RUHR-UNIVERSITÄT BOCHUM

RUB

DAS SENEREKO-PROJEKT

Netzwerkanalyse für die Religionsgeschichte

Zielsetzung und Methoden

In der Religionswissenschaft gibt es ein wachsendes Interesse an relationalen Perspektiven auf religionsgeschichtliche Phänomene. Das Projekt „Semantisch-soziale Netzwerkanalyse als Instrument zur Erforschung von Religionskontakten“ (SeNeReKo) wendet Konzepte aus der Netzwerkforschung an, um Relationen in religionshistorischem Textmaterial sichtbar zu machen. Dabei werden nicht nur soziale Netzwerke, also Beziehungen zwischen Akteuren, sondern insbesondere auch semantische Strukturen untersucht.

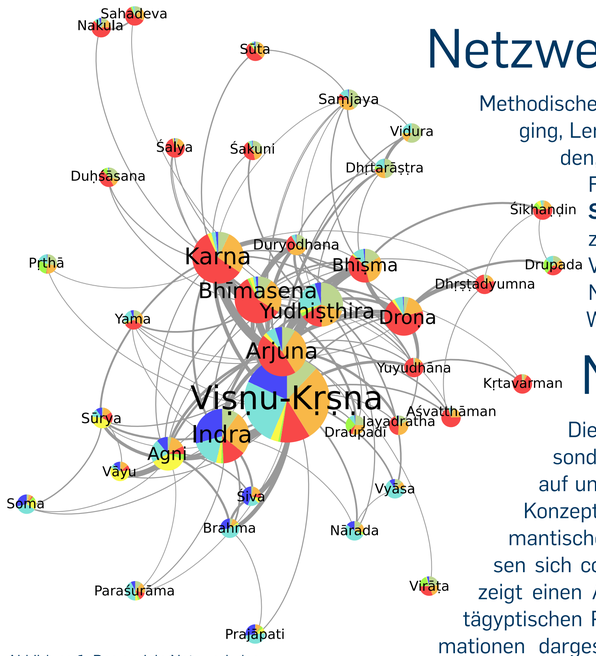


Abbildung 1: Das soziale Netzwerk des Mahābhārata mit Topic-Verteilung (Ausschnitt)

Netzwerk-Extraktion

Methodische Grundlage des Projekts ist die Extraktion von Netzwerkdaten aus den untersuchten Textcorpora. Dafür ist eine linguistische Annotation (etwa Part-of-Speech-Tagging, Lemmatisierung und Named-Entity-Recognition) des Textmaterials erforderlich. Diese Annotationen können teilweise aus den verfügbaren Corpora übernommen werden, teilweise werden sie mithilfe computerlinguistischer Werkzeuge ergänzt.

Für die Extraktion von Netzwerken aus dem Textmaterial werden verschiedene Methoden angewandt, die unterschiedliche Strukturaspekte der Corpora hervorheben. **Soziale Netzwerke** zeigen die Relationsstrukturen der in den Textcorpora auftretenden Personen, Helden und Götter. **Semantische soziale Netzwerke** ergänzen soziale Netzwerke um Informationen über die semantischen Konnotationen von Entitäten und Relationen (Abbildung 1). Hierfür wird ein relationales Topic-Modeling-Verfahren angewandt[1]. **Semantisch/soziale Netzwerke** bringen die semantische Ebene in das Netzwerkmodell ein, indem die Kontexte von Personennennungen als Netzwerk von Entitäten und Lemmata modelliert werden (Abbildung 2). **Textnetzwerke** schließlich stellen ganze Texte als Netzwerk von Wörtern dar. Hierfür können Wortkookkurrenzen[2], syntaktische Dependenzrelationen[3] oder Verteilungsähnlichkeiten[4] zugrunde gelegt werden.

Netzwerkanalyse

Die extrahierten Netzwerke können mithilfe üblicher Verfahren analysiert werden. Angewandt werden insbesondere Zentralitätsmaße und Clustering-Algorithmen (community detection). Analysen können sich dabei auf unterschiedliche Ebenen beziehen: Auf der Textebene können thematische Strukturen und zentrale Konzepte visualisiert werden. Auf Corpusebene können übergreifende Strukturen sozialer und semantischer Kontexte analysiert werden. Sowohl für einzelne Akteure als auch für Lemmata lassen sich corpusweit zentrale kookkurrente Wörter und Wortgruppen identifizieren. Abbildung 2 zeigt einen Ausschnitt aus dem Kontext-Netzwerk des Gottes „Horus“ aus dem Corpus der altägyptischen Pyramidentexte. Es werden sowohl Entitäten, Toponyme als auch semantische Informationen dargestellt. In ähnlicher Weise ist das Verhältnis der altägyptischen Konzepte „Maat“ (Ordnung, Gerechtigkeit) und „Heka“ (Kraft, Magie) analysiert worden[5].

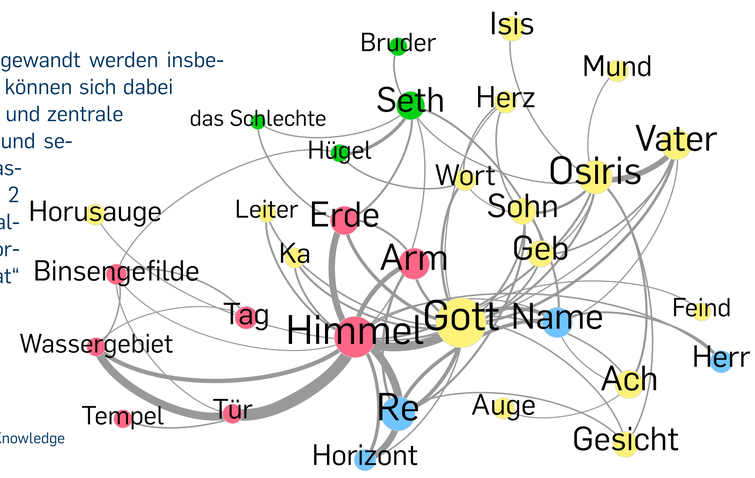


Abbildung 2: Sozialer und semantischer Kontext des altägyptischen Gottes Horus in den Pyramidentexten

Literatur

[1] Chang, Jonathan, Jordan Boyd-Graber, and David M. Blei. 2009. 'Connections between the Lines: Augmenting Social Networks with Text'. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 169–78. ACM.
[2]Paranyushkin, Dmitry. 2011. Identifying the Pathways for Meaning Circulation Using Text Network Analysis. Nodus Labs. http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/.
[3] Elwert, Frederik. 2014. 'Network Analysis between Distant Reading and Close Reading'. In Reading Historical Sources in the Digital Age. Selected Papers from the DHLU2013 Conference. Luxembourg (im Erscheinen).
[4] McClure, David. 2014. '(Mental) Maps of Texts'. http://dclure.org/essays/mental-maps-of-texts/.
[5] Hofmann, Beate, and Frederik Elwert. 2014. „Heka Und Maat. Netzwerkanalyse Als Instrument Ägyptologischer Bedeutungsanalyse“. In „Vom Leben Umfangen“. Ägypten, Das Alte Testament Und Das Gespräch Der Religionen. Gedenkschrift Für Manfred Görg, hg. von Georg Gafus and Stefan Wimmer, 235–45. Ägypten Und Altes Testament 80. Münster: Ugarit.

Corpora

Die Grundlage des Projekts bilden drei religionshistorische Corpora. Dadurch können unterschiedlichen Fragestellungen nachgegangen werden. Zudem wird sichergestellt, dass die entwickelten Methoden sprachübergreifend anwendbar sind.

Altägyptisches Corpus

Grundlage der altägyptischen Texte bildet die Datenbank „Thesaurus Linguae Aegyptiae“ der Berlin-Brandenburgischen Akademie der Wissenschaften. Das Textmaterial aus über 2000 Jahren Geschichte ist sowohl lemmatisiert als auch morphosyntaktisch annotiert und besteht aus unterschiedlichen Textgattungen wie religiösen, medizinischen oder auch literarischen Schriften. Durch ein solch heterogenes Corpus kann den verschiedensten Fragestellungen nachgegangen werden. Darunter fallen beispielsweise Kontextanalysen von religiösen Begriffen, aber auch Untersuchungen zum semantischen Umfeld religiöser Entitäten.

Stele 116, Louvre, Paris

CC By-SA Guillaume Blanchard, http://commons.wikimedia.org/wiki/File:Egypte_louvre_116_stele.jpg

Thai-Manuskript eines kleinen Abschnitts aus dem Pāli-Kanon

CC By-SA http://commons.wikimedia.org/wiki/File:Tipitaka_scripture.jpg

Pāli-Kanon

Der Pāli-Kanon ist eine mehr als zwei Millionen Wörter umfassende Sammlung buddhistischer Texte in der mittelindischen Sprache Pāli. Entstanden in Nordindien im Laufe der ersten Jahrhunderte v.u.Z. und weitgehend kodifiziert im heutigen Sri Lanka um die Zeitenwende herum, liegt der Kanon mittlerweile in digitalisierter Form vor. In ihm werden zahlreiche Situationen geschildert, in denen der Buddha und seine Anhänger auf rivalisierende religiöse Gruppen, individuelle Kritiker und fremde Lehren reagieren, so dass er eine reiche Quelle für die interreligiöse Dynamik im damaligen Südasien darstellt. Da der Pali-Kanon für traditionelle Analysemethoden zu umfangreich ist, arbeiten wir an computergestützten Verfahren, um ein klares Bild der in den Texten enthaltenen sozialen und semantischen Strukturen zu erhalten.

Mahābhārata

Das altindische Sanskrit-Epos Mahābhārata ist ein ebenso umfangreicher (über 1,5 Mio. Wörter) wie gehaltvoller Text. Es ist das Produkt eines mehrere Jahrhunderte (ca. 5. Jh. v. – 5. Jh. n.u.Z.) umfassenden Entstehungsprozesses und enthält vielfältige Informationen über die altiindsche Weltanschauung und Gesellschaft sowie über die Herausbildung des klassischen Hinduismus. Netzwerkanalytische Verfahren erlauben es uns, Beziehungen zwischen zentralen Konzepten und Akteuren zu visualisieren und so neue Einblicke in die semantische Struktur des damaligen Denkens zu gewinnen.

Kīcaka und Bhīmasena. Blatt aus einer Mahābhārata-Handschrift. 1670. Brooklyn Museum, Geschenk von Cynthia Hazen Polsky

Tools & Technologie

Wo möglich nutzt das Projekt etablierte Standards und stellt seine entwickelte Software unter einer Open-Source-Lizenz bereit. Die Werkzeuge des Projekts sind grundsätzlich sprachunabhängig und können so auch für andere Projekte genutzt werden.

Für weitere Informationen und Downloads siehe https://github.com/SeNeReKo.

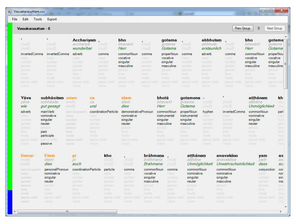
Datenverwaltung und -verarbeitung

Die verwendeten Korpora werden in unterschiedlichen, nicht-standardisierten Formaten bereitgestellt. Zunächst werden sie in ein gemeinsames TEI-Schema konvertiert und in einer XML-Datenbank abgelegt. Die Werkzeuge zur weiteren linguistische Verarbeitung und zur Netzwerkextraktion nutzen das linguistische Annotationsformat TCF zum Datenaustausch. Sie sind so mit der CLARIN WebLicht-Infrastruktur kompabitel. Die im Rahmen des Projekts entwickelte Python-Bibliothek TCFlib ermöglicht es, TCF-Dateien zu verarbeiten und WebLicht-kompatible Services aufzurufen und zu erstellen.

Extraktion der Netzwerke

Im Rahmen des Projekts wurden verschiedene Methoden zur Netzwerkextraktion wie Kookkurrenz- und Dependenznetzwerke implementiert. Sie stehen im TCFnetworks-Paket zur Verfügung. Durch die Nutzung der WebLicht-Infrastruktur oder anderer TCF-kompatibler Systeme können Texte annotiert werden und als Eingabe für die Netzwerkerzeugung dienen.

Text-Annotation



Um NLP-Tools trainieren zu können, müssen zuerst Daten manuell annotiert werden. Für diese Aufgabe nutzen wir aus Effizienzgründen unsere speziell zu diesem Zweck entworfene Software, welche auch spezielle Eigenheiten des Pāli berücksichtigt.

Wörterbuch-Server

Für NLP-Aufgaben ist ein gutes Wörterbuch unverzichtbar. Für Pāli gibt es jedoch kein geeignetes digitales Wörterbuch. Daher wurden unterschiedliche Wörterbuchquellen im Rahmen des Projekts aufbereitet und in einen speziellen Wörterbuch-Server integriert. Dieser Server ist eine NodeJS/MongoDB-Software und stellt ein REST-API bereit. Diese Infrastruktur ermöglicht eine algorithmische und manuelle Datenverarbeitung der Wörterbucheinträge, um die Daten so mit maschinenlesbaren Informationen anreichern zu können.

http://senereko.ceres.rub.de/

CERES

Prof. Dr. Volkhard Krech
Center for Religious Studies
Ruhr-University Bochum
Team: Frederik Elwert, Marek Firlej, Simone Gerhards, Beate Hofmann, Manuel Pachurka, Sven Sellmer, Ayleen Winkler, Sven Wortmann

Prof. Dr. Claudine Moulin
Trier Center for Digital Humanities
Trier University
Team: David Alfter, Thomas Burch, Jürgen Knauth

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung