Nachhaltige Erschließung umfangreicher handschriftlicher Überlieferungen. Ein Fallbeispiel

Faßhauer, Vera
fasshauer@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main, Deutschland

Angesichts stetig wachsender Kapazitäten zur Speicherung großer Datenmengen nutzen Bibliotheken und Archive zunehmend die Möglichkeit, ihre Sammlungen zu digitalisieren und die Faksimiles online bereitzustellen. Rein konservatorischen Erwägungen folgend, belassen sie es dabei häufig bei der Erfassung der Metadaten und verzichten auf die weiterreichende inhaltliche Erschließung des Materials. So bleibt es oftmals allein dem Nutzer überlassen, sich einen Zugang zu den Inhalten der Sammlungen zu verschaffen.

Sofern es sich dabei um Druckwerke handelt, ist dieses Vorgehen durchaus hinreichend, zumal die Fähigkeit zur Lektüre von Antiqua- und Frakturdrucken zumindest im deutschsprachigen Raum allgemein vorausgesetzt werden kann. Da mit Hilfe der OCR-Technologie inzwischen selbst bei der automatischen Erkennung der Frakturschrift sehr gute Ergebnisse erzielt werden können, werden digitalisierte Druckwerke auch jenseits der genauen inhaltlichen Erfassung nutzbar, indem sie durch *distant reading* und statistische Zugänge erschlossen werden können.

Anders verhält es sich bei historischen Handschriften: Da heutzutage nur sehr wenige Personen über hinreichende paläographische Kenntnisse verfügen, stellen digitale Reproduktionen handgeschriebener Dokumente für den größten Teil des Publikums nicht viel mehr als bloße Abbildungen historischer Artefakte dar, die in ihrer Materialität zwar eine ganz bestimmte Oberflächenstruktur aufweisen, aber die darin transportierten Inhalte nur wenigen erfahrenen Lesern preisgeben. Zusätzlich erschwert wird die Lektüre im Fall von Tagebuchaufzeichnungen oder Notizbüchern, die nur selten auch für fremde Augen bestimmt waren.

Die Gewährleistung eines unbeschränkten und langfristigen Zugangs zu digitalisierten historischen Handschriftenarchiven ist also nicht *per se* gleichbedeutend mit einer unbegrenzten Zugänglichkeit, Nutzbarkeit und Weiterverwertbarkeit ihrer Inhalte. Eine wichtige Aufgabe der *Digital Humanities* bei der nachhaltigen Pflege des kulturellen Erbes ist deshalb eine über die bloß konservierende Ablichtung hinausgehende Erschließung der in diesen Textbeständen enthaltenen Informationen. Die Fragestellung ist also: Wie lassen sich diese Daten erfassen und für *close*- wie auch für *distant reading-Prozesse* aufbereiten? Lässt sich ein Zugang schaffen, ohne den gesamten Bestand manuell zu bearbeiten? Und inwieweit kann die klassische paläographische Hand- und Kopfarbeit durch automatisierte Prozesse ersetzt werden? Der Beitrag stellt diese Problemlage zunächst am Fallbeispiel der Senckenberg-Tagebücher exemplarisch dar und zeigt anschließend eine Lösungsstrategie auf, bei der manuelle und digitale Methoden kombiniert zum Einsatz kommen und bereits vorhandene, frei zugängliche Software verwendet wird.

Der Frankfurter Arzt Johann Christian Senckenberg (1707–1772) hinterließ handschriftliche Aufzeichnungen im Umfang von 53 Quartbänden mit je etwa 700 Seiten. Während die späteren Bände einesteils in ausführlichen ärztlichen Fallstudien und anderenteils in kritischen Bemerkungen über die sittlichen Missstände der Reichsstadt bestehen, befassen sich die mit *Observationes in me ipso factae* übertitelten ersten dreizehn Jahrgänge hauptsächlich mit dem Schreiber selbst. Da Senckenberg dem radikalen Pietismus nahestand und sich ganz aus dem kirchlichen Gemeindeleben zurückgezogen hatte, erfüllten die frühen Tagebücher hauptsächlich die Funktion eines religiösen Gewissensspiegels. Darüber hinaus notierte er über Jahrzehnte hinweg täglich seinen Speiseplan, sein Bewegungspensum und seine Stoffwechselaktivität ebenso detailliert wie die jeweilige Wetterlage, die Umgebungstemperatur und den Luftdruck, die er mit den wechselnden Zustände seines Gemüts und mit äußeren Umwelteinflüssen in Beziehung setzte. Der Zweck dieser akribischen Beobachtungen war seine diätetische und moralische Selbstoptimierung, welche sowohl eine untadelige Lebensführung im Diesseits als auch seine Erlösung im Jenseits gewährleisten sollte. Zugleich dienten sie der Erfassung und Deutung von Korrelationen zwischen Vorgängen in Leib, Seele, Natur und Kosmos.

Diese Aufzeichnungen stellen sich nicht nur dem heutigen Publikum als Big Data dar, sondern wurden bereits von ihrem Autor als riesiger Datenpool konzipiert: Zeitweise brachte er täglich bis zu 5000 Wörter in deutscher und lateinischer Sprache zu Papier, so dass er in manchen der insgesamt 43 Jahrgänge ca. 2600 Seiten sehr eng mit jeweils etwa 900 Wörtern beschrieb. Zugleich pietistisches Selbstzeugnis und wissenschaftliche Aufzeichnungsform, ist dieser schriftlich fixierte und weltweit einzigartige Erfahrungsschatz eine Fundgrube für die Erforschung der frühneuzeitlichen Religions- und Wissenschaftsgeschichte. Darüber hinaus wirft er neue historische Schlaglichter auf die aktuell diskutierten Möglichkeiten und Grenzen der Nutzung großer Datensammlungen und ihr Verhältnis zur Theorie (vgl. Anderson 2008; boyd et al. 2012, Rosenberg 2014).

Mit Förderung durch die Dr. Senckenbergische Stiftung wurden die insgesamt ca. 40.000 Quartseiten in hochaufgelöster Form digitalisiert und von der Universitätsbibliothek Frankfurt unter Open Access-Bedingungen online zur Verfügung gestellt (UB Frankfurt

2013–2016). Am Frankfurter Institut für Deutsche Literatur und ihre Didaktik entsteht derzeit eine TEI/XML-basierte Online-Edition der Aufzeichnungen, welche gleichfalls von der durch den Autor selbst begründeten Stiftung finanziert wird. In Anbetracht ihres riesigen Umfangs und der schwer entzifferbaren Handschrift Senckenbergs ist eine zeitnah fertigstellbare Volltextedition des Gesamtbestandes schwerlich möglich und wäre aufgrund der bei dieser Aufzeichnungspraxis naturgemäß häufig auftretenden inhaltlichen Redundanzen auch nicht sinnvoll. Aus diesem Grund wurde im Vorfeld eine repräsentative Bandauswahl getroffen, welche nach den Maßgaben der historischen Signifikanz, der thematischen Vielfalt und der größtmöglichen Vermeidung von Redundanzen erfolgte. Die inhaltliche Komplexität und die auf schnelle Erfassung großer Datenmengen ausgerichtete Schreibroutine des Autors machen zudem eine Transkriptionsweise erforderlich, die weit über die diplomatisch-zeichengetreue Textwiedergabe hinausgeht: Abgesehen von der Tatsache, dass es sich um einen halb frühneuhochdeutschen und halb lateinischen Text handelt und der Schreiber oftmals mehrfach in einem Satz zwischen beiden Sprachen hin- und herwechselt, sind viele der Sätze so komplex, dass der Leser zum Verständnis auf alle verfügbaren grammatischen Merkmale angewiesen ist. Vor allem die morphologischen Merkmale sind aber im Deutschen wie auch im Lateinischen hauptsächlich in eben jenen Wortendungen enthalten, welche häufig durch Abkürzung entfallen. Ein ähnliches Problem besteht auch hinsichtlich der Symbole, die größtenteils dem alchemistischen Kontext entstammen: Sie können ein ganzes Wort oder auch nur einen Teil davon ersetzen, bis zu vier verschiedene Wortbedeutungen und noch viel mehr grammatische Formen repräsentieren und in völlig verschiedenen semantischen Umgebungen erscheinen. Um dem Leser einen hinreichenden Zugang zum Sprachgebrauch des Autors zu bieten und ein Textverständnis überhaupt erst zu ermöglichen, müssen Abkürzungen und Symbole ihrem kontextspezifischen Zusammenhang entsprechend aufgelöst und sowohl semantisch als auch grammatikalisch in den Text eingepasst werden.

Auf den ersten Blick scheinen digitale Methoden hier kaum weiterzuhelfen: Zu wenig deutlich ist die Schrift, zu komplex die Inhalte, zu spezifisch das Vokabular und zu mehrdeutig die einzelnen Zeichen. Hinzu kommt noch, dass sich sowohl Senckenbergs Handschrift als auch die Inhalte seiner Aufzeichnungen im Verlauf von vier Jahrzehnten stark veränderten und mithin ganz neue graphische Muster hervorbrachten. Wenngleich die Transkription der Texte nur händisch erfolgen kann, wird dadurch doch ihre Maschinenlesbarkeit überhaupt erst gewährleistet und damit die grundlegende Voraussetzung für automatisierte Prozesse sowie die Anwendung, Weiterentwicklung und Schulung der sie ermöglichenden Technologien geschaffen. So erfordert das Training des Tools Transkribus (Universität Innsbruck o.J.) zunächst einmal eine ausreichende Menge an manuell erzeugten und präzisen Texttranskriptionen und die anschließende händische Überarbeitung des Outputs (vgl. Transkribus Wiki o.J.). Aufgrund der wachsenden Nachlässigkeit der Handschrift und der inhaltlichen Heterogenität der drei Unterbestände muss der Lernprozess für jeden Teilbestand separat erfolgen. Nach Abschluss dieses Lernprozesses ist jedoch zumindest eine halbautomatische Texterfassung möglich. Der erkannte Text kann anschließend elektronisch durchsucht und wissenschaftlich ausgewertet werden.

Ein ähnliches Verhältnis zwischen manuellen und automatisierten Prozessen besteht hinsichtlich der inhaltlichen Erschließung der Texte. Da sie von einem einzigen Schreiber mit umfassender grammatischer Bildung stammen, liegt nur eine geringe orthographische Varianz bei der Schreibung ein- und desselben Wortes vor. Anders als in heterogenen Korpora, die Texte mehrerer Schreiber mit unterschiedlichem Bildungshintergrund und sprachgeografischer Herkunft versammeln, ist deshalb eine vorherige händische Normierung der Grafie nicht notwendig (vgl. demgegenüber Faßhauer et al. 2013, Faßhauer et al. 2014). Mit Hilfe der vorliegenden Transkriptionen kann deshalb ein effizientes Training des Tools TreeTagger (Schmid 1994-) für das Frühneuhochdeutsche und Neulateinische vorgenommen werden. Die halbautomatisch generierten Lemmata und Part-of-Speech-Tags, welche sowohl für die manuellen Transkriptionen als auch für die automatisch erfassten Texte erstellt wurden, werden anschließend in den Partitur-Editor der Software EXMARaLDA (Hedeland et al. o.J.) eingespielt. Mit dem zugehörigen Analysetool EXAKT werden per RegEx-Suche auf der Lemmaspur zunächst alle Nomina herausgefiltert und in einem manuellen Prozess Schlagwörter ausgewählt (ähnlich auch Biehl et al. 2015). Aus der Untermenge der großgeschriebenen Substantive, die sich mittels der automatischen Sortierfunktion der Trefferliste leicht ermitteln lassen, werden alle Personen- und Ortsnamen entnommen. Anhand dieser Recherchezugänge kann nun das gesamte Korpus systematisch recherchiert werden. Die von EXAKT angebotenen Anfragen über RegEx und Levenshtein-Distanzen ermöglichen dabei eine schreibweisentolerante Begriffsermittlung, wodurch mancher HTR-Lesefehler überwunden werden kann.

Bibliographie

Anderson, Chris (2008): „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete“, in: *Wired Magazine* http://www.wired.com/2008/06/pb-theory/

Biehl, Theresia / Lorenz, Anne / Osierenski, Dirk (2015): „Exilnetz33. Ein Forschungsportal als Such- und Visualisierungsinstrument“, in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities* (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1).

Boyd, Danah / Crawford, Kate (2012): „Critical Questions for Big Data“, in: *Information,*

Communication & Society 15 (5): 662–679 10.1080/1369118X.2012.678878.

Fasshauer, Vera / Lühr, Rosemarie / Prutscher, Daniela /Seidel, Henry (2013): Dokumentation der Annotationsrichtlinien für das Korpus *Frühneuzeitliche Fürstinnenkorrespondenzen im mitteldeutschen Raum*. .

Fasshauer, Vera / Lühr, Rosemarie / Prutscher, Daniela /Seidel, Henry (2014): *Fürstinnenkorrespondenz* (version 1.1), Universität Jena, DFG. LAUDATIO Repository. http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm .

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (o.J.): *EXMARaLDA. Werkzeuge für mündliche Korpora* http://www.exmaralda.org/ [letzter Zugriff 21. März 2016].

Rosenberg, Daniel (2014): „Daten vor Fakten“, in: Reichert, Ramón (ed.): *Big Data: Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie*. Bielefeld: transcript Verlag 133–156.

Schmid, Helmut (1994-): *TreeTagger. A part-of-speech tagger for many languages*. http://www.cis.unimuenchen.de/~schmid/tools/TreeTagger/ [letzter Zugriff 20. August 2016].

Transkribus Wiki (o.J.): *Transkribus-Benutzeranleitung*. https://transkribus.eu/wikiDe/index.php/Hauptseite [letzter Zugriff 20. August 2016].

UB Frankfurt =Universitätsbibliothek Frankfurt am Main (2013-2016): *Nachlass Johann Christian Senckenberg*. http://sammlungen.ub.uni-frankfurt.de/senckenberg/nav/index/all [letzter Zugriff 20. August 2016].

Universität Innsbruck (o.J.): *Transkribus*. https://transkribus.eu/Transkribus/ [letzter Zugriff 20. August 2016].