hat die Abmessungen Breite=308 & Höhe=105.. .33

Abbildung 13: Ausschnitt des von B.I.T.-Alpha generierten ALTO-XML (Positionen).....33

Abbildung 14: Ausschnitt des von B.I.T.-Alpha generierten ALTO-XML (Dateiname).....33

Abbildung 15: Gesamtübersicht der Erkennungsergebnisse des dritten Korpus.....38

Abbildung 16: Intrinsische Problem-Features in der Ergebnis-Übersicht.....39

Abbildung 17: Extrinsische Problem-Features in der Ergebnis-Übersicht.....39

Abbildung 18: Gesamtübersicht der OCR-Ergebnisse von JPEG- und TIFF-Vergleichen.....40

Abbildung 19: JPEG-Qualitäten und die intrinsischen/extrinsischen Ergebnisse der OCR-Verfahren.....41

Abbildung 20: TIFF-Qualitäten und die intrinsischen/extrinsischen Ergebnisse der OCR-Verfahren.....41

Abbildung 21: Layoutanalyse mit dem Algorithmus von Manmatha und Rothfeder.....58

6. Literaturverzeichnis

Angadi, S. A., and M. M. Kodabagi. 2014. „A Robust Segmentation Technique for Line, Word and Character Extraction from Kannada Text in Low Resolution Display Board Images". In: *Proceedings 2014 Fifth International Conference on signal and Image Processing (ICSIP 2014)*, ed. Juan Guerrero. 42-49. Piscataway, New Jersey: IEEE. Accessed August 30, 2016. doi: 10.1109/ICSIP.2014.11.

Ávila, B.T. and R. D. Lins. 2005. „A fast orientation and skew detection algorithm for monochromatic document images". In: *Proceedings of the 2005 ACM Symposium on Documemt Engineering*, ed. Anthony Wiley. 118-126. New York: Association for Computing Machinery (ACM). Accessed August 30, 2016. doi: 10.1145/1096601.1096631.

Biemann, C. and S. Teresniak, 2005. "Disentangling from Babylonian Confusion – Unsupervised Language Identification". In: *Computational Linguistics and Intelligent Text Processing*, ed. Alexander Gelbukh. Berlin, Heidelberg: Springer. Accessed April 06, 2014. http://www.springerprofessional.de/087---Tarsenglied-from-babylonian-confusion-unsupervised-language-identification/1229120.html.

Blanke, Tobias, Michael Bryant, and Marc Hedges. 2012. "Ocropodium: open source OCR for small-scale historical archives." Journal of Information Science 38 (1): 76–86. Accessed August 30, 2016. doi: 10.1177/0165551511429418.

Blanke, Tobias, Michael Bryant, and Mark Hedges. 2012. "Open source optical character recognition for historical research." Journal of Documentation 68 (5): 659–83. Accessed August 30, 2016. doi: 10.1108/00220411211256021.

Blanke, Tobias, Michael Bryant, and Reto Speck. 2012. "Information Extraction on Noisy Texts for Historical Research." In: *Digital Humanities 2012*, ed. Jan Christoph Meister. Hamburg: Hamburg University Press. 117-118. Accessed August 30, 2016. URN: urn:nbn:de:gbv:18-3-1257.

Breuel, Thomas. 2010. "The hOCR Embedded OCR Workflow and Output Format". Accessed August 30, 2016. https://docs.google.com/document/d/1QQnIQtvdAC_8n92-LhwPcjtAUFwBlzE8EWnKAxlgVf0/mobilebasic?pli=1&viewopt=127.

Breuel, T. M. 2003. „An algorithm for finding maximal whitespace rectangles at arbitrary orientations for document layout analysis". In: *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003)*. 66-70. Piscataway, New Jersey: IEEE. Accessed August 30, 2016. doi: 10.1109/ICDAR.2003.1227629.

Breuel, Thomas, Joost van Beusekom, and Faisal Shafait, eds. 2009. „Resolution Independent Skew and Orientation Detection for document images". In: *Proc. SPIE7247, Document Recognition and Retrieval XVI, 72470K*, eds. Kathrin Berkner, Laurence Likforman-Sulem. Bellingham: International Society for Optics and Photonics. Accessed August 30, 2016. doi: 10.1117/12.807735.

Breuel, Thomas. 2009. „Recent progress on the OCRopus OCR system". In: *MOCR '09 Proceedings of the International Workshop on Multilingual OCR*, ed. Venu Govindaraju. New

York: Association for Computing Machinery (ACM). Accessed August 30, 2016. doi: 10.1145/1577802.1577805.

Breuel, T. M. 2002. „Two geometric algorithms for layout analysis". In: *Document Analysis Systems V., 5th International Workshop, DAS 2002 Princeton, NJ, USA, August 19–21, 2002 Proceedings*, eds. Daniel Lopresti, Jianying Hu and Kashi Ramanusan. Berlin, Heidelberg: Springer. 188-199. Accessed August 30, 2016. doi: 10.1007/3-540-45869-7_23.

Breuel, Thomas. 2003. „High performance document layout analysis". In: *Proceedings of the Symposium on Document Image Understanding Technology (SDIUT'03)*. University of Maryland. 209-218. Accessed August 30, 2016. http://lamp.cfar.umd.edu/meetings/Proceedings /SDIUT03.pdf.

Breuel, Thomas M. 2008. „The OCRopus open source OCR system". In: *Proc. SPIE6815, Document Recognition and Retrieval XV, 68150F*, eds. Kathrin Berkner, Berrin A. Yanikoglu. Bellingham: International Society for Optics and Photonics. Accessed August 30, 2016. doi: 10.1117/12.783598.

Breuel, Thomas M. 2009. Training, Adaptation, and Semi-Supervised Learning in a Real-World OCR System. Kaiserslautern: Deutsches Forschungsinstutut für künstliche Intelligenz und Universität Kaiserslautern.

Buckley, Robert. 2013. Using lossy jpeg 2000 compression for archival master files 12 - Prepared for the Library of Congress Office of Strategic Initiatives, Version. Vol. 1.". Washington, DC.

Clausner, Christian, Stefan Pletschacher, and Apostolos Antonacopoulos. 2013. „The Significance of Reading Order in Document Recognition and Its Evaluation". In: *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013)*. Piscataway, New Jersey: IEEE. 688-693. Accessed August 30, 2016. doi: 10.1109/ICDAR.2013.141.

Compo, Gilbert P., Jeffrey S. Whitaker, and Prashant D. Sardeshmukh. 2006. "Feasibility of a 100-year reanalysis using only surface pressure data.". Colorado. Bulletin of the American Meteorological Society 87 (2): 175–90.

Coyle, Karen. 2006. "Mass digitization of books." Amsterdam. The Journal of Academic Librarianship 32 (6): 641–45.

De Castro Polastroa, Mateus, and Nalvo F. Almeidac. 2011. "OCR errors and their effects on computer forensics.". In: *ICoFCS – The Sixth International Conference on Forensic Computer Science, ICoFCS 2011*. Brasília: Associação Brasileira De Especialistas Em Alta Tecnologia (ABEAT). 115-121.

Déjean, Hervé, and Jean-Luc Meunier. 2009. "On tables of contents and how to recognize them." International Journal of Document Analysis and Recognition (IJDAR) 12 (1): 1–20. Heidelberg: Springer.

DFG. 2013. "Praxisregeln 'Digitalisierung'". http://www.dfg.de/formulare/12_151/12_151_de.pdf. Accessed April 06, 2014.

Dhiman, Shivani, and A. J. Singh. 2013. "Tesseract Vs Gocr A Comparative Study.". New York. International Journal of Recent Technology and Engineering (IJRTE) 2 (4): 80-83.

Does, Jesse de, and Katrien Depuydt. 2013. „Lexicon-supported OCR of eighteenth century Dutch books: a case study". In: *Proc. SPIE8658, Document Recognition and Retrieval XX, 86580L*, eds. Richard Zanibbi, Bertrand Coüasnon. Bellingham: International Society for Optics and Photonics. Accessed August 30, 2016. doi: 10.1117/12.2008423.

Dais, Dr. Siegfried, and Prof. Henning Kagerman. "Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0." Accessed April 05, 2014. http://www.forschungsunion.de/pdf/industrie_4_0_umsetzungsempfehlungen.pdf.

Eislöffel, J., and S. Röser. 1987. "Ein Astrometrie-Software-Paket für die PDS-1010A AM MPI für Astronomie." Mitteilungen der Astronomischen Gesellschaft 70: 296. Hamburg.

El Abed, Haikal, and Volker Märgner. 2007. „Reject rules and combination methods to improve Arabic handwritten word recognizers". In: Document Analysis and Recognition (ICDAR): 21. Curitiba.

Federbusch, Polzin, and Thomas Stäcker. 2013. "Volltext via OCR - Möglichkeiten und Grenzen. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz." Accessed April 06, 2014. http://www.personalschriften.de/aktuelles/artikelansicht/details/volltext-via-ocr-moeglichkeiten-und-grenzen.html.

Friske, Mario. 2013. "Fehlermanagement in Großprojekten–Erfahrungen und Best Practices." Siegen. Softwaretechnik-Trends 33 (4).

Heliński, Marcin, Miłosz Kmieciak, and Tomasz Parkoła. 2012. Report on the comparison of Tesseract and ABBYY FineReader OCR engines. Den Haag: Project IMPACT.

Jannidis, Fotis and Christof Schöch. 2013. "Quantitative Text Analysis for Literary History". DARIAH-DE Working Papers Nr. 2. Göttingen: DARIAH-DE. Accessed August 30, 2016. URN: urn:nbn:de:gbv:7-dariah-2016-1-2.

Konya, Iuliu Vasile, Stefan Eickeler, and Christoph Seibert. 2010. „Fast seamless skew and orientation detection in document images". In: 2010, 20th I*nternational Conference on Pattern Recognition (ICPR 2010)*. Piscataway, New Jersey: IEEE. 1924-1928. Accessed August 30, 2016. doi: 10.1109/ICPR.2010.474.

Konya, Iuliu Vasile, Christoph Seibert, Stefan Eickeler, and Sebastian Glahn. 2009. „Constant-time locally optimal adaptive binarization". In: *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*. Piscataway, New Jersey: IEEE. 738-742. doi: 10.1109/ICDAR.2009.105.

Konya, Iuliu Vasile. 2013. Adaptive Methods for Robust Document Image Understanding. Dissertation, Bonn: Universität Bonn.

Lidke, Jakub, Christian Thurau, and Christian Bauckhage. 2010. „The Snippet Statistics of Font Recognition". In: 2010, 20th I*nternational Conference on Pattern Recognition (ICPR 2010).*

Piscataway, New Jersey: IEEE. 1868-1871. Accessed August 30, 2016. doi: 10.1109/ICPR.2010.461.

Lin, Xiaofan. 2001. „Reliable OCR solution for digital content re-mastering". In: *Proc. SPIE 4670, Document Recognition and Retrieval IX, 223*, eds. Paul B. Kantor, Tapas Kanungo, and Jiangying Zhou. Bellingham: International Society for Optics and Photonics. Accessed August 30, 2016. doi: 10.1117/12.450731.

Liu, Mohan, Iuliu Konya, Jan Nandzik, Nicolas Flores-Herr, Stefan Eickeler, and Patrick Ndjiki-Nya. 2012. "A new quality assessment and improvement system for print media." EURASIP Journal on Advances in Signal Processing 2012 (1): 1–17. Heidelberg: Springer.

Lopresti, Daniel, ed. 2008. Measuring the impact of character recognition errors on downstream text analysis. In: Proc. *SPIE 6815, Document Recognition and Retrieval XV, 68150G*, eds. Kathrin Berkner and Berrin A. Yanikoglu. Bellingham: International Society for Optics and Photonics. Accessed August 30, 2016. doi: 10.1117/12.767131.

Manmatha, R., and Jamie L. Rothfeder. 2005. "A scale space approach for automatically segmenting words from historical handwritten documents." Piscataway, New Jersey: IEEE. Transactions on Pattern Analysis and Machine Intelligence 27 (8): 1212–25. Accessed August 30, 2016. doi: 10.1109/TPAMI.2005.150.

Moyle, Martin, Justin Tonra, and Valerie Wallace. 2011. "Manuscript transcription by crowdsourcing: Transcribe Bentham." Den Haag. Liber Quarterly 20 (3/4): 347–56. Accessed August 30, 2016. doi: 10.18352/lq.7999.

Pfleger, Regina. 2014. "„Zählen, wiegen, messen" – IT-Transformationen erfolgreich steuern." Heidelberg: Springer. HMD Praxis der Wirtschaftsinformatik 51 (2): 164–74. Accessed August 30, 2016. doi: 10.1365/s40702-014-0016-1.

Pratikakis, Gatos, and Ntirogiannis. 2013. „ICDAR 2013 document image binarization contest (DIBCO 2013)". In: 2013, *12th International Conference on Document Analysis and Recognition (ICDAR 2013)*. Piscataway, New Jersey: IEEE. 1471-1476. Accessed August 30, 2016. doi: 10.1109/ICDAR.2013.219.

Ratinov, Lev, Dan Roth, Doug Downey, and Mike Anderson. 2011. „Local and global algorithms for disambiguation to wikipedia" In: HLT'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies 1 : 1375-1384. Stroudsburg: Association for Computational Linguistics.

Rosey Holley. 2009. „How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". D-Lib Magazine 15 (3/4). Reston, Virginia, U.S.A : Corporation for National Research Initiatives. Accessed August 30, 2016. http://www.dlib.org/dlib/march09/holley/03holley.html.

Sadri, J, and Cheriet, eds. 2009. A New Approach for Skew Correction of Documents Based on Particle Swarm Optimization. In: *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*. Piscataway, New Jersey: IEEE. 1066-1070. Accessed August 30, 2016. doi: 10.1109/ICDAR.2009.268.

Sayre, Kenneth M. 1973. "Machine recognition of handwritten words: A project report." Pattern recognition 5 (3): 213–28. Amsterdam: Elsevier.

Schlarb, Sven, and Clemens Neudecker. 2012. „A Heuristic Measure for Detecting Influence of Lossy JP2 Compression on Optical Character Recognition in the Absence of Ground Truth". In: *Archiving 2012 : preservation strategies and imaging technologies for cultural heritage institutions and memory organizations, June 12 -15, 2012, Copenhagen, Denmark, final programm and proceedings*. Springfield, Virginia: Society for Imaging Science and Technology. 250-254.

Sezgin, Mehmet. 2004. "Survey over image thresholding techniques and quantitative performance evaluation." Journal of Electronic imaging 13 (1): 146–68.

Shafait, Faisal, Daniel Keysers, and Thomas M. Breuel. 2008. "Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms." IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6): 941–54. Piscataway, New Jersey: IEEE. Accessed August, 30, 2016. doi: 10.1109/TPAMI.2007.70837.

Sharma, Om P., M. K. Ghose, Krishna B. Shah, and Benoy K. Thakur. 2013. "Recent trends and tools for feature extraction in OCR technology". International Journal of Soft Computing and Engineering (IJSCE) 2. Bhopal, Madhya Pradesh.

Smith, Ray. 2007. „An Overview of the Tesseract OCR Engine". In: *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Volume 2*. Piscataway, New Jersey: IEEE. 629-633. Accessed August 30, 2016. doi: 10.1109/ICDAR.2007.4376991.

Smith, Raymond W.. 2009. „Hybrid Page Layout Analysis via Tab-Stop Detection". In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009). Piscataway, New Jersey: IEEE. 241-245. Accessed August 30, 2016. doi: 10.1109/ICDAR.2009.257.

Stäcker, Thomas. 2006. "Das Zentrale Verzeichnis Digitalisierter Drucke (zvdd)-Ein Baustein zur Digitalisierung des gedruckten kulturellen Erbes." AKMB-news: Informationen zu Kunst, Museum und Bibliothek 12 (1): 32–34. Heidelberg.

Still Image Working Group. 2010. "Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files. Federal Agencies Digitization Guidelines Initiative (FADGI)." Accessed April 06, 2014. http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf.

Strohmaier, Christian M. 2005. Methoden der lexikalischen Nachkorrektur OCR-erfasster Dokumente. Dissertation, München: LMU Fakultät für Sprach- und Literaturwissenschaften. Accessed August 30, 2016. URN: urn:nbn:de:bvb:19-36743.

Torabi, Katayoun, Jessica Durgan, and Bryan Tarpley. 2013. „Early modern OCR project (eMOP) at Texas A&M University: using Aletheia to train Tesseract". In: *DocEng'13 Proceedings of the 2013 ACM symposium on Document engineering*. New York: Association for Computing Machinery (ACM).

Virtual Identity AG - Deutsche Nationalbibliothek. 2014. "Ein Portal für Kultur und Wissenschaft: Netzwerk, Suchmaschine, Gedächtnis? - Anforderungen und Erwartungen an die „Deutsche Digitale Bibliothek"". Accessed August 30, 2016. https://www.deutsche-digitale-bibliothek.de/static/de/sc_documents/div/nutzerstudie.pdf.

Volk, Martin, Torsten Marek, and Rico Sennrich. 2010. „Reducing OCR errors by combining two OCR systems". In: *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), August 16, 2010, Lisbon, Portugal*. Lisbon: University of Lisbon. 61-65. Accessed August 30, 2016. http://ilk.uvt.nl/LaTeCH2010/LPF/ws16.pdf.

Weymann, Anna, and Ibero-Amerikanisches Insitut. 2010. Einführung in die Digitalisierung von gedrucktem Kulturgut. Berlin: Ibero-Amerikanisches Institut, Stiftung Preußischer Kulturbesitz.

7. Webreferenzen

- http://www.monasterium.net
- http://www.numeric.ws
- http://nestor.sub.uni-goettingen.de
- http://www.jstor.org
- http://www.goethe.de/wis/bib/dib/de3616481.htm
- http://www.deutschestextarchiv.de
- http://www.impact-project.eu
- http://www.succeed-project.eu
- http://emop.tamu.edu
- http://www.deutsches-textarchiv.de
- http://chroniclingamerica.loc.gov/ocr
- http://eebo.chadwyck.com
- http://corpusdelespanol.org/x.asp
- http://portal.clarin.inl.nl/vu-dnc
- http://frakturschrift.de/de:start
- http://bit.dyndns.biz
- http://code.google.com/p/tesseract-ocr
- http://code.google.com/p/ocropus
- http://arstechnica.com/information-technology/2007/10/hands-on-with-googles-ocropus-open-source- scanning-software
- http://docs.google.com/document/d/1QQnIQtvdAC_8n92-LhwPcjtAUFwBlzE8EWnKAxlgVf0/preview
- http://dariah.hki.uni-koeln.de
- http://www.gesamtkatalogderwiegendrucke.de/
- http://idhmcmain.tamu.edu/projects/Mellon/Gantt.pdf
- http://bit.dyndns.biz/
- http://mike2.openmethodology.org/wiki/Big_Data_Definition
- http://impactocr.wordpress.com/2010/05/07/an-overview-of-technical-solutions-in-impact

8. Anhang

Korpus 1 (500 Seiten aus 100 Signaturen der Helmstedter Drucke der HAB)

Jahr	Signatur	Seitenanzahl	Druckort	Selektierte Seiten
1659	1096-helmst-dr-10s	23	Helmstedt	00008;00013;00014;00018;00021
1667	123-helmst-dr-16s	12	Helmstedt	00004;00005;00006;00007;00008
1667	126-helmst-dr-4s	12	Helmstedt	00003;00004;00005;00006;00007
1670	126-helmst-dr-9s	12	Helmstedt	00003;00004;00005;00006;00008
1663	131-helmst-dr-65s	12	Helmstedt	00003;00004;00006;00007;00008
1665	131-helmst-dr-76s	12	Helmstedt	00003;00004;00005;00006;00008
1666	131-helmst-dr-79s	12	Helmstedt	00003;00004;00005;00006;00007
1669	136-helmst-dr-28s	12	Helmstedt	00003;00004;00005;00006;00007
1666	136-helmst-dr-37s	12	Helmstedt	00004;00007;00008;00005;00006
1662	137-helmst-dr-7s	52	n/a	00008;00016;00024;00028;00046
1672	167-6-hist	436	Helmstedt	00032;00134;00314;00326;00426
1666	172-helmst-dr-1s	12	Helmstedt	00003;00004;00006;00007;00008
1660	199-helmst-dr-1s	48	Helmstedt	00011;00017;00025;00030;00038
1649	223-2-quod-10s	52	Helmstedt	00003;00013;00021;00023;00045
1649	223-2-quod-13s	32	Helmstedt	00006;00009;00022;00024;00027
1650	223-2-quod-23s	40	Helmstedt	00006;00024;00026;00031;00032
1664	26-helmst-dr-8s	8	Helmstedt	00003;00004;00005;00007;00008
1656	30-helmst-dr-21s	8	Helmstedt	00003;00004;00008;00006;00007
1665	30-helmst-dr-36s	8	Helmstedt	00003;00004;00006;00007;00008
1662	32-4-phys-10s	28	Helmstedt	00011;00015;00019;00020;00024
1665	32-4-phys-18s	44	Helmstedt	00010;00015;00016;00027;00034
1663	32-4-phys-31s	28	Helmstedt	00006;00010;00013;00018;00023
1661	32-4-phys-32s	32	Helmstedt	00005;00007;00019;00024;00025
1665	35-helmst-dr-21s	16	Helmstedt	00005;00006;00009;00011;00013
1662	401-helmst-dr	238	Helmstedt	00019;00122;00165;00167;00206
1659	435-16-2-theol	398	Helmstedt	00118;00184;00259;00353;00378
1636	459-6-theol-7s	60	Helmstedt	00013;00033;00037;00046;00056
1648	531-helmst-dr-15s	156	Helmstedt	00052;00068;00107;00125;00132
1647	531-helmst-dr-7s	60	Helmstedt	00022;00024;00040;00048;00050

Jahr	Signatur	Seitenanzahl	Druckort	Selektierte Seiten
1648	531-helmst-dr-9s	44	Helmstedt	00005;00015;00022;00033;00038
1640	536-helmst-dr-10s	16	Helmstedt	00004;00008;00009;00010;00012
1641	536-helmst-dr-14s	24	Helmstedt	00005;00007;00008;00009;00022
1643	536-helmst-dr-27s	20	Helmstedt	00004;00005;00010;00016;00017
1668	539-helmst-dr-23s	36	Helmstedt	00005;00006;00010;00011;00018
1661	716-helmst-dr-6s	36	Helmstedt	00020;00021;00024;00032;00033
1662	738-helmst-dr-2s	54	Helmstedt	00024;00040;00042;00032;00038
1646	772-helmst-dr-3s	60	Helmstedt	00006;00013;00022;00042;00045
1653	862-helmst-dr-1s	88	Helmstedt	00004;00007;00008;00017;00039
1669	920-helmst-dr-12s	72	Helmstedt	00018;00024;00025;00051;00057
1665	923-helmst-dr-1s	34	Helmstedt	00005;00010;00030;00040;00046
1666	926-helmst-dr-9	36	n/a	00008;00014;00017;00024;00029
1655	936-helmst-dr-17s	28	Helmstedt	00007;00014;00018;00020;00021
1645	936-helmst-dr-1s	38	Helmstedt	00006;00011;00025;00030;00031
1658	936-helmst-dr-23s	68	Helmstedt	00034;00036;00044;00047;00058
1658	936-helmst-dr-26s	44	Helmstedt	00005;00015;00021;00028;00037
1645	936-helmst-dr-3s	48	Helmstedt	00009;00013;00016;00035;00038
1647	936-helmst-dr-6s	28	Helmstedt	00006;00010;00013;00017;00019
1644	953-helmst-dr-3s	48	Helmstedt	00004;00022;00032;00036;00044
1661	alv-t-140-5s	80	Helmstedt	00024;00035;00036;00049;00062
1668	alv-x-150-94s	36	Helmstedt	00007;00012;00015;00018;00020
1651	g-180-4f-helmst-19s	28	Helmstedt	00005;00009;00018;00019;00023
1669	g-180-4f-helmst-8s	176	Helmstedt	00009;00040;00068;00072;00102
1663	g-181-4f-helmst-18s	20	Helmstedt	00003;00005;00007;00011;00012
1661	g-181-4f-helmst-21s	28	Helmstedt	00007;00009;00012;00014;00018
1663	g-181-4f-helmst-27s	44	n/a	00005;00007;00015;00020;00029
1660	g-181-4f-helmst-28s	28	Helmstedt	00004;00008;00010;00013;00014
1662	g-181-4f-helmst-5s	32	Helmstedt	00005;00012;00016;00018;00019
1664	g-182-4f-helmst-6s	130	Helmstedt	00010;00023;00068;00083;00106
1668	g-184-4f-helmst-2-10s	20	Helmstedt	00003;00006;00012;00013;00014
1676	g-184-4f-helmst-2-3s	20	n/a	00004;00006;00008;00013;00016
1668	g-184-4f-helmst-2-4s	20	Helmstedt	00004;00006;00007;00014;00016
1652	g-187-4f-helmst-3s	84	Helmstedt	00009;00021;00026;00047;00052
1656	g-187-4f-helmst-7s	52	Helmstedt	00024;00025;00038;00041;00048
1660	gl-1130	96	Helmstedt	00028;00058;00080;00085;00088
1608	gn-4760-1b	968	Helmstedt	00018;00023;00029;00034;00040
1654	h-304-4f-helmst-3s	124	n/a	00021;00079;00091;00099;00109
1654	lg-1877	632	Helmstedt	00062;00146;00313;00502;00598
1669	li-1532	552	Helmstedt	00125;00302;00394;00450;00467
1697	li-2133	1012	Helmstedt	00187;00202;00661;00716;00838

Jahr	Signatur	Seitenanzahl	Druckort	Selektierte Seiten
1697	li-2133	1012	Helmstedt	00187;00202;00661;00716;00838
1662	li-sbd-115-2s	40	Helmstedt	00008;00009;00012;00013;00033
1666	li-sbd-231-5s	190	Helmstedt	00004;00012;00013;00022;00026
1661	li-sbd-240-2s	280	n/a	00068;00090;00102;00219;00239
1661	n-218c-4f-helmst-11s	12	n/a	00002;00003;00005;00007;00008
1654	p-529-4f-helmst-3s	144	Helmstedt	00026;00065;00079;00081;00114
1668	quh-104-6-5s	196	n/a	00018;00097;00108;00148;00159
1650	qun-166-7-14s	276	Helmstedt	00014;00109;00211;00222;00242
1633	qun-166-7-1s	60	Helmstedt	00006;00023;00026;00037;00044
1639	qun-167-11-7-3s	124	Helmstedt	00009;00034;00078;00088;00112
1654	qun-167-13-25s	110	Helmstedt	00022;00043;00081;00045;00074
1651	qun-167-5-19s	44	Helmstedt	00005;00008;00022;00023;00026
1657	qun-167-5-2-36s	12	Helmstedt	00003;00004;00005;00007;00008
1657	qun-167-8-2s	36	Helmstedt	00003;00006;00019;00029;00031
1652	qun-167-8-4s	40	Helmstedt	00009;00014;00023;00024;00031
1655	qun-167-8-5s	86	Helmstedt	00016;00042;00055;00057;00061
1652	qun-167-8-7s	20	Helmstedt	00003;00004;00006;00007;00011
1671	qun-225-11s	36	Helmstedt	00006;00014;00017;00018;00020
1667	qun-23-9s	32	Helmstedt	00004;00005;00018;00006;00010
1643	qun-299-2-24s	56	Helmstedt	00010;00013;00023;00024;00034
1666	qun-697-2s	100	Magdeburg; Helmstedt	00028;00037;00038;00066;00084
1659	rd-205-2b	1276	Helmstedt	00436;00599;00761;00842;01010
1656	re-240-1s	788	Helmstedt	00388;00538;00623;00668;00752
1671	re-65-5s	50	Helmstedt	00004;00014;00016;00017;00019
1664	sch-t-76-22s	12	Helmstedt	00003;00004;00005;00007;00008
1671	t-853-2f-helmst-2s	44	Helmstedt	00013;00016;00022;00033;00037
1760	to-279	870	Halle;Saale; Helmstedt	00013;00016;00022;00033;00037
1661	xb-1431	100	Helmstedt	00012;00046;00055;00072;00079
1698	xd-4f-30	1960	Jena	00305;00446;01334;01646;01654
1672	xd-8f-627	576	n/a	00056;00116;00125;00237;00315
1672	yx-29-8f-helmst-21s	28	Helmstedt	00003;00017;00020;00021;00023
1653	yx-59-8f-helmst-4s	224	Helmstedt	00017;00102;00139;00180;00185

Korpus 2 (~2900 Seiten aus 10 Signaturen der BSB aus dem Google-Projekt)

Signatur BSB	Titel	bin	Seiten	Jahr	Verlag	Ort	Autor/Hrsg.	Margi-Nalien	Größe	MB/Bild	Resolution
bsb10139422	En Damvs Diodori Sicvli Historici Graeci, Qvae Nvnc Qvidem Extare Noscvntvr Opera : Nempe, De Illustrium Regum Philippi, & Alexandri ... praeclarè factis, Bartholomaeo Cospo Bononiensi interprete, Lib. II. De fabulosis Ioannis Monachi, ex libris historiarum suarum de uita Alexãdri ab eodem Bartholomaeo uersum, Lib. I. Aegyptiorum gestis ... à Pogio Florentino latinitate donatos, Lib. VI. Bocatij aliquot insignium foeminarum ... Historias, ex eiusdem Lib. X.	Nein	398	1531	Petri	Basel	Diodorus, Siculus; Boccaccio, Giovanni; Johannes Zonaras	ja	9000	22.61	2300*3266
bsb10157993	Oratio de studio sacrarum literarum : in Academia Lipsensi recitata a Jacobo Andreae XX. Novembris Anno 1576	Nein	25	1577	Gruppenbachius	Tubingae	Andreä, Jacob	ja	323	11.53	1700*2200
bsb10160279	Hyperaspistes Tōn Prolegomenōn Ioannis Brentii, In Apologam Confessionis Ill. Ducis Vuirtenbergensis	Ja	192	1558		Francoforti	Andreä, Jacob		766	3.98	1685*2306
bsb10165677	Poetae christiani veteres	Ja	152	1502	Aldus Manutius	Venedig			568	3.73	1656*2314
bsb10168344	Friderici Nauseae Blancicampiani, Ecclesiastae Moguntini, Miscellanear[um] libri duo : quoru[m] prior est pro Horis Canonicis, alter vero pro sacrosancta Missa Apologeticus	Ja	216	1531	Quentell	Coloniae	Nausea, Friedrich	ja	872	4.03	1700*2323
bsb10171607	M. T. Ciceronis Orationvm Volvmina Tria	Nein	712	1574	Rihelius	Argentorati	Cicero, Marcus Tullius		5300	7.44	1293*1958
bsb10175361	Brevis Et Catholica Symboli Apostolici Explicatio	Nein	144	1566	Plantin	Antverpiae	Sturm, Johannes	ja	953	6.61	1139*1888
bsb10178305	Princeps Iuventutis Panegyricus Ferdinando Austriaco Dicatus	Ja	275	1557	Oporinus	Basileae	Herold, Johannes		674	2.45	1240*1960
bsb10187183	Elvcidatio In Confiteor Gvillelmi Pepin : In Qua Qvaecvnque Ad Confessionem, ad Peccata Cordis, Oris, Operis, & Omissionis ... accuratè explicantur ... Cvm Indice Locvpletissimo	Nein	760	1588	Bertani	Venetiis	Pépin, Guillaume		5200	2.45	1200*1860
bsb10203117	Elegiae duae Academiae Viennens. ad Maximilianum Bohemorum regem	Nein	44	1552		Viennae	Corsi, Pietro ;Aldus, David		471	10.70	1600*2210

Korpus 3 (76 Seiten zur qualitativen Untersuchung von Problem-Features aus dem Bestand der HAB und der SUB)

- 107-40-eth Seiten: 0071 0074
- 1224-20-theol Seiten: 0080 0081 0127
- 134-13-phys Seiten: 0017 0021
- 196-3-hist-2s Seite: 0131
- 22-musica-helmst-4s Seite: 0077
- 263-1-hist-2f Seiten: 0052 0054
- 32-1-geom-2f-4 Seiten: 0040 0073
- 48-6-pol-1s Seiten: 0052 0056
- 83-1-quod-3s Seiten: 0055 0128
- SUB Seite: 01-8 H GERM VIII, 681-5-00000056
- SUB Seite: 02-8 P GERM II, 8294 RARA-00000109 00000192 00000203
- SUB Seite: 05-8 ARS MIL 1106 3 (1)-00000006 00000017
- SUB Seite: 07-8 FAB II, 617 RARA-00000133
- SUB Seite: 08-8 FAB VI 1909-00000062 00000270
- SUB Seite: 10-8 H GERM VIII, 681-2-3-00000346 00000386
- SUB Seite: 12-8 H GERM VIII, 681-5-00000056
- SUB Seite: 13-8 MATH I. 4052 13 28
- SUB Seite: 15-8 P GERM II. 8365-00000076
- ac-95 Seiten: 0072 0082
- g-717-4f-helmst-6s Seiten: 0062 0076
- ga-67 Seite: 0128
- gl-488 Seiten: 0017 0095
- gm-2093 Seite: 0116
- gn-10332 Seiten: 0149 0169 0171
- jb-4f-24-2b-2s Seiten: 0019 0021
- lo-29 Seiten: 0014 0019
- lo-6724 Seiten: 0051 0139
- lo-677-5-1 Seite: 0518
- lo-677-6 Seiten: 0016 0017
- lo-7872 Seiten: 0059 0237
- lo-8303 Seiten: 0056 0122
- od-215 Seiten: 0079 0097
- s-88-2f-helmst Seite: 0042
- sc-9 Seiten: 0024 0027 0077
- sf-3 Seite: 0086
- te-60-2s Seiten: 0052 0056
- th-160 Seite: 0736
- vb-638 Seite: 0129
- wa-1284-1 Seiten: 0086 0224
- wa-6098 Seiten: 0020 0038
- xb-4f-111 Seiten: 0068 0115
- xb-6428 Seiten: 0090 0091
- xb-6616-2s Seiten: 0038