

Texterkennung und Frakturschrift

Wie Rechner alte Bücher erschließen – und was daran vielleicht trügerisch ist

Historische Texte durch Digitalisierung zu bewahren und weiterzuverbreiten, trägt zur Rettung unseres kulturellen und wissenschaftlichen Erbes bei, stärkt die Bildung und bringt frühere Gedanken in neue Diskussionen. Weil die Verdrängung der Frakturschrift Millionen Menschen im deutschen Sprachraum von einem wesentlichen Teil unserer schriftlichen Überlieferung abschneidet und weil heute mit Computern und digitalen Textbeständen gearbeitet wird, ist die Übertragung alter Texte unumgänglich. Sie bringt aber einiges mit sich, das zu hinterfragen sich lohnt.

Bevor in diesem Beitrag die grundsätzliche Vorgehensweise bei der maschinellen Texterkennung dargestellt wird, soll es um einige Begleitfragen rund um die Texterkennung gehen, die mit dem Wesen der Schrift und mit der kognitiven Leistung des Textverständnisses zu tun haben, nicht zuletzt aber auch mit der Auswahl und Filterung von Information durch Rechner, Netz-Suchmaschinen und nicht zuletzt durch den Leser selbst.

Das Lesen und Verstehen von geschriebenem Text ist eine Kulturleistung, die Intelligenz und Kulturgebundenheit voraussetzt, also die Kenntnis der Textsprache und der Eigenheiten der verwendeten Schrift, das Verständnis für das „Gemeinte“ (Hermeneutik) wie auch für Ironie. Schon das analoge Lesen durch den Menschen selbst erfüllt nicht immer alle Anforderungen, die daran zu stellen sind. Beispielsweise das Wissen um das Lang-s in der lateinischen Schreibschrift, das dem Rurrent=h gleicht, aber nachdem es außer Gebrauch gekommen war, mit eben diesem „h“ verwechselt wurde. So wurde dann z. B. der Name „Weiss“ fälschlich „Weihs“ geschrieben. Zu den wichtigsten Eigenschaften der Frakturschriften gehört die Tatsache, daß sie mit ihren Ligaturen und Sonderzeichen etliche Wörter als Schrift-Bilder wiedergeben, und zwar in erheblich größerem Ausmaß, als es die heutigen Antiqua- und sämtliche Festbreitenschriften vermögen. Frakturschriften brauchen dadurch alles in allem weniger Raum, in ihnen ist die Information also dichter gepackt, aber sie werden eben wortweise anstatt buchstabenweise gelesen. Hinzu kommen bestimmte typographische Feinheiten wie das runde r, auf die hier nicht weiter einzugehen ist.

Diese Hinweise laufen auf die Feststellung hinaus, daß das menschliche, kognitive Lesen und Verarbeiten zeit-

genössischer Texte ein grundlegend anderer Vorgang im Vergleich zum digitalisierenden, zerstückelnden und sortierenden Erschließen durch elektronische Geräte ist. Man kann ihnen zwar beibringen, „auf was hinaus“ sie einen vorgegebenen Frakturtext verarbeiten sollen, aber sie können das immer nur auf ihre Weise machen, die aber zu dem alten Text von seinem Wesen und Erscheinungsbild her nicht paßt. Aus diesem einen Passungsfehler heraus ergeben sich sämtliche Probleme, die im Zusammenhang mit der Digitalisierung von Frakturschriften auftreten: So etwa das richtige Erkennen des Lang-s, von Ligaturen, Sonderzeichen sowie Initialen, des weiteren die richtige Bewertung von Seitenkustoden, Signaturen, Bignetten und anderen Zierteilen. Da ältere Texte auch einen älteren Sprachgebrauch wiedergeben, müßte die Texterkennung auch auf dessen Eigenheiten Rücksicht nehmen. Es ist dabei mit der richtigen Erkennung der Trennweise von Silbengelenken in der klassischen Rechtschreibung (ck | k=k) nicht getan, sondern müßte auch Wörter- und Schreibvarianten bei der späteren Stichwortsuche erschließen. Das bedeutet, daß die Aufgaben, die sich aus diesem Übertragungsvorgang ergeben, erheblich anwachsen.

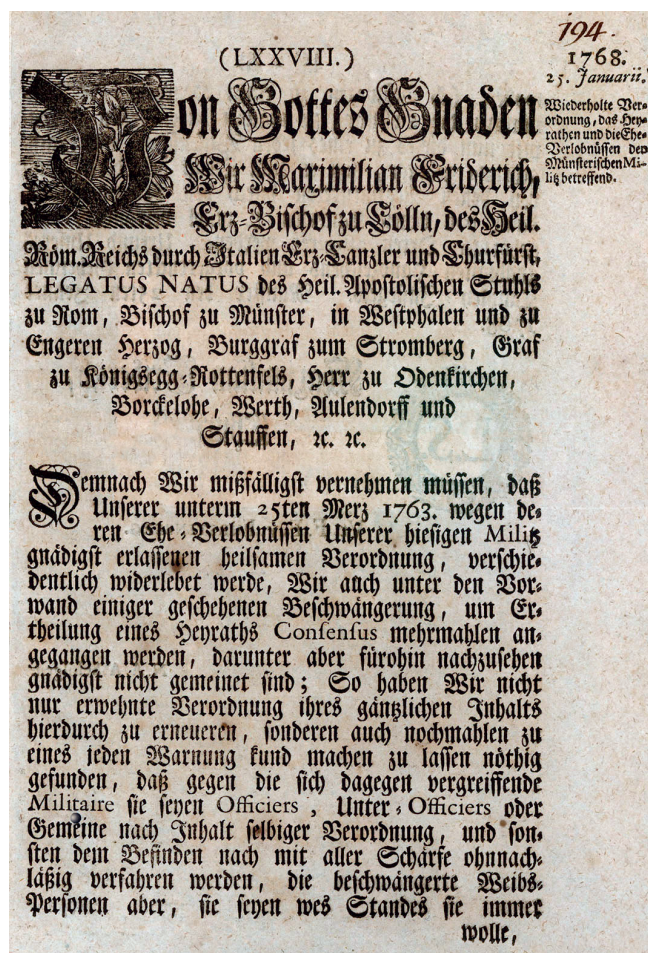
Es ist dabei auch wesentlich, daß heute in ganz anderer Weise geforscht und gearbeitet wird als noch vor drei Jahrzehnten: Früher erschloß man sich einen ersten Zugang zum Thema, indem man sich einen ersten Überblick über die Literatur- und Quellenlage verschaffte. Die so gefundenen Bücher fand man in Bibliotheken, wo man sie aufschlug und wo sie normalerweise zwei inhaltliche Zugänge eröffnen, nämlich vorne das Inhaltsverzeichnis und hinten das Register. Über den einen oder anderen Eingang konnte man dann finden, was das jeweilige Buch zur Frage zu sagen hat. Heute hingegen – man mag es besser oder schlechter finden – schreibt man einen oder mehrere Begriffe in ein Suchfeld und drückt eine Taste. Damit erhält man eine Trefferliste mit verschiedenen, mehr oder weniger brauchbaren Fundstellen, an denen die gesuchte Sequenz gefunden wurde. Handelt es sich um mehr als ein Dutzend Treffer, hängt die Rezeption der Ergebnisse maßgeblich davon ab, an welcher Position auf der Liste sie sich befinden.

Natürlich hat das viele Vorteile, möglicherweise findet man damit auch mehr – vielleicht auch nicht, wenn etwa die eine oder andere Textstelle nicht richtig übertragen wurde. Aber der wesentliche Unterschied liegt darin, daß die ältere Technik eine Intelligenzleistung ist und die zweite nicht. Wer selbst in Büchern sucht, braucht dafür ein gutes Maß an Auswahl- und Bewertungskompetenz, an Urteilsvermögen und kritischem Verstand. Wer das Suchen einem Computer überläßt, dem kann der Fehler unterlaufen, seine Belege nach Zufallsprinzip auszuwählen. Erheblich gefährlicher ist es, wenn zwischen Fundstellen und Forscher auch noch obskure Algorithmen privatwirtschaftlich organisierter Internetkonzerne eingezogen werden, die für den Suchenden eine Vorauswahl treffen. Dieser Punkt ist angesichts der weltweiten Digitalisierungsprojekte des Google-Konzerns von besonderer Bedeutung.

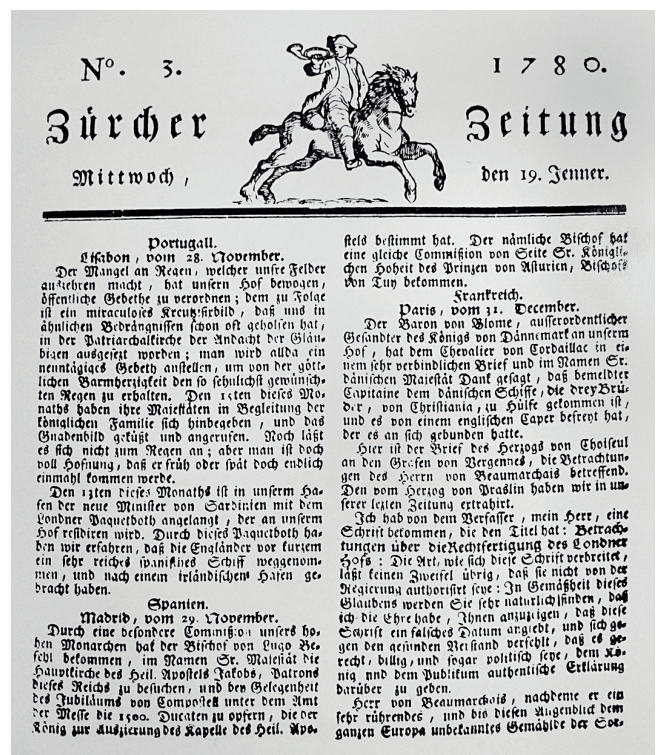
Für den kompetenten Forscher alter Schule wäre die Digitalisierung zu guten Bilddateien, mit grober Verschlagwortung versehen, völlig ausreichend, für die Wissenschaftler bis Mitte des 20. Jahrhunderts sozusagen der Himmel auf Erden. Sie würde die Originale sichern und

schonen, ihren Inhalt verbreiten und sogar Nachdrucke ermöglichen, ohne dafür das jeweilige Original zu benötigen. Für die heutige Zeit kann man das nicht ernsthaft fordern, ohne gleichzeitig Trakturschrift flächendeckend in die Schulen zu bringen. Folgerichtig laufen alle „richtigen“ Digitalisierungsprojekte heute auf die Texterkennung und Übertragung zu digitalen Dateien hinaus.

Die Frage, nach welchen Kriterien die Texte durch den Rechner erschlossen werden, wurde bereits gestreift. Das neue Paradigma, alles überall und jederzeit verfügbar haben zu müssen, stellt sich als trügerisch heraus. Denn bis vor einiger Zeit lag der Fehleranteil in Texten, die aus Trakturschriften übertragen wurden, noch bei 20 bis 30 %. Die Chance, annähernd alle gesuchten Textstellen zu finden, ist also nicht besonders hoch – trotzdem traut man dieser Technik eine weitreichende Erschließung des Inhalts zu. Die Beurteilung, wie gut die Texte übertragen worden sind, kann vom Computer ebenfalls nur mit einer bestimmten Restfehlerrate erfolgen. Für wirklich sichere Ergebnisse braucht man kompetente Menschen. Angesichts der schieren Massen bereits mangelhaft übertragener Texte ist das eine Aufgabe, deren Bewältigung noch einige Jahrzehnte in Anspruch nehmen wird. Die Annahme, daß alles, was jetzt schon digitalisiert wurde, auch schon in hoher Übertragungs- und Erschließungsqualität zur Verfügung steht, ist sicher falsch.



Wiederholte Verordnung, das Heyrathen und die Ehe-Verlobnüssen der Münsterischen Militz betreffend, 1768.



Dritte Ausgabe der Zürcher Zeitung, 1780; heute: Neue Zürcher Zeitung.

Wer gelegentlich historische Vorlagen für Nachdrucke oder neu gesetzte Ausgaben vorbereitet, kennt einige der Schwierigkeiten rund um die Digitalisierung schon aus eigener Erfahrung. Bereits das Aufbereiten alter Seiten als Bilddateien erfordert besondere Aufmerksamkeit und Arbeit. Obwohl einschlägige Programme eine Menge davon erledigen können, muß man das Ergebnis immer Seite für Seite durchgehen. Der Computer kommt in der Regel nicht gut mit Abweichungen vom Satzspiegel zurecht, wenn etwa ein Bild über den Text hinausragt oder wenn ein Kapitelbeginn mit mehr freier Fläche einhergeht als eine normale Textseite. Solche Dinge müssen immer in Handarbeit nachgebessert werden. Schwierig wird es auch bei verzerrten Seitenabbildungen. Sie können zwar auf verschiedene Art entzerrt werden, doch ist auch hier der kritische Blick des Benutzers erforderlich. Weitere Erfordernisse betreffen die Umstellung von Graustufen auf Schwarz=weiße Wiedergabe, das Ausmaß automatischer Entfleckung und die richtige Auswahl des Satzspiegels. Befindet sich außerhalb des Satzblocks ein Krümel oder eine Anmerkung, zieht der Computer den Rahmen bis dorthin. Das ist bei Arbeiten dieser Art ebenfalls Seite für Seite durchzugehen und zu berichtigen.

Sofern der Computer auch die Texterkennung bewerkstelligen soll, ist zu beachten, daß diese Aufgabe in mehrere Teilaufgaben zerlegt wird. Zunächst wird die Abbildung „binarisiert“, also in reine Schwarzweiß=Wiedergabe gebracht. Dabei ist natürlich die wichtigste Frage, welche Grauwerte noch zu „Weiß“ genommen werden und welche zu „Schwarz“. Wird diese Grenze geschickt gewählt, können Flecken, durchschlagende Schrift usw. ausgeschieden werden. Farbige Flecken kann man übrigens in einem vorangehenden Durchgang entfernen, indem man den dazugehörigen Farbkanal herunterregelt. So kann man zum Beispiel alles, was rot ist, vom Blatt verschwinden lassen. Allerdings auch Rotdruck – also Vorsicht.

Im nächsten Schritt werden unerwünschte Punkte und Bildfehler entfernt. Danach wird die Seitenabbildung entzerrt, was für die Texterkennung besonders wichtig ist, weil der Computer sich nur anhand gerade laufender Textzeilen richtig orientieren kann. Danach werden verschiedene Regionen auf dem Blatt bestimmt und klassifiziert. Dabei legt der Computer fest, in welchem Bereich sich Text befindet und wo ein Bild. Doch schon bei Initi-

alen kommen die Computer nicht mehr weiter – sie ordnen sie normalerweise als Bilder ein und übertragen dann nicht den damit dargestellten Buchstaben. Weitere Entwicklungsarbeiten sollen die Fehleranfälligkeit in diesem Bereich vermindern. Ob das gelingen wird, bleibt abzuwarten. Die Idee steht und fällt nämlich mit der Frage, ob der Computer der Aufgabe gewachsen sein wird, in einem vergleichsweise komplexen Bild die Grundform eines Buchstabens zu erkennen. Nach der Festlegung von Regionen werden die einzelnen Zeilen festgelegt, und das gelingt nur dann richtig gut, wenn die Zeilen gerade laufen. Wer selbst mit OCR (Abkürzung für optical character recognition = optische Zeichenerkennung) arbeitet, kann bestätigen, daß dieses Verfahren mit Schreibmaschinen=schrift am besten funktioniert, auch wenn es bei schwach getippten Buchstaben immer wieder zu Fehlern kommt.

Erst wenn all diese Vorarbeiten erledigt sind, beginnt die eigentliche Texterkennung. Idealerweise soll der Text so ausgegeben werden, wie er sich in der Vorlage befindet. Bei alten Buchseiten mit Frakturschrift tauchen immer wieder Antiquabuchstaben auf (bei Fremdwörtern und fremdsprachigen Zitaten etwa), die Ligaturen und Sonderformen wurden schon erwähnt. Für die OCR=Programme besteht die Schwierigkeit also nicht allein darin, ein a richtig als a wiederzugeben, sondern auch zu zeigen, was für ein a es ist. Weil es da viele Möglichkeiten gibt, kann es an jeder Stelle, wo der Buchstabe in der Vorlage nicht gut gedruckt ist oder wo sich etwas Schmutz befindet, zu Übertragungsfehlern kommen. Man kann sich also nicht darauf verlassen, daß ein Text, der von einem OCR=Programm ausgegeben wurde, in jeder Hinsicht einwandfrei ist.

Der Traum der Bibliothekare ist es, hinterher einen völlig originalgetreuen Text zu erhalten, der digital weiterverarbeitet, das heißt durchsucht und indiziert werden kann. Es gibt viele Beispiele dafür, daß das sehr gut gemacht sein kann, wenn etwa hinter der Abbildung des Originals (in Frakturschrift) der digitale Text läuft und man daher gleichsam im Original nach Wörtern suchen kann. Das sind technische Errungenschaften, die eine Texterschließung auf vorher nicht gekannte Art erlauben. Der Weg dahin ist allerdings recht weit und man sollte sich vor der Annahme hüten, wir wären auf ihm schon weit vorangeschritten.

