RUHR-UNIVERSITÄT BOCHUM

RUB

NETWORK ANALYSIS FOR THE HISTORY OF RELIGIONS

The SeNeReKo project

Corpora & Resources

The basis of the project are three different corpora of historical religious texts. This ensures that the developed methods are applicable to texts in different languages.

Ancient Egyptian

The database Thesaurus Linguae Aegyptiae, provided by the Berlin-Brandenburg Academy of Sciences and Humanities, represents an annotated selection of ancient Egyptian texts from different genres and periods with more than 1.100.000 text words. In our case it is used to analyse religious dynamics in the history of Ancient Egypt. An application of network analysis shall open up another perspective for an interpretation of interreligious contacts between Egypt and for instance the Near East. For this purpose the relations between actors and concepts are extracted from syntactic and semantic structures.

Stele 116, Louvre, Paris

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

Thai manuscript of a small part of the Pali canon

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

Pali Canon

The Pali Canon is a huge collection of Buddhist texts in the middle Indic language Pali. It has been composed around the Common Era in what is now Northern India and Sri Lanka and is nowadays available in a digital version, provided by the Vipassana Research Institute. The Canon contains a lot of narratives in which the Buddhist authors depict the Buddha and his followers in relation to competing religious groups, individuals and doctrines. Therefore the Pali Canon is a rich source for the analysis of interreligious dynamics in ancient South Asia. As the Canon is too huge for manual analysis, computational analysis is needed to get a clear picture of the social-semantic structures inherent in these narratives.

Mahābhārata

The ancient Indian Sanskrit epic Mahābhārata is a text of enormous length and richness. It grew over the course of several centuries (ca. 5th century BCE—5th century CE) and contains a wealth of information on the development of Hinduism. Network analysis allows to visualize connections between central concepts and actors and so offers new insights into the semantic structure of ancient Indian throught.

Kichaka and Bhimasena, Folio from a Dispersed Mahabharata Series, 1670. Brooklyn Museum, Gift of Cynthia Hazen Polsky

Research Question & Methods

Theoretical approaches in the study of religions increasingly stress the importance of interreligious contacts for the formation and development of religious traditions. Building on these developments, concepts from network theory are used in order to analyse the context and interdependence of religious actors and concepts. The project “social and semantic network analysis as a means to study religious contact” (SeNeReKo) applies methods from network analysis to the study of historical religious texts. In contrast to approaches in historical network research that focus on social actors alone, the project also takes semantic structures into account.

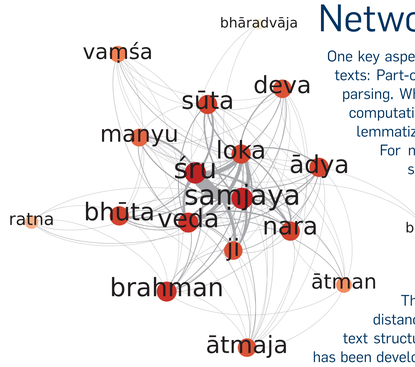

Figure 1: Central concepts of the first chapter of the Mahabharata, by community centrality.

Network Extraction

One key aspect of the project is the extraction of networks from texts. This step builds upon the linguistic analysis of the texts: Part-of-speech annotation, lemmatization, named entity identification, anaphora resolution and optionally syntax parsing. Where these annotations are not available in the original corpora, they are added by applying methods from computational linguistics. For the Pali canon, which is only sparsely annotated, part-of-speech tagging and lemmatization routines are developed.

For network extraction, one has to define criteria for the identification of both nodes and edges. **Nodes:** For semantic analysis, lemmata are used as nodes in the network, excluding stop words. For social analysis, social entities are identified. For this purpose, names and references that appear in the texts are assigned to unique social actors. This ensures that actors with multiple names and titles appear only once in the network, and that different actors with the same name are disambiguated.

Edges: Different network extraction methods are applied to create links between network entities. Currently, shifting-window (n-word) and fixed-window (sentences) approaches are used [1,2]. These methods both approximately consider words in a given distance to each other as related. To get a more precise view on actual text structures, a new method based on syntactic dependency structures has been developed [3]. Since it depends on syntactic information, this has to be added to the corpora, which requires further efforts for the target languages.

Network Analysis

The extracted networks can be analysed using standard network analytical procedures, especially centrality measures and community detection algorithms. This allows for analyses on different levels: On a text level, it can be used for text exploration and the identification of central concepts and semantic clusters (figure 1). On the corpus level, it can be used to study the semantic context of individual lemmas [4]. For this purpose, a subnetwork is extracted containing only nodes related to the lemma under study. Community detection algorithms can then be used to identify semantic context clusters and give hints about different contexts of use of a given lemma (figure 2). In a first publication, this methods has been used to study the relation of the two Ancient Egyptian concept “maat” (order, justice) and “heka” (often translated as magic, power) [5].

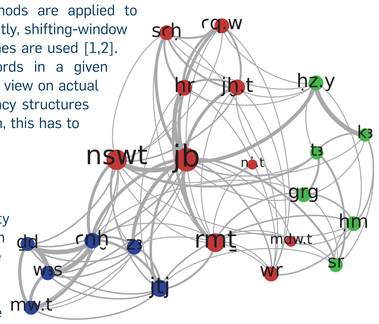

Figure 2: Semantic contexts of the Ancient Egyptian term “maat” (schematic representation)

References

[1] Paranyushkin, Dmitry. Identifying the Pathways for Meaning Circulation Using Text Network Analysis. Nodus Labs, 2011, http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/.
[2] Biemann, Christian et al. ‘Language-Independent Methods for Compiling Monolingual Lexical Data’. Computational Linguistics and Intelligent Text Processing. Ed. Alexander Gelbukh. Berlin, Heidelberg: Springer, 2004. 217–228. Lecture Notes in Computer Science 2945.
[3] Elwert, Frederik. ‘Network Analysis between Distant Reading and Close Reading’. Reading Historical Sources in the Digital Age. Selected Papers from the DHLU2013 Conference (Forthcoming).
[4] Dorow, Beate, and Dominic Widdows. ‘Discovering Corpus-Specific Word Senses’. Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. 79–82.
[5] Hofmann, Beate, and Frederik Elwert. ‘Heka Und Maat. Netzwerkanalyse Als Instrument Ägyptologischer Bedeutungsanalyse’. “Vom Leben umfangen.” Ägypten, das Alte Testament und das Gespräch der Religionen. Gedenkschrift für Manfred Görg. Ed. Georg Gafus and Stefan Wimmer. Wiesbaden: Harrassowitz, 2014. Ägypten Und Altes Testament 80 (Forthcoming).

Tools & Technology

Where possible, the project makes use of established standards and provides its software under Open Source licenses. Efforts are made to ensure the usefulness of project tools for other researchers and make them language independent.

Visit https://github.com/SeNeReKo for details.

Data Management & Processing

The corpora used are available from their respective providers in different, non-standard file formats. As a first step, they have been converted into a common TEI schema and are managed in an eXist XML database.

The linguistic annotation and network extraction tools are designed to be compatible with the CLARIN WebLicht infrastructure. Therefore, they are implemented as independent services that use the TCF format for data exchange. TCFlib, a python library for the easy creation of TCF compatible services and service chains, is made available on GitHub.

Network Extraction

Network extraction is implemented as a collection of WebLicht compatible services using TCFlib, available in the TCFnetworks package. The aim is to integrate them into the WebLicht environment and make them available for widespread use. Using a standard WebLicht annotation chain for part-of-speech tagging, lemmatization and optionally dependency parsing, TCF files suitable as input for the network extraction services can be generated.

Text Annotation

In order to train NLP tools for the automated annotation of the Pali corpus, training data need to be manually annotated. For this task, we use our own tagging software, which is specifically designed for efficiency and handles some peculiarities of the Pali language.

Dictionary Server

For NLP tasks a good dictionary is essential. For Pali, no suitable digital dictionary is available. Thus, available digitized dictionaries are integrated in a custom dictionary server. The server is built on a NodeJS/MongoDB combination and provides as REST API. This allows for algorithmic and manual processing of the free-text dictionary entries in order to extract machine readable information. This way we get dictionary data suitable for the required NLP tasks.

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

SeNeReKo

सेनेरेको

SPONSORED BY THE

Federal Ministry of Education and Research