

Gods, graves and graphs – social and semantic network analysis based on Ancient Egyptian and Indian corpora

Frederik Elwert, Simone Gerhards & Sven Sellmer

Abstract: In this paper, the authors show the application and use of automated text network analysis based on ancient corpora. The examples draw from Ancient Egyptian sources and the Indian Mahābhārata. Different text-based network generation algorithms like “Nubbi” or “Textplot” are presented in order to showcase alternative methodological approaches. Visualizations of the generated networks will help scholars to grasp complex social and semantic text structures and serve as a starting point for new research questions. All tools for applying the methods to ancient corpora are available as open source software.

1. Introduction

Network analysis as an analytical approach is quite popular in the digital humanities. Its methodological foundation, mathematical graph theory, is a generic way of modelling entities (nodes) and their relations (edges), and can be adopted to a variety of application domains. For humanists, inspiration often stems from the sociological branch of network research, Social Network Analysis (SNA). Modelling the connections between historical¹ as well as fictional² persons is an obvious application of SNA as a means to study historical eras and literary works in a distant reading fashion. But network analysis is not limited to studying personal networks. For philological research, approaches borrowed from computational and corpus linguistics can also help to highlight the connections between concepts in a text or a corpus.

The aim of this paper is to give an overview of some of the very different ways in which we can use network analysis in order to study ancient texts. Given the nature of such a methodological overview, we will be able to only briefly touch on the details of the different approaches. Examples from our own research on Ancient Egyptian and Indian corpora will be used to show the practical value of the different approaches and highlight their differences. Our paper will follow a path from social network analysis on the one end, which is probably better known, to semantic text network analysis on the other end, and various combinations in between.

1 Gramsch (2013).

2 Moretti (2011).

2. The Project

The research presented in this paper is based on our work in the project “semantic and social network analysis as a means to study religious contact” (SeNeReKo). SeNeReKo was a joint project between the Center for Religious Studies in Bochum and the Trier Center for Digital Humanities from 2012 till 2015.³ The team comprised one computer scientist, one Egyptologist, one Indologist, and one scholar of religion. The aim of the project was to develop and apply new computational methods for the study of religious history. One major issue when applying computational text analysis methodology to historical corpora is that we work with languages for which relatively few tools and linguistic resources are available, compared to modern languages. This poses a challenge when relying on techniques from computational linguistics for distant reading approaches. At the same time, this allows us to evaluate the requirements for a real-world application (given the issues mentioned) of different methods for the study of historical corpora.

3. The Sources

On the Indian side we worked with two large corpora of a little more than 1.5 million lexical units⁴ each:

1. a collection of Buddhist canonical texts composed in the Middle Indian language Pāli, the so-called Pāli Canon;
2. the Sanskrit epic *Mahābhārata*

Here, we will focus on the latter text. It is traditionally counted as an “epic”, but differs in several respects from the European representatives of this genre. It is not only much longer than the Homeric epics (ten times the length of the Iliad), but also considerably more diverse: apart from the main plot (a family feud) and lengthy battle descriptions, it also features numerous stories and tens of thousands of lines containing philosophical and ethical teachings.

Coming to the technical side, both corpora are available in digital form, but only as plain text files – which means that they have to be pre-processed (lemmatised etc.) before they can be used for analyses of the presented type. This is a difficult task, because in addition to the problems connected with flective languages in general, Sanskrit and Pāli pose further, very substantial ones, especially the phonetic changes called *sandhi* (see fn. 4). Luckily, for the *Mahābhārata* we were able to use a lemmatised text (partly with part of speech information) that was prepared by the pioneering computer program SanskritTagger of our kind colleague Oliver Hellwig.⁵ For the Egyptian material we used the text corpus from the database of the Thesaurus Linguae Aegyptiae from the Berlin-Brandenburgische Akademie der Wissenschaften. They kindly provided us their digitized and annotated database, which contains more than

³ The SeNeReKo project was funded by the German Federal Ministry of Education and Research, project number: 01UG1242A. The authors of this paper are responsible for its content.

⁴ Speaking of “lexical units”, we refer to the surface forms of words, which in Sanskrit differ according to the phonetic context. E.g., the nom. masc. sg. form devaḥ (“god”) may appear in the text as: devaḥ, devas, devas, devaś, devo, or deva. This phenomenon is known as “sandhi”.

⁵ Hellwig (2010).

one million lexical units. The texts go back as far as the third millennium BCE. The corpus consists of different genres like religious hymns, biographic inscriptions or literary and medical texts. The length of the texts varies from a few words to more than 8 000, so that one can speak of a heterogeneous text corpus. Originally, the texts were written on papyrus, stone, ostraca or, for instance, temple walls.

4. An Introduction to Network Analysis

Applying network analysis requires building a network model. A network model consists of *nodes* and *edges*. As with any kind of modelling, this requires an abstraction from the given data and reducing it to its core features.⁶ In the most common case of social networks, nodes are people, and edges are relations between them. In the case of online social network platforms like twitter, registered users are nodes. And if user A follows user B, that constitutes a relation, or edge. In this case, a certain network model is already given and technically enforced when using the platform's functionality. But of course, social networks as a social phenomenon pre-date the corporate adaptation of that term. Network ties are also constituted by letters between people,⁷ or by political alliances,⁸ or by mutual support⁹. But networks in the humanities don't have to consist of people. Everything that can be described as a set of relations between units of some sort is a network. In the remainder of the article, we will start with social networks, i.e. networks between people. Due to the project's background of religious studies, our personal networks will also include gods. From there, we will then move more and more away from pure social networks to other types of networks that one might study in the humanities, like conceptual and text networks. While the former allow us to study the use of single words, the latter represent complete texts as networks.

Whatever our nodes and relations are, the formal language of mathematical graph theory allows us to mathematically analyse a given network, and ask questions like: "Which node is the most central or important one in the network?" Or, "Which smaller groups can I identify in my network?" A network model is attractive for humanities research, since it highlights the relational aspects of our data: The importance of a person in relation to others, the constitution of the meaning of a word from its relation to other words, and the meaning of text as an emergent property of the relation of the words it is made of. Relational modes of thinking resonate with recent theoretical developments in different areas of the humanities. However, network models are not the only kind of model that captures relationality, and a network might not be the best kind of model for any given research question. Other methods like topic models or neural word embeddings can be used for similar purposes and might be more appropriate when large amounts of textual data are available.

6 What core features are is not absolute: It depends on the type of model (e.g., a directed or an undirected network), but also on the research question. In this sense, core features for a network model are much more constructed than found in the data.

7 Winterer (2012).

8 Gramsch (2013).

9 Düring (2015).

5. Social Network Analysis

To start with an example of social network analysis as applied to literary texts, we strived to obtain an overview of the important persons in the *Mahābhārata* and their mutual relations. Here, we had to solve two problems:

1. How to identify persons?
2. How to define their relations?

Regarding the first problem, it was impossible to take into account coreferences because this would have required an enormous amount of analysis by human experts. But even when looking only at names, the task is more tricky than one might expect, because important figures tend to appear not only under their primary names, but also under (in many cases multiple) epithets; in addition, sometimes the same epithet is used for several persons. These problems had to be, therefore, tackled with the help of manual checking.

As to the second question, it is clear that, for the present purpose, a relation must be recognizable by a computer, so we chose a very technical criterion: a relation (represented by an edge in our social network) exists where two persons appear in the same verse (i.e., in the case of the *Mahābhārata*, mostly in the space of two lines).

On this basis a social network can be quite easily constructed once the data are prepared. Figure 1 shows the result for the most frequent persons of the *Mahābhārata*.

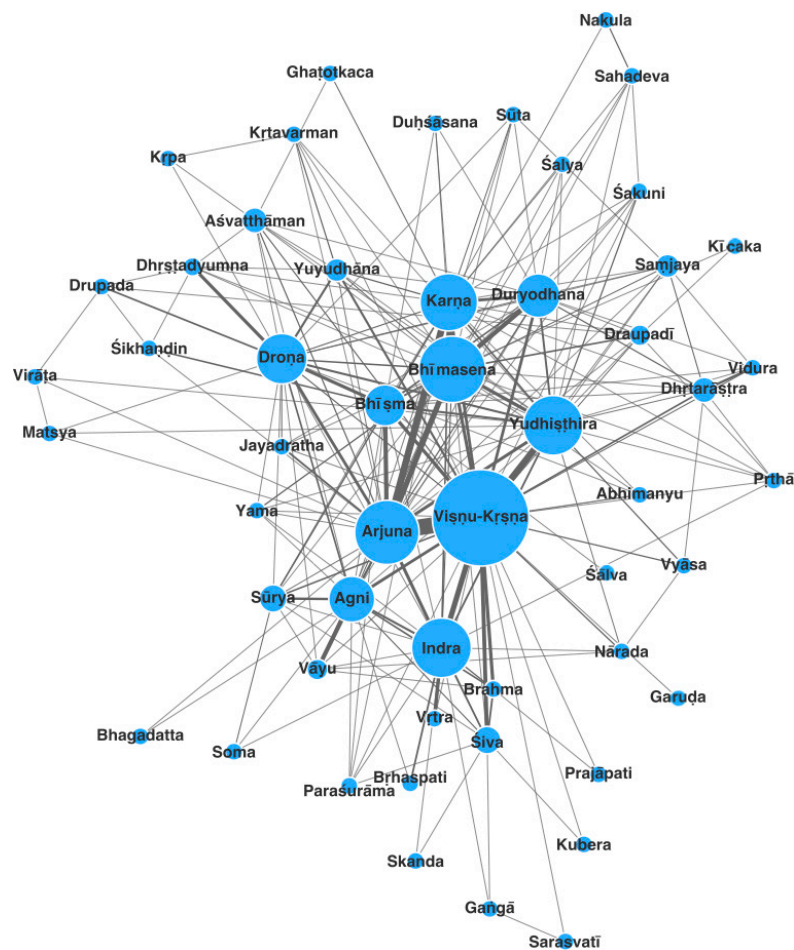


Figure 1: Social network of the main protagonists in the *Mahābhārata*.

This graph is mainly meant to demonstrate the ability of the computer to automatically detect so-called “communities”, i.e. clusters of nodes/persons that are more closely connected with each other than with the rest of the network. This becomes nicely visible in figure 2.

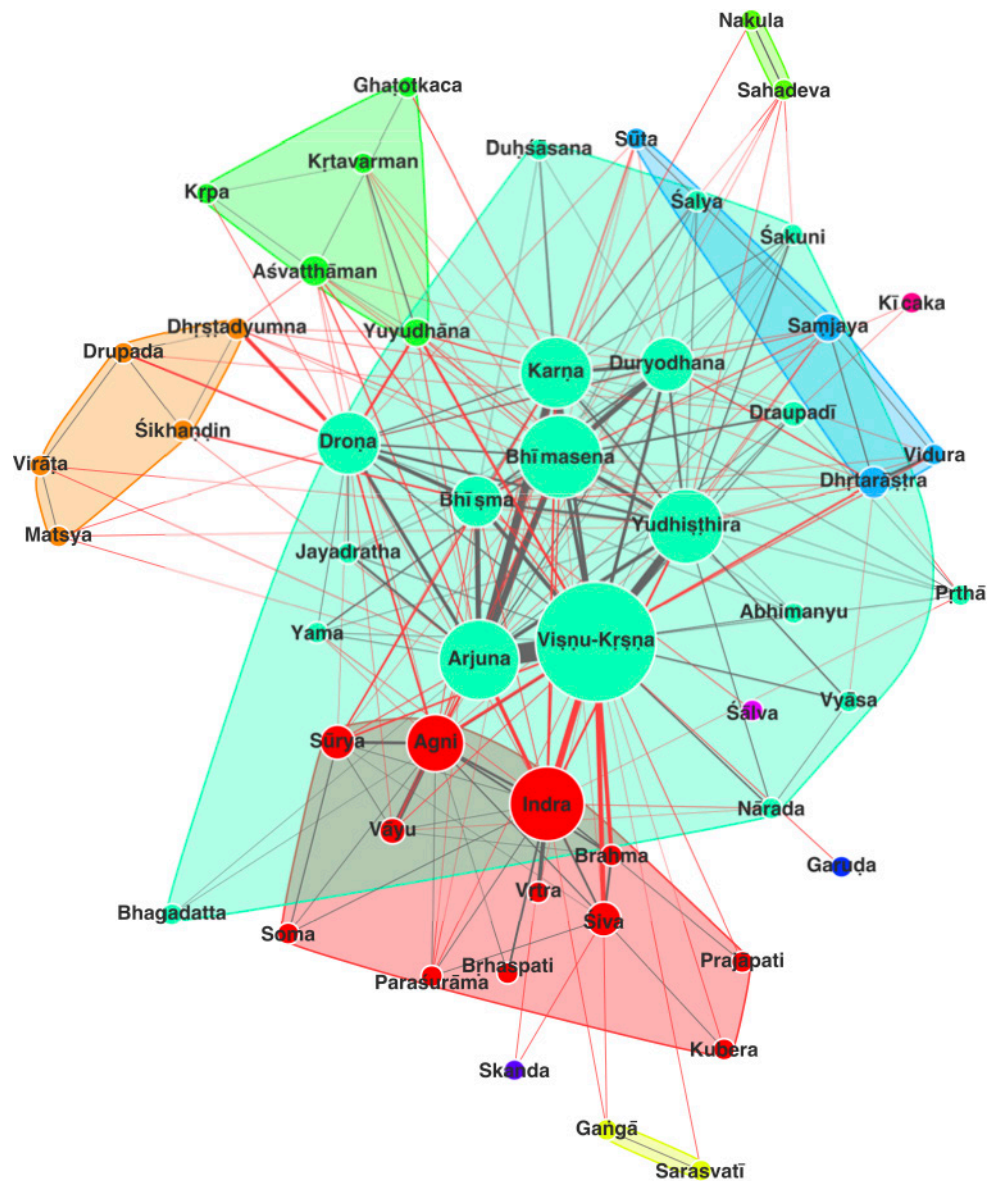


Figure 2: Communities in the *Mahābhārata*.

Here, the software (automatically!) groups the gods together (red), with the exception of Yama, the god of death, who belongs to the central green community, which comprises most of the principal actors. The small subgroups consist of heroes that are particularly closely linked among themselves. The singletons represent such persons that do not form part of the main plot but play an important role in one specific episode.

6. Semantic social networks

The previous example provides a good impression of the social structure of the *Mahābhārata*. But social structure is not everything that we want to study. Especially in the humanities, but of course also in the social sciences, we are interested in content, in semantics: What is this network actually about? Who are these people, and of what quality are their relations? Instead of the simple kind of network model introduced above, a semantic social network can be used that contains additional information about the connotations associated with its elements.

One way to add semantic information to networks is what we call the typed edges model:¹⁰ The network model can be enriched by adding information about the kind of relations that we observe. So we could distinguish between friendship, teacher-student-relations, and enmity. This usually requires to build a typology of relations that guides data collection. The researcher has to decide which kinds of relations are taken into account and define criteria for their identification. Depending on the research philosophy, this deductive approach can be an issue. In our case of ancient cultures, we did not want to impose a given typology (which might be derived from modern-day western concepts) on our material. Instead, we were interested in discovering the differences in relations that were expressed by the sources themselves. As a consequence, we followed a more inductive way of creating a typed model of the *Mahābhārata*'s social network.

The basis for this is an algorithm called “Nubbi”.¹¹ Nubbi utilizes topic modelling, a machine learning technique that allows to identify latent semantic structures in texts.¹² A topic in this sense is expressed by a list of thematically related words. Topic models use the word distribution across documents (or document sections) as information to automatically assign words to topics. A typical word list produced by a topic-modelling algorithm might consist of the words “ratha” (chariot), “śara” (arrow), “raṇa” (battle) and “han” (to kill). Labelling such topics is always an interpretative act. The labels used here, e.g. “fighting” for the given word list have been assigned manually.

Nubbi extends this model by distinguishing between topics that describe entities, or nodes, and topics that describe relations, or edges. For this purpose, it makes use of the text that surrounds the occurrence of an entity or a relation in the corpus. When a single entity is found in the text, the surrounding text is used for finding entity topics. When multiple entities are found, the text contributes to the topics describing the relations between those entities.¹³ We assume that the words near the mentioning of a relation in the text can be used to infer the quality of the relation, and of the entities that are part of it. This translates roughly to the idea of different types of actors and types of relations in a network, but it is more like a thematic connotation. Figure 3 illustrates this.

10 In network analysis terminology, this is often called a multiplex network, if multiple relations of different kinds are allowed between two nodes.

11 Chang et al. (2009).

12 Brett (2012).

13 The actual model treats relation documents as mixtures of entity and relation topics: In addition to the relation topics, also the entity topics of both individual entities contribute to the word distribution.

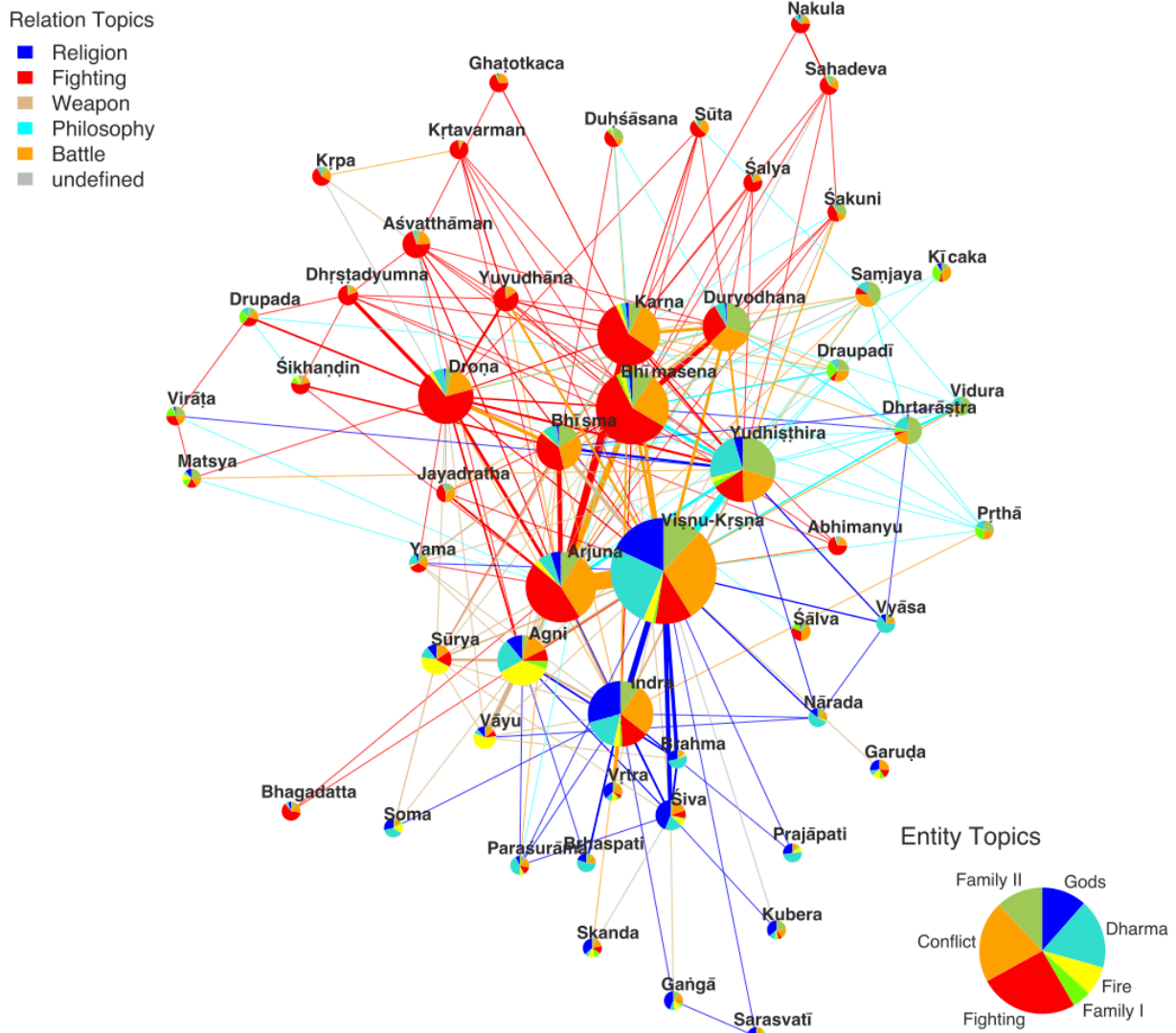


Figure 3: Semantically enriched social network of the *Mahābhārata*.

Here we have the same network as before, but with the information generated by Nubbi added through colouring the nodes and edges. Let us start with the persons that are now represented by pie diagrams. The size of the differently coloured sectors corresponds to the percentage with which the persons are associated with the single topics. The general distribution of the topics is visible on the pie in the bottom right corner. It is important to note that our domain expert has added the names of the topics. The program only gives a list of words that belong to (or constitute) a topic. Sometimes these “computer topics” are rather surprising or even unintelligible to the human interpreter, but in the present case they were more or less humanly understandable, so it was possible to attach a kind of “title” to each of them.

As just mentioned, Nubbi also extracts topics that are characteristic for the *relation* of two entities (therefore we call them “relation topics” – see left upper corner). They are represented by the colour of the edges. Because most edges are rather small, we chose to use only the colour of the predominant topic. So, for example, looking at the red connections – which symbolise fighting – one can easily identify the main enemies.

Now, one could continue to enumerate the pieces of information hidden in this graph – mostly intuitively convincing for an expert of the *Mahābhārata*, but sometimes also astonishing and intriguing – but it seems better to point to two general observations we made in the present context:

1. Persons may appear both as actors and in other, more figurative functions. E.g., when it is said that a hero “shines brightly like Sūrya” (the sun god) or that one warrior sends another “to Yama’s abode” (= to the god of death, i.e., kills him), then Sūrya and Yama have a very different role from the usual ones of agent or patient. Since these metaphorical uses do not constitute interpersonal relations in a classical sense, they might be undesirable. (Indeed we decided to remove the Yama verses of that type from the network, keeping only the relation between the killer and the killed.)
2. Certain kinds of relation topics are structurally underrepresented, especially the philosophical ones, because often lengthy philosophical instructions are prompted by a simple question, but follow only in the subsequent lines and are therefore not recognized as belonging to the relation questioner – answering person.

These, and cognate, phenomena call for future improvements, but even now we hope to have shown that by refining simple co-occurrence networks it is possible to model and visualize the semantic aspect of social relations (as reflected in textual content) to a useful and, according to our impression, sometimes astonishing degree.

7. Semantic Context Networks

In the previous case of the semantic social network, a network model is used only to capture the interpersonal relations. The semantic dimension is analysed using a different kind of model, here a topic model. Both kinds of models are integrated in a way that the social network informs the topic-modelling algorithm. Topic-modelling is also relational in the sense that it is based on co-occurrence of words that form the context of the elements of the network. The words that surround a reference to a person or a relation – or any other entity – in a text are not purely coincidental; we can assume they have some sort of semantic relationship to that entity. However, the internal state of the model is somewhat opaque, making it difficult to analyse these co-occurrences on a local level. For a specific entity, we get only a list of words associated with that entity as a result of the learning process, but we cannot inspect the nature and form of those associations. But if we regard these context words as related to the entity in question, then we can describe them as a network as well. We call this kind of network representation of semantics the “semantic nodes model”: Here, not only social actors, but also words are nodes of the network. The semantic information is thus contained in the network structure itself. This is suitable for examining the semantic context in more detail than what a topic model allows.

To build the context network, we used a co-occurrence based algorithm.¹⁴ These algorithms assume words to be related if they appear close to each other, possibly adding extra weight to the edges based on the proximity of the words. Then, directly neighboured words would have a heavier connection than words in greater distance. Such algorithms have been used to model whole texts.¹⁵ In our research, we found that these networks are often difficult to interpret on the global level once the underlying texts become too large or too diverse. This makes the method less suitable for studying medium to large corpora. However, they are useful for modelling the local neighbourhood of words. Thus, we apply co-occurrence networks mainly to study the semantic context of individual words or entities. This use resembles techniques used for word sense induction in computational linguistics, but with a different research question. Here, we

¹⁴ Paranyushkin (2011).

¹⁵ Lietz (2007).

In order to identify different contexts in which the lemma is used, a community detection algorithm was applied, assigning the nodes to different groups. This structure, therefore often called a community structure, describes how the network is compartmentalized into sub-networks (see figure 4). As a result, all words¹⁸ are marked by one of six colours, every colour standing for a different context. As for the size of the nodes, the bigger a node, the more central it is for the context of Horus. So the words “god”, “heaven”, “Re”, “name”, “Osiris”, “father”, “arm” and “Seth” have the highest degree. Here, the question may arise: “Where is Horus represented in the network?” The answer is, “nowhere”. Because he is by definition connected to all words, it is not necessary to show Horus explicitly in the network. Now we take a look at the details. First, we will have a look at the violet community, which consists most of the other gods and of information about their relationship to Horus. Osiris is the father of Horus, so Horus is his son. Isis, in turn, is the mother of Horus, and the sister of Nephthys. Furthermore, we find the divine siblings Geb and Nut, and Atum. The node “child” leads to a second community of relatives, the orange one of Horus’ children (Hapi, Duamutef, Kebekhsenuf and Amset). In the pyramid texts, one of their main purposes is to supply the descendent with food, as visible in the network. So they do not occur in the same context as the other gods. In the centre of the network are body parts that are semantically connected to Horus like “arm”, “mouth”, and “heart”. The blue community shows Horus’ affiliation to heaven and afterlife. Here appears the sky divinity Re, and the heavenly region, but also words that show the way to the afterlife like “door” or “ladder”. The last community to be mentioned here is the red one, which belongs to the god Seth. He is the uncle of Horus, but also his competitor. In Egyptian mythology, Seth is portrayed as the usurper who killed and mutilated his own brother Osiris. Horus sought revenge upon Seth, and the myths describe their conflicts. In the pyramid texts, Seth occurs as the (evil) counterpart who needs to be defeated.¹⁹ To sum up the main points, the network shows us the words most relevant for Horus in the pyramid texts. In 1916 Thomas George Allan wrote his egyptological Dissertation about “Horus in the Pyramid Texts”²⁰ where he analysed on a semantical basis the relations of Horus to other divinities, body parts or, for instance, the king. It is very interesting that he obtained the same results that our network shows at a glance: Horus’ strong connection to heaven, his function of helping the dead king going to “heaven”, the connection to his father Osiris and the connection to body parts like “arm”, “mouth” and “heart”, and the important role of Seth.

8. Text Networks (textplot)

The previous example used a co-occurrence algorithm to model the semantic context of an actor, the god Horus. But the same technique can be used to study the context of any given word. Following this path, the network model used no longer resembles to a social network, but a word network. The application of this kind of networks is not limited to the study of word contexts, but can also be used to model larger units like entire texts. Of course, a network model of a text is an abstraction, and one loses a lot of detail that is contained in the syntactic structure of the sentences. But following the idea of distant reading,²¹ a text network should highlight some information that is harder to grasp otherwise.

18 The English words are based on the TLA translation of the lemma entries.

19 Meurer (2002), 99 passim.

20 Allen (1916).

21 Moretti (2013).

As mentioned above, co-occurrence networks have indeed been used to this end.²² However, we found that the basic principles of co-occurrence do not scale well: The larger the text, the more blurred and hard to interpret the results become. While they work very well to capture local structures, they are less suitable to express the macro-structure of larger textual units. This problem is tackled by a different approach of building text networks called textplot.²³ It uses an intermediary abstraction to highlight global word relations instead of local phenomena. The basic idea is to add an edge between two words if they appear in the same passages throughout the text. To this end, it defines relations between words in terms of similarity of their distribution across the whole text.

Technically speaking, a kernel density estimate is used to model the distribution of a word as a smoothed curve. Then, the overlap between every pair of curves is calculated. For each word, a link to a given number (here: ten, following the original paper) of words with the most similar distribution curve is created. The resulting network is a suitable representation of the broad thematic structure of the text.

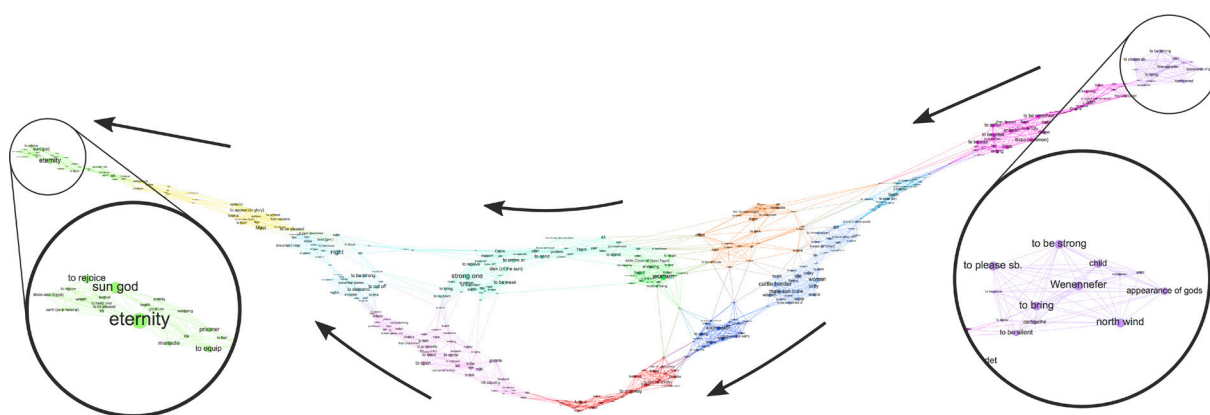


Figure 5: Text network of “the contendings of Horus and Seth”.

The network represents the text of “the contendings of Horus and Seth”.²⁴ It is a good example for the textplot method, because it is with 4820 lexical units the longest, coherent narrative in the TLA database and therefore provides enough data for testing the algorithm. The text deals with the battles between Horus and Seth for the succession to the throne of Osiris (see the remarks about the struggle between Horus and Seth above). The specific time of the contendings is a period during which the fighting has temporarily stopped and Seth and Horus have brought their case before the tribunal of the divine ennead. Throughout the story, Horus and Seth compete in several ways in order to find out who will be king.

This elongated network visualizes the process of the story very well (see figure 5):²⁵ The beginning of the story is a sort of a trial when Seth and Horus plead their cases, gods appear and the divine judges state their opinion. At the End of the story, the trial starts up again between Horus and Seth and finally, the situation is resolved when Horus is determined to be rightful king of Egypt. So beginning and end of the story are similarly structured.

But in the middle of the story something happens: The upper part represents the sections in which the gods are discussing who should be the next king and heir of the crown. The lower part of the network visualizes all the little sub-stories where the goddess Isis is involved and

²² Lietz (2007).

²³ McClure (2014).

²⁴ pChester Beatty I, recto (Dublin, Chester Beatty Library).

²⁵ Based on TLA data of the text.

where she tries to manipulate the action-packed battles between Horus and Seth. The textplot algorithm is able to recognize that the story consists of these two different “plots”, which are not that clearly structured in the actual story.

To sum up, constructing networks with textplot is a useful method for distant reading and for making visible the inner text structure.

9. Conclusion

Network models are an interesting approach to capture the relational nature of many of the phenomena that are of interest for humanities research. As a formal tool, they can be used to model anything that can be expressed as a set of nodes and edges. Picking the right network model and deciding what these nodes and edges are in a specific case has to be guided by the research question. In this paper, we presented a series of network modelling techniques that are suitable for studying ancient texts. Starting with social network analysis of literary characters, we showed how the semantic dimension of text, i.e., its content, could also be modelled as a network.

The findings presented here stem from research by the SeNeReKo project. Using ancient corpora as differing as the Indian *Mahābhārata* and Egyptian pyramid texts, we evaluated several network creation techniques. The tools we created to apply these methods to historical corpora are available as open source software.²⁶

As can be expected, the methods described in this paper and similar ones proved to be particularly helpful in the case of large texts because firstly, the manual gathering of, e.g., the data used in the networks based on the *Mahābhārata* would have required a virtually unmanageable amount of human work; but more importantly, graphical representations of such data (as shown above) enable the scholar to grasp complex social and semantic structures at a glance that could not – or only very imperfect – be noticed by traditional reading. Our research taught us that graphs of that kind, as a rule, do not provide final answers by themselves, but trigger new questions and are excellent starting points for further research.

²⁶ <https://github.com/SeNeReKo>.

10. References

- Allen (2015): Allen, James P., *The ancient Egyptian pyramid texts (Writings from the ancient world 38)*, Second edition, Atlanta, GA.
- Allen (1916): Allen, Thomas George, *Horus in the Pyramid Texts*, Dissertation, University of Chicago.
- Brett (2012): Brett, Megan R., “Topic Modeling: A Basic Introduction”, *Journal of Digital Humanities* 2/1, <http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>.
- Chang et al. (2009): Chang, Jonathan, Jordan Boyd-Graber und David M. Blei, “Connections between the lines: augmenting social networks with text”, in: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, New York, 169–178.
- Düring (2015): Düring, Marten, *Verdeckte soziale Netzwerke im Nationalsozialismus. Berliner Hilfsnetzwerke für verfolgte Juden*, Berlin.
- Gramsch (2013): Gramsch, Robert, *Das Reich als Netzwerk der Fürsten: politische Strukturen unter dem Doppelkönigtum Friedrichs II. und Heinrichs (VII.) 1225–1235*, Ostfildern.
- Hellwig (2010): Hellwig, Oliver, “Performance of a Lexical and POS Tagger for Sanskrit”, in: Girish Jha (ed.): *Sanskrit Computational Linguistics, Lecture Notes in Computer Science*, Berlin/Heidelberg, 162–172.
- Lietz (2007): Lietz, Haiko, “Mit neuen Methoden zu neuen Aussagen: Semantische Netzwerkanalyse am Beispiel der Europäischen Verfassung”, <http://www.haikolietz.de/docs/verfassung.pdf> (7 March 2017).
- McClure (2014): McClure, David, “(Mental) maps of texts”, <http://dclure.org/essays/mental-maps-of-texts/> (7 March 2017).
- Meurer (2002): Meurer, Georg, *Die Feinde des Königs in den Pyramidentexten (Orbis biblicus et orientalis 189)*, Freiburg, Schweiz.
- Moretti (2011): Moretti, Franco, “Network Theory, Plot Analysis”, *New Left Review* 68, 80–102.
- Moretti (2013): Moretti, Franco, *Distant Reading*, London.
- Paranyushkin (2011): Paranyushkin, Dmitry, “Identifying the Pathways for Meaning Circulation using Text Network Analysis”, <http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/> (7 March 2017).
- Winterer (2012): Winterer, Caroline, “Where is America in the Republic of Letters?”, *Modern Intellectual History* 9/3, 597–623.

Authors²⁷

Frederik Elwert, Dr.

Frederik Elwert, Dr.
Center for Religious Studies (CERES)
Ruhr University Bochum
Universitätsstr. 90a
D-44789 Bochum

Email: frederik.elwert@rub.de

Simone Gerhards M.A.

Institute for Ancient Studies
Egyptology, FB 07
Johannes Gutenberg University Mainz
D-55099 Mainz

Email: gerhards@uni-mainz.de

Sven Sellmer, Dr. habil.

Chair of Oriental Studies
Adam Mickiewicz University
ul. 28 czerwca 1956 nr 198
61-485 Poznań
POLAND

Email: sven@amu.edu.pl

²⁷ The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC-BY 4.0 International.