OCR

Günter Mühlberger

Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Struktur-erkennung (OCR)

Foto: privat

Günter Mühlberger

Die OCR Erkennung ist eine Schlüsseltechnologie, an der man bei der systematischen Digitalisierung von historischen Zeitungen nicht vorbeikommen wird. Obwohl vielfach nur eine Wortgenauigkeit von 80 % oder weniger für Zeitungen des 19. und 20. Jahrhunderts zu erzielen sein wird, bietet dieser fehlerhafte Volltext trotzdem die Grundlage für eine ganze Reihe interessanter Anwendungen – von der Volltextsuche, über die Indexierung durch Suchmaschinen bis zur Online-Korrektur durch Benutzer. Der Einsatz der OCR erfordert allerdings sowohl bei der Projektplanung, der Gestaltung des Workflows, der Durchführung der Qualitätskontrolle als auch der Konzeption der Langzeitarchivierung und der Präsentation im Internet ein Umdenken gegenüber herkömmlichen Digitalisierungsprojekten.

OCR recognition is a key technology which cannot be circumvented when systematically digitizing historical newspapers. Although often achieving a word accuracy of only 80 % or less for newspapers of the 19th and early 20th century, these imperfect files nevertheless provide a basis for a number of interesting applications – from full-text searching to indexing by search engines and online correction by users. However, in comparison to traditional digitization projects, the use of OCR requires a fundamental change of thinking during the project planning, the design of the workflow, the implementation of quality control, and in the designing of long-term preservation and presentation of digitized material on the Internet.

Einleitung

Problemfeld historische Schriften

Optical Character Recognition (OCR) hat unter Digitalisierungsmanagern und Bibliothekaren einen schlechten Ruf: Die Erkennungsgenauigkeit insbesondere bei historischen Schriften wird vielfach als zu gering, die Kosten hingegen werden als zu hoch empfunden.[1] Auch 2010 gehört somit die systematische OCR Erkennung historischer Drucke noch keineswegs zum Standardrepertoire von Digitalisierungsprojekten in Deutschland und Europa. Lässt man Benutzer zwischen unterschiedlichen Lieferungsformaten und Zugriffsmöglichkeiten wählen, dann zeigen Untersuchungen, dass das durchsuchbare PDF, das in einem ersten Layer das Bild einer Seite, im zweiten Layer den Volltext enthält, an allererster Stelle steht – weit vor allen anderen Formaten und sogar dem Besitz des Originals vorgezogen wird.[2] Auch in den Papieren zur Deutschen Digitalen Bibliothek wird die Texterkennung als erstrebenswert angesehen: »In Zukunft wird durch die zu erwartenden Fortschritte im OCR-Bereich die Anzahl der Volltexte sprunghaft ansteigen. Diese Volltexte sollten ebenfalls für die DDB nutzbar sein, mindestens zum Aufbau von Suchindizes, besser noch auch für die Anzeige und Nutzung.«[3]

Anzahl der Volltexte wird in Zukunft sprunghaft ansteigen

Der vorliegende Aufsatz geht von der Überzeugung aus, dass ein Vorhaben zur systematischen Digitalisierung historischer Zeitungen die Texterkennung in die Planungen mit einbeziehen muss. Die besondere Herausforderung liegt allerdings darin, dass die OCR Erkennung nicht einfach ein Modul ist, das ganz am Ende eines Digitalisierungsworkflows angehängt werden kann, sondern dass die Entscheidung für die OCR Erkennung auf allen Ebenen Konsequenzen nach sich zieht: Von der Auswahl des Materials, über die Art des Scannens, der Qualitätskontrolle bis zur Gestaltung der Such- und Navigationsmöglichkeiten in einer digitalen Bibliothek. Aufgrund dieser Überlegungen orientiert sich dieser Aufsatz am Workflow eines typischen Digitalisierungsprojekts und stellt anhand der einzelnen Schritte die wichtigsten Implikationen dar, die mit der Entscheidung für eine OCR Erkennung einher gehen. Einen besonderen Schwerpunkt bilden dabei auch die neuesten Entwicklungen aus dem EU Projekt IMPACT, das sich mit der Texterkennung historischer Schriften beschäftigt.[4]

Auswahl des Materials

Bei Zeitungen handelt es sich um besonders sensibles Material: Sie wurden oftmals auf minderwertigem Papier gedruckt und die maschinell erstellten Bleilettern »bis zum Anschlag« ausgereizt. Die ursprünglich gefalteten Ausgaben wurden dann innerhalb der Bibliotheken in Jahrgängen gebunden und manchmal – sofern die alte Bindung gebrochen war – durch eine neuerliche Bindung »zu Tode restauriert«, sodass sie sich kaum mehr öffnen lassen. Hinzu kommt, dass viele Zeitungen in einem bemitleidenswerten Zustand sind, da sie zweifelsohne zu den am stärksten nachgefragten historischen Beständen einer Bibliothek gehören und sie somit besonderen Belastungen durch die Benutzer ausgesetzt sind.

Für die Digitalisierung von Zeitungen stehen – im Gegensatz etwa zu Büchern – oftmals drei Alternativen zur Verfügung, und die Entscheidung für eine der drei Optionen ist fundamental für die Qualität der Texterkennung:

1. Scannen vom gebundenen Original, so wie es im Magazin einer Bibliothek vorgefunden wird.
2. Scannen von einem Mikrofilm, wie er vielfach in den

letzten Jahrzehnten auf der Basis von Sicherungsmaßnahmen erstellt wurde.

3. Scannen vom Original, dessen Rücken aber entfernt wurde, sodass es sich um einzelne, lose Blätter handelt.

Das Scannen vom gebundenen Original aus den Regalen einer Bibliothek ist auf den ersten Blick die einfachste und bequemste Methode. Die Bibliothek besitzt die volle Kontrolle und kann sofort mit der Digitalisierung beginnen. Die Nachteile können jedoch gravierend sein: In vielen Fällen ist die Bindung so eng, dass der Band nur unter erschwerten Bedingungen zu 180 Grad geöffnet werden kann. Auch wenn die Hersteller von Scannern mit einigem Aufwand an Entzerrungsmethoden arbeiten und Google etwa genau aus diesem Grund ein eigenes Patent entwickelt hat, so handelt es sich doch immer um einen Kompromiss, der für einen Teil des Materials ohne Zweifel eine Verbesserung schafft, der aber für die Massenproduktion nur bedingt tauglich ist.

Aus den genannten Gründen werden bei der Zeitungsdigitalisierung deshalb oftmals Mikrofilme verwendet, die in einer nicht unbedeutenden Zahl ohnehin in Bibliotheken vorhanden sind. Gegenüber dem Scannen des gebundenen Bandes ist vor allem ein klarer Preisvorteil gegeben, der mit einem Faktor 5–10 bezeichnet werden kann. Auch die Digitalisierung selbst erzielt bei Mikrofilmen gute Ergebnisse. Bei einer bisher nicht veröffentlichten Studie im Rahmen des IMPACT Projekts wurde die Erkennungsgenauigkeit verglichen, die sich beim Scannen vom Original im Vergleich zum Scannen vom Mikrofilm ergab. Der Mikrofilm stammte aus einer Kampagne der British Library und wurde vor wenigen Jahren angefertigt, die Scans vom Original wurden extra für den Vergleich mit einem gängigen Buchscanner erstellt. Zugrunde gelegt wurden mehrere zufällig ausgewählte Seiten aus dem *Examiner* (1829) und *Reynold's Newspaper* (1881). Ein Vergleich der Ergebnisse zeigt keine Unterschiede in der Erkennungsgenauigkeit, eine Erfahrung, die auch viele Scandienstleister bestätigen können, die bis vor wenigen Jahren bei Digitalisierungsaufträgen vorher einen Mikrofilm anfertigten und diesen dann erst mittels Mikrofilmscanner in ein Digitalisat verwandelten.

Doch leider lässt sich dieses Einzelbeispiel nicht so einfach generalisieren: Beim oben gewählten Beispiel handelt es sich um einen Zeitungsband, der sich weitgehend problemlos öffnen ließ und der mit einer modernen Mikrofilmkamera von einem sorgfältig arbeitenden Dienstleister angefertigt wurde. Diese drei Faktoren treffen jedoch für viele ältere Mikrofilme, die in Archiven oder Bibliotheken lagern, leider nicht zu. Man muss sich vielmehr vor Augen halten, dass die technischen Probleme, wie sie für das Scannen gebundener Ausgaben gelten, auch für die Mikroverfilmung gegolten haben: Vielfach treten bei der Mikroverfilmung Wölbungen und Verzerrungen im Falz durchgängig auf. Hinzu kommt, dass viele Sicherungskopien unter großem Zeit- und Kostendruck entstanden sind und die Qualitätskontrolle durch den Auftraggeber zu wünschen übrig ließ. Erst in den letzten 20–30 Jahren kann davon ausgegangen werden, dass die Filme über eine gleichbleibend gute Qualität verfügen.[5] Somit gilt: Alle Probleme, die beim Scannen von gebundenen Vorlagen zu beobachten sind, treten auch bei der Mikroverfilmung auf, es kommen aber noch zusätzliche Probleme mit älteren Mikrofilmen hinzu. Handelt es sich jedoch um eine »saubere Verfilmung«, die mit einem »problemlosen« Band durchgeführt wurde, dann ist die Digitalisierung des Mikrofilms ein echter und kostengünstiger Ersatz für das Scannen vom Original und mit keinen Einbußen bei der OCR Erkennung im Vergleich zur OCR Erkennung direkt vom Scan des Originals verbunden.

Scannen von gebundenen Vorlagen oder Mikrofilmen

Die dritte Alternative, nämlich das Zerlegen bzw. Aufschneiden des Bandes in Einzelblätter, mag für viele Bibliothekare auf den ersten Blick keine ernstzunehmende Option darstellen, doch lohnt sich ein näherer Blick. Erstens ist anzumerken, dass Zeitungen niemals für das Binden vorgesehen waren, d.h. ihre Bindung ist das Resultat rein praktischer Gründe, gehört aber nicht zum ursprünglichen Erscheinungsbild einer Zeitung. Zweitens liegen die Vorteile eines zerlegten Bandes auf der Hand: Die Probleme mit dem Falz und der Wölbung fallen gänzlich weg. Es lassen sich Dokumentenscanner, Overheadscanner mit Glasplatte oder sogar großflächige Flachbettscanner verwenden. Ein Blick auf den Workflow der Medienbeobachter, bei denen mehrere tausend Seiten pro Tag gescannt und recherchiert werden müssen, zeigt, dass auch hier selbstverständlich die aktuellen Ausgaben der Zeitungen zerlegt und mittels spezieller Dokumenten- oder Flachbettscanner verarbeitet werden.[6]

Studie im Rahmen des IMPACT Projekts

Image Capturing

Während beim herkömmlichen Scannen vor allem visuelle Kriterien zählen und der Text für den Menschen lesbar sein soll, müssen für eine optimale OCR Erkennung eine Reihe weiterer Kriterien beachtet werden. Die meisten dieser Kriterien klingen selbstverständlich, doch sind sie in der Praxis oftmals kaum zu erreichen.

Kriterien für die OCR Erkennung

- Die Vorlage sollte sich vollständig öffnen lassen bzw. sollte es sich um Einzelseiten handeln, die plan aufgelegt werden können. Dieses Kriterium ist bei Zeitungen am schwierigsten zu erreichen. Kleinere

Wölbungen können durch verschiedene Gegenmaßnahmen ausgeglichen werden (Glasplatte, Entzerrung durch Scanner bzw. Software); bei eng gebundenen Bänden lässt sich hingegen oftmals echter Informationsverlust nicht vermeiden.

Qualität der Scans ist entscheidend

- Die Vorlage sollte nicht verschmutzt sein oder Unterstreichungen und Anmerkungen von Benutzern enthalten.
- Unterschiedliche Lichtverhältnisse innerhalb der Seite, etwa durch die bei älteren Vorlagen typische Welligkeit, treten praktisch bei allen älteren Exemplaren auf und können je nach dem Grad der Feuchtigkeit und der Lagerung im Magazin erheblich sein.
- Die Textzeilen sollen im digitalen Bild möglichst gerade verlaufen, d. h. keinerlei Wölbungen aufweisen und parallel zum oberen und unteren Rand sein. Dieses Kriterium ist wohl am schwierigsten zu erreichen, jedoch ebenfalls entscheidend für die Güte der OCR Erkennung.
- Der Unterschied zwischen Vorder- und Hintergrund sollte möglichst deutlich sein und ein Durchscheinen der Buchstaben von der Rückseite möglichst vermieden werden. Auch dieses Kriterium ist oftmals nur durch Mühe zu erreichen und widerspricht den Anforderungen für Bilder (Fotos), die bei geringerem Kontrast eine bessere optische Wirkung erzielen.

Man muss sich vor Augen halten, dass die Qualität des Scans für die weitere OCR Verarbeitung von entscheidender Bedeutung ist. Andere Faktoren, über die oftmals wesentlich mehr diskutiert wird, sind demgegenüber als weniger wichtig einzustufen.

Auflösung und Erkennungsgenauigkeit

Welchen Einfluss besitzt nun die Auflösung für die Erkennungsgenauigkeit? Im Gegensatz zur industriellen Dokumentendigitalisierung, bei der üblicherweise mit 300 ppi gearbeitet wird, wird in Bibliotheken unter dem Einfluss der richtungsweisenden Veröffentlichungen von Anne R. Kenney[7] oftmals mit 600 ppi gescannt. Macht man einige Tests mit der ABBYY FineReader Software wird man erkennen, dass eine höhere Auflösung jedoch keineswegs automatisch zu besseren Ergebnissen führt. Vielmehr treten z. T. andere Fehler auf als etwa bei 300 ppi, und das Gesamtresultat bleibt oftmals fast gleich.

Projektergebnisse zur Binarisierung

Anders hingegen bei kleineren Schriften, die schon bei 400 ppi deutlich besser erkannt werden. Ein einfacher Test kann diese Aussage illustrieren: Gescannt wurde eine Vorlage, bei der 107 der häufigsten Wörter der deutschen Sprache mit 4 Punkt Größe auf einem gängigen Laserdrucker ausgedruckt und mit 300 bzw. 400 ppi mittels BookEye 3 Scanner digitalisiert wurden. Das Ergebnis ist eindeutig: Bei 300 ppi waren 31 Wörter falsch, bei 400 PPI nur 13 Wörter. Allerdings waren schon bei 5 Punkt Größe keine Unterschiede mehr feststellbar. Somit ergibt sich als Grundregel: Weist eine Vorlage eine große Menge relativ kleiner Schrift auf, dann sind 400 oder gar 600 ppi tatsächlich empfehlenswert und werden zu einer verbesserten Erkennung der kleineren Schriften führen, ansonsten sind jedoch auch 300 ppi ausreichend. Bei Zeitungen, die oftmals über relativ kleine Schriften verfügen, wird man daher im Zweifelsfall eher mit 400 ppi arbeiten.

Eine andere oftmals gestellte Frage ist jene nach der Farbtiefe. Hier gibt es eine Untersuchung von Tracy Powell und Gordon Paynter, die zu dem Schluss kommen: »There was no evidence of any improvement in OCR accuracy from greyscale digitisation«.[8] Das Ergebnis kam zustande, indem mehrere Mikrofilme sowohl bitonal als auch in 8 Bit Graustufen gescannt und das Ergebnis der OCR Erkennung verglichen wurde.

Um diese Ergebnisse jedoch angemessen interpretieren zu können, muss man sich zunächst klar machen, dass die am Markt befindlichen OCR Programme ausschließlich auf der Basis von bitonalen Bildern arbeiten. Mit anderen Worten: Auch Farb- oder Graustufenbilder werden zuerst zu einem bitonalen Bild konvertiert (= Binarisierung). Die entscheidende Frage ist daher, wie diese Konvertierung erfolgt. Bei der im obigen Beispiel angeführten Versuchsanordnung steht die interne Binarisierung des Mikrofilmscanners gegen die interne Binarisierung des OCR Programms. D. h. auch der Mikrofilmscanner tastet zuallererst die Vorlage mit einer höheren Informationstiefe ab und errechnet daraus dann ein bitonales Bild. Da der Mikrofilmscanner den ursprünglichen Datenstrom zur Verfügung hat und diesen sehr spezifisch interpretieren kann, ist davon auszugehen, dass die Ergebnisse dieses Binarisierungsprozesses sehr gut sind. Umgekehrt verfügt die OCR Software über eine wesentlich eingeschränktere Informationsgrundlage, kann aber andererseits wieder die Binarisierung in Relation zur Texterkennung durchführen. Welcher Faktor nun stärker zum Tragen kommt, ob die interne Binarisierung oder die Binarisierung der OCR, kann im Vorhinein nicht entschieden werden. Es ist aber davon auszugehen, dass wenn es sich um homogenes Material handelt – wie eben einem Mikrofilm – keine Unterschiede zur OCR internen Binarisierung zu erwarten sind.

Anders hingegen sieht es bei »schwierigem« Material aus: dort kann eine adaptive Binarisierung durch die OCR Software große Vorteile erzielen, wie das folgende Beispiel zeigt: In einem Digitalisierungsprojekt an der Abteilung für Digitalisierung und elektronischen Archivierung der Universitätsbibliothek Innsbruck wurden 800.000 Zeitungsausschnitte der Jahre 1960 bis 2000 digitalisiert. Diese Zeitungsausschnitte

waren sehr heterogen: Es handelte sich um Artikel aus mehr als 30 verschiedenen Zeitungen, die teilweise stark vergilbt oder ausgeblichen waren und zudem noch schief auf einem A4 Blatt aufgeklebt wurden. Bei der Digitalisierung mit einem Dokumentenscanner der Marke Kodak i260 wurden aus praktischen Gründen zwei Ausgabeformate gewählt: Einmal die vom Scanner binarisierten Bilder, einmal relativ schwach komprimierte 24 Bit JPEG Farbbilder, die auf 90 % ihrer ursprünglichen Information komprimiert wurden. Vergleicht man die OCR Erkennung dieser Zeitungsausschnitte, dann zeigt sich ein gravierender Unterschied zwischen den beiden Qualitäten: Das bitonale Bild ergibt eine Wortgenauigkeit von ca. 96 %, die Erkennung auf Grundlage der 24 Bit Datei hingegen eine Wortgenauigkeit von ca. 99 %. Betrachtet man die Unterschiede genauer, dann zeigt sich, dass bei den unproblematischen Textabschnitten keinerlei Unterschiede vorhanden sind; dort hingegen, wo Benutzer Unterstreichungen einfügten, oder ein Durchscheinen von der Rückseite gegeben ist, oder das Papier durch eine Faltung beschädigt war, die Erkennungsrate von der Binarisierung durch die OCR profitierte. Auch im IMPACT Projekt werden derartige Überlegungen verfolgt und ein adaptiver Binarisierungsalgorithmus entwickelt, der direkt mit der Zeichenerkennung verknüpft ist. Erste Ergebnisse dieser Forschungsarbeit sollen bereits in die nächsten Produktversionen des Abbyy FineReader einfließen.

Damit stellt sich eine letzte Frage, nämlich, ob denn der Komprimierungsfaktor einen Einfluss auf die OCR Erkennung hat. Mit 400 ppi und 24 Bit gescannte Zeitungsseiten in TIFF RGB unkomprimiert abzulegen führt auch heute noch zu erheblichen Speicherproblemen. Viele Dokumentenscanner können zudem dieses Format gar nicht bedienen. In der Praxis wird man also auf JPEG oder JPEG2000 Dateien ausweichen, die nur einen geringen Informationsverlust aufweisen, jedoch meist nur ein Zehntel des ursprünglichen Speicherbedarfs benötigen. Auch hierzu wurde an der Abteilung für Digitalisierung und elektronische Archivierung ein kleiner Test mit dem ScanRobot von Treventus durchgeführt und mehrere Seiten mit 300 ppi gescannt und als TIFF unkomprimiert, sowie als JPEG mit 90, 75, 50 und 25 % Komprimierungsrate gespeichert. Die Ergebnisse zeigen das erwartete Resultat: Zwischen TIFF unkomprimiert und JPEG mit 90 % Komprimierungsfaktor gibt es keine signifikanten Unterschiede. Sogar relativ starker Informationsverslust, wie er bei der Einstellung von 25 % gegeben ist, verändert die OCR Qualität nicht. Die unkomprimierte Speicherung von 24 Bit Farbbildern ist daher für Zeitungsseiten aus Gründen der OCR Erkennung nicht notwendig.

Ergebnisse aus der Praxis der Zeitungsdigitalisierung

Zusammenfassend lässt sich somit sagen, dass mit der Digitalisierung von losen Zeitungsseiten mit 400 ppi und 24 Bit Farbtiefe die optimalen Voraussetzungen für eine OCR Erkennung geschaffen werden. Allerdings sollte man unbedingt am konkreten Material einen repräsentativen Test durchführen, denn auch mit gebundenen Zeitungsbänden, die sich gut öffnen lassen und über eine ausreichend große und saubere Schrift sowie gutes Papier verfügen, oder mit sorgfältig erstellten Mikrofilmen, die ebenfalls auf gutem Material basieren, lassen sich auch mit 300 ppi und bitonalen Bildern nahezu gleichwertige Ergebnisse erzielen.

Texterkennung

Im Gegensatz zum Markt für Scanner, auf dem sich eine große Anzahl von Firmen tummelt, ist bei der OCR Software in den letzten 10–15 Jahren ein starker Konzentrationsprozess zu beobachten. Im Wesentlichen dominieren nur noch wenige Firmen die Szene: Nuance mit OmniPage, ABBYY mit FineReader und IRIS mit Readiris. ABBYY wiederum ist die einzige Firma, die sich seit mittlerweile zehn Jahren auch um historische Schriften und Dokumente bemüht[9] und mit der Frakturerkennung ein Alleinstellungsmerkmal besitzt. Einigen Wind aufgewirbelt hat auch die Ankündigung von Google im Jahr 2006, dass man eine freie OCR Engine (Tesseract) zur Verfügung stellen werde.[10] Unter Leitung von Thomas Breuel vom Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Kaiserslautern wird nunmehr seit 2007 das Programm OCRopus entwickelt, das diesen Ansatz aufgreift und ebenfalls unter einer Open Source Lizenz zugänglich gemacht wird.[11] Allerdings befindet sich OCRopus noch in einem sehr frühen Stadium, sodass eine seriöse Abschätzung noch nicht getroffen werden kann, ob damit eine ernsthafte Alternative zu den kommerziellen Softwarepaketen gegeben sein wird. Immerhin handelt es sich sowohl bei Nuance als auch bei ABBYY um weltweit agierende Unternehmen, die auch Kyrillisch, Chinesisch, Hebräisch, Arabisch und knapp 200 weitere Schriften unterstützen.

Einfluss des Komprimierungsfaktors

Für ein konkretes Zeitungsdigitalisierungsprojekt im mitteleuropäischen Raum bedeutet dies, dass man kaum eine ernsthafte Alternative zur Frakturerkennung der ABBYY Software finden wird und sich daher mit den Möglichkeiten und Tücken dieses Programms etwas intensiver auseinander setzen sollte – auch um die Angaben von Dienstleistern besser nachprüfen zu können.

Frakturerkennung

Die eigentliche OCR Erkennung kann grob vereinfacht in drei Schritte aufgeteilt werden, wobei alle drei Schritte innerhalb der OCR Engine mehrfach wiederholt werden und aufeinander abgestimmt sind:

- Layouterkennung und Segmentierung
- Anwendung von Klassifikatoren auf Worte und Buchstaben
- Abgleich der erkannten Worte mit einem Wörterbuch

Layouterkennung

Die Layouterkennung ist der erste und grundlegendste Schritt: Hier entscheidet das Programm, was überhaupt als »Inhalt« einer Seite gelesen werden soll. Schwarze Ränder, Schmutz, Anmerkungen von Benutzern sollen gar nicht erst der Schrifterkennung vorgelegt werden. Sind die mit »Inhalt« gefüllten Regionen erkannt, so muss noch entschieden werden, ob es sich dabei um Text, Tabellen, Bilder, Barcode oder vielleicht graphische Elemente wie Schmuckzeichen oder Linien handelt. Innerhalb eines Textblocks müssen dann noch Linien und Worte erkannt werden.

Anwendung von Klassifikatoren

Der zweite Schritt ist die Anwendung von Klassifikatoren auf die ermittelten Wörter und Buchstaben. Diese Klassifikatoren sind vorher erstellte Muster, die einen bestimmten Buchstaben repräsentieren. Das bedeutet allerdings, dass das System auf alle vorkommenden Buchstaben trainiert sein muss. Bei seltenen Schriften, wie sie in Zeitungen z. B. als Auszeichnungsschriften in Titeln oder in Annoncen vorkommen, führt dies zu Problemen bei der Erkennung. Der Erkennungsvorgang selbst lässt sich durch den Benutzer nur insofern beeinflussen, als die Möglichkeit besteht, einzelne Buchstaben mittels eines speziellen Editors besonders zu trainieren. Dies sollte allerdings nur sehr vorsichtig durchgeführt werden, da es zu unerwünschten Nebeneffekten kommen kann.

Problem der Segmentierung

Empfehlenswerter scheint es zu sein, direkt mit ABBYY in Kontakt zu treten und die Optimierung der Software für diese Buchstaben in Auftrag zu geben. Auf der Grundlage von mindestens 100 Beispielen pro Buchstabe kann eine derartige Optimierung durchgeführt werden. Dies hat unter anderem die Estnische Nationalbibliothek für einige seltene Buchstaben der estnischen Fraktur getan und erzielt damit deutlich bessere Ergebnisse als vorher.[12]

Anwendung eines Wörterbuchs

Der letzte Schritt ist schließlich die Anwendung eines Wörterbuchs auf die von der reinen Schrifterkennung vorgegebenen Buchstabenkombinationen. Auch dabei handelt es sich um einen komplexen Rückkoppelungs- und Optimierungsprozess. Wie groß der Einfluss des Wörterbuchs auf die Erkennungsgenauigkeit ist, zeigt ein Test, den die CIS Gruppe an der Ludwig Maximilian Universität München durchgeführt hat: Ein Text wurde nur auf Basis der Zeichenerkennung, also ohne jede Unterstützung durch ein Wörterbuch erkannt. In einem zweiten Schritt wurden verschiedene Wörterbücher angewendet, und es zeigte sich eine Fehlerreduktion von 30 % bis 60 % gegenüber den ursprünglichen Ergebnissen.[13] Dieses Wissen um die Bedeutung von Wörterbüchern war auch der Ausgangspunkt für die außerordentlich starke Beteiligung von Linguisten am IMPACT Projekt. Sowohl historische Wörterbücher als auch die linguistischen Tools zur Erstellung dieser Wörterbücher sollen interessierten Forschern und Bibliotheken frei zur Verfügung gestellt werden.

Artikelseparierung

Eine ganz besondere Herausforderung bei der Digitalisierung historischer Zeitungen stellt die Artikelseparierung dar: Einzelne Artikel und ihre Elemente, wie Titel, Untertitel, Fließtext und die Abfolge des Textes sollen korrekt erkannt werden. Die einfache Anwendung der OCR kann allerdings zu eklatanten Problemen bei der Segmentierung führen. Entweder werden einzelne Spalten nicht ordentlich getrennt, oder aber die Leserichtung der Blöcke wird nicht korrekt wiedergegeben. Der Text ist dann zwar immer noch für eine Suchmaschine brauchbar, für viele andere Anwendungen, wie etwa die Anzeige des Artikels, ist das Resultat weitgehend wertlos.

Daraus ergibt sich bei der Digitalisierung von historischen Zeitungen eine schwierige Situation: Will man einen Workflow, bei dem die OCR Erkennung völlig automatisiert abläuft, dann wird man eine Reihe von Fehlern bei der Layoutanalyse in Kauf nehmen müssen. Das hat zwar nur einen geringen Einfluss auf die Güte der Texterkennung, aber bei einer allfälligen Nachkorrektur durch den Benutzer sind diese Fehler äußerst lästig. Das Korrekturinterface müsste entsprechend komplex aufgebaut sein, um auch eine nachträgliche Änderung von Segmenten, etwa das Teilen, Zusammenfügen oder die Änderung der Reihenfolge von Blöcken gewährleisten zu können. All diese Operationen sind kompliziert und bedeuten einen hohen Aufwand.

Aus diesem Grund wird die Artikelseparierung oftmals an externe Dienstleister vergeben. In Deutschland sind hier etwa CCS GmbH, das Frauenhofer Institut für Intelligente Analyse- und Informationssysteme oder PrePress Systeme GmbH zu nennen.[14] Man muss sich aber trotzdem vor Augen halten, dass trotz aller Automatisierung eine manuelle Inspektion jeder einzelnen Seite und der allfälligen Korrektur der Segmentierung trotzdem notwendig ist und zu einer wesentlichen Verteuerung des Projekts führt. Die Verbesserung der Segmentierung sowie die Entwicklung einer generischen Plattform zur automatisierten Strukturerkennung sind auch im IMPACT Projekt wesentliche Arbeitspakete und sollen ganz konkret zu einer verbesserten Erkennung von Zeitungsseiten führen.

Erkennungsgenauigkeit

Soll im Rahmen eines Digitalisierungsprojekts eine Abschätzung gemacht werden, welche Erkennungsraten konkret zu erwarten sind, so lässt sich dies nur auf der Basis einer Stichprobe ermitteln. Hat ein durchschnittlicher Zeitungstitel z. B. 250.000 Seiten, so erlauben 500 mittels Zufallsgenerator ausgewählte Testseiten einen guten Rückschluss auf den gesamten Bestand. In diesem Fall würde man die 500 Seiten digitalisieren und mittels OCR erkennen. Die Auszählung selbst erfolgt üblicherweise manuell, wobei man wiederum nicht die vollständige Seite, sondern innerhalb der Seite einzelne Ausschnitte wählen wird. Im Gegensatz zur Auszählung von Buchstaben, die mühsam, zeitintensiv und auch fehleranfällig ist, kann die Auszählung auf Wortebene wesentlich einfacher und verlässlicher durchgeführt werden. Um den ganzen Vorgang zu vereinfachen und zu objektivieren, wird im IMPACT Projekt ein Evaluationsprogramm entwickelt, das den automatischen Abgleich zwischen einer exakten Vorlage und der tatsächlichen OCR Erkennung ermöglicht. Das Kompetenzzentrum zur Textdigitalisierung wird derartige Testprojekte ab 2012 auch als Dienstleistung für Bibliotheken anbieten.

Die größte inhaltliche Untersuchung zum Thema der OCR Genauigkeit bei historischen Zeitungen wurde 2009 von Simon Tanner im Auftrag der British Library durchgeführt. Als Grundlage dienten ca. 20.000 willkürlich ausgewählte Zeitungsseiten aus der Newspaper Online Collection der British Library, die insgesamt rund zwei Millionen Seiten enthält und Zeitungen des 17., 18. und 19. Jahrhunderts umfasst. Von diesen 20.000 Seiten wurden jeweils einige Ausschnitte willkürlich ausgewählt und der korrekte Text mit dem OCR Text verglichen. Ausgewertet wurde nicht nur die Wortgenauigkeit, sondern auch die Buchstabengenauigkeit sowie die Genauigkeit bezogen auf bedeutungstragende Wörter. Die Ergebnisse, die sich hier allerdings nur auf die Zeitungen des 19. Jahrhunderts beziehen, sind sehr aufschlussreich und können wie folgt zusammengefasst werden: Erstens besteht zwischen Wort- und Buchstabengenauigkeit ein klarer Zusammenhang, d. h. es genügt eine der beiden Maßzahlen, um die Genauigkeit der OCR zu bestimmen. Zweitens sind die Unterschiede zwischen den einzelnen Zeitungstiteln signifikant und reichen von 60 % bis zu 95 % Wortgenauigkeit. Drittens ist das Erscheinungsjahr der Zeitung keineswegs der alles entscheidende Faktor, vielmehr zeigt sich ein komplexer Verlauf, der stark von der Entwicklung der Drucktechnik, der Papierqualität, dem Layout und den jeweils aktuellen Schriften beeinflusst ist.

Für das gesamte Sample englischer Zeitungen des 19. Jahrhunderts ermittelte Tanner eine Wortgenauigkeit von 78 %, die Verteilung bezogen auf Zeitungstitel stellt sich wie folgt dar:[15]

Wortgenauigkeit	60–69 %	70–79 %	80–89 %	>90 %
Zeitungstitel	20 %	27 %	51 %	2 %

Tabelle: Wortgenauigkeiten pro Zeitungstitel – British Library 19th Century Newspaper Project

Vergleichbare Resultate hat auch ein Test erbracht, der an der Abteilung für Digitalisierung und elektronische Archivierung durchgeführt wurde. Zugrunde gelegt wurden 30 zufällig ausgewählte Zeitungsseiten, die im Rahmen von Digitalisierungsaufträgen in den letzten Jahren erstellt wurden. Insgesamt wurden fünf verschiedene deutschsprachige Zeitungstitel in den Test mit einbezogen; die Jahrgänge reichen von 1813 bis 1924. Für die Texterkennung wurden jeweils kurze Textabschnitte mit ca. 100 Wörtern ausgewählt, insgesamt über 3.300 Wörter wurden ausgezählt. Die Texte stammen aus dem redaktionellen Teil, Annoncen und Werbungsseiten wurden nicht berücksichtigt. Die Texte wurden mittels ABBYY FineReader Recognition Server 2.0 erkannt. Die Ergebnisse sind überraschend gut: Auch wenn es sich durchgehend um Frakturschrift handelt, sind nur 16 % der Wörter falsch. Das gute Ergebnis kommt vor allem durch die Zeitungen des frühen 20. Jahrhunderts zustande, bei denen teilweise Erkennungsraten von 98 % und mehr erreicht werden.

Unterschiede in der Erkennungsgenauigkeit signifikant

Archivierung

Das Resultat der OCR Erkennung ist ein mehr oder weniger korrekter Text; doch dieser kann in vielen unterschiedlichen Formaten und mit den unterschiedlichsten Metadaten vorliegen. Das einfachste Ausgabeformat ist eine TXT Datei, wie sie etwa für die Volltextsuche benötigt wird. Ein bei Benutzern wesentlich beliebteres Format ist ein PDF Dokument, das sowohl das Bild der gescannten Seite als auch den Volltext enthält. Der Volltext kann durchsucht oder auch in andere Anwendungen kopiert werden, das Bild der Seite gewährleistet hingegen den Blick auf das Original.

unterschiedliche Formate und Metadaten

Darüber hinaus gibt es aber noch wesentlich mehr Informationen, die bei der OCR Erkennung anfallen: Handelt es sich um einen Text-, Bild- oder Tabellenblock? Wurde der Buchstabe sicher erkannt, oder wird er als möglicherweise falsch eingestuft? Wurde das Wort in einem Wörterbuch gefunden? Diese Informationen stehen typischerweise als XML Datei zur Verfügung und können für weitere Anwendungen wertvolle Dienste leisten. Darüber hinaus gibt es aber auch

noch Daten, die typischerweise mit spezieller Software bzw. von Dienstleistern erstellt werden, also z. B. die Artikelstruktur, Überschriften, Bildunterschriften und Ähnliches. Eine vollständig gescannte Zeitung mit z. B. 250.000 Seiten ergibt somit ein äußerst komplexes digitales Objekt, das mehr als eine Million Einzeldateien und tausende von Ordnern umfassen wird.

Frage der Langzeitarchivierung

Dementsprechend stellt sich auch die Frage der Langzeitarchivierung. Für die OCR Daten selbst wurde im Rahmen des METADATA ENGINE Projekts das Format ALTO (Analysed Layout and Text Object) entwickelt, das mittlerweile von einigen größeren Bibliotheken benutzt und seit August 2009 offiziell von der Library of Congress betreut wird.[16] ALTO erlaubt die Speicherung von Blöcken, Wortkoordinaten, aber auch die Versionierung, sodass die verschiedenen »Spuren«, die z. B. durch eine teilweise Korrektur des Textes entstanden sind, in der XML Struktur abgebildet werden können.

ALTO wurde in engem Zusammenhang mit dem ebenfalls von der Library of Congress gehosteten METS (Metadata Encoding and Transmission Standard) Format entwickelt. Ein digitales Zeitungsobjekt könnte demnach wie folgt gegliedert sein:

Korrektur im Social Web

Eine übergeordnete METS Datei enthält die allgemeinen deskriptiven, technischen und rechtlichen Informationen zur Zeitung und zum Scanprojekt sowie die Links auf die untergeordneten METS Dateien, die pro Jahrgang vorhanden sind. Diese METS Jahrgangsdateien enthalten detaillierte deskriptive, technische und administrative Daten zur Zeitung, sowie eine »Structural Map« mit einem Verzeichnis aller Nummern, dem Datum und eventueller Parallelausgaben, sowie die Links auf alle Dateien, die in verschiedenen Gruppen abgelegt sind. Diese bestehen einerseits aus den Masterimages, eventuellen Derivaten wie Thumbnails und einer auf die Größe der Zeitung optimierten Leseausgabe sowie den ALTO Dateien (oder dem nativen ABBYY XML File) und den PDF Dateien mit hinterlegtem Volltext. Liegen noch zusätzlich Daten aus der Strukturerkennung vor, die sich auf Artikel und Überschriften etc. beziehen, dann können diese entweder zusätzlich im METS File kodiert werden, oder aber in einer zusätzlichen TEI Light Datei, wie sie etwa von der Bayerischen Staatsbibliothek oder der SUB Göttingen verwendet wird.[17]

OCR Korrektur

Die Vorstellung, dass man ernsthaft darüber diskutiert, wie zehntausende oder gar Millionen von automatisch erkannten Textseiten manuell korrigiert werden sollten, wäre noch vor wenigen Jahren als völlig absurd eingestuft worden. Doch die Entwicklung des Internet zu einem alltäglichen Gebrauchsmittel – Stichwort »soziales Web« – ermöglicht völlig neue Anwendungen, darunter auch die massenhafte Korrektur von OCR erkannten Texten.

So hat die Australische Nationalbibliothek ein sehr einfaches Korrekturmodul eingerichtet, bei dem Benutzer neben dem Bild eines Artikels auch den (fehlerhaften) OCR Text sehen. Dieser Text kann dann direkt Zeile für Zeile verbessert werden. Obwohl dieser Vorgang recht mühsam ist, haben doch seit August 2008 mehr als 6.000 Benutzer davon Gebrauch gemacht und bisher ca. sieben Millionen Zeilen in 318.000 Artikeln korrigiert. Immerhin sind dies 15 % aller Artikel, die bisher online angeboten werden. Rechnet man diese Leistung auf einer Basis von 50 Zeichen pro Zeile hoch und vergleicht dies mit den derzeit üblichen Preisen fürs Korrekturtippen der Dienstleister (ca. 0,6 EUR pro 1.000 Zeichen), so kommt man immerhin auf eine Summe von rund 200.000 EUR, die durch die Benutzer in weniger als zwei Jahren freiwillig geleistet wurden. Besonders positiv zu bewerten ist allerdings auch die Tatsache, dass diese Benutzer sehr bewusst ein Angebot ihrer Nationalbibliothek unterstützen und so einen Beitrag zur Verbesserung des Kulturangebots ihres Landes leisten.

Es geht also nicht so sehr darum, von Anfang an die perfekte Transkription des Textes zur Verfügung zu haben, sondern dem Benutzer die Chance zu geben, jene Artikel, die ihm aus welchen Gründen auch immer besonders am Herzen liegen, zu korrigieren und damit für ihn und andere besser (be)nutzbar zu machen. Korrigierte Artikel werden bei der Volltextsuche besser gefunden, man kann ihren Text kopieren und weiterverarbeiten – der Nutzen ist somit für den Benutzer

Richtigstellung

Im ZfBB-Themenheft Zeitungen (2010) wird in meinem Beitrag »Zeitungen in deutschen Bibliotheken, Archiven und Museen: ein Überblick zur aktuellen Situation« auf S. 148 im Zusammenhang mit der Sammlung Deutscher Drucke (SDD) und dem Aufbau einer nationalen Zeitungssammlung irrtümlich der Eindruck erweckt, als läge die Hauptverantwortung für den beschriebenen Mangel beim Institut für Zeitungsforschung in Dortmund sowie bei der Staats- und Universitätsbibliothek Bremen als ehemalige DFG-Sondersammelgebietsbibliothek für das Fach Publizistik. Dies entspricht nicht den Tatsachen, stattdessen hat die AG Sammlung Deutscher Drucke das Defizit erkannt und bemüht sich im Rahmen ihrer Aufgabenstellung um eine baldige Lösung.

Christoph Albers,
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

direkt gegeben. Der Bibliothek kommt hier eine neue Rolle zu, nämlich ganz bewusst das freiwillige Engagement ihrer Nutzer zu unterstützen und technische Möglichkeiten bereit zu stellen, die dann zu jedem beliebigen Zeitpunkt und von jeder beliebigen Person für ihre eigenen Zwecke genutzt werden können, aber auch langfristig anderen Lesern dienen.

Das IMPACT Projekt wird im Bereich dieser kollaborativen Textkorrektur ganz wesentliche Neuerungen bringen. Gleich zwei Projektpartner beschäftigen sich mit der systematischen Verbesserung von OCR Texten. Die bereits genannte Forschungsgruppe um Klaus Schulz an der LMU München entwickelt ein Postkorrektur-Modul, das auf Basis einer linguistischen Analyse eine besonders effektive Korrektur erlauben soll. Einen ganz anderen Ansatz hat die Forschungsgruppe am IBM Research Lab Haifa gewählt: Hier werden alle Buchstaben und Worte innerhalb eines Dokuments miteinander verglichen und ein Ähnlichkeitsmaß erstellt. Durch den Vergleich mit einer OCR Engine bzw. durch die Eingabe eines Benutzers wird ein selbstlernendes System in Gang gesetzt, das nun die Buchstaben und Worte immer genauer klassifizieren kann und damit sowohl den durch den Benutzer benötigten Input minimiert als auch die Erkennung verbessert. Das Interface für dieses System benötigt nur ein geringes Maß an Training und sollte weitgehend selbsterklärend sein. Im Vergleich zur zeilenbasierten Korrektur sollte das Tool eine signifikante Steigerung der Produktivität um mindestens den Faktor Fünf erlauben.[18] Davon sollten klarerweise nicht nur Benutzer, die an einer systematischen Verbesserung eines Textes interessiert sind, profitieren. Vorstellbar sind auch völlig neue Dienste, die von Dienstleistern im Hintergrund umgesetzt werden, wie etwa ein »Korrektur-on-Demand« Service, bei dem auf Anfrage hin bestimmte Artikel kostengünstig korrigiert werden.

Suche

Die vorhergehenden Ausführungen haben deutlich gemacht, dass in vielen Fällen die Digitalisierung von historischen Zeitungen nicht über eine Wortgenauigkeit von 80–90 % hinauskommen wird. Wie ist dieses Ergebnis nun zu bewerten? Ist das Glas halb leer oder halb voll?

Die Antwort darauf hängt unmittelbar damit zusammen, wie der fehlerhafte Text im Rahmen einer digitalen Bibliothek präsentiert wird und welche Services daran geknüpft sind. Man sollte sich vor Augen halten, dass mit der Dominanz von Suchmaschinen das Kriterium der »Vollständigkeit« ohnehin einer Uminterpretation unterliegt: Niemand erwartet, dass er bei der Suche nach »Goethe Wilhelm Meister« in einer Internet Suchmaschine eine vollständige Liste jener Dokumente erhalten wird, die im Internet zu diesem Buch veröffentlicht wurden, sondern man gibt sich oftmals mit den relevantesten Dokumenten zufrieden – sofern sie jene Informationen enthalten, die man benötigt oder die man grundsätzlich erwartet hat. Ähnliches gilt nun auch für OCR erkannte Texte: Selbstverständlich wird eine Suche in einem Text mit 80 % Wortgenauigkeit längst nicht alle relevanten Treffer produzieren, aber der Benutzer wird trotzdem wesentlich »mehr« finden, als ohne diese Volltextsuche.

wesentliche Neuerungen im Bereich der Korrektur zu erwarten

Wesentlich ist allerdings, den »schlechten« Text dem Benutzer nicht zu verbergen, sondern ihn ausdrücklich sichtbar zu machen, indem er etwa direkt neben das gescannte Bild gestellt wird oder die Suche nicht mit Bildausschnitten sondern mit Textausschnitten arbeitet. Nur so kann der Benutzer lernen, dass die Volltextsuche auf einem fehlerhaften Text beruht und daher ganz gewiss nicht ein vollständiges Resultat liefern wird. Und gleichzeitig sollte genau dieser fehlerhafte Text auch »der Stachel im Fleisch« sein, der letztlich die Benutzer motiviert, an einer kollaborativen Verbesserung teilzunehmen.

Obwohl systematische OCR Erkennung zweifellos die größte Auswirkung auf die Gestaltung einer digitalen Zeitungsbibliothek besitzt, kann auf diesen Aspekt hier nur sehr kurz eingegangen werden, da er den Rahmen dieses Aufsatzes sprengen würde.

OCR Text und die Zeitungssuche im Internet

Von ganz entscheidender Bedeutung ist der (fehlerhafte) OCR Text auch für die Auffindbarkeit der Zeitung im Internet. Zum Einen kann der Text mittels einer SiteXML direkt zur Indexierung durch eine Suchmaschine vorgelegt werden. Benutzer werden so über Google oder andere Suchmaschinen auf einen relevanten Artikel verwiesen. Und es ist davon auszugehen, dass dies viele Benutzer auf die digitalisierten Zeitungsseiten leiten wird, die ansonsten keine Notiz von der Verfügbarkeit einer digitalisierten historischen Zeitung genommen hätten. Zum Zweiten kann aufgrund der Volltextsuche auch ein wesentlich reicheres Interface im Rahmen einer digitalen Bibliothek aufgebaut werden. Die Volltextsuche kann etwa mit den vorhandenen Metadaten facettiert werden, d. h. die Suche nach »Hofmannsthal Schnitzler« könnte einerseits die Treffer im Volltext zeigen, andererseits aber Facetten aufweisen, in welchen Jahren, in welchen Zeitungstiteln, aber auch – sofern vorhanden – in welchen Artikeltypen, Sprachen, Erscheinungsorten der Zeitungen, und ähnliches sich diese Kombination befindet.

Schließlich müssen in einer künftigen digitalen Zeitungsbibliothek auch die oben beschriebenen Korrekturmöglichkeiten integriert werden, wobei die von

den Benutzern korrigierten Texte möglichst in Echtzeit zu einem Update des Volltextindex führen sollten, sodass bereits die nächste Volltextsuche von dem verbesserten Text profitieren kann. Es ist klar, dass die bisher gängigen digitalen Bibliotheken mit diesen Herausforderungen noch nicht oder nur unzureichend umgehen können. Besonders im Bereich der Webpräsentation sind zudem Ideen und innovative Ansätze gefordert, um das eigentliche Ziel jedes Digitalisierungsprojekts erreichen zu können: Benutzern unterschiedlichster Herkunft eine attraktive und wertvolle Informationsquelle im Internet bieten zu können.

Ideen und innovative Ansätze gefordert

1 Vgl. etwa: Wegstein, Werner u. a.: TextGrid Report 4.1 : Digitalisierung von Primärquellen für die TextGrid-Umgebung: Modellfall Campe-Wörterbuch. Version: 12. Oktober 2009. www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf. – [Stand: 15.04.2010] Die DFG Praxisregeln erwähnen OCR Erkennung als eine Option, machen jedoch – im Gegensatz zur manuellen Erstelllung des Volltextes – keine Aussage darüber, welche Erkennungsgenauigkeit erwartet wird. Vgl.: DFG-Praxisregeln »Digitalisierung« zu den Förderprogrammen der Wissenschaftlichen Literaturversorgungs- und Informationssysteme. Stand April 2009. www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung.pdf

2 Mühlberger, Günter; Gstrein, Silvia: eBooks on Demand (EOD) : a European digitization service. In: IFLA Journal 2009, S. 39. http://ifl.sagepub.com/cgi/content/abstract/35/1/35. – [Stand: 15.04.2010]

3 Bund-Länder-Fachgruppe »Deutsche Digitale Bibliothek« (Hrsg.): Fachkonzept zum Aufbau und Betrieb einer »Deutschen Digitalen Bibliothek«. Entwurf: 16.02.2008. www.deutsche-digitale-bibliothek.de/doc/fachkonzept_160208_ohne_organisation.doc

4 EU Projekt IMPACT. www.impact-project.eu/ – [Stand: 15.04.2010]

5 Vgl. die Aussagen, die auf der Tagung »Verfilmung und Digitalisierung: Bestandserhaltung schriftlicher Dokumente für die Informationsgesellschaft. Abgehalten vom Forum Bestandserhaltung an der Bayerischen Staatsbibliothek München, 15.–16. November 2007« getroffen wurden. In: Zeitschrift für Bibliothekswesen und Bibliographie. 2008, Heft 3/4, S. 207–212.

6 Mündliche Auskunft von Claus Gravenhorst, Mitarbeiter der Firma CCS GmbH., die seit mehr als zehn Jahren Clipping Software für Medienbeobachter erstellt.

7 Chapman, Stephen; Kenney, Anne R.: Digital Conversion of Research Library Materials. A Case for Full Informational Capture. In: D-Lib Magazine, Oktober 1996 ISSN 1082–9873. www.dlib.org/dlib/october96/cornell/10chapman.html – [Stand: 15.04.2010]

8 Powell, Tracy; Paynter, Gordon: Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images. In: D-Lib Magazine. March/April 2009, Volume 15, Number 3–4 ISSN 1082–9873. www.dlib.org/dlib/march09/powell/03powell.html – [Stand: 15.04.2010]

9 Ausgangspunkt war das von der EU geförderte Projekt METADATA ENGINE, bei dem eine erste Version einer allgemeinen Frakturerkennung entwickelt wurde. http://meta-e.aib.uni-linz.ac.at/ – [Stand: 15.04.2010]

10 Announcing Tesseract OCR. Wednesday, August 30, 2006 Post by Luc Vincent, Uber Tech Lead. http://googlecode.blogspot.com/2006/08/announcing-tesseract-ocr.html – [Stand: 15.04.2010]

11 Google Code. http://code.google.com/p/ocropus/ – [Stand: 15.04.2010]

12 Estonian National Library Runs ABBYY Recognition Server for Gothic Text OCR. www.abbyy.com/Default.aspx?DN=74456adc-f778-4c2f-add4-ee865a355393 – [Stand: 15.04.2010]

13 Gotscharek, Annette; Reffle, Ulrich; Ringlstetter, Christoph: On Lexical Resources for Digitization of Historical Documents. In: Document Engineering. Proceedings of the 9th ACM symposium on Document engineering. 2009, Tabelle 4, S. 197.

14 CCS GmbH (www.ccs-digital.info/). CCS ist besonders im internationalen Raum erfolgreich und kann auf Projekte in Norwegen, Finnland, den USA und Australien verweisen. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) hat bereits 2005 die OCR Erkennung und Segmentierung für die Neue Zürcher Zeitung vorgenommen und nach eigenen Angaben das Projekt erfolgreich abgeschlossen. Vgl.: www.iais.fraunhofer.de/nzz.html. PPS PrePress-Systeme GmbH ist besonders in Deutschland stark vertreten und hat eine Reihe regionaler und überregionaler Zeitungen und Zeitschriften digitalisiert: www.prepress-systeme.de/.

15 Tanner, Simon; Muñoz, Trevor; Ros, Pich Hemy: Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. In: D-Lib Magazine. July/August 2009. Volume 15, Number 7/8 ISSN 1082–9873 – [Stand: 15.04.2010] Vgl. auch die Vortragsfolien anläßlich der IMPACT OCR Konferenz, in denen auch die Aussagen zur historischen Verteilung der Erkennungsgenauigkeit im 19. Jahrhundert enthalten sind. Tanner, Simon: Measuring the OCR Accuracy across The British Library's 2 Million Page Newspaper Archive. Vortrag gehalten anlässlich der Konferenz: IMPACT Conference 2009. OCR in Mass Digitisation. Challenges between Full Text, Imaging and Language. www.impact-project.eu/news/ic2009/presentations/ – [Stand: 15.04.2010]

16 ALTO Website der Library of Congress. www.loc.gov/standards/alto/news.php. – [Stand: 15.04.2010]

17 Mahnke, Christian: OCR Renderfarmen und TEI. Vortrag beim Deutschen Bibliothekartag 2010. www.opus-bayern.de/bib-info/volltexte/2010/863/pdf/bibtag2010-mahnke.pdf. – [Stand: 15.04.2010]

18 Balk, Hildelies; Ploeger, Lieke: IMPACT: working together to address the challenges involving mass digitization of historical printed text. In: OCLC Systems & Services. 2009/4, S. 242 f.

Der Verfasser

Dr. Günter Mühlberger, Universitäts- und Landesbibliothek Tirol, Abteilung für Digitalisierung und elektronische Archivierung, Innrain 52, A–6020 Innsbruck, Tel.: ++43-(0)512 – 507-8454, Mail: guenter.muehlberger@uibk.ac.at