EIGENE ERHEBUNG ODER «FERTIGE» DATEN? ZU MÖGLICHKEITEN UND GRENZEN DER VISUELLEN DARSTELLUNG STATISTISCHER DATEN

Katrin Henzel, Stefan Walter

There are two options for data visualization: Either the use of ready-made statistics or one's own data collection. Both approaches have their advantages and disadvantages which will be discussed using the example of the Google service "Ngram" respectively a statistical data collection of autograph books. Whereas the advantage of "Ngram" is seen in its stimulation of further research, self-made statistical data collections offer a variety of visualization options which go far beyond the mere presentation. In this respect, we seek to encourage researchers to collect and analyze their own data.

1. Visualisierungstechniken in der Geschichtswissenschaft

Die Geschichtswissenschaft ist eine Fachdisziplin, die sich seit jeher auch der visuellen Wissensvermittlung bedient.[1] Gerade beim Nachvollziehen umfangreicher und komplexer Zusammenhänge erhöht der Einsatz von Bildmedien die Erkenntnis – man denke etwa an die bereits im Schulunterricht verwendeten Hilfsmittel wie Landkarten[2] oder Zeitstrahl.

Dabei ist der Rückgriff auf die visuelle Sichtbarmachung des Zeit-Raum-Verhältnisses nicht nur in vermittelnder Funktion, sondern auch gegenstandsbezogen nur konsequent: Bedienten sich doch die Geschichtsschreiber von Beginn an einprägsamer (Sprach-)Bilder, um dem Erzählten Lebendigkeit und Tiefe zu verleihen, aber auch, um eine (Deutungs-) Perspektive vorzugeben.[3]

Wert und Funktion der Bilder für Geschichtsschreibung, -wissenschaft und -vermittlung sowie die kritische Hinterfragung scheinbar objektiver Bilder sind in der jüngeren Forschung erkannt und ausgeführt worden.[4] Dies trifft unseres Erachtens jedoch weniger auf den Teilbereich innerhalb der Visual History zu, der die bildhafte Darstellung von Daten und Wissenszusammenhängen als Hilfsmittel oder Ergebnis historischer Forschungen beinhaltet.[5] Woher rührt dieses Defizit? Zwei mögliche Ursachen sind zu vermuten: Erstens ist die Förderung kognitiver Erkenntnis durch bildliche Veranschaulichung mittlerweile so selbstverständlich, dass eine theoretische Auseinandersetzung speziell in diesem Bereich nicht zu lohnen scheint; ja, dass der Einsatz von Kurven und Diagrammen und ähnlichen visuellen Instrumenten in der Geschichtswissenschaft eventuell sogar als unangemessen missverstanden werden könnte.[6] Zweitens besteht ein Defizit an bestimmten methodischen Zugängen zum Quellenmaterial. Erst bei der Untersuchung grosser Datenmengen stellt sich in der Regel die Frage der adäquaten visuellen Veranschaulichung der analysierten Ergebnisse. Die vorherrschende Nichtbeschäftigung mit der Visualisierung von Datenergebnissen kann als ein Indiz für die offensichtlich unterrepräsentierte quantitative Forschung innerhalb der Geschichtswissenschaft gesehen werden. Diese Aussage bezieht sich freilich auf den *allgemeinen* Stand der Geschichtswissenschaft, nicht einzelne Fachdisziplinen. Hier sind bekanntermassen stark quantitativ ausgerichtete Bereiche seit einigen Dezennien vertreten und bereichern die Geschichtswissenschaft um methodische Zugänge und neue Forschungsergebnisse. Zu nennen sind hier in erster Linie die Historische Statistik und die Historische Demographie. Dabei gilt, was Manfred Thaller treffend formuliert, eben nicht nur für die spezifisch quantitativ ausgerichteten Teildisziplinen: «Manche Arten historischer Quellen sind nur quantitativ zu interpretieren. Werden sie dies nicht, entstehen Artefakte.»[7]

Vom Unbehagen gegenüber quantitativen Methoden in der akademischen Geschichtsdidaktik

Obwohl die Bedeutung quantitativer Quellen seit langer Zeit bekannt ist, lässt sich die Vermittlung entsprechender Methoden in der allgemeinen Ausbildung für Historiker an den Hochschulen hingegen weiterhin als defizitär beschreiben. Ein Blick in Einführungen[8] in die Geschichtswissenschaft wirkt ernüchternd: Nur vier von insgesamt zehn untersuchten Einführungen setzen sich überhaupt mit quantitativen Methoden auseinander, und nur eine einzige – Ernst Opgenoorths und Günther Schulz' «Einführung in das Studium der Neueren Geschichte» – führt auch tatsächlich in diesen Bereich umfassend ein[9] und betont den komplementären Charakter quantitativer und qualitativer Verfahren.[10] Ansonsten wird andernorts, wenn überhaupt von quantitativen Methoden die Rede ist, in der Regel lediglich der Dualismus qualitativer und quantitativer Verfahren konstatiert und sogar gelegentlich auf die Gefahren der einseitigen Nutzung quantitativer Methoden hingewiesen.[11] Möglicherweise scheinen statistische Verfahren methodisch für die Geschichtswissenschaft wenig relevant. Die seit dem Historismus methodisch traditionell in den Geisteswissenschaften stark verankerte und damit bevorzugte Anwendung (rein) hermeneutischer Analyseverfahren, insbesondere in der deutschen Geschichtswissenschaft, scheint hierfür ein plausibler Grund zu sein.[12]

Gegenwärtige Tendenzen der Visualisierung in den Sozialwissenschaften

Fragen der Visualisierung von Forschungsergebnissen betreffen heute in gleichem Masse alle mit umfangreichem Datenmaterial, insbesondere quantitativ arbeitenden Fachbereiche. Zu nennen wären die Netzwerkforschung (vorrangig qualitativ arbeitend),[13] die Historische Geographie[14] und die Historische Linguistik. So liesse sich die Aufreihung fortsetzen. Was dabei deutlich wird: Es handelt sich um Spezialgebiete, die in ausserordentlichem Masse als interdisziplinär zu bezeichnen sind und sich in der Regel nur schwer einem wissenschaftlichen Fach klar zuordnen lassen.

Aus diesen Beobachtungen zum Verhältnis der Geschichtswissenschaft zur visuellen Aufbereitung erhobener und analysierter Daten lässt sich die These ableiten, dass Fachbereiche, die von jeher stark interdisziplinär ausgerichtet waren, stärker mit Daten arbeiten, deren Ergebnisse fachübergreifend und allgemeinverständlich zu präsentieren sind, da sie sich zwischen mindestens zwei eigenständigen traditionellen Fächern bewegen.

Die Visualisierung in den Digital Humanities

Ein Blick auf die 2012 in Hamburg stattgefundene DH-Tagung «Digital Diversity: Cultures, Languages and Methods»[15] bestärkt diesen Eindruck. Geistes- und Sozialwissenschaftler haben bezogen auf die Datenerhebung und -auswertung sowie deren Darstellung unabhängig von ihren fachspezifischen Fragestellungen ähnlich gelagerte Probleme zu lösen; und so ist es nur folgerichtig, diese gemeinsam anzugehen.[16] Sinnvoll ist daher die fachübergreifende Bündelung wie beispielsweise in der zum «Network for Digital Methods in the Arts and Humanities» (NeDiMAH) gehörenden Arbeitsgruppe «Information Visualisation».[17] Gerade im Bereich der Textwissenschaften/Editionen werden zunehmend Institutionen gegründet und Portale eingerichtet, die als Austauschplattform dienen, die gemeinsame Entwicklung von Tools und dabei auch die Nutzung von Software für die Visualisierung ermöglichen.[18] Die digitale Verfügbarkeit von Texten macht es nun zunehmend möglich, diese für quantitative Analysen zu nutzen und für die Visualisierung zu erschliessen.

Zwei mögliche Strategien des Historikers

Von unseren Überlegungen zum Gebrauch von Visualisierungstechniken in der Geschichtswissenschaft leiten wir nun den Blick auf zwei Möglichkeiten, die sich dem Historiker im Umgang mit Daten bieten:

1. Eine Möglichkeit besteht darin, auf den vornehmlich hermeneutischen Ansätzen zu verharren. Demgemäss erfolgt ein Rückgriff auf «fertige» Daten lediglich in unterstützender Funktion. Hierbei können grundsätzlich auch Daten von kommerziellen Anbietern wie zum Beispiel Google mit dem «Google Ngram Viewer» herangezogen werden.[19] Die Möglichkeiten der Visualisierung sind dabei allerdings aufgrund fester Vorgaben eingeschränkt. Es lässt sich vermuten (und wird weiter unten zu zeigen sein), dass der Interpretationsspielraum der Daten sowie deren Formen der Visualisierung damit ebenfalls begrenzt sind.
2. Die andere Möglichkeit stellt den unvoreingenommenen Zugriff auf geeignete Methoden der Sozialforschung dar. Wie schon deutlich gemacht wurde, stellen Datenerhebung und -auswertung keine an Einzeldisziplinen gebundenen Methoden dar. Statistische Auswertungsmethoden finden sich sowohl in den Natur- als auch den Geisteswissenschaften. Quantitative und qualitative Methoden sind jeweils auch für historische Fragestellungen nutzbar, ihre Anwendung ist an die Frage der Quellen-/Datenlage gekoppelt.[20] Entsprechend lassen sich die Möglichkeiten der Visualisierung und des Interpretationsspielraums erweitern.

Im Folgenden sollen beide Vorgehensweisen an Beispielen diskutiert werden. Wir vertreten hierbei die Ansicht, dass es für Historiker bei bestimmten Fragestellungen durchaus sinnvoll sein kann, eher eigene statistische Analysen durchzuführen als auf «fertige» Daten zurückzugreifen. Statistische Analysen sind mit einer Vielzahl von Visualisierungsoptionen verbunden, die weit über die Möglichkeiten der Darstellung extern bereitgestellter Daten hinausgehen. Sie dienen nicht nur der unterstützenden Präsentation, sondern zugleich der Analyse des erhobenen Datenmaterials. Dadurch können sie zu überraschenden Erkenntnissen führen und neue Fragestellungen generieren. Historiker sollten deshalb aus unserer Sicht keine Scheu haben, sich auch quantitativer Methoden und der daran geknüpften Visualisierungstechniken zu bedienen.

2. Der «Ngram Viewer» als «Werkzeug des Historikers»?

Am Beispiel des von Google bereitgestellten Dienstes «Ngram Viewer»[21] soll ein typischer Fall extern bereitgestellter Daten, die vom Nutzer also nicht selbst erhoben werden, erörtert werden.[22] An diese Vorstellung knüpft sich die Frage, inwieweit der Ngram-Viewer als ein «Werkzeug des Historikers» gelten kann.[23] Dieses von Google bereitgestellte Tool und dessen Datengrundlage (Google Books) ist mit seinen Voraussetzungen und Funktionen schon umfangreich vorgestellt und zusammengefasst worden.[24] Es sollen trotzdem die wichtigsten Punkte, sofern sie für die Argumentation dienlich sind, noch einmal zusammengefasst werden. Am besten lässt sich dies mithilfe eines Beispiels erläutern. Google Books hatte im Jahr 2011 über 15 Millionen Bücher digitalisiert, von denen *circa* fünf Millionen[25] mit einem Umfang von über 500 Millionen Wörtern aus sieben Sprachen[26] für den Ngram-Viewer genutzt werden.[27]

Eine Abfrage funktioniert dergestalt: «Usage frequency is computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year.»[28] In der Linguistik bezieht sich das N-Gram-Modell auf einzelne Zeichen/-ketten (vor allem Buchstaben) aus einer Sequenz, die die Grundlage für statistische Analysen bilden.[29] Mit seiner Anwendung soll das N-Gram-Modell Vorhersagen über die Umgebung eines Buchstabens errechnen.[30] Der Ngram-Viewer von Google begrenzt die Zerlegung von Sätzen auf den Wortumfang, sodass sich Wörter und Wortgruppen in einem vom Nutzer bestimmbaren Zeitraum abfragen lassen.

Illustrieren wir eine typische Abfrage mit Google Ngram anhand eines Beispiels: In die Maske werden die Autorennamen «Marx» und «Goethe» eingegeben. Die von Google eingestellte Default-Annahme der «Smooth»-Stufe 3 wird auf 0 herabgesetzt,[31] als Korpus wird das deutschsprachige im Zeitraum von 1800 bis 2000 (ebenfalls als Standard voreingestellt) ausgewählt.

Der Ngram-Viewer von Google gibt dabei das relative Vorkommen eines N-Grams («Marx», «Goethe») in einer Auswahl von Publikationen eines Jahres wieder, wobei die berechneten Daten mit Hilfe eines Liniendiagramms visualisiert werden. An «Goethe» lässt sich dabei sehr schön zeigen, dass es hier um einen berühmten Autor geht, bei dem erwartungsgemäss an Jahrestagen deutliche Ausschläge der Kurve aufgrund vermehrter Publikationen festzustellen sind (so beispielsweise 1932, dem 100. Todestag, oder 1949, dem 200. Geburtstag des Dichters). Bei «Marx» sind die «Peaks» ebenfalls vorhanden (beispielsweise im Todesjahr 1883 und genau 100 Jahre später sowie 1972 mit Erscheinen des Probebandes

ABBILDUNG 1
Abfrage der Begriffe «Marx» und «Goethe» im Google-Ngram-Viewer aus dem Korpus «German» im Zeitraum «1800 bis 2000» mit «Smooth»-Stufe 0

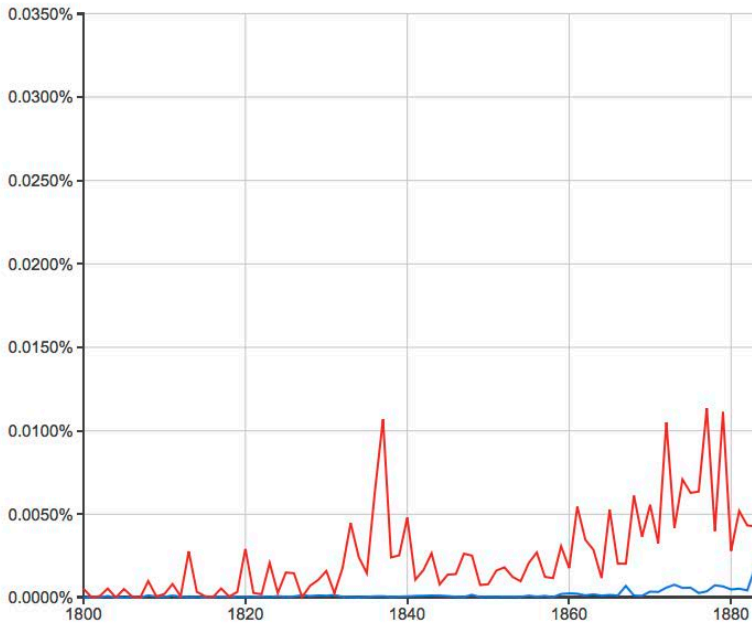
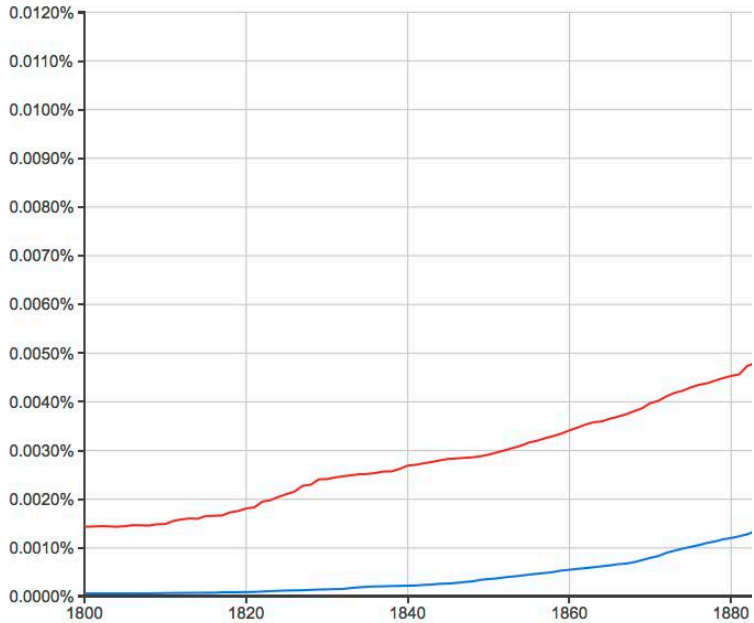
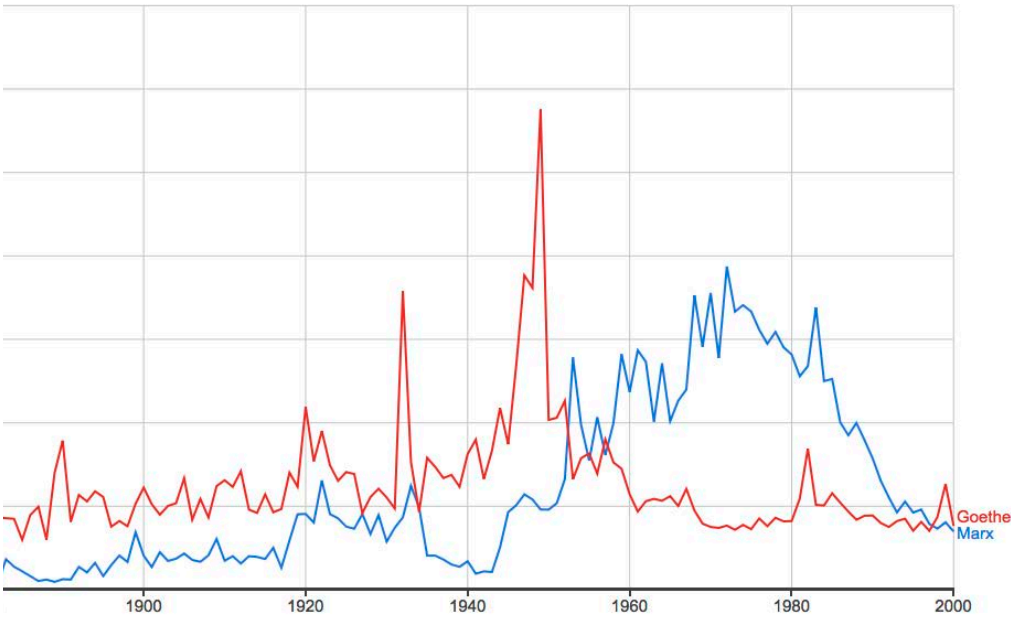
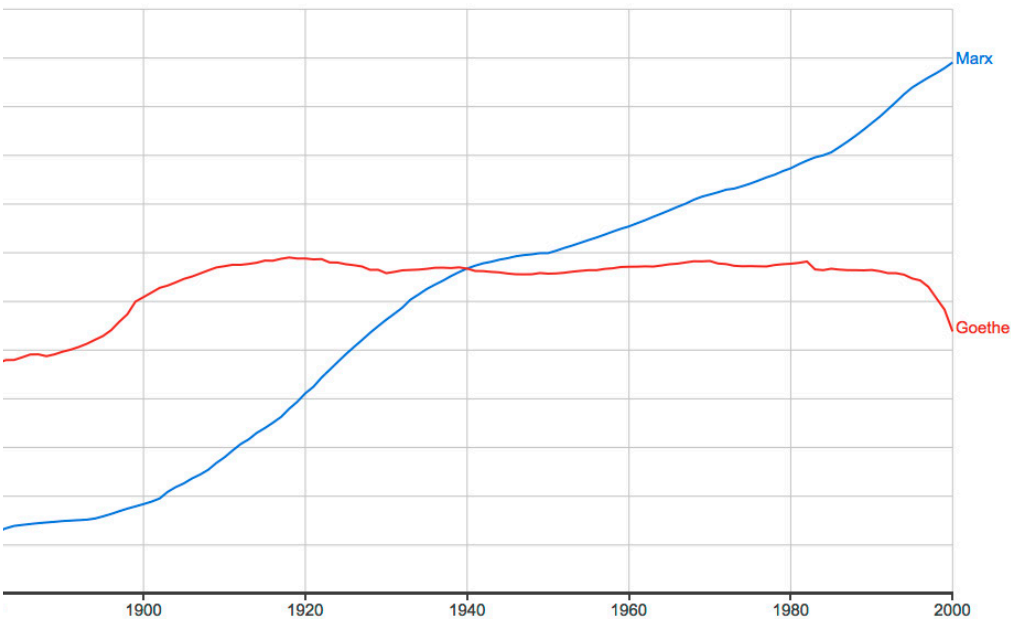


ABBILDUNG 2
Abfrage der Begriffe «Marx» und «Goethe» im Google-Ngram-Viewer aus dem Korpus «German» im Zeitraum «1800 bis 2000» mit «Smooth»-Stufe 50




Goethe
Marx
1900
1920
1940
1960
1980
2000


Marx
Goethe
1900
1920
1940
1960
1980
2000

der deutschen Karl-Marx-und-Friedrich-Engels-Gesamtausgabe), wenngleich in weniger grossen Ausschlägen; hier verläuft die Kurve glatter.

Diese ersten Beobachtungen geben nun Anlass, sich genauer mit den möglichen Gründen für diese Kurven, aber auch den Abfragemodi auseinanderzusetzen.[32] Die Deutung der «Peaks» als Hinweise auf Jubiläen sind zunächst einmal nur plausibel erscheinende Hypothesen, die weiter erforscht werden müssen. Vertiefende Analysen sind allerdings mit dem Google-Programm selbst – zumindest bisher – nicht möglich. So wissen wir nicht, ob ein Mann oder eine Frau, ein Professor oder ein Parteifunktionär, ob ein Publizist aus der DDR oder aus Österreich jeweils der Autor ist, ob das Medium ein wissenschaftlicher Aufsatz in einem Sammelband oder ein fiktionaler Roman ist, ob ein grösserer oder ein kleiner Verlag das Buch gedruckt hat, und so weiter. Insofern ist es auch kaum möglich, mit dem Google Ngram-Viewer Informationen über die Verbreitung von interessierenden N-Grams bei einer Bevölkerungsgruppe zu gewinnen. Denn aus einer Publikation über Marx, etwa in den wortreichen SED-Konvoluten oder in einem abseitigen Artikel, folgt nicht «automatisch» deren breite Rezeption. Auch die vielfach von Michel et al. vorgebrachte Argumentation, der Ngram-Viewer eigne sich zur Identifizierung zensierter Werke,[33] ist aus unserer Sicht nicht haltbar.

Eine weitere Unklarheit liegt in der Korpusauswahl. Derzeit stehen zwei unterschiedliche Korpora, nämlich «German» und «German 2009», zur Verfügung, die den kontinuierlich fortgesetzten Digitalisierungsaktivitäten und dem entsprechenden Anstieg des verfügbaren Materials Rechnung tragen. Worin sich beide Korpora konkret unterscheiden, ist jedoch nicht erkennbar. Auch ist eine Zufallsauswahl der gesamten Texte aufgrund der weitgehend homogenen Herkunft (Universitätsbibliotheken ohne Bestandsaufschlüsselung) und der erfolgten Vorauswahl (keine Zeitschriften) nicht gegeben.[34] Repräsentative Aussagen über den allgemeinen Sprachgebrauch innerhalb einer Gesellschaft sind deshalb nicht möglich.[35] Zudem ist das angewandte OCR-Scanverfahren fehleranfällig und reduziert das Korpus – die Ergebnisse werden verwässert.[36]

Ein weiteres Problem, welches allerdings eher beim Nutzer als beim Datenanbieter liegt, ist der Umstand, dass das Tool eine simple Bedienung suggeriert (die es zweifelsohne auch hat), der Nutzer allerdings Gefahr läuft, «schnelle», also flüchtige und damit fehleranfällige Abfragen durchzuführen. So gibt es bei den Suchbegriffen «Marx» und «Goethe» natürlich auch die naheliegenden Möglichkeiten der Suche samt Vornamen und weiterer Varianten (um beispielsweise Namensvetter auszuklammern). Jede abweichende Abfrage erzeugt dabei auch andere Ergebnisse. Der

Nutzer muss sich also im Klaren sein, wie seine Fragestellungen lauten, wie das Korpus und die Suchoptionen beschaffen sind.[37]

Einen weiteren Kritikpunkt stellt neben der intransparenten Datenauswahl und -bereitstellung die Produktion von «Gefälligkeitsergebnissen» dar. Diese ergeben sich durch die «Smooth»-Funktion. Hierbei wird der Wert X mit X+n und X-n zu einem Mittelwert zusammengefasst. Der Nutzer kann beim «Smoothen» zwischen den Stufen 0 bis 10, dann in Zehnerschritten auf einer Skala bis 50 wählen. Je höher die Stufe ist, desto glatter werden die Ergebnisse und desto flacher die Kurven.

Stellt man in einer erneuten Abfrage mit den Kriterien, die auch für ABBILDUNG 1 gelten («Marx» und «Goethe», Korpus «German», Zeitraum «1800 bis 2000)», nun aber statt der Stufe 0 die höchste Stufe (50) ein, bleibt zwar der grundsätzliche Kurvenverlauf (die Grundtendenz, dass «Goethe» bis Anfang der 1950er Jahre weitaus häufiger vorkommt als «Marx», sich dann «Marx» vor «Goethe» absetzt und die Häufigkeit von «Goethe» ab diesem Punkt relativ konstant bleibt) erhalten. Doch gehen durch die Glättung wichtige Detailinformationen, auf denen sich Interpretationen stützen, verloren: Das wichtige Jahr 1972 (siehe oben) ist nicht mehr als «Peak» zu erkennen, die Kurve verläuft stattdessen kontinuierlich «sanft» nach oben.

Glättungsfunktionen wie die «Smooth»-Funktion sind im Rahmen graphischer Datenanalyse nicht ungewöhnlich. Sie werden angewandt, um verdeckte, bisher unbekannte Strukturen in den erhobenen Daten zu entdecken, wodurch neue Fragestellungen aufgeworfen werden können. Allerdings kann man diese Fragestellungen mit dem Google Ngram-Viewer kaum weiter vertiefen. Vielmehr liegt die Versuchung nahe, mit der «Smooth»-Funktion erwünschte und somit «gefällige» Strukturen, die als Beleg für eine Hypothese dienen könnten, durch Auswahl eines geeigneten Glättungsgrades erst zu erzeugen. Der Nutzen des Tools scheint uns deshalb vor allem in der *Anregung*, nicht aber im *Beleg* für eine Hypothese zu liegen. Möchte man eine aufgeworfene Fragestellung indes weiter vertiefen, reicht ein Rückgriff auf «fertige», aggregierte Daten kaum aus.

3. Eine zweite Strategie: Eigene quantitative Analysen

Damit rückt eine zweite Strategie in den Fokus des Historikers, bei der er selbst geeignete schriftliche Dokumente heranzieht und eigene quantitative Analysen durchführt. Hierbei hat er die Möglichkeit, über die Visualisierung zeitlich gereihter Daten, wie sie der Google Ngram-Viewer letztendlich nur bietet, hinauszugehen. Statistik-Software bietet hierzu einerseits standardmässig eine Vielzahl von Graphikoptionen für eine ansprechende, die Interpretation unterstützende *Präsentation* der Daten.[38] Andererseits offerieren dieselben Statistik-Tools zusätzliche Visualisierungsmöglichkeiten, die allein der *Analyse* und weniger der späteren Präsentation der Daten vorbehalten sind. In diesem Zusammenhang ist zu bemerken, dass fast immer die Beschäftigung mit quantitativ erhobenen Daten zunächst mit einer visuellen Inspektion der erhobenen Merkmale beginnt. Man visualisiert hierzu die Verteilungen, untersucht Häufigkeiten, sucht nach groben, die späteren Schätzungen womöglich beeinflussenden «Ausreissern» in den erhobenen Merkmalen. Über diese univariate Diagnostik mithilfe geeigneter Visualisierungsformen (Histogramme, Boxplots, Fehlerbalken, ...) wird allerdings in der Regel in den späteren Publikationen nicht berichtet. Sie erfüllt üblicherweise lediglich den Zweck zu entscheiden, welche statistischen Berechnungen später durchgeführt werden können. Noch häufiger als die univariaten Darstellungsformen kommen allerdings Streudiagramme (Scatterplots) als Hilfsmittel für die graphische Datenanalyse zum Einsatz. Hierbei werden in einem Diagramm die Wertpaare zweier Variablen x und y gegeneinander abgetragen, so dass visuell eine Punktwolke (x_i, y_i, für i = 1...n) entsteht. Anschliessend wird versucht, mögliche Zusammenhänge zwischen den beiden Variablen zu entdecken.

Wir können im Rahmen dieses Aufsatzes nicht die verschiedenen graphischen Darstellungsmöglichkeiten, die sich im Zusammenhang statistischer Analysen ergeben können, erläutern und verweisen auf die relevante Literatur.[39] Wir wollen jedoch im Folgenden am Beispiel einer quantitativen Dokumentenanalyse zeigen, wie sich *Präsentation* und *Analyse*, tabellarische und visuelle Darstellung auch im Rahmen historisch angelegter Forschung sinnvoll für die Interpretation der Befunde ergänzen lassen. Dabei wollen wir auch auf *Grenzen* der Visualisierung statistischer Daten eingehen, die sich insbesondere beim Einsatz multivariater statistischer Berechnungen ergeben.

Goethe und Marx im Poesiealbum zwischen 1949 und 1989

In der Zeitgeschichtsforschung wird häufig nach der ideologischen Beeinflussung der Bürger in den beiden deutschen Diktaturen des 20. Jahrhunderts gefragt, jedoch auf den Mangel an entsprechenden Studien verwiesen.[40] Eine Möglichkeit, mehr über eine vom Staat erwünschte ideologische Beeinflussung in einer Bevölkerung zu erfahren, könnte aus unserer Sicht darin bestehen, Poesiealben von Heranwachsenden auf Indizien einer ideologischen Beeinflussung hin zu untersuchen. Gemäss einer volkskundlichen Definition handelt es sich bei einem Poesiealbum um ein Buch, «in das Freunde ihren Namen besonders in Verbindung mit einem Spruch und allerlei Auszierden, so Handzeichnungen u.a.m. eintragen».[41] Hervorgegangen aus der ursprünglich unter Erwachsenen verbreiteten Stammbuchsitte, werden Poesiealben seit Mitte des 19. Jahrhunderts hauptsächlich von Heranwachsenden und hierbei insbesondere von Mädchen geführt. Noch heute ist die Sitte vielerorts im deutschsprachigen Raum zu beobachten.[42] Die Auswertung von Poesiealben als Gegenstand staatlicher Beeinflussung kann dabei mit hermeneutischen Methoden erfolgen, aber auch mit quantitativ-statistischen.

Als Einträger in Poesiealben fungieren hauptsächlich gleichaltrige Freunde und Mitschüler, die wir im Folgenden mit Hinweis auf die sozialwissenschaftliche Forschungstradition als Peergroup bezeichnen. Daneben werden jedoch auch Lehrer, Familienangehörige sowie Personen des erweiterten ausserschulischen Bekanntenkreises um Inskriptionen gebeten. Für die Erforschung der Poesiealben folgt hieraus die Forderung, die Zugehörigkeit eines Inskribenten zu einer Einträgergruppe nach Möglichkeit zu klären. Inhaltlich kommen in den eingetragenen Texten oftmals bestimmte Wertvorstellungen zum Ausdruck. Hierzu wird in der Regel auf vorhandene «Sprüche» oder Zitate zurückgegriffen. Diese können zwar der Albumtradition entstammen, werden jedoch häufiger auch ganz anderen Quellen entnommen. Ein Indiz für eine aus Sicht des DDR-Staates als erwünscht erscheinende Beeinflussung könnte der Rückgriff auf einen entsprechend erwünschten Autor in einem Poesieeintrag sein. Dies scheint besonders dann der Fall, wenn durch Nennung des Autorennamens der eingetragene Text als Zitat ausgewiesen wurde. Da wir in der Diskussion des Google Ngram-Viewers nach Marx und Goethe gefragt hatten, wollen wir erneut auf diese Autoren zurückgreifen. Wurde in ein Album ein Marx-Zitat eingetragen, kann dies als Indiz einer staatlich erwünschten Beeinflussung des Einträgerverhaltens interpretiert werden. Wurde ein Eintrag als von Goethe stammend gekennzeichnet, wird dies nicht als staatlich erwünschte Beeinflussung gedeutet, da Goethe vermutlich im gesamten deutschsprachigen Raum allgemeine Wertschätzung geniesst.[43]

Umfang und Zeitraum der Erhebung der Poesiealben

Für die nachfolgenden Analysen greifen wir auf die Untersuchung von 2863 Einträgen in Poesiealben zurück, die im Rahmen eines Dissertationsprojekts im Zeitraum zwischen Mai 2009 und Mai 2011 erhoben wurden.[44] Insgesamt nahmen an der Studie 65 Albumbesitzer teil (58 weibliche und sieben männliche), die zusammen 84 Poesiealben geführt haben. 32 Personen stammen aus der ehemaligen DDR und führten insgesamt 45 Alben. 33 Personen kamen aus den alten Bundesländern und führten insgesamt 39 Alben.[45] Wir möchten darauf hinweisen, dass hierzu Albumhalter aus dem persönlichen Umfeld, über einen Mailverteiler sowie über Archive und Sammler gezielt angesprochen worden sind, so dass hier keine zufallsbasierte, sondern eine bewusste Auswahl der Alben vorliegt. Trotz der hohen Anzahl an untersuchten Einträgen beanspruchen wir deshalb auch keine Repräsentativität unserer Daten. Wir können nur Hinweise auf mögliche Trends geben, die durch repräsentative Erhebungen geprüft werden sollten.

Befund und Visualisierung I: Allgemeine Häufigkeiten

Fragen wir zunächst ganz allgemein, wie häufig in den untersuchten, zwischen 1949 und 1989 geführten Alben Einträge gekennzeichnet als von Marx, Goethe oder einem anderen Autor stammend vorkommen. Wir bilden hierzu eine nominale Variable, die vier Merkmalsausprägungen besitzt: Das verwendete Zitat stammt von 1 = Marx, 2 = Goethe, 3 = einem anderen Autor, 4 = keine Angabe des Autorennamens. TABELLE 1 gibt hierüber Auskunft. Es zeigt sich, dass 3,8% aller untersuchten Einträge als Goethe-Zitate ausgewiesen worden sind, jedoch nur 0,3% als von Karl Marx stammende. 12,2% der Texte wurden als Zitat eines anderen Autors gekennzeichnet.

TABELLE 1
Allgemeine Häufigkeit der Autorenangaben in Poesiealben 1949–1989

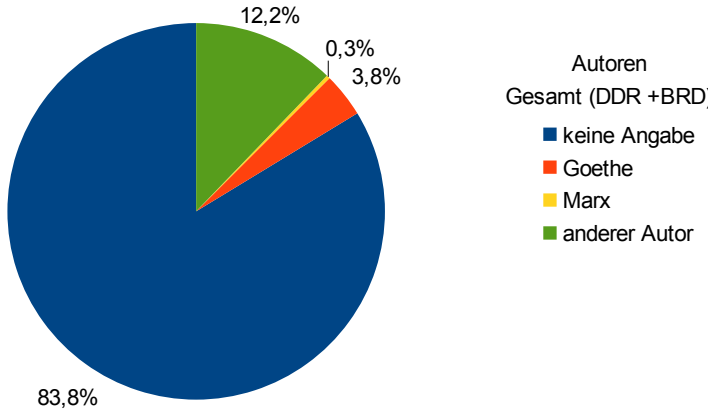
Deskriptive Statistiken; Quelle: hier wie im Folgenden Datenkorpus der erhobenen Poesiealben

ZITATKENNZEICHNUNG	HÄUFIGKEIT	PROZENT
Karl Marx	8	0,3
Joh. Wolfgang Goethe	100	3,8
Anderer Autor	323	12,2
Keine Autorenangabe	2222	83,8
Gesamt	2653	100,0

Da sich die Ausprägungen der Variable sinnvoll auf 100% kumulieren lassen, eignet sich zur visuellen Darstellung ein Kreisdiagramm. Diese Form der Darstellung trägt als Präsentationsgraphik zur Veranschaulichung

der Tabellenwerte bei, indem sie die Aufmerksamkeit zum einen auf den Verzicht der Autorenangabe lenkt (83,8%), zum anderen in der Darstellung die proportional starke Präsenz der Goethe-Zitate im Vergleich zu Marx sowie zu den übrigen zitierten Autoren augenfällig macht.

ABBILDUNG 3
Visuelle Darstellung der Daten aus Tabelle 1



Befund und Visualisierung II: Einträgergruppen im Ost-West-Vergleich

Vertiefen wir nun die Analyse und fragen, ob es Ost-West-Unterschiede bei der Autorenauswahl gibt.[46] Hierzu teilen wir die Einträger nach Inskriptionen in ein Album der ehemaligen DDR beziehungsweise der alten Bundesrepublik auf. TABELLE 2 gibt Auskunft über die erzielten Befunde.

TABELLE 2
Autorenangaben nach Ost und West

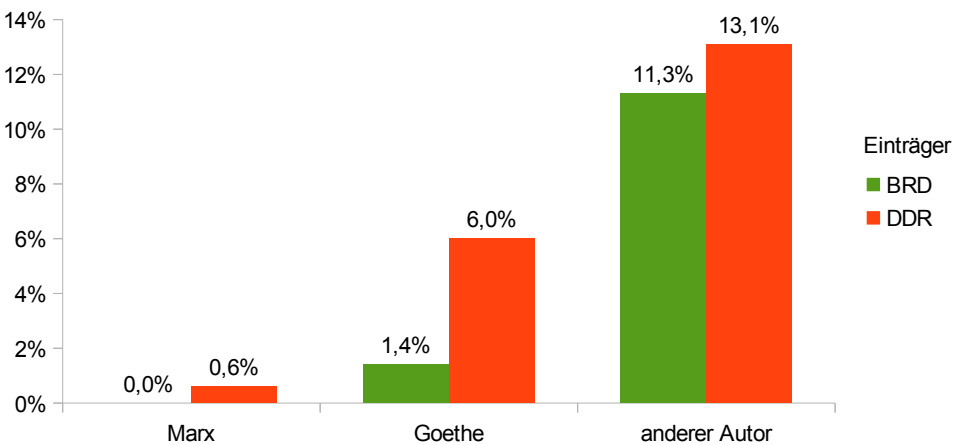
Deskriptive Statistiken, Kreuztabelle; keine Berücksichtigung von Panel-Einträgen

ZITATKENNZEICHNUNG		EINTRÄGER AUS BRD	EINTRÄGER AUS DDR	GESAMT
Joh. Wolfgang Goethe	N	19	81	100
	%	1,4%	6,0%	3,8%
Karl Marx	N	0	8	8
	%	0,0%	0,6%	0,3%
Anderer Autor	N	148	175	323
	%	11,3%	13,1%	12,2%
Keine Autorenangabe	N	1147	1075	2222
	%	87,3%	80,3%	83,8%
Gesamt	N	1314	1339	2653
	%	100,0%	100,0%	100,0%

Die verfeinerte Analyse anhand des biographischen Ost-/West-Hintergrundes eines Einträgers bringt bedeutsame Unterschiede. Marx wird nur von Einträgern in der DDR verwendet. Dies geschieht jedoch sehr selten, was für eine geringe staatliche Beeinflussung in der DDR spricht. Überraschend ist, dass vor allem DDR-Einträger Goethe-Zitate verwendet haben. Man kann daraus schlussfolgern, dass im DDR-Alltag womöglich die Weimarer Klassik der Klassenkampf-Rhetorik vorgezogen wurde.

Die Tabellenwerte können mithilfe eines gruppierten Balkendiagramms, wie in ABBILDUNG 4 geschehen, visuell dargestellt werden. Für die Kategorienachse dient uns dabei unsere nominale Variable. Wir haben sie jedoch für die visuelle Darstellung leicht modifiziert, indem wir auf die Einträger ohne Autorenangabe verzichtet haben. Die Berücksichtigung dieser Balken würde das Gesamtbild der visuellen Darstellung dominieren. Durch das Weglassen wird die Konzentration auf die uns interessierenden Autoren gelenkt, wodurch die Ost-West-Unterschiede in den jeweiligen Kategorien (insbesondere bei Marx und Goethe) eine grössere Betonung erfahren.

ABBILDUNG 4
Visuelle Darstellung der Daten aus Tabelle 2



Befund und Visualisierung III: Multivariate Prüfung der Goethe-Zitate

Man sollte an diesem Punkt die statistische Analyse nicht abbrechen. Vielmehr ist stets eine multivariate Prüfung anzustreben, die weitere relevante Einflussfaktoren berücksichtigt bzw. die bisher berücksichtigten Faktoren auf mögliche Scheinkorrelationen kontrolliert.[47] So kann zum Beispiel plausibel angenommen werden, dass die Beliebtheit der Goethe-Zitate unabhängig von der Staatsangehörigkeit vor allem altersabhängig ist und besonders gern von Lehrern verwendet wurde. Wir haben allerdings bisher nur das Vorkommen in der Gesamterhebung sowie anschliessend das Vorkommen in den Teilstichproben der DDR- und BRD-Alben analysiert. Die hierbei gefundenen Ost-West-Unterschiede könnten dabei allerdings auf unsere Art der Erhebung zurückgehen, die nicht zufallsbasiert ist. Wie Google vornehmlich die Spezialbestände von Universitätsbibliotheken eingescannt hat und womöglich dadurch bestimmte Fachdiskurse über- und andere Populärdiskurse unterrepräsentiert in den Ngrams wiedergibt, könnten auch in unserer Gesamtstichprobe aufgrund der nicht zufallsbasierten Erhebungsweise bestimmte Einträgergruppen und damit ihre Lieblingsautoren über- beziehungsweise unterrepräsentiert vorkommen. Die oben abgebildeten Ost-West-Unterschiede könnten folglich nur scheinbar existieren. Aus diesem Grund sollte stets versucht werden, einen möglichen Zusammenhang multivariat zu kontrollieren, was wir anhand der Einträge mit Goethe-Zitaten demonstrieren möchten.

Die Gewinnung zusätzlicher relevanter Variablen, von denen wir annehmen, dass sie einen Einfluss auf das Zitierverhalten ausüben, ist durch die weitere Dokumentenanalyse, aber auch durch externe Informationsbeschaffung möglich. Die abhängige Variable liegt allerdings lediglich in dichotomer Form vor (ein gekennzeichnetes Goethe-Zitat liegt im Eintrag vor = 1, liegt nicht vor = 0). Als multivariate Analysemethode kommt deshalb für uns die binär-logistische Regressionsanalyse (Logistische Regression) als geeignetes statistisches Verfahren infrage. Dieses Verfahren schätzt bedingte Wahrscheinlichkeiten für das Eintreten eines Sachverhalts (zum Beispiel Goethe-Zitat = 1) auf Basis der Maximum-Likelihood-Methode. Die Schätzung erfolgt dabei unter Verwendung der sogenannten logistischen Funktion, die einen nicht-linearen Zusammenhang zwischen der Eintrittswahrscheinlichkeit der dichotom abhängigen Variable und den unabhängigen Variablen unterstellt.[48] Speziell für unseren Fall schätzen wir die Wahrscheinlichkeit, mit der ein Eintrag eines Goethe-Zitats in Abhängigkeit der uns relevant erscheinenden Einflussfaktoren (Ost-/West-Herkunft, Geschlechtszugehörigkeit, Grösse des Heimatortes als Ausdruck für Stadt-Land-Unterschiede, Zugehörigkeit zu einer Einträgergruppe, Alter und Bildungsgrad des Einträgers) zu erwarten ist.

Die Interpretation der im Rahmen dieses Verfahrens berechneten Koeffizienten ist nicht intuitiv, da ein nicht-linearer Zusammenhang berechnet wird. Zur Abschätzung des Einflusses der einzelnen Faktoren auf die Eintrittswahrscheinlichkeit des interessierenden Sachverhalts (Goethe-Zitat wurde eingetragen) wird üblicherweise auf den «Effektkoeffizienten» zurückgegriffen, der auch als «odd ratio» bezeichnet wird. Ähnlich wie bei einer Pferdewette gibt der Effektkoeffizient für jeden Faktor ein Chancenverhältnis wieder, das jeweils die Chance für das Eintreffen des interessierenden Sachverhalts (Goethe wurde zitiert) ausdrückt. Um möglichst robust erklärende Einflussfaktoren zu erhalten, berechnen wir insgesamt vier Modelle und variieren dabei die Anzahl der einbezogenen Faktoren. Modell 1 bezieht alle Einträger ein und nimmt zunächst folgende relevante Variablen als erklärende Faktoren in die Schätzung auf: Geschlecht des Inskribenten, Jahr des Eintrags, Wohnortgrösse des Einträgers.[49] In Modell 2 kommt eine abhängige Variable hinzu, mit der die Zugehörigkeit eines Inskribenten zu einer bestimmten Einträgergruppe erfasst wurde.[50] In Modell 3 wird zusätzlich zu den bisherigen Faktoren auch die Ost-/West-Herkunft des Einträgers berücksichtigt. Da wir speziell für die Angehörigen der Peergroup auch das Alter eines Inskribenten zum Eintragszeitpunkt annähernd bestimmen können, berechnen wir mit diesem zusätzlichen Faktor ein viertes Modell, welches allerdings nur die Einträger aus der Peergroup berücksichtigt.[51]

TABELLE 3
Logistische Regression: Kennzeichnung eines Goethe-Zitats

UNABHÄNGIGE VARIABLEN		ABHÄNGIGE VARIABLE: ZITAT ALS VON JOHANN WOLFGANG GOETHE STAMMEND GEKENNZEICHNET			
		1	2	3	4
Geschlecht des Einträgers (weiblich = 1)		0,696	0,765	0,843	0,811
Jahr des Eintrags (Kohorteneffekt)		0,995	0,992	1,001	1,004
Ortsgrösse des Einträgers		1,070	1,058	1,068	1,195**
Einträgergruppe	Peer		Ref.**	Ref.**	
	Familie		1,045	1,152	
	Lehrer		4,422**	4,610**	
	Sonstige		1,541	1,786	
Einträger aus DDR/BRD (DDR = 1)				4,773**	11,6**
nur Peer: Alter bei Eintrag					1,057
Pseudo-R^2 (Nagelkerke)		0.007	0.053**	0.112**	0.122**

Logistische Regression; odds ratio, ** signifikant <1%, * signifikant <5%, + signifikant <10%
Modell 2-3: Referenzkategorie Einträgergruppe: Peer (/CONTRAST (Einträgergruppe) = Indicator(1)
Modell 1-3: nur singulärer Eintrag bzw. 1. Eintrag eines Einträgers (Ausschluss der Panel-Einträge)
Modell 4: nur singulärer Eintrag bzw. 1. Eintrag sowie nur Peer-Einträger berücksichtigt

TABELLE 3 gibt die berechneten Effektkoeffizienten sowie in der Fusszeile den berechneten Nagelkerke-Wert der Pseudo-R^2-Statistik wieder. Der Nagelkerke-Wert dient uns als Mass für die Güte der Anpassung des jeweils berechneten Modells an die Empirie. Dieser Wert ist zwar vorsichtig zu interpretieren, denn er erhöht sich schätzungsbedingt mit jedem zusätzlich aufgenommenen Faktor quasi «automatisch». Dennoch lässt er einen groben Vergleich der jeweiligen Modellgüte zu.[52]

Was ist der Tabelle zu entnehmen? Modell 1 ist schlecht angepasst, das drückt sich im sehr geringen Nagelkerke-Koeffizienten aus. Keiner der in Modell 1 einbezogenen Faktoren hat einen signifikanten (und somit systematischen und nicht zufälligen) Einfluss auf die Eintragswahrscheinlichkeit eines Goethe-Zitats ausgeübt. Durch Hinzunahme des Faktors Einträgergruppe in Modell 2 wird die Modellgüte allerdings signifikant verbessert und der Nagelkerke-Koeffizient steigt deutlich. Das bedeutet, dass die Zugehörigkeit zu einer bestimmten Einträgergruppe einen signifikanten Effekt darauf ausgeübt hat, ob ein Goethe-Zitat eingetragen wurde. Im Vergleich zur Peergroup des Halters, die uns hier als Referenzkategorie dient, sind es wie vermutet die Lehrer, die mit signifikant höherer Wahrscheinlichkeit Goethe-Zitate in ihren Albumeinträgen verwendeten. Zwar deutet die Richtung der Effektkoeffizienten darauf, dass auch die Verwandten eines Halters und auch die sonstigen Einträger ebenfalls der Tendenz nach mit höherer Wahrscheinlichkeit Goethe-Zitate eingetragen haben als die Peergroup des Halters, allerdings werden die geschätzten Koeffizienten als nicht signifikant ausgewiesen. Das heisst, Peergroup, Verwandte und Sonstige haben sich hinsichtlich der Wahrscheinlichkeit der Verwendung eines Goethe-Zitats wohl eher nicht unterschieden.

Durch die Hinzunahme des DDR-/BRD-Faktors in Modell 3 verbessert sich erneut der Nagelkerke-Koeffizient deutlich, was auf eine weitere Verbesserung der Erklärungskraft des Modells deutet. Mit weniger als 1% Irrtumswahrscheinlichkeit lässt sich dabei ein systematischer Einfluss des Ost-/West-Faktors annehmen. Die Chance, dass ein DDR-Einträger Goethe eingetragen hat, war im Vergleich zu einem BRD-Einträger etwa 4,8-fach höher. Damit werden die bereits aus der Deskription bekannten DDR-BRD-Unterschiede auch unter multivariaten Bedingungen erhärtet. Auch die Annahme, dass Lehrer bevorzugt Goethe-Zitate eingetragen haben, wird in Modell 3 weiter gestützt. Für alle weiteren Faktoren lässt sich hingegen erneut kein systematischer Einfluss auf die Wahrscheinlichkeit eines Goethe-Zitats erkennen. Modell 4, welches schliesslich nur die Peergroup des Halters berücksichtigt, zeigt auf, dass sich Mitschüler bzw. gleichaltrige Freunde eines Albumhalters in Ost und West bezüglich der Verwendung eines Goethe-Zitats deutlich unterschieden haben. Die

Chance, dass ein Jugendlicher in der DDR Goethe zitiert hat, war 11,6-fach höher als bei einem Jugendlichen in der Bundesrepublik. Zudem deuten sich in Modell 4 leichte Stadt-Land-Unterschiede an. Je grösser der Ort, umso wahrscheinlicher hat ein Peer Goethe zitiert.

Multivariate Analysen sind notwendig, um statistische Zusammenhänge durch Einbezug anderer denkbarer Einflussfaktoren zu kontrollieren. Allerdings stösst hier die Visualisierung an ihre Grenzen. Zwar gibt es die technische Möglichkeit, mehr als zwei Variablen in einem Diagramm darzustellen, allerdings tritt schnell eine Überforderung bei der Interpretation des Dargestellten ein.[53] Rainer Schnell formuliert das Problem, das bei der Darstellung mehrdimensionaler Daten eintritt, wie folgt: «Das zentrale Problem der Darstellung multivariater Daten besteht nicht darin, möglichst viele Variablen ‹irgendwie› simultan darzustellen, sondern eine Darstellung zu finden, die kognitiv verarbeitbar ist.»[54] Es sind also vor allem die kognitiven Grenzen menschlicher Wahrnehmungsfähigkeit, die einer Visualisierung multivariater Daten hinderlich sind. Hier bleibt man somit wie in unserem Zusammenhang auf die Interpretation der geschätzten Koeffizienten in der Tabelle angewiesen.

Befund und Visualisierung IV: Empfohlene Visualisierung bei Prüfung auf Interaktionseffekte

Wir wollen jedoch abschliessend zeigen, dass es insbesondere bei der Prüfung von vermuteten Interaktionseffekten mithilfe einer Logistischen Regression für die sinnvolle Interpretation sogar erforderlich erscheint, die berechneten Wahrscheinlichkeiten zu visualisieren.[55] Wir wollen dies abschliessend an einem Beispiel illustrieren.

Modell 4 von TABELLE 3 kann entnommen werden, dass DDR-Jugendliche mit deutlich grösserer Wahrscheinlichkeit Goethe-Zitate inskribiert haben als westdeutsche Jugendliche. Allerdings könnten wir vermuten, dass sich dies im Verlauf der Adoleszenz geändert habe. So könnten westdeutsche Jugendliche im Verlauf der Adoleszenz Goethe zunehmend für sich entdeckt und dies auch immer häufiger in den Alben durch Zitatkennzeichnung bekundet haben. Demgegenüber könnten sich ostdeutsche Jugendliche im Verlauf der Adoleszenz zunehmend anderen Autoren zugewandt haben. Wir vermuten also eine intervenierende Wirkung des Alters auf das Eintragsverhalten in DDR und BRD, kurz eine Interaktion zwischen Alter und Ost-/West-Herkunft.

Um diese Interaktion zu prüfen, haben wir im Rahmen einer Logistischen Regression zwei Modelle berechnet. Zunächst haben wir in einem ersten Modell nur die Faktoren Ost-/West-Hintergrund und Alter beim Eintrag berücksichtigt. Im zweiten Modell wurde zusätzlich ein

multiplikatives Interaktionsterm aus beiden Faktoren – Herkunft und Alter – in die Schätzung einbezogen.[57] TABELLE 4 gibt die Befunde der Schätzungen wieder:

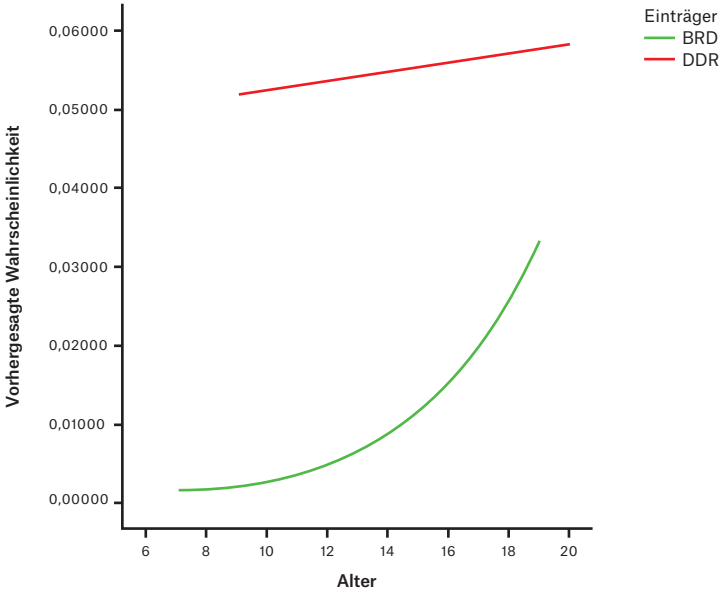
TABELLE 4
Goethe-Zitat nach Alter in DDR und BRD

UNABHÄNGIGE VARIABLEN	ABHÄNGIGE VARIABLE	
	1	2
Peer aus DDR/BRD (DDR = 1)	11,947**	11,029**
Alter des Peer	1,033	1,309
DDR/BRD* Alter (zentriert)		0,542
Pseudo-R^2 (Nagelkerke)	0,095**	0,099**

Logistische Regression; odd-ratios, ** signifikant <1%, * signifikant <5%, + signifikant <10%
Filter: nur singulärer Eintrag bzw. 1. Eintrag und nur Einträge von Peer

Der Tabelle kann entnommen werden, dass nur der bereits bekannte starke Effekt des Ost-/West-Hintergrundes einen signifikanten und somit systematischen Einfluss ausübt. Die Pseudo-R^2-Statistik (Nagelkerke) verbessert sich nur geringfügig bei Hinzunahme des Interaktionsterms in Modell 2. Das Interaktionsterm (DDR/BRD*Alter) wird auch nicht als signifikant ausgewiesen; laut Tabelle deutet also einiges darauf hin, dass zwischen Alter und Hintergrund keine Interaktion vorliegt. Allerdings ist darauf hinzuweisen, dass die Deutung von Effektkoeffizienten von Interaktionstermen nicht eindeutig ist, so dass sich immer auch eine *Visualisierung* der vorgesagten Wahrscheinlichkeit empfiehlt. Diese ist mit einem Streudiagramm (Scatterplot) problemlos möglich, da lediglich zwei Variablen – Alter (x) und vorhergesagte Wahrscheinlichkeit des Goethe-Zitats (y) – abgebildet werden, wobei jedoch die Ost-/West-Herkunft in der Darstellung berücksichtigt wird. ABBILDUNG 5 gibt den entsprechenden Scatterplot wieder. Ursprünglich erfolgte die Darstellung in Form der Wertpaare, also in Punktform. Um den Zusammenhang zwischen den beiden Variablen aber noch deutlicher hervortreten zu lassen, haben wir die Wertpaare durch eine Interpolationslinie miteinander verbunden.

ABBILDUNG 5
Wahrscheinlichkeit eines gekennzeichneten Goethe-Zitats nach Alter in Ost/West[58]



Was kann der Abbildung entnommen werden? Die Wahrscheinlichkeit, dass ein Ost-Jugendlicher ein Goethe-Zitat verwendete, war – wie auch die Effektkoeffizienten in der Tabelle auswiesen – stets höher als bei einem West-Jugendlichen. Allerdings haben sich gemäss der Abbildung die Wahrscheinlichkeiten für Ost- bzw. West-Jugendliche auf sehr verschiedene Weise entwickelt. Dies geht *erst* aus der Visualisierung und nicht aus den geschätzten Effektkoeffizienten der Tabelle hervor. Bei den Ost-Jugendlichen stieg die Wahrscheinlichkeit zwischen dem *circa* neunten und 19. Lebensjahr nur sehr moderat und dabei eher linear an. Bei den West-Jugendlichen stieg die Wahrscheinlichkeit eines Goethe-Zitats im selben Lebensabschnitt von einem sehr geringen Ausgangsniveau viel stärker und zugleich nicht-linear an. Das deutet auf ziemlich unterschiedliche Rezeptionsweisen von Goethe durch die Jugendlichen während ihrer Schulzeit hin. Hier deckt also erst die Visualisierung Effekte auf, die zu neuen Fragen führen. Die Funktion der Visualisierung geht hier über die Präsentation oder Veranschaulichung hinaus. Sie wird hier zum notwendigen Analyseinstrument und ermöglicht erst eine vertiefende Interpretation.

4. Resümee

Mit unserem Beitrag haben wir versucht, auf folgende Punkte aufmerksam zu machen:

1. Die Diskussion des Einsatzes von Visualisierungstechniken in der Geschichtswissenschaft zeigt, dass Visualisierungstechniken eher dann eingesetzt werden, wenn auch quantitative Methoden Anwendung finden.
2. Obschon deren Anwendung nicht an eine spezifische Wissenschaftsdisziplin gebunden ist, wird auf quantitative Methoden – abgesehen von einigen wenigen Forschungsfeldern – in der Geschichtswissenschaft allerdings noch zu selten zurückgegriffen.
3. Um auf die Visualisierung grosser Datenmengen nicht verzichten zu müssen, kann ein Historiker auf verarbeitete («fertige») Daten zurückgreifen und diese bildhaft darstellen. Als ein relevanter kommerzieller Anbieter von «fertigen» Daten und Visualisierungen kann der Ngram-Viewer von Google angesehen werden. Über die Funktion der Präsentation der erzeugten zeitlich gereihten Häufigkeiten kommt der Google Ngram-Viewer (derzeit) nicht hinaus. Den Nutzen dieses Google-Dienstes sehen wir deshalb vornehmlich in der Anregung für weitere Forschungsanstrengungen.
4. Eine andere Möglichkeit besteht darin, eigene quantitativ-statistische Auswertungen vorzunehmen. Hier eröffnet sich dem Historiker eine weitaus grössere Palette an Visualisierungsmöglichkeiten. Am Beispiel der Analyse von Poesiealben haben wir dabei gezeigt, dass zum einen auch im Rahmen statistischer Analysen die Graphiken häufig die Funktion der Präsentation übernehmen. Daten, die sich auch tabellarisch darstellen lassen, werden hierbei in geeignete Diagrammtypen visuell umgesetzt und unterstützen die Interpretation. Es wurde weiterhin darauf aufmerksam gemacht, dass multivariate statistische Prüfungen sinnvoll sind, deren Visualisierung allerdings an kognitive Grenzen des Betrachters stossen. Schliesslich haben wir am Beispiel der Darstellung eines nicht-linearen Zusammenhangs gezeigt, dass Visualisierungen im Rahmen statistischer Datenanalyse auch die Funktion der Analyse übernehmen.

1 Verweisen möchten wir in diesem Zusammenhang beispielhaft auf Kathrin Maurers Studie zur Bedeutung des Bildes (Panorama) und von Bildmedien (Photographie, Buchillustration und historische Kartographie) im deutschen Historismus: Kathrin Maurer, Visualizing the past. The power of the image in German Historicism (Interdisciplinary German Cultural Studies, Bd. 13), Berlin 2013.

2 Dass das Lesen der Karten verschiedene Grundfertigkeiten voraussetzt (insbesondere räumliche Vorstellung, Mathematik, geographisches Wissen) und unter fachlicher Anleitung erlernt werden muss, macht folgende Studie deutlich: Michael Sauer, Zur «Kartenkompetenz» von Schülern. Ergebnisse einer empirischen Untersuchung, in: Geschichte in Wissenschaft und Unterricht 61/4, 2010, S. 234–248. Sauers Resümee zum Einsatz von Kartenmaterial im Unterricht fällt dabei eher nüchtern aus: Karten würden als Arbeitsinstrumente zu wenig genutzt und/oder die Kompetenz der Schüler überschreiten, vgl. S. 244.

3 Der Gebrauch von Bildern als sprachlichen Konstrukten innerhalb von Narrativen, auf den hier lediglich am Rande aufgrund seiner traditionellen Rolle in der Geschichtsschreibung und seiner mnemotechnischen und erkenntnisfördernden Eigenschaften hingewiesen wird, ist vom Bildbegriff abzugrenzen, wie wir ihn im vorliegenden Beitrag verwenden wollen: als Techniken der Visualisierung zwecks Darstellung und Erläuterung komplexer Zusammenhänge innerhalb der Wissenschaften. Zur Ergänzung sei an dieser Stelle auf das Bild in einer dritten Bedeutung verwiesen: als Objekt/Quelle der Geschichtswissenschaft. Zur Rolle der (Bild-)Kunst in der und für die Geschichts-, Kunst- und Kulturwissenschaften s. Bernd Roeck: Visual turn? Kulturgeschichte und die Bilder, in: Geschichte und Gesellschaft 29/2, 2003, S. 294–315.

4 Siehe hierzu insbesondere die Berichte der Sektionen 5.7 und 5.13 des 46. Deutschen Historikertags in: Clemens Wischermann, Armin Müller, Rudolf Schlögl, Jürgen Leipold (Hg.), GeschichtsBilder. 46. Deutscher Historikertag vom 19. bis 22. September in Konstanz. Berichtsband, Konstanz 2007. Bezogen auf die kritische Prüfung scheinbar objektiver Bilder ist massgeblich: David Gugerli, Barbara Orland (Hg.), Ganz normale Bilder. Historische Beiträge zur visuellen Herstellung von Selbstverständlichkeit (Interferenzen, Bd. 2), Zürich 2002. Zur potentiellen Kraft der Veränderung realer Gegebenheiten durch Bilder s. Peter Geimer (Hg.), Ordnungen der Sichtbarkeit. Fotografie in Wissenschaft, Kunst und Technologie, Frankfurt am Main 2002, sowie Ana Karaminova, Wirklichkeit und Fiktion der Bilder, in: Dies., Martin Jung (Hg.), Visualisierungen des Umbruchs. Strategien und Semantiken von Bildern zum Ende der kommunistischen Herrschaft im östlichen Europa, Frankfurt am Main 2012, S. 21–45.

5 «Der Begriff Visual History umschreibt […] drei Ebenen: die Erweiterung der Untersuchungsobjekte der Historiker in Richtung der Visualität von Geschichte und der Historizität des Visuellen, das breite Spektrum der Erkenntnismittel im Umgang mit visuellen Objekten sowie schliesslich die neuen Möglichkeiten der Produktion und Präsentation der Forschungsergebnisse.» Gerhard Paul, Von der Historischen Bildkunde zur Visual History. Eine Einführung, in: Ders. (Hg.), Visual History, Göttingen 2006, S. 7–36, S. 27. Um die dritte Ebene geht es uns hier.

6 Es sei an dieser Stelle auf ein Podiumsgespräch von Christopher Clark im Rahmen der Frankfurter Buchmesse am 10. Oktober 2013 in der Deutschen Nationalbibliothek Frankfurt am Main verwiesen: In diesem hatte Clark bezüglich der komplexen Struktur seiner Publikation «Die Schlafwandler» auf selbst erstellte Diagramme verwiesen, die ihm bei der Systematisierung als Instrument dienten, die er aber vor der Veröffentlichung wieder entfernt hatte, um sich bei den Kollegen «nicht zu blamieren».

7 Manfred Thaller, Historische Datenbanken. Vorteile und Probleme, in: Geschichte und Informatik (= Histoire et informatique), 11, 2000, S. 7–24, S. 17, PURL: http://dx.doi.org/10.5169/seals-8924). Noch immer grundlegend für die quantitative Arbeit des Historikers ist Manfred Thaller, Numerische Datenverarbeitung für Historiker. Eine praxisorientierte Einführung in die quantitative Arbeitsmethode und in SPSS (Materialien zur Historischen Sozialwissenschaft, Bd. 1), Wien 1982.

8 Hierfür wurde nur eine Auswahl an Einführungen, die jedoch durchaus als etablierte Kompendien oder doch wenigstens als einschlägige Empfehlungen für Studienanfänger angesehen werden können, berücksichtigt (in alphabetischer Reihung): Ahasver von Brandt, Werkzeug des Historikers. Eine Einführung in die Historischen Hilfswissenschaften (Kohlhammer Urban Taschenbücher, Bd. 33), 18. Auflage, Stuttgart 2012; Birgit Emich, Geschichte der Frühen Neuzeit studieren, Konstanz 2006; Nils Freytag, Wolfgang Piereth, Kursbuch Geschichte. Tipps und Regeln für wissenschaftliches Arbeiten (UTB, Bd. 2569), 4., aktualisierte Aufl., Paderborn 2009; Hans-Werner Goetz, Proseminar Geschichte. Mittelalter (UTB, Bd. 1719), 3., überarbeitete Auflage, Stuttgart 2006; Rosmarie Günther, Einführung in das Studium der Alten Geschichte (UTB, Bd. 2168), 3., überarbeitete und aktualisierte Auflage, Paderborn 2009; Martha Howell, Walter Prevenier, Werkstatt des Historikers. Eine Einführung in die historischen Methoden (UTB, Bd. 2524), hg. von Theo Kölzer, Köln 2004; Martin Lengwiler, Praxisbuch Geschichte. Einführung in die historischen Methoden (UTB, Bd. 3393),

Zürich 2011; Ernst Opgenoorth, Günther Schulz, Einführung in das Studium der Neueren Geschichte (UTB, Bd. 1553), 7., vollständig neu bearbeitete Auflage, Paderborn 2010; Anette Völker-Rasor (Hg.), Frühe Neuzeit, 3. Auflage, München 2010; Barbara Wolbring, Neuere Geschichte studieren, Konstanz 2006.

9 Lediglich Opgenoorth/Schulz (Anm. 8, ebenda auch alle folgenden Genannten) behandeln die Grundlagen der Statistik inklusive graphischer Darstellungsformen ausführlich, s. das Kapitel «Serien von Daten» oder auch «Der Umgang mit Serien von Daten», das seit der 3. Auflage aus dem Jahr 1989 (seit Aufnahme in die UTB-Reihe, zu der Zeit ist Opgenoorth noch alleiniger Verfasser) vorzufinden ist. Wolbring äussert sich immerhin in einem Überblick (der Anlage des Buchs geschuldet) kurz dazu, gleichermassen Goetz, S. 316–318 als auch Howell/Prevenier. Keine Bedeutung kommt den quantitativen Methoden in den Einführungen von Emich, Günther, Lengwiler sowie Völker-Rasor zu; v. Brandt und Freytag/Piereth legen bewusst den Akzent ausschliesslich auf die hermeneutische Methode.

10 S. Opgenoorth/Schulz (Anm. 8), S. 36–37; S. hierzu auch Kersten Krüger, Historische Statistik, in: Hans-Jürgen Goertz (Hg.), Geschichte. Ein Grundkurs, Reinbek 1998, S. 59–82, S. 77.

11 So zum Beispiel Howell/Prevenier (Anm. 8), S. 68. Der Aspekt der Abschreckung spielt womöglich auch eine Rolle bei der Gestaltung der Einführungsseminare, in denen quantitative Analysemethoden in der Regel selten oder gar nicht thematisiert werden. Eine genaue Untersuchung zur Praxis von Einführungsseminaren in der Geschichtswissenschaft brächte hierüber Klarheit.

12 Ein weiterer Vorbehalt gegenüber der Anwendung quantitativer Methoden in der Geschichtswissenschaft könnte innerhalb des Fachs darin gesehen werden, dass es sich hierbei nicht um genuin geschichtswissenschaftliche, sondern eher um eine aus der Soziologie stammende Analysestrategie handelt. Zum Wechselverhältnis von Geschichte und Soziologie s. (eine Auswahl): Thomas Mergel, Geschichte und Soziologie, in: Hans-Jürgen Goertz (Hg.), Geschichte. Ein Grundkurs, Reinbek 1998, S. 621–651; Nina Baur, Was kann die Soziologie methodisch von der Geschichtswissenschaft lernen?, in: Historical Social Research (= Historische Sozialforschung) 33, 2008, S. 217–248. Heinrich Best, Führungsgruppen und Massenbewegungen im historischen Vergleich. Der Beitrag der Historischen Sozialforschung zu einer diachronen Sozialwissenschaft, Köln 2008.

13 S. hierzu insbesondere den Sammelband: Michael Schönhuth, Markus Gamper, Michael Kronenwett, Martin Stark (Hg.), Visuelle Netzwerkforschung. Qualitative, quantitative und partizipative Zugänge, Bielefeld 2013.

14 S. hierzu den kompakten und informationsreichen Artikel von Klaus Fehn, der auch mit einigen Vorurteilen der (allgemeinen) Geschichtswissenschaft gegenüber der Methodik der Historischen Geographie aufräumt: Klaus Fehn, Historische Geographie, in: Hans-Jürgen Goertz (Hg.), Geschichte. Ein Grundkurs, Reinbek 1998, S. 394–407.

15 S. URL: http://www.dh2012.uni-hamburg.de/ (Zugriff am 6.10.2013). Ein Blick in die umfangreiche Publikation der Konferenz-Abstracts verdeutlicht das Bedürfnis, Daten zu visualisieren, eindrucksvoll. Jan Christoph Meister (Hg.), Digital Humanities 2012. Conference Abstracts. University of Hamburg, Germany, July 16–22, 2012, Hamburg 2012 (PURL: http://hup.sub.uni-hamburg.de/HamburgUP/DH2012_Book_of_Abstracts). Jedoch steht die Thematisierung der Datenvisualisierung noch relativ am Anfang, wie die eher geringe Anzahl an einschlägigen Vorträgen hierzu zeigt.

16 Dieser Vorgang wird in den Medien und den einzelnen Fachdisziplinen nicht selten vermengt mit Vorwürfen der Beliebigkeit und Auflösung fachbezogener Fragestellungen, um die Angst vor den Digital Humanities als unmittelbare Bedrohung der etablierten hermeneutischen Einzelforschungen in den Geisteswissenschaften zu schüren.

17 URL: http://www.nedimah.eu/workgroups/information-visualisation (Zugriff am 6.10.2013).

18 Beispielsweise COST Action Interedition (URL: http://www.interedition.eu/), DARIAH (URL: http://www.dariah.eu/) oder Voyant Tools (http://docs.voyant-tools.org/). (Zugriff jeweils am 6.10.2013).

19 Philipp Sarasin beispielsweise plädiert für einen vorurteilsfreien Umgang mit diesem Tool und hebt das Charakteristische hervor: das Wecken des Spieltriebs des Historikers, aus dem sich überhaupt erst neue Perspektiven und Fragestellungen ergeben können. Vgl. Philipp Sarasin, Sozialgeschichte vs. Foucault im Google Books Ngram Viewer. Ein alter Streitfall in einem neuen Tool, in: Pascal Maeder, Barbara Lüthi, Thomas Mergel (Hg.), Wozu noch Sozialgeschichte? Eine Disziplin im Umbruch. Festschrift für Josef Mooser zum 65. Geburtstag, Göttingen 2012, S. 151–174.

20 Die immer behauptete Trennung in «rein» quantitative (statistische) und qualitative (interpretierende) Methoden rührt eher aus der Konkurrenz von Theorieansätzen, die sich damit stärker begründen lassen, als von den Methoden selbst. Stets ist im Prozess der Anwendung «quantitativer» Methoden eine Vielzahl von Interpretationen notwendig. Umgekehrt gilt dies auch für qualitative Methoden, wenn etwa Interpretationen auf impliziten Häufigkeitsauszählungen von Sinngehalten basieren.

21 URL: http://books.google.com/ngrams (Zugriff am 12.10.2013).

22 Die gängige der beiden Ausgangssituationen für die Visualisierung von Daten ist diejenige, dass der Wissenschaftler die Daten selbst erhebt und analysiert, statt sie zu übernehmen. Achim Bühl, SPSS 14. Einführung in die moderne Datenanalyse, 10., überarbeitete und erweiterte Auflage, München 2006, S. 726.

23 Die Paraphrase bezieht sich natürlich auf die gleichnamige Einführung in die Historischen Hilfswissenschaften von Brandts (Anm. 8). Diese geht allerdings in keiner Weise auf analytische Verfahren in der Geschichtswissenschaft ein. Die Provokation zielt damit auch in diese Richtung.

24 Massgeblich sind die Darlegungen der Mitarbeiter selbst: Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden, Material and Methods. Supporting Online Material for Quantitative Analysis of Culture Using Millions of Digitzed Books, veröffentlicht am 16. Dezember 2010 auf Science Express (DOI: 10.1126/science.1199644); Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden, Quantitative Analysis of Culture Using Millions of Digitized Books (DOI: 10.1126/science.1199644), in: Science 331/176, 2011, S. 176–182. Hilfreich darüber hinaus sind die Darstellungen aus Historikerperspektive von Sarasin (Anm. 19), vor allem S. 154; Peter Haber, Digital Past. Geschichtswissenschaft im digitalen Zeitalter, München 2011, S. 113–115.

25 2011 betrug die konkret ermittelte Zahl 5.195.769 Bücher, Michel et al., Quantitative Analysis of Culture (Anm. 24), S. 176.

26 Englisch (361 Mrd. Wörter), Französisch und Spanisch (je 45 Mrd. Wörter), Deutsch (37 Mrd. Wörter), Chinesisch (13 Mrd. Wörter), Russisch (35 Mrd. Wörter), Hebräisch (2 Mrd. Wörter). Michel et al., Quantitative Analysis of Culture (Anm. 24), S. 176.

27 2013 war die Zahl der von Google digitalisierten Bücher bereits doppelt so hoch, s. Robert Darnton, The National Digital Public Library is launched! in: The New York Review of Books, 25. April 2013. URL: http://www.nybooks.com/articles/archives/2013/apr/25/national-digital-public-library- launched/ (Zugriff am 24.3.2014). Leider waren hierzu keine aufgeschlüsselten Angaben zu Wort- und Sprachumfang und Weiterverwendung für den Ngram-Viewer zu finden, sodass wir uns hier nur auf die 2011 publizierten Informationen beziehen können.

28 Michel et al., Quantitative Analysis of Culture (Anm. 24), S. 176.

29 «An *N-gram* is a subsequence of *N* items from a given sequence. *N*-grams are widely used in statistical natural language processing. *N*-grams containing 1, 2, 3, 4, or *N* characters are referred to as a *unigram*, a *bigram*, (or a *digram*), a *trigram*, and an *N*-gram, repectively. […] the source could be a text written in a natural language, a continous information source (e.g., speech), represented in a discrete manner, or an abstract stochastic process which produces a sequence of symbols.» Alexander Bolshoy, Zeev (Vladimir) Volkovich, Valery Kirzhner, Zeev Barzily, Genome Clustering. From linguistic models to classification of genetic texts (Studies in Computational Intelligence, Bd. 286), Berlin 2010, S. 26. Dort findet sich auch der Hinweis, dass der Begriff des N-Grams von C. E. Shannon (1948) eingeführt wurde, der möglicherweise auch die erste Anwendung durchgeführt hat.

30 «[T]he *N*-gram model is intended to predict the identity of the next letter», Bolshoy et al. (Anm. 29), S. 26.

31 Die «Smooth»-Funktion wird weiter unten noch erläutert. Wichtig für obigen Zusammenhang ist das Ausschalten der Funktion (Stufe 0), um die abgefragten Daten möglichst in ihrem ursprünglichen Zustand zur Verfügung zu haben.

32 Genau dieses Argument, den Wissenschaftler erst auf neue Wege und zu neuen Fragestellungen zu führen, betont auch Philipp Sarasin in seinem Beitrag, s. Sarasin (Anm. 19).

33 Zum Beispiel in Michel et al., Quantitative Analysis of Culture (Anm. 24), S. 181.

34 «Most books were drawn from over 40 university libraries around the world», Michel et al., Quantitative Analysis of Culture (Anm. 24), S. 176. Die in diesem Abschnitt folgenden Verweise beziehen sich alle auf diese Seite.

35 Zum Problem der Repräsentativität s. Andreas Diekmann, Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, 4. Auflage, Reinbek 2010, S. 373–374; sowie Rainer Schnell, Paul B. Hill, Elke Esser, Methoden der empirischen Sozialforschung, 7. Auflage, München 2008, S. 304–305.

36 Um eine höchstmögliche Datenqualität zu gewährleisten, wurden Bände von geringer

OCR-Qualität ausgeschlossen, s. dazu und auch zu den folgenden Angaben Michel et al. Materials and Methods (Anm. 24), S. 7. (Man kann erahnen, was dies etwa für DDR-Verlagsprodukte zu Zeiten des knappen Papierkontingents hinsichtlich der Datenauswahl für den Ngram-Viewer bedeutet.) Für Sprachen mit lateinischem Alphabet wurde die Qualitätsgrenze auf ein Minimum von 80% gesetzt. Bei den chinesischen und russischen Büchern hingegen wurde aufgrund des geringen Materialumfangs kein OCR-Filter eingesetzt. Bei den hebräischen wurde der Filter auf 50% herabgesetzt. Für die englischen wurde bei US- und UK-spezifischen Korpora die Auswahl-Grenze ebenfalls herabgesetzt (auf 60%). Dies erschwert eine Vergleichbarkeit der Daten.

37 Mit den jüngsten Neuerungen kann der Nutzer zudem grammatische Veränderungen (konkret: Flexionen) in der Suche berücksichtigen und als Kurven anzeigen; auch wurde die Unterscheidung in Schreibvarianten (konkret: Gross-/Kleinschreibung) auf Nutzerseite vereinfacht, s. URL: http://techcrunch.com/2013/10/17/google-updates-ngram-viewer-with-improved-search-tools/ (Zugriff am 19.12.2013).

38 S. die umfangreichen Visualisierungstechniken von Statistiksoftware, wie etwa von SPSS, so in: Bühl, SPSS 14 (Anm. 22), S. 973–1020; sowie Felix Brosius, SPSS 21, Heidelberg u.a. 2013, S. 893–1008.

39 Rainer Schnell, Graphisch gestützte Datenanalyse, München 1994; sowie Horst Degen, Graphische Datenexploration, in: Christof Wolf, Henning Best (Hg.), Handbuch der sozialwissenschaftlichen Datenanalyse, Wiesbaden 2010, S. 91–116.

40 So Heinz-Elmar Tenorth, Politische Okkupation in der Schule und die Eigenlogik von Bildungsprozessen. Erziehung und Bildung im Transformationsprozess, in: Humboldt-Spektrum 6/3, 1999, S. 38–43.

41 Alfred Fiedler, Vom Stammbuch zum Poesiealbum. Eine volkskundliche Studie (Kleine Beiträge zur Volkskunstforschung, Bd. 7), Weimar 1960, S. 9.

42 Ausführlicher zur Poesiealbumsitte: Jürgen Rossin, Das Poesiealbum. Studien zu den Variationen einer stereotypen Textsorte, Frankfurt am Main 1985; Gertrud Angermann, Stammbücher und Poesiealben als Spiegel ihrer Zeit. Nach Quellen des 18.-20. Jahrhunderts aus Minden-Ravensberg (Schriften der Volkskundlichen Kommission des Landschaftsverbandes Westfalen-Lippe, Bd. 20), Münster 1971.

43 Wir können also zwei dichotome Variablen bilden, die wie folgt kodiert werden: 1. Variable: ein Zitat von Goethe liegt vor = 1; liegt nicht vor = 0; 2. Variable: ein Zitat von Karl Marx liegt vor = 1; liegt nicht vor = 0. Eine dritte Variable bilden wir, wenn ein Eintrag von einem anderen Autor stammt: 3. Variable. Ein Zitat eines anderen Autors liegt vor = 1; liegt nicht vor = 0.

44 Das laufende Dissertationsprojekt Stefan Walters trägt den Arbeitstitel «Der Staat und die Werte. Zum Einfluss des Staates auf die Werthaltung von Heranwachsenden in DDR und Bundesrepublik am Beispiel der Untersuchung von Einträgen in Poesiealben zwischen 1949 und 1989» und wird von Kurt Mühler am Institut für Soziologie der Universität Leipzig betreut.

45 Die 2863 erhobenen Einträge stammen von insgesamt 2653 verschiedenen Einträgern. 210 Einträge stammen somit von Inskribenten, die wiederholt in eines der untersuchten Alben eingetragen haben. Diese «Panel-Einträge» sind einerseits durch die Erhebung der Alben innerhalb des Bekanntenkreises verursacht. Andererseits kommen sie auch zustande, wenn ein Albumbesitzer zu einem späteren Zeitpunkt ein weiteres Album anlegt und dieselben Einträger aus einem früheren Album um eine Inskription bittet. Panel-Einträge können zu Verzerrungen der Befunde führen. So könnte ein und derselbe Einträger wiederholt ein bestimmtes Zitat für seinen Eintrag verwendet haben. Insofern sind sie gesondert zu behandeln. Für die folgenden Analysen haben wir deshalb nur die zeitlich zuerst erfolgten Einträge berücksichtigt.

46 Die Zuordnung nach Ost und West ergibt sich aus dem Eintrag selbst, in dem in der Regel der Wohnort angegeben wird («Leipzig, den …»). Hat ein Einträger auf eine Lokalisierung verzichtet, haben wir den Halter des Albums nach dem Eintragsort (Wohnort des Einträgers) gefragt. Für 50,5% der Einträger wurde auf diesem Weg eine DDR-Herkunft, für 49,5% der Einträger eine Herkunft aus den alten Bundesländern ermittelt.

47 Zum Problem der Scheinkorrelation siehe Andreas Diekmann, Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, Reinbeck 1995, S. 603.

48 Wir können dieses Verfahren an dieser Stelle nur skizzieren und verweisen stellvertretend auf Dieter Urban, Jochen Mayerl, Regressionsanalyse. Theorie, Technik und Anwendung, Wiesbaden 2011.

49 Die *Geschlechtszugehörigkeit* eines Einträgers wurde durch die Analyse der Widmungen und Subskriptionen erhoben und mit 0 = männlich (22,1%), 1 = weiblich (76,1%) dichotomisiert. Das *Jahr des Eintrags* wurde durch die Datierungsangaben im Eintrag ermittelt. Bei fehlender Datierung wurde der Halter des Albums gebeten, das Jahr des Eintrags zu ergänzen (arithmetischer Mittelwert: 1969; Standardabweichung: 12,8). Die *Wohnortgrösse* dient der Ermittlung von möglichen Stadt-Land-Unterschieden. Ausgangspunkt war die Erhebung des im Eintrag angegebenen Ortes. Fehlte die Ortsangabe, wurde der Albumhalter nach dem Wohnort des Einträgers befragt. Anschliessend wurde die aktuelle Einwohnerzahl des jeweiligen Ortes anhand aktueller Einwohnerstatistiken auf den offiziellen Webseiten der Kommunen bzw. durch Angaben der Landesstatistikämter ermittelt. Für die multivariate Analyse wurde die Einwohnerzahl kategorisiert, wobei sich bei der Bildung der Kategorien an den Ausprägungen der entsprechenden Variable in der «Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften» (ALLBUS) orientiert wurde: bis 1999 = 1; 2000–4999 =2; 5000–19999 = 3; 20000–49999 = 4; 50000–99999 = 5; 100000–499999 = 6; 500000 und mehr = 7 (arithmetischer Mittelwert: 3,64; Standardabweichung: 2,27). Da hier nur nach tendenziellen Stadt-Land-Unterschieden gefragt wird, erscheint dieses Vorgehen als hinreichend.

51 Die unabhängige Variable *Einträgergruppe* wurde durch Befragung des Halters ermittelt. Hierbei konnte der Halter aus verschiedenen Vorgaben wählen und den Einträger einer Kategorie zuordnen. War eine Befragung des Halters nicht möglich, wurde eine Eingruppierung anhand der inskribierten Widmungsformeln (z.B. «Dein Mitschüler») vorgenommen. In der multivariaten Analyse wurde in Peergroup (Mitschüler, gleichaltrige Freunde des Halters = 73,2%), Familie (Verwandte des Halters = 9,5%), Lehrer (9,2%) sowie sonstige Einträger (Personen des erweiterten persönlichen Netzwerkes des Halters = 3,5 %) unterschieden.

52 Das ungefähre *Alter* eines Peer zum Zeitpunkt des Eintrags wurde mit folgender Formel ermittelt: Alter des Halters bei Albumbeginn + Jahr bei Albumbeginn - Jahr des Eintrags. Hintergrundannahme war, dass der Halter des Albums und ein Mitschüler bzw. ein gleichaltriger Freund mehr oder weniger gleichaltrig sind (arithmetischer Mittelwert: 11,9 Jahre; Standardabweichung: 2,38).

53 Anders formuliert: Je höher dieser Wert und je geringer die Anzahl der dabei einbezogenen Variablen ist, umso besser erklärt er das Modell. Da es das Ziel ist, gute Erklärungen zu bekommen, sind besser an die Empirie angepasste Modelle vorzuziehen.

54 Zu Möglichkeiten der Darstellung mehrdimensionaler Daten s. Schnell (Anm. 39), S. 125–162.

55 Schnell (Anm. 39), S. 162.

56 Wir verweisen zugleich auf Empfehlungen von Henning Best, Christof Wolf, Logistische Regression, in: Wolf/Best (Anm. 39), S. 827–854.

57 Zuvor sind metrische Variablen, wie in unserem Fall das Alter, zur Verringerung der Multikollinearität (Korrelation zwischen den erklärenden Variablen) durch Z-Transformation zu standardisieren, S. Best/Wolf (Anm. 53), S. 841.

58 Vorhergesagte Wahrscheinlichkeit für gekennzeichnetes Goethe-Zitat bei Berechnung nach Tabelle 4.