Aufsätze

Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift *Die Grenzboten*

Manfred Nölte, Staats- und Universitätsbibliothek Bremen
Jan-Paul Bultmann, Universität Bremen
Maik Schünemann, Universität Bremen
Martin Blenkle, Staats- und Universitätsbibliothek Bremen

Zusammenfassung:
Den Geisteswissenschaften stehen nach und nach mehr computerbasierte Werkzeuge und Infrastrukturen der Digital Humanities zur Verfügung, für die die Existenz und weitere Erstellung von Volltext mit guter Qualität eine unabdingbare Voraussetzung ist. Der Bedarf nach qualitativ hochwertigem Volltext aus Retrodigitalisierungsprojekten steigt daher ständig an. Der zu Frakturschrift berechnete OCR-Volltext hat eine deutlich schlechtere Qualität als von Antiqua-Schrift berechneter. Daher ist für das wissenschaftliche Arbeiten unkorrigierter und unstrukturierter OCR-Volltext von Frakturschrift häufig wertlos. Da eine bedarfsgerechte Erzeugung von Volltext in der Größenordnung von mehreren Millionen Seiten in Bezug auf Aufwand und Kosten effizient sein sollte, wird hier eine möglichst weitgehende Automatisierung der Nachbearbeitung von OCR-Volltext vorgestellt.
An der Staats- und Universitätsbibliothek Bremen (SuUB) wurde dazu ein Ansatz entwickelt, der sich durch Einfachheit auszeichnet: Eine Liste historischer bzw. dialekt- oder fachspezifischer Wortformen – eine der Voraussetzungen dieses Ansatzes – ist verhältnismäßig leicht erstellbar. Ein effizienter Algorithmus leistet den Abgleich von hier ca. 1,7 Millionen Wortformen gegen bei der Zeitschrift *Die Grenzboten* knapp 80 Millionen enthaltenen Wörtern und lässt sich auf verständliche und nachvollziehbare Art und Weise parametrisieren, d.h. auf die spezifischen Eigenschaften des jeweiligen Volltextprojektes einstellen. Die erreichbaren Ergebnisse sind stark abhängig von der Ausgangsqualität des Volltextes sowie von dem Umfang und der Qualität der Liste der historischen Wortformen und dem verwendeten Fehlermodell. So können beispielsweise bestimmte Fehler nur mit einem den Kontext berücksichtigenden Ansatz korrigiert werden. Weiterhin wurde zusammen mit der Firma ProjectComputing mit Sitz in Canberra, Australien, der cloud service overProof[1] um die Funktionalität der Nachkorrektur deutschsprachiger Frakturschrift erweitert.
In einem Ausblick werden Bedarfe und Möglichkeiten für die Zukunft aufgezeigt.

Summary:
Gradually, the humanities are provided with a number of computer based tools and scientific infrastructures of the digital humanities. As digital full text is strongly needed for these tools and infrastructures, the demand for high-quality full texts is constantly rising. OCRed full text from Gothic typeface texts is of considerably worse quality than OCRed full text from Antiqua. The value

1 OverProof, zuletzt geprüft am 01.02.2016, http://overproof.projectcomputing.com/.

of uncorrected and unstructured OCR full text is fairly low. As multiple millions of pages need to be processed, the method should be efficient with respect to expenditure and costs. Therefore, we introduce an almost fully automated approach for the post correction of OCR full text. The approach developed at the Staats- und Universitätsbibliothek Bremen (SuUB) is a straightforward one. One of the requirements, a list of historical word forms, was easily generated. An efficient algorithm carries out the matching of 1,7 million word forms against almost 80 million words taken from the historical journal *Die Grenzboten*. The parametrization of the algorithm, i.e. the adaption to the specific requirements of the full text project, is comprehensible and easy to understand. The results which can be achieved strongly depend on the initial quality of the full text, the dimension and quality of the list of historical word forms and the error model applied. For example, specific types of errors can only be corrected by taking context information into account. Furthermore, the cloud service overProof was enhanced by the ability to correct German Gothic typeset. This was done in a cooperation with the Australian company ProjectComputing.
In the discussion, requirements and options for the future are presented.

Zitierfähiger Link (DOI): http://dx.doi.org/10.5282/o-bib/2016H1S32-55
Autorenidentifikation: Nölte, Manfred GND 124205860; Bultmann, Jan-Paul GND 1081923962; Schünemann, Maik GND 1081924071; Blenkle, Martin GND 172847575
Schlagwörter: Digitalisierung; Retrodigitalisierung; OCR

1. Einleitung

Allen Geisteswissenschaftlern stehen nach und nach mehr computerbasierte Analysemöglichkeiten für Textkorpora und Infrastrukturen der Digital Humanities zur Verfügung, für die die Existenz und weitere Erstellung von digitalem Volltext mit guter Qualität eine unabdingbare Voraussetzung ist.[2] Der Bedarf nach qualitativ hochwertigem Volltext aus Retrodigitalisierungsprojekten steigt daher ständig an, was sich bereits in verschiedenen Projekten widerspiegelt.[3] Dies gilt nicht nur für Historiker, Computerphilologen und Sprachwissenschaftler. Insbesondere die jüngeren Wissenschaftlergenerationen haben eine hohe Affinität für digitale Inhalte.

Aus Kostengründen wird in Digitalisierungsprojekten häufig auf eine Volltexterfassung durch Doublekeying verzichtet und stattdessen die *Optical Character Recognition* (OCR) eingesetzt. Der Einsatz

2 Thomas Stäcker, „Konversion des kulturellen Erbes für die Forschung: Volltextbeschaffung und -bereitstellung als Aufgabe der Bibliotheken“, *o-bib* 1, Nr. 1 (2014): 223, http://dx.doi.org/10.5282/o-bib/2014H1S220-237.

3 John Evershed und Kent Fitch, „Correcting Noisy OCR: Context Beats Confusion“ in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (New York: ACM, 2014), 45–51, http://dx.doi.org/10.1145/2595188.2595200; Maria Federbusch und Christian Polzin, *Volltext via OCR – Möglichkeiten und Grenzen*, Beiträge aus der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz (Berlin: Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, 2013), zuletzt geprüft am 01.02.2016, http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf; Lenz Furrer und Martin Volk, „Reducing OCR Errors in Gothic-Script Documents“, *ERCIM News* 86 (2011): 29–30, http://dx.doi.org/10.5167/uzh-49203; Günter Mühlberger, „Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Strukturerkennung (OCR)“, *Zeitschrift für Bibliothekswesen und Bibliographie* 58, Nr. 1 (2011): 10–18, http://dx.doi.org/10.3196/186429501158135.

von OCR zu Frakturschrift ist im Vergleich zur Antiqua-Schrift deutlich problematischer.[4] Für das wissenschaftliche Arbeiten ist diese, insbesondere bei der OCR von Frakturschrift, häufig wertlos. Ebenfalls sollte eine bedarfsgerechte Erzeugung von nachkorrigiertem und strukturiertem Volltext in der Größenordnung von mehreren Millionen Seiten in Bezug auf Aufwand und Kosten effizient sein. Daher wird hier eine möglichst weitgehende Automatisierung der Nachkorrektur von OCR-Volltext vorgestellt.[5]

Zu der hier beschriebenen OCR-Nachkorrektur gab es den folgenden Kontext. Über 185.000 Seiten in Fraktur gesetzter Schrift wurden an der Staats- und Universitätsbibliothek Bremen (SuUB) im Rahmen zweier DFG-geförderter Projekte digitalisiert und im Volltext erschlossen. Dazu wurde die OCR ABBYY FineReader Version 9 verwendet. Sie stellte die Dateien im Format FineReader8-schema-v2[6] zur Verfügung.

Durch Segmentierung des OCR-Textes auf Wortebene ergaben sich beim *Grenzboten* knapp 80 Millionen Wörter. Diese wurden mit einer Liste von ca. 1,7 Millionen historischen Wortformen abgeglichen und auf der Grundlage ausgewählter Heuristiken[7] korrigiert.[8] Ein effizienter Algorithmus leistet diesen Abgleich und berücksichtigt dabei die im Projekt identifizierten, für Frakturschrift typischen OCR-Fehler.

Die DFG-Praxisregeln „Digitalisierung" sehen inzwischen vor, dass „nach dem Stand der gegenwärtigen Technik […] eine OCR-Erkennung bei Drucken der Maschinenpressenzeit ab 1850 verpflichtend" ist.[9] Dies entspricht der minimalen Funktionalitätsanforderung, um eine Volltextrecherche zu ermöglichen. Motivation und Ziel eines dieser Publikation zugrundeliegenden durch die DFG geförderten Projektes war, den OCR-Volltext, der zunächst nur als sogenannte „schmutzige OCR", d.h. ausschließlich als Basis für eine Suchfunktionalität erstellt wurde, weiter aufzubereiten, sodass dieser geeigneter für eine wissenschaftliche Beforschung zur Verfügung steht. Einerseits ist der Volltext als Basis für eine reine Suchfunktionalität viel zu wertvoll, als dass man diesen nur als Suchindex innerhalb eines Digitalisierungsportals verwendet, andererseits gibt es weitere Anforderungen für die Integration von Volltext in eine virtuelle Forschungsumgebung oder eine Forschungsinfrastruktur, wie z.B. dem CLARIN-D.[10] Insgesamt geht daher die genannte Motivation und Zielsetzung über die reine Zeichenkorrektur hinaus. So wurde für die geplante Integration in CLARIN-D eine textformale Auszeichnung bis zur Generierung eines Korpus im TEI P5 Format erstellt. Damit ist die Auszeichnung

4 Bettina Kann und Michael Hintersonnleitner, „Volltextsuche in historischen Texten - Erfahrungen aus den Projekten der Österreichischen Nationalbibliothek", *BIBLIOTHEK – Forschung und Praxis* 39, Nr. 1 (2015): 76, http://dx.doi.org/10.1515/bfp-2015-0004.

5 Federbusch und Polzin, *Volltext*, 32.

6 Schema: http://www.abbyy.com/FineReader_xml/FineReader8-schema-v2.xml.

7 Der in der Informatik verwendete Begriff der Heuristik entspricht dem Einsatz von Grundannahmen in einem gegebenen Kontext. Im Abschnitt 3. Zielsetzung und der Bremer Ansatz werden die verwendeten Heuristiken näher erläutert.

8 Federbusch und Polzin, *Volltext*, 130.

9 DFG-Praxisregeln Digitalisierung, DFG-Vordruck 12.151 – 02/13, 44, zuletzt geprüft am 01.02.2016, http://www.dfg.de/formulare/12_151/12_151_de.pdf.

10 CLARIN-D – eine „digitale Forschungsinfrastruktur für Sprachressourcen in den Geistes- und Sozialwissenschaften", zuletzt geprüft am 01.02.2016, http://www.clarin-d.de/de/.

von Seitenelementen wie Absätze, Rubrikenüberschriften, Abbildungen und Fußnoten gemeint. Ein wesentliches Ziel ist dabei, einen durchgängigen Fließtext zu erhalten, der nicht durch die übrigen Seitenelemente unterbrochen wird. Stäcker stuft die strukturelle Auszeichnung als für die digitale Nachnutzung unverzichtbar ein.[11]

Die erwähnten computerbasierten Werkzeuge und Infrastrukturen, jede Benutzung und Verarbeitung von Volltext sind stark von der Qualität bzw. Textgenauigkeit des entsprechenden Volltextes abhängig. Hier ist damit zunächst die Quote korrekter Zeichen bzw. Wörter gemeint. Bei einem etwas weiter gefassten Begriff von Volltextqualität kann man die Auszeichnung von Strukturmerkmalen wie z.B. Absätze und Fußnoten mit hinzuziehen. Jede einzelne Elimination eines Zeichenfehlers (und davon gab es beim *Grenzboten* schätzungsweise ca. 8,65 Millionen[12]) verbessert die Qualität der Ergebnisse zahlreicher computerbasierter Verarbeitungsmethoden, wie z.B. Volltextrecherche, unscharfe/fuzzy Suche (rechtstrunkierte Suche, Flexionsformen berücksichtigende Suche, Alinierung, Suche mit regulären Ausdrücken), POS Tagging, Entity Recognition, „quantitative Auswertungen im Rahmen von Text- oder Datamining“[13], Identifikation von Textsorten[14], Topic Modelling und semantische Verarbeitung von Text[15].

Bei der Bewertung von Zeichenerkennungsquoten darf man sich nicht durch Zahlen über 90 % täuschen lassen. Bei 90 % ist noch jedes 10-te Zeichen fehlerhaft. Das wären beim *Grenzboten* fast 7 Zeichen pro Zeile. Sicher mögen schon Texte um 90 % Zeichengenauigkeit einen wissenschaftlichen Nutzen haben, wenn die Fragestellung entsprechend ist.[16] Im Kontext wissenschaftlicher Analysen mit einer Abhängigkeit der Fragestellung von stark eingrenzenden Suchkriterien, wie z.B. die Suche nach Entitäten bzw. Wortgruppen mit überwiegend geringer Häufigkeit im zu analysierenden Text, sind zuverlässige Ergebnisse mit 95 % nur schwer möglich. Erste gesicherte Analysen mit z.B. stark eingrenzenden Suchkriterien sollten ab einer Quote von 99,5 % möglich sein. Erst ab 99,95 % bezeichnen die DFG Praxisregeln einen Volltext als wissenschaftlich zuverlässig.[17] Damit eignet sich ein Volltext für die „ausschließende Suche“, d.h. für den Test, ob ein Suchwort tatsächlich nicht in dem Volltext vorkommt. Ein zu 100 % fehlerfreies großes Textkorpus wird es kaum geben, da die Originalvorlagen für große Digitalisierungsprojekte selbst auch fehlerbehaftet sind. Die folgenden Abschnitte werden aufzeigen, mit welchen Aufwänden eine Verbesserung der Zeichenerkennungsquote von 98,28 % auf 98,83 % erreicht wurde und wie diese Korrektur von 32 % aller Fehler zu bewerten ist. Das selbst gesteckte Ziel einer Zeichenerkennungsquote von 99,5 % wurde im Rahmen des vorgestellten Volltextprojektes nicht erreicht.

11 Stäcker, „Konversion“, 227.
12 Die Zeitschrift *Die Grenzboten* enthält ca. 500 Millionen Zeichen; bei 98,28 % Zeichenerkennungsquote ergeben sich hochgerechnet ca. 8,6 Millionen Zeichenfehler.
13 DFG-Praxisregeln Digitalisierung, DFG-Vordruck 12.151.
14 Kerry Kilner und Kent Fitch, „Discovering and Rediscovering Full Text: Unearthing and Refactoring“, zuletzt geprüft am 01.02.2016, http://dh2015.org/abstracts/xml/KILNER_Kerry_Discovering_and_Rediscovering_Full_T/KILNER_Kerry_Discovering_and_Rediscovering_Full_Text__U.html.
15 Stäcker, „Konversion“, 231–233.
16 Zum Beispiel bei statistischen Analysen über hinreichend große Textmengen zusammen mit Analysekriterien, die hinreichend große Mengen von Daten ergeben und damit statistisch signifikante Aussagen ermöglichen.
17 DFG-Praxisregeln Digitalisierung, DFG-Vordruck 12.151, 32.

Oft wird diskutiert, ob die Buchstabengenauigkeit bzw. Zeichenerkennungsquote oder die Wortgenauigkeit bzw. Worterkennungsquote das bessere Maß für die Dokumentation der Textqualität ist. In Abschnitt 5. Ergebnisse werden beide Zahlenwerte angegeben. Mit dem Fokus auf OCR-Nachkorrektur hat die Zeichenerkennungsquote hier die höhere Priorität. Die Korrigierbarkeit eines Textes oder Wortes sowie die Leistung eines Korrekturansatzes sind mit der Zeichenerkennungsquote besser messbar. So ist das dem OCR-Text entnommene Wort „crsüiltc“ für „erfüllte“, das aus 8 Zeichen besteht, mit 4 Zeichenfehlern gerade noch korrigierbar, mit 5 oder 6 Fehlern aber nicht mehr. Werden nur 3 der 4 Zeichenfehler korrigiert, dann registriert die Worterkennungsquote diese Verbesserung nicht, obwohl eine fehlertolerante Suche hier deutlich bessere Chancen hat, das Wort zu finden. Ebenso hat das menschliche Auge es leichter, z.B. „erfüilte“ als das Wort „erfüllte“ zu erkennen.

Der auf mehreren Veranstaltungen der EU Projekte IMPACT und SUCCEED vorgestellte Stand der Forschung ging in das hier vorgestellte OCR-Nachbearbeitungsprojekt ein. Auf der Konferenz DATeCH-2014[18] Digitization Day 2014 in Madrid wurde der cloud service overProof[19] vorgestellt.[20] Dieser webbasierte Korrekturservice für englischsprachige OCR-Texte wurde im Rahmen einer Zusammenarbeit mit der SuUB Bremen für die Korrektur von OCR-Text deutschsprachiger Frakturtexte weiterentwickelt. Die Ergebnisse dazu werden in Abschnitt 5.2. vorgestellt. Mit dem am *Center for Information and Language Processing*[21] der Ludwig-Maximilians-Universität München entwickelten PoCoTo (CIS-LMU Post Correction Tool[22]) steht darüber hinaus ein Softwaretool zur Verfügung, das neben automatisierten Korrekturen auch eine optische Kontrolle anbietet.

2. Ausgangslage – Der unkorrigierte Volltext

Die nationalliberale Zeitschrift *Die Grenzboten* ist zwischen 1841 und 1922 wöchentlich, zeitweise zweiwöchentlich erschienen, sie hatte das Ziel, die gesamte bürgerliche Lebenswelt abzubilden. Ab 1871 erhält die Zeitschrift den Untertitel: „Zeitschrift für Politik, Literatur und Kunst“. Beiträge aus Politik und Geschichte über Wirtschaft bis hin zu schöngeistigen Themen aus Kunst, Musik und Literatur zeigen die umfangreiche Themenvielfalt der Zeitschrift, die das bürgerliche Leben im langen 19. Jahrhundert durch eine politisch und kulturell wechselhafte Zeit begleitete. Die lange Erscheinungsdauer der *Grenzboten* ermöglicht die Analyse von Verstetigung und Wandel kultureller Werte sowie medialer Strukturen des deutschen Nationalismus. Als eines der bedeutendsten Periodika des Jahrhunderts sind die *Grenzboten* somit eine herausragende Quelle für Geschichtswissenschaft, Germanistik, Kulturwissenschaften (bspw. Literatur-, Kunst- und Musikgeschichte), Pressegeschichte und weitere Fachwissenschaften.

18 DaTeCH; Digitisation Days; koordiniert vom IMPACT Centre of Competence, zuletzt geprüft am 01.02.2016, http://www.digitisation.eu/blog/digitisation-days-19-20-may-2014-2/.

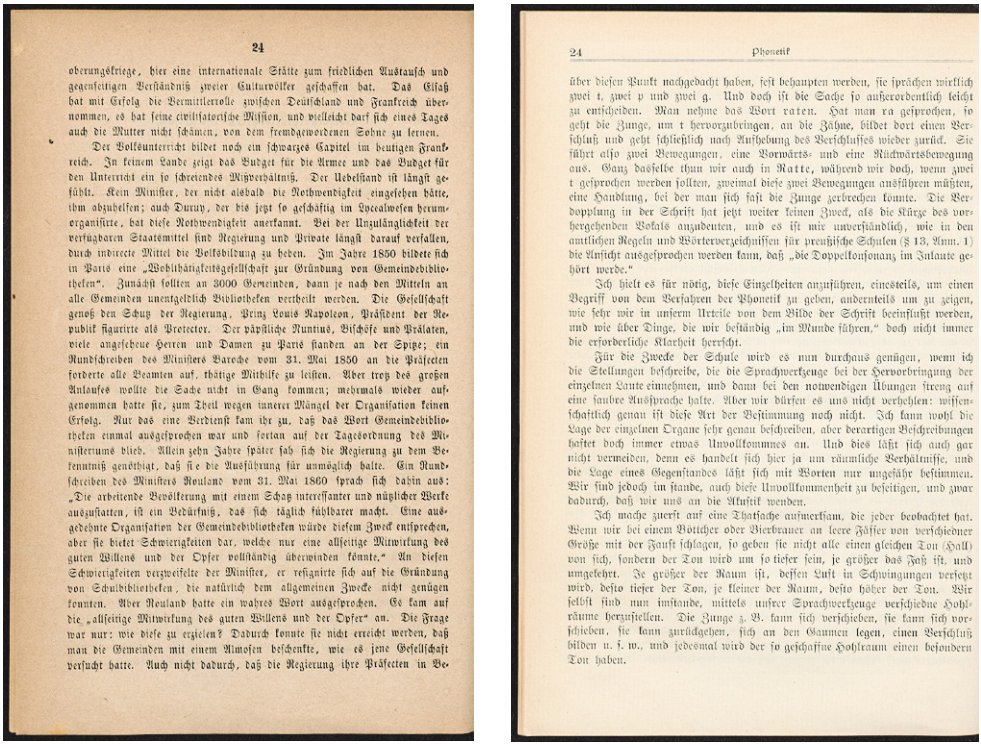
19 Cloud service “overProof – Automatic Correction of OCR”, zuletzt geprüft am 01.02.2016, http://overproof.projectcomputing.com/.

20 Evershed und Fitch, “Correcting Noisy ORC”.

21 CIS, Center for Information and Language Processing, Ludwig-Maximilians-Universität München, , zuletzt geprüft am 01.02.2016, http://www.cis.uni-muenchen.de/.

22 CIS LMU Post Correction Tool, zuletzt geprüft am 01.02.2016, http://www.digitisation.eu/tools-resources/tools-for-text-digitisation/cis-lmu-post-correction-tool-pocoto/.

Das Schriftbild sowie die Seitenstruktur des *Grenzboten* sind weitgehend homogen. Die überwiegende Anzahl der Seiten besteht aus einspaltigem Blocksatz mit Paginierung, im späteren Erscheinungsverlauf auch mit Rubrikenüberschrift in der Kopfzeile. Die folgende Seite stellt das Erscheinungsbild des *Grenzboten* prototypisch links vor 1887 und rechts mit Rubrikenüberschrift nach 1887 dar. Auf Seiten, die von diesem Erscheinungsbild abweichen, wie z.B. Titel- und Anzeigenseiten, Inhaltsverzeichnisse und besonders gesetzte Seiten (Tabellen, Fußnoten und Einrückungen), geht der Abschnitt 6.1. ein.



24

oberungskriege, hier eine internationale Stätte zum friedlichen Austausch und gegenseitigen Verständniß zweier Culturvölker geschaffen hat. Das Elsaß hat mit Erfolg die Vermittlerrolle zwischen Deutschland und Frankreich übernommen, es hat seine civilisatorische Mission, und vielleicht darf sich eines Tages auch die Mutter nicht schämen, von dem fremdgewordenen Sohne zu lernen.

Der Volksunterricht bildet noch ein schwarzes Capitel im heutigen Frankreich. In keinem Lande zeigt das Budget für die Armee und das Budget für den Unterricht ein so schreiendes Mißverhältniß. Der Uebelstand ist längst gefühlt. Kein Minister, der nicht alsbald die Nothwendigkeit eingesehen hätte, ihm abzuhelfen; auch Duruy, der bis jetzt so geschäftig im Localwesen herumorganisirte, hat diese Nothwendigkeit anerkannt. Bei der Unzulänglichkeit der verfügbaren Staatsmittel sind Regierung und Private längst darauf verfallen, durch indirecte Mittel die Volksbildung zu heben. Im Jahre 1850 bildete sich in Paris eine „Wohlthätigkeitsgesellschaft zur Gründung von Gemeindebibliotheken". Zunächst sollten an 3000 Gemeinden, dann je nach den Mitteln an alle Gemeinden unentgeltlich Bibliotheken vertheilt werden. Die Gesellschaft genoß den Schutz der Regierung, Prinz Louis Napoleon, Präsident der Republik figurirte als Protector. Der päpstliche Nuntius, Bischöfe und Prälaten, viele angesehene Herren und Damen zu Paris standen an der Spitze; ein Rundschreiben des Ministers Baroche vom 31. Mai 1850 an die Präfecten forderte alle Beamten auf, thätige Mithilfe zu leisten. Aber trotz des großen Anlaufes wollte die Sache nicht in Gang kommen; mehrmals wieder aufgenommen hatte sie, zum Theil wegen innerer Mängel der Organisation keinen Erfolg. Nur das eine Verdienst kam ihr zu, daß das Wort Gemeindebibliotheken einmal ausgesprochen war und fortan auf der Tagesordnung des Ministeriums blieb. Allein zehn Jahre später sah sich die Regierung zu dem Bekenntniß genöthigt, daß sie die Ausführung für unmöglich halte. Ein Rundschreiben des Ministers Rouland vom 31. Mai 1860 sprach sich dahin aus: „Die arbeitende Bevölkerung mit einem Schatz interessanter und nützlicher Werke auszustatten, ist ein Bedürfniß, das sich täglich fühlbarer macht. Eine ausgedehnte Organisation der Gemeindebibliotheken würde diesem Zweck entsprechen, aber sie bietet Schwierigkeiten dar, welche nur eine allseitige Mitwirkung des guten Willens und der Opfer vollständig überwinden könnte." An diesen Schwierigkeiten verzweifelte der Minister, er resignirte sich auf die Gründung von Schulbibliotheken, die natürlich dem allgemeinen Zwecke nicht genügen konnten. Aber Rouland hatte ein wahres Wort ausgesprochen. Es kam auf die „allseitige Mitwirkung des guten Willens und der Opfer" an. Die Frage war nur: wie diese zu erzielen? Dadurch konnte sie nicht erreicht werden, daß man die Gemeinden mit einem Almosen beschenkte, wie es jene Gesellschaft versucht hatte. Auch nicht dadurch, daß die Regierung ihre Präfecten in Be-

24 Phonetik

über diesen Punkt nachgedacht haben, seit behaupten werden, sie sprächen wirklich zwei t, zwei p und zwei g. Und doch ist die Sache so außerordentlich leicht zu entscheiden. Man nehme das Wort raten. Hat man ra gesprochen, so geht die Zunge, um t hervorzubringen, an die Zähne, bildet dort einen Verschluß und geht schließlich nach Aufhebung des Verschlusses wieder zurück. Sie führt also zwei Bewegungen, eine Vorwärts- und eine Rückwärtsbewegung aus. Ganz dasselbe thun wir auch in Ratte, während wir doch, wenn zwei t gesprochen werden sollten, zweimal diese zwei Bewegungen ausführen müßten, eine Handlung, bei der man sich fast die Zunge zerbrechen könnte. Die Verdopplung in der Schrift hat jetzt weiter keinen Zweck, als die Kürze des vorhergehenden Vokals anzudeuten, und es ist mir unverständlich, wie in den amtlichen Regeln und Wörterverzeichnissen für preußische Schulen (§ 13, Anm. 1) die Ansicht ausgesprochen werden kann, daß „die Doppelkonsonanz im Inlaute gehört werde."

Ich hielt es für nötig, diese Einzelheiten anzuführen, einesteils, um einen Begriff von dem Verfahren der Phonetik zu geben, anderenteils um zu zeigen, wie sehr wir in unserm Urteile von dem Bilde der Schrift beeinflußt werden, und wie über Dinge, die wir beständig „im Munde führen," doch nicht immer die erforderliche Klarheit herrscht.

Für die Zwecke der Schule wird es nun durchaus genügen, wenn ich die Stellungen beschreibe, die die Sprachwerkzeuge bei der Hervorbringung der einzelnen Laute einnehmen, und dann bei den notwendigen Übungen streng auf eine saubre Aussprache halte. Aber wir dürfen es uns nicht verhehlen: wissenschaftlich genau ist diese Art der Bestimmung noch nicht. Ich kann wohl die Lage der einzelnen Organe sehr genau beschreiben, aber derartigen Beschreibungen haftet doch immer etwas Unvollkommnes an. Und dies läßt sich auch gar nicht vermeiden, denn es handelt sich hier ja um räumliche Verhältnisse, und die Lage eines Gegenstandes läßt sich mit Worten nur ungefähr bestimmen. Wir sind jedoch im stande, auch diese Unvollkommenheit zu beseitigen, und zwar dadurch, daß wir uns an die Akustik wenden.

Ich mache zuerst auf eine Thatsache aufmerksam, die jeder beobachtet hat. Wenn wir bei einem Böttcher oder Bierbrauer an leere Fässer von verschiedner Größe mit der Faust schlagen, so geben sie nicht alle einen gleichen Ton (Hall) von sich, sondern der Ton wird um so tiefer sein, je größer das Faß ist, und umgekehrt. Je größer der Raum ist, dessen Luft in Schwingungen versetzt wird, desto tiefer der Ton, je kleiner der Raum, desto höher der Ton. Wir selbst sind nun imstande, mittels unsrer Sprachwerkzeuge verschiedne Hohlräume herzustellen. Die Zunge z. B. kann sich verschieben, sie kann sich vorschieben, sie kann zurückgehen, sich an den Gaumen legen, einen Verschluß bilden u. s. w., und jedesmal wird der so geschaffne Hohlraum einen besondern Ton haben.

Abb. 1: Prototypisches Erscheinungsbild des Grenzboten links vor 1887 und rechts mit Rubrikenüberschrift nach 1887

Überschriften weichen gelegentlich von dem Schriftschnitt des Fließtextes stark ab. Siehe in Abb. 2 einen direkten Vergleich der Buchstaben R, E und Z in verschiedenen Schriftschnitten.[23]

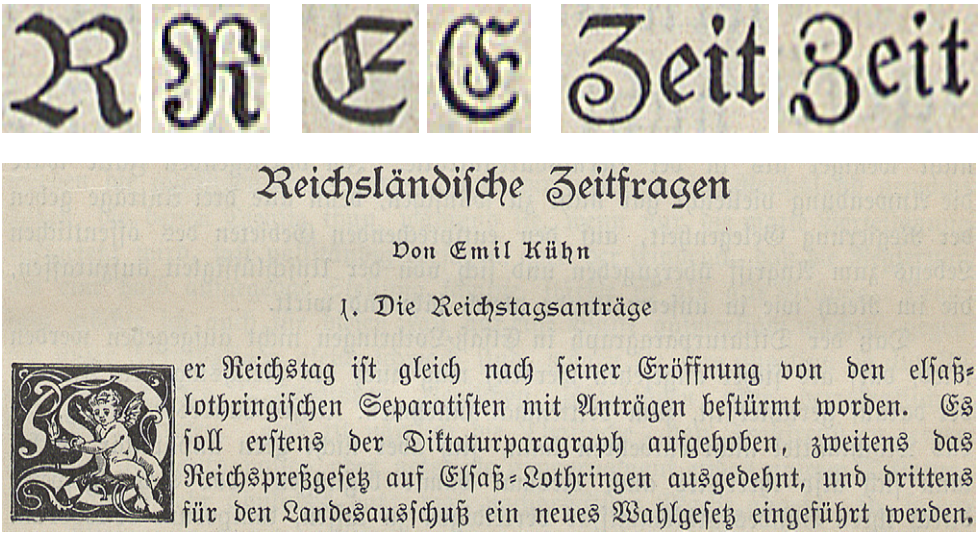


Abb. 2: Vergleich der Buchstaben R, E und Z in verschiedenen Schriftschnitten

In Inhaltsverzeichnissen, Fußnoten, eingerücktem Text sowie Bogensignaturen wird eine kleinere Schriftgröße verwendet:[24]

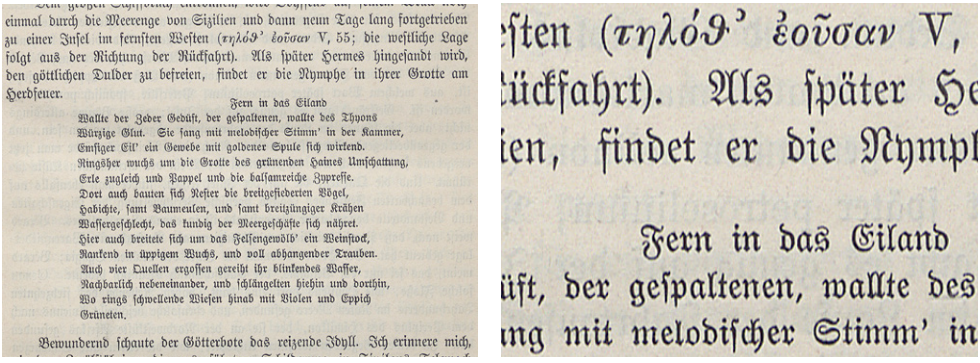


Abb. 3: Verschiedene Schriftgrößen

23 http://brema.suub.uni-bremen.de/grenzboten/periodical/pageview/226917.
24 http://brema.suub.uni-bremen.de/grenzboten/periodical/pageview/87570.

Die Konzeption des Digitalisierungsprojektes erfolgte in Zusammenarbeit mit Wissenschaftlerinnen und Wissenschaftlern aus verschiedenen geisteswissenschaftlichen Disziplinen. Bei dem Bedarf nach möglichst fehlerfreiem Volltext wurde der Einsatz von Double Keying erwogen. Aus Kostengründen sollten die ca. 187.000 Seiten in 270 Bänden jedoch per OCR im Volltext erschlossen werden. Dazu wurde die Software ABBYY FineReader in der Version 9 eingesetzt.

Auch bei der Zeitschrift *Die Grenzboten* machten sich die in der Literatur beschriebenen Probleme bei der Erkennung von Frakturschrift im OCR-Ergebnis bemerkbar.[25] Die Zeichenerkennungsrate nach Abschluss des Digitalisierungsprojektes betrug 98,28 %. Zusätzlich wurde eine Statistik der häufigsten Fehler erstellt.

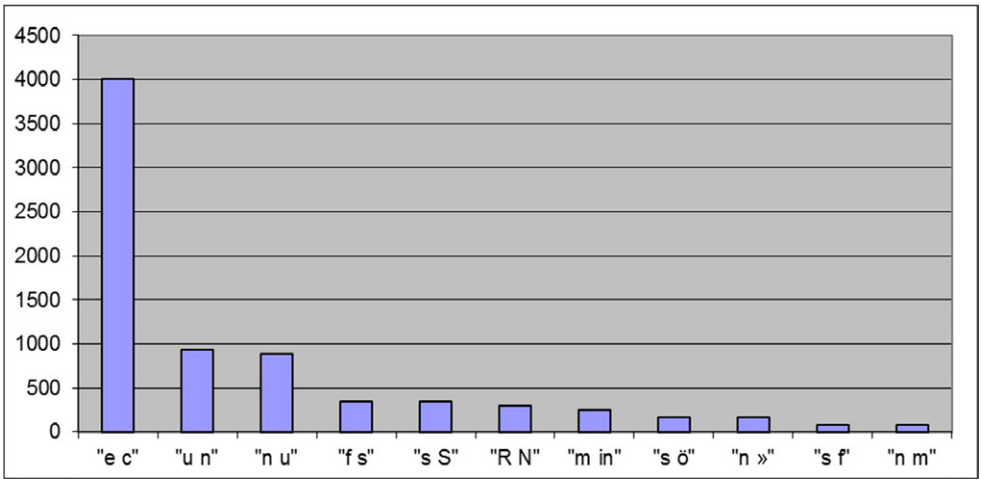


Abb. 4: Statistik der häufigsten Zeichenfehler

3. Zielsetzung und der *Bremer Ansatz*

Der *Bremer Ansatz* zur Nachkorrektur von OCR-Volltext basiert auf der Annahme, dass überwiegend OCR-spezifische Zeichenfehler, wie z.B. Verwechslungen der Buchstaben e/c, u/n, s/f usw. auftauchen; und dies wurde bei der Erstellung der Fehlertypstatistiken bestätigt. Weiterhin zeichnet sich der Ansatz durch Einfachheit aus; eine Liste historischer bzw. dialekt- oder fachspezifischer Wortformen ist verhältnismäßig leicht erstellbar, wenn auf entsprechende Korpora zurückgegriffen werden kann. Ein effizienter Algorithmus leistet den Abgleich von hier ca. 1,7 Millionen Wortformen gegen beim *Grenzboten* knapp 80 Millionen enthaltenen Wörtern und lässt sich auf nachvollziehbarer Art und Weise parametrisieren, d.h. auf die spezifischen Eigenschaften des jeweiligen Volltextprojektes einstellen. So konnten hier die bei Fraktur so typischen oben erwähnten Zeichenvertauschungen e/c, u/n und s/f mit einem entsprechend stark angepassten Parameter problemlos berücksichtigt werden.

25 Federbusch und Polzin, *Volltext*, 21; Furrer und Volk, „Reducing OCR Errors"; Steffen Wawra und Silke Wündrich, „OCR für Frakturschriften?", *Bibliotheksdienst* 33 (1999): 2110–2117.

Die in der Einleitung erwähnten Heuristiken steuern die Effizienz und auch das Auftreten von mehr oder weniger falsch-positiven Korrekturen. Eine falsch-positive Korrektur[26] ist die fehlerhafte Veränderung eines korrekten Wortes. In diesem Bezeichnungsschema sind die gewünschten Korrekturen richtig-positiv und alle unverändert belassenen Wörter teilen sich auf in falsch-negativ und richtig-negativ, d.h. in nicht korrigierte OCR-Fehler und unverändert belassene korrekte Wörter.

Hervorzuhebende Heuristiken sind die folgenden:

1. Es werden keine Wortformen korrigiert, die selber in der Liste historischer Wortformen enthalten sind.
2. Ein potenziell fehlerhaftes Wort des OCR-Textes besteht ausschließlich aus einer gegebenen Menge von Zeichen.[27] Es werden nur die so identifizierten Wortformen bearbeitet und die Wortzwischenräume unverändert gelassen.
3. Die sogenannte Liste der Zeichensubstitutionen entspricht den tatsächlich auftretenden OCR-Fehlern. Da die verwendete OCR-Software manche Fehler öfter generiert, werden die Zeichensubstitutionen gewichtet. Dies entspricht einem der verwendeten OCR-Software spezifisch angepassten Fehlermodell.

Noch grundsätzlicher als Heuristiken sind die folgenden Entscheidungen und intrinsischen Effekte:

1. Es werden keine Leerzeichen entfernt oder eingefügt, d.h. Wortformen werden weder zusammengefügt noch getrennt.
2. Eine Korrektur findet nur dann statt, wenn eine Wortform der Liste der historischen Wortformen durch eine Anzahl der gegebenen Zeichensubstitutionen erreichbar ist (z.B. werden Wortformen mit Ziffern nicht korrigiert, da hierfür keine sinnvollen Zeichensubstitutionen identifiziert wurden).
3. Die Korrektur wird Wort für Wort und ohne Verwendung von Kontext vorgenommen. Der algorithmische Ansatz wird im Abschnitt 4. beschrieben. Über den *Bremer Ansatz* hinausführende Ansätze, wie z.B. die Berücksichtigung von Wortkontexten werden im Abschnitt 5.2. sowie in der Diskussion beleuchtet.[28]

26 Der Begriff „falsch-positiv“ wird bei der Beurteilung von Klassifikatoren definiert (https://de.wikipedia.org/wiki/Beurteilung_eines_Klassifikators); bei der OCR-Nachkorrektur müsste exakt z.B. von „falsch-positiven Modifikationen“ gesprochen werden, da im strengeren Sinne nur der Fall der „richtig-positiven Modifikationen“ als „Korrekturen“ zu bezeichnen wären. Die übrigen drei Fälle entsprächen dann „Verschlimmbesserungen“, „ausgelassenen Korrekturen“ und den hoffentlich überwiegenden „unveränderten korrekten Wörter“.

27 Bei genauerer Betrachtung ist diese Tokenisierung exakt zu spezifizieren. Tatsächlich ist die Erstellung einer Liste von Wortzeichen eine Parametrisierung für jedes Volltextprojekt; vgl. den Abschnitt 3.2. Parametrisierung des Bremer Ansatzes.

28 Evershed und Fitch, „Correcting Noisy ORC“.

3.1. Wie wurden OCR-Fehler gezählt?

Basierend auf 370 Seiten Ground Truth Text[29] konnten OCR-Fehler automatisiert gezählt und typisiert werden. Dazu wurde ein an der SuUB Bremen entwickeltes Softwaretool OCR- Visualizer[30] eingesetzt. Der OCR-Visualizer aligniert den Volltext seitenweise. Dabei werden zunächst verschiedene Typen von Textabweichungen identifiziert:

- Einfügungen (Insertion)
- Ersetzungen (Substitution)
  - Mehrzeichensubstitutionen
- Löschungen (Deletion)

Sehr hilfreich sind die mit dem *Bremer Ansatz* eingeführten Mehrzeichensubstitutionen. Diese mehrere Zeichen betreffenden Textabweichungen, wie z.B. rn/m, im/un und iii/m werden nicht nur als Kombination von Einfügungen (Insertion), Ersetzungen (Substitution) und Löschungen (Deletion) identifiziert, sondern als Mehrzeichensubstitution („many-to-one" bzw. „one-to-many" Substitution)[31] zusammengefasst.

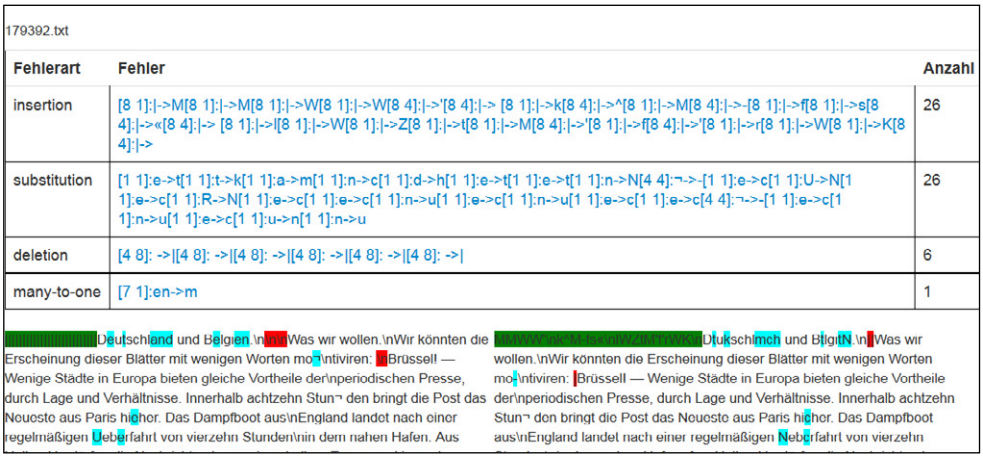
Für die Bewertung eines OCR-Textes werden bei Mehrzeichensubstitutionen stets die fehlerhaften Zeichen im Ground Truth Text gezählt. So würde die Mehrzeichensubstitution (Ground Truth Text, OCR-Text)=(rn, m) als zwei Fehler gezählt werden, hingegen (m, rn) als ein Fehler. Weiterhin dokumentiert der OCR-Visualizer, ob es sich bei den betroffenen Zeichen einer Substitution um Buchstaben, Ziffern, Sonderzeichen (Diakritika, Abkürzungszeichen), Satzzeichen oder Leerzeichen (bzw. Whitespace) handelt. Die Mehrzeichensubstitutionen repräsentieren direkt eine spezifische Eigenart von OCR-Fehlern und finden daher ebenfalls bei der Konzeption des Korrekturalgorithmus Berücksichtigung.

Bei der automatischen Analyse sind verschiedene Szenarien denkbar: Berücksichtigung von Groß-/Kleinschreibung, Ziffern, Satzzeichen, Sonderzeichen und Leerzeichen (Whitespace). Der OCR-Visualizer identifiziert die entsprechenden Zeichentypen und ermöglicht somit, verschiedene Zählszenarien darzustellen. Von streng (Berücksichtigung aller Fehlertypen) bis zur Groß-/Kleinbuchstabenunabhängigen Stichwortsuche sind verschiedene Szenarien abbildbar. Darüber hinaus leistete das Tool gute Beiträge bei der Parametrisierung des *Bremer Ansatzes* sowie bei der Quantifizierung der erzielten Ergebnisse. Bei den im Abschnitt 5. angegebenen Ergebnissen wurde das Szenario Stichwortsuche verwendet. Dabei wurden keine Ziffern, Sonderzeichen, Satzzeichen oder Leerzeichen (bzw. Whitespace) berücksichtigt.

29 Der Ground Truth Text (eine per Abschrift und manueller Nachkorrektur erstellte fehlerfreie Version des betrachteten OCR-Textes) wurde im Deutschen Textarchiv erstellt. Bei dem Zugriff auf die folgenden URLs ist eine kostenlos erhältliche Registrierung notwendig. http://www.deutschestextarchiv.de/dtaq/book/show/grenzboten_179382_282158;
http://www.deutschestextarchiv.de/dtaq/book/show/nn_charaktere01_1848;
http://www.deutschestextarchiv.de/dtaq/book/show/nn_charaktere02_1848;
http://www.deutschestextarchiv.de/dtaq/book/show/gutzkow_patkul_1842.

30 https://github.com/suub/ocr-visualizer.

31 Hier kam eine Heuristik zum Einsatz, die es erlaubte, auf der Basis einer kodierten Charakteristik jedes Buchstabens der Frakturschrift automatisiert zu entscheiden, ob mehrere aufeinanderfolgende Buchstaben potenziell einem anderen Buchstaben ähneln.



179392.txt

Fehlerart	Fehler	Anzahl
insertion	[8 1]:|->M[8 1]:|->M[8 1]:|->W[8 1]:|->W[8 4]:|->'[8 4]:|-> [8 1]:|->k[8 4]:|->^[8 1]:|->M[8 4]:|->-[8 1]:|->f[8 1]:|->s[8 4]:|->«[8 4]:|-> [8 1]:|->|[8 1]:|->W[8 1]:|->Z[8 1]:|->t[8 1]:|->M[8 4]:|->'[8 1]:|->f[8 4]:|->'[8 1]:|->r[8 1]:|->W[8 1]:|->K[8 4]:|->	26
substitution	[1 1]:e->t[1 1]:t->k[1 1]:a->m[1 1]:n->c[1 1]:d->h[1 1]:e->t[1 1]:e->t[1 1]:n->N[4 4]:¬->-[1 1]:e->c[1 1]:U->N[1 1]:e->c[1 1]:R->N[1 1]:e->c[1 1]:e->c[1 1]:n->u[1 1]:e->c[1 1]:n->u[1 1]:e->c[1 1]:e->c[4 4]:¬->-[1 1]:e->c[1 1]:n->u[1 1]:e->c[1 1]:u->n[1 1]:n->u	26
deletion	[4 8]: ->|[4 8]: ->|[4 8]: ->|[4 8]: ->|[4 8]: ->|[4 8]: ->|	6
many-to-one	[7 1]:en->m	1

Deutschland und Belgien.\nWas wir wollen.\nWir könnten die Erscheinung dieser Blätter mit wenigen Worten motiviren: Brüssell — Wenige Städte in Europa bieten gleiche Vortheile der\nperiodischen Presse, durch Lage und Verhältnisse. Innerhalb achtzehn Stun¬ den bringt die Post das Neueste aus Paris highor. Das Dampfboot aus\nEngland landet nach einer regelmäßigen Ueberfahrt von vierzehn Stunden\nin dem nahen Hafen. Aus

Deutschland und BelgiN.\nWas wir wollen.\nWir könnten die Erscheinung dieser Blätter mit wenigen Worten mo-\ntiviren: Brüssell — Wenige Städte in Europa bieten gleiche Vortheile der\nperiodischen Presse, durch Lage und Verhältnisse. Innerhalb achtzehn Stun¬ den bringt die Post das Neueste aus Paris highor. Das Dampfboot aus\nEngland landet nach einer regelmäßigen Neberfahrt von vierzehn

Abb. 5: OCR-Visualizer

3.2. Parametrisierung des *Bremer Ansatzes*

Der *Bremer Ansatz* entspricht keinem fertigen und allgemein verwendbaren Softwaresystem, sondern muss jeweils den spezifischen Eigenheiten eines Digitalisierungsprojektes angepasst werden. Neben der Sprache und der Schriftart sind dies weitere Eigenschaften, die den Charakter eines Digitalisierungsprojektes bestimmen: Erscheinungsjahrhundert, Anteile von Fremdsprachen oder Dialekte, Homogenität des Materials in Bezug auf das Erscheinungsbild und Texteigenschaften, Fachsprache, Qualität der Digitalisierung (Scans) sowie die verwendete OCR-Software. Diese Liste ließe sich sicherlich erweitern. Die Parametrisierung eines Softwaresystems, d.h. die vollständige oder teilweise Berücksichtigung dieser Eigenschaften mittels Basisdaten oder Einstellmöglichkeiten, ist einerseits Verpflichtung und Aufgabe, andererseits bietet sie jedoch auch die Möglichkeit, andere Eigenschaften speziell oder zusätzlich zu berücksichtigen.

Die Zeitschrift *Die Grenzboten* kann in Bezug auf diese Eigenschaften als weitgehend unproblematisch bezeichnet werden, sie repräsentiert bezüglich ihrer Eigenschaften sicher einen großen Teil der Titel des 19. Jahrhunderts. Ihr entsprechen die folgenden drei Angaben und somit ist der *Bremer Ansatz* vollständig parametrisiert.

1. Definition und Bereinigung der Liste der historischen Wortformen

Die Liste der historischen Wortformen wurde in Zusammenarbeit mit dem Zentrum Sprache der BBAW (Berlin-Brandenburgischen Akademie der Wissenschaften) an dem das Deutsche Textarchiv (DTA) angesiedelt ist, erstellt. Sie wurde als ein Auszug laufender Wortformen aus dem DTA entnommen und enthält ca. 1,7 Millionen Wortformen. Zusätzlich ist mit Frequency die Häufigkeit der jeweiligen Wortform in dem Korpus angegeben. Die folgende Tabelle zeigt die Liste der historischen Wortformen ausschnittsweise; sie kann in GitHub[32] vollständig heruntergeladen werden.

32 https://github.com/suub/bote/blob/99d845dd390c668e3b47813059a8da22d77f1e0c/resources/current-params/dict.fuwv.

Tab. 1: Auszug aus der Liste der historischen Wortformen

Frequency	Surface Form	Transliteration into the subset of ISO-8859–1 (Latin-1)	Modern Form
278187	und	und	und
239487	der	der	der
233389	die	die	die
...			
28268	iſt	ist	ist
14885	ſo	so	so
...			
600	Aehnlichkeit	Aehnlichkeit	Ähnlichkeit
322	Säugethiere	Säugethiere	Säugetiere
319	theilt	theilt	teilt
...			
6	ältlicher	ältlicher	ältlicher
1	̊ältlicher	ältlicher	ältlicher

Die *Surface Form* (direkte Übernahme aller Schriftzeichen aus der Originalvorlage) und die *Modern Form* (aktuelle Schreibung) spielten bei der OCR-Nachkorrektur des *Grenzboten* keine Rolle, da mit dem ABBYY Finereader kein langes s generiert wurde. Zeichen wie „å̊“ kommen beim *Grenzboten* nicht vor.

Die laufenden Wortformen großer Textkorpora enthalten naturgemäß einen Anteil fehlerhafter oder ungewöhnlicher Zeichenketten, die keinem korrekt geschriebenen Wort entsprechen. Insbesondere im *Long Tail*[33], *d.*h. in dem großen Bereich von Wortformen mit niedriger Häufigkeit, ließen sich Zeichenketten wie „20pferdekräftiger“, „1874–86“, „ἀλφη-“, „partic“ und „essayaient“ finden. Der *Bremer Ansatz* ist tolerant gegen diese Zeichenketten, da mehrere Kriterien zusammenkommen müssen bis sich eine darin begründete falsch-positive Korrektur ergibt. Deutlich kritischere Wörter, wie z.B. „uud“ und „Dentschland“ in der ursprünglichen Liste der historischen Wortformen, mussten jedoch dringend entfernt werden, da sonst zahlreiche Fehler in dem OCR-Text nicht korrigiert würden (der Fall falsch-negativ).

2. Festlegung von gewichteten Zeichensubstitutionen

Die sogenannte Liste der Zeichensubstitutionen soll den tatsächlich auftretenden OCR-Fehlern der verwendeten OCR-Software entsprechen. Hier gehen auch die oben erwähnten Mehrzeichensubstitutionen ein. Da die verwendete OCR-Software manche Fehler öfter generiert (vgl. Abb. 4), werden die Zeichensubstitutionen grob gewichtet. Auch verschiedene Textmaterialien (Antiqua, Fraktur, Handschrift, Schriftschnitt, Schriftgröße) führen zusammen mit verschiedenen Volltexterfassungsansätzen (OCR, manuelle Nachkorrektur, Doublekeying) zu veränderten Fehlerausprägungen, d.h. auch in diesen Fällen soll und kann die Liste der Zeichensubstitutionen angepasst werden.

33 https://de.wikipedia.org/wiki/The_Long_Tail.

Während des OCR-Nachkorrekturprojektes wurden zahlreiche Listen von Zeichensubstitutionen und verschiedene Gewichtungen durch den automatisierten Abgleich gegen den Ground Truth Text bewertet.[34] Eine automatisiert berechnete Liste von gewichteten Zeichensubstitutionen war geringfügig schlechter als eine manuell nachbearbeitete Liste. Die zuletzt verwendete Liste kann auf GitHub[35] eingesehen werden.

3. Parametrisierung der Tokenisierung (Textsegmentierung auf Wortebene)
Hier wird die Zeichenmenge definiert, aus dem potenziell fehlerhafte Wortformen im OCR-Volltext bestehen dürfen. Neben allen Buchstaben wurde beim *Grenzboten* erwogen, zusätzlich die Zeichen „«“ und „»“ (eine im deutschen Sprachraum des 19. Jahrhunderts ungebräuchliche Variante eines Anführungszeichens) mit als Wortzeichen zu verwenden, da es dazu eine relevante Anzahl von Zeichenfehlern mit der Substitution „»/n“ gab (Beispiel „folge» / folgen“). Das bedeutet, das Korpus eines Volltextprojektes sollte gut bekannt sein. Ein nicht zu heterogen gewähltes Korpus erlaubt zudem, sich für eine Liste von Wortzeichen sicher entscheiden zu können.

4. Algorithmus des *Bremer Ansatzes*

Der Algorithmus zur eigentlichen OCR-Nachkorrektur soll hier grob skizziert werden. Algorithmen sind Handlungsvorschriften zur Lösung von Problemen. Dabei werden bestimmte Eingaben in berechnete Ausgaben überführt. Zur Auflistung der Eingaben des OCR-Nachkorrekturproblems wird hier nochmal zusammengefasst, was in den Abschnitten 1. Einleitung, 3. Zielsetzung und der *Bremer Ansatz* und 3.2. Parametrisierung des Bremer Ansatzes bisher erwähnt wurde:

- eine Liste von historischen Wortformen inklusive der Frequenz bzw. Worthäufigkeit der jeweiligen Wortform
- die Liste der Zeichensubstitutionen und deren Gewichtung (der Häufigkeit der jeweiligen Substitution entsprechend)
- ein zu korrigierendes Wort aus dem OCR-Volltext

Das Problem besteht darin, zu dem gegebenen Wort aus dem OCR-Volltext das nächstgelegene Wort in der Liste der historischen Wortformen zu finden. Dazu wird ein Wortabstand bestimmt, der durch eine Zahl repräsentiert wird. Abweichend von Wortabstandsfunktionen, die über Editier-Operationen definiert sind,[36] werden hier ausschließlich typische OCR-Fehler und deren Gewichtung berücksichtigt. Bei der Entscheidung für ein von gegebenenfalls mehreren naheliegenden Wörtern soll darüber hinaus die Worthäufigkeit berücksichtigt werden. Dabei ergibt sich annäherungsweise ein OCR-Fehler berücksichtigender Wortabstand.[37]

34 Aus Kostengründen wurde auf eine Evaluation mit einem Evaluations-Ground-Truth verzichtet. Während des Projektes bestätigte jedoch die Parametrisierung auf einer Teilmenge des Ground Truth mit Evaluation auf der Ground-Truth-Restmenge die Verallgemeinerbarkeit der gefundenen Parametrisierung.

35 https://github.com/suub/bote/blob/99d845dd390c668e3b47813059a8da22d77f1e0c/resources/current-params/substs.edn.

36 Zum Beispiel die Levenshtein-Distanz: https://de.wikipedia.org/wiki/Levenshtein-Distanz.

37 Da diese Operationen (Substitutionen) für verschiedene Richtungen nicht zwangsläufig einen identischen Beitrag zur Distanz liefern müssen, erfüllt der so definierte Wortabstand streng mathematisch nicht die für eine Abstandsfunktion notwendige Symmetrie-Bedingung.

Die Handlungsvorschrift wurde in der Programmiersprache clojure[38] funktional kodiert. Die wie folgt skizzierte Liste von einzelnen Schritten wurde beim *Grenzboten* für jedes der knapp 80 Millionen Wörter (186.740 Dateien bzw. *Grenzboten*-Seiten) mit einem Zeitaufwand von insgesamt 4 Stunden und 15 Minuten durchlaufen.[39]

1. das gegebene Wort wird mit der Liste der ca. 1,7 Millionen historischen Wortformen abgeglichen
2. ist dieses Wort in dem Wörterbuch nicht vorhanden, werden mit der Liste der Zeichensubstitutionen Wörter erzeugt, die potenziell korrekt sein können (im Sinne von Kandidaten)
3. die gefundenen Verbesserungskandidaten werden nach Wortabstand[40] sortiert
4. eine Bewertungsfunktion berücksichtigt diese Sortierung sowie die Worthäufigkeit des gefundenen Wortes. Der danach beste Kandidat wird als korrigiertes Wort ausgewählt

Am Beispiel des Wortes „gelden", das selbst eine historische Schreibung des Wortes „gelten" darstellt, soll die Liste der sortierten Verbesserungskandidaten veranschaulicht werden: gelten, gelben, gelden, gelder, gelbem, gelteu, gelber, getten, geiten.[41] Zum Ende der Liste nimmt die Anzahl der Zeichensubstitutionen zu und es werden im Sinne der Gewichtung der Zeichensubstitutionen „unwahrscheinlichere" Substitutionen verwendet. Wenn „gelden" selbst nicht in der Liste der historischen Wortformen vorkommt (eine Entscheidung auf der Basis des Erscheinungsjahrhunderts des Textes), würde eine Korrektur zugunsten von „gelten" vorgenommen werden.

Die an der SuUB Bremen prototypisch entwickelte Software wurde in GitHub[42] unter https://github.com/suub/ eingestellt. Die Software besteht aus drei Komponenten error-codes[43], ocr-visualizer[44] und bote[45]. Die Komponente Error-Codes berechnet Zeichen- und Worterkennungsquoten im Vergleich mit dem Ground Truth Text. Der OCR-visualizer ist eine webbasierte Visualisierung der OCR-Fehler, die ebenfalls auf dem Vergleich mit dem Ground Truth Text basiert. Die eigentliche Nachverbesserung des OCR-Volltextes mit dem vorgestellten Algorithmus ist in der Softwarekomponente „Bote" vorhanden. Diese prozessiert alle ABBYY-xml-Dateien eines gegebenen Verzeichnisses, wendet den beschriebenen Korrekturalgorithmus auf jedes Wort an und gibt wiederum Dateien im ABBYY-xml-Format aus. Dabei werden umgebrochene Wörter als vollständiges Wort bearbeitet und anschließend wieder umgebrochen.

38 Clojure Website, zuletzt geprüft am 01.02.2016, http://clojure.org/; ein Auszug aus https://de.wikipedia.org/wiki/Clojure: „Clojure ist ein moderner Lisp-Dialekt, der interaktive Entwicklung unterstützt. […] Clojure läuft in der Java Virtual Machine …".

39 Verwendet wurde ein aktueller Desktop-PC (4 Prozessoren, Intel® Core™ i5, 3 GHz, 8 GB Hauptspeicher). Die Software des *Bremer Ansatzes* wurde für Parallelprozessierung entwickelt, d.h. alle verfügbaren Prozessoren werden beim Korrekturlauf gleichermaßen genutzt.

40 Beim Wortabstand geht das Produkt der Gewichte aller Zeichensubstitutionen ein.

41 Die Liste der sortierten Verbesserungskandidaten inklusive Bewertung: ([„gelten" 1913/29969870] [„gelben" 211/12844230] [„gelden" 1/8990961] [„gelder" 13/89909610] [„gelbem" 53/29969870] [„gelteu" 1/89909610] [„gelber" 1/89909610] [„getten" 1/44954805] [„geiten" 1/89909610])

42 https://github.com/suub/bote/tree/master.

43 https://github.com/suub/error-codes.

44 https://github.com/suub/ocr-visualizer.

45 https://github.com/suub/bote.

Weiterhin wurde ein GUI-Tool OCR-Evaluator[46] entwickelt, das die oben erwähnten Komponenten bündelt. Es ermöglicht die einfache Bestimmung von Zeichen- und Worterkennungsquoten, deren Visualisierung und Nachkorrektur auf Plain-Text-Basis. Alle Softwarekomponenten sind uneingeschränkt nutzbar.[47]

5. Ergebnisse

Die Tab. 2 veranschaulicht einzelne Korrekturbeispiele anhand des Druckbildes, dem Ground Truth Text, der Ausgabe der OCR und der Korrektur bzw. nicht-Korrektur durch den *Bremer Ansatz*. So wurden die Worte „Ader“ und „dein“ im OCR-Text nicht zu „Aber“ und „dem“ korrigiert (der Fall falsch-negativ), da jeweils beide Wörter in der Liste der historischen Wortformen enthalten sind (vgl. die im Abschnitt 3. angegebene Heuristik 1 „Es werden keine Wortformen korrigiert, die selber in der Liste historischer Wortformen enthalten sind“). Die nächsten vier Beispiele entsprechen dem Fall einer richtig-positiven Korrektur. Bei dem Wort „Entwickelung“ wurde der Multizeichenfehler „im/un“ korrigiert. Das Wort „Sciu“ in der letzten Zeile wurde im ersten Zeichen „geringfügig“ verschlechtert (der Fall falsch-positiv) und an zwei weiteren Zeichen verbessert (richtig-positiv). Bereits die gebräuchlichste Form fehlertoleranter Suche – die Suche ohne Unterscheidung von Groß- und Kleinschreibung – würde die fehlerhafte Kleinschreibung des Wortes „sein“ tolerieren.

Tab. 2: Korrekturbeispiele

Druckbild	Ground Truth	OCR-Text	Bremer Ansatz
	Aber	Ader	Ader
	dem	dein	dein
	Zeitschrift	Zeitschrist	Zeitschrift
	gewonnen	gcwouneu	gewonnen
	Entwickelung	Eutwicklung	Entwickelung
	Herz der begeisterten	Herz dcr bcgeisterteu	Herz der begeisterten
	Sein Wille erfüllte	Sciu Willc crsültc	sein Wille erfüllte

Die Ergebnisse in Bezug auf Zeichen- und Worterkennungsquoten für verschiedene Abschnitte aus verschiedenen Jahrgängen vor und nach der Korrektur sind in der folgenden Tab. 3 angegeben. Sie basieren auf den insgesamt 370 Seiten Ground Truth Text[48] und wurden mit dem an der SuUB Bremen entwickelten Softwaretool OCR-Evaluator berechnet.

46 https://github.com/suub/ocr-evaluator.

47 Software-Repositories im Filehosting-Dienst GitHub dokumentieren die die Nutzung regelnde Lizenz in der LICENSE Datei eines Software-Entwicklungsprojekts. Die LICENSE Datei beispielsweise für die Komponente OCR-Evaluator basiert auf der MIT-Lizenz: https://github.com/suub/ocr-evaluator/blob/master/LICENSE.

48 Es wurde angenommen, dass der Ground Truth Text für den *Grenzboten* repräsentativ ist. Sollte dies nicht der Fall sein, wären die Ergebnisse geringfügig schlechter. Dazu müssten sich jedoch in den bei der Erstellung des Ground Truth Textes unberücksichtigten Abschnitten Zeichensubstitutionen überproportional häufen. Aus Kostengründen wurde auf die Erstellung weiterer Ground Truth Texte zur ausschließlichen Evaluation, d.h. ohne Verwendung bei der Parametrisierung, verzichtet.

Tab. 3: Zeichen- und Worterkennungsquoten für verschiedene Abschnitte aus verschiedenen Jahrgängen vor und nach der Korrektur

Jahrgang	Seitenanzahl	Ausgangsquoten		nach Korrektur	
		Zeichen	Wörter	Zeichen	Wörter
1841 + 1842	352	98,27 %	94,82 %	98,81 %	97,23 %
1870	11	99,42 %	98,26 %	99,52 %	98,32 %
1900	9	97,52 %	92,51 %	98,82 %	96,45 %

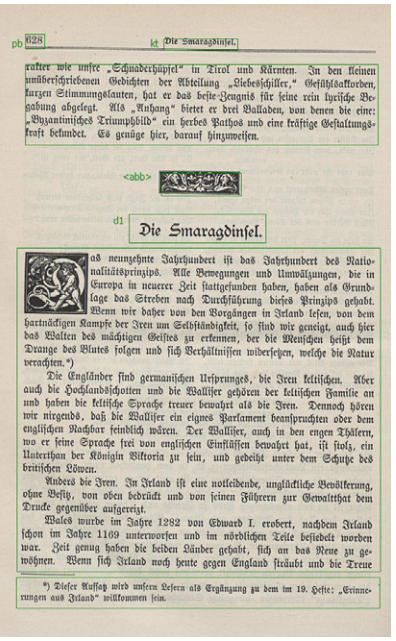
Bezogen auf den gesamten Ground Truth Text wurde ausgehend von einer Zeichenerkennungsquote von 98,28 % eine Erkennungsquote von 98,83 % erreicht. D.h. 32 % aller Fehler wurden eliminiert und in dem aus 517 Millionen Zeichen bestehenden *Grenzboten* wurden 2,84 Millionen Zeichenfehler automatisiert korrigiert. Die folgende Abbildung veranschaulicht ein beispielsweise gutes Korrekturergebnis auf einer Grenzboten-Seite.[49] Die Anzahl hellblau markierter Zeichenfehler im OCR-Text (links) wurde hier wesentlich reduziert.

Abb. 6: Veranschaulichung der OCR-Nachkorrektur

49 http://brema.suub.uni-bremen.de/periodical/pageview/179393

5.1. Konvertierung des *Grenzboten*-Korpus in das TEI P5 Datenformat

Zu allen über 185.000 Seiten wurde als Ergebnis des OCR-Nachbearbeitungsprojektes eine textformale Auszeichnung erstellt. Damit ist die Auszeichnung von Seitenelementen wie Absätze, Rubrikenüberschriften, Abbildungen und Fußnoten gemeint. Ein wesentliches Ziel ist dabei, einen durchgängigen Fließtext zu erhalten, der nicht durch die übrigen Seitenelemente unterbrochen wird. Diese Auszeichnung ist die Grundlage für die Konvertierung des *Grenzboten*-Korpus in das auf TEI P5 Richtlinien basierende DTA-Basisformat.[50]


pb 628 lt Die Smaragdinsel.

rakter wie unsre „Schnaderhüpfel" in Tirol und Kärnten. In den kleinen unüberschriebenen Gedichten der Abteilung „Liebesschiller," Gefühlsakkorden, kurzen Stimmungslauten, hat er das beste Gengnis für seine rein lyrische Begabung abgelegt. Als „Anhang" bietet er drei Balladen, von denen die eine: „Byzantinisches Triumphbild" ein herbes Xathos und eine kräftige Gestaltungskraft bekundet. Es genüge hier, darauf hinzuweisen.

<abb>

d1 Die Smaragdinsel.

as neunzehnte Jahrhundert ist das Jahrhundert des Nationalitätsprinzips. Alle Bewegungen und Umwälzungen, die in Europa in neuerer Zeit stattgefunden haben, haben als Grundlage das Streben nach Durchführung dieses Prinzips gehabt. Wenn wir daher von den Vorgängen in Irland lesen, von dem hartnäckigen Kampfe der Iren um Selbständigkeit, so sind wir geneigt, auch hier das Walten des mächtigen Geistes zu erkennen, der die Menschen heißt dem Drange des Blutes folgen und sich Verhältnissen widersetzen, welche die Natur verachten.*)

Die Engländer sind germanischen Ursprunges, die Iren keltischen. Aber auch die Hochlandsschotten und die Walliser gehören der keltischen Familie an und haben die keltische Sprache treuer bewahrt als die Iren. Dennoch hören wir nirgends, daß die Walliser ein eignes Parlament beansprüchten oder dem englischen Nachbar feindlich wären. Der Walliser, auch in den engen Thälern, wo er seine Sprache frei von englischen Einflüssen bewahrt hat, ist stolz, ein Unterthan der Königin Viktoria zu sein, und gedeiht unter dem Schutze des britischen Löwen.

Anders die Iren. In Irland ist eine notleidende, unglückliche Bevölkerung, ohne Besitz, von oben bedrückt und von seinen Führern zur Gewaltthat dem Drucke gegenüber aufgereizt.

Wales wurde im Jahre 1282 von Edward I. erobert, nachdem Irland schon im Jahre 1169 unterworfen und im nördlichen Teile besiedelt worden war. Zeit genug haben die beiden Länder gehabt, sich an das Neue zu gewöhnen. Wenn sich Irland noch heute gegen England sträubt und die Treue

*) Dieser Aufsatz wird unsern Lesern als Ergänzung zu dem im 19. Hefte: „Erinnerungen aus Irland" willkommen sein.

Abb. 7: Beispiel einer textformalen Auszeichnung

Dazu wurden vom Deutschen Textarchiv (Zentrum Sprache an der BBAW[51]) auf Bildkoordinaten basierte Rahmen zu Absätzen, Fußnoten, etc. positioniert (vgl. Abb. 7) und entsprechend semantisch ausgezeichnet. In einem automatisierten Prozess wurden diese Bildkoordinaten mit den im ABBYY-xml Format enthaltenen Koordinaten ebenfalls am Zentrum Sprache vereinigt. Auf diese Weise konnten *Grenzboten*-Seiten im DTA-Basisformat erstellt werden. Das in Bezug auf Zeichenfehlerquote und Textstruktur optimierte *Grenzboten*-Korpus soll in CLARIN-D frei von Urheberrechten unter Public Domain Mark 1.0 verfügbar gemacht werden.

5.2. Der cloud service overProof und das Post Correction Tool – PoCoTo

Mit dem Vortrag „*Correcting Noisy OCR: Context beats Confusion*"[52] wurde auf der Konferenz DATeCH-2014[53] in Madrid der cloud service overProof[54] von John Evershed und Kent Fitch (Firma ProjectComputing, Canberra, Australien) vorgestellt. Dieser webbasierte Korrekturservice für englischsprachige OCR-Texte wurde im Rahmen einer Zusammenarbeit mit der SuUB Bremen für die Korrektur von OCR-Text deutschsprachiger Frakturtexte weiterentwickelt.

Der algorithmische Ansatz des cloud service overProof enthält Wortkontext berücksichtigende Korrekturkriterien. Dieser Ansatz eliminiert die bisher erwähnten Nicht-Korrekturen durch fehlenden Kontext (der Fall falsch-negativ). Weiterhin besitzt overProof Entscheidungskriterien, Leerzeichen einzufügen bzw. zu löschen, d.h. Wörter zu verbinden bzw. zu trennen. Auch darin liegt ein Potenzial, weitere Nicht-Korrekturen zu eliminieren, aber es erscheinen auch neue Typen von falsch-positiven

50 DTA-Basisformat, zuletzt geprüft am 01.02.2016, http://www.deutschestextarchiv.de/doku/basisformat.
51 BBAW, Zentrum Sprache, zuletzt geprüft am 01.02.2016, http://www.bbaw.de/forschung/zentren/sprache. Das Deutsche Textarchiv (DTA) ist hier angesiedelt.
52 Evershed und Fitch, „Correcting Noisy OCR".
53 Siehe Anm. 18.
54 Siehe Anm. 19.

Korrekturen: Fehlerhaft verbundene bzw. fehlerhaft getrennte Wörter. Die Korrekturergebnisse durch overProof sind im Vergleich mit dem *Bremer Ansatz* als mindestens gleichwertig zu bewerten.[55] Sie mögen die untere Grenze des Erreichbaren sein, da inzwischen der cloud service overProof kontinuierlich weiterentwickelt wurde.

Ein Gesamtkorrekturdurchlauf aller ca. 187.000 Dateien benötigte einen Zeitaufwand von 3 Tagen und 21,6 Stunden. Die Bearbeitung durch overProof erfolgte im Rahmen des Projektes ohne Kostenberechnung und wurde als Teststellung betrachtet. Für die knapp 200.000 Seiten der Zeitschrift *Die Grenzboten* hätte die OCR-Korrektur nach aktuellen regulären Firmen-Konditionen Kosten in Höhe von ca. 820$ verursacht.[56] Das marktfähige System vermeidet als Webservice Aufwände beim Auftraggeber in Bezug auf Servermanagement, Installation, Konfiguration und Systempflege. Als Standard für OCR-Volltext arbeitet overProof mit Dateien im ALTO-Format. Im Rahmen der Kooperation wurde für dieses Projekt eine Anpassung für Dateien im ABBYY-XML-Format vorgenommen. Verfügbar ist derzeit ein Dienstleistungsangebot von overProof mit Ablaufbeschreibung, Beispielen und Kostenrahmen für englischsprachige Nachkorrektur.[56]

Ein weiteres vielversprechendes Open Source Softwaretool zur OCR-Nachkorrektur ist das am *Center for Information and Language Processing*[57] der Ludwig-Maximilians-Universität München entwickelte Softwaretool PoCoTo.[58] Es lässt den Grad zwischen voller optischer Kontrolle und automatisierten Korrekturen frei wählen. Durch die Systemarchitektur (Trennung von serverbasiertem Text and Error Profiler sowie der Grafischen Oberfläche (GUI) PoCoTo) hat das Tool sehr gute Anlagen, sich problemlos in die Workflows von OCR-Projekten zu integrieren. Zum Zeitpunkt der Nachkorrektur des *Grenzboten* stand PoCoTo nicht zur Verfügung.

6. Weitere in Digitalisierungsprojekten einsetzbare Ergebnisse

Digitalisierungsprojekte bieten ein großes Potenzial für den Einsatz von automatisierter Unterstützung bei zahlreichen Arbeitsgängen, wie Strukturierung, Lückenidentifikation und Qualitätssicherung. In Sommer et al. wurde mit Methoden der Bildverarbeitung ein automatisches Verfahren für eine Vorsegmentierung zur Unterstützung der Strukturierung angewendet.[59] Hier soll für die Lückenidentifikation

55 Die Unsicherheit in dieser Aussage begründet sich in den automatisiert ermittelten Zeichenerkennungsquoten. Im Falle der Ergebnisse von overProof wurden zahlreiche Korrekturfehler von einem Satzzeichen zu einem Leerzeichen bzw. von mehrere Leerzeichen zu einem identifiziert und mitgezählt, deren Umfang bis Projektende nicht abgeschätzt werden konnte.

56 Dienstleistungsangebote von overProof, zuletzt geprüft am 01.02.2016, http://overproof.projectcomputing.com/about. Der Umfang der Dienstleistung wird als Anzahl hochgeladener Wörter pro Monat definiert. Dabei ist der Service für Massendigitalisierung ausgelegt. Das dokumentierte Kostenmodell reicht in den Bereich von einer Milliarde Wörtern, was am Beispiel des *Grenzboten* ungefähr 2,3 Millionen Seiten entspräche. Die Kosten für den Seitenumfang der Zeitschrift *Die Grenzboten* (ca. 187.000 Seiten bzw. ca. 80 Million Wörter) würden sich auf ungefähr 820$ beziffern.

57 Siehe Anm. 21.

58 Florian Fink, *Postcorrection Tool (PoCoTo) Manual*, Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig-Maximilians-Universität München, zuletzt geändert am 25.08.2015, zuletzt geprüft am 23.02.2016, https://github.com/cisocrgroup/Resources/blob/master/manuals/.

59 Dorothea Sommer, Kay Heiligenhaus, Carola Wippermann und Manfred Pankratz, „Zeitungsdigitalisierung: eine neue Herausforderung für die ULB Halle“, *ABI Technik* 34, Heft 2 (Juli 2014): 81, http://dx.doi.org/10.1515/abitech-2014-0013.

und die halbautomatisierte Qualitätssicherung exemplarisch verdeutlicht werden, welchen weiteren Nutzen Digitalisierungsprojekte aus den Informationen ziehen können, die in OCR-Volltext enthalten sind. Die Unterstützung von Arbeitsgängen der Digitalisierung mit Hilfe von OCR-Text ist nichts Neues. So gibt es bereits Managementsoftware für Digitalisierung, die beispielsweise OCR-basiert die Paginierung[60] unterstützt.

6.1. Durch OCR-Volltext unterstützte Lückenidentifikation und Qualitätssicherung

Am Beispiel der Lückenidentifikation und der halbautomatisierten Qualitätssicherung im Rahmen von Digitalisierungsprojekten wurde die Nutzbarkeit von OCR-Volltext konzeptionell und prototypisch evaluiert. Die Qualität digitalisierter Images wurde dabei automatisiert bewertet. Die Qualitätsbewertung basiert auf der Annahme, dass eine hohe Wortfehlerquote auf ein mangelhaftes Digitalisierungsergebnis hindeutet. Diese Fehlerquote wurde mit Hilfe der bereits erwähnten Liste historischer Wortformen ermittelt.

Es konnten in der Zeitschrift *Die Grenzboten* die in Abb. 8 dargestellten auffälligen und in Bezug auf Volltextqualität problematischen Typen von Seiten identifiziert werden.

Abb. 8: Beispiele automatisch identifizierter auffälliger Seiten

Anzeigenseiten waren überwiegend in Antiqua gesetzt. Anders als der aktuelle ABBYY Recognition Server war der im Jahr 2012 verwendete FineReader 9 nicht in der Lage zwischen Frakturschrift- und Antiquaschrift-Erkennung umzuschalten.[61] Bei Inhaltsverzeichnissen und Tabellen wurden kleine Schriftgrößen verwendet, was ebenfalls auffällig hohe Wortfehlerquoten generiert hat. Bei

60 Paginierung: Zuordnung von Seitenzahlen zu digitalisierten Bilddateien.

61 Bei einem mit dem ABBYY RecognitionServer durchgeführten Test zu einer *Grenzboten*-Seite mit gemischter Fraktur- und Antiquaschrift wurden die Antiqua-Anteile zwar deutlich besser erkannt, die Frakturschrift-Erkennung hat sich jedoch verschlechtert.

Fremdsprachen, durchscheinender Schrift, schlechtem Schriftbild, kaputten Seiten und geringem Kontrast kann eine OCR-Software selbstverständlich keine guten Ergebnisse hervorbringen. Die möglichst präzise und vollständige automatisierte Identifikation dieser Fehlerquellen würde es einem Digitalisierungsprojekt jedoch ermöglichen, entsprechende Maßnahmen vorzunehmen: Digitalisierung mit höherer Auflösung, Einsatz einer OCR für die entsprechende Sprache, Verwendung einer problemspezifischen Bildvorverarbeitung[62] oder Lückenergänzung.

Leicht umsetzbar wäre es beispielsweise auch, doppelt eingescannte Seiten automatisiert mit Hilfe von OCR-Text zu identifizieren. Die Analyse von OCR-Volltext ist weniger rechenaufwändig als vergleichbare auf den Bilddateien basierende Verfahren.

7. Diskussion

Hier soll diskutiert werden, wie der *Bremer Ansatz* und die Ergebnisse zu bewerten sind und was das Fazit ist. Es soll betrachtet werden, welche Schlüsse man für weitere Vorhaben ziehen kann und welche Desiderate sich aus diesem und anderen Projekten ergeben.[63]
Mit dem *Bremer Ansatz* wurde ein Softwareprototyp zur Korrektur von OCR-Volltext vorgestellt, der deutliche Verbesserungen der Textqualität erzielt. Die erreichte Erkennungsquote von 98,83 % entspricht einer Korrektur von 32 % aller Fehler im Gesamtkorpus. Das selbst gesteckte Ziel einer Zeichenerkennungsquote von 99,5 % wurde jedoch nicht erreicht. In den folgenden Faktoren liegt die Begründung dafür, dass das erzielte Ergebnis nicht besser ausgefallen ist:

1. Ausgangsqualität des OCR-Volltextes[64]
2. Leistungsfähigkeit des verwendeten Algorithmus bzw. des gesamten Ansatzes
3. Umfang bzw. Vollständigkeit und Qualität der bei der Parametrisierung verwendeten Angaben
   3.1. Liste der historischen Wortformen
   3.2. Liste der Zeichensubstitutionen (das Fehlermodell)
   3.3. Parametrisierung der Tokenisierung

In diesem Abschnitt soll kurz angerissen werden, an welchen Stellen das Potenzial einer Leistungssteigerung sowie einer Qualitätsverbesserung liegt bzw. welche vielversprechenden im Projekt nicht verfolgten Ansätze existieren. Die Leistungsfähigkeit des *Bremer Ansatzes* zusammen mit zwei der

62 Beispielsweise Verwendung einer angepassten Binarisierung (Konvertierung eines Bildes in ein bitonales Bild) bei durchscheinender Schrift bzw. bei Seiten mit geringem Kontrast.

63 Evershed und Fitch, „Correcting Noisy OCR"; Federbusch und Polzin, *Volltext*; Furrer und Volk, „Reducing OCR Errors"; Mühlberger, „Digitalisierung historischer Zeitungen"; Sommer, Heiligenhaus, Wippermann und Pankratz, „Zeitungsdigitalisierung"; Stäcker, „Konversion"; Maria Wernersson, „Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung", *ABI Technik* 35, Heft 1 (2015): 23–35, http://dx.doi.org/10.1515/abitech-2015-0014.

64 Das Korrekturpotenzial kann bei verschiedenen Niveaus von Zeichenfehlerquoten höher oder niedriger ausfallen. Die Ergebnisse zu verschiedenen Abschnitten aus dem *Grenzboten* haben gezeigt, dass das Korrekturpotenzial bei steigender Zeichenfehlerquote steigt. So konnte zu einem Abschnitt aus dem Jahrgang 1900, mit der Zeichenfehlerquote 2,48 %, mehr als die Hälfte der Fehler korrigiert werden. Hingegen konnte ein Abschnitt aus dem Jahr 1870, mit der geringsten Zeichenfehlerquote 0,58 %, lediglich auf ein Niveau von 0,48 % korrigiert werden, d.h. nur ca. 17 % der Fehler.

verwendeten Parametrisierungen (Faktoren 3.2 und 3.3) wurde mit einer experimentellen Analyse bewertet. Mit einem hypothetisch bestmöglich erfüllten Faktor 3.1, d.h. mit einer vollständigen und fehlerfreien Liste der historischen Wortformen wurde, basierend auf 370 Seiten Ground Truth Text, eine Zeichenerkennungsquote von 99,22 % erzielt.[65] Das bedeutet, dass die Liste der historischen Wortformen einen wesentlichen Einfluss auf die erreichbare Korrekturqualität hat, aber sie kann nicht alleine weitere Qualitätssteigerungen darüber hinaus darstellen.

Weitere Komponenten mit Potenzial für eine Leistungssteigerung wurden bereits erwähnt: das Einfügen/Entfernen von Leerzeichen sowie Wortkontext berücksichtigende Ansätze. Die folgenden sieben Beispiele ausgelassener Leerzeichen sollen verdeutlichen, dass die jeweils getrennten Wörter mit dem *Bremer Ansatz* fast vollständig hätten korrigiert werden können:

erstauntenDeutschlaude	Deutschlaudeinen
Dentschlandsmüsse	Dentschlandkann
Dentschlandssorgen	EntKicklungDentschlands
CentralisationDentfchlands	

So ist beispielsweise bei „EntKicklungDentschlands“ das Wort „Dentschland“ korrigierbar, da sich Deutschland in der Wortformenliste befindet und die Zeichensubstitution n/u existiert. Das Wort „EntKicklung“ wäre nicht korrigiert worden, da die Zeichensubstitution K/w nicht in der Liste der Zeichensubstitutionen[35] enthalten ist. Eine Abschätzung für das darin liegende weitere Korrekturpotenzial wurde nicht ermittelt. Die Komplexität eines Korrekturalgorithmus, der die Fälle verschmelzender und aufgeteilter Wörter berücksichtigt, wäre nicht unerheblich und birgt die Gefahr eines neuen Typs von falsch-positiven Korrekturen.

Beim *Bremer Ansatz* wurden nur Wortformen korrigiert, die selbst nicht in der Liste der historischen Wortformen enthalten waren. Damit ergeben sich ausgelassene Korrekturen (der Fall falsch-negativ), wie „Aber/Ader“, „dem/dein“, „Hans/Haus“ etc., die nur über einen Wortkontext berücksichtigenden Ansatz bewältigt werden können. Der oben erwähnte cloud service overProof adressiert beide hier erwähnten Möglichkeiten der Leistungssteigerung.

Im Folgenden werden Bedarfe und Möglichkeiten für zukünftige OCR-Nachkorrekturprojekte aufgezeigt. Bei allen zur Verfügung stehenden Ressourcen und Werkzeugen ist es sehr wichtig, im Blick zu behalten, mit welchem Aufwand ein gegebener OCR-Volltext von einer Ausgangserkennungsquote x auf eine Quote y verbessert werden kann. „Gerade die Kostenfrage dürfte ein Schlüsselelement in der Bewertung von OCR-Verfahren bilden, denn offenbar gibt es in diesem Bereich einen *Pareto-Effekt*, der dazu führt, dass der Wunsch nach hoher Textgenauigkeit die Kosten gerade für den Schritt zu sehr guten Texten sprunghaft ansteigen lässt.“[66] Wie bereits erwähnt, kann bei Volltextprojekten

65 Es wurde ausschließlich der Ground Truth Text mit einer aus dem Ground Truth Text selbst zusammengestellten Wortformenliste korrigiert und mit dem oben vorgestellten OCR-Evaluator bewertet. Das Ergebnis 99,22 % stellt eine theoretische Schätzung dar. Eine Substitutionsliste mit Stand vom Januar 2015 konnte noch geringfügig verbessert werden, würde diese Schätzung jedoch nur minimal betreffen.

66 Stäcker, „Konversion“, 129–130.

in der Größenordnung von mehreren Millionen Seiten in Bezug auf Aufwand und Kosten nur eine möglichst weitgehende Automatisierung der Nachbearbeitung von OCR-Volltext zielführend sein. In dem hier vorgestellten Volltextprojekt ist sicher deutlich geworden, dass Softwaretools manuelle und intellektuelle Anteile nicht auf null reduzieren. Schlussfolgerungen aus der Kenntnis des betrachteten Volltext-Korpus, die Auswahl und Parametrisierung der Softwaretools sowie die Konzeption des gesamten Projektes bleiben nicht-automatisierbare Anteile des gesamten Vorhabens. Da es sich um einmalige Aufwände handelt, ergibt sich zumindest die Möglichkeit möglichst große homogene Korpora bzw. Sammlungen für eine Volltexterfassung vorzusehen. So erhält man ein optimales Verhältnis von Aufwand und Nutzen.[67]

Weiter verbesserte OCR-Systeme werden es erlauben, die OCR-Textqualität auch ohne OCR-Nachkorrektur zu steigern. Nimmt man jedoch steigende Anforderungen an Volltextqualität und den Bedarf nach Volltextprojekten zu Material mit schlechterem Schriftbild (z.B. bei Handschriftenerkennung[68]) an, dann wird sich eine Nachfrage nach OCR-Korrektur erhalten. Sicher wird sich die Ausprägung der OCR-Fehler ändern und erfordert angepasste Fehlermodelle sowie Parametrisierungen der Korrektursysteme.

Folgende Desiderate für OCR- und OCR-Nachkorrektursysteme ergeben sich aus den Erfahrungen des hier vorgestellten Projektes:

1. Eine Umschaltung zwischen Fraktur und Antiqua ohne Nachteile bei der Erkennungsqualität.
2. Die Identifikation von abweichenden Schriften (wie z.B. griechisch) und Sprachen.
3. Die pragmatische Zusammenstellung und freie Verfügbarkeit von Wortformenlisten für verschiedene Jahrhunderte sowie für verschiedene Sprachen und Dialekte.

Es wäre wünschenswert, wenn Bibliotheken, weitere Forschungsinfrastrukturen und Crowdsourcing-Projekte (wie z.B. Wikisource) intensiv zusammenarbeiten würden, um diese Ziele anzugehen.[69] Die Existenz des aktuell laufenden DFG-geförderten Koordinierungsprojektes „Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR)“[70] beweist, dass die Volltexterstellung mit OCR weiterhin ein zentrales Thema für Digitalisierungsvorhaben sein wird.

67 Wawra und Wündrich, „OCR für Frakturschriften?“, 2117.

68 Vgl. die Präsentation des OCR-Programms *Transkribus* zur Handschriftenerkennung, vorgestellt von Günter Mühlberger im Rahmen des internationalen Workshops „Digitizing German-Language Cultural Heritage from Eastern Europe“ (27.-28.04.2015, Regensburg), zuletzt geprüft am 01.02.2016, http://minorecs.hypotheses.org/137: „Günter Mühlberger (University Library Innsbruck) introduced the workshop's participants to a milestone of digitization technology: with the OCR program *Transkribus* it is possible for the first time to automatically recognize handwritten texts. The program is still in development and in need of the input of interested users: after registration the user can upload digitized materials into the program which serves as a training for improving the recognition accuracy. *Transkribus* needs about 100 pages in order to learn the individual traits of a handwriting and read it properly. The user can correct the recognized text and thus emend the further recognition of additional texts. It is easy to foresee that *Transkribus* will become in the near future a standard tool for the deeper indexing of medieval and early modern handwritten documents.“

69 Vgl. das Fazit aus Stäcker, „Konversion“, 235.

70 Projekt „OCR-D“ und Projekt „Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR)“, zuletzt geprüft am 01.02.2016, http://www.ocr-d.de/ und http://gepris.dfg.de/gepris/projekt/274863866.

7.1. Fazit

- Mit dem *Bremer Ansatz* wurde ein Softwareprototyp zur Korrektur von OCR-Volltext vorgestellt, der deutliche Verbesserungen der Textqualität erzielt.
- Eine in Bezug auf Korrektheit und Umfang optimierte Liste von historischen Wortformen würde das Ergebnis weiter verbessern.
- Mit dem kostenpflichtigen cloud service overProof und dem Open Source Tool PoCoTo stehen weitere Angebote zur OCR-Nachkorrektur zur Verfügung.
- Das im TEI P5 Datenformat in CLARIN-D integrierte Grenzboten-Korpus ermöglicht eine bestmögliche Beforschung.

Literaturverzeichnis

– Eversheed, John und Kent Fitch. „Correcting Noisy OCR: Context Beats Confusion.“ In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 45–51. New York: ACM, 2014. http:dx.doi.org/10.1145/2595188.2595200.

– DFG-Praxisregeln Digitalisierung. DFG-Vordruck 12.151 – 02/13. Zuletzt geprüft am 01.02.2016. http://www.dfg.de/formulare/12_151/12_151_de.pdf.

– Federbusch, Maria und Christian Polzin. *Volltext via OCR – Möglichkeiten und Grenzen*. Beiträge aus der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz 43. Berlin: Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, 2013. Zuletzt geprüft am 01.02.2016. http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf.

– Fink, Florian. *Postcorrection Tool (PoCoTo) Manual*. Centrum für Informations- und Sprachverarbeitung (CIS), Ludwig-Maximilians-Universität München. Zuletzt geändert am 25.08.2015. Zuletzt geprüft am 23.02.2016. https://github.com/cisocrgroup/Resources/blob/master/manuals/.

– Furrer, Lenz und Martin Volk. „Reducing OCR Errors in Gothic-Script Documents.“ *ERCIM News* 86 (2011): 29–30. http:dx.doi.org/10.5167/uzh-49203.

– Kann, Bettina und Michael Hintersonnleitner. „Volltextsuche in historischen Texten – *Erfahrungen aus den Projekten der Österreichischen Nationalbibliothek*.“ *BIBLIOTHEK – Forschung und Praxis* 39, Heft 1 (2015): 73–79. http:dx.doi.org/10.1515/bfp-2015-0004.

– Kilner, Kerry und Kent Fitch. „Discovering and Rediscovering Full Text: Unearthing and Refactoring.“ Zuletzt geprüft am 01.02.2016. http://dh2015.org/abstracts/xml/KILNER_Kerry_Discovering_and_Rediscovering_Full_T/KILNER_Kerry_Discovering_and_Rediscovering_Full_Text__U.html.

- Mühlberger, Günter. „Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Strukturerkennung (OCR).“ *Zeitschrift für Bibliothekswesen und Bibliographie* 58, Nr. 1 (2011): 10–18. http://dx.doi.org/10.3196/186429501158135.

- Sommer, Dorothea, Kay Heiligenhaus, Carola Wippermann und Manfred Pankratz. „Zeitungsdigitalisierung: eine neue Herausforderung für die ULB Halle.“ *ABI Technik* 34, Heft 2 (Juli 2014): 75–85. http://dx.doi.org/10.1515/abitech-2014-0013

- Stäcker, Thomas. „Konversion des kulturellen Erbes für die Forschung: Volltextbeschaffung und -bereitstellung als Aufgabe der Bibliotheken.“ *o-bib* 1, Nr. 1 (2014): 220–237. http://dx.doi.org/10.5282/o-bib/2014H1S220-237.

- Wawra, Steffen und Silke Wündrich. „OCR für Frakturschriften?“ *Bibliotheksdienst* 33 (1999): 2110–2117.

- *Wernersson, Maria. „Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung“. ABI Technik* 35, Heft 1 (2015): 23–35. http://dx.doi.org/10.1515/abitech-2015-0014.