

OCR für Frakturschriften?

Neues aus dem Bereich automatischer Schrifterkennung

Steffen Wawra, Silke Wünderich

1. Kurze Entwicklungsgeschichte der Fraktur

Die Fraktur ist eine deutsche Renaissanceschrift, die zu Volksschrift wurde. Die Fraktur ist die letzte bedeutenden Schriftschöpfung der Frühzeit und entstand zu Beginn des 16. Jahrhundert und steht am Ende der Kette der Entwicklung der *gebrochenen Schriften* (Gotisch, Rundgotisch, Schwabacher und Fraktur).

Die Entwicklung dieser vier Arten von Schrifttypen, denen eine *fraktale* (*gebrochene*) Ausprägung eigen ist, vollzog sich als Reflex auf die Erfindung des Buchdruckes in der historisch erstaunlich kurzen Zeit von nur 70 Jahren, um sich dann in den Grundformen - trotz mannigfacher Abwandlungen - bis ins 20. Jahrhundert nicht mehr zu verändern.

Über vier Jahrhunderte lang war Fraktur die meistverbreitete Schrift in Deutschland.

Neben den populären und wissenschaftlichen Büchern, den Tageszeitungen und Zeitschriften wurden auch *Akzidenzien* (Geschäft- und Kleindrucksachen) in Fraktur gesetzt.

Die neue Type Fraktur - die sich aus der schmalen *Gotisch*, der schwungvollen *Schwabacher* und der in den Kanzleien geschriebenen *Kurrent* formte - hatte sich bis 1600 vollständig gegen die alteingesessenen Schriften durchgesetzt und wurde seitdem als eigentliche deutsche Schrift empfunden. Im Gegensatz zum in Fraktur gesetzten deutschsprachigen Schrifttum wurde für lateinische Texte seit etwa Mitte des 16. Jahrhundert *Antiqua* verwendet.

Der Formencharakter der Fraktur wird bei den Kleinbuchstaben durch den Wechsel von bogenförmig geschwungenen und geraden Schäften bestimmt, ihre Großbuchstaben sind bauchig und geschwungen. Inzid der Herkunft aus der Schreibschrift sind die kleinen Schnörkel, die sog. „*Elefanterrüssel*“. Dabei erscheinen die schlanken Klein- und die opulenten Großbuchstaben harmonisch aufeinander abgestimmt und ergeben ein flüssiges und gut lesbares Schriftbild.

Der Gesamtcharakter der Schrift erscheint als Produkt jener Zeit dem barokkem Empfinden zu entsprechen: er ist unruhig, schwellend und dunkel.

Jedoch nicht nur im deutschen Sprachraum, sondern auch in Skandinavien, in den baltischen Ländern, in Polen und den slawischen Gebieten der habsburgischen Monarchie hielt die Fraktur Einzug.

Unter dem Einfluss der Kupferstichtchnik geriet die Fraktur jedoch in Verfall: auch fand sich der zeitgenössische Geschmack - vom Klassizismus zunehmend bestimmt - in *Antiqua* weitaus mehr angenommen, die sich daraufhin anschickte, auch im Bereich der deutschsprachigen Drucke führend zu werden.

Wie die Fraktur im 16. Jahrhundert die Elemente ihrer Vorgängerschriften weiterentwickelte, so wurde der nun Versuch unternommen, sie zum Ende des 18. Jahrhunderts ebenfalls dem Zeitgeist anzupassen: in der von *UNGER* geschaffenen Type der Schrift „*Die neue Cecilia*“ von K. Ph. Moritz (1794) gelang dies über eine leichtere und hellere Gestaltung der Fraktur. Der allgemeine künstlerische Verfall im 19. Jahrhundert führte jedoch zu einer starken Hinwendung zu Formen des Klassizismus, so dass diesen Versuchen der Adaption kein Erfolg beschieden war. Die *Antiqua* gewann auch das wissenschaftliche Buch.

Herausragende Zeitgenossen, deren Wirken eng mit der Entwicklung der deutschen Sprache überhaupt verbunden war (so die *Gebrüder Grimm*) traten entschieden für *Antiqua* ein.

Zum Bedeutungsverlust beim Setzen von wissenschaftlichen Werken des 19. Jahrhunderts trat ein allgemeiner Bedeutungsverlust im 20. Jahrhundert hinzu: blieben nach dem 1. Weltkrieg auch die gebrochenen Schriften im Gebrauch, so setzte sich *Antiqua* als weltoffenerer Typographie zunehmend durch und nahm dabei auch Einflüsse der Moderne (Bauhaus, Neue Sachlichkeit und Funktionalismus) auf.

Das Schicksal der Fraktur wurde unter der NS-Diktatur endgültig besiegelt: wurden zunächst Gotisch, Schwabacher und Fraktur als „arteigene deutsche Schriften“ klassifiziert, so wurde diese Haltung im Kriegsjahr 1941 vollständig revidiert. Als Grund für die über einen „Führerbefehl“ völlig überraschend erfolgte Umstellung sämtlicher Druckerzeugnisse innerhalb einer stark beanspruchten Kriegswirtschaft wird die propagandistische Wirkung vermutet, die mit einer besseren Lesbarkeit von Dekreten etc. in den besetzten Gebieten erhofft wurde.

2. Entwicklung einer OCR für Frakturschriften

Seltene und wertvolle Schriften des 16.-19. Jahrhunderts liegen häufig nur noch in wenigen Exemplaren vor und unterliegen aufgrund ihres Alters Benutzungseinschränkungen.

Sammlungen dieses Zeitraumes wurden in der Vergangenheit bis zur Gegenwart vornehmlich über die verdienstvollen Mikrofiche-Editionen von Verlagen - etwa die des Verlages Georg Olms, des Harald Fischer Verlages oder des K.G. Saur Verlages im Volltext erschlossen. Diese Produkte sind jedoch aufgrund ihrer Preisgestaltung nur in ausgewählten Bibliotheken nutzbar, zudem gibt es innerhalb der Nutzergemeinde differenzierte Aussagen über den Gebrauch des Mikrofiches für die wissenschaftliche Arbeit.

Aber auch eine Digitalisierung von Dokumenten zeitigt nicht notwendigerweise einen höheren Gebrauchswert: das „einfache“ Ablegen einer Dokumentenseite als Image (Bild- oder Graphikdatei) ergibt noch keinen wesentlichen Vorteil gegenüber der Volltextverfilmung per Mikrofiche und schon gar nicht gegenüber der Originalvorlage: es wird kein Mehrwert angeboten oder geschöpft, die Beschäftigung mit dem Text ist in den Formen Original, Fiche und Image von etwa gleicher Schwierigkeit - ohne jetzt hier auf die kulturhistorische Debatte eingehen zu wollen.

Das DFG-Schwerpunktprogramm „Retrospektive Digitalisierung von Bibliotheksbeständen“ wird hier mit dem Angebot von verteilten digitalen Dokumenten einen wesentlich höheren Mehrwert schaffen: nicht die 1:1-Abbildung eines Textes in einem anderen Medium, sondern das Angebot von Navigationshilfen und Recherchetools, die diesen Text erschließen, einschließlich der Verknüpfung mit anderen digitalisierten Dokumenten zu einer Virtuellen Forschungsbibliothek bestimmen die künftige Entwicklung.

Der entscheidende Schritt, um die Arbeit mit digitalisierten Dokumenten zu erleichtern und effektiv zu gestalten, ist die *Volltexterfassung* des Textes, d.h. die automatische Zeichenerkennung der auf diesen Textseiten verwendeten Schriftzeichen.

Solange diese Möglichkeit nicht gegeben ist - das Dokument ist also noch nicht recherchierfähig - stellt das klassische Register in Buchform die immer noch überlegenere, weil komfortablere Form dar.

Nach einer Volltexterfassung ist das Werk in einem beliebigen Textformat (ASCII, WORD, SGML, RTF, XML, PDF, HTML, o.a.) verfügbar. Damit ist es möglich,

- automatisch zu indexieren,
- mit Metadaten (Dublin Core, HTML Meta-Tags, o.a.) zu versehen,
- in ein WWW-fähiges Format zu transformieren oder im WWW über ein Plug-In anzubieten,
- in eine Datenbank einzubinden,
- oder mit anderen Dokumenten inhaltlich zu verknüpfen.

Die Volltextfassung ist für moderne Texte inzwischen Standard. Es gibt zahlreiche Anbieter von OCR (Optical Character Recognition)-Software, die bei den heute gängigen Schriftfonts hinreichend gute Ergebnisse liefern. Diese Programme versagen jedoch völlig bei Fraktur-Schriften.

3. Grundlagen der Schrifterkennung und der Fraktur-OCR

Gegenüber der Volltextfassung bei modernen Vorlagen, die im wesentlichen aus den beiden Schritten

1. Scannen der Vorlagen
2. OCR-Erfassung

bestehen, werden bei der Bearbeitung bibliophiler Texte zusätzliche Arbeitsschritte erforderlich. So können die alten und genauso wertvollen wie zerbrechlichen Werke nicht mit einem Standard-Scanner bearbeitet werden. In der Regel werden spezielle Buchscanner eingesetzt, um das übliche Aufbiegen der Bücher bei Standard-Scannern zu vermeiden.

Nach dem Scannen liegen die Seiten als Bilder im Rechner vor. Eine direkte OCR-Erfassung ist aber aufgrund der oft mangelhaften Druck- und Papierqualität nicht möglich. Es ist eine aufwendige Vorbearbeitung zum Entfernen von Schmutzstellen, zum Kontrastausgleich bei vergilbten Seiten, dem Begraden der Zeilen in der Nähe des Buchfalzes und dem Erkennen und Extrahieren von Zeichnungen und kalligraphischen Textstellen notwendig. Das sind nur einige notwendige Vorarbeiten.

Nach diesen Schritten greift das OCR-Programm auf die einzelnen Zeichen des Textes zu.

Erst jetzt beginnt die eigentliche Volltextfassung. Die Komplexität der im Hintergrund des Programmes ablaufenden Funktionalität ist dabei nicht sichtbar: die derzeitige Version Fraktur 3.2 bearbeitet über 200 Zeichen je Sekunde, so dass selbst Werke mit mehreren hundert Seiten innerhalb kürzester Zeit erfasst werden können. Wie das geschieht und welche Vorgänge dabei im Hintergrund ablaufen, soll im folgenden kurz erläutert werden.

Bei allen Verfahren zur Schrifterkennung werden für die einzelnen Schriftzeichen zunächst eine Reihe von Merkmalen bestimmt. Mit Hilfe dieses Merkmalsvektors wird danach eine Klassifizierung des Buchstabens durchgeführt. Die herkömmlichen Verfahren zur Schrifterkennung scheitern bei Frakturschriften, da diese eine starke Fluktuation innerhalb desselben Zeichensatzes aufweisen und die charakteristischen Merkmale dieser „alten“ Schriftfonts von denen moderner Schriften stark abweichen.

Bei der verwendeten Software wurde auf der Grundlage von modernen mathematischen Methoden (Wavelet-Algorithmen und nichtlineare Ausgleichsrechnung als Filter zur Bestimmung von Zeichencharakteristika) ein Verfahren zur optimierten Bestimmung von Merkmalsvektoren von Fraktur-Schriften entwickelt.

Mit Testmaterial, bestehend aus den ersten Seiten des vorliegenden Werkes, wird ein neuronales Netz trainiert, das letztendlich automatisch einen binären Entscheidungsbaum zur Klassifizierung der Fraktur-Zeichen generiert. Dabei wird jedem Zeichen zusätzlich eine Erkennungswahrscheinlichkeit zugeordnet. Kritische Zeichen werden farblich für eine manuelle Nachbearbeitung markiert.

Ob und in welchem Umfang eine Nachbearbeitung erforderlich ist, hängt entscheidend von zwei Faktoren ab:

1. *Qualität der gescannten Vorlagen*

Je nach Güte des Drucks können Erkennungsraten von über 99% erzielt werden.

2. *Art der Nutzung des erfassten Textes*

Soll der erfasste Text anschließend in einem modernen Schriftfont ausgedruckt oder in ähnlicher Art und Weise weiter verarbeitet werden, so ist eine Erkennungsrate von über 99,5% notwendig. Wird jedoch lediglich gefordert, dass der Text nach Schlagwörtern durchsucht werden kann, so ist eine Erkennungsrate von ca. 95% für eine komplexe Suchmaschine völlig ausreichend, bei einer Anzeige der gesuchten Textstelle wird dann auf die gescannten Originalseiten zurückgegriffen.

Nach dem Scannen der Vorlagen läuft die automatische Erfassung der Texte insgesamt also in einem vierstufigen Prozess ab:

1. Training der OCR
2. OCR-Erfassung der gescannten Seiten
3. Manuelle Nachkorrektur wenn erforderlich
4. Konvertierung in das gewünschte Format.

Im Bedarfsfall können sich Schritte zur Archivierung (auch Langzeitspeicherung), zur Integration in Datenbanken oder zum Speichern auf CD anschließen.

4. Beispielprojekt

Für die Berlin-Brandenburgische Akademie der Wissenschaften war das Goethe-Jubiläum im Jahre 1999 Anlass, um mit einer Reihe von Aktivitäten des Auswärtigen Mitgliedes der Preußischen Akademie der Wissenschaften Jo-

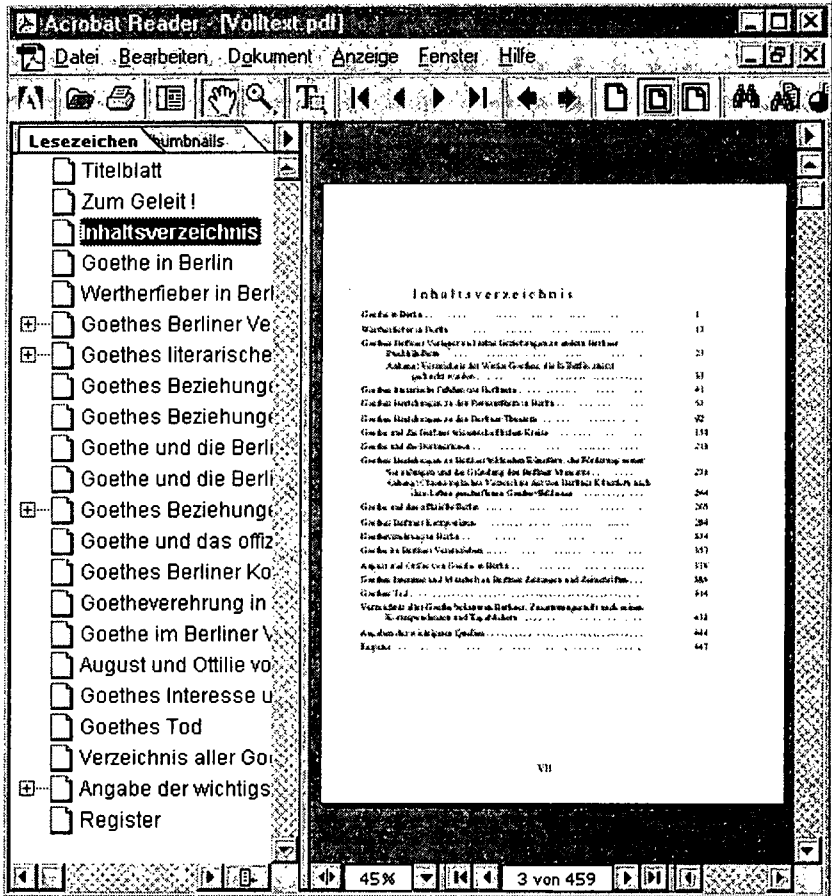
hann Wolfgang Goethe (gewählt im Sommer 1806) zu gedenken und die an der Akademie lebendige Goethe-Forschung zu dokumentieren.

Eine von Archiv und Bibliothek realisierte Ausstellung „*Goethe und die Berliner Wissenschaftsakademie. Eine Spurensuche in Archiv und Bibliothek*“, gezeigt vom 28. August bis zum 3. September 1999 in der Kleinen Bibliothek im Akademiegebäude am Gendarmenmarkt, illustrierte die Mitgliedschaft Goethes und stellte ausgewählte seltene und kostbare Buchausgaben aus, die sich in Akademiebesitz befinden.

Parallel zur Ausstellung wurde ein elektronisches Angebot der Ausstellung im Web realisiert. (<<http://bibliothek.bbaw.de:76/Goethe/home.htm>>)

Um das zweifelsfrei nicht spannungsfreie Verhältnis Goethes zu Berlin und der Berliner Akademie zu dokumentieren, wurde das 1925 im Leopold Klotz Verlag, Gotha erschienene Buch von E. Arnhold: „Goethes Berliner Beziehungen“ von der Firma *WiSenT GmbH* mit der von ihr entwickelten Fraktur-OCR erfasst. Bereits nach der automatischen Erfassung lag die Erkennungsrate über 99%, die durch die nachfolgende manuelle Nachkorrektur auf über 99,9% gesteigert wurde.

Um einen wie eingangs geschilderten Mehrwert zu realisieren, bestand die Forderung an das Produkt im Angebot einer datenbankähnlichen Lösung und der Möglichkeit der Präsentation im Web. Dieser ist die Firma *WiSenT* durch Verwendung des PDF-Dateiformats nachgekommen: die Lösung ist recherchierbar mit dem Acrobat Reader Search 4.0. Dieses frei verfügbare Programm ermöglicht eine digitalisierte Darstellung des Objektes, welche zum einen traditionelle Lesegewohnheiten (z. B. Suche im Register, Aufsuchen der betreffenden Textstelle im Dokument) über ein Inhaltsverzeichnis und ein Register abbildet - welche Links zum Volltext anbieten - und zum anderen eine komfortable Stichwortsuche einschließt, die auch eine Suche mit Operatoren (UND, ODER, * etc.) über das gesamte Volltextdokument ermöglicht.



Ansicht des Werkes mit Hilfe des Acrobat Reader Search 4.0 von Adobe

Das Produkt ist zu begutachten unter:

<http://bibliothek.bbaw.de:76/Goethe/GoethesBeziehungen/Volltext.pdf>

5. Fazit

Es wurde deutlich, dass der Einsatz von automatischer Schrifterkennung auch im Bereich von Frakturschriften möglich ist. Der Aufwand, welcher in die notwendige Vorarbeiten investiert werden muss, relativiert sich im Hinblick auf

den erreichten Mehrwert. Kontraproduktiv wirkt die große Anzahl von verwendeten Fonts, da die Merkmalsvektoren im ungünstigsten Fall neu berechnet werden müssen. Die Erfassung eines einzelnen Stückes wird sicher zu den Ausnahmen zählen, ein optimales Verhältnis von Aufwand und Nutzen wird am ehesten in der Volltexterfassung von Sammlungen (auch Periodica) entstehen.

Literatur:

Funke, Fritz: Buchkunde: Ein Überblick über die Geschichte des Buch- und Schriftwesens.- Leipzig: Verlag für Buch- und Bibliothekswesen, 1963.- S. 200ff.

Lexikon des gesamten Buchwesens / Hrsg. von Severin Corsten, Günther Pflug und Friedrich Adolf Schmidt-Künsemüller. - Stuttgart: Anton Hiersemann, 1991

Rehse, E. -G.: Gebrochene Schriften: Schaubuch, Nachschlagewerk und Hilfsbuch für den Umgang mit gebrochenen Schriften. - Itzehoe: Verl. Beruf u. Schule, 1998

