Humboldt-Universität zu Berlin
Philosophische Fakultät I
Institut für Geschichtswissenschaften
Forschungsseminar: Ökonomie der Zerstörung – Wirtschaft und Diktatur im 20. Jahrhundert
Dozent: Prof. Alexander Nützenadel

Die Zeit und die Ökonomisierung

Betrachtung der Werkzeuge der Digital Humanities und Entwicklung einer robusten Analyse

Vorgelegt von:
Florian Müller
E-Mail: mullerfl@hu-berlin.de
MA Moderne Europäische Geschichte

Inhalt

Einleitung

Die Politik und die Ökonomie sind in der Konzeption von Michelle Foucault zwei Gegenpole der Macht.[1] Erst die Ökonomie mit ihrem Konstrukt der Bevölkerung konnte dem Souverän und dem Polizeistaat ein geeignetes Werkzeug entgegenstellen und leitete so das Zeitalter der Gouvernementalität ein.[2] Eng verbunden mit dem Aufstieg der Gouvernementalität war die Frage nach der Verteilung von Macht innerhalb eines Staates, später gar innerhalb einer Nation. So ist die Gouvernementalität und mit ihr die Ökonomie nicht ohne den Liberalismus mit seinem Konzept der Freiheit für Foucault zu denken.[3]

Auch wenn Foucault die Konzeption der Bevölkerung als die wichtigste Errungenschaft der Ökonomie feiert, gilt im Allgemeinen der Markt mit seinem Güteraustausch als das wesentliche Element der Ökonomie. Auf diesem werden Güter ausgetauscht und der Wert der Güter richtet sich (theoretisch) nach dem Angebot und der Nachfrage. Der so gebildete Preis dient den Marktteilnehmern als Indikator für ihre Kaufentscheidung.

Zeitungen und ihr Inhalt sind unter diesen Aspekten eine besondere Erfindung, bringen sie doch den Markt mit der Äußerung von Meinungen und Berichten und damit den Liberalismus zusammen. Bereits 1881 bemerkte Charles Dudley Warner in einem Vortrag zum amerikanischen Zeitungsmarkt, dass es für keinen Zeitungsverleger Sinn ergibt, eine Zeitung herauszugeben, die defizitär wirtschaftet, dass aber zugleich trotz der verzerrten Darstellung der Realität eine Zeitung über eine höhere Moral verfügt als die Gesellschaft, die sie umgibt.[4]

Die Frage, ob ein freier Markt politische Freiheit impliziert, wurde insbesondere nach den verheerenden Auswirkungen des Zweiten Weltkriegs und in der Zeit des Kalten Krieges intensiv diskutiert.[5] Für Ökonomen muss dabei für eine ideale und effektive Preisbildung Wissen und Meinungen frei verfügbar sein, Schumpeter vergleicht den politischen Wahlkampf gar mit dem täglichen Gang in den Supermarkt, bei dem die Hersteller mittels der Reklametechnik versuchen, den Konsumenten für sich zu gewinnen.[6] Daher sind die Verfügbarkeit freier Informationen und die damit verbundene Unterbindung von Zensur für eine freie

[1] Vgl. Michel Foucault, Sicherheit, Territorium, Bevölkerung: Vorlesung am Collège de France, 1977-1978, Frankfurt am Main, 2015, 4. Auflage, S. 77.
[2] Vgl. ebenda, S. 119.
[3] Vgl. ebenda, S. 506.
[4] Vgl. Charles Dudley Warner, American Newspaper, online verfügbar unter: http://www.gutenberg.org/cache/epub/3110/pg3110-images.html . Zuletzt geprüft am: 13.9.2017.
[5] Vgl. Milton Friedman, Kapitalismus und Freiheit, München, 2014, 9. Aufl., S. 32–34.
[6] Vgl. Joseph Alois Schumpeter, Kapitalismus, Sozialismus und Demokratie, Tübingen, 1993, 7., erw. Aufl., S. 418.

Marktwirtschaft von großer Bedeutung. Besonders in Deutschland über das gesamte 20. Jahrhundert hinweg befindet sich der Zeitungsmarkt jedoch im ständigen Wechsel zwischen freien und kontrollierten Märkten. Nach dem Zweiten Weltkrieg besteht sogar eine direkte Konkurrenz zwischen beiden Arten des Zeitungsmarktes. Daher könnte vor allem die Analyse des deutschen Zeitungsmarktes in Zeiten des Kalten Krieges Aufschluss über die Zusammenhänge von Macht, Freiheit und Ökonomie geben.

Doch die Menge an Artikeln in Zeitungen symbolisieren ein Problem der Geschichtswissenschaften, das laut Hayden White zu einer regelrechten Krise führe: Die Krise der Beliebigkeit von Quellen. Durch die Flut an neuen historischen Quellen werden diese beliebig, da es dem Historiker nicht mehr gelingen kann, alle zu einem Gebiet gehörenden Schriften und Quellen zu überblicken. Der Historiker kann sich dadurch, so Hayden, eine Kette der Geschichte selber schmieden, in Abhängigkeit seiner persönlichen Meinung und Absichten. Durch die damit verbundene Beliebigkeit der Quellen wird aus der historischen Arbeit eine Erzählung und es überzeugt nicht der Wissenschaftler, der die beste Quellenarbeit leistet, sondern der Historiker, der weiß, das Publikum mit seinen Geschichten mitzureißen.[7]

Die Forschung scheint die Bedeutung der Suche nach einer neuen Möglichkeit der Quellenauswertung erkannt zu haben, denn nicht nur die DFG, sondern sogar das BMBF fördern mittlerweile große Verbundprojekte, die sich mit der Erforschung des sogenannten Distant Readings, d.h. der automatisierten Auswertung von Texten, widmen. Augenscheinlich ist, dass mit der wachsenden Begeisterung und Neugier der Forschung in Bezug auf das Distant Reading und der Analyse natürlichsprachlicher Texte (NPL) Zeitungen als Quellen eine enorme Aufwertung erfahren haben. Zahlreiche Tages- und Wochenzeitungen wurden in den letzten Jahren digitalisiert und als Korpus bereitgestellt. Wahrscheinlich bieten sich Zeitungen besonders für die Erprobung des Distant Readings an, da in der Kombination mit den neuen Werkzeugen sie erstmals als Einheit auswertbar sind und im großen Maße verglichen werden können.

Diese Arbeit soll zunächst einen Überblick über den aktuellen Stand der Forschung im Bereich des NLP (Natural Language Processing) in den Digital Humanities[8] geben. Dazu werden

[7] Vgl. Hayden White, Der historische Text als literarisches Kunstwerk, in: Christoph Conrad und Martina Kessel (Hrsg.), Geschichte schreiben in der Postmoderne. Beiträge zur aktuellen Diskussion, Stuttgart, 1994, S. 123–157, hier S. 143–144.

[8] Es werden bewusst die gesamten Digital Humanities, also die Geisteswissenschaften und nicht nur die digitalen Geschichtswissenschaften betrachtet, da das Feld der Digital History aus der Sicht des Autors noch zu klein für eine ausführliche Betrachtung und Bewertung ist.

verschiedene, vom Autor ausgewählte Projekte vorgestellt und die Besonderheiten der einzelnen Projekte aufgezeigt. Im Anschluss sollen anhand eines Fallbeispiels Möglichkeiten aufgezeigt werden, hermeneutische Fragen mit den bereits heute zur Verfügung stehenden Werkzeugen und Eigenentwicklungen zu beantworten und nach Möglichkeiten der Verbesserungen gesucht werden.

Werkzeuge und Methoden der Digital Humanities

Zeitungen sind neben den neuen Medien eines der beliebtesten Testobjekte für Versuche des NLP in den Digital Humanities. Dies ist vermutlich der Größe des verfügbaren Materials sowie der Verfügbarkeit geschuldet. Zudem erlauben Zeitungen, große Zeitreihen betrachten zu können und damit Veränderungen im Denken und Fühlen einer Gesellschaft aufzuweisen.

Dabei stellt sich im Deutschen noch mehr als im Englischen die Frage, wie zuverlässig die Werkzeuge der Linguisten und Informatiker sind und ob sie verlässliche und reproduzierbare Ergebnisse gewährleisten können. Während es im englischen Sprachraum dank internationaler Verteilung der geistigen Kräfte verschiedene Projekte gibt[9], sind die Entwicklungen für deutschsprachige Analysen meist auf den deutschsprachigen Raum beschränkt.[10]

Ziel dieses Kapitels ist es, einen Überblick über die derzeit verfügbaren Werkzeuge und Projekte zu geben und deren Stärken, aber auch ihre Schwächen aufzuzeigen, um den Leser einen Eindruck über den Stand der Forschung im Bereich des NLP in den Digital Humanities zu geben und möglicherweise für das nächste Kapitel Anregungen zu finden, an welchen Stellen noch vom Wissenschaftler Eigenleistungen gefordert sind und an welchen Stellen sich der Wissenschaftler schon auf vorhandene Projekte und Tools beziehen kann.

DiaCollo[11]

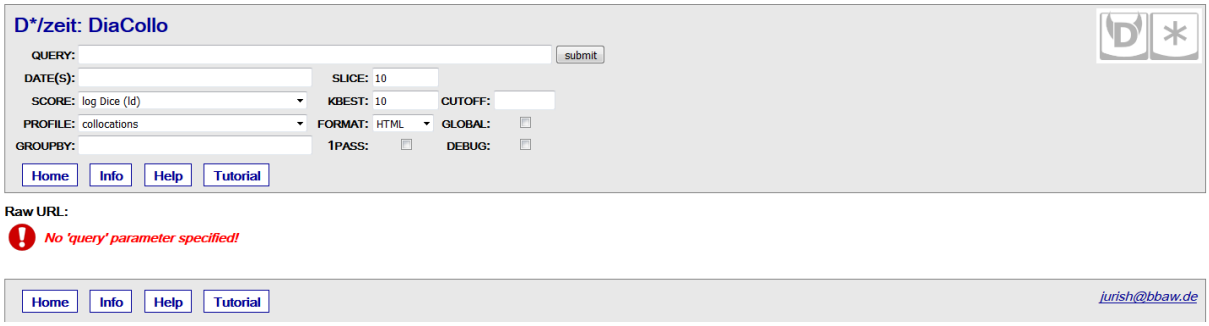

Abbildung 1 Startansicht DiaCollo Zeitkorpus

„DiaCollo ist ein Werkzeug für das Auffinden von typischen Wortverbindungen (Kollokationen) zu einem Stichwort in einem bestimmten Zeitraum und die visuell aufbereitete Darstellung der Ergebnisse. […] Das Zentrum Sprache bietet hier verschiedene

[9] Vgl. Marenglen Biba und Fatos Xhafa, Learning structure and schemas from documents, Berlin, 2011, S. XV–XVIII.
[10] Vgl. Alexander Mehler und Kai-Uwe Kühnberger et al., Modeling, Learning, and Processing of Text Technological Data Structures, Berlin, Heidelberg, 2012, S. XIII–XV.
[11] Erreichbar unter http://kaskade.dwds.de/dstar/zeit/diacollo/, zuletzt geprüft am 11.10.2017, 13:35 Uhr

Textsammlungen an, deren Entstehung sich über einen jeweils längeren Zeitraum erstreckt, unter anderem:

- dta+dwds: Die Referenzkorpora des Deutschen Textarchivs und des DWDS (Zeitraum von 1650 bis 2000).
- dta: Das Deutsche Textarchiv ausschließlich (1650–1900)
- kern: Das DWDS-Kernkorpus ausschließlich (1900–2000)
- zeit: Die Wochenzeitung "Die Zeit" (1946–2015)
- zeitungen: Eine Zusammenstellung verschiedener deutscher/österreichischer/schweizer Tages- und Wochenzeitungen (1946–2015)"[12]

DiaCollo ist ein Beispiel für ein Webtool, das auf ein professionell aufbereitetes, geschlossenes Korpus zurückgreift. Für die Digital Humanities besonders interessant sind dabei Korpora wie „zeit" und „zeitungen", da diese, wie es die Namen schon erahnen lassen, auf einem breiten Bestand von Zeitungsartikeln basieren.[13]

DiaCollo verfügt über eine Reihe von Vorzügen. So kann die Abfrage sehr komplex formuliert werden und erlaubt sogar die Verwendung von Regular Expressions. Zudem lassen sich die Ergebnisse nicht nur durch Zahlen, sondern können auch graphisch durch Wortwolken oder andere Darstellungsarten visualisiert werden.

Leider ist die Oberfläche aber für einen Einsteiger auf dem ersten Blick sehr unübersichtlich und überladen (s. Abbildung 1). Die Menüführung ist komplett in englischer Sprache gehalten und es ist einem Laien nicht in jedem Fall ersichtlich, welche Bedeutung die einzelnen Funktionen und Einstellungen haben. Zwar wurde von Seiten des Anbieters darauf reagiert, indem eine ausführliche Dokumentation bereitgestellt wurde, den Berührungsängsten vieler Nutzer dürfte diese aber nicht entgegenwirken. Ferner ist das Korpus geschlossen und kann nicht durch eigene Daten erweitert werden. Es ist nicht einmal möglich, verschiedene Korpora miteinander zu kombinieren. Wurde die Abfragemöglichkeit zwar einerseits gelobt, hat sie leider andererseits auch entscheidende Schwächen. Das Volumen an gleichzeitigen Abfragen, gerade mit Blick auf die Anzahl an Suchbegriffen, ist stark eingegrenzt, d.h. es wäre unrealistisch, die Zusammenhänge von 100 Begriffen über einen historischen Zeitraum mit DiaCollo nachvollziehen zu wollen.

[12] Bryan Jurish, DiaCollo Tutorial: Einführung, online verfügbar unter: http://kaskade.dwds.de/diacollo-tutorial/#introduction.html . Zuletzt geprüft am: 12.10.2017
[13] Unter http://kaskade.dwds.de/dstar/ findet man eine Auflistung aller verfügbaren Korpora

ePol[14] und der Leipzig Corpus Miner[15]

Das ePol Projekt ist ein an Verbundprojekt der Universität Leipzig und der Helmut-Schmidt-Universität Hamburg. „Die dem Projekt zugrunde liegende Hypothese lautet, dass sich im Zuge einer Verbreitung politischer Leitideen des Neoliberalismus der Chicago School, die von der Postdemokratiedebatte als neoliberale Hegemonie beschrieben wird, in zunehmendem Maße ökonomie- und marktaffine Begründungen für kollektiv bindende Entscheidungen zur Implementierung beliebiger policies in immer mehr Politikfeldern durchgesetzt haben und nach wie vor durchsetzen.“[16]

Die Untersuchung dieser Hypothese beruht auf der Auswertung von ca. 3,5 Millionen Zeitungsartikeln aus der „Frankfurter Allgemeinen Zeitung“, der „tageszeitung“, der „Süddeutschen Zeitung“ und „Die Zeit“.[17] Für die technische Auswertung der Artikel wurde an der Universität Leipzig der Leipzig Corpus Miner entwickelt. Dieser bietet neben der reinen Auszählungen von Häufigkeiten auch Kollokationsanalysen wie auch eine Schlagwortanalyse an. Zwar ist das Produkt noch in einer Alpha- Phase und so manche Funktionen funktionieren noch nicht, trotzdem erscheint der Aufbau bereits jetzt für die praktische Arbeit von Geisteswissenschaftler angenehmer als die von DiaCollo. Die Dokumentation ist jedoch derzeit noch recht dürftig. Sehr gut ist, dass das Werkzeug nicht auf ein festes Korpus begrenzt ist, sondern dass beliebige Korpora importiert werden können, sofern sie einem vorgegebenen Format entsprechen. Das Tool wird unter einer LGPL Lizenz[18] als eine Software-as-a-Service Lösung angeboten und ist kostenlos nutzbar, allerdings ist eine virtuelle Maschine und Linux notwendig. Das Importieren der Daten kann je nach Datengröße sehr langsam sein und mehrere Tage dauern. Die Entwickler geben an, für das Hochladen des Korpus des ePol Projektes drei Tage benötigt zu haben.[19] Dafür scheint es aber laut der Dokumentation keine Beschränkung im Größenumfang zu geben. Wünschenswert wäre es, wenn das Programm auch ohne eine virtuelle Maschine und lokal nutzbar wäre, damit der Nutzer bei komplexen Analysen nicht auf die Rechenkapazitäten des Rechenzentrums in Leipzig angewiesen ist und er Bedarf das Programm an seine Anforderungen anpassen könnte. Zudem können so die Werkzeuge nicht

[14] Erreichbar unter http://www.epol-projekt.de/, zuletzt geprüft am 11.10.2017, 13:35 Uhr
[15] http://lcm.informatik.uni-leipzig.de/
[16] Gregor Wiedemann und Matthias Lemke, Einleitung: Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, in: Matthias Lemke und Gregor Wiedemann (Hrsg.), Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, Wiesbaden, 2016, 1. Aufl. 2016, S. 1–16, hier S. 5
[17] Vgl. ebenda, S. 7.
[18] Für weitergehende Informationen zu dieser Lizenzart: https://www.gnu.org/licenses/lgpl-3.0.de.html , zuletzt geprüft am 15.10.2017, 10:20 Uhr
[19] Vgl. http://lcm.informatik.uni-leipzig.de/download.html, zuletzt geprüft am 15.10.2017, 10:20 Uhr

auf Anforderungen anderer Projekte angepasst werden. Trotzdem ist es natürlich erfreulich, dass das ePol Projekt seine Werkzeuge auch anderen Projekten der Digital Humanities bereitstellt.

Weblicht[20]

Das an den Universitäten Tübingen gehostete Projekt Weblicht hat das Ziel, verschiedenste Werkzeuge zur automatisieren Auszeichnungen von Textkorpora zu bündeln und als virtuelle Umgebung Forschern bereitzustellen. Die Entwicklung von Weblicht startete bereits im Oktober 2008 als Verbundprojekt des deutschen Ablegers des Forschungsverbundes CLARIN[21]. WebLicht ist für alle Nutzer im Bereich des DFN Netzwerkes frei zugänglich.[22]

Die Stärke von WebLicht liegt in seiner freien Zugänglichkeit und seinem Umfang an hochwertigen Analysetools. Vor allem deutsche Nutzer werden es zudem zu schätzen wissen, dass WebLicht die wichtigsten deutschen Werkzeuge bündelt und somit auch für deutsche Texte eine qualitativ hochwertige automatisierte Korpusauszeichnung vornimmt. Die Quellen können dabei in allen gängigen Formaten vorliegen und alle Ergebnisse stehen als XML-Datei zum Herunterladen bereit.

Leider sind allerdings die Kapazitäten von WebLicht stark begrenzt. Zum einen dürfen die Quelldateien nicht größer als 5MB sein, was im Rahmen einer Zeitungsanalyse nicht einmal einer Ausgabe entsprechen kann. Diese Beschränkung besteht leider auch, wenn man WebLicht nicht über das Webinterface, sondern per Konsole als eine Software-as-a-Service Anwendung nutzt. Zudem erweitert jeder Analyseschritt den Datenumfang, weshalb es selbst bei kleineren Korpora häufig zu Fehlern und Abbrüchen bei der Bearbeitung kommt, da die einzelnen Tools an ihre Leistungsgrenze stoßen.

[20] https://weblicht.sfs.uni-tuebingen.de/, zuletzt geprüft am 13.10.2017, 13:24 Uhr
[21] https://www.clarin.eu/, zuletzt geprüft am 13.10.2017, 13:32 Uhr
[22] Vgl. Main Page: Welcome to WebLicht! Erschienen am 09.06.2017, online verfügbar unter: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page . Zuletzt geprüft am: 13.10.2017.

Entwicklung einer robusten Analyse

Da viele der verfügbaren Anwendungen sich nach wie vor in einem Versuchsstadium befinden oder gar noch nur in einer frühen Planungs- bzw. Entwicklungsstufe und dadurch sämtliche Ergebnisse einer automatisierten Analyse leicht durch Kritiker hinterfragt werden könnten, soll in diesem Kapitel betrachtet werden, ob es auch mit den einfachen Werkzeugen bereits möglich ist, die von Eijnatten et al. geforderte Beantwortung einer hermeneutischen Frage[23] auf Grundlage einer automatisierten Auswertung beliebiger digitalisierter Quellen nachzugehen. Als Material und Anwendungsbeispiel sollen dabei ein selbsterstelltes Korpus und eine selbstgewählte Fragestellung dienen.

Das Korpus bildet die Zeitung „Die Zeit“ der Jahrgänge von 1969 bis 1989. Die Zielstellung soll es sein, ähnlich wie beim ePol Projekt, eine Ökonomisierung der Sprache festzustellen. Das Vorgehen zu ePol unterscheidet sich dabei insofern, dass sich bei der Analyse auf die Nomen konzentriert werden soll, da diese das wichtigste Wort in der deutschen Sprache darstellen[24] und für die meisten Werkzeuge gut erkennbar sind. Ferner soll ein reines Distant Reading vorgenommen werden und nicht das von ePol vorgeschlagene Blended Reading, welches eine Mischung aus Distant und Close Reading vorsieht.[25] Die Ökonomisierung wird mithilfe des Auftretens von etwa 3600 ökonomischen Begriffen untersucht, die als Teil der ökonomischen Fachsprache angenommen werden. Die zu untersuchende These lautet, dass es zu einer Ökonomisierung gekommen ist, wenn es in den wirtschaftsfernen Ressorts und damit in der nichtökonomischen Fachsprache zu einer Zunahme dieser ökonomischen Begriffe gekommen ist. Die These beruht auf der Beobachtung, dass unter der Ökonomisierung ein Vorgang verstanden wird, durch den es zu einer zu einer „Aufwertung ökonomischer Handlungsprinzipien etwa in der Kunst, im Journalismus, im Gerichtswesen oder in den Hochschulen“[26] kommt.

[23] Vgl. Joris van Eijnatten, Toine Pieters und Jaap Verheul, Big Data for Global History: The Transformative Promise of Digital Humanities, in: BMGN - Low Countries Historical Review 128 (2013), Nr: 4, S. 55–77, hier S. 75–76.
[24] Vgl. Holger Becker, Die Wirtschaft in der deutschsprachigen Presse: Sprachliche Untersuchungen zur Wirtschaftsberichterstattung in der Frankfurter Allgemeinen Zeitung, der Neuen Zürcher Zeitung, der Presse und im Neuen Deutschland, Frankfurt am Main, New York, 1995, S. 133.
[25] Vgl. Matthias Lemke und Alexander Stulpe, Blended Reading: Theoretische und praktische Dimensionen der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung, in: Matthias Lemke und Gregor Wiedemann (Hrsg.), Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, Wiesbaden, 2016, 1. Aufl. 2016, S. 17–62, hier S. 43–53.
[26] Vgl. Uwe Schimank und Ute Volkmann, Ökonomisieurng der Gesellschaft, in: Andrea Maurer (Hrsg.), Handbuch der Wirtschaftssoziologie, Wiesbaden, 2008, 1. Aufl., S. 382–393, hier S. 382.

Zum Aufbau des Zeitungskorpus wurde aus allen Artikeln die Nomen mithilfe eines Part-of-Speech-Taggers[27] (kurz POS-Tagger) extrahiert und in einer Tabelle abgelegt. Dabei wird zu jedem extrahierten Nomen notiert, aus welcher Jahr, welcher Ausgabe, welchem Ressort und welchem Artikel es stammt. Ein beispielhafter Eintrag würde daher folgendermaßen aussehen:

Kapital,1969,1,Wirtschaft,1

Die ökonomischen Begriffe entstammen Glossaren und Indizes verschiedenster ökonomischer Lehrbücher. Die Annahme ist, dass diese Lehrbücher im starken Maße Fachsprache enthalten und daher die Begriffe aus diesen beiden Sektionen ökonomische Fachbegriffe darstellen. Diese Annahme kann als valide vorausgesetzt werden, da, wie Abbildung 2 entnommen werden kann, bei der Betrachtung des Vorkommens der Begriffe in den verschiedenen Ressorts eine deutliche Betonung des Wirtschaftsressorts festgestellt werden kann.

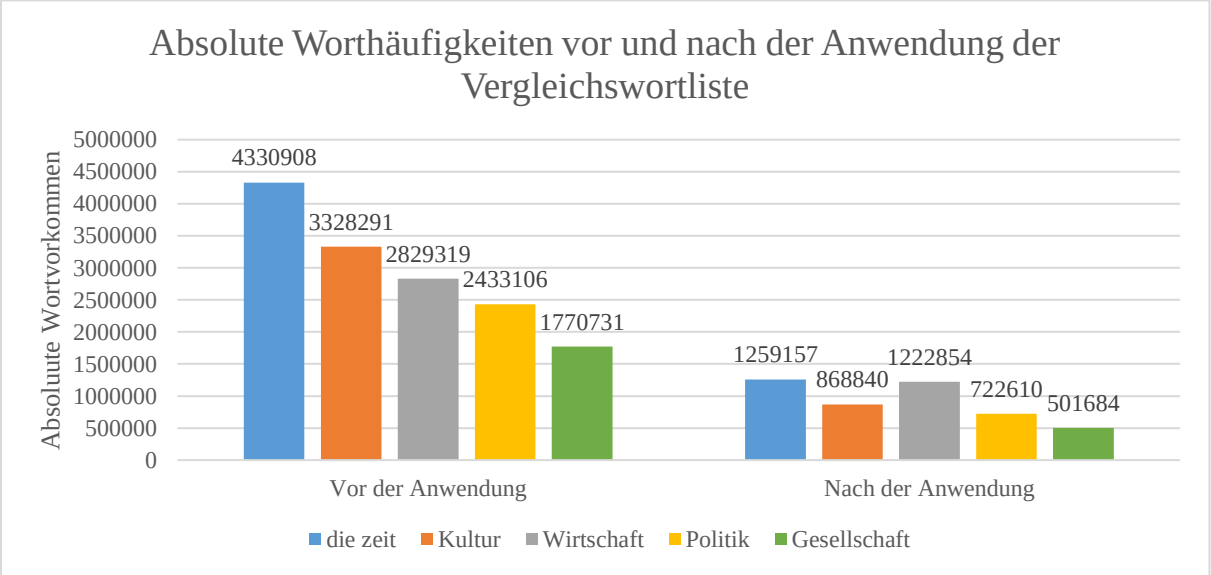

Abbildung 2 Vergleich Vorkommen aller Begriffe und ökonomischer Begriffe in den 5 größten Ressorts

Auch wenn die von der Zeitung „Die Zeit“ digitalisierten Jahrgänge einige Fehler bezüglich der Zeichenerkennung enthalten, kann die Qualität des Korpus hier als ausreichend angesehen werden. Für die Ressortzuweisung trifft dies allerdings nicht zu, wie man der Abbildung 3 entnehmen kann.

Aus den Abbildungen 2 und 3 lässt sich erschließen, dass eine Aussage über die Ökonomisierung der Sprache des betrachteten Korpus stark davon abhängig ist, ob das Ressorts „die zeit“ als wirtschaftsnah oder wirtschaftsfern ausgewiesen werden kann. Eine Möglichkeit wäre nun,

[27] Verfügbar unter http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger, zuletzt geprüft am 15.10.2019, 10:28 Uhr.

die Artikel des Ressorts den anderen bestehenden Ressorts zuzuordnen. Dazu müssten die Begriffe, die im Ressort „die zeit“ zu finden sind, auf ihre ökonomische Herkunft hin untersucht werden. Um die Analyse zu vereinfachen, bräuchte man dabei lediglich zu entscheiden, ob ein Artikel ökonomischer Natur ist oder einem wirtschaftsfernen Ressort zugeordnet werden kann.

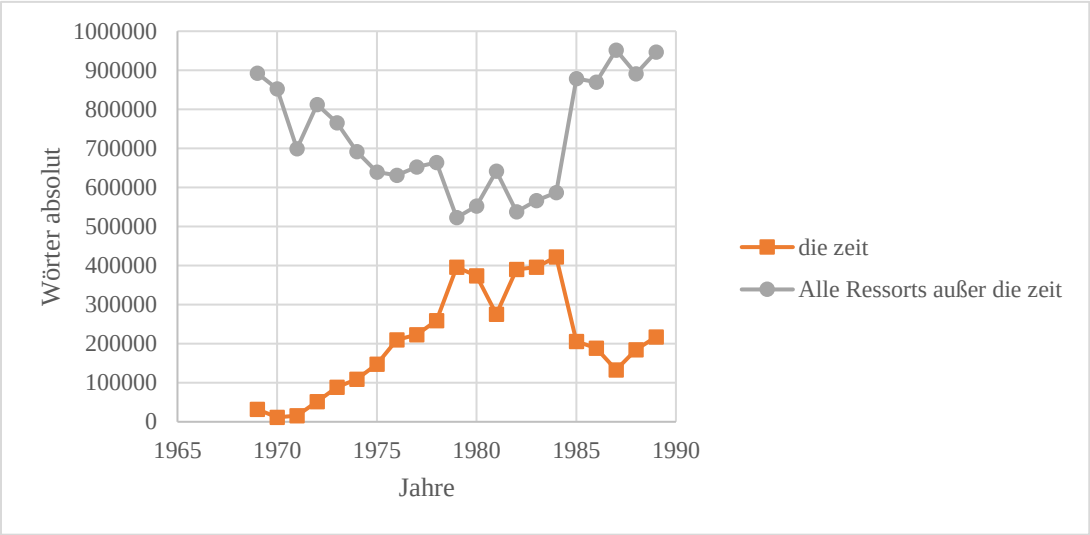

Abbildung 3 Wörter im Ressort "die zeit" absolut im Vergleich mit allen anderen Ressorts

Das vorgeschlagene Vorgehen hat jedoch zwei Fehler unbekannter Art und unbekannter Ausmaße, die sich in die Analyse einschleusen würden:

1. Es ist nicht bekannt, ab welchen Wert von einem wirtschaftsnahen Artikel gesprochen werden kann
2. Es ist nicht bekannt, wie viele andere Ressorts eigentlich wirtschaftsnah sind, aber fälschlicherweise als wirtschaftsfern angenommen wurden

Der erste Fehler kann das Ergebnis der Analyse in beide Richtungen verändern. Je nachdem, ab welchem Verhältnis von Wirtschaftsbegriffen im Artikel zu Wörtern im Artikel von einem wirtschaftsnahen Artikel die Rede sein kann, nimmt der Gesamtanteil der wirtschaftsfernen Artikel am Gesamtumfang einer Jahresausgabe zu oder ab. Auch wenn es möglich wäre, den durchschnittlichen Anteil an wirtschaftsnahen Begriffen in Artikeln des Wirtschaftsressorts als Schwellenwert zu nehmen, wäre dies nur eine weitere Annahme, die in Kombination mit dem zweiten Fehler dem Ergebnis seine Aussagekraft nehmen würde. Im schlimmsten Fall ließe sich gar eine Annäherung des Anteils der ökonomischen Begriffe an unseren selbstgewählten Grenzwert feststellen.

Die Annahme des zweiten Fehlers beruht auf der Beobachtung, dass Ressortnamen wie „Gesellschaft“[28], „Zeit-Serie“, „Serie“ oder ähnliche Namen keine Aussage darüber machen, ob es sich hierbei um ein wirtschaftsfernes oder wirtschaftsnahes Ressort handelt. Zusätzlich wurde bisher darauf verzichtet, Ressorts wie „Grundstücksmarkt“ oder „Immobilien“ dem Wirtschaftsressort zuzuordnen, obwohl diese dem Namen nach als wirtschaftsnah angenommen werden können. Zwar stellt sich insbesondere bei dem Ressort „die zeit“ die Frage, zu welchem Ressorts die Artikel wirklich zugeordnet werden können, da der Umfang dieses Ressorts bis zu 50% aller Begriffe eines Jahrgangs ausmacht, aber auch die anderen Ressorts könnten aufsummiert zu einer Verzerrung des Ergebnisses führen.

Die Hoffnung, auf eine externe Einteilung der Artikel durch die Redaktion der Zeit zurückgreifen zu können, muss damit als hinfällig angesehen werden. Statt auf die bisherige Zuordnung mit teilweise unklaren Namen zu bauen, könnte eine eigene eindeutige Einteilung in Kategorien wie „Politik“, „Sport“, „Feuilleton“, „Bildung“ und „Wirtschaft“ vorgenommen werden. Als Unterscheidungsmerkmal könnten die Strukturen der bereits bestehenden Ressorts mit eindeutiger Zuordnung genutzt werden, um unterscheidbare Strukturen für jedes Cluster zu bilden.

Der vorstellbare Ablauf, wie diese Cluster anschließend gebildet werden, könnte zunächst sehr allgemein formuliert wie folgt aussehen:[29]

1. Es werden zufällig so viele Elemente als Startpunkte für die Cluster ausgewählt, wie Cluster gebildet werden sollen.
2. Eine Distanz[30] zwischen den Clustern wird jeweils ermittelt
3. Ist die Distanz innerhalb eines bestimmten Bereichs, so dass von einer zu starken Nähe der Mittelpunkte einzelner Cluster ausgegangen werden muss, wird Schritt 1 wiederholt
    a. Wenn dieser Punkt x-mal wiederholt wurde, breche den Vorgang ab und verringere die zu erstellende Clusterzahl
4. Es wird das nächste Element gewählt und seine Distanz zu den bisherigen Clustern ermittelt
5. Das Element wird dem Cluster zugeordnet, zu dem er die geringste Distanz aufweist
6. Das so gebildete Cluster bekommt einen neuen Mittelpunkt
7. Wiederhole Schritt 3 bis alle Artikel zugeordnet wurden

[28] Das Magazin Capital wirbt z.B. mit dem Slogan: „Wirtschaft ist Gesellschaft“!
[29] Vgl. Jurij Leskovec, Anand Rajaraman und Jeffrey D. Ullman, Mining of massive datasets, Cambridge, 2015, 2nd ed., 1st repr, S. 233–236.
[30] Um was für eine Distanz es sich handelt und wie diese gebildet wird, müsste separat betrachtet werden.

Durch die Schritte zwei und drei kann zunächst einer zu starken Auswirkung der zufälligen Verteilung der Startpunkte auf das Endergebnis entgegengewirkt werden. Zwar müsste der Bereich in Schritt drei erst durch andere Überlegungen ermittelt werden, aber zumindest könnte so das Risiko einer möglichen Diskriminierung eines Clusters stark vermindert werden, da eine zu starke Nähe zweier Startpunkte eine Neuverteilung der Startpunkte bedeuten würde. Die Bedingung 3a ist ferner notwendig, da eine zu große Zahl von gewünschten Clustern bei zu geringer Zahl an Elementen zu einer endlosen Ermittlung von Startpunkten führen könnte.

Trotz dieser Vorkehrungen impliziert dieses Vorgehen eine Reihe von offenen Fragen. So wird nicht vorgegeben, wie der Algorithmus reagieren soll, wenn die Distanzen zwischen einem Element und verschiedenen Clustern gleich groß sind. Sollte dann eine Art Münzwurf die Zuordnung regeln oder sollte das Element allen in Frage kommenden Clustern zugeordnet werden? Zudem ist es aufgrund der zufälligen Verteilung der Startpunkte denkbar, dass die Ergebnisse nicht reproduzierbar sind. Würde dies Auswirkungen auf die Aussage der Analyse haben?

Zur Vereinfachung soll nun angenommen werden, dass es gelingen würde, ein Maß für eine Distanz zu schaffen, das die Artikel eindeutig trennt und eine Einteilung in die Cluster reproduzierbar und eindeutig möglich macht. Da jedoch keine klare Aussage darüber gemacht werden kann, wie stark der Anteil des Wirtschaftsteils in der ursprünglichen Grundgesamtheit war –erinnert sei hier die beiden Fehler unbekannter Art – lässt eine solche Einteilung keine Aussage über die Ökonomisierung der Sprache im wirtschaftsfremden Ressorts zu! Vielmehr könnte man im Falle einer tatsächlichen Ökonomisierung nur feststellen, dass die Anzahl der Artikel im Wirtschaftsressort zunehmen. Da aber ja untersucht werden soll, ob es besonders wirtschaftsnahe Artikel in wirtschaftsfernen Ressorts gibt, würde eine solche Einteilung keine Aussage über eine mögliche Ökonomisierung zulassen.

Eine trennscharfe Einteilung der Artikel in verschiedene Cluster erbringt keinen Mehrwert bei der Beantwortung der These. Daher muss die These mit einem Verfahren untersucht werden, welches die Antwort unabhängig macht von der Ressorteinteilung der Artikel. Es bietet sich dabei an, die gesamte Sprache der Zeitung als eine eigene Fachsprache anzunehmen, die sich aus den Fachsprachen der einzelnen Ressorts zusammensetzt. Dann wäre die ökonomische Fachsprache nur ein Bestandteil unter vielen, den man aber mit Hilfe des ökonomischen Sprachkorpus zum Vorschein bringen könnte. Daraus lässt sich wiederum die These ableiten, dass eine zunehmende Ökonomisierung sich in einer stärkeren Präsenz der ökonomischen Sprachanteile in der Zeitungssprache bemerkbar machen müsste.

Die zu Grunde liegende Überlegung des weiteren Vorgehens ist, dass ein Bericht über die Schönheit der Aufführung in der Oper eine starke thematische Distanz zur Entscheidung der Bundesbank über die Zinspolitik aufweist. Ziel soll es daher sein, diese Distanz technisch zu modellieren und vor allem ihr eine Bedeutung zuzuweisen.

Ein Maß für eine Distanz muss zunächst vier wesentliche Eigenschaften erfüllen:[31]

1. Eine Distanz ist immer positiv
2. Eine Distanz ist nur dann null, wenn beide Messpunkte gleich sind
3. Es ist egal, von welchen der beiden Punkte man die Distanz misst
4. Es gibt keinen kürzeren Weg zwischen den Punkten (die sogenannte Dreiecksungleichung)

Das klassische Maß für eine Distanz für Vektoren in einem Raum ist die euklidische Norm,[32] allerdings lassen sich die Artikel einer Zeitung und die darin enthaltene Sprache nur schwer über Koordinaten in einem Raum darstellen. Im Bereich der Big Data und Data Warehouse Analyse haben sich aber auch weitere Distanzmaße wie die Hamming Distanz, die Edit Distanz oder die Jaccard Distanz etabliert.[33]

Sowohl die Hamming-, wie auch die Edit-Distanz sind nicht für die Ziel dieser Arbeit konzipiert, da diese beiden Maße vor allem Wortähnlichkeiten messen. Bei der Edit-Distanz wird dazu gezählt, wie viele Veränderungen an einem Element vorgenommen werden müssen, damit es einem anderen Element ähnelt.[34] Die Hamming-Distanz misst dagegen einfach die Unterschiede aller Eigenschaften zweier Elemente. Je mehr Eigenschaften abweichen, desto größer sind die Elemente voneinander entfernt.[35]

Die Jaccard Distanz hingegen arbeitet mit der Jaccard Ähnlichkeit von Elementen, welche für die hier betrachtete Analyse von Interesse ist. Dazu werden die Eigenschaften zweier Elemente miteinander verglichen und das Verhältnis von übereinstimmenden Eigenschaften zu der Summe aller Eigenschaften gebildet.[36] Im vorliegenden Fall würde die Ähnlichkeit gemessen, in dem gezählt wird, wie viele Elemente im Vergleichskorpus enthalten sind und wie viele Elemente in einem einzelnen Zeitungsartikel. Anschließend müsste nun gezählt werden, wie viele der Elemente des Vergleichskorpus auch in der Menge des Artikels wiedergefunden

[31] Vgl. ebenda, S. 87.
[32] Vgl. Gerhard Jank und Hubertus Theodorus Jongen, Höhere Mathematik für Maschinenbauer: Skript zur Vorlesung, 1999, S. 55.
[33] Vgl. Leskovec et al. Mining of massive datasets, S. 87–91.
[34] Vgl. ebenda, S. 90.
[35] Vgl. ebenda, S. 91.
[36] Vgl. ebenda, S. 69.

werden können. Diese Menge teilt man abschließend durch die Summe der zuvor ermittelten Menge maximal möglicher Elemente. Die Distanz berechnet sich im Anschluss aus ***1-sim(x,y)***[37]. Dadurch wird gewährleistet, dass bei einer vollständigen Übereinstimmung der Abstand der beiden Elemente null misst.

Die wichtigste Größe in unserem vorliegenden Fall bleibt der Anteil der ökonomischen Begriffe verglichen mit der Gesamtzahl der Begriffe eines Artikels, d.h. die Dichte an Vergleichswörtern in einem Artikel. Je mehr Wörter eines Artikels dabei aus dem Vergleichskorpus stammen würden, desto größer wäre seine Dichte. Eine Dichte von Null entspräche demnach, um beim oben genannten Beispiel zu bleiben, einem Artikel über die Opernvorführung bzw. jedem anderen Artikel, der keinen der gesuchten ökonomischen Begriffe enthält. Je größer die Dichte werden würde, desto größer ist der Anteil an ökonomischen Begriffen und desto eher kann davon ausgegangen werden, dass es sich bei dem Artikel um einen ökonomischen Artikel handelt. Interessant wäre es daher, wie stark die Dichte an ökonomischen Begriffen in den Artikeln sich über die Jahre verändert.

Neben der reinen Dichte ökonomischer Begriffe im Jahresschnitt könnte mit Hilfe einer Jaccard- Distanz zusätzlich die Variabilität untersucht werden. Vermutet wird, dass die Entfernungen der Artikel vom Ursprung, d.h. dem ökonomischen Korpus, bei einer Ökonomisierung der Sprache über die Jahre hinweg geringer werden und damit die Anzahl verschiedener Begriffe pro Artikel zugenommen hat. Problematisch könnte dabei jedoch werden, dass die Ergebnisse sehr eng beieinander liegen werden, da die Menge der Begriffe in einem Artikel deutlich kleiner ist als die Menge an ökonomischen Begriffen im Vergleichskorpus.

Der große Nachteil dieses Vorgehens ist, dass anstelle jedes Ressorts eines Jahres jeder Artikel eines Jahres angefasst werden müsste. Dies bedeutet einen ungemein höheren Rechenaufwand, der die Analyse auch mit modernen Rechenkapazitäten sehr langwierig macht. Je mehr Schleifen hintereinander durchlaufen werden müssen, desto mehr Zeit wird für eine Berechnung benötigt. Daher muss bei der Implementierung der Datenstruktur Wert darauf gelegt werden, dass im besten Falle nur eine Widerholungsschleife benötigt wird. Für das Jahr 1969 enthält das Korpus 5785 Artikel. Geht man davon aus, dass die Artikel mit der Zeit nicht weniger werden, so müssen mindestens 20x5785 Schleifendurchläufe getätigt werden. Ohne

[37] Vgl. ebenda, S. 88–89.

weitere Optimierung dauert so die Analyse eines Jahrgangs über 3 Stunden, die Analyse von 20 Jahrgängen also fast drei Tage!

Führt man die Berechnungen für die Jahre 1969 und 1989 durch, so erhält man ein interessantes Ergebnis. Entgegen der Vermutung, dass die Gesellschaft und mit ihr ihre Sprache, ökonomisiert wurde, gilt es festzuhalten, dass unter Berücksichtigung aller Artikel eines Jahrgangs sich die Variabilität der ökonomischen Begriffe kaum gewandelt hat, sie ist sogar leicht zurückgegangen. Die Dichte an ökonomischen Begriffen in den Artikeln ist sogar spürbar zurückgegangen, wobei jedoch im Gegenzug die Anzahl an Artikeln, in denen kein einziger der gesuchten Begriffe gefunden wurde, mit einem Rückgang von 12 auf 6 sich halbiert hat.

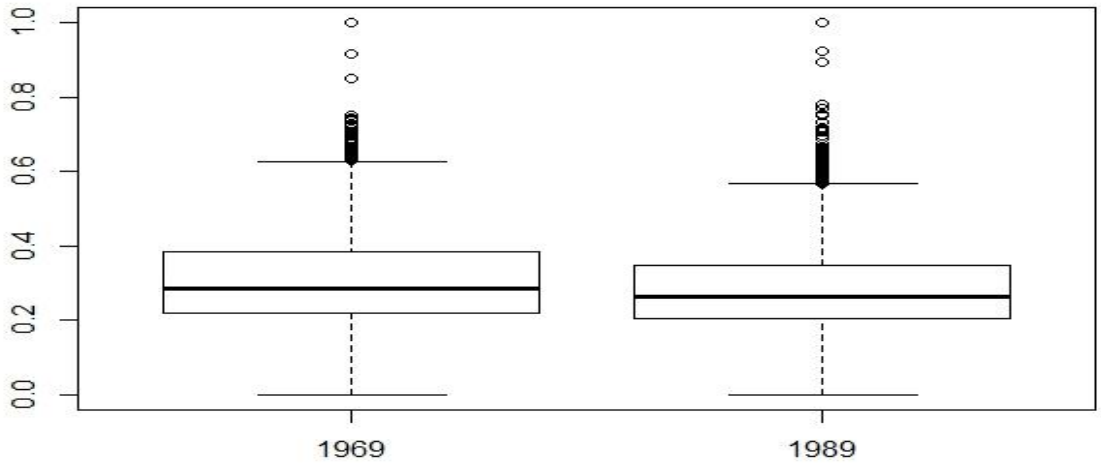

Abbildung 4 Boxplot der Dichtenerteilung

Abbildung 4 verbildlicht sehr eindeutig, dass die Begriffe der Ökonomie 1989 im Vergleich zu 1969 an den Rand geschoben wurden. Bei einer Ökonomisierung wäre eine Verschiebung in Richtung Zentrum zu erwarten gewesen.

Bewahrheitet hat sich leider, dass die Jaccard-Distanz keine valide Aussage über die Variabilität der ökonomischen Begriffe in den Jahren pro Artikel gestattet. Wie Abbildung 5 entnommen werden kann, liegen die Werte zu nah beieinander. Es müsste also für die Variabilität ein neues Maß berechnet werden. An dem Ergebnis ändern dürfte es hingegen nichts.

Auch wenn es sich bei dieser Analyse „nur“ um das Zählen von Wörtern gehandelt hat, liefert sie ein interessantes Ergebnis, welches es Wert ist, in einer weiteren Arbeit genauer betrachtet zu werden. Diese kann dann auch wieder auf Grundlange von Methoden des selektiven Close Readings oder gar des Blended Readings geschehen. Das Ergebnis impliziert aber auch eine Hinterfragung des genutzten Ökonomisierungsbegriffes. So hat das ePol Projekt z.B. sich auf

die Terminologie und Themen des Neoliberalismus ausgerichtet[38] und nicht auf die ökonomische Sprache allgemein. Bedeutet aber eine Zunahme von Diskussionen um eine neoliberale Wirtschaftspolitik eine zunehmende Ökonomisierung der Gesellschaft? Die hier gemachten Beobachtungen, sofern sie sich auch mit Blick auf den gesamten Zeitraum bestätigen sollten, widersprechen dieser These. Es zeigt aber, dass es am Ende vielleicht weniger die Werkzeuge sind, die einer Verbreitung der Digital Humanities im Wege stehen, sondern mehr die nicht eindeutigen Definitionen rund um die zu beantwortenden Fragen. Dies würde schlussendlich wieder zu Hayden White aus der Einleitung zurückführen, der den Geschichtswissenschaften vorwirft, keine feste, fachspezifische Terminologie zu besitzen und in Ermangelung dieser keine wissenschaftliche Disziplin darstellen könne.[39]

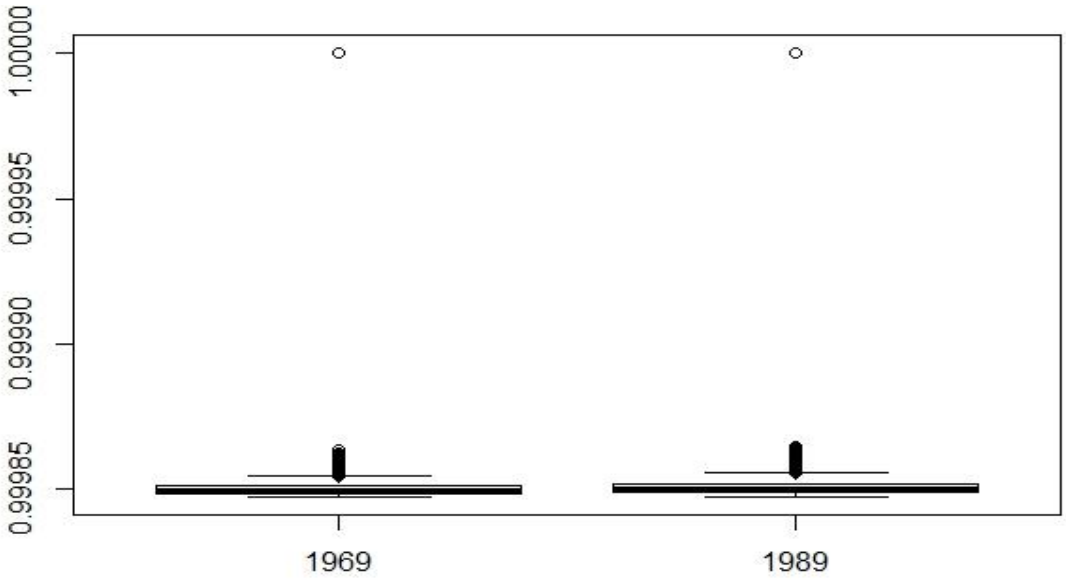

Abbildung 5 Boxplot Jaccard Distanzen[40]

[38] Vgl. Wiedemann et al. Einleitung: Text Mining in den Sozialwissenschaften, S. 5.
[39] Vgl. White Der historische Text als literarisches Kunstwerk, S. 148.
[40] Bei Betrachtung der Graphik ist zu beachten, dass der Minimumwert der Skala um die 0.9985 liegt. D.h. auch wenn der Boxplot suggeriert, dass die Werte mit einer vollen Distanz von 1 sehr weit weg vom Rest der Werte ist, beträgt er in Wirklichkeit nur 0.0015 Längeneinheiten!

Fazit

Die Digital Humanities in ihrer gesamten Breite und ihre Werkzeuge befinden sich noch in einer frühen Phase ihrer Etablierung. Diese Arbeit ging daher der Frage nach, wie weit Werkzeuge und Methoden bereits für den praktischen, wissenschaftlichen Alltag einsatzfähig sind.

Es musste dabei festgestellt werden, dass die Werkzeuge sich meist noch in einer sehr frühen Entwicklungsphase befinden. Daher wäre es im Hinblick auf die schnellere Weiterentwicklung wünschenswert, dass die Werkzeuge für die Open Source Gemeinde freigegeben werden würden. Zwar ist z.B. der in dieser Arbeit vorgestellte LCM unter der LGPL (Lesser General Public Licence) verfügbar und viele Programme kostenlos als eine Software-as-a-Service Lösung erhältlich, aber gerade die Software-as-a-Service Lösungen verhindern eine Anpassung der bestehenden Werkzeuge an die sich derzeit noch ständig wechselnden Anforderungen. Zudem schränken solche Lösungen aufgrund der Anforderungen, die Daten auf einen externen Server schicken zu müssen, die Größe der zu verarbeitenden Datenmenge meist ein. Wenn nicht jedes Projekt wieder von vorne beginnen müsste, sondern die Projekte auf den Erfahrungen von anderen Forschungsteams aufbauen können, würde dies der weiteren Entwicklung sehr zugutekommen. Verständlicherweise haben die Forscher dabei wahrscheinlich Bedenken, dass ihre geistige Arbeit von anderen geklaut werden könnte.

Trotz der Einschränkungen bei den Werkzeugen ist es dennoch möglich und notwendig, bereits jetzt mit Grundlagen der Statistik, Mathematik und Informatik erste tiefergehende Analysen zu tätigen, die über die Suche nach für Fragen relevante Textstellen hinausgehen. Forscher auf dem Feld der Digital Humanities dürfen dabei nicht den Fehler begehen, nur Erfolge aufweisen zu wollen. Jeder Versuch, eine Frage gewissenhaft zu beantworten, führt zu neuen, wichtigen Erkenntnissen, die in weitere Forschungsprojekte einfließen kann. Daher ist jeder Versuch, auch wenn er aus der Sicht der Disziplin zu keinem befriedigenden Ergebnis geführt hat, eine wichtige Erfahrung, auf die künftige Forscher aufbauen können. Ein unbedingter Erfolgsdruck würde nur dazu führen, dass Analysen und ihre Ergebnisse solange angepasst werden, bis die Aussagen sich mit den Erwartungen des Forschers oder gar der Gesellschaft decken. Dies wäre aber nichts anderes als das Erzählen einer Geschichte, wie es Hayden White den Historikern vorwirft. Statt daher auf Ergebnisse zu schauen, sollte sichergestellt werden, dass die Methodik eindeutig ist.

Die Hoffnung bleibt zunächst, dass auch mit dem „einfachen Zählen von Wörtern“ bereits erste Erkenntnisse zu gewinnen sind, die die Motivation erzeugen, weitere Untersuchungen und Versuche zu unternehmen. So ist das in dieser Arbeit vorgestellte Fallbeispiel bereits zu großen Teilen realisiert worden und bedarf nun des Feinschliffs. Abschließend festzuhalten ist aber auch, dass die Methoden der Analysen nur so genau sein können, wie es die Definitionen der zu untersuchenden Vorgänge sind. Sollte also jede Forschungsgruppe unter einem Vorgang oder Begriff etwas anderes verstehen, dann können die Korpora und Werkzeuge noch so genau sein, die Ergebnisse würden stark variieren. Daher ist es, gerade im Zeitalter der großen Datenanalysen, unabdingbar, festzuhalten, was unter bestimmten Begrifflichkeiten innerhalb der Geschichtswissenschaften zu verstehen ist. Dadurch werden auch die Historiker in den Forschungsdiskurs miteinbezogen, die sich nicht mit der Erstellung von Algorithmen und der Auswertung statistischer Materialien beschäftigen möchten.

Literaturverzeichnis

Becker, Holger, Die Wirtschaft in der deutschsprachigen Presse: Sprachliche Untersuchungen zur Wirtschaftsberichterstattung in der Frankfurter Allgemeinen Zeitung, der Neuen Zürcher Zeitung, der Presse und im Neuen Deutschland, Frankfurt am Main, New York, 1995.

Biba, Marenglen und Xhafa, Fatos, Learning structure and schemas from documents, Berlin, 2011.

Foucault, Michel, Sicherheit, Territorium, Bevölkerung: Vorlesung am Collège de France, 1977-1978, Frankfurt am Main, 2015.

Friedman, Milton, Kapitalismus und Freiheit, München, 2014.

Jank, Gerhard und Jongen, Hubertus Theodorus, Höhere Mathematik für Maschinenbauer: Skript zur Vorlesung, 1999.

Jurish, Bryan, DiaCollo Tutorial: Einführung, online verfügbar unter: http://kaskade.dwds.de/diacollo-tutorial/#introduction.html . Zuletzt geprüft am: 12.10.2017.

Lemke, Matthias und Stulpe, Alexander, Blended Reading: Theoretische und praktische Dimensionen der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung, in: Matthias Lemke und Gregor Wiedemann (Hrsg.), Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, Wiesbaden, 2016, S. 17–62.

Leskovec, Jurij, Rajaraman, Anand und Ullman, Jeffrey D., Mining of massive datasets, Cambridge, 2015.

Mehler, Alexander, Kühnberger, Kai-Uwe und Lobin, Henning et al., Modeling, Learning, and Processing of Text Technological Data Structures, Berlin, Heidelberg, 2012.

Schimank, Uwe und Volkmann, Ute, Ökonomisieurng der Gesellschaft, in: Andrea Maurer (Hrsg.), Handbuch der Wirtschaftssoziologie, Wiesbaden, 2008, S. 382–393.

Schumpeter, Joseph Alois, Kapitalismus, Sozialismus und Demokratie, Tübingen, 1993.

van Eijnatten, Joris, Pieters, Toine und Verheul, Jaap, Big Data for Global History: The Transformative Promise of Digital Humanities, in: BMGN - Low Countries Historical Review 128 (2013), Nr: 4, S. 55–77.

Warner, Charles Dudley, American Newspaper, online verfügbar unter: http://www.gutenberg.org/cache/epub/3110/pg3110-images.html . Zuletzt geprüft am: 13.9.2017.

White, Hayden, Der historische Text als literarisches Kunstwerk, in: Christoph Conrad und Martina Kessel (Hrsg.), Geschichte schreiben in der Postmoderne. Beiträge zur aktuellen Diskussion, Stuttgart, 1994, S. 123–157.

Wiedemann, Gregor und Lemke, Matthias, Einleitung: Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, in: Matthias Lemke und Gregor Wiedemann (Hrsg.), Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, Wiesbaden, 2016, S. 1–16.

Abbildungsverzeichnis