

## Kontextbasierte Zitationsanalyse soziologischer Klassiker im Verlauf von 100 Jahren

**Messerschmidt, Reinhard**

reinhard.messerschmidt@uni-koeln.de  
Universität zu Köln, Deutschland

**Mathiak, Brigitte**

mathiak@gmail.com  
Universität zu Köln, Deutschland

Bibliometrische Zitationsanalyse ist in den Naturwissenschaften allgemein üblich geworden, in den Geistes- und Sozialwissenschaften jedoch mit Problemen hinsichtlich der Datenbasis und unterschiedlicher Zitationsweisen konfrontiert. So betonen Sula und Miller (2013), dass verschiedene Referenzkontexte nicht ignoriert werden dürfen, da intellektuelle Dispute zum geisteswissenschaftlichen Kern gehören. Für drei Klassiker der Soziologie haben wir daher die Zitationen in ihrem Zitationskontext, sowie im zeitlichen Verlauf analysiert und verschiedene gängige Hypothesen zu Trends in diesem Bereich statistisch überprüft.

### Datenbasis und Methode

Das zentrale Textkorpus besteht aus digitalisierten Tagungsbänden („Verhandlungen“) der Deutschen Gesellschaft für Soziologie (DGS) von 1910 bis 2010 und umfasst 6869 Dokumente, sowohl direkt konvertiert aus dem Ausgangsmaterial, als auch OCR-behandelte Scans. Beide Dokumenttypen wurden zunächst nach diktions- und n-gram basierter Vorbereitung (Saad und Mathiak 2013) in reinen Text konvertiert. Wie in den Geisteswissenschaften üblich, sind zwar viele wichtige Akteure im Korpus präsent, aber deren Hauptwerke üblicherweise Monografien, für die zusätzliche Quellkorpora konstruiert wurden. Ausgewählt wurden mit Karl Marx, Max Weber und Theodor W. Adorno drei Klassiker, deren gesammelte bzw. ausgewählte Schriften digital in hoher Qualität vorliegen und für die Soziologie selbst eine entscheidende Rolle spielen. Weber ist dabei mit Abstand der am häufigsten zitierte und gilt als Ahnherr der deutschen Soziologie. Marx wurde (gemeinsam mit Friedrich Engels) stark und insbesondere auch kontrovers diskutiert. Dem Werk Adornos kommt in der Soziologie und Sozialphilosophie der 1960er Jahre eine herausragende Stellung zu. Aufgrund des selektiven Charakters des DGS-Korpus bezüglich jeweiliger Tagungsthemen wurde zur Ergänzung ein Korpus aus seit 1949 digital verfügbaren Fachzeitschriften erstellt und annotiert.

Der Fokus des Projekts lag auf der Analyse von Text-ReUse sowie Sentiments in Zitationen und Paraphrasen. Vorhandene Ontologien wie CiTO erwiesen sich aufgrund vieler für unsere Zwecke irrelevanter Kategorien als zu komplex und zeitaufwändig. Aus Effizienzgründen und auch um Erkenntnisse in Bezug auf Sentimentpolarität (Boland et al. 2013) nutzen zu können, haben wir uns primär auf drei Ausprägungen letzterer konzentriert: positiv, negativ und neutral. Weitere im Rahmen von Sentimentanalysen übliche Differenzierungen hinsichtlich z.B. gradueller Abstufung und Subjektivität (Pang und Lee 2008) wurden bewusst ausgeblendet. Nach ersten Annotationsversuchen wurde allerdings klar, dass zusätzliche Kategorien für Ambivalenz und Negationsstrukturen notwendig sind. Insgesamt 3382 Codes wurden dabei interpretativ und kontextbezogen von soziologischen Experten in MaxQDA auf Basis des zuvor beschriebenen Codeschemas annotiert. Im Verlauf dieses Prozesses wurden spezifische Charakteristika des Korpus deutlich, die bei der Analyse zu berücksichtigen waren:

Erstens zeigten sich im Datensatz strukturelle Brüche hinsichtlich der Dokumentanzahl je Tagung. Während diese von 1910 bis 1979 bereits zwischen 9 und 94 variiert, existieren ab 1980 abgesehen von einer Ausnahme zwischen 240 und 675 Dokumente. Dadurch sind absolute Zahlen von Autorennamen nicht vergleichbar (Abb. 1).

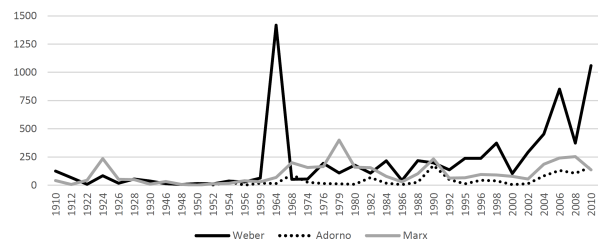


Abbildung 1: absolute Ergebnisse pro Jahrgang (Maximum bei Weber 1964 bedingt durch auf ihn bezogenes Thema der Tagung, Einbruch im Jahr 2000 dagegen durch außergewöhnlich niedrige Zahl von Dokumenten, siehe Abb. 4)

Zweitens ist, wie Sula und Miller (2013) bereits betonten, die Abgrenzung einer Zitation insbesondere in älteren Dokumenten nicht immer klar. Die zunächst simple Keyword-Suche nach Autorennamen führt zu systematischer Überschätzung der Referenzen aufgrund 1) Erwähnung von Autoren in anderem Kontext z.B. von Zusammenfassungen wie „klassische Autoren (insbesondere Simmel, aber auch Marx und Weber)“ oder 2) Biografischen Darstellungen sowie 3) Literaturverzeichnissen. Zusätzlich zeigt sich insbesondere bei Weber das Problem der Autorendisambiguierung, denn nicht jeder Weber ist Max: angefangen von seinem Bruder Alfred über seine Frau Marianne bis zu insgesamt 30 weiteren – teilweise in denselben Dokumenten. Zusätzlich müssen unterschiedliche Zitationsstile sowie

die parallele Zitation mehrerer Werke berücksichtigt werden, da andernfalls eine Unterschätzung vorliegt. Spezifische Abkürzungen wie z.B. siebzig Mal „MWG“ für Webers gesammelte Schriften oder 102-fach „MEW“ für ausgewählte Werke von Marx/ Engels und vergleichsweise weit von der Erwähnung des Autorennamens positionierte Referenzen würden nur allzu leicht übersehen, wenn nicht auch explizit nach diesen gesucht wird.

Dies wirft einerseits die Frage nach der angemessenen Definition einer Referenz im Kontext dieses Projekts auf sowie andererseits danach, welche entsprechende Darstellung im zeitlichen Verlauf adäquat ist. Hinsichtlich ersterer fiel die Entscheidung zugunsten maximierter Offenheit und Inklusion, was insbesondere bei Weber, welcher oft nur mittels Erwähnung des Namens ohne genauen Werkbezug referenziert wird, viele Fälle kanonischer Zitation einschließt und sogar Fälle indirekter Zitation, in welchen andere Autoren referenziert wurden z.B. „in Webers Terminologie (vgl. Habermas 1982)“. Der Grund dafür ist, dass andernfalls unter ausschließlicher Berücksichtigung nur formal korrekter Referenzen deren Fallzahl massiv abnimmt (Abb. 2).

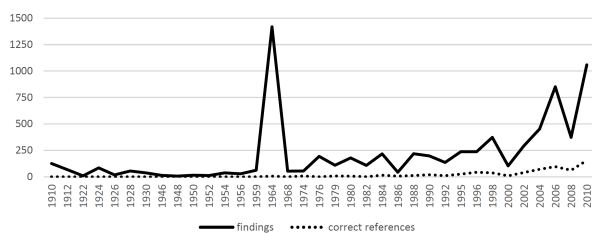


Abbildung 2: absolute Ergebnisse pro Jahrgang vs. Formal korrekte Referenzen für Weber

Im Fall Max Webers stehen nur 640 korrekte Referenzen insgesamt 7381 Suchergebnissen gegenüber, was einem Verlust von 91,3% entspricht, welcher insbesondere darauf hinausläuft, dass alle Referenzen von den 1970er Jahren auf Grund formaler Defizite entsprechend heutiger Standards verloren gehen. Daran wird die Bedeutung sogar der lockeren Erwähnungen sichtbar. Die zweite zuvor aufgeworfene Frage nach der adäquaten Repräsentation erwies sich als kompliziert und wird im folgenden Abschnitt diskutiert.

### Bibliometrische und wissenschaftsgeschichtliche Ergebnisse

Im langfristigen Trend ließe die Interpretation von Abbildung 2, abgesehen vom themenbedingten Ausreißer des Jahres 1964, auf ein zunächst zurückgehendes und dann eine in den 1960er und 70er Jahren anfangs langsame Renaissance hinaus, welche sich im neuen Jahrhundert intensiviert. Dieser Effekt einer sogenannten Weber-Renaissance (Glassman 1983, Hinz 1966) ist auch aus der Fachliteratur bekannt. Werden die Ergebnisse aber in Relation zu den Dokumenten pro Tagung berechnet, wird

es schwieriger, von einer solchen Renaissance zu sprechen, wenngleich der Ausreißer von 1964 bleibt (Abb. 3).

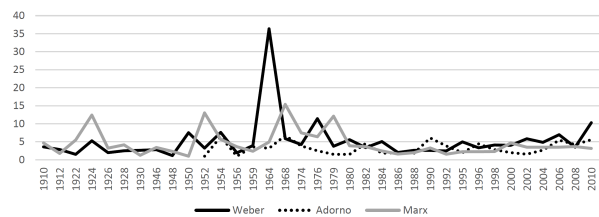


Abbildung 3: durchschnittliche Referenzen pro Dokument (Verhältnis Ergebnisse/ Anzahl der ergebnisbeinhaltenden Dokumente)

Beim Blick auf die Dokumentanzahl pro Jahrgang (Abb. 4) zeigt sich die eingangs erwähnte Heterogenität sowie ein generell zunehmender Trend, was den Unterschied der relativen zur absoluten Darstellung erklärt.

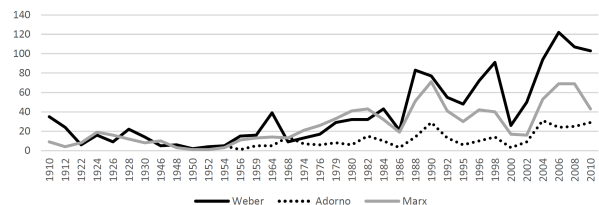


Abbildung 4: absolute Dokumentanzahl pro Jahrgang

Wenn weiterhin diese absolute Dokumentanzahl pro Jahrgang ins Verhältnis zur Gesamtzahl jährlicher Dokumente gesetzt wird (Abb. 5), zeigt sich vielmehr ein abnehmender Trend für alle Autoren.

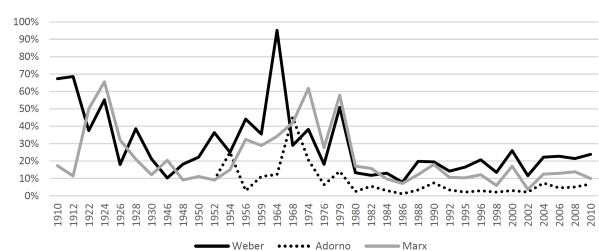
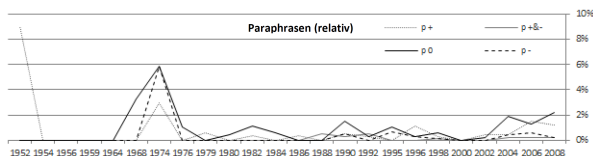
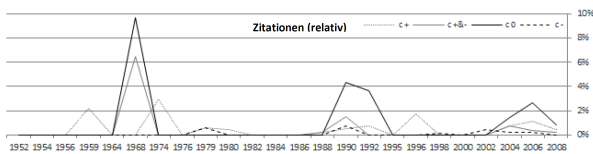


Abbildung 5: Prozentualer Anteil an Dokumenten pro Jahrgang (Verhältnis Anzahl der ergebnisbeinhaltenden Dokumente/ Gesamtzahl der Dokumente jeweils eines Jahrgangs)

Alle gezeigten Grafiken sind Konstruktionen unter Hervorhebung verschiedener Aspekte, wenngleich die letzten beiden hinsichtlich der Intensität zu verschiedenen Jahrgängen informativer erscheinen als die vorangegangenen. Trotz der immer noch vagen Verbindung jeweiliger Zahlen von Referenzen angesichts der zuvor dargestellten

Definitionsprobleme, lässt sich jedoch kein guter Grund für die Präferenz einer der in den letzten beiden Abbildungen dargestellten Berechnung finden. Nichtsdestotrotz zeigen beide Optionen keinesfalls eine Weber-Renaissance.

Demgegenüber zeigt sich bei der für Adorno durchgeführten vertieften kontext- und sentimentbezogenen Analyse ein klar abnehmender Trend. Angesichts der im Vergleich zu Weber viel geringeren Anzahl an Referenzen welche erst ab 1952 auftreten konnten wir die Sentimentpolarität (positiv, ambivalent, neutral, negativ) von Zitationen (Abb. 6) und Paraphrasen (Abb. 7) detailliert annotieren.



Abbildungen 6, 7: Sentimentpolarität für Zitationen und Paraphrasen

(relative Häufigkeiten in Bezug auf Dokumentanzahl pro Jahrgang)

Zunächst zeigen sich in beiden Abbildungen lokale Maxima um 1968, welche angesichts Adornos enormer Rezeption im Kontext der 68er-Bewegung kaum überrascht – im Gegensatz zur leichten Verzögerung bei Paraphrasen, welche möglicherweise durch die zunehmende Bekanntheit seiner Werke bedingt ist. Möglicherweise könnte ein solches Schema generell im Hinblick auf die Entstehung zukünftiger Klassiker auftauchen, was genauer zu untersuchen wäre. Die zweite überraschende Beobachtung besteht in der trotz damals hochgradiger Polarisierung der Disziplin (z.B. im „Positivismusstreit“) starken Häufigkeit neutraler Referenzen. Diese ist jedoch vielfach durch eine spezifische Argumentationsstruktur bedingt, in der nach vielen neutral-deskriptiven Aussagen letztendlich nur wenige polarisierte verwendet werden.

Abschließend können wir berichten, dass die Schwierigkeiten, Geisteswissenschaften durch die „positivistische“ bibliometrische Tradition adäquat abzubilden auch konzeptuell real sind und nicht nur der schwierigen Datenlage angelastet werden können. Ein mehr an derartiger Analyse, die sich jedoch nur auf die kleinen Einheiten der Auseinandersetzung konzentriert, verfehlt das Gesamtbild. Trotzdem kann sie als Hilfsmittel eingesetzt werden, um neue Wege der wissenschaftsgeschichtlichen Annäherung zu eröffnen,

darunter insbesondere die Auseinandersetzung im distant reading (Moretti 2013).

## Fußnoten

1. Soziale Welt, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Deutsche Zeitschrift für Philosophie
2. <http://purl.org/spar/cito/>
3. [http://cceh.uni-koeln.de/share/annotation\\_soc\\_classics.zip](http://cceh.uni-koeln.de/share/annotation_soc_classics.zip)
4. <http://www.maxqda.de/>
5. Die in der Bibliometrie übliche auf Textlänge basierende Berechnung erwies sich angesichts diesbezüglicher Heterogenität des DGS-Korpus als nicht anwendbar.

## Bibliographie

**Adorno, Theodor W.** (2004): „Theodor W. Adorno, Gesammelte Schriften“, in: *Digitale Bibliothek* 97.

**Boland, Katarina/ Wira-Alam, Andias/ Messerschmidt, Reinhard** (2013): „Creating an annotated corpus for sentiment analysis of German product reviews“, in: *GESIS-Technical Reports* 2013/05 [http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis\\_reihen/gesis\\_methodenberichte/2013/TechnicalReport\\_2013-05.pdf](http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2013/TechnicalReport_2013-05.pdf) [letzter Zugriff 19. August 2016].

**Glassman, Ronald** (1983): „The Weber renaissance“, in: *Current Perspectives in Social Theory* 4: 239–271.

**Hinz, Horst** (1966): „Max-Weber-Renaissance?“, in: *Vierteljahreshefte zur Wirtschaftsforschung* 4: 454–479.

**Marx, Karl / Engels, Friedrich** (2004): „Marx, Engels, ausgewählte Werke“, in: *Digitale Bibliothek* 11.

**Moretti, Franco** (2013): *Distant Reading*. London: Verso.

**Pang, Bo / Lee, Lillian** (2008): „4.1.2 Subjectivity Detection and Opinion Identification“, in: *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

**Sula, Chris Alen / Miller, Matt** (2013): „Citation studies in the humanities“, in: *DH2013: Conference Abstracts* <http://dh2013.unl.edu/abstracts/ab-353.html> [letzter Zugriff 19. August 2016].

**Saad, Farag / Mathiak, Brigitte** (2013): „Revised mutual information approach for german text sentiment classification“, in: *WWW '13 Companion. Proceedings of the 22nd international conference on World Wide Web* 579–586. <http://dl.acm.org/citation.cfm?id=2487788.2487997> [letzter Zugriff 19. August 2016].

**Weber, Max** (2004): „Max Weber, gesammelte Werke“, in: *Digitale Bibliothek* 58.