„Korporatheken“: Die digitale und verdatete Bibliothek

Noah Bubenhofer[1] und Klaus Rothenhäusler[2]

[1]*Universität Zürich, Institut für Computerlinguistik,*
bubenhofer@cl.uzh.ch
[2]*Universität Zürich, Institut für Computerlinguistik,*
rothenha@cl.uzh.ch

Abstract

Mit dem digitalen Zeitalter ergeben sich für die Geisteswissenschaften neue methodologische Möglichkeiten, deren Wurzeln jedoch weit zurück reichen. Am Anfang steht der Index, der inzwischen mit viel weniger Aufwand als früher zum Volltextindex wird. Mit der anschliessenden „Verdatung“ werden Sprachdaten verrechenbar und damit anders nutzbar. Welche Rolle können Bibliotheken in einer verdateten Welt spielen? Der folgende Beitrag betont nicht nur die technischen Möglichkeiten, sondern auch die Probleme, spezifisch aus geistes- und kulturwissenschaftlicher Sicht.

The digital era produces a whole range of new approaches for research in the humanities whose origins, however, can be traced far back. Their very source is the index, which can take the form of a full text index much easier than before. By turning text into data it becomes computationally tractable and can be used in novel ways. Which role can libraries play in a digitized world? This paper explores technical possibilities but it also points to problems specific for arts and social sciences.

1 Einleitung

Es ist eine Trivialität, auf das grosse Interesse an quantitativen Textanalyseverfahren im digitalen Zeitalter hinzuweisen. Ebenso, dass dieses Interesse sowohl Chancen als auch Gefahren für die klassischen Institutionen der Textverwaltung, die Bibliotheken, zeitigen. Zu den Chiffren dieser Zeit gehören nicht nur die Suchmaske von Google, Google Books, die Timeline bei Facebook und Snowden und die Geheimdienste, sondern in der Akademia Schlagworte wie Big Data, Data Mining, Visual Analytics oder Digital Humanities. Diese Chiffren stehen für die Bedrohungen der klassischen Bibliothek, da sie neue Strategien der Suche, Organisation und Analyse von Daten repräsentieren.

Auch in den Geistes-, Kultur- und Sozialwissenschaften entwickelten sich in den letzten Jahrzehnten neue Herangehensweisen an digitale (oder digitalisierbare) Daten. Das Label „Digital Humanities“ vereinigt eine breite Palette solcher Methoden und Theorien, wobei die disziplinären Unterschiede mitunter gross sind. In der Linguistik beispielsweise hat sich mit der (modernen) Korpuslinguistik in den letzten Jahrzehnten eine Methodologie entwickelt, die nicht nur die neuen Möglichkeiten der computerisierten Textanalyse nutzt, sondern mit

der auch eine neue Modellierung des Untersuchungsgegenstands selbst, der Sprache, einher geht. Es ist dies, kurz gefasst, ein neues Interesse für die sprachliche Oberfläche (Sprachgebrauch statt Sprachsystem) und für die Zusammenhänge zwischen Sprachgebrauch und sozialem Handeln (Feilke 1996; Feilke und Linke 2009). Diese Interessen ermöglichen einen neuen Zugriff auf Sprachgebrauch: Einen quantitativen Zugriff, mit dem aus der Analyse von grossen Korpora geschriebener oder gesprochener Sprache datengeleitet typische Muster des Sprachgebrauchs berechnet werden können. Diese Muster, typische Formulierungsmuster, Floskeln, Wortverwendungen etc. können sodann als Spuren sprachlichen Handelns gedeutet werden (Bubenhofer 2009).

Wir möchten nun im Folgenden aus einer spezifisch korpuslinguistischen Perspektive auf Bibliothekskulturen schauen. Denn: Welche Bedürfnisse hat eine digitale Geisteswissenschaftlerin an eine Bibliothek? Und welche Forschungsinteressen, welches Text- und Analyseverständnis bringt sie mit? Und ergeben sich daraus sogar Anregungen für eine andere Bibliothekskultur, für andere Funktionen und Rollen von Bibliotheken?

2 Trails

1945 schrieb Vannevar Bush, damals Direktor des Büros für „Scientific Research and Development“ der US-Regierung, den viel beachteten Artikel „As We May Think“, in dem er aus ingenieurtechnischer Sicht die Fortschritte von analogen Rechnern voraussagte. Darin beschrieb er „Memex“, eine flexible Privatbibliothek:

> „A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.“ (Bush 1945:121)

Der Memex würde mit Hilfe von Mikrofilm und einer komplexen Mechanik in der Lage sein, beliebige Dokumente zu finden und untereinander oder mit weiteren Notizen zu verknüpfen. Das Ergebnis wären „Trails“, also Pfade, die das menschliche Lektüreerlebnis, das Ergebnis des hermeneutischen Prozesses, speichern und für andere Memex-Nutzer anwendbar machen würde. Bush gilt mit seiner Vision als einer der Begründer des Data Minings und des Hypertextes, der aber bereits während des Zweiten Weltkrieges Maschinen erschuf, die tatsächlich in diese Richtung gingen. So beispielsweise den „Rapid Selector“ zur Verwaltung grosser Datenbestände, dessen Verwendung auch Roberto Busa, Theologe und Linguist, erwog, um den Index Thomisticus zu erstellen (Busa 1951).

Vielleicht muss Bush gar nicht so sehr als einer der Gründerväter des Data Minings angesehen werden, schliesslich war er ganz in der Welt der analogen Maschinen verhaftet, sondern steht in einer langen Tradition *diagrammatischer* Experimente mit Daten, die er zum Äussersten der feinmechanischen Möglichkeiten trieb. Im Zentrum dieser Tradition befindet sich das Diagramm:

> „Diagramme sind, so könnte man vielleicht sagen, graphische Abkürzungsverfahren für komplexe Schematisierungen. Sie bewahren ein Minimum ästhetischer Anschauung das wir benötigen, um zu verstehen, wovon die Rede ist, vor allen um uns von abstrakten Sachverhalten in buchstäblichem Sinn ein Bild machen zu können.“ (Stetter 2005)

Diagramme können sehr komplex sein, erscheinen uns oft in wohlbekannten Formen wie Streu-, Balken-, Torten- und Liniendiagrammen, Karten etc., aber auch in weniger offensichtlichen Varianten: als textstrukturelle Merkmale wie Listen (Steinseifer 2013), als Zahlenstrahl und Sternbilder (Krämer 2013) oder als Partituren (Bubenhofer im Druck[b]). Gemeinsam ist Diagrammen, dass sie Gebrauchsbilder sind, mit denen operiert werden kann (Krämer 2009). Wissen wird in eine diagrammatische Form überführt (eine Karte zeichnen, Summen als Balken darstellen), die dabei hilft, neue Erkenntnisse zu finden, indem mit dem Diagramm gearbeitet wird (Karte lesen → Weg finden, Balken nach Grösse ordnen → Zusammenhänge erkennen). Bushs Memex erlaubt es, Tausende von Dokumenten gleichzeitig griffbereit zu haben (sie sind über Siglen ansprechbar und werden durch Mikrofilme repräsentiert) und Verknüpfungen darin zu erstellen, die wiederum abgespeichert werden können. Einen ähnlichen Versuch unternahm Agostino Ramelli 1588 mit der Skizze eines Leserades (Ramelli 1588:317), das einem Mühlrad gleicht, auf dessen Schaufeln aufgeschlagene Bücher abgelegt werden. Vor dem Rad sitzend kann gleichzeitig gelesen und durch einfaches Drehen am Rad zum nächsten Buch gesprungen werden.

Über viele Stationen hinweg (Zettelkasten, verschiedene Formen von Registern und Indizes, Visualisierungen hierarchischer Strukturen etc.) lässt sich eine Linie bis hin zur modernen Korpuslinguistik und dem Paradigma des Data Minings ziehen (Bubenhofer im Druck[b]):

> „Die [...] praktisch geübte und methodisch reflektierte Isolation einzelner Textpartien zum locus bedeutet, die Vielfalt tradierter Texte als einen riesenhaften Speicher zu behandeln, der in sich die Summe des verfügbaren Wissens inkorporiert." (Siegel 2009:37)

Und mehr noch: die Einheit des Einzeltextes wird aufgelöst und der neue locus, also das Ensemble einzelner Textpartien, ermöglicht eine neue Sicht auf die Texte. Es ist, als ob die Ordnung des klassischen Bibliothekskatalogs sabotiert würde, der ja qua Metadaten zu jedem Titel den Text zur kleinsten zu ordnenden Einheit macht. Um andere Ordnungen zu ermöglichen, dienen diagrammatische Methoden: Immer geht es darum, eine Art Diagramm zu entwickeln, das Texte oder Textteile repräsentiert, jedoch kraft seiner Zeichenhaftigkeit Operationen erlaubt, die mit den Texten selber nicht möglich wären. Insofern ist die Bibliothek, auch mit klassischem Bibliothekskatalog, eine riesige diagrammatische Maschine, die eine Vielzahl von Diagrammen nutzt (angefangen beim Zettelkatalog), um neue Textordnungen zu ermöglichen. Mit den „Trails" in Bushs Memex kommt aber eine Idee hinzu, die mit der modernen Deutung als „Hypertext" nur unzureichend abgedeckt ist. Trails haben das Potenzial, mehr zu sein als über Hyperlinks verkettete Texte, nämlich dann, wenn Texte digital verrechenbar werden und die Arbeit mit Daten vergemeinschaftet wird.

3 Volltextindex

Voraussetzung für die digitale Verrechenbarkeit ist jedoch die Digitalisierung, um aus Büchern sogenannte „Volltexte" zu erzeugen. Die Bibliotheken sind seit einiger Zeit an vorderster Front dabei, Bücher zu digitalisieren, wie auch Google mit dem Google Books-Projekt. Das Scanning alleine ermöglicht jedoch noch keine Erzeugung von „Volltext"; dafür müssen die Texte in digital als Buchstabenzeichen repräsentierbare Symbole codiert werden, was entweder über Abschreiben oder OCR geschieht. Dieser Schritt wird beschrieben als Entwicklung von der Digitalisierung zur Verdatung (engl. „Datafication"), als Grundlage für „Big

Data“-Anwendungen, die ändern würden, wie wir „live, work and think“ (Mayer-Schönberger und Cukier 2013).

Zumindest für den Umgang mit Text begann „Verdatung“ mit Busas Pionierarbeiten zum bereits oben erwähnten Index Thomisticus lange vor den von den Big-Data-Apologeten angeblich revolutionären Google Books-Projekt. Busa wollte als Theologe Texte besser *verstehen*, indem er die alte diagrammatische Umformung des Textes in Wortindizes und Konkordanzen nutzte:

> „To break up an author's text into phrases, transcribe them on cards, retranscribe each card as many times as there are words in the phrase, put them into alphabetical order, so as to have in a card file, in the proper order, all, I repeat, materially all the prepositions, conjunctions, adverbs, adjectives, nouns, verbs – for example all the *in*, all the *not*, all the *here* all the *now*, which have dripped from the author's pen, and so on for each single word of his vocabulary [. . .].“ (Busa 1951:10)

Die Bedeutung dieses Wunsches liegt bei der Systematizität und Vollständigkeit des Vorhabens, mit dem die Textoberfläche als Ausgangspunkt für Deutung verstanden wird. Um eine Sammlung von 10 Millionen Karten in Zettelkästen, wie etwa beim Thesaurus Linguae Latinae tatsächlich vorhanden, vermeiden zu können, behalf sich Busa als erster mit IBM Lochkartenrechnern und erstellte so den ersten digital repräsentierbaren *index verborum* (Busa 1951; Bonfanti 2012).

Der Wunsch, grosse Textsammlungen so aufzubereiten, dass ihr Wortmaterial auf eine Art repräsentiert wird, mit dem beliebige diagrammatische Operationen ermöglicht werden, hat also eine lange Tradition, die weit in die Zeit vor dem Computer hineinreicht. Natürlich erleichtert der Computer auf der Basis entsprechend codierter Daten (digitalisiert und verdatet) diese Operationen ungemein. Im Bibliothekswesen zeugen die Digitalisierungsbemühungen davon, dass die Digitalisierung selbst allerdings schon aufwändig genug ist, so dass die eigentliche Verdatung überhaupt nicht in Angriff genommen werden kann. Diese wäre aber die Voraussetzung, um viel mehr zu erreichen als „nur“ Volltextindizes, nämlich die Verrechenbarkeit von sprachlichen Ausdrücken und Texten.

4 Verrechenbarkeit

> „One of the meanings of *ass* is its habitual collocation with an immediately preceding *you silly*, and with other phrases of address or of personal reference. Even if you said ‚An ass has been frightfully mauled at the Zoo‘, a possible retort would be, ‘What on earth was he doing?’“ (Firth 1957:195)

Die Semantik eines Ausdrucks ergibt sich aus dem distributionellen Verhalten in Texten. Diese Erkenntnis ist ein wichtiger Schlüssel, um mit quantitativen Methoden Texte zu „verstehen“. Die dafür nötige diagrammatische Operation sieht dabei vor, das Distributionsverhalten eines Wortes in einer Matrix zu fassen, mit der die Häufigkeit des Kovorkommen des Wortes mit anderen Wörtern aufgeführt wird. Dadurch ergibt sich ein „Kollokationsprofil“ des Wortes, ausgedrückt alleine durch die Häufigkeiten, mit denen das Wort zusammen mit allen anderen Wörtern auftritt. Dieses Profil ist mathematisch gesehen ein mehrdimensionaler Vektor, eine Zahlenreihe, in einem Vektorraum. Die Ähnlichkeit von Wörtern kann danach

geometrisch gemessen werden als Abstand zwischen zwei Vektoren im Vektorraum. Der Vektor als Zahlenreihe ist das Ergebnis eines „transkriptiven Verfahrens“ (Jäger 2008), mit dem die Semantik des Ausdrucks in ein anderes symbolisches System überführt worden ist und nun diesen Regeln gehorcht: Der Vektor kann mit anderen Vektoren verrechnet werden; so ist es z.B. möglich, die Distanz zwischen den Vektoren von „Italien“ und „Rom“ zu „Frankreich“ hinzuzurechnen, um als passendsten Ausdruck dieses Vektors „Paris“ zu finden (solche Modelle werden in der Computerlinguistik als „Word Embeddings“ bezeichnet, vgl. z.B. Mikolov u. a. (2013)).

Dies funktioniert ähnlich für gesamte Texte, beispielsweise beim populären Verfahren des Topic Modellings, wo eine Textmenge maschinell in Cluster von ähnlichen Texten gegliedert wird, die sich aufgrund des Wortmaterials ähneln. Die Texte selbst werden dabei ebenfalls durch Vektoren repräsentiert (Graham, Weingart und Milligan 2012).

Das grundlegende Potenzial diagrammatischer Operationen, kombiniert mit der digitalen Repräsentation von Text, ermöglicht die Verdatung von Sprachgebrauch, macht Sprachgebrauch verrechenbar – auch mit nicht-sprachlichen Daten, um beispielsweise Korrelationen zwischen Sprachgebrauch und nicht-sprachlichen Entitäten zu finden. Möglich macht dies das „Metamedium“ Computer (Kay und Goldberg 1977; Manovich 2014) mittels der drei Grundfunktionen Codierung, Algorithmisierung und Formatierung (Heilmann 2012). Dabei wird jedoch auch deutlich, dass der Gewinn dieses Settings von diagrammatischen Operationen, Digitalisierung und Verdatung besonders aus geisteswissenschaftlicher Sicht dann besonders gross ist, wenn die Flexibilität des Metamediums Computer möglichst ausgereizt wird. Es ist eben gerade der Charme dieses Settings, immer wieder andere Daten immer wieder anders verrechnet und formatiert in diagrammatische Anordnungen zu bringen, um völlig unterschiedliche Rekontextualisierungen der Daten zu erzeugen und damit unterschiedliche Interpretationsmöglichkeiten zu eröffnen. Aus linguistischer Perspektive argumentiert Lauersdorf (im Druck) etwa dafür, die immer gegebene Ambiguität und möglichen Interpretationsvarianten von wissenschaftlichen Visualisierungen als Ergebnis von Datenanalysen gezielt noch zu potenzieren: „Use all the data! [...] View all the data! [...] View all the combinations! [...] View all angles! [...] Use all the techniques!“ (Lauersdorf im Druck)

Bereits die grundlegenden diagrammatischen Transformationen haben gezeigt, dass dadurch die Texteinheit aufgelöst wird. Im Modus der Verrechnung verdateter Sprache gelangen so Textpassagen in egalitäre Positionen: Die Fussnote in Text A kann den ähnlichen semantischen Gehalt haben wie die Einleitung in Text B und stellt sie somit potenziell der Einleitung in Text B gleich, obwohl sich vielleicht die Metadaten von Text A und B stark unterscheiden. Diese potenzielle Egalisierung von beliebigen Texteinheiten bietet die Chance, immer wieder andere Trails in den Daten zu finden, die die Texteinheiten sprengen.

Methodologisch gesehen vereinfacht die digitale Verrechenbarkeit induktive, datengeleitete Zugänge, mit denen vorerst ohne ausgearbeitete Hypothesen in den Daten auffällige Muster entdeckt werden können. Erst in einem zweiten Schritt werden diese Muster interpretiert und kategorisiert und dienen so der Generierung von neuen Hypothesen (Tognini-Bonelli 2001; Perkuhn 2007; Scharloth und Bubenhofer 2011; Bubenhofer, Scharloth und Eugster 2014). Während solche Ansätze nicht nur in der Korpuslinguistik, sondern auch im Data Mining, den Visual Analytics und verwandten Disziplinen für verschiedene Zwecke eingesetzt werden, scheinen die Anwendungsmöglichkeiten dafür bei den Bibliotheken vielleicht weniger offen-

sichtlich. Von Interesse wären aber gerade für die wissenschaftliche Literatur umfassendere Analysen von expliziten und impliziten Zitationen bis hin zu Paraphrasierungen, um einerseits im weitesten Sinne wissenschaftsgeschichtliche Fragen verfolgen zu können, andererseits aber auch einen Nutzen für die Bibliotheksrecherche anbieten zu können. Auf der Hand liegt die Auswertung der expliziten Zitation über die Literaturverzeichnisse, wie sie die verschiedenen Zitationsindizes (z.B. CiteSeer) bereits leisten. Interessanter wäre aber ein automatisches Entdecken von (wortwörtlich oder ähnlich) zitierten Passagen, wie es z.B. im kleinen Rahmen für Texte der Antike bereits gemacht wird.[1]

Ganz grundsätzlich könnten also verschiedene Methoden der maschinellen Datenanalyse klassische Katalogdaten anreichern; keine neue Idee, und da und dort auch bereits praktiziert. Besonders interessant dabei ist jedoch, dass dadurch die viel zu kurz greifende Rollenverteilung von Bibliothek als Dienstleister und Wissenschaft als Nutzniesser in Frage gestellt wird. Denn Inhalte, die von der Bibliothek verwaltet werden, werden damit zu Untersuchungsgegenständen, Methoden der Wissenschaft finden Eingang in Bibliotheken, und das wiederum könnte die Grundlage für einen regen Austausch über Theorien von Text, Textverwaltung, Zitation, Wissensgenerierung etc. darstellen.

5 Crowd

Bedeutend an einer Bibliothek ist selbstredend, dass sie mit wissenschaftlichen Communities interagiert. In Verbindung mit Web 2.0 ergeben sich daraus eine Reihe von Möglichkeiten, Interaktionen zwischen Community und Bibliothek und innerhalb der Communities selbst in Bezug auf die Bibliothek(en) möglich zu machen, auszuwerten, zu nutzen und für alle sichtbar zu machen. Darauf muss an dieser Stelle nicht weiter eingegangen werden, da diese Möglichkeiten (und Grenzen) insbesondere auch in dieser Zeitschrift schon mehrfach diskutiert wurden (Herrlich, Ledl und Tréfás 2013).

Vielleicht lohnt es sich aber an dieser Stelle nochmals klar zu machen, dass in Vannevar Bushs Idee der Trails die soziale Komponente des Teilens bereits vorhanden ist:

> „In fact, he [the owner of the memex] has a trail on it [a finding]. A touch brings up the code book. Tapping a few keys projects the head of the trail. A lever runs through it at will, stopping at interesting items, going off on side excursions. It is an interesting trail, pertinent to the discussion. So he sets a reproducer in action, photographs the whole trale out and passes it to his friend for insertion in his own memex, there to be linked into the more general trail. Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified.“ (Bush 1945:124)

Deutlich wird dabei aber auch, dass in der Wissenschaft nicht Trails der Art „Leser/innen, die Buch A gelesen haben, haben auch Buch B gelesen“ interessant sind, sondern solche, die beliebig kleine Einheiten von Wörtern, Absätzen, Passagen, Kapiteln, Bildern, Werken etc. miteinander verknüpfen und kommentieren. Die „Crowd“ hätte dann die Möglichkeit, ihr Wissen auf all diesen Ebenen für die ganze Community sichtbar zu machen. Und die

[1] Vgl. dazu das gerade gestartete Projekt „Digital Plato“ von Scharloth, Sier, Schubert und Molitor (https://forschung-sachsen-anhalt.de/project/digital-plato-tradition-reception-19278).

individuellen Trails könnten in aggregierter Form sichtbar gemacht werden, um die häufig und weniger häufig begangenen Wege zu zeigen.

6 Unberechenbarkeit

Quantitative Verfahren, Crowd, typische Korrelationen, häufig begangene Wege: Diese Stichworte verweisen bereits auf die problematische Seite der oben skizzierten Verfahren. Sie sind von komplexen Algorithmen abhängig, die nicht immer transparent sind, und sie begünstigen oft Korrelationen, die statistisch gesehen signifikant sind, also nicht unbedingt per se frequent, jedoch bezüglich bestimmten Vergleichsgrössen häufiger, als erwartet. Bei der Sortierung von Suchergebnissen beispielsweise sind Ranking-Algorithmen im Spiel, deren genaue Funktionsweise oft opak ist (Spinnler-Dürr 2013).

Web 2.0- und Crowd-Sourcing-Anwendungen folgen einer libertären Ideologie, die zwar den Beitrag des Individuums aufwertet, jedoch durch „such notions as collective intelligence, or through the idea of a long-tail economy, or the wisdom of crowds“ (Berry 2011:60). Der Beitrag des Einzelnen, der abweichende Trail, geht also unter.

Besonders aus geisteswissenschaftlicher Sicht ist Kritik gegenüber opaken Algorithmen („black boxes“, vgl. Rieder und Röhle (2012)), aber auch den in der Informatik herrschenden idealistischen und utilitaristischen Topoi (Fuller 2003:15; Goffey 2014:21; Ramsay 2011) angebracht. Erhellend ist beispielsweise die Genese der Visual Analytics, einer Analysemethode, die grosse Datenmengen visualisiert, um sie überhaupt analysierbar zu machen. Das sind Methoden, die natürlich auch mit Gewinn für Bibliotheksbestände eingesetzt werden (können). Allerdings erkauft man sich mit der Nutzung solcher Methoden und Tools auch bestimmte methodologische und theoretische Prämissen, die man dem Wissenschaftstheoretiker Ludwik Fleck folgend auch als Denkstil innerhalb dieses Denkkollektivs bezeichnen kann (Fleck 1980). Die Methoden der Visual Analytics sind zunächst im Lichte der Anschläge in den USA vom 11. September 2001 entwickelt worden (Thomas und Cook 2005), um in grossen Datenmengen versteckte terroristische Aktivitäten zu entdecken. Auch jetzt dienen solche Methoden, wie auch Data Mining generell, dazu, in den Daten versteckte „Schätze“ (kommerzielle Interessen) oder „Monster“ (Geheimdienste) zu finden (Bubenhofer 2016). Aus geistes- und kulturwissenschaftlicher Sicht ist die Annahme eines Schatzes oder Monsters in den Daten fragwürdig, egal ob man poststrukturalistischen, dekonstruktivistischen, systemtheoretischen oder anderen Grosstheorien folgt. Das heisst nicht, dass solche Methoden nicht auch für geistes- und kulturwissenschaftliche Forschungsinteressen eingesetzt werden sollen – im Gegenteil –, es bedeutet aber, dass für diese Disziplinen eine Bibliothek, die Google nacheifert und solche Methoden theoriefrei und rein utilitaristisch umsetzt, wenig attraktiv ist.

7 Wünsche

Im letzten Kapitel formulieren wir vor dem Hintergrund eines korpuslinguistischen Verständnisses Wünsche an die Bibliothek. Übergeordnetes Ziel wäre es, dass diese Anregungen einen Mehrwert für die ganzen textorientierten Digital Humanities schaffen, aber auch inspirierend für die Bibliotheken wirken.

7.1 Digitalisierung von Beständen

Voraussetzung jeder digitalen Verrechenbarkeit von Texten und damit oberstes Desiderat der Korpuslinguistik ist die digitale Verfügbarkeit von Volltexten. Die erste Forderung datenorientiert arbeitender humanistischer Forschung an die Bibliothek der Zukunft besteht also in der Digitalisierung der bis dato lediglich in Printform vorliegenden Bestände. Im Gegensatz zu Bemühungen, die hier bereits von zahlreichen Bibliotheken unternommen werden, steht für eine korpuslinguistisch interessierte Forschung dabei nicht ein möglichst perfekt reproduziertes digitales Abbild eines Originals im Vordergrund, sondern der digital zugängliche Text. Statistische Analyseverfahren erweisen sich gegenüber verrauschten Daten, die etwa durch automatische Texterkennungsverfahren entstehen können, mitunter als recht fehlertolerant. Den Bibliotheken kommt aber als Trägerinnen eines öffentlichen Bildungsauftrags zudem eine spezielle Rolle zu, da sie keine kommerziellen Interessen verfolgen müssen, die den Digitalisierungsbestrebungen von grossen Konzernen zu Grunde liegen.

Während Bibliotheken sich dem Open Access Gedanken auf Ebene der Daten bereits weitgehend verpflichtet fühlen, fehlt es an Konsequenz, wenn es darum geht den algorithmischen Umgang mit den Daten transparent und selbstbestimmt zu gestalten. Datengeleitete Forschung in den Humanities sieht sich deshalb immer wieder in der prekären Lage, auf Datenmaterial wie Google N-Grams angewiesen zu sein, deren wissenschaftlicher Wert zweifelhaft und deren Reproduzierbarkeit nicht gegeben ist (Koplenig 2015). Die Wissenschaft und insbesondere die Geisteswissenschaft muss in diesem Bereich Kontrolle und Unabhängigkeit in grossen Teilen erst wieder zurück gewinnen. Erschwerend kommt hinzu, dass die Humanities selbst da meist auf die technische Expertise von Informatikern und Naturwissenschaftlern angewiesen sind, wo freier Zugang zu den Algorithmen der Datenverarbeitung prinzipiell gegeben ist. Die Trennung von Entwicklern und Expertennutzern von Informationssystemen ist der meist beschrittene Weg in der humanistischen Forschung mit maschinellen Textanalysemethoden. Auf der einen Seite steht die technisch, informatisch geschulte Programmiererin, die die ihr zur Verfügung gestellten Daten aufbereitet, und auf der anderen Seite die inhaltlich versierte Humanistin, die praktisch als Kundin die zur Verfügung gestellte Ware entgegen nimmt und zu deuten vermag (Bubenhofer im Druck[a]; Bubenhofer 2016). Von keiner der beiden Seiten wird erwartet, dass sie versteht, was die andere genau macht. Insbesondere für die digitale Humanistin entsteht so die Situation, dass die algorithmische Manipulation als black box-Verfahren undurchschaubar und nur auf Umwegen beeinflussbar bleibt (Rieder und Röhle 2012). Letztlich ändert sich das Verhältnis, das sich durch die Verwendung der Daten von kommerziellen Anbietern in grösserem Massstab ergibt, nicht wesentlich. Es liegt in der Verantwortung der Forschenden, ihre digitale Unmündigkeit wenigstens soweit zu überwinden, dass sie den teils sehr verschiedenen Paradigmen, die in anderen Wissenschaften gelten, selbstbewusst und kritisch gegenüber treten können.

Ein korpuslinguistischer Zugang zu den Texten von Bibliotheken sollte einem Minimalansatz folgen: Wesentlich mehr als eine grundlegende Datenstruktur, auf deren Basis Analyseverfahren für Textdaten entwickelt werden können, muss gar nicht zu Verfügung gestellt werden. Über die spezielleren Anforderungen muss im Forschungsprozess entschieden und allenfalls müssen Datenrepräsentationen generiert werden, die für bestimmte Arten von Analysen besser geeignet sind.

Im Gegensatz zur robusten Einschlitzsuche definiert sich die Nutzerfreundlichkeit für eine Schnittstelle zur Forschung mit grossen Textmengen an Hand ganz anderer Kriterien. Da es sich um einen spezialisierten Zugriff handelt, darf der Nutzerin als Expertin mehr abverlangt und ihr die Verantwortung über die korrekte Nutzung der angebotenen Schnittstelle weitgehend übertragen werden. Der Fokus sollte mehr auf Flexibilität als auf Einfachheit in der Benutzung liegen.

Um von wissenschaftlichem Nutzen für die datengeleitete humanistische Forschung zu sein, bedarf es einer Plattform, die ganz unterschiedlichen Forschungsfragen dienen kann. Die zu Grunde liegenden Datenstrukturen müssen daher vielseitig sein. Moderne Suchindizes erfüllen sicherlich die meisten Bedürfnisse. Mit der Verbreitung von Resource Discovery Systemen haben die technischen Grundlagen für Volltextindizierung in Form von Suchmaschinentechnologie in zahlreichen Bibliotheken bereits Einzug gehalten. Für die im weit verbreiteten VuFind[2] verwendeten Open Source Technologien existieren beispielsweise Erweiterungen, die speziell entwickelt wurden, um auf den zu Grunde liegenden Indizes effizient Data Mining Algorithmen betreiben zu können[3]. Allerdings steigen die infrastrukturellen Anforderungen bei einer Ausweitung von Katalog- auf Volltextindizes um etliche Grössenordnungen, und ohne eine Zusammenarbeit mit technischen Diensten dürfte das jede Bibliothek überfordern.

Benutzerfreundlichkeit für algorithmisch betriebene Forschung lässt sich weiterhin dadurch erhöhen, dass Schnittstellen für eine breite Anzahl von Programmiersprachen zur Verfügung gestellt werden. Moderne webbasierte Programmierschnittstellen helfen den Aufwand zu verringern, indem sie weitgehend sprachunabhängig genutzt werden können.

7.2 Personalisierung des Bibliothekszugangs

Die Fragstellungen, die sich mit den in den Bibliotheken vorhandenen Textmengen für datengeleitete Analysen adressieren liessen, sind immens, aber kaum je zwei unterschiedliche Fragestellungen dürften mit der gleichen Auswahl an Textmaterial beantwortet werden können. Eine freie und dabei auf unterschiedliche Granularitäten (ganze Bücher, Kapitel, Absätze) ausgelegte Auswahl von Dokumenten zu einem Korpus ist deshalb ebenfalls Voraussetzung für eine erkenntnisversprechende Verwertung der Daten. Im Hintergrund steht dabei eine umfassende Personalisierung des Bibliothekszugangs in Form einer individuellen Indexerstellung, die teils manuell, teils automatisch, aber auch crowd-getrieben vonstatten gehen kann. Nutzerinnen sollten in die Lage versetzt werden, untereinander ihre Rechercheergebnisse zu teilen und anderen auf diesem Weg Zugriff auf spezialisierte Indizes zu erlauben. Diese Forderungen entsprechen bekannten Überlegungen zur Bibliothek 2.0 und sind in einer früheren Ausgabe dieser Zeitschrift bereits diskutiert worden (Herrlich, Ledl und Tréfás 2013).

7.3 Automatisierung und Kuratierung

Erscheint der Zugang für korpuslinguistische Forschung auf den ersten Blick vielleicht wie eine Nischenanwendung, verspricht er dennoch einen Nutzen, der über die unmittelbare Forschung weit hinaus geht. Auf Grundlage der Textdaten in Bibliotheken Verfahren zu entwickeln, die es ermöglichen, intertextuelle Beziehungen herzustellen, ohne auf explizite

[2] http://vufind-org.github.io/vufind/
[3] https://mahout.apache.org/

Verweise oder Übereinstimmung des Wortmaterials angewiesen zu sein, ermöglicht ganz neue Sichten auf die Wissensbestände, die in den Regalen der Bibliotheken aufgehoben sind. Sie sind Grundlage für Forschungsfragen, die wie oben beschrieben die Einheit von Texten transzendieren. Auf Basis solcher Forschungsergebnisse könnten alternative Indizes entwickelt werden, die den Nutzern wieder zur Verfügung gestellt werden. Für die Wissenschaft gälte es die Verbindungen zwischen den errechneten Indizes herzustellen, Sinn zu erzeugen und ihre Erkenntnis in die Arbeit von Bibliothekarinnen einfliessen zu lassen, ohne deren Kuratierung der Nutzen automatisch erzeugter Indizes mit Sicherheit beschränkt bliebe.

8 Fazit

Zentrale Forderung ist: Die Bibliotheken müssen ihre Bestände im Volltext, nicht nur deren Metadaten digital verfügbar machen, so dass mit computergestützten Verfahren darauf Forschung betrieben werden kann. Im Vergleich zu anderen (kommerziellen) Angeboten wären Bibliotheken in der Lage, über ihre Metadatenspeicher einen wissenschaftlichen Mehrwert zu diesen Rohdaten zu liefern, durch den sie sich ein Alleinstellungsmerkmal sichern und ihre Position bei der Informations- und Literaturversorgung wieder stärken könnten. Sie wären also nicht nur Bibliothek, sondern auch „Korporathek“. Dass es für Hochschulbibliotheken nicht einfach sein dürfte, Verlage davon zu überzeugen, generell Volltextzugriff zu ermöglichen, steht ausser Frage, in konzertierter Aktion könnte aber auch das erreichbar sein. Und mit dem Ausbau von Open Access Angeboten steht ein nicht unerhebliches Druckmittel bereit.

Die Bibliotheken spielen ihre Rolle dann gut aus, wenn sie möglichst viele diagrammatische Transformationen der Inhalte, die sie beherbergen, ermöglichen. Dass die digitale und verdatete Bibliothek dafür Grundvoraussetzung ist, steht ausser Frage und dürfte wenig Zweifel aufwerfen – auch wenn die Realisierung nicht trivial ist. Allerdings sind die Hochschulbibliotheken der ganzen disziplinären Breite der Wissenschaften verpflichtet, und es existieren deshalb sehr unterschiedliche theoretische und methodologische Prämissen des Umgangs mit Digitalität. Im besten Fall entwickeln sich die Bibliotheken zu einer Plattform, auf der alle diese Fragen um Daten und Deutungen kritisch verhandelt werden können.

Literatur

Berry, D. M. (2011). *The Philosophy of Software. Code and Mediation in the Digital Age*. Palgrave Macmillan. DOI: 10.1057/9780230306479.

Bonfanti, C. (2012). Roberto Busa (1913-2011), Pioneer of Computers for the Humanities. In: *Reflections on the History of Computing. Preserving Memories and Sharing Stories*. Hrsg. von A. Tatnall. Heidelberg: Springer, S. 57–61.

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Sprache und Wissen 4. Berlin, New York: de Gruyter.

— (2016). Drei Thesen zu Visualisierungspraktiken in den Digital Humanities. In: *Rechtsgeschichte – Legal History* 24, S. 351–355. DOI: 10.12946/rg24/351-355.

— (im Druck[a]). The Linguistic Construction of World – an Example of Visual Analysis and Methodological Challenges. In: *Quantifying Approaches to Discourse for Social Scientists*. Hrsg. von R. Scholz. Basingstoke: Palgrave Macmillan.

Bubenhofer, N. (im Druck[b]). Visual Linguistics: Plädoyer für ein neues Forschungsfeld. In: *Visual Linguistics*. Hrsg. von N. Bubenhofer und M. Kupietz. Heidelberg: Heidelberg University Publishing.

Bubenhofer, N., Scharloth, J. und Eugster, D. (2014). Rhizome digital: Datengeleitete Methoden für alte und neue Fragestellungen in der Diskursanalyse. In: *Zeitschrift für Diskursforschung* Sonderheft Diskurs, Interpretation, Hermeneutik.1, S. 144–172. DOI: 10.5167/uzh-111258.

Busa, R. (1951). *Sancti Thomae Aquinatis Hymnorum ritualium varia specimina concordantiarum: primo saggio di indici di parole automaticamente composti e stampi da macchine IBM a schede perforate = A 1st example of word index automatically compiled and printed by IBM punched card machines*. Archivum philosophicum Aloisianum 2. Milano: Bocca.

Bush, V. (1945). As We May Think. In: *The Atlantic*. URL: http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Feilke, H. (1996). *Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik*. Frankfurt am Main: Suhrkamp.

Feilke, H. und Linke, A., Hrsg. (2009). *Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt*. Berlin, New York: de Gruyter.

Firth, J. R. (1957). Modes of Meaning. In: *Papers in Linguistics 1934–1951*. London: Oxford University Press, S. 190–215.

Fleck, L. (1980). *Entstehung und Entwicklung einer wissenschaftlichen Tatsache: Einführung in die Lehre vom Denkstil und Denkkollektiv*. Hrsg. von L. Schäfer und T. Schnelle. 10. Aufl. Suhrkamp.

Fuller, M. (2003). *Behind the Blip: Essays on the Culture of Software*. Autonomedia.

Goffey, A. (2014). Technology, Logistics and Logic: Rethinking the Problem of Fun in Software. In: *Fun and Software: Exploring Pleasure, Paradox and Pain in Computing*. Hrsg. von O. Goriunova. New York: Bloomsbury Academic, S. 21–40.

Graham, S., Weingart, S. und Milligan, I. (2012). *Getting Started with Topic Modeling and MALLET*. URL: http://programminghistorian.org/lessons/topic-modeling-and-mallet.

Heilmann, T. A. (2012). *Textverarbeitung. Eine Mediengeschichte des Computers als Schreibmaschine*. MedienAnalysen 10. Bielefeld: Transcript.

Herrlich, B., Ledl, A. und Tréfás, D. (2013). Editorial. In: *Zeitschrift für Bibliothekskultur* 1.1, S. 1–4. DOI: 10.12685/027.7-1-1-17.

Jäger, L. (2008). Transkriptive Verhältnisse. Zur Logik intra- und intermedialer Bezugnahmen in ästhetischen Diskursen. In: *Transkription und Fassung in der Musik des 20. Jahrhunderts. Beiträge des Kolloquiums in der Akademie der Wissenschaften und der Literatur, Mainz, vom 5. bis 6. März 2004*. Hrsg. von G. Buschmeier, U. Konrad und A. Riethmüller. Stuttgart: Franz Steiner, S. 103–134.

Kay, A. und Goldberg, A. (1977). Personal Dynamic Media. In: *Computer* 10.3, S. 31–41. DOI: 10.1109/C-M.1977.217672.

Koplenig, A. (2015). The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets – Reconstructing the Composition of the German Corpus in Times of WWII. In: *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqv037.

Krämer, S. (2009). Operative Bildlichkeit. Von der ‚Grammatologie' zu einer ‚Diagrammatologie'? Reflexionen über erkennendes Sehen. In: *Logik des Bildlichen. Zur Kritik der ikonischen Vernunft*. Hrsg. von M. Heßler und D. Mersch. Metabasis 2. Bielefeld: Transcript, S. 94–123.

Krämer, S. (2013). Diagrammatisch. In: *Rheinsprung. Zeitschrift für Bildkritik* 11, S. 162–174. URL: https://rheinsprung11.unibas.ch/ausgabe-05/glossar/diagrammatisch.html.

Lauersdorf, M. R. (im Druck). Linguistic visualizations as objets d'art? In: *Visual Linguistics*. Hrsg. von N. Bubenhofer und M. Kupietz. Heidelberg: Heidelberg University Publishing.

Manovich, L. (2014). Software is the Message. In: *Journal of Visual Culture* 13.1, S. 79–81. DOI: 10.1177/1470412913509459.

Mayer-Schönberger, V. und Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

Mikolov, T., Chen, K., Corrado, G. und Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. URL: https://arxiv.org/abs/1301.3781.

Perkuhn, R. (2007). "Corpus-driven": Systematische Auswertung automatisch ermittelter sprachlicher Muster. In: *Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache*. Hrsg. von H. Kämper und L. M. Eichinger. Studien zur deutschen Sprache 40. Tübingen: Narr, S. 465–491. URL: http://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4777.

Ramelli, A. (1588). *Le diverse et artificiose machine del capitano Agostino Ramelli: nellequali si contengono varij et industriosi movimenti, degni digrandissima speculatione, per cavarne beneficio infinito in ogni sorte d'operatione: composte in lingua Italiana et Francese*. DOI: 10.3931/e-rara-8944.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. 1. Aufl. Urbana: University of Illinois Press.

Rieder, B. und Röhle, T. (2012). Digital Methods: Five Challenges. In: *Understanding Digital Humanities*. Hrsg. von D. M. Berry. Basingstoke: Palgrave Macmillan, S. 67–84.

Scharloth, J. und Bubenhofer, N. (2011). Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen*. Hrsg. von E. Felder, M. Müller und F. Vogel. Berlin, New York: de Gruyter, S. 195–230.

Siegel, S. (2009). *Tabula: Figuren der Ordnung um 1600*. Berlin, Boston: Akademie-Verlag.

Spinnler-Dürr, A. (2013). Die Diktatur der Suchmaschinen. In: *Zeitschrift für Bibliothekskultur* 1.2, S. 58–66. DOI: 10.12685/027.7-1-2-31.

Steinseifer, M. (2013). Texte sehen – Diagrammatologische Impulse für die Textlinguistik. In: *Zeitschrift für germanistische Linguistik* 41.1, S. 8–39. DOI: 10.1515/zgl-2013-0002.

Stetter, C. (2005). Bild, Diagramm, Schrift. In: *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*. Hrsg. von G. Grube, W. Kogge und S. Krämer. Kulturtechnik 4. München: Wilhelm Fink Verlag, S. 115–136.

Thomas, J. J. und Cook, K. A., Hrsg. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization und Analytics Ctr.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Studies in Corpus Linguistics 6. Amsterdam: John Benjamins Publishing Company.

