

Vom Schürfen und Knüpfen – Text Mining und Netzwerkanalyse für Historiker_innen

Veranstalter: Centrum für Religionswissenschaftliche Studien (CERES), Ruhr-Universität Bochum (RUB)

Datum, Ort: 10.04.2015–12.04.2015, Bochum

Bericht von: Peter Fleer, Dienst Historische Analysen, Schweizerisches Bundesarchiv, Bern

Unter dem Titel „Vom Schürfen und Knüpfen – Text Mining und Netzwerkanalyse für Historiker_innen“¹ hatte das Centrum für Religionswissenschaftliche Studien (CERES)² der Ruhr-Universität Bochum (RUB) vom 10. bis 12. April 2015 zu einem internationalen Workshop eingeladen. Der Workshop fand im Rahmen der Reihe „Historische Netzwerkforschung“³ statt. Die Vorträge erörterten einerseits neueste Forschungen im Bereich der historischen Religionswissenschaft. Andererseits wurde in *hands-on* Workshops die Möglichkeit geboten, IT-Werkzeuge kennenzulernen. Nachfolgend präsentieren wir einen Überblick über die Präsentationen und Diskussionen in den einzelnen Vorträgen und Workshops.

Das SeNeReKo-Projekt

Der Workshop stand im Zusammenhang mit dem Projekt SeNeReKo (Semantisch-soziale Netzwerkanalyse als Instrument zur Erforschung von Religionskontakten)⁴. SeNeReKo ist ein gemeinsames Forschungsvorhaben des CERES und des Trier Center for Digital Humanities (TCDH). Das Projekt führt im Kontext der „Digital Humanities“ Geistes- und Sozialwissenschaften einerseits und Informatik andererseits zusammen. Ziel ist es, die vorhandenen digitalen Materialien für Fragestellungen der Religionswissenschaft fruchtbar zu machen und in diesem Zuge auch neue methodische Verfahren zu entwickeln.

Die Historische Netzwerkforschung beschäftigt sich in erster Linie mit der Vernetzung und Interaktion historischer Akteure (Personen, Organisationen, Gesellschaften, Nationen). Sie wendet dabei moderne computergestützte netzwerkanalytische Methoden an. Analysiert und interpretiert werden vielfältige Merkmale von meist sozialen Netz-

werken. Im Vordergrund stehen dabei nicht die einzelnen Akteure, sondern deren Beziehungen untereinander. Beziehungen stellen in diesem Sinn die kleinste Analyseeinheit der Netzwerkanalyse dar.

Vom Schürfen und Knüpfen

In ihrer Keynote „Developing Practical Solutions to Real World Problems by Going from Words to Networks“⁵ stellte JANA DIESNER⁶ (Illinois) eines ihrer Projekte vor. Dabei geht es darum, computergestützte Methoden zu entwickeln, die die Wirkung von sozialen und politischen Kampagnen und investigativen Medienbeiträgen messen. Dies wird gemäß der Referentin in Zukunft wichtiger, da die Geldgeber dieser Projekte wissen wollen, ob und inwiefern die von ihnen unterstützten Projekte die kritisierten Zustände verbessert und zu den intendierten Veränderungen geführt haben. Diesner und ihr Team werteten dazu große Mengen an digital verfügbaren Medienberichten und Social Media-Quellen mit elaborierten Textmining- und Netzwerkanalyseverfahren aus, um aus den Rohdaten handlungsrelevantes Wissen zu generieren. Der *Impact Assessment Approach* von Diesner hat gezeigt, dass damit die Zielerreichung von politischen Kampagnen differenziert gemessen werden kann. Voraussetzung dazu ist die klare Zieldefinition durch die Sponsoren und Kontrollorgane. Die Weiterentwicklung des Verfahrens, so die Referentin, hängt weniger von technischen Fortschritten in der Algorithmus-Programmierung ab als von der theoretischen Verfeinerung des Modells.

ANDREAS KUCZERA (Gießen) demonstrierte in seinem Vortrag „Digitale Farbenspiele oder nützliches Werkzeug – Visualisierung von Netzwerken aus den Registern von Editions- und Regestenwerken“ das Poten-

¹ <<http://senereko.ceres.rub.de/de/hnrws2015/>>.

² <<http://www.ceres.ruhr-uni-bochum.de/>>.

³ <<http://historicalnetworkresearch.org/workshop-series/>>.

⁴ <<http://www.ceres.rub.de/de/project/forschungsprojekte/senereko/>>.

⁵ < http://www.senereko.ceres.ruhr-uni-bochum.de/static/uploads/senereko/hnrws15/diesner.practical_solutions_hnrws2015.pdf >.

⁶ <<http://people.lis.illinois.edu/~jdiesner/index.html>>.

tial der Open-Source-Visualisierungssoftware Gephi zur Visualisierung von Netzwerkstrukturen im Personenregister eines Urkundenbuchs. Dabei ging es zunächst darum, den Text aus den Regesten mittels Textmining-Verfahren in maschinenlesbare Formate (XML und/oder Tabellen) zu transformieren. Dazu importierte der Referent zunächst den Regesten-Volltext in eine Graphendatenbank. Mithilfe dieser Datenbank liessen sich strukturierte Subsets der Daten erstellen (sogenannte Subgraphen), die anschliessend nach Gephi exportiert und dort weiterbearbeitet und in aussagekräftige Netzwerkgrafiken transformiert wurden. Kuczera zeigte auf, dass die computergestützte Netzwerk-Visualisierung von Regesten insbesondere heuristischen Mehrwert bietet, indem sie den Blick auf interessante Zusammenhänge in den Quellen lenken, welche vorher auf Grund der Datenmasse nicht sichtbar gemacht werden konnten. Allerdings müsse vor vorschnellen inhaltlichen Schlussfolgerungen gewarnt werden. Entscheidend sei in jedem Fall die Datenqualität. So liessen etwa im Fall von Registern gemeinsame Nennungen von Personen in Urkunden oder Regesten nur begrenzt Schlüsse auf deren soziale Beziehungen zu.⁷

Am Beispiel der Online-Plattform Trismegistos⁸, einer Datenbank mit persönlichen Textzeugnissen aus Griechenland und Ägypten im Zeitraum zwischen 800 vor und 800 nach Christus, illustrierte SILKE VANBESELAERE (Leuven / London) in ihrem Workshop-Beitrag „Digital Prosopography and Network Analysis“, wie mit Hilfe von Sozialer Netzwerkanalyse historische Akteure in einem großen Datenset eindeutig identifiziert werden können. Vanbeselaere kombiniert traditionelle prosopographische Forschungsmethoden mit digitalen Verfahren zur Netzwerkgenerierung und -interpretation. Zur Visualisierung setzt auch sie Gephi ein. Die Netzwerk-Visualisierungen dienen ihr zum Erkennen von Gemeinschaften und zur Analyse der sozialen Beziehungen innerhalb dieser Gemeinschaften. Zum Beispiel lassen sich damit Kernfamilienbeziehungen von weiteren Verwandtschaftsbeziehungen differenzieren oder Verwandtschaftsgruppen von wirtschaftlichen oder politischen Netzwerken unterscheiden.

DANIEL REUPKE (Bayreuth / Saarbrücken) machte sich unter dem Titel „HNF2.0?! – Überlegungen zu Vergangenheit und Zukunft eines modischen Forschungsansatzes“⁹ Gedanken zu den Entwicklungen im Bereich der Historischen Netzwerkforschung (HNF). Anhand eines Vergleichs eines 2005 abgeschlossenen Forschungsprojekts über Netzwerke der Kreditvergabe in einer ländlichen Grenzregion des 19. Jahrhunderts mit einem geplanten Projekt über Musiker- und Musiknetzwerke zeigte er den Kontrast zwischen den teilweise gescheiterten Ansätze von Netzwerkanalysen in der von ihm so benannten Phase HNF1.0 und den neu erwachsenen Möglichkeiten der HNF2.0. Diente HNF1.0 in erster Linie als Organisations- und Ordnungssystem für Informationen sowie zur Visualisierung der Erkenntnisse und zur Mustererkennung, ermögliche HNF2.0 nun die Bearbeitung von großen digitalen Datenbeständen, automatische Extraktionsverfahren und die maschinelle Netzwerkgenerierung. Dies schaffe Raum für neue Erklärungsansätze und Möglichkeiten für die Nutzung der generierten Daten. Am Beispiel der Korrespondenz von Georg Philipp Telemann liessen sich so dessen persönliches und geografisches Netzwerk darstellen. Geplant ist weiter, dessen Musiknetzwerk aufgrund von digitalisierten Notenmanuskripten und der maschinellen Erkennung von Akkordmustern und Harmoniestrukturen aufzuschlüsseln.

In seinem Referat „Transforming Indexes Locorum into Citation Networks“ analysierte MATTEO ROMANELLO (London) Zitierhinweise auf kanonische Texte. Diese Hinweise in der Sekundärliteratur seien wichtig, indem sie Beziehungen zwischen Texten herstellten, die im selben Zusammenhang zitiert wurden. Gegenüber der manuellen Indexierung dieser Zitierungen böten digitale Verfahren die Möglichkeit, viel größere Informationsmen-

⁷ Vgl. auch Andreas Kuczera, Digitale Farbenspiele oder nützliches Werkzeug – Visualisierung von Netzwerken aus den Registern von Editions- und Regestenwerken, in: Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte, 8. Januar 2015, <<http://mittelalter.hypotheses.org/5089>> (ISSN 2197-6120) (20.07.2015).

⁸ <<http://www.trismegistos.org/>>.

⁹ <<http://www.senereko.ceres.ruhr-uni-bochum.de/static/uploads/senereko/hnrws15/reupke.hnf20.hnrws2015.pdf>>.

gen zu bearbeiten. Der Autor hat dazu Verfahren zur automatischen Extraktion von Zitationen als *Named Entities* entwickelt. Das Verfahren verläuft in drei Hauptschritten: 1. Extraktion und Klassifizierung der Bestandteile einer Zitation (*named entity extraction*), 2. Extraktion der Beziehungen zwischen den Zitations-Bestandteilen (*relations detection*), 3. Eindeutige Identifizierung der zitierten Autoren (*named entity disambiguation*).

Der Autor sieht den Mehrwert, den die Darstellung von Zitationsindices als Netzwerke bietet, darin, dass die zitierten Autoren, Werke und Textstellen nicht isoliert dastehen, sondern deren Beziehungen untereinander gemessen, gefiltert und visualisiert werden können. Es geht ihm somit nicht bloss um eine neue Art der Repräsentation, sondern um einen radikalen Wandel in der Art und Weise wie die Information aus Indices für die Analyse von intertextuellen Zusammenhängen genutzt werden können.

ISTVAN CZACHESZ (Heidelberg) illustrierte in seinem Vortrag „The Study of Word Co-occurrence Networks in the Greek New Testament“¹⁰ anhand einer Fallstudie, wie Netzwerktheorie auf die Analyse historischer Texte angewendet werden kann. Er zeigte, dass Zentralitäts-Messungen von Knoten in Wort-Co-occurrence-Netzwerken von Bibelstellen interessante Informationen über die Kernthematik und die Gestaltungseigenschaften dieser Stellen liefern, die über die Erkenntnisse hinausgehen, die mit traditionellen Methoden gewonnen werden können. Zentralität gibt insbesondere Aufschluss über indirekte Beziehungen zwischen Nachbarn zweiten oder dritten Grads, die keine direkten Kanten (Beziehungen) untereinander aufweisen.

Der Beitrag von DARIYA RAFIYENKO (Leipzig) „Gicht und ihre Formalisierung“¹¹ ging von der Hypothese aus, dass das Hauptanliegen der linguistischen Semantik die Erforschung der Bedeutung von sprachlichen Zeichen und Zeichenketten ist, die historische Semantik dagegen einen interdisziplinär und kulturwissenschaftlich angelegten Semantikansatz verfolgt. Dieser zeichne sich dadurch aus, dass er die Genese, Entwicklung und das Verschwinden von Strukturen semantischen Wissens diachron zu entdecken versuche. Die

historische Semantik gehe davon aus, dass die Gleichförmigkeit des immer wieder Gesagten einen bestimmten Diskurs innerhalb der Gesellschaft widerspiegeln. Bezogen auf antike Texte, so die Referentin, ist der Imperativ des Wiederlesens der antiken Texte ein integraler Bestandteil des historischen Wahrheitsfindungsprozesses.

Anhand von Methoden der Statistik und Computerlinguistik untersuchte Rafiyenko über einen Zeitraum von 2000 Jahren (8. Jh. v. Chr. bis 15. Jh. n. Chr.) Texte, die im Zusammenhang mit Gicht stehen. Dabei stellten sich erhebliche Probleme bei der Formalisierung der Begrifflichkeit. Zur Lösung dieser Probleme setzte die Autorin insbesondere auch komplexe mathematische, wahrscheinlichkeitstheoretische Verfahren ein.

LIISI LAINESTE und MARI SARV (Tartu) stellten in ihrem Vortrag „Folklore and Social Networks“¹² eine Netzwerk-Untersuchung der Kommunikationsmuster innerhalb der estnischen Folkloreszene vor. Unter anderem konnten sie damit geografische Muster der Ausbreitung bestimmter kultureller Praktiken aufzeigen. Die Autorinnen analysierten etwa die Häufigkeit von bestimmten Kategorien in Volksliedern (Menschen, konkrete Objekte, Tiere, abstrakte Phänomene, übernatürliche Wesen). Auf dieser Grundlage konnten sie mittels Netzwerkanalyse unter anderem drei volksmusikalische Kulturregionen in Estland unterscheiden.

FREDERIK ELWERT (Bochum) gab im Rahmen eines Workshops eine Einführung in grundlegende Konzepte und Funktionalitäten von Gephi¹³, eine der gegenwärtig meistverbreiteten Open-Source-Netzwerkanalyse-Anwendungen. Damit können auch große, komplexe Netzwerke erstellt werden. Eine Stärke von Gephi liegt in den vielseitigen Visualisierungsmöglichkeiten.

Gephi erlaubt den Import von CSV-Files

¹⁰ < http://www.senereko.ceres.ruhr-uni-bochum.de/static/uploads/senereko/hnrws15/czachesz_networks_nt_hnrws2015.pdf>.

¹¹ < http://www.senereko.ceres.ruhr-uni-bochum.de/static/uploads/senereko/hnrws15/rafiyenko_kahl.gicht_hnrws2015.pdf>.

¹² < http://www.senereko.ceres.ruhr-uni-bochum.de/static/uploads/senereko/hnrws15/sarv_laineste.folklore.hnrws2015.pdf>.

¹³ < <https://gephi.github.io/>>.

z.B. aus ConText. Benötigt werden in der Regel zwei Tabellen, eine für die Knoten (*nodes*), eine für die Kanten (*edges*). Diese können im Data Laboratory von Gephi zusätzlich weiter aufbereitet werden. Zur Netzwerkvisualisierung stellt Gephi Statistik-, Filter-, Partitionierungs- und Ranking-Instrumente mit unzähligen Optionen zur Verfügung. Daher ist die Netzwerkerstellung mit Gephi, trotz der einfachen Bedienbarkeit, nicht ohne vertiefte Kenntnis theoretischer Netzwerkkonzepte und der darauf basierenden Programm-Funktionalitäten und deren Abhängigkeiten möglich.

Unter dem Workshop-Titel „Analyzing Words and Networks with ConText“¹⁴ stellte JANA DIESNER (Illinois) das von ihr und ihrem Team entwickelte Tool ConText¹⁵ vor. ConText strukturiert den Workflow der Datenaufbereitung und Visualisierung und bietet Unterstützung bei der sogenannten *relation extraction*, das heißt bei der Herstellung von netzwerkfähigen Daten aus strukturierten und unstrukturierten Texten sowie bei der Analyse von Text- und Netzwerkdaten.

Der von ConText unterstützte Workflow umfasst unter anderem die Datenerhebung aus unterschiedlichen Quellen, die Erstellung von bereinigten Datenkorpora, Pre-processing-Techniken (*stemming*, *parts-of-speech-tagging*), Verdichtungstechniken (*topic modelling*, Termgewichtungstechniken), *Sentiment Analysis*, Termkategorisierungsverfahren (*entity detection*), Verfahren zur Beziehungs-Extraktion, Visualisierung von Text-mining-Resultaten (zum Beispiel Export nach Gephi).

Datengrundlage können Daten aus dem Internet oder selber generierte Text-, und CSV-Files sein. Damit können Inputdaten und zur Bearbeitung nötige Daten wie Codebooks bereitgestellt und mit ConText verarbeitet werden.

Der Workshop „Text Network Analysis“¹⁶ unter der Leitung von FREDERIK ELWERT (Bochum) stellte sodann die Plattform WebLicht¹⁷ vor, die im Rahmen des CLARIN-D-Projekts¹⁸ aufgebaut wird. CLARIN-D ist eine web- und zentrenbasierte Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften. Sie verfolgt das Ziel linguistische Daten, Werkzeuge und Dienste in ei-

ner integrierten, interoperablen und skalierbaren Infrastruktur für die Fachdisziplinen der Geistes- und Sozialwissenschaften bereitzustellen. Das Projekt wird vom Bundesministerium für Bildung und Forschung (BMBF) gefördert (Laufzeit bis 31. 5. 2016). Der Zugang zur Forschungsinfrastruktur ist nicht *open-access*, sondern nur via dedizierte Institutionen möglich.

Im Unterschied zur sozialwissenschaftlichen Netzwerkanalyse, die auf Personen oder Gruppen als zentrale Einheiten gerichtet ist und soziale Strukturen untersucht und die Texte als Informationsquellen über Akteure und ihre Beziehungen nutzt, fokussierte der Workshop auf die Texte selbst. In dieser Perspektive werden Wörter (nicht Akteure) zu Knoten. Beziehungen zwischen Wörtern werden von gemeinsamen Kontexten oder von ähnlichen Gebrauchsmustern im Text abgeleitet. Netzwerke können so verwendet werden, um Textstrukturen, semantische Beziehungen oder Bedeutungsfelder zu analysieren und zu visualisieren.

Der Workshop zeigte verschiedene Ansätze der Text-Netzwerkanalyse auf. Er demonstrierte anhand der Infrastruktur von WebLicht, wie Verfahren des *natural language processing* für die Aufbereitung von Texten eingesetzt werden, wie diese Texte anschließend in Netzwerke konvertiert und wie diese Netzwerke visualisiert und analysiert werden können.

Fazit

Insgesamt bot der Anlass wertvolle Anregungen zu Verfahren und Tools zur computergestützten Netzwerkanalyse von historischen Texten. Interessant war der Ansatz des Workshops, die Verfahren zur Analyse von sozialen Beziehungen in Gemein- oder Gesellschaften auf die Beziehungen zwischen Elementen (in erster Linie Wörter) innerhalb von Texten anzuwenden.

¹⁴ < http://www.senereko.ceres.ruhr-uni-bochum.de/static/uploads/senereko/hnrws15/diesner.context_hnrws2015.pdf>.

¹⁵ < <http://context.lis.illinois.edu/>>.

¹⁶ < <http://senereko.ceres.rub.de/de/hnrws2015/programm/text-network-analysis/>>.

¹⁷ < <https://weblicht.sfs.uni-tuebingen.de/WebLicht4/>>.

¹⁸ < <http://www.clarin-d.de/de/home.html>>.

Verschiedene Aspekte der Netzwerkanalyse kamen zur Sprache. Eine theoretische Vertiefung in die Netzwerkanalyse fand indes nicht statt. Die Visualisierung und Interpretation von Netzwerken blieb dementsprechend im Beispielhaften. Der Schwerpunkt lag auf praktisch orientierten Präsentationen von Tools und Verfahren, wobei hier die Datenaufbereitung den größten Raum einnahm. Vereinfacht ging es dabei darum, Wörter und deren Beziehungen aus mehr oder weniger strukturierten Fliesstexten in einer geeigneten Tabellenform darzustellen. Grundsätzlich entstehen dabei zwei Tabellen, eine für Knoten (meist Wörter repräsentierend) und eine für Kanten (aus den angewandten Textanalyseverfahren gewonnene Beziehungen repräsentierend).

Bezüglich des Mehrwerts der maschinellen gegenüber herkömmlichen Verfahren wurden in den vorgestellten Fallbeispielen und Workshops vor allem zwei Aspekte immer wieder erwähnt: die Möglichkeit, gegenüber herkömmlichen Methoden viel größere Datenmengen (Stichwort „Big Data“) zu bewältigen, und der heuristische Nutzen, indem neue Sichten und Fragestellungen entstehen. Betont wurde zudem mehrmals, dass es wichtig sei, die Verfahren (inklusive der Datenaufbereitung) zur Netzwerkgenerierung und zur Interpretation von Visualisierungen auf ein solides theoretisches Fundament zu stellen. Impliziter kam zum Ausdruck, dass die Möglichkeiten der Netzwerkanalyse weniger von der Technik als von den Daten und deren Qualität begrenzt werden.

Konferenzübersicht:

KEYNOTE

Jana Diesner (University of Illinois Urbana-Champaign), Developing Practical Solutions to the Real World Problems by Going from Words to Networks

WORKSHOP SESSIONS

Sven Sellmer (Ruhr-Universität Bochum), Introduction to R

Frederik Elwert (Ruhr-Universität Bochum), Introduction to Python

Frederik Elwert (Ruhr-Universität Bochum), Introduction to Gephi

Jana Diesner (University of Illinois Urbana-Champaign), Analyzing Words and Networks with ConText

Silke Vanbeselaere (Katholieke Universiteit Leuven & King's College London), Digital Prosopography and Network Analysis

Frederik Elwert (Ruhr-Universität Bochum), Text Networks

PAPER SESSIONS

Andreas Kuczera (Justus-Liebig-Universität Gießen), Digital Color Play or Useful Tool – Visualisation of Networks out of Registers of Editions and Regests

Daniel Reupke (Universität Bayreuth & Universität des Saarlandes), HNF2.0?! – Überlegungen zu Vergangenheit und Zukunft eines modischen Forschungsansatzes

Matteo Romanello (Deutsches Archäologisches Institut & King's College London), Transforming Indexes Locorum into Citation Networks

Istvan Czachesz (Universität Heidelberg), The Study of Word Co-Occurrence Networks in the Greek New Testament

Dariya Rafiyenko (Universität Leipzig), Gicht und ihre Formalisierung

Liisi Laineste & Mari Sarv (beide Eesti Kirjandusmuuseum, Tartu), Folklore and Social Networks

Tagungsbericht *Vom Schürfen und Knüpfen – Text Mining und Netzwerkanalyse für Historiker_innen*. 10.04.2015–12.04.2015, Bochum, in: H-Soz-Kult 17.09.2015.