Das kleine Digitale: Ein Plädoyer für Kleinkorpora und gegen Grossprojekte wie Googles Ngram-Viewer

Hodel, Tobias

Preprint: Nach Feierabend. Zürcher Jahrbuch für Wissensgeschichte 9 (2013), S. 103-119.

Das kleine Digitale

Ein Plädoyer für Kleinkorpora und gegen Großprojekte wie den Google nGram-Viewer

Tobias Hodel

Wie schön wäre es doch, wenn sich alles Wissen mit wenigen Tastaturanschlägen durchforsten ließe; eine Maschine spuckt alle relevanten Materialien zum nachgefragten Thema aus. Das (lineare) Lesen von Büchern würde der Vergangenheit angehören, künftig wäre das Abmühen mit schwerfällig aufgebauten Kapiteln und unleserlichen Satzkonstruktionen überflüssig. Reine Daten lieferten Erkenntnisse, wenn möglich in Form von Diagrammen und Graphen, die eine Interpretation sofort bildlich sichtbar machen würden.[1]

Erreichen die Geisteswissenschaften eine Form der Industrialisierung, wie sie Hegel in seiner Realphilosophie für die englische Manufaktur beschrieb, und erniedrigt eine solche den Forschenden, wie sie bereits die Fabrikarbeiter im 18. Jahrhundert zu Zuarbeitern, zu Rädchen im Getriebe degradierte?[2] Mahner, die das Ende der Geisteswissenschaften aufgrund von rechnergestützter Informationsbeschaffung und Auswertung sehen, sind leicht zu finden.[3] Das weltweite Netz im Allgemeinen und der Suchmaschinengigant Google im Speziellen gaukeln ihren Nutzern vor, umfassende Erkenntnis – alles Wissen der Welt – auf den Bildschirm zu zaubern. Ob tausendjährige Manuskripte, hundertjährige Zeitschriften oder eben erst erschienene Bestseller: Die Maschine spuckt unter Zuhilfenahme der richtigen Suchbegriffe das *Wissen der Menschheit* jederzeit und überall aus.[4]

Der Anspruch von Google erschöpft sich jedoch nicht in der Suche nach im Netz verfügbaren Ressourcen. Seit 2004 verfolgt die erfolgreichste Suchmaschine des 21. Jahrhunderts das Ziel, das Buch ins Internet zu bringen, »die Informationen der Welt zu organisieren«.[5] Aufbauend auf dem Wortbestand der von Google gescannten Bücher kam dann 2010 die Idee auf, Wort- und Phrasenhäufigkeiten zu bestimmen. Mittels eines einfachen Interfaces ist es seither möglich, ausgewählte Bestände in Sekundenschnelle zu durchsuchen und die Suchergebnisse grafisch darzustellen. Dahinter liegt die Überlegung, dass es anhand quantitativer Auszählung von unglaublich vielen (momentan circa sechs Prozent der jemals produzierten) Büchern möglich sein müsste, Aussagen über das Wissen einer Zeit, den diachronen Wandel im Wortgebrauch oder etwa auch über Moden im Phrasengebrauch zu machen.

Wer den Ausführungen der beteiligten Wissenschaftler folgt,[6] könnte zu dem Schluss kommen, dass den Geisteswissenschaften ein Wandel bevorsteht, wie er nach dem *cultural turn* Anfang der 1990er Jahre nicht mehr für möglich gehalten worden war: Ist die umfassende »Quantifizierung von Kultur« etwa bereits im Gange? Im Gegensatz zu den Hochzeiten der Sozialgeschichte zwischen 1960 und 1970, die mit der Auszählung von Daten jeglicher Art eine objektive Geschichte von gesellschaftlichen Schichten und Gruppen zu schreiben versuchte, sind vierzig Jahre und einige kulturwissenschaftliche Grabenkämpfe später nicht zuletzt die Wissenshistoriker am Zug, die versuchen, mittels einer gigantischen Menge von Text Diskurse und Diskursbrüche (im Sinne Foucaults) festzumachen: »*nGram-Viewer*« nennt sich das Kind von Google, das Wissenschaftsgeschichte schreiben und »*cultural treasures*« sichtbar machen soll.[7] Das Scanning- und Auswertungsunternehmen passt hervorragend ins Unternehmensprofil: Bei der herkömmlichen Google-Suchabfrage werden innert kürzester Zeit die für den betreffenden Nutzer wichtigsten Links auf eine Abfrage generiert (und dabei möglichst viele Klicks auf den Anzeigenteil stimuliert). Der nGram-Viewer vergleicht ebenfalls in Sekundenschnelle Begriffe und Phrasen und wertet diese als »Daten« aus.[8] Ziel ist die *effiziente* Auswertung des bislang (in Buchform) produzierten Wissens.[9] Die vermeintliche Evidenz von Kurven und Zahlen verleiht dem nGram-Viewer einen Anstrich von Wissenschaftlichkeit (im Sinne von *scientific*) und erhöht den Reiz des Werkzeugs.[10] Gepaart mit der aktuellen Konjunktur der »*Digital Humanities*« lebt die Idee wieder auf, dass geisteswissenschaftliche Untersuchungsgebiete mit »Zahlen« unterfüttert werden müssen.

Nicht nur das Privatunternehmen Google beschäftigt sich intensiv mit der Digitalisierung von Büchern aller Art, auch auf Ebene der EU, einzelner Staaten und Institutionen werden Ressourcen elektronisch zur Verfügung gestellt. Projekte wie die *Europeana* (eigentlich eine Metasuchmaschine über nationale Sammlungen) »stellen Sammlungen zusammen, die es erlauben die Europäische Geschichte von der Antike bis in die Gegenwart zu entdecken.«[11] Die Ursprünge der *Europeana* und des französischen Digitalisierungsprojekts *Gallica* liegen im Widerstand, der sich in den Reihen der Bibliothekare, Verleger und teilweise auch der Wissenschaftler formierte, als Google sein Buchscanprojekt präsentierte.[12] Aus dieser Perspektive kann dem Engagement des Suchmaschinenriesen durchaus Positives abgewonnen werden, besteht doch mittlerweile die einhellige Meinung, dass Digitalisieren besser ist als untätig zu bleiben. Dieses Ziel wird seit kurzem durch die *Digital Public Library of America*

auch in den USA von offizieller Seite verfolgt.

Durch die globale Verknüpfung von Institutionen im Prozess der Buchdigitalisierung entsteht der Anschein einer durchorchestrierten Infrastruktur, die nationalstaatliche Grenzen überwunden hat und eine »vollständige« Erfassung des Buchwissens erreicht. Der »Mythos des universellen Wissens«,[13] der häufig mit dem Phänomen Google in Verbindung gebracht wird, projiziert sich dadurch, mindestens für den Raum Europa, auf die Europeana.[14] Dabei fällt auf: Geldmangel scheint es in diesem Bereich der Geisteswissenschaften – rechnet man denn die Zurverfügungstellung von Digitalisaten als Teil des Wissenschaftsbereichs – nicht zu geben. Neben regulären Fördermitteln findet sich hier eine Vielzahl von alternativen Finanzierungsquellen, die sonst kaum für die Förderung »herkömmlicher« Forschungsprojekte gewonnen werden könnten. Die Aufbereitung von Text, Bild und Ton scheint aufgrund der vermeintlichen politischen Neutralität förderungswürdiger zu sein, als langwierige und kritische Denkunternehmungen, die zudem noch scheitern können. Die Parallele zu den Anfängen der großen Editionsunternehmen, die im 19. Jahrhundert gestartet wurden, fällt deutlich ins Auge.[15] Neben der zeitlichen Folge – die (meist staatlichen) Institutionen begannen sich nach dem Start des Google-Buchprojekts mit den Möglichkeiten auseinanderzusetzen – fällt auf, dass sich die (teilweise universitären) Projekte zunehmend am marktwirtschaftlich operierenden Unternehmen orientieren. Es ist daher ist zu erwarten, dass Produkte wie der nGram-Viewer in naher Zukunft auch von weiteren Institutionen übernommen und angeboten werden (wie dies die Plattform Arxiv.org für naturwissenschaftliche Publikationen bereits demonstriert). Bekannt gemacht wurden die nGrams von Google durch einen Artikel in der Zeitschrift *Science*, der breit rezipiert wurde.[16] In diesem Artikel führten von Google gesponserte Wissenschaftler den Begriff »*culturomics*« (analog zu *genomics*) ein. Dank der Analyse einer möglichst großen Anzahl von Texten versprechen sie sich die Entschlüsselung des »kulturellen Genoms« der Menschheit. Durch die Analyse des Wortpools von Google Books soll eine Annäherung an historische Entwicklungen und Brüche möglich sein. *Culturomics* ist denn auch nicht weniger als eine neue Methode, welche die Geisteswissenschaften verbessern oder gar revolutionieren soll, denn die Häufigkeit eines bestimmten Wortgebrauchs ist durch die breite Datenmenge valide quantifizierbar.

Es ist wenig erstaunlich, dass die Idee zu *Culturomics* von Linguisten mitgeprägt wurde, verorten sich doch diese in einer besonderen geisteswissenschaftliche Disziplin, welche mit

naturwissenschaftlichen Auswertungsmöglichkeiten operiert und ihre Modelle und Theorien anhand von Sprachsammlungen quantitativ zu validieren versteht. Die Erstellung und Auswertung von Korpora, wie sie letztlich auch Google Books zusammenfügt, ist ein Verfahren, das an der Schwelle zwischen unterschiedlichen Formen von »Wissenschaftlichkeit« entstanden ist und durch die Anwendung von statistischen Auswertungen auf Kulturphänomene eine Sonderrolle einnimmt.

Auf den folgenden Seiten soll anhand des nGram-Viewers ein kritischer Blick auf die Digitalisierungsbegeisterung des 21. Jahrhunderts geworfen und im Umgang mit digitalen Großdatenbanken zur Vorsicht gemahnt werden. Meine Kritik erfolgt nicht aus einem konservativen Habitus heraus, der die Geisteswissenschaften in einem dem humboldtschen Ideal verpflichtenden Zustand verharrt sehen will und der hermeneutische Ansätze unangemessen überhöht. Es geht mir darum, auf relevante Probleme aufmerksam zu machen, die der Einsatz von computerisierter Technik mit sich bringt. Es ist nicht mein Ziel, quantifizierende Ansätze und neue Formen der Visualisierung *per se* zu verdammen, sondern im Gegenteil ihre sinnvollen Anwendungsweisen aufzuzeigen und kritisch zu diskutieren. Die Möglichkeiten des Umgangs mit großen Datenmengen erlauben, ja zwingen uns zur Auseinandersetzung mit Text- und anderen Korpora und deren Einbindung in Forschungsdesigns. Diese Überlegungen folgen aus der Beschäftigung mit linguistischen Verfahren zur Auswertung von Textkorpora, wobei nicht die Sprachgeschichte, sondern die Geschichte und die Historizität von Schrift und Schriftschaffen für mich im Mittelpunkt stehen.

Der vorliegende Essay besteht aus zwei Teilen: Während der erste Teil der kritischen Auseinandersetzung mit Großprojekten gewidmet ist, versteht sich der zweite als eine Skizze der Möglichkeiten sorgfältig erarbeiteter kleiner Korpora, die neue Erkenntnisse dank neuer digitaler Methoden versprechen.

Die Welt von Google Books und Culturomics

Der nGram-Viewer: Was kann dieses neue Werkzeug und wie nutzt man es? Mit dem nGram-Viewer kann das Vorkommen von bis zu fünf Wörtern im Korpus von Google Books (circa fünf Millionen gescannte Bücher) abgefragt werden.[17] Ebenso ist es möglich, die Häufigkeitsverteilungen mehrerer Wörter oder Phrasen im Verhältnis anzeigen zu lassen. Das Resultat sind immer ein oder mehrere Graphen, eine Auswertung in Zahlen wird nicht angeboten. Die rohen *grams*, das heißt, die Worte in ihrer deklinierten beziehungsweise konjugierten Form

(also nicht die Lexeme) mit der Angabe ihrer Häufigkeit, werden aber auch zum Download angeboten, so dass es grundsätzlich möglich ist, eigene Untersuchungen anzustellen, die nicht zwangsläufig in einen Graphen münden müssen. Interessanterweise finden sich weder in Zeitungsartikeln noch in den konsultierten Blogeinträgen Untersuchungen, die diese Möglichkeit wahrgenommen haben. Im Netz gibt es diverse Anleitungen, wie der nGram-Viewer zu bedienen ist, auch in fachwissenschaftlichen Publikationen finden sich bereits Ansätze.[18]

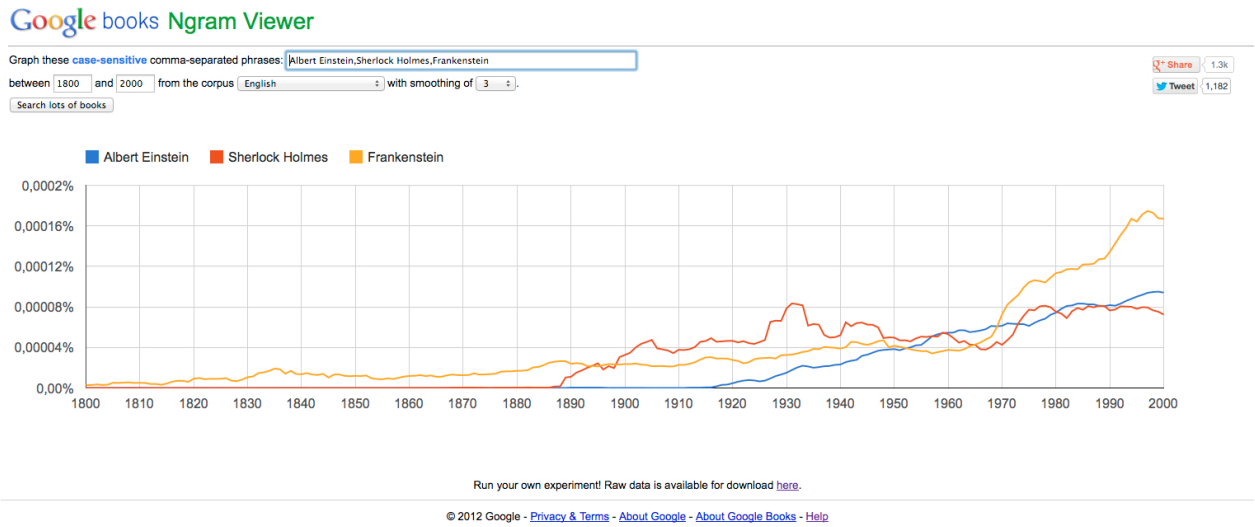

Abbildung 1: »Albert Einstein«, »Sherlock Holmes«, »Frankenstein«, Korpus »English«, 1800-2000 (gleitender 3-Jahresdurchschnitt) Quelle: Google Books nGram-Viewer [11. Dezember 2012]. Eingabefeld und Ansicht des Google Books nGram-Viewers.

Die Idee von *Culturomics* ist analog zum hergestellten Graphen recht simpel: Mittels des Abfragens von Worten und Phrasen sollen Trends in der Geschichte sichtbar gemacht werden. Das Paradebeispiel, das in besagtem *Science*-Aufsatz prominent platziert wurde, behandelt die Zensur berühmter Persönlichkeiten während der Herrschaft der Nationalsozialisten in Deutschland.[19] Anhand des Vorkommens des *bi-gram* »Marc Chagall« im englischen und deutschen Buchkorpus wird demonstriert,[20] dass der Künstler zwischen 1933 und 1945 im deutschen Korpus seltener vorkommt, als der Vergleich mit den Vorkommnissen vor 1933 und nach 1945 erwarten ließen. Im englischen Korpus hingegen entsprach die Häufigkeit der Zeichenfolge »Marc Chagall« in derselben Zeit den Erwartungen. Anhand der Wortkombination soll folglich »Zensur« sichtbar gemacht werden.

Weitere Beispiele sind etwa die bi-gramme »Henri Matisse« und »Pablo Picasso«:

Aufgrund von Häufigkeiten – wiederum das deutsche mit dem englischsprachigen Korpus vergleichend – wird ein sogenannter *»suppression index«* erstellt. Der Index quantifiziert den Unterschied zwischen erwarteter und tatsächlicher Häufigkeit des Vorkommens der bi-gramme in der Zeit zwischen 1933 und 1945. Während sich im englischsprachigen Korpus die Häufigkeit den Erwartungen entsprechend entwickelt, bewegen sich im deutschen Teil die Ausschläge in Richtung von »Häufigkeit tiefer als erwartet«. Als Kontrollgruppe wird eine ungenannte Anzahl an Personen – tituliert als »Nazis« – ins Feld geführt, deren Vorkommen zwischen 1934 und 1943 um einiges höher ausfällt, als aufgrund des Vergleichs zu erwarten wäre.[21] Das Potenzial der Methode scheint verlockend und ähnelt naturwissenschaftlichen Versuchsanordnungen, die – weit entfernt von geisteswissenschaftlicher Forschung – objektive Resultate versprechen. Ein Blick in den Maschinenraum des nGram-Viewers trübt jedoch den Eindruck und macht Probleme sichtbar.

Text, Metadaten und Materialität – Problemfelder des nGram-Viewers

Der Ansatz *Culturomics* verspricht nichts weniger als *»the application of high-throughput data collection and analysis to the study of human culture«*.[22] Obwohl sich dieses Ziel ausgereift anhört, birgt das Projekt eine Vielzahl von technischen Mängeln, die hier kurz zu rekapitulieren sind:

Der Grundpfeiler der nGrams sind gescannte und danach durch Texterkennung (sogenannte *OCR*: *Optical Character Recognition*) verarbeitete Worte. Ausgehend von gescannten Seiten, die natürlich qualitativ besser oder schlechter sein können, versucht ein auf der Erkennung von Mustern und auf Abgleich mit Wörterbüchern basierender Automatismus sinnvolle Worte zusammen zu setzen.[23] OCR-Verfahren werden seit einigen Jahren mit einer Selbstverständlichkeit eingesetzt, die fast schon als fahrlässig bezeichnet werden muss. Betrachtet man die Resultate im Detail, wird schnell klar, wie unsauber die Erkennung funktioniert, wie häufig Wörter falsch oder unkorrekt gelesen werden. Als User des nGram-Viewer bemerkt man dies jedoch nicht, da bei einer Volltextsuche nur Treffer angezeigt werden, Falschlesungen verschwinden im Rauschen des Textes.[24]

Die Fehlerhaftigkeit an sich ist ein Problem der Technik und für die Auswertung nicht *per se* problematisch, auch deklarieren die meisten OCR-Engines ihre Genauigkeit (häufig um 99,99 %, 1 Fehler pro 10’000 Zeichen). Es ist hingegen bedenklich, dass die Mehrheit der Institutionen und Firmen, die in großer Zahl scannen, der Erkennrate keine entsprechende

Relevanz beimessen. Obwohl die Intransparenz auch bei einem privaten Unternehmen wie Google störend ist, so lässt sich dagegen nur wenig ins Feld führen. Allerdings ist es alles andere als redlich, wenn eng an den Wissenschaftsbetrieb angeschlossene Unternehmungen es nicht für nötig erachten aufzuzeigen, wie häufig die Resultate mittels Volltexterkennung verfälscht werden. Deren Praxis des Scannens und Aufbereitens orientiert sich blind am Vorbild von Google. Die Lage ist umso prekärer, als die Anzeige häufig ein Bild (eine »Fotografie«) des Scans anzeigt, ohne darzulegen, welche Worte wie erkannt wurden und als Text »hinter« dem Bild mitgeführt werden. Dem unbedarften Nutzer zeigt sich ein perfekt erscheinender Scan und da nur richtig erkannte Treffer gezeigt werden (fälschlicherweise als Treffer erkannte Resultate, sog. *false-positive* sind bei OCR-Engines eher selten), wiegt der Nutzer sich in falscher Sicherheit.

Aus der Google-Kurve ist natürlich auch nicht ersichtlich, wo das gesuchte nGram gefunden wurde. Handelt es sich um einen Begriff im Anhang eines Buches, etwa im Literaturverzeichnis oder gar in einer Verlagswerbung, die auf den letzten Seiten des Buchs zu finden ist? Die sorgfältig durch die Autoren oder Herausgeber erstellte Hierarchisierung und Unterscheidung zwischen Text und Paratext wird konsequent überlesen und als gleichwertig eingestuft. Für die Zählung in der Datensammlung von Google macht es keinen Unterschied, ob ein Wort oder eine Phrase Teil der Legende, einer Fußnote oder des Fließtextes ist. Ebenso spielt es überhaupt keine Rolle, ob das gefundene nGram einmal oder vielfach in einem Buch vorkommt, gezählt wird jeweils ein Treffer in der Statistik. Im nGram-Viewer spielt es zudem keine Rolle, ob es sich um einen Werbeprospekt, einen mehrsprachigen (!) Sammelband oder um eine wissenschaftliche Quellenedition handelt: Alle Publikationsformen werden in einzelne Worte zerlegt und gleichwertig behandelt.

Die Entwickler und Auszähler der nGrams haben mit Problemen der Unschärfe gerechnet und eine Untergrenze festgelegt. Nur Zeichenkombinationen, die häufiger als vierzig Mal in einem Jahr vorkommen, werden überhaupt als »existent« erfasst. Ähnlich problematisch gestaltet sich der Umgang mit der Qualität von OCR: Von allen gescannten Büchern entsprach nur etwa ein Drittel den Anforderungen, die restlichen Bücher wurden aus dem Datensatz entfernt. Welche Standards waren maßgebend für den Einbezug? Nirgends finden sich Angaben dazu. Fest steht: Eine diffuse Definition von »Datenqualität«[25] bestimmt, was in das Korpus kommt.
Und wie wird bestimmt, welches Buch welchem Korpus zugeordnet wird? Anwort auf die Frage sind die sogenannten Metadaten, die auf nicht nachvollziehbaren Wegen ins Google-Books-Programm kommen. Bei der ersten Veröffentlichung des nGram-Viewers erntete das Google-

Team für derartige Fehler auf verschiedensten Blogs Hohn und Spott:[26] Sei es, weil Mark Twain als Autor von Büchern angegeben wurde, die vor seiner Geburt erschienen, oder weil laut Google-Books bereits im 17. Jahrhundert Teenie-Vampir-Geschichten auf Hochglanzpapier und als Bücherserie verbreitet worden sein sollen,[27] oder pikanterweise der Begriff »google« hundert Jahre alt sein müsste: »Google« ist eine Falschschreibung von »googol« (Ausdruck für 10^{100}). Mitte September 1997 wurde die Domain »Google.com« registriert. Trotzdem taucht der Begriff im Google Korpus vor diesem Datum auf.

Abbildung 2: »google«, Korpus »English«, 1800-2000 (gleitender 3-Jahresdurchschnitt) Quelle: Google Books nGram-Viewer [06. Juli 2013].

Noch komplizierter gestaltet sich der Einbezug von Zeitschriften: Teils fließen diese in die verschiedenen Google-Korpora ein, teils nicht – dies scheint dem Umstand geschuldet zu sein, dass die Entwickler der Korpora explizit von Buchabfragen ausgehen und (noch) nicht von Zeitschriften.[28] Zuweilen werden daher gesamte Zeitschriftenbestände auf die Erstausgabe datiert, was zu schwerwiegenden Mengenfehlern führt, oder aber es werden nur einzelne Ausgaben aufgenommen, so dass sich die Frage stellt, was denn nun eigentlich das Korpus bzw. die Korpora des nGram-Viewers ausmacht.

Handelt es sich bei diesen Problemen, wie man einwenden könnte, nur um Kinderkrankheiten und wird der nGram-Viewer in Zukunft technisch »perfektioniert«? Sogar wenn dies der Fall wäre, stellen sich weitere Fragen.

Ausgewählt: Das Korpus und seine »Kultur«

Lenkt man den Blick vom einzelnen Buch auf eine höhere Stufe, den Korpora, und bewegt sich weg von der digitalen Maschinerie hin zu den Praktiken der Vorauswahl in Bibliotheken, so multiplizieren sich die Probleme. Obwohl sich Google gerne mit seinen universitären und institutionellen Partnerschaften brüstet, wird nirgends offen gelegt, welche Bücher letztlich gescannt wurden. Im deutschsprachigen Raum stellen die Bayerische Staatsbibliothek (BSB) und die Österreichische Nationalbibliothek (ÖNB) ihre Bestände Google zur Verfügung.[29] Diesen gegenüber stehen unter anderen die grossen angloamerikanischen Bibliotheken von Harvard, Princeton, Stanford und Oxford, die wohl auch nicht wenige Werke zum Korpus »German« beigetragen haben dürften. Die Beschränkung auf renommierte Institutionen ist einerseits verständlich, da sie eine Vielzahl von »wichtigen« Werken versammeln, andererseits wird damit einer Kanonisierung des Wissens Vorschub geleistet, das sich an Hochschulrankings orientiert.

Bei der Bayerischen Staatsbibliothek, Hauptlieferant des deutschsprachigen Korpus, geht das Problem noch einen Schritt weiter: Gemäß Sammlungsauftrag gelten hier neben der »Bavarica« zum Beispiel auch die »Altertumswissenschaften«, »Buch-, Bibliotheks- und Informationswissenschaften« und »Musik« von Haus aus als relevante und daher zu sammelnde Bestände. Aufgrund der Kooperation mit Google erhalten diese Schwerpunktprogramme eine zusätzliche Relevanz, da diese Bücher direkt in die Korpora von Google einfließen. Seit die Verlage ab 2004 direkt mit Google kooperieren können, nimmt die verhältnismäßige Häufigkeit des *grams* »Bayern« interessanterweise und erwartungsgemäß rapide ab.

Eine weitere Problematik der BSB ist die Verlustrate während der Zeit zwischen 1939 und 1945. Laut Schätzung von Experten sollen vier Millionen Einheiten nach oder während des Zweiten Weltkriegs aus der Münchener Bibliothek verschwunden sein.[30] Diese Menge an Bücher muss zweifelsohne einen Einfluss auf die Ergebnisse des nGram-Viewers haben.

Zu guter Letzt ein kulturtheoretisches Argument: Was genau mit »*culture«* (immerhin ein wichtiger Wortbestandteil von *Culturomics*) gemeint ist, wird nirgends gesagt. Im Gegenteil: In dem genannten *Science*-Aufsatz wird eine Vielzahl von Begriffen und Wortkombinationen unreflektiert durch die Vergleichsmaschine gejagt. Dass das Ganze für einen konzeptionellen Aufsatz, der auf die Möglichkeiten der gesammelten nGramme aufmerksam machen soll, nicht unproblematisch ist, liegt auf der Hand. Man befindet sich in einem Zirkelschluss: Worte – und damit Sprache – werden durch Grams definiert (vierzig Treffer pro Jahr berechtigen zum Eintritt

in die Welt der Sprache). Gleichzeitig soll die so »zugelassene« Sprache Kultur darstellen und zwar für denjenigen Teil der Menschheit (mit *Culturomics* wird schließlich *human culture* »genomisiert«),[31] die einen Sprachkorpus verwenden. Gemäß *Culturomics* ist kulturell wichtig, was – überspitzt formuliert – häufig als Wort oder Phrase zu einem gegebenen historischen Moment vorkommt. Kultur wird reduktionistisch als Sammlung von Daten aufgefasst, durchaus nicht weit entfernt von der zurecht oft kritisierten geertzschen Metapher von Kultur als »Ensemble von Texten«. Eine solche reduktionistische Definition wird im Übrigen auch von vielen Naturwissenschaftlern nicht geteilt.[32]

Zusammenfassend muss dem *Culturomics*-Projekt von Google ein klar ungenügendes Zeugnis ausgestellt werden. Die Resultate sind durch methodische Fehler verfälscht und unzuverlässig. Auslesevorgänge und Qualitätsstandards werden nicht offen gelegt und werden wie ein Betriebsgeheimnis behandelt. Produktions- und Überlieferungsbedingungen von Büchern können nicht differenziert betrachtet werden, sondern vermengen sich bei der Datenausgabe.[33] Schließlich ist auch die Langfristigkeit der Bereitstellung der Daten nicht garantiert. Das Ende des beliebten Google Readers legt diesen Schluss nahe.

Die im Anfangsabschnitt beschriebene hegelsche Fabrikmaschine kommt einem wieder in den Sinn. Die automatisierte Abfrage mit einer Kurve als Resultat zeugt von einer Entfremdung und Entmündigung des Anwendenden zugleich.[34] Die Materialität der erforschten Erzeugnisse und die Nähe zum Text verschwinden in der Weichzeichnung des Graphen. Dennoch werden sich die Geisteswissenschaften jetzt und in Zukunft mit solchen Korpora beschäftigen müssen, denn der nGram-Viewer von Google ist nur die Speerspitze. Digitalisiert wird auf allen Ebenen und die Aufbereitung von großen Textmengen ist die logische Folge: Mittels News-Aggregaten wird etwa versucht, politische Ereignisse zu prognostizieren.[35] *Preprints* wissenschaftlicher Artikel werden direkt als nGrams zur Verfügung gestellt.[36] Forschungsdesigns werden auf Grundlage von großen und ungenau definierten Korpora erstellt.[37]

Die Arbeit mit großen Datenmengen (neudeutsch: *big data*) ist heute nicht nur Ausdruck von Internet-Gigantismus im Stile von Google, sondern auch in den geisteswissenschaftlichen Fächern eine Tatsache. Erstaunlich ist dabei jedoch die Anlehnung der Wissenschaftler an die Praktiken privatwirtschaftlicher Großunternehmen, ohne hinreichend zu reflektieren, was dort genau wie und weshalb getan wird. Der Maschinerie Google wird in vielerlei Hinsicht nachgeeifert, während Transparenz, kritische Distanz und letztlich die genuinen Interessen

geisteswissenschaftlicher Forschung auf der Strecke bleiben.

Sinnvolle Methoden mit stimmigen Korpora: Eine Ideenskizze

Nach all der Kritik bleibt die Frage: Sollen die Geisteswissenschaften methodisch bei der Kulturanalyse mittels qualitativer Tiefenbohrung verbleiben? Ja, aber bitte nicht ausschließlich! Auch wenn die Resultate des nGram-Viewers als problematisch zu erachten sind, formiert sich hier ein Methodenapparat, der auch in vorwiegend qualitativ arbeitenden Wissenschaftszweigen nutzbar gemacht werden kann (und muss). Ein passend unpassender Ansatz dazu aus den Sportstatistik verliebten USA:

Im Jahr 2002 erreichten die *Oakland Athletics*, ein Team der nordamerikanischen Baseball-Liga, das Viertelfinale der nationalen Meisterschaft. Trotz der mit Abstand tiefsten Lohnkosten und der Abgänge wichtiger Spieler vor Saisonbeginn schafften die *A's*, woran 22 andere Mannschaften scheiterten: Die Qualifikation für die Playoffs. Auf dem Weg zu diesem Erfolg wurden sie während zwanzig Spielen in Folge nicht geschlagen, bis heute ein Rekord in der über hundertjährigen Geschichte der *American League*. Die mittlerweile sogar mit Brad Pitt verfilmte Erfolgsgeschichte ist dabei eng mit dem Begriff *Sabermetrics* (von *SABR: Society for American Baseball Research*) verknüpft.[38] *Sabermetrics* bezeichnet die Suche nach objektivem Wissen über Baseball.[39] Ursprünglich dienten die Spielstatistiken im Baseball dazu, um zu diskutieren, welche Spieler in die Riege der Besten, in die *Hall of Fame* aufgenommen werden sollten. Sie weisen darüber hinaus aber auch ein prognostisches Potenzial auf. Mithilfe von *Sabermetrics* lässt sich etwa die Wahrscheinlichkeit vorausberechnen, wie viele Läufe ein Spieler pro Spiel erzielen wird (wahlweise unter Berücksichtigung, ob der Werfer links- oder rechtshändig ist) oder dass ein Werfer gegenüber einem bestimmten Schlagmann seine Würfe viermal außerhalb der *Strike*-Zone anbringt. Unter relativ strikter Anwendung von Daten und Vergleichsmaterialien gelang es dem Management der *Oakland Athletics* so, ein Team zusammenzustellen, das aus *sabermetrischer* Perspektive optimal und dennoch preisgünstig war, da den meisten verpflichteten Spielern nur wenig von den anderen Teams zugetraut wurde. Die Siegesserie und das Erreichen der sogenannten *postseason* war das Resultat.

Dem Westküsten-Team ist es zwar bislang nicht gelungen, die Meisterschaft, die *World Series*, zu gewinnen, doch hat sich die Anwendung von *Sabermetrics* auch in anderen Teams eingebürgert. Ein streng kontrolliertes Korpus eignet sich offensichtlich als gewinnbringendes Analysetool. Doch auch im Sport gibt es Manager, die sich gegen den Einsatz statistischer

Methoden sträuben und sich lieber auf Augenschein und Erfahrung verlassen. Ähnlich verhält es sich in aktuellen Diskussionen der Geisteswissenschaften gegen den Einsatz von quantitativen Methoden.

Bei aller Verschiedenheit von Sport und Wissenschaft unterliegt der Einsatz eines Datenkorpus zum Zwecke einer Auswertung einer wichtigen Voraussetzung: Vergleichbarkeit muss gegeben sein. Oder anders gesagt: Die Zahlengrundlage muss nachvollziehbar sein. Sportanalysten – und damit verlassen wir die Welt des Baseballs wieder – stützen ihre Analyse auf eine »vollständige« Datengrundlage ab. Dem Geisteswissenschaftler bleibt hingegen nichts anderes übrig, als seine Daten, seien es Texte, Zahlen oder Bilder, mühsam und quellenkritisch zusammen zu stellen sowie Vorbehalte bezüglich Überlieferungsverlusten und anderen Aspekten in Fußnoten anzubringen. Welche Optionen und Ansätze gibt es hier?

Kleine Korpora/Maschinen: Umgang mit Text und Daten im digitalen Zeitalter

Meine Vorschlag zur Lösung des dargelegten Problems ist es, dass die Geisteswissenschaften sich bewusst auf kleine Korpora beschränken sollten, um diese jeweils mit einem spezifischen Set an Methoden analysieren zu können.[40] Im Gegensatz zum Google-Books-Korpus wird so statt eines unüberschaubaren Bestandes an Büchern und Texten eine bewusste für die gewählte Fragestellung relevante Auswahl in Betracht gezogen, die als solche auch ausgewiesen und reflektiert wird.

Eine beispielhaftes Vorgehen findet sich im Aufsatz von Ludolf Kuchenbuch zu einem Traktat des 12. Jahrhunderts (*De diversis artibus*).[41] Mit einem Ansatz, den Kuchenbuch weniger als Begriff denn als Hilfsgriff »mikrosemantisch« nennt, versucht er das Traktat unter Einbezug von korpuslinguistisch geprägten Denkanstößen zu verstehen. Konkret sucht Kuchenbuch im Werk des Theophilus, ausgehend von einem Kapitel über ein Buch hin zum ganzen Werk, nach dem »skelettierten Norm-Satz« und dem »regierenden Verb«. Unter Zuhilfenahme von Auszählungen und Positionsanalysen von Verben, Metaphern und Synonymen innerhalb des Werkes vermag Kuchenbuch – verkürzt gesagt – darzulegen, dass Theophilus sein Werk und die Arbeit daran als Gottesdienst versteht.[42] Zudem stellt er Bedeutungsunterschiede in der Verwendung der semantisch verwandten Verben *laborare*, *operare*, *facere* und *formare* fest. Er knüpft damit zwar an eine bestehende Deutungstradition an, jedoch aus neuer methodischer Perspektive. Die Vorgehensweise liefert demnach sowohl zum Aufbau und zeitgenössischen Wortgebrauch als auch zu einem alternativen Verständnis des Werkes diverse neue Ansätze.

Neben derartigen analytischen Mischformen zwischen Hermeneutik und Korpuslinguistik sind vielfältige Einsatzmöglichkeiten von gut aufbereiteten Korpora denkbar: Um Wortgebrauch und semantischen Wandel nachzuvollziehen, braucht es jedoch mehr als bi-grams, die zufällig nebeneinander stehen. Der sogenannte Ko-Text wird relevant, also die Wörter im näheren und weiteren Umfeld der untersuchten Begriffe.[43] Dies führt zu weitergehenden Fragen: In welchen Kombinationen werden Wörter benutzt? Handelt es sich dabei um feste Ausdrücke oder nicht? Über die Bedeutung der im Mittelalter und in der Frühen Neuzeit häufig auftretende Paarformel »Schutz und Schirm« tobte Ende der 1990er Jahre ein akademischer Streit. Dabei waren Interpretationen von Quellenstellen und Häufigkeiten eminent wichtig.[44] Entsprechende Volltextsuchen mit Informationen zum Herstellungshintergrund der Quellen könnten derartige Diskussionen heute wesentlich breiter abstützen.

Darüber hinaus können Untersuchungen zu Wortfeldern angestellt werden, wie etwa zur Frage: Bedeutet der Begriff »Wein« in einem tausendjährigen Stück überhaupt dasselbe wie heute oder ist er ein »falscher Freund«?[45] Ebenso kann danach gefragt werden, ob der Ko-Text über die Zeit stabil bleibt oder sich Verschiebungen feststellen lassen?[46] Mittels nGrams wird zuweilen auch versucht, Dokumente zu datieren. Mit derartigen statistischen Annäherungen können allein basierend auf Wort- und Phrasenbestand Hypothesen aufgestellt werden, wann eine Urkunde erstellt wurde.[47] Die Erarbeitung von Wissensnetzwerken ist denkbar. In der gelungenen Anwendung eAQUA wird etwa »strukturiertes Wissen aus Antiken Quellen« extrahiert. [48] Aus der Kookkurrenz von Wörtern in Texten können so *mental maps* erstellt werden.[49]

Derartige Fragestellungen und Methoden stellen mögliche Ausgangspunkte dar, um »soziale« und «kulturelle« Phänomene zu untersuchen. Sie ersetzen die etablierten Herangehensweisen der Geisteswissenschaftler nicht, sondern fordern vielmehr deren Weiterentwicklung und Reflexion. Die Vorteile kleiner Korpora liegen auf der Hand: Auch wenn keine Vollständigkeit erreicht werden kann, so vermittelt deren lokale Abgeschlossenheit eine hinreichende Rahmung für Forschung. Auch die in den letzten Jahren so vielbeachtete Bedeutung der Materialität von Texten kann berücksichtigt werden ebenso wie die Vielfalt der Publikationsformen: Man könnte etwa entscheiden, ob man Werbeprospekte neben kanonischen Werken in das eigene Forschungskorpus aufnehmen möchte, oder nicht. Das Korpusdesign kann den jeweiligen Ansprüchen der konkreten Forschungsfrage angepasst werden und wäre produktiver als die Arbeit mit einem diffusen, alles umfassenden Korpus.

Großprojekte und ihre Chancen

Der Google-nGram-Viewer ist zwar höchst problematisch, er eröffnet zugleich aber, auch dank der hohen Publizität, Chancen. Die Öffentlichkeit, und damit indirekt auch die geisteswissenschaftliche Forschung, wird gezwungen, sich mit dem Phänomen *big data* auseinanderzusetzen – hier besteht aus geisteswissenschaftlicher Sicht in der Tat Nachholbedarf. Die Frage der Darstellungsweise, Visualisierungen, Graphen oder andere Formen der statistischen Auswertungen, die bisher vor allem in der Wissenschaftsforschung und Wissensgeschichte problematisiert wurde, wird für die Geisteswissenschaften insgesamt relevant. Zugleich ermöglicht diese Auseinandersetzung eine bessere Orientierung in einer immer stärker computerisierten und datenanhäufenden Welt. Die Auswertung von Textkorpora ist ein zukunftsträchtiger Teil der Geisteswissenschaften und muss aus kritischer Distanz, aber auch mittels praktischer Erprobung geprüft und in Forschung integriert werden. Weder eine grundsätzlich ablehnende Haltung noch eine technikverliebte Naivität werden dem digitalen Wandel gerecht, vielmehr müssen konkrete Anforderungen (nicht *best* sondern *necessary practices*) formuliert und umgesetzt werden:
Erstens sollten große Korpora die Auswahl einzelner oder mehrerer Kleinkorpora erlauben; zweitens muss die Weiterverarbeitung erstellter Korpora (etwa die Beigabe von Metadaten oder die textkritische Auszeichnung) durch außenstehende Nutzer möglich sein; drittens werden Standards offen kommuniziert und bleiben langfristig nachvollziehbar; viertens werden alle Daten konsequent für alle offen gelegt (im Sinne von Open Access).[50]

Alle diese Punkte betreffen offenkundig nur die Aufbereitung (und eventuelle Weiterverarbeitung) von Daten. Die Auswertung muss nicht zwangsläufig in statistische Validierung münden, unterschiedliche Vorgehensweisen der Textanalyse sind möglich: Die Auswertung von großen Beständen genauso wo eine detailscharfe Tiefenanalyse. Wichtig ist die Legitimierung der gewählten Ansätze vor dem Hintergrund des jeweiligen Erkenntnisinteresses. Forschende dürfen nicht zu reinen Inputgebern einer Maschinerie werden. Es gilt, die Komponenten (etwa das Korpus) und die angewandte Methode zu durchschauen und in die Analyse mit einzubeziehen.

Auch in Zukunft werden sich die Geisteswissenschaften weiter mit Texten auseinandersetzen, sie auf unterschiedlichste Art und Weise interpretieren, sie be- und weiterverarbeiten – nicht zuletzt in Form großer und kleiner Korpora, die wieder und wieder neu

konfiguriert werden müssen.

---

[1] Ich danke Nadja Schorno für die Hilfe bei der Textgestaltung, Peter Dürmüller, Nanina Egli, Petra Hornung, Simon Teuscher und Christoph Stätzler für vielseitige Gedankenanstöße sowie für wichtige Anregungen aus der anonymen Begutachtung. Erste Überlegungen dieses Beitrages habe ich auf meinem Blog veröffentlicht: http://solascriptum.wordpress.com

[2] Vergleiche Georg Wilhelm Friedrich Hegel: *Jenenser Realphilosophie. Natur- und Geistesphilosophie*, Leipzig 1931, S. 239.

[3] Am prominentesten: Nicholas G. Carr: *The Shallows. What the Internet Is Doing to Our Brains*, New York 2011, innovativer und weniger pessimistisch Ken Auletta: *Googled. The End of the World as We Know It*, London 2010.

[4] Eine hervorragende Einordnung zu Google als Linksammlung mit Filterfunktion leistet Ulrike Bergermann: »Linkspeicher Google. Zum Verhältnis von PageRank und Archäologie des Wissens«, in: Thomas Weitin und Burkhardt Wolf (Hg.): *Gewalt der Archive*. Paderborn 2012, S. 371–391.

[5] Aussage von Sergei Brin, im *Associated Press Worldstream*, 2004, zitiert nach Pia Dietrich u. a.: *Google Buchsuche. Chance oder Gefahr für die Bibliothekswelt*, Unveröffentlichte Diplomarbeit, Luzern 2006.

[6] Jean-Baptiste Michel u.a.: »Quantitative Analysis of Culture Using Millions of Digitized Books«, in: *Science* 331 (6014), 2011, S. 176–182. Vergleich auch Erez Lieberman u.a.: »Quantifying the Evolutionary Dynamics of Language«, in: *Nature* 449 (7163), 2007, S. 713–716.

[7] Steven Cherry: »The Cultural Treasures in Google Ngram«, in: *ieee Spectrum*, 9. Juli 2012, http://spectrum.ieee.org/podcast/geek-life/profiles/the-cultural-treasures-in-google-ngram (aufgerufen: 3. 1. 2013). Die ständige Verbesserung des Tools bleibt eines der Ziele von Google, so wartete der Konzern 2012 mit einer neuen Version auf. Die Neuerungen betreffen vor allem linguistische Abfragemöglichkeiten, vergleiche Yuri Lin u.a.: »Syntactic Annotations for the Google Books Ngram Corpus«, in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju 2012, S. 169–174.

[8] Zur Geschichte der Datenbank und insbesondere der Auswertung von Text als Daten vergleiche David Gugerli: »Die Welt als Datenbank. Zur Relation von Softwareentwicklung, Abfragetechnik und Deutungsautonomie«, in: *Nach Feierabend* 3, 2007, S. 11–36.

[9] Carr: *The Shallows,* a.a.O., S. 150ff. Kritisch gegenüber der Datenerhebung vergleiche Stephen Marche: »Literature Is Not Data: Against Digital Humanities«, in: Los Angeles Review of Books, 28. Oktober 2012, http://lareviewofbooks.org/article.php?id=1040 (aufgerufen: 8. 1. 2013].

[10] Leicht kritisch zum Umgang mit Google nGram, insbesondere zur Auswertung in Form von Kurven ist Philipp Sarasin: »Sozialgeschichte vs. Foucault im Google Books Ngram Viewer: Ein alter Streitfall in einem neuen Tool«, in: Pascal Maeder, Barbara Lüthi und Thomas Mergel (Hg.): *Wozu noch Sozialgeschichte*? Göttingen 2012, S. 151–174.

[11] *»Assemble[...] collections [that] allow you to explore Europe's history from ancient times to the modern day.« zit. nach* »Europeana. Think Culture«, URL: http://www.europeana.eu/portal/aboutus.html (aufgerufen: 27. 11. 2012).

[12] Mehrere institutionell gut vernetzte Persönlichkeiten wehrten sich gegen das Scanvorhaben, vergleiche Jean Noël Jeanneney: *Google and the Myth of Universal Knowledge: A View from Europe*, Chicago 2007; Bruno Racine: *Google et le nouveau monde*, Paris 2010; Alain Jacquesson: *Google Livres et le futur des bibliothèques numériques*, Paris 2010.

[13] Jean-Noël Jeanneney: *Google and the Myth of Universal Knowledge. A View from Europe*,

Chicago, IL 2007.
[14] Eine Übersicht der Schweizer Digitalisierungsprojekte: https://www.digicoord.ch/index.php/Digitalisierungsprojekte (aufgerufen: 27. 11. 2012).
[15] Vergleiche MGH (Hg.): *Zur Geschichte und Arbeit der Monumenta Germaniae Historica. Ausstellung anläßlich des 41. Deutschen Historikertages München, 17.-20. September 1996. Katalog*, München 1998.
[16] Michel: »Quantitative Analysis of Culture« in: *Science*, a.a.O.
[17] Kurzbeschreibung von *Google Books* siehe Carr: *The Shallows*, a.a.O., S. 161–166; Selbstbeschreibung bei Google: »Über die Google Buchsuche«, http://books.google.ch/intl/de/googlebooks/about.html (aufgerufen: 27. November 2012).
[18] Vergleiche Sarasin: »Sozialgeschichte vs. Foucault«, in: *Wozu noch Sozialgeschichte*?, a.a.O., S. 151–153.
[19] Hier und im Folgenden nach Michel: »Quantitative Analysis of Culture« in: *Science*, a.a.O., S. 180–182.
[20] Ein *gram* ist eine unbestimmte Einheit; mit *n-gram* wird eine unbestimmte Anzahl von *grams* bestimmt; ein *bi-gram* ist folglich zwei von der Einheit. *Grams* stehen bei Google für Wörter. Ein *bi-gram* ist also die Kombination von zwei Wörtern in fester Reihenfolge.
[21] Alle genannten Beispiele beziehen sich ebd., S. 180 (Abbildungen 4A, 4E und 4F).
[22] Ebd., S. 181.
[23] Für harsche Kritik am OCR, vergleiche Danny Sullivan: »When OCR Goes Bad. Google's Ngram Viewer & The F-Word«, 19. Dezember 2010, http://searchengineland.com/when-ocr-goes-bad-googles-ngram-viewer-the-f-word-59181 (aufgerufen: 28. 11. 2012).
[24] Zur Problematik von OCR-Engines mit Verbesserungsansätzen vergleiche Michael Piotrowski: *Natural Language Processing for Historical Texts*, San Rafael 2012, S. 25–48.
[25] »*We selected a subset of over 5 million books for analysis on the basis of the quality of their OCR and metadata*«, aus Michel: »Quantitative Analysis of Culture« in: *Science*, a.a.O., S. 176.
[26] Weiterführende Verweise in Tim Carmody: »Scholar To Google. Your Metadata Sucks«, 30. August 2009, http://snarkmarket.com/2009/3281 (aufgerufen: 7. 1 2013).
[27] Das Erscheinen des Buches Ellen Schreiber: *Vampire Kisses 9. Immortal Hearts*, New York 2012, ist bei Google auf 1612 datiert.
[28] Michel: »Quantitative Analysis of Culture« in: *Science*, a.a.O., S. 176 und S. 181.
[29] Einschränkend ist zu bemerken, dass die ÖNB nur ihre Bestände bis zum 19. Jahrhundert ins Google Programm aufnehmen ließ.
[30] Vergleiche Klaus Haller: »Bewegte Geschichte«, in: Rolf Griebel und Klaus Caynowa (Hg.): *Information – Innovation – Inspiration. 450 Jahre Bayerische Staatsbibliothek*, München 2008, S. 127–164, hier S. 145.
[31] Michel: »Quantitative Analysis of Culture« in: *Science*, a.a.O., S. 181.
[32] Vergleiche etwa die Kulturvorstellung von Hillis, in: John Brockman: *Die dritte Kultur. Das Weltbild der modernen Naturwissenschaft*, München 1996, S. 534.
[33] Zum Beispiel problematisiert für die mittelalterliche Überlieferung von Simon Teuscher: »Document Collections, Mobilized Regulations, and the Making of Customary Law at the End of the Middle Ages«, in: *Archival Science* 10 (3), 2010, S. 211–229.
[34] Vergleiche Hegel: *Jenenser Realphilosophie*, a.a.O., S. 239.
[35] Kalev H. Leetaru: »Culturomics 2.0. Forecasting Large-Scale Human Behavior Using Global News Media Tone in Time And Space«, in: *First Monday Online* 16 (9), 2011.
[36] http://arxiv.culturomics.org (aufgerufen: 3. 1 2013).

[37] Vergleiche Bernhard Jussen: »›Ordo‹ zwischen Ideengeschichte und Lexikometrie. Vorarbeiten an einem Hilfsmittel mediävistischer Begriffsgeschichte«, in: Bernd Schneidmüller und Stefan Weinfurter (Hg.): *Ordnungskonfigurationen im hohen Mittelalter*, Ostfildern 2006, S. 227–256.
[38] Vergleiche Michael Lewis: *Moneyball. The Art of Winning an Unfair Game*, New York 2003.
[39] Übersetzung des Autors nach Gabriel B. Costa, Michael R. Huber und John T. Saccoman: *Understanding Sabermetrics. An Introduction to the Science of Baseball Statistics*, Jefferson, NC 2008, S. ix.
[40] Vergleiche die Forderung nach Methodenvielfalt, insbesondere der *sémantique historique*: Alain Guerreau: *L'avenir d'un passé incertain. Quelle histoire du Moyen Âge au XXIe siècle*?, Paris 2001, S. 191–237.
[41] Ludolf Kuchenbuch: *Reflexive Mediävistik. Textus, Opus, Feudalismus*, Frankfurt am Main 2012, S. 341–401.
[42] Zitiert nach ebd., S. 349f.
[43] Eine Suchmöglichkeit, die vom Google-nGram-Viewer zwar angeboten wird, ohne jedoch genaue Definitionsmöglichkeiten zuzulassen, wie weit entfernt die Worte stehen können beziehungsweise müssen.
[44] Vergleiche Christine Reinle: *Bauernfehden. Studien zur Fehdeführung Nichtadliger im spätmittelalterlichen römisch-deutschen Reich, besonders in den bayerischen Herzogtümern*, Stuttgart 2003.
[45] Guerreau ist der Meinung, dass stark zwischen dem mittelalterlichen und modernen Verständnis unterschieden werden muss, vergleiche ders.: *L'avenir d'un passé incertain*, a.a.O., S. 195–207.
[46] Aus einer technischen Perspektive: Kalev Leetaru: *Content Analysis. A Data Mining and Intelligence Approach*, London 2012.
[47] Gelila Tilahun, Andrey Feuerverge und Michael Gervers: »Dating Medieval English Charters«, in: *Annales of Applied Statistics* 6 (4), 2012, S. 1615–1640.
[48] http://www.eaqua.net/index.php (aufgerufen: 17. 1 2013).
[49] Vergleiche Roxana Kath: »Das Konzept des ›einfachen Lebens‹ in der Antike. Ein Beispiel für die Anwendung von Textmining-Verfahren in der Geschichtswissenschaft«, in: Charlotte Schubert und Gerhard Heyer (Hg.): *Das Portal eAQUA*, (Working Papers CONTESTED ORDER, Bd. 1), Leipzig 2010, S. 71–90.
[50] Denkbar wäre die Verwendung von *Creative Commons* Lizenzen, vergleiche http://creativecommons.org/ (1. 1. 2013].