Theo Röhle

»Grand games of solitaire«. Textuelle Ordnungen in den *Digital Humanities*

Ende 2010 stellte Google den *Ngram Viewer* vor, ein experimentelles Projekt aus den unternehmenseigenen *Labs*, mit dem sich Worthäufigkeiten im gesamten Bestand der für Google Books eingescannten Bücher visualisieren lassen. Die Präsentation des Tools fand ein großes Echo in den Medien, nicht zuletzt aufgrund seiner einfachen Bedienbarkeit und der umfangreichen Datenbasis, die angeblich vier Prozent aller jemals gedruckten Bücher abdecken soll. Weniger Beachtung als das Tool selbst fand ein Artikel in der Zeitschrift *Science*, der im Zusammenhang mit der Freischaltung des *Ngram Viewers* für die Öffentlichkeit erschien. In diesem Artikel beschreibt das Entwickler-Team hinter dem Projekt, das aus insgesamt 16 Personen aus verschiedenen Fachbereichen sowie dem Google Books Team besteht, dessen Anwendungsmöglichkeiten in den Kulturwissenschaften (Michel et al. 2011).

Statt in kulturwissenschaftlicher Tradition auf Fragen des Kontexts, der Narration oder der Intertextualität einzugehen, interessieren sich die hauptsächlich naturwissenschaftlich orientierten Entwickler des *Ngram Viewers* allein für die quantitativen Aspekte der analysierten Texte. So zeigen sie zum Beispiel an einem Graphen auf, dass von 1800 bis 2000 der Name ›Freud‹ wesentlich häufiger im allgemeinen Textkorpus zu finden ist als ›Darwin‹, ›Einstein‹ oder ›Galileo‹. Zwar warnen die AutorInnen davor, aus diesen Ergebnissen vorschnelle Schlüsse zu ziehen, dennoch ist die Interpretation der quantitativen *evidence* schnell zur Hand: »›Galileo‹, ›Darwin‹, and ›Einstein‹ may be well-known scientists, but ›Freud‹ is more deeply ingrained in our collective subconscious« (ebd., 182). Auch vor der Anwendung der quantitativen Verfahren auf sehr komplexe historische Themen scheuen die AutorInnen nicht zurück: »For instance, Nazi censorship of the Jewish artist Marc Chagall is evident by comparing the frequency of ›Marc Chagall‹ in English and in German books« (ebd., 181). Der Anspruch des Projekts ist entsprechend hoch gesteckt: Es soll um die Entschlüsselung eines ›kulturellen Codes‹ gehen, der sich – analog zum Mapping von Gensequenzen – quantitativ erfassen lassen soll. In Anlehnung zum Fachgebiet der *genomics*, aus dem ein Teil der EntwicklerInnen stammt, wird daher versucht, den Begriff *culturomics* zu etablieren.

Ein solch affirmativer Bezug auf quantitative Verfahren, wie er in dem *culturomics*-Projekt vertreten wird, mag in kulturwissenschaftlichen Zusammenhängen zunächst erstaunen. Erweitert man das Blickfeld etwas, so wird jedoch deutlich, dass computergestützte Verfahren in den Geisteswissenschaften insgesamt auf dem Vormarsch sind. Unter dem Stichwort *Digital Humanities* fließen wachsende Summen an Fördergeldern in entsprechende Projekte (Duwe/Meffert 2008), an vielen Universitäten werden so genannte *Labs* gegründet, wie zum Beispiel das *Literary Lab* in Stanford, das von dem Literaturwissenschaftler Franco Moretti geleitet wird◂1 oder das *HumLab* im nordschwedischen Umeå, das im europäischen Rahmen eine wichtige Rolle spielt. In Deutschland wurden in den letzten Jahren erste Studiengänge unter der Bezeichnung *Digital Humanities* eingerichtet, an der Universität Göttingen wurde im Sommer 2011 das *Centre for Digital Humanities* eröffnet.

Es fällt auf, dass die Bandbreite der Projekte, die unter den Begriffen *Digital Humanities*, *eHumanities* oder früher auch *Humanities Computing* gefasst werden, enorm groß ist. Es geht dabei sowohl um Digitalisierung, Archivierung und Editionen-Erstellung als auch um methodologische Fragestellungen bis hin zum *E-Learning* (Schreibman 2004). Erweitert man den Blickwinkel in zeitlicher Hinsicht, so wird zudem deutlich, dass der Einsatz computergestützter Verfahren in den Geistes- und Kulturwissenschaften kein neues Phänomen ist. Gerade in Deutschland gibt es eine Reihe von Einrichtungen, die in diesem Bereich schon mehrere Jahrzehnte kontinuierlich tätig sind, dabei allerdings meist unter weniger einschlägigen Namen firmieren, wie zum Beispiel das Kölner Institut für Historisch-Kulturwissenschaftliche Informationsverarbeitung oder das Zentrum für Datenverarbeitung der Universität Tübingen. Gerade in den Philologien gab es zudem wechselnde Konjunkturen computergestützter Verfahren, die sich jedoch eher in Nischen etabliert haben, als eine komplette methodologische Umorientierung dieser Bereiche anzustoßen (vgl. Müller 2004; Meister 2005).

Im Verlauf dieser Konjunkturen wird der Datenbank immer wieder eine umwälzende Rolle für die Entwicklung der Geisteswissenschaften zugeschrieben. Während der Einsatz von Datenbanken in der Naturwissenschaft spätestens seit ihrem Wandel zur industriell geprägten *Big Science* in den 1950er Jahren zum Alltag der Forschungspraxis gehört (Bowker 2006) und der später diagnostizierte Übergang zur *Computational Science* eher eine Wende von einer statistisch-deskriptiven zu einer Simulationswissenschaft markiert (Gramelsberger 2011), scheint die Rede von einem *Computational Turn* in den Geisteswissenschaften immer wieder neu in der Lage einerseits Kontroversen zu entfachen, andererseits aber auch Hoffnungen darauf zu wecken methodologisch

neue Wege einschlagen zu können. Der *Economist* brachte diese Hoffnungen 1995 auf den Punkt: »Databases are transforming scholarship in the most conservative corners of the academy, forcing technological choices even on to the humanities« – ein Zitat, auf das der Informatiker Michael Fraser ein Jahr später mit dem ironischen Kommentar: »A similar headline could have appeared in 1974 or even 1964« verwies (Fraser 1996). Die Geschichte computergestützter Verarbeitungsverfahren in den Geisteswissenschaften scheint somit regelmäßigen Faszinationskonjunkturen unterworfen zu sein.

Wenn daher heute ein weiteres Mal die Rede davon ist, dass die Geisteswissenschaften sich zunehmend an Google orientieren sollten (Parry 2010) oder dass man insgesamt in den Kulturwissenschaften, wie der Historiker Tom Scheinfeldt meint, in ein »post-theoretical age« eingetreten sei, eine Phase, in der die Empirie in Form sehr großer Datensätze wieder im Fokus steht (vgl. Cohen 2010),◂2 scheint es umso wichtiger, diese Aussagen historisch zu kontextualisieren. Dieser Beitrag unternimmt einen Schritt in diese Richtung, indem er die wissenschafts- und technikhistorische Auseinandersetzung mit den Vorläufern der aktuellen computergestützten Methoden sucht. Eine medienwissenschaftliche Auseinandersetzung mit den *Digital Humanities* hat – trotz der verstärkten Hinwendung der Medienwissenschaft zu wissenschaftshistorischen Themen – bisher kaum stattgefunden. Dies ist insofern erstaunlich, wie die Bearbeitung methodologischer und theoretischer Fragestellungen innerhalb dieses Bereichs auf das Engste mit allgemeinen medientechnischen Entwicklungen verwoben ist.◂3 Die umfangreichsten Schnittstellen zwischen diesen Diskursen dürften sich in den Hypertext-Debatten der 1990er Jahre finden, wo dem Computer beziehungsweise der Datenbank eine entscheidende Rolle für neue Formen der nicht-linearen Textorganisation und -rezeption zugeschrieben wurde (Aarseth 2003). Kontrastiert man diese – weitgehend gescheiterten – Visionen einer neuen Unmittelbarkeit und Offenheit im Umgang mit Text mit den Zielen des aktuellen *culturomics*-Projekts, so wird deutlich, dass der Datenbank und den mit ihr verbundenen Verfahren des Sortierens, Sammelns, Suchens und Spielens sehr unterschiedliche Effekte auf die Organisation, Erschließung und Rezeption von Text zugeschrieben werden. An einem frühen Beispiel aus dem Bereich der *Digital Humanities*, der automatisierten Erstellung einer Konkordanz der Schriften von Thomas von Aquin, werde ich im Folgenden diskutieren, welche Arten der Texterschließung hier erprobt wurden und in welchem Verhältnis diese zu nachfolgenden medientechnischen Entwicklungen stehen.

Die Datenbank als »Oder-Medium«

Um diesen Fragen nachzugehen, scheint es zunächst sinnvoll, eine Arbeitsdefinition dessen zu erstellen, was eine Datenbank aus medienwissenschaftlicher Sicht ausmacht. Dafür soll an dieser Stelle weniger auf die Spezifika einzelner Technologien eingegangen werden, im Vordergrund stehen vielmehr die Prinzipien beziehungsweise die ›Logik‹ der Datenbank – das, was Manovich (1999) in seinem einschlägigen Text als »symbolische Form« kennzeichnet. Manovich baut seine Argumentation auf der von Ferdinand Saussure vorgenommenen semiotischen Unterscheidung zwischen Syntagma und Paradigma auf. Die syntagmatische Dimension entspricht einer linearen Anordnung von Elementen, zum Beispiel als Narrativ mit kausalen Zusammenhängen, die materiell manifestiert (*in praesentia*) und damit erkennbar ist. In der paradigmatischen Dimension ist die Anordnung der Elemente nicht unmittelbar erkennbar, da sie keine lineare oder zeitliche Abfolge beinhaltet und nicht materiell, sondern ausschließlich in den Köpfen der ProduzentInnen beziehungsweise RezipientInnen vorhanden ist (*in absentia*). Manovich argumentiert, dass Datenbanken dieses Verhältnis umkehren, indem sie die paradigmatische Dimension privilegieren und die syntagmatische vernachlässigen; so wenden sie auch das Verhältnis der Sichtbarkeiten um: »Database (the paradigm) is given material existence, while narrative (the syntagm) is de-materialised« (ebd., 90).
Hartmut Winkler (2003) bindet diese Argumentation in eine Diskussion der Auswirkungen von Video-on-Demand-Angeboten auf die Fernsehrezeption ein, nimmt dabei jedoch eine terminologische Justierung vor. Da sich die von Manovich diagnostizierte Umkehrung der Sichtbarkeiten durch Saussures Terminologie schlecht wiedergeben lässt, schlägt er stattdessen eine Unterscheidung zwischen Und-Medien und Oder-Medien vor:

> »Und-Medien wären Medien, die auf die syntagmatische Folge setzen, auf Anreihung, Kontinuität und Gleiten, auf räumliche Nähe ohne markierte Grenzen und auf den kontinuierlichen Fluss der Zeit. Oder-Medien wären solche, die eine Entscheidung fordern, so dass im Fortgang nur eine der gestellten Alternativen wirksam bleiben kann« (ebd., 326).

Obwohl sich Winklers Diskussion hauptsächlich im Bereich von Film und Fernsehen bewegt, ist sein Exkurs in das »Bücheruniversum« der Bibliotheken für den hier anvisierten Zusammenhang relevant. Hier existieren laut Winkler auf verschiedenen Ebenen sehr unterschiedliche Verhältnisse zwischen ›Und‹ und ›Oder‹:

»Die einzelnen Buchstaben, dies wäre die erste Ebene, gehorchen einer Logik von Auswahl und Substitution. Sie werden durch Leerräume voneinander abgetrennt; dass nur 26 Alternativen zur Wahl stehen und dass Gutenberg die Lettern mechanisch austauschbar in Blei gegossen hat, macht diesen Zug zusätzlich deutlich. Schon die zweite Ebene, die Reihung der Buchstaben in der einzelnen Zeile allerdings verfährt kontinuierlich-linear. [...] Buchseiten – die dritte Ebene – gibt es erst, seit sich der Kodex gegen die Schriftrolle durchgesetzt hat« (ebd., 327).

Das Medium Buch lässt sich nach Winkler demnach nicht eindeutig einer der beiden Dimensionen zuordnen, es verschränkt vielmehr die Ordnungssysteme auf verschiedenen Ebenen miteinander. Wie diese Ebenen genau konstituiert sind, ist jedoch nicht ein für allemal gegeben, sondern, genau wie ihr Verhältnis untereinander, einer medientechnischen Entwicklung unterworfen. Wenn also die Art und Weise, wie Texterschließung medientechnisch organisiert wird eine wichtige Rolle dafür spielt, welche Ordnung auf den verschiedenen Ebenen jeweils privilegiert wird, muss der Fokus der medienwissenschaftlichen Analyse auf den spezifischen historischen Umorganisationen dieser Arten der Texterschließung liegen.

Die Konkordanz als »Oder-Medium«

Für die Frage, wie das Verhältnis von Datenbanken und Texterschließung zu charakterisieren ist und welchen historischen Veränderungen es unterworfen ist, eignet sich ein Blick auf solche Vorhaben, in denen heutige technische Verfahren für die automatisierte Textverarbeitung noch nicht zur Verfügung standen. An der konkreten technischen Entwicklungsarbeit, die an diesen Stellen verrichtet wird, lassen sich Übergänge und Bruchstellen zwischen zwei Arten der Textorganisation und -erschließung besonders gut herausstellen. Im Fokus dieses Beitrags steht die Erstellung einer Konkordanz, ein heute kaum noch geläufiges Verfahren der Textorganisation, dessen Spezifika zunächst einer kurzen Erläuterung bedürfen.
Konkordanzen stellen vor der Einführung elektronischer Suchverfahren neben dem Register eine der wichtigsten Formen der Texterschließung dar. Etymologisch geht der Begriff auf das lateinische *concordantia* (Übereinstimmung) zurück, die Konkordanz listet – wie das Register – die graphisch übereinstimmenden Wörter eines Textes auf und enthält Verweise auf die Textstellen, an denen diese Wörter zu finden sind. Im Unterschied zum Register sind Konkordanzen jedoch nicht notwendigerweise alphabetisch angeordnet, sondern können nach beliebigen Kriterien, wie zum Beispiel Worthäufigkeiten, Wor-

tendungen oder thematischen Kategorien sortiert sein. Zudem werden in Konkordanzen die Wörter meist zusammen mit ihrem unmittelbaren Kontext, das heißt der sie umgebenden Wörter, angegeben.
Als früheste fertiggestellte Konkordanz gilt die *Concordantia breves*, die der Dominikanermönch Hugo von St. Cher in den Jahren 1230 bis 1244 unter Mitarbeit zahlreicher Mönche seines Ordens als Werkzeug für die Erschließung des lateinischen Bibeltextes erstellte. Bibelkonkordanzen etablierten sich schnell als Hilfsmittel für die exegetische Arbeit und scholastischen Disputationen, da sich mit ihrer Hilfe relevante Textstellen zu bestimmten Themenkomplexen über eine Stichwortsuche finden ließen. Zudem war es für die Erstellung einer Konkordanz unabdinglich eine Einteilung des Textes vorzunehmen, um die Referenzierung der einzelnen Wörter zu ermöglichen. Der Konkordanz von Hugo von St. Cher wird daher eine zentrale Rolle dafür zugeschrieben, dass sich die 1205 von Stephen Langton vorgenommene und bis heute verbindliche Kapiteleinteilung der *Vulgata* durchsetzte (Calwer Verlag 2001, Einleitung o. S.).

Die Automatisierung der Konkordanzerstellung

Aus einer scholastischen Dissertation, in der Konkordanzen eine entscheidende Rolle spielten, ist in den 1940er Jahren ein Projekt entstanden, das aufgrund seines Umfangs und seines Charakters vielfach als Gründungsakt der *Digital Humanities* angeführt wird (Burton 1981a; Hockey 2004). Die tragende Rolle in diesem Projekt spielte der italienische Geistliche Pater Roberto Busa, der ab 1941 an der Päpstlichen Gregorianischen Universität in Rom an einer Dissertation in Thomistischer Philosophie arbeitete. Die Fragestellung dieser Dissertation zielte auf das Konzept der Präsenz in den Schriften Thomas von Aquins ab. Bei der Konsultation der verfügbaren Konkordanzen zu diesen Werken erwies es sich jedoch als problematisch, dass in den Verzeichnissen häufig verwendete Wörter, wie zum Beispiel Konjunktionen, nicht aufgelistet wurden. Verweise für die Begriffe *praesens* und *praesentia* reichten hier nicht aus, da auch die Präposition ›in‹ in bestimmten Fällen wichtige Aufschlüsse über das Konzept der Präsenz erlaubte.
Um seine Fragestellung adäquat bearbeiten zu können, betrachtete es Busa daher als notwendig sämtliche Erwähnungen des Wortes ›in‹ in den von ihm untersuchten Texten zu analysieren. Dafür übertrug er die Sätze, in denen dieses Wort enthalten war, manuell auf Karteikarten, wodurch letztlich eine Sammlung von circa 10.000 Karten entstand. Anhand dieser Karten war es möglich, die Sätze nach verschiedenen Kriterien zu sortieren – »Grand games

of solitaire«, wie es Busa (1980) in einem späteren Rückblick beschreibt. Medientechnisch betrachtet schafft die Übertragung der Sätze auf Karten, das heißt die Loslösung aus ihrem linear-syntagmatischen Zusammenhang, somit die Voraussetzung dafür, dass in einem nächsten Schritt paradigmatische Ordnungen im wahrsten Sinne des Wortes ›durchgespielt‹ werden können. Gegenüber den existierenden Konkordanzen hatten Busas Karten zum einen den Vorteil der Vollständigkeit – zumindest was das Wort ›in‹ betraf, auf das es ihm ankam – zum anderen boten ihm die Karten wesentlich flexiblere Kombinationsmöglichkeiten als die Auflistungen in den bis dato verfügbaren Konkordanzen, die auf wenigen vorgegebenen Sortierkriterien basierten.◂4

Der Anspruch, dass die Texterschließung vollständig zu erfolgen habe, das heißt, dass sämtliche Vorkommen eines Wortes zu berücksichtigen seien und perspektivisch auch sämtliche verwendeten Wörter auf diese Weise erfasst werden sollten, war für Busa nicht nur im Hinblick auf die spezifische Fragestellung seiner Dissertation zentral. Die Möglichkeit, den Wortgebrauch eines Autors über die gesamte Breite des Textes nachzuvollziehen, betrachtete er vielmehr als Voraussetzung für die Identifizierung immanenter Widersprüche und damit als Grundlage für die Formulierung von Kritik an dessen Werk:

»In the works of every philosopher there are two philosophies: the one which he consciously intends to express and the one he actually uses to express it. The structure of each sentence implies in itself some philosophical assumptions and truths. In this light, one can legitimately criticize a philosopher only when these two philosophies are in contradiction« (ebd., 83).

Die Entwicklung von Verfahren der Texterschließung wird somit bei Busa von der Vorstellung angetrieben, eine Ebene der praktischen Formulierungsarbeit am Text analysieren zu können, auf der sich gegebenenfalls Widersprüche zur theoretisch intendierten Position des Autors finden lassen. Das von ihm entwickelte Verfahren ermöglichte dies insofern, als dass es einerseits flexible paradigmatische Anordnungen erlaubt, andererseits aber – in Form der auf den Karten verzeichneten Sätze – die syntagmatische Ordnung zum Teil intakt lässt. Die Produktivität dieser spezifischen Verschränkung von ›Und‹ und ›Oder‹ für das eigene Forschungsvorhaben veranlasste Busa ein weitaus größeres und langwierigeres Projekt, den *Index Thomisticus*, in Angriff zu nehmen, durch das er nachfolgenden Theologen ein ebenso effektives Verfahren der Texterschließung zur Verfügung stellen wollte. Im Folgenden werden die technischen Details dieses Projekts erläutert, um den Blick dafür zu schärfen, welche Anpassungsleistungen notwendig waren, um diese Art des Textzugangs zu realisieren.◂5

Der Index Thomisticus

Der *Index Thomisticus* sollte ein Verzeichnis sämtlicher Wörter umfassen, die in den Schriften Thomas von Aquins und verwandter Werke enthalten sind. Ausgehend von den Erfahrungen aus Busas erstem Projekt war es das erklärte Ziel, eine möglichst hohe Bandbreite von Sortiermöglichkeiten abzudecken. Auch wenn als Endprodukt eine gedruckte Konkordanz entstehen sollte, schien die Übertragung des Textes auf Karten als geeignetes Verfahren für die Herstellung eines solchen Verzeichnisses. Eine manuelle Erstellung dieser Karten für alle enthaltenen Wörter kam jedoch aufgrund des umfangreichen Textkorpuss' nicht in Frage. Auf der Suche nach automatisierten Verarbeitungsmöglichkeiten kam Busa im Jahr 1949 mit IBM-Gründer Thomas J. Watson Sr. in Kontakt. Mit technischer und personeller Unterstützung von IBM sowie finanzieller Unterstützung von einer Reihe italienischer Geistlicher und Industrieller wurde schließlich die Arbeit am *Index Thomisticus* 1951 aufgenommen.
Das Projekt hatte außergewöhnliche Dimensionen: Über mehrere Jahre hinweg waren über 60 Vollzeit-Arbeitskräfte damit beschäftigt, die insgesamt über 10,6 Millionen Wörter in maschinenlesbare Form zu übertragen. Erst 16 Jahre nach Beginn des Projekts war dieser erste Verarbeitungsschritt komplett abgeschlossen. Die Sortierung der Karten für die Erstellung der ersten Konkordanzen nahm weitere sechs Jahre in Anspruch. Ab 1973 wurde der *Index Thomisticus* schließlich sukzessive in insgesamt 31 Bänden veröffentlicht. Diese Art der Veröffentlichung erscheint aus heutiger Sicht widersprüchlich, werden hier doch die hinzugewonnenen (Um-)Ordnungsmöglichkeiten zugunsten einer linearen Anordnung im Buch wieder aufgegeben. Betrachtet man die insgesamt sechs veröffentlichten Konkordanzen,◂**6** so wird allerdings deutlich, dass es sich hierbei um ein Verweissystem handelt, das nur sehr bedingt einer syntagmatischen ›Und‹-Ordnung folgt, sondern vielmehr – ähnlich wie ein Bibliothekskatalog – die verschiedenen Auswahlmöglichkeiten auf der ›Oder‹-Ebene präsentiert. Insofern stellt die Konkordanz eine Proto-Datenbank dar: Sie ersetzt die linear-syntagmatische Ordnung durch eine bestimmte Auswahl paradigmatischer Ordnungskriterien. Konsequenterweise erscheint die Konkordanz 1992 zusätzlich auf CD-ROM (*Thomae Aquinatis Opera Omnia cum hypertextibus in CD-ROM*) und später auch als Online-Version.◂**7**
Der zeitliche und personelle Umfang des Projekts macht bereits deutlich, dass auf dem Weg zu einer automatisierten Verarbeitung großer Textmengen eine Reihe von Hürden zu überwinden war. Eine der Hauptaufgaben des Projekts bestand laut Busa in der Vorbereitung des zu verarbeitenden Textes. Zum einen musste ein System gefunden werden, um die Position jedes Wortes im Gesamt-

text angeben zu können, zum anderen musste eine Codierung für die Art der Wörter und Satzteile gefunden werden. Eine Besonderheit des *Index Thomisticus* bestand zudem darin, dass die enthaltenen Wörter nicht allein auf ihre graphische Übereinstimmung überprüft werden sollten, sondern dass sinnverwandte Wörter zudem bestimmten Lemmata zugeordnet werden sollten, damit der Zusammenhang zwischen diesen Wörtern aus der entsprechenden Konkordanz unmittelbar hervorgeht. Zu diesem Zweck entwickelte Busa zusammen mit zwei Mitarbeitern ein Lexikon, das sowohl eine Zuordnung der Flexionsformen eines Wortes zu ihrem Lemma als auch eine thematische Zuordnung ermöglichte.

An den Beschreibungen des Projekts wird augenfällig, welche Diskrepanzen zwischen der technischen Infrastruktur, die Anfang der 1950er Jahre zur Verfügung stand, und den Zielsetzungen des Projekts bestanden. Die von IBM zur Verfügung gestellten Rechner waren, wie es der Unternehmensname ›International Business Machines‹ andeutet, auf Geschäftsprozesse ausgerichtet. Dies hatte zur Folge, dass sowohl die verwendeten Codierungen als auch die Benutzerschnittstellen der Hardware fast ausschließlich auf die Verarbeitung numerischer Informationen ausgelegt und für die Eingabe längerer Texte kaum geeignet waren.

Als Speichermedium standen für das Projekt anfangs noch ausschließlich Lochkarten zur Verfügung. Aufgrund der Ausrichtung dieser Karten auf den geschäftlichen Bereich war weder Interpunktion vorgesehen, noch gab es Möglichkeiten, Groß- und Kleinschreibung zu markieren oder gar nicht-lateinische Alphabete einzustanzen (siehe Abbildung 1). Zur Erstellung des *Index Thomistiscus* wurden daher elaborierte Codierungsverfahren entwickelt, wobei auf Grundlage der Standardcodierung eine zweite Code-Ebene eingefügt wurde. So verminderte sich zwar die Anzahl der verfügbaren Zeichen pro Karte, dafür konnten auf dieser Ebene (zum Beispiel durch die Kombination bestimmter finanzieller Symbole) Informationen wie Satzzeichen, aber auch ›Meta-Daten‹ wie Wortposition oder Worteigenschaften gespeichert werden.

Folgt man Busas Ausführungen zum Projekt, so hatten sich die Anforderungen der Geschäftswelt jedoch nicht nur in Form der ursprünglich für seine Zwecke unpassenden Kartencodierungen und Nutzerschnittstellen in die Technik eingeschrieben. Auch die menschlichen Operateurinnen hatten, so Busa, Schwierigkeiten, sich von den etablierten Codes aus der Geschäftswelt zu lösen: »[T]hose girls first trained in business key punching are unable to grasp literary punching« (Busa 1964, 67). In einer 1954 eigens für das Projekt eingerichteten Schule wurden daher Frauen ohne Stanzerfahrung in der Verwendung der speziell entwickelten Codiersysteme ausgebildet. Laut Busa war die dabei erlern-

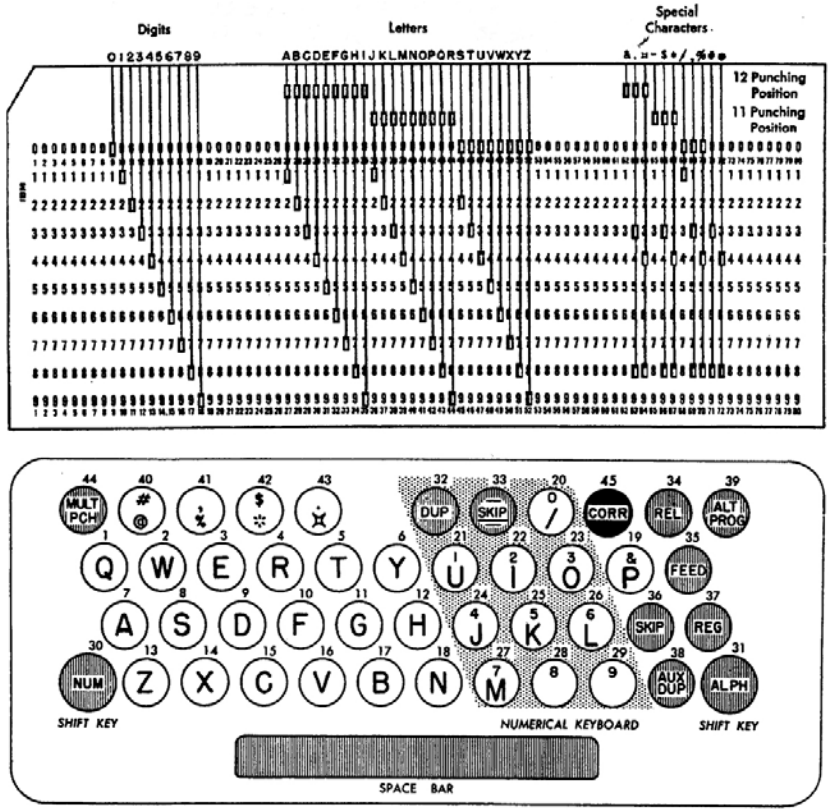

Abb. 1: Die 1949 eingeführte erste Standard-Codierung der IBM-Lochkarten und die Tastaturbelegung des IBM 026-Stanzgeräts, das diese Karten verarbeitete.

te Flexibilität im Umgang mit Codes umgekehrt in der Geschäftswelt sehr gefragt: »They quickly learn our complicated alphabetical coding and those who have left our Center and have been employed elsewhere for business punching appear to be very adept at numerical key punching« (ebd.).

Die Übertragung des Textes auf Lochkarten erfolgte zunächst Satz pro Satz. Die so entstandenen ›Satzkarten‹ wurden im nächsten Schritt von einem IBM 858 Cardatype verarbeitet, der die darauf enthaltenen Wörter automatisch erfasste und Wort für Wort auf neue Karten übertrug. Zusätzlich zum Wort selbst wurden hier Angaben zur Position des Wortes im Text und zur Art des Wortes (Zitat, Ortsangabe usw.) gespeichert. Auf die Rückseite der Karten wurde zwischen die Lochzeilen der Satz gedruckt, in dem sich das entsprechende Wort befand,

damit die Karten auch ohne Rechner verwendbar waren. Am Ende dieser Verarbeitungsprozesse stand eine Sammlung, die für jedes der 10,6 Millionen in den Texten enthaltenen Wörter eine individuelle Karte enthielt. Diese Karten konnten nun automatisiert neu kombiniert und nach bestimmten Kriterien sortiert werden, um auf dieser Grundlage Konkordanzen zu erstellen.

Die Verarbeitung umfangreicher Texte durch Computer stellte Anfang der 1950er somit keine triviale Aufgabe dar, sondern erforderte eine ganze Reihe von Adaptionen und Umbauten. Sowohl auf Seiten des textuellen Materials, auf Seiten der Technik, als auch auf Seiten der menschlichen Akteure mussten beträchtliche Anpassungsleistungen erbracht werden, um ein Zusammenspiel zu ermöglichen. Die größte Herausforderung stellte dabei die Herauslösung des Computers aus dem Bereich rein numerischer Datenverarbeitung dar. Schenkt man damaligen Vertretern des Fachs Glauben, so reichen die Konsequenzen dieser Adaptionen weit über die Grenzen des fachspezifischen »literary data processing« hinaus. Laut IBM-Mitarbeiter James A. Painter, der sich auf einer Konferenz im Jahr 1964 zu diesem Thema äußerte, spielten derartige Projekte, nicht zuletzt durch die Weiterentwicklung der Codierungsverfahren, eine zentrale Rolle dafür, dass der Computer überhaupt zu einer »general purpose«-Maschine werden konnte (Painter 1964, 169f.). Anschließend an diese Beobachtung stellt sich die Frage, woran sich entsprechende Übergänge auf andere Bereiche konkret festmachen lassen und ob sich hier weitere Veränderungen in der Weise erkennen lassen, wie Texterschließung durch Datenbanken organisiert wird.

Von der Theologie zur Philologie

Außerhalb theologischer Zusammenhänge erlangte Busas Projekt vor allem dadurch Bekanntheit, dass er es auf einer Vielzahl interdisziplinärer Tagungen präsentierte (Burton 1981a). Gleichzeitig fiel das Projekt in eine Phase, in der automatisierte Verfahren der Konkordanzerstellung im Bereich der Literaturwissenschaften insgesamt an Popularität gewannen (Burton 1981b). So wurde 1957 das Projekt der so genannten *Cornell Concordances* ins Leben gerufen, für die in den darauffolgenden Jahren die Werke von Matthew Arnold, William Butler Yeats, Emily Dickinson, William Blake und Jean Racine in maschinenlesbare Form übertragen wurden. Zudem wurden Anfang der 1960er Jahre an vielen Universitäten Zentren eröffnet, die sich der Erstellung von Konkordanzen und der rechnergestützten Analyse von Texten widmeten, zum Beispiel das einflussreiche *Centre for Literary and Linguistic Computing* in Cambridge und

eine Gruppe um Wilhelm Ott in Tübingen, deren Textanalyse-Software *TuStep* bis heute eine zentrale Rolle in diesem Bereich spielt (ebd.).

Innerhalb dieser Entwicklungen in der Literaturwissenschaft ist jedoch gleichzeitig eine Tendenz zu verzeichnen, die sich in Busas Projekt noch nicht entfaltet hatte: Der *Index Thomisticus* hatte noch größtenteils den Charakter eines Verweissystems. Die Konkordanz diente dazu bestimmte Textstellen zu finden, die dann herkömmlich linear rezipiert und interpretiert werden können. Gleichzeitig erlaubt der erste Schritt der Konkordanzerstellung – der Abgleich diskreter Wortformen auf graphische Übereinstimmung und die Herauslösung dieser Wortformen aus ihrem syntagmatischen Zusammenhang – zusätzlich die Sortierung nach quantitativen Kriterien, zum Beispiel nach Worthäufigkeiten. Während diese Sortierung im *Index Thomisticus* als *eine* Möglichkeit der paradigmatischen Ordnung neben anderen existiert, lässt sich im Bereich der Literaturwissenschaften beobachten, dass diese quantitative Dimension im Zuge der Ausbreitung von Konkordanzen immer stärker hervortritt.

Ein Grund für diese Privilegierung der quantitativen Dimension dürfte darin zu suchen sein, dass die Konkordanzerstellung in den Literaturwissenschaften auf eine Tradition stößt, in der quantitative Formen der Textanalyse schon länger gepflegt wurden. So unternahm der Physiker Thomas Corwin Mendenhall 1901 mit dem Artikel *A Mechanical Solution of Literary Problems* in der Zeitschrift *The Popular Science Monthly* den Versuch, einen Grundstein für das neue Fachgebiet der »Stylometrics« zu legen. Hierfür wurde die Länge aller in einem Text verwendeten Wörter ermittelt, um die Häufigkeit zu bestimmen mit der AutorInnen Wörter bestimmter Längen verwenden. Auf der Basis solcher Profile sollten dann Texte verglichen werden, um Kontroversen über deren VerfasserInnen zu entscheiden. Inspiriert wurde Mendenhall hierzu durch die Arbeiten des Mathematikers Augustus de Morgan, der bereits Mitte des 19. Jahrhunderts versucht hatte, anhand der Analyse von Wortlängen Hinweise auf die Autorschaft der Paulus-Briefe zu erlangen (Lord 1958; Hockey 2004, 5). Im Unterschied zu de Morgans griff Mendenhall jedoch bei seinen Untersuchungen bereits auf mechanische Hilfsmittel zurück: Statt Strichlisten zu führen, bedienten zwei Frauen, die mit der Ermittlung der Wortlängen beauftragt waren, eine eigens entwickelte »counting machine« (Mendenhall 1901, 102). Die »Stilometrie« erlebte in den 1950er und 1960er Jahren im Zuge der vermehrt verfügbaren Konkordanzen einen neuen Aufschwung. Diese erlaubten es nicht nur die Häufigkeit von Wörtern bestimmter Länge, sondern nun auch die Häufigkeit bestimmter Wörter automatisiert zu ermitteln. Die Verwendung spezifischer Vokabulare sollte auf diese Weise quantitativ erfasst und bestimmten Autoren zugeordnet werden.◂**8**

Es lässt sich somit nachvollziehen, dass eine Übertragung medientechnischer Verfahren aus dem theologischen Bereich, in dem Busa mit seinem Projekt angesiedelt war, zur allgemeinen Analyse von literarischen Texten in der Philologie stattgefunden hat. Gleichzeitig findet sich hier jedoch eine Tendenz zur Privilegierung der quantitativen Dimension, die in Busas als Verweissystem konzipierten Konkordanzen noch nicht angelegt war. Man kann daher fragen, ob sich vergleichbare Übertragungen und Tendenzen auch in der allgemeinen medialen Praxis wiederfinden. Konkret lässt sich eine solche Übertragung an den Arbeiten eines weiteren IBM-Mitarbeiters, Hans-Peter Luhn, festmachen, der Ende der 1950er Jahre Vorarbeiten für die Entwicklung späterer Volltext-Suchverfahren leistete.◂9

Von der Konkordanz zur Volltextsuche

Konzeptueller Ausgangspunkt für Luhn ist, dass zwischen der Häufigkeit mit der Wörter in einem Text verwendet werden, und der Rolle dieser Wörter im Rahmen einer wissenschaftlichen Argumentation ein Zusammenhang besteht. Die Ermittlung von Worthäufigkeiten eignet sich aus seiner Sicht daher nicht nur für die Bestimmung des Vokabulars eines Autors, sondern die Frequenz eines Wortes lässt sich als Indikator dafür interpretieren, wie wichtig dieses Wort im Kontext des entsprechenden Textes ist. Auch bei der Erstellung eines Lexikons können Worthäufigkeiten als Kriterium dafür dienen, ob die Lemmata adäquat gewählt wurden und der fachspezifischen Sprache gerecht werden, die im Text verwendet wird.

Quantitäten kommen bei Luhn somit, ähnlich wie bei den literaturwissenschaftlichen Ansätzen der »Stilometrie«, eine eigene Bedeutung zu; sie dienen nicht, wie noch bei Busa, lediglich als Sortierkriterium. Während Luhn anfangs noch eine Kombination aus manuell vorgenommenen Klassifikationen und automatisierten quantitativen Analysen für die Entwicklung von Text-Suchverfahren befürwortet (Luhn 1957), kommen spätere Ansätze komplett ohne menschliche Beteiligung aus. So beschreibt Luhn (1958) die Entwicklung eines Programms, das der automatischen Erstellung von *Abstracts* wissenschaftlicher Artikel dient. In den *Abstract* aufgenommen werden die Sätze, die eine Kombination verschiedener hochfrequenter Wörter in geringer Distanz zueinander enthalten. Denn, so Luhn, »wherever the greatest number of frequently occurring different words are found in greatest physical proximity to each other, the probability is very high that the information being conveyed is most representative of the article« (ebd., 160f.).

Der Zusammenhang zwischen Worthäufigkeiten und Relevanz, den Luhn in seinen Arbeiten systematisiert, findet schließlich in Form des Maßes *Term Frequency* Eingang in die textstatistischen Verfahren, die im *Information Retrieval* weiterentwickelt und später in die Online-Suche übernommen werden (Stock 2006, 321ff.).◂10 Entscheidend ist, dass hier zwei Arten des Zugangs zusammenfallen: Einerseits werden die vorhandenen Texte – weiterhin dem Prinzip der *concordantia* folgend – nach einer Übereinstimmung zwischen Suchbegriff und in den Texten enthaltenen Wörtern durchsucht.◂11 Dieser Abgleich allein produziert bei einem großen Textkorpus allerdings eine sehr lange Liste gefundener Dokumente. Es bedarf daher eines automatisiert bestimmbaren Sortierkriteriums, anhand dessen eine solche Liste nach Relevanz geordnet werden kann. Ein solches Kriterium stellt die *Term Frequency* dar,◂12 deren Vorgänger in der Analyse von Worthäufigkeiten bei Luhn zu finden sind.

Schluss

Winkler spricht in seinem oben erwähnten Aufsatz von einem Umschlagspunkt, den die Organisation audiovisueller Elemente in Form der Datenbank mit sich bringt:

> »Je komplexer, umfangreicher und leistungsfähiger diese Datenbank wird, desto mehr werden die einzelnen audiovisuellen Materialien, um deren Erschließung es geht, zu einem Anhängsel dieser Makro-Struktur werden. Der Modus des Zugriffs wird auch hier die Kohärenz des linearen Syntagmas antasten« (Winkler 2003, 329).

Im Gegensatz zum audiovisuellen Bereich, in dem diese Entwicklung bisher nur absehbar ist, beschreibt er den Umschlag im Textuniversum als bereits erfolgt: »Je mehr wir nicht mehr mit Texten, sondern nun mit Textstellen arbeiten und je leichter Suchmaschinen uns den Zugriff machen, desto mehr gewinnt der Zugriff selbst Gewicht, und zwar gegen die Linearität der Zeile« (ebd., 328).
Die Übertragung von automatisierten Verfahren der Textverarbeitung aus dem spezialisierten Kontext der theologischen und philologischen Projekte in den Bereich des *Information Retrieval*, dem die Online-Suche als eine zentrale Komponente heutiger medialer Praktiken entstammt, scheint den von Winkler diagnostizierten Umschlag von ›Und‹ zu ›Oder‹ zu bestätigen. Das Projekt des *Index Thomisticus*, so außergewöhnlich es sich in seiner spezifischen Ausformung darstellt, war kein isoliertes Phänomen, sondern trug durch einen langen Prozess von Rekonfigurationen dazu bei, die medientechnischen Vo-

raussetzungen für diesen späteren Umschlag zu schaffen. Die Übertragung des Textes auf maschinenlesbare Karten bricht die syntagmatische Ordnung auf und bildet, in Kombination mit der Einkodierung umfangreicher ›Meta-Daten‹ in diese Karten, die Voraussetzung für Zugänge, die auf unterschiedlichen Sortierkriterien basieren. Im weiteren Verlauf – dies ist die zweite wichtige Beobachtung – steht allerdings nicht mehr die Ausschöpfung dieser Bandbreite von Sortierverfahren im Vordergrund, vielmehr setzen sich in der Literaturwissenschaft und im *Information Retrieval* vornehmlich quantitativ orientierte Sortierkriterien durch.

Diese Durchsetzung bestimmter Arten von Sortierkriterien legt nahe, dass man die Ordnungen des ›Und‹ und des ›Oder‹ nicht so dichotom gegenüberstellen kann, wie es bisher erfolgt ist. Statt einen allgemeinen Umschlag von ›Und‹ nach ›Oder‹ zu diagnostizieren, scheint es nötig, auf der ›Oder‹-Seite der Privilegierung *spezifischer* paradigmatischer Ordnungen nachzugehen. Konkret auf die hier beschriebenen Entwicklungen bezogen kann man somit zwar konstatieren, dass durch die Volltextsuche die linear-syntagmatische Ordnung von Texten ins Wanken gerät, hat damit aber über die Art der Auswahlmöglichkeiten auf der paradigmatischen Ebene noch keine Aussage getroffen. Akzeptiert man nun die These, dass sich in den skizzierten Entwicklungen eine Reduktion auf die quantitative Dimension abzeichnet, so stellt sich im nächsten Schritt die Frage, woher diese Reduktion rührt. Lässt sich die zunehmende Privilegierung der quantitativen Ebene im Umgang mit Text auf bestimmte medientechnische Charakteristika zurückführen? Schreibt sich die numerische Herkunft des Computers letztlich – trotz aller Anpassungsleistungen aus geisteswissenschaftlicher Richtung – in die wissenschaftlichen und medialen Praxen ein?

Das *culturomics*-Projekt, das den Ausgangspunkt dieses Beitrags bildete, lässt sich sicherlich in dieser Hinsicht interpretieren. Es wäre dann als vorläufiger Höhepunkt einer Entwicklung zu betrachten, in dem die quantitative Dimension beim Umgang mit digitalen Texten *notwendigerweise* immer mehr an Gewicht gewinnt. Stützen lässt sich eine solche Lesart durch programmatische Aussagen von Vertretern der *Digital Humanities*, die auf Aspekte wie Exaktheit◂**13** und Restlosigkeit◂**14** fokussieren und damit an quantitativ orientierte Rationalitätsvorstellungen anknüpfen. Die Datenbank wäre in diesem Zusammenhang allererst ein wissenschaftliches Werkzeug der Effizienz und des distanzierten Überblicks, das – dann tatsächlich analog zu den *genomics* – einen vermeintlich ›objektiven‹ Blick auf das Material erlaubt.

Es gibt allerdings auch völlig andere Stimmen aus den *Digital Humanities*. So vertritt beispielsweise Steven Ramsay (2010) einen »screwmeneutical impera-

tive«, durch den vielfältige Zugänge zum Material geschaffen werden sollen. Der Datenbank fällt hier nicht nur die Rolle zu, das lineare Syntagma zu durchbrechen, sondern auch auf der paradigmatischen Ebene immer wieder neue Auswahlmöglichkeiten ins Spiel zu bringen. In eine ähnliche Richtung argumentiert auch Willard McCarty (2009) in seiner Antrittsrede als Professor für *Humanities Computing* am Londoner King's College:

»Not push a button and wait for the answer; not follow links; not work within a system of tags established canonically for us by an expert, or a committee, or a consensus of the great and the good; but on the spur of the moment, try things out and see what happens, try things out and model our way experimentally toward a better knowing«.

Im Hinblick auf die anfangs gestellte Frage, wieso die Kombination aus geisteswissenschaftlichen Fragestellungen und computergestützter Verfahren in regelmäßigen Abständen in der Lage ist, Kontroversen zu schüren, Hoffnungen zu wecken und Faszinationspotentiale zu aktualisieren, ist es interessant zu vermerken, dass sich beide Arten von Visionen – man könnte sie als ›objektive Übersicht‹ und ›kreative Rekombination‹ bezeichnen – in völlig unterschiedlichen Phasen der *Digital Humanities* finden lassen.◂15 Das Gebiet als Ganzes scheint daher keiner übergreifenden Entwicklungslogik – von offen zu geschlossen oder andersherum – zu folgen, stattdessen werden die Grenzen von Formalisierung und Quantifizierung immer wieder zum Gegenstand expliziter Auseinandersetzungen gemacht. Gerade im Aufeinandertreffen von Computern und geisteswissenschaftlichen Fragestellungen scheinen Ordnungssysteme ihre Bruchstellen zu offenbaren. Dies erhöht die Chance, dass die methodologische Privilegierung bestimmter Dimensionen des Untersuchungsmaterials, wie zum Beispiel die quantitativen Aspekte von Texten, nicht unter der Wahrnehmungsschwelle reproduziert wird, sondern in fachspezifischen Debatten explizit thematisiert wird, was der Aushandlung neuer Zugänge zum Material Vorschub leisten könnte. In diesem Sinne kann man die interpretativen Schwächen des quantitativ orientierten *culturomics*-Projekts durchaus als positives Signal werten, lassen sie doch aktive Gegenbewegungen gegen dessen überdimensionierte Erklärungsansprüche erwarten.

Anmerkungen

01► Franco Moretti fordert schon seit geraumer Zeit ein so genanntes »Distant Reading« ein, bei dem es nicht mehr um einen hermeneutischen Zugang zu einzelnen Texten gehen soll, sondern um die quantitative Auswertung großer Textmengen. Ziel ist in diesem Fall zum Beispiel die Analyse von Genreverschiebungen oder auch ein historischer Vergleich des

Produktionsvolumens von Büchern in verschiedenen Ländern (Moretti 2009).

02► Unter dem Stichwort »post-theoretical age« hatte sich 2009 auch schon Brian Eno im britischen *Prospect Magazine* in ähnlicher Weise zu Wort gemeldet: »In the absence of data, you theorise. In an abundance, you just need to do the maths. And, because of all those super-efficient search engines, we share more and more data. Data dissolves ideology« (Eno 2009).

03► Beispielsweise spielte die aus den *Digital Humanities* hervorgegangene Text *Encoding Inititiative*, die Markupverfahren für digitale Texte entwickelt, eine zentrale Rolle bei den Vorarbeiten zur Standardisierung von XML (deRose 1999).

04► Diesen Vorzug der Karte beziehungsweise des Zettels gegenüber dem Buch stellt Krajewski (2002) ausführlich am Beispiel der Bibliothekskataloge dar. Auf einer allgemeinen Ebene streicht Latour (2009) in der Erläuterung seines Konzepts der »immutable mobiles« die Eigenschaft der Mobilität von Papieraufzeichnungen heraus und zeigt am Beispiel von Mendelejews Periodensystem auf, inwiefern diese als Grundlage experimenteller paradigmatischer Ordnungen betrachtet werden können. Auch Latour zieht hier den Vergleich zum Patience-Spiel: »Jedes Element befindet sich nun auf einer neuen Papierform auf dem Schnittpunkt eines Längen- und Breitengrads; diejenigen, die sich auf der gleichen horizontalen Linie befinden, sind durch ihr Atomgewicht nahestehend, obwohl sie durch ihre chemischen Eigenschaften unterschieden sind; diejenigen, die sich auf derselben vertikalen Linie befinden, sind durch ihre Eigenschaften ähnlich, obwohl sie sich in ihrem Atomgewicht mehr und mehr voneinander entfernen. Auf diese Weise wurde lokal ein neuer Raum geschaffen; neue Verbindungen von Distanz und Nähe, neue Nachbarschaften, neue Familien wurden entwickelt: Eine Periodizität (daher der Name der Tabelle) wird sichtbar, die bis dahin im Chaos der Chemie unsichtbar war« (ebd. 141f.).

05► Die Rekonstruktion stützt sich hautpsächlich auf Busa (1964; 1980), Winter (1999) und Hockey (2004).

06► Diese umfassten 1. Eine alphabetische Liste aller Wörter mit Angabe der Häufigkeiten, 2. Ein Register der Lemmata, 3. Ein umgekehrtes Register der Lemmata mit Angabe der Häufigkeiten, 4. Eine Liste der Wortformen inkl. Häufigkeiten, nach Lemmata sortiert, 5. Ein Ortsregister und 6. Eine Konkordanz, in der jedes Wort der Hymnen im Kontext, das heißt in der jeweiligen Verszeile, angegeben ist.

07► Abrufbar unter [http://www.corpusthomisticum.org/it/]; letzter Abruf: 19.12.2011.

08► Eine interessante Randnotiz aus medienwissenschaftlicher Sicht ist ein Plädoyer für den Einsatz statistischer Verfahren in der Stilforschung von Norbert Bolz (1984). Im engeren Bereich des *Humanities Computing* scheinen Konkordanzen ab Mitte der 1960er etabliert, der Zenit der Faszination für diese Umordnung textuellen Materials allerdings auch schon überschritten. So liefert die erste Ausgabe der Zeitschrift »Computers and the Humanities« 1966 ihren LeserInnen zwar noch eine Liste aktueller Programme zur Erstellung von Konkordanzen. In der Ausschreibung eines Preises für innovative Projekte wird jedoch aus-

drücklich darauf hingewiesen, dass der Bedarf an Konkordanz-Programmen gesättigt ist und diese daher keine Chance auf eine Auszeichnung haben (Lieb 1966).

09► Luhn (1957, 314) bezieht sich explizit auf die Arbeiten am *Index Thomisticus*, wie auch Busa (1990, 341) sich umgekehrt auf Luhn bezieht.

10► In aktuellen kulturwissenschaftlichen Beiträgen zu Ranking-Verfahren in der Online-Suche entsteht mithin der Eindruck, das Google-Ranking würde ausschließlich auf *PageRank* basieren, das heißt es würde nur die Analyse der ein- und ausgehenden Links einer Seite für die Relevanzbewertung herangezogen. Ein solcher Fokus auf die Linktopologie trägt zwar dazu bei, dass die Vorläufer von *PageRank* in der Sozio- und Bibliometrie inzwischen besser beleuchtet werden (Donner 2010, Mayer 2009). Allerdings droht dabei in Vergessenheit zu geraten, dass auch *PageRank* nur ein Faktor unter vielen Rankingkriterien ist, wobei die Textstatistik weiterhin eine wichtige Rolle einnimmt. In ihrer ursprünglichen Beschreibung von Google schreiben Brin/Page zum Beispiel ausdrücklich, dass die Häufigkeit von Wortnennungen zusammen mit Informationen zur jeweiligen Textformatierung in einen so genannten »IR score« einfließt, der anschließend mit *PageRank* kombiniert wird (Brin/Page 1998).

11► Seit 1972 übernimmt diese Aufgabe standardmäßig die Software ›grep‹, auf deren Rolle Duguid (2009) ausführlicher eingeht.

12► Beziehungsweise später darauf aufbauende relative Häufigkeitsmaße wie *Within Document Frequency* oder *Inverse Document Frequency* (Stock 2006, 321ff.).

13► »In the early literature of humanistic computing, one finds great hope that the computer would make the information it yielded more reliable than that gathered directly by human effort. There was a notion that what computers could not do, editors should not do, lest they detract from that objectivity or fail to take full advantage of the computer's capabilities« (Burton 1982, 195).

14► »Scientists talk about Big Science. I am proposing a Big Humanities. I would venture to say that digitizing (with interoperability and universal access) the entire record of human expression and accomplishment would be as significant and as technologically challenging an accomplishment of the information age as sequencing the human genome or labeling every visible celestial object« (Davidson 2008, 714).

15► So schreibt Irwin Lieb bereits in der ersten Ausgabe der Zeitschrift »Humanities Computing« 1966: »In the first stage, scholars of pythagorean mood who knew what mathematicians, logicians and linguists were doing with computers, thought and asked how some of their techniques could be applied to the materials with which they were themselves most concerned to work. The computer replaced their card files and the ingenuities of indexing which they had devised with colored tabs and codes. The techniques of counting, sorting, storing and recovering were practiced, extended, and refined. Now, though, at the start of what may be a second stage, we are trying to set aside the image of the file (as well as some of the calculator images) and trying to imagine computers on different models, we

are not sure what – the puzzle, the trip, module constructions, a dozen other schemes« (Lieb 1966, 9).

Literatur

Aarseth, Espen J. (2003) Nonlinearity and Literary Theory. In: *The New Media Reader.* Hrsg. v. Nick Montfort & Noah Wardrip-Fruin. Cambridge, MA: MIT Press. S, 761-780.

Bolz, Norbert (1984) Gewinnung und Auswertung quantitativer Merkmale in der statistischen Stilforschung. In: Methoden der Stilanalyse. Hrsg. v. Bernd Spillner. Tübingen: Gunter Narr Verlag, S. 193-222.

Bowker, Geoffrey C. (2006) Memory practices in the sciences. Cambridge, MA: MIT Press.

Brin, Sergey / Page, Lawrence (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine [http://infoab.stanford.edu/~backrub/google.html]; letzter Abruf: 10.7.2011.

Burton, Dolores M. (1981a) Automated Concordances and Word-Indexes: The Fifties. In: *Computers and the Humanities* 15, 1, S. 1-14.

Burton, Dolores M. (1981b) Automated Concordances and Word-Indexes: The Early Sixties and The Early Centers. In: Computers and the Humanities 15, 2, S. 83-100.

Burton, Dolores M. (1982) Automated Concordances and Word-Indexes: Machine Decisions and Editorial Revisions. In: *Computers and the Humanities* 16, 4, S. 195-218.

Busa, Roberto (1964) An Inventory of Fifteen Million Words. In: *Literary data processing. Conference proceedings.* Hrsg. v. Jess B. Bessinger, Stephen M. Parrish & Harry F. Arader. White Plains, N.Y: IBM, S. 64-79.

Busa, Roberto (1980) The Annals of Humanities Computing: The Index Thomisticus. In: *Computers and the Humanities* 14, 2, S. 83-90.

Busa, Roberto (1990) Informatics and new philology. In: *Computers and the Humanities* 24, 5/6, S. 339-343.

Calwer Verlag (2001) Große Konkordanz zur Lutherbibel. Stuttgart: Calwer Verlag.

Cohen, Patricia (2010) Digital Keys for Unlocking the Humanities' Riches. In: *Times.* 17.11.2010 [http://www.nytimes.com/2010/11/17/arts/17digital.html]; letzter Abruf: 4.01.2012.

Davidson, Cathy (2008) Humanities 2.0: Promise, Perils, Predictions. In: *PMLA* 123, 3, S.707-717.

DeRose, Steven (1999) XML and the TEI. In: Computers and the Humanities 33, 1/2, S. 11-30.

Duguid, Paul (2009) Die Suche vor grep. Eine Entwicklung von Geschlossenheit zu Offenheit? In: Deep Search. Die Politik des Suchens jenseits von Google. Hrsg. v. Felix Stalder & Konrad Becker. Innsbruck: Studienverlag, S. 15-36.

Duwe, Janina / Meffert, Katja (2008) State-of-the-Art Analyse E-Humanities [http://www.textgrid.de/fileadmin/TextGrid/konferenzen_vortraege/eHumanities_Junio8/SotAA_1.1.pdf]; letzter Abruf: 10.7.2011.

Donner, Martin (2009) Rekursion und Wissen. Zur Emergenz technosozialer Netze. In: Rekursionen. Hrsg. v. Philipp von Hilgers & Ana Ofak. München: Fink, S. 77-115.

Eno, Brian (2010) The post-theoretical age. In: Prospect Magazine. [http://www.prospectmagazine.co.uk/2009/10/the-post-theoretical-age/]; letzter Abruf: 04.01.2012.

Fraser, Michael (1996) A Hypertextual History of Humanities Computing: Convergence and Collaboration. [http://users.ox.ac.uk/~ctitext2/history/converge.html]; letzter Abruf: 04.01.2012).

Gramelsberger, Gabriele (2011) From Science to Computational Sciences. A Science History and Philosophy Overview. In: From Science to Computational Sciences. Studies in the History of Computing and its Influence on Today's Sciences. Hrsg. v. Gabriele Gramlesberger. Zürich/Berlin: Diaphanes, S.19-44.

Hockey, Susan (2004) The History of Humanities Computing. In: A companion to digital humanities. Hrsg. v. Susan Schreibman, Ray Siemens & John Unsworth. Malden, MA: Blackwell, S. 3-19.

Krajewski, Markus (2002) Zettelwirtschaft. Die Geburt der Kartei aus dem Geiste der Bibliothek. Berlin: Kadmos.

Latour, Bruno (2009) Die Logistik der immutable mobiles. In: Mediengeographie. Theorie-Analyse-Diskussion. Hrsg. v. Jörg Döring & Tristan Thielmann. Bielefeld: Transcript, S. 67-110.

Lieb, Irwin (1966) The ACLS Program for Computer Studies in the Humanities: Notes on Computers and the Humanities. In: Computers and the Humanities 1, 1, S. 7-11.

Lord, R.D. (1958) Studies in the History of Probability and Statistics: VIII. De Morgan and the Statistical Study of Literary Style. In: Biometrika 45, 1/2, S. 282.

Luhn, Hans-Peter (1957) A statistical approach to mechanized encoding and searching of literary information. In: IBM Journal of Research and Development 1, 4, S. 309-317.

Luhn, Hans-Peter (1958) The automatic creation of literature abstracts. In: IBM Journal of Research and Development 2, 2, S. 159-165.

Manovich, Lev (1999) Database as Symbolic Form. In: Convergence. The International Journal of Research into New Media Technologies 5, 2, S. 80-99.

Mayer, Katja (2009) Zur Soziometrik der Suchmaschinen. In: Deep Search. Politik des Suchens jenseits von Google. Hrsg. v. Felix Stalder & Konrad Becker. Innsbruck: Studienverlag, S. 64-83.

McCarty, Willard (2009) Attending from and to the machine [http://staff.cch.kcl.ac.uk/~wmccarty/essays/McCarty,%20Inaugural.pdf]; letzter Abruf: 04.01.2012.

Meister, Jan Christoph (2005) Projekt Computerphilologie. Über Geschichte, Verfahren und Theorie rechnergestützter Literaturwissenschaft. In: Digitalität und Literalität. Zur Zukunft der Literatur. Hrsg. v. Harro Segeberg & Simone Winko. München: Fink, S. 315-341.

Mendenhall, Thomas C. (1901) A Mechanical Solution of a Literary Problem. In: The Popular Science Monthly 60, 7, S. 97-105.

Michel, Jean-Baptiste et al. (2011) Quantitative Analysis of Culture Using Millions of Digitized Books. In: Science 331, 6014, S. 176-182.

Moretti, Franco (2009) Kurven, Karten, Stammbäume. Abstrakte Modelle für die Literaturgeschichte. Frankfurt/Main: Suhrkamp.

Müller, Oliver (2004) Messbare Dichtung? Eine Feldstudie zur exakten Literaturwissenschaft in den 1960er Jahren. In: Soziale Räume und kulturelle Praktiken. Über den strategischen Gebrauch von Medien. Hrsg. v. Georg Mein & Markus Rieger-Ladich. Bielefeld: Transcript, S. 149-180.

Painter, James A. (1964) Implications of the Cornell Concordances for Computing. In: Literary data processing. Conference proceedings. Hrsg. v. Jess B. Bessinger, Stephen M. Parrish & Harry F. Arader. White Plains, N.Y: IBM, S. 160-170.

Parry, Marc (2010) The Humanities Go Google In: The Chronicle of Higher Education. [http://chronicle.com/article/The-Humanities-Go-Google/65713/]; letzter Abruf:04.01.2012.

Ramsay, Steven (2010) The Hermeneutics of Screwing Around; or What You Do with a Million Books [http://www.playingwithhistory.com/wp-content/uploads/2010/04/hermeneutics.pdf]; letzter Abruf:04.01.2012.

Schreibman, Susan / Siemens, Ray / Unsworth, John (Hrsg.) (2004) A companion to digital humanities. Malden, MA: Blackwell.

Stock, Wolfgang G. (2006) Information Retrieval. Informationen suchen und finden. München: Oldenbourg Wissenschaftsverlag.

Winkler, Hartmut (2003) Video on Demand. Zugriff auf bewegte Bilder. In: Medien und Ästhetik. Festschrift für Burkhardt Lindner. Hrsg. v. Harald Hillgärtner & Thomas Küpper. Bielefeld: Transcript, S. 318-331.

Winter, Thomas N. (1999) Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance. In: The Classical Bulletin 75, 1, S. 3-20.