Frederik Elwert and Simone Gerhards

Tracing Concepts – Semantic Network Analysis as a Heuristic Device for Classification

Summary: Since the digital age is affecting more and more aspects of historical studies, the need to explore and evaluate new computational methods for the constantly growing amount of digital resources has become greater than ever. Due to the huge amount of digitized text, new methods may help to answer innovative research questions, challenge existing theoretical patterns and generate new approaches and perceptions. The research field emerging as a result of these reflections, the so-called digital humanities, might increasingly be incorporated into the established humanities disciplines. The authors present in this paper a heuristic method for the study of textual sources that allows the tracing of ancient Egyptian concepts and classification schemes by uncovering the relationships of words. By using automated network analysis in combination with a close reading of the sources, we can trace the outlines of the semantic field of a word; the ancient Egyptian expression for *day* serves as an example and will be studied in detail with this method.

1 Introduction: what is semantic network analysis and what is it good for?

One growing field in which digital approaches are applied is historical network analysis. Originating from the social sciences, network analysis is a method to analyze relations between units, mostly entities like people or groups, and to reconstruct and visualize underlying patterns. A related field is semantic network analysis, a method that allows relationships between words within one or more textual sources to be analyzed.

But the term *semantic network analysis* does not have the same meaning in every field of research.[1] In this paper, we define it as follows: Semantic network analysis understands text as a complex semantic system that consists of different lexical units, such as nouns, verbs, and proper names. These are in a certain relationship which is defined through closeness and distance to each other, not only through frequency. These relationships can be used to extract information about the concepts behind

1 For an overview of different definitions of semantic network analysis, see van Atteveldt 2008: 2–6; 13–61.

DOI 10.1515/9783110538779-012

them, but also to reconstruct hidden patterns of meaning. It captures information about the context of a term as a network that can be visualized in order to guide interpretation.[2]

The paper is based on research in the context of the SeNeReKo project. SeNeReKo is a joint research project of the Center for Religious Studies in Bochum (CERES) and the Trier Center for Digital Humanities (TCDH), Germany.[3] The project aims to utilize new methodological approaches for the study of digitized historical corpora. This approach was first applied in Egyptology as a method for the reconstruction of the ancient Egypt concepts in a paper on the terms *heka* and *maat* (Hofmann and Elwert 2014).

This paper describes a method for the study of textual sources that allows the tracing of ancient Egyptian concepts by investigating the relationships of words. Through the use of automated network analysis as a heuristic device in combination with a close reading of the sources, the ancient Egyptian expression for *day* will be studied in detail. Different implicit conceptualizations connected with the term are detected by looking at its context. We argue that clusters of textual units, detected through analyzing relations between words, can serve as indicators for the classification of typical usage patterns of key terms in historical corpora. The results highlight which conceptualizations are specific for a certain ancient Egyptian time period, and which of them have a broader scope.

2 Theory: distant reading and close reading

Traditionally, textual scholarship focuses on extensive reading of relevant source texts. This *close reading* allows us to grasp details in style or linguistic structure that can only become visible against the backdrop of scholarly expertise in the field under study. While these approaches allow a thorough reconstruction, they are also limited in their scope. In his widely discussed book, Franco Moretti argues for *distant reading*, and he provides two main arguments. First, he mentions sampling. Close reading depends on the selection of works that are included in the study, while others – often a majority – are left out. He criticizes this kind of canon building that he observes in literary studies: 'you invest so much in individual texts only if you think that very few of them really matter' (Moretti 2013: 48). With increasing digitization and digital pro-

2 This definition consists of ideas based on van Atteveldt 2008: 66; Lietz 2007: 1 f.; Rieger 1988: 44–46.

3 The acronym SeNeReKo stands for the German title *Semantisch-soziale Netzwerkanalyse als Instrument zur Erforschung von Religionskontakten*. The project was funded by the German Federal Ministry of Education and Research (BMBF) as an *eHumanities* project under the project number 01UG1242A until June 30th 2015. For further information see http://senereko.ceres.rub.de/. The authors are responsible for the content of this chapter.

duction, canonization becomes more and more a matter of choice, not of necessity. It is possible to study texts on a much larger scale now, also including works outside the classical canon. At the same time, this kind of large-scale work requires computational methods. Even if larger and larger collections are available in digital form, time is still a limiting factor for close reading. Only computer-assisted techniques, which utilize quantitative analysis, allow us to make use of the available source material. This is when *text* becomes *data*.

This might explain the vivid interest in distant reading methodology in the study of early modern and modern literature. But for ancient history, sampling depends on many other factors besides canonization. In fact, physical preservation might be the most limiting factor in the selection of source material. However, there is a second argument for distant reading, namely theory. In the context of social science research, Strauss and Corbin differentiate theory from description by its use of concepts, and its making statements about the relations of these concepts (Strauss and Corbin 1996: 13). Moretti makes a similar point in favour of distant reading. Only abstraction enables insight on a systematic level, although it comes at the price of losing detail (Moretti 2013: 49). For him, this lack of detail is not a disadvantage, quite the contrary: when it comes to theory, 'distance [...] *is a condition of knowledge*' (ibid., 48 [emphasis in original]).

But conceptualization is not only a scholarly activity; it is also part of social interaction and expressed through language. Corpus linguistics as an emerging field can be seen as a result of the empirical turn in linguistics. Instead of seeing category building as a theoretical exercise, it aims at describing acts of categorization as an empirical phenomenon:

> It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach. (Teubert 2005: 4).

In this regard, network analysis can serve as a bridge between linguistics and cultural studies.

Network analysis seeks to uncover the implicit structures hidden in textual data (Bubenhofer 2009: 100). These structures represent meaning. In this sense, meaning is expressed as well as constituted through repetition. Only the previous use of a word informs statements about its meaning, and the continued use of a word constitutes its meaning in the future (Rieger 1988: 44; Teubert 2005: 3).[4] Empirical, corpus-driven methods then allow us to trace the acts of conceptualization that are part of the texts we study. As a practice in distant reading, the single text itself disappears, while the

4 For a detailed elaboration of the semiotic theory that informs this approach, see Rieger 1989.

patterns of language use start to show, providing hints about the concepts that stand behind the text.

The use of network metaphors in the study of language goes back as far as Wittgenstein (2008 [1953]: 57). On a methodological level, the meaning – or better yet, the meanings – of a word can be seen as a network of interwoven dependencies to other words. When operationalizing this basic idea that meaning is constituted through language use, the context in which a word appears in the corpus is an important indicator. The interplay of a given word with the words surrounding it indicates the concepts behind the use of the word. In computational linguistics, a network model has been used as one approach to solve the problem of polysemy. When a word has multiple meanings, the different contexts of its use should become visible in the network structure of interrelated words (Biemann 2006; Dorow and Widdows 2003). Here, we apply a similar approach to trace ancient Egyptian concepts.

Still, a corpus-driven approach can only serve as a first step. The patterns that emerge from the computational analysis require interpretation. The interpretation cannot be based on the discovered patterns alone; it requires taking the text back into account (Bubenhofer 2009: 103). In a second step, the network can and should be used as an index to the text. It allows us to spot interesting structural phenomena and trace their origins in the corpus (Elwert 2016). Semantic network analysis as we understand it is merely a heuristic device that guides the work of the scholar, namely: the reconstruction of underlying phenomena and concepts.

3 Definition: what is a network and what does it look like

Before we start discussing the method, it is necessary to give a definition of the technical terms used in this paper. The idea of networks is almost ubiquitous in different fields of the humanities and social sciences. Generally, it simply expresses the idea of relationality. Formal network analysis as a specific form of computational analysis, however, is based on a very narrow idea of a network, rooted in mathematical graph theory.

A *network* (also called graph) consists of *nodes* (also called vertices) and *edges* (also called links). Since this model is of a completely abstract type, there is no limitation as to the nature of the nodes or edges. Anything that can be described as a set of entities (nodes) and their relations (edges) can be modelled as a network. In our case of semantic networks, words are the nodes, while the edges describe a shared context (for details about network extraction, see below). The network model used here has two additional properties: it is *undirected* (in contrast to a directed network), which means that all relations between words are treated as symmetrical. And it is *weighted*, so that the frequency of relations can be taken into account.

This abstract graph model allows us to mathematically uncover structural properties of the network and of individual nodes. Many network metrics and network analysis algorithms are available that can be used to describe network characteristics. Here, we mainly use two kinds of algorithms.

Centrality describes which nodes are more important than others (Freeman 1978: 215–217; Borgatti and Everett 2006: 466–468). There are different possibilities to define importance, so consequently there are several ways to compute centrality. The plainest of the centrality measures is the *degree*. Simply put: in an undirected network, the degree is the number of edges a node has (Freeman 1978: 218). The more nodes a node is directly connected to, the more central it is, following this definition of centrality. An extension of the degree measure is the weighted degree. It also takes into account the weight of edges, which in our case represent the frequency of a relation. A single edge with a weight of 2 contributes as much to the weighted degree as two edges with a weight of 1.

While centrality measures give information about single nodes, they reveal only little about the internal structure of the network as a whole. A more complex set of analytical algorithms deals with the problem of community detection. In network terms, communities are clusters of nodes that are relatively tightly connected, while having only a few connections to nodes outside the community (Blondel et al. 2008: 2). In many cases, there is no single solution to this problem, or it is computationally too complex to be solved in a reasonable time. As a consequence, several different strategies are applied to find an approximate solution in a reasonable time. Additionally, many of these algorithms have a random component, so even running the same algorithms twice does not necessarily return exactly the same communities. While this often seems unsatisfying, in a way it mirrors the nature of humanities research. In contrast to simple statistics, these questions are so complex that there is no single correct answer. Rather, the result of the computation is just a heuristic device that highlights one possible view of the structural properties of the object under study. At the same time, these results are not arbitrary, as they do represent one possible division of the network into groups.

4 Material: big data and small data in egyptology

Our starting point for the network analysis of ancient Egyptian terms is the Thesaurus Linguae Aegyptiae,[5] a database and publication platform by the Ancient Egyptian Dictionary Project at the Berlin-Brandenburg Academy of Sciences and Humanities.[6] The Thesaurus contains a digital corpus of Egyptian texts that supports computer-assisted research. The TLA kindly provide the SeNeReKo project with their digitized and linguistically annotated data, which contain more than one million lexical units. Every unit is annotated with part of speech information like *noun*, *preposition*, etc. For names, additional information about entity classes such as *king*, *god* or *toponym* are given. This very detailed annotation schema is extremely helpful for research methods like network analysis and allows results of a high quality to be obtained. German and partly English translations exist for the ancient Egyptian words and texts, but the networks are generated directly from the Egyptological transliteration of the textual sources. The digitized texts that are part of the TLA collection go back as far as the third millennium BC. The corpus consists of different text genres, such as religious hymns, biographical inscriptions, or literary and medical texts. The length of the texts varies from a few words to more than 8,000. In this regard, it is a very heterogeneous text corpus. Although it consists of more than a million lexical units, it cannot necessarily be classified as big in the context of computational analysis. One million sounds like a rather big amount, but it is a) still only a small excerpt of all ancient Egyptian texts existing and b) still very small compared to big data in the usual sense of data with a size beyond the ability of standard software tools (Snijders, Matzat and Reips 2012: 1). But by Egyptological standards it can be considered a reasonably big database with quite a large number of sources. Therefore, it is a very good starting point for our approach of network analysis of ancient Egyptian terms.

5 For further information see http://aaew.bbaw.de/tla/index.html. Hereafter referred to as TLA.

6 Contributions have been made by project groups at the Berlin-Brandenburg Academy of Sciences and Humanities in Berlin, the Saxonian Academy of Sciences and Humanities in Leipzig, the Demotic Text Database Project of the Academy of Sciences and Literature Mainz, the Book of the Dead Project of the North Rhine-Westphalian Academy of Sciences, Humanities and the Arts (Bonn), the project *Digital Heka* Leipzig, the Leuven Online Index of Ptolemaic and Roman Hieroglyphic Texts Project of the Katholieke Universiteit Leuven and the Edfu Project.

5 Use case: the egyptian day represented by the term *hrw*

The ancient Egyptian common term *hrw* (WB II, 498–500.24) for *day* has been chosen as example to show the method of semantic network analysis because it is frequently used during all epochs and in different contexts. To date, there is no scholarly publication which examines the semantics and contextual fields of *hrw* in detail. By offering a diachronic perspective on the context of *hrw*, network analysis provides a new approach to classifying and categorizing its use (cf. the categorization scheme: day – night; day – month – year; etc.). However, because only a limited sample of texts is available in the TLA, it is not possible to reconstruct all uses of the term. Therefore, the statements in the following examples correlate just to the references that are listed in the TLA database and cannot be generalized for all of ancient Egypt. The analyses presented here should be seen as a starting point and can be extended once larger digital corpora are available.

5.1 Method: strategies for network extraction

In order to extract a network representation of the term *hrw*, one has to specify rules that define *nodes* and *edges* with regard to words in a corpus. The basic unit for defining the nodes are words. However, not all words are relevant for semantic analysis, and different surface forms of a single lexeme tend to distort the analysis. So in order to identify the nodes of the network representation, words are first mapped to their lexical base form or lemma. This step is called lemmatization in computational linguistics (Manning and Schütze 1999: 132). The second step is to remove so-called stop words. Stop words are frequent terms that – of themselves – carry only little meaning, like prepositions, conjunctions, auxiliary verbs, and other frequently used, generic words (Ganesan 2014). In certain syntactic arrangements, these words can influence semantics of other words or of larger phrases. But since network analysis removes the words from their syntactic context, these functions of prepositions and other stop words are discarded. Additionally, in the TLA corpus, words are assigned discrete lemmata in case prepositions significantly alter the word sense. They can be removed without affecting the semantic information. Removing these words can help to reduce the amount of noise and usually improves the results of further analyses. For every network of *hrw,* the same stop word list was used in order to ensure stable conditions for analysis. In addition to a fixed stop word list, words can also be selected based on

part-of-speech information. In the case of the analyses presented below, we built the networks only from nouns, verbs and adjectives.[7]

The edges in the network representation approximately express semantic relations. Corpus linguistics and computational linguistics usually define word associations as co-occurrences.[8] If two words regularly appear in the same context, they will likely have a semantic relationship (Manning and Schütze 1999: 185; Bubenhofer 2009: 115). In order to build a network model of these semantic relations, it is necessary to specify a formal criterion for linking two words. A usual procedure is to define a window of a given size: all the words that appear around the target term within the specified distance will be linked to that term (Bullinaria and Levy 2007: 513). Here, we are interested not only in the links of our target term *hrw* itself, but also in the links between the words surrounding *hrw*. Thus, also the contexts of all related terms are taken into account when determining the edges in the network. The result is an ego-network of *hrw*, that is the network consisting of all nodes connected directly to *hrw* (its neighbourhood) and the edges between these nodes (the *alteri*).

A straightforward procedure would be to collect the edges between all words in the corpus, and then select only the ego network of *hrw*. However, this approach would also include those edges between *alteri* that do not occur in the context of *hrw*, but elsewhere in the corpus. The resulting network would then reflect, to a large extent, the general semantic structure of the corpus, but not necessarily that of *hrw*. Considering this, we take only those edges between *alteri* into account that can be identified in the context of *hrw*. Our first step for network extraction is to generate a sub-corpus that contains all sentences[9] in which the lexeme *hrw*[10] occurs. Experiments show that using sentences as the delimiting text unit for *alter-alter* relations produces better results than using larger units. Additionally, sentences are taken into account when determining word co-occurrences. Words are only linked by an edge if the window does not cross a sentence boundary. Our tests showed that sentences seem to approximately express semantic contexts, while word relations across sentences are often less interpretable.

Linguistic analysis of co-occurrences usually takes the frequency of co-occurrences into account. More frequent co-occurrences will express a stronger semantic relation. Instead of using raw frequency counts (Bullinaria and Levy 2007: 513), statistical measures such as likelihood ratios (Manning and Schütze 1999: 172) or pointwise mutual information (ibid, 178; Bullinaria and Levy 2007: 514) are used. Another

7 Nouns include the names of gods and goddesses. The definition of the grammatical terms is based on TLA annotation schema.

8 The term co-occurrence denotes two words appearing in a defined context, e.g. a sentence or a certain number of subsequent words, Bubenhofer 2009: 113.

9 The Egyptian writing systems under study usually do not mark word or sentence boundaries. Our use of *word* and *sentence* follows the editorial decisions expressed in the TLA encoding.

10 TLA Lemma No. 99060.

approach is to use word proximity to define edge weights. Not all words within the selected window size are treated as equally connected; instead, words appearing closer to the target term are linked with a higher weight. An innovative solution in the domain of text network analysis has been described by Paranyushkin (2011). This procedure uses a rolling window of five, i.e. selecting the first five words and linking them, then selecting words two to six and linking them, and so on. If a link already exists, its weight is incremented by one. Then, the procedure is repeated with a window of two, i.e. linking directly neighbouring words again. As a result, words closer to each other are linked through edges with a higher degree compared to words with a higher distance in the text. Our implementation of this approach differs from the original one in that we stop at sentence boundaries and not only at paragraph boundaries. Fig. 1 shows this principle using only the nouns of the sentence:

> Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. (from Herman Melville's *Moby Dick* [1851: 1]).

Our experiments comparing this approach to a statistical significance measure (log-likelihood ratio) show that this kind of weighted window produces results that are more interpretable in the context of network analysis.

Fig. 1: Word network of nouns in a single sentence

In the corpus, there are 678 different[11] references to *hrw*, with the earliest originating from the Old Kingdom[12] and the latest from the Greco-Roman Period[13]. Every unique sentence that contained the lemma *hrw* was collected into a subcorpus. In a second step, two additional subcorpora were extracted for a diachronic perspective. The networks of the Old Kingdom (2700–2200 BC) and New Kingdom (1550–1070 BC) have been selected because they contained enough references to allow for a more detailed analysis. The assignment of texts to periods follows the metadata contained in the TLA database.

11 Text passages that occur more than once were removed manually.
12 Tomb of Rawer in Giza, from the 5th dynasty BC, PM III, 269.
13 pFlorenz PSI inv. I 72 from the 1st century CE.

The co-occurrence networks were created from the corpus using the TCFnetworks package (Elwert 2014). The networks were then imported into Gephi (Bastian, Heymann, and Jacomy 2009) for analysis and visualization.

5.2 Network analysis: Detecting semantic clusters

Following the idea that the context of a word is indicative for its meaning, we assume that usage, and consequently meaning, of a word are not uniform. Meaning can change not only diachronically as semantic shift, but also synchronically, as different meanings of a word (or nuances of a word's meaning) can be highlighted, depending on the situation.

In computational linguistics, this assumption informs a wide range of research in the domain of word sense disambiguation. For text processing, it is often necessary to identify the meaning of an ambiguous word. The basic approach tries to assign an occurrence of a word to one of its senses as listed in a lexicon (Knopp, Völker, and Ponzetto 2013: 97). Since the corpus of the TLA already contains disambiguated lemma information for each word, this is not relevant to our research. Instead, we are interested in more subtle shifts of meaning that are not discriminated in a dictionary. To this end, techniques of word sense induction can be applied, since they aim at discovering senses of a word from a given corpus without using additional resources like a lexicon (ibid).

Semantic network analysis can be applied for the task of word sense induction (Dorow and Widdows 2003). A semantic network allows the identification of context clusters, and consequently of distinct meanings of a word. Using techniques of community detection (see above), the semantic network around a given term can be partitioned into sub-networks. Since the network edges represent co-occurrences of words in the context of the term under study, in this case *hrw*, these communities indicate relatively distinct contexts in which *hrw* is used.

5.3 Network representation: opportunities and challenges

In the following section we will discuss some exemplary networks in order to show the method with its opportunities as well as some challenges it poses. It should not be seen as an approach to generate new Egyptological knowledge, but as a heuristic instrument that is able to indicate the places in the original textual sources, which can be consulted to reconstruct ancient Egyptian concepts and/or classification schemes.

Firstly we will discuss the entire network made of all references for the lemma *hrw*. It is suitable in order to get a general overview of the words, their relational structure, and the context information, because it spans all time periods and text categories. But considered only by itself, it is not possible to draw conclusive statements from it, precisely because it spans different time periods and text categories. Thus, in a second step, the entire network will be compared with the networks of the Old and New Kingdom. This will provide a diachronic perspective and allow for more specific conclusions. In comparison to the full network, it is interesting to observe how communities change their size and content, and how new communities arise. This may indicate time dependent conceptualizations of *hrw*.

The entire network, constructed on the basis of all 678 references, consists of 1368 nodes and 8194 edges. In order to clarify the structure of the network, we selected only nodes directly connected to *hrw*, but removed the node *hrw* itself, because it has a relation to all other words in the network by definition. In this process, single nodes and small groups that have no connection to the majority of the nodes – the so-called giant component – are also removed. This yields a smaller graph consisting of *merely* 963 nodes and 5267 edges. However, this is still too big to perceive all relevant information at a glance (for an excerpt of the network see fig. 2). Therefore, we will discuss the community structure that was detected by the above-mentioned algorithm by Bondel et al. (2008). We can identify 15 conceptual communities within the network, the largest of which consists of 10.49% of all nodes and the smallest of 3.32%. None of the communities is clearly dominant. In the network every community is visualized using a different colour (see fig. 2). The central nodes of the network are: *grḫ* (TLA no. 167920), *jwi* (TLA no. 21930), *nṯr* (TLA no. 90260), *jb* (TLA no. 23290), *Rˁw* (TLA no. 400015), *ḥm* (TLA no. 104690), and *pri* (TLA no. 60920). The interpretation of the results will be carried out in relation to two different levels. On the one hand it is visually supported by the two-dimensional network and on the other hand by different centrality measurements and algorithmic calculations.

In the following tables, four exemplary communities and their first members are listed. For easy reference the unique TLA number is added in brackets. The translation corresponds to the one given in the TLA. In addition, you will find both the count of a certain word in the extracted text corpus and its degree in the particular subnetwork. Although count and degree correlate with each other, the degree can be seen as a relational measure of the importance of a word with regard to *hrw*.

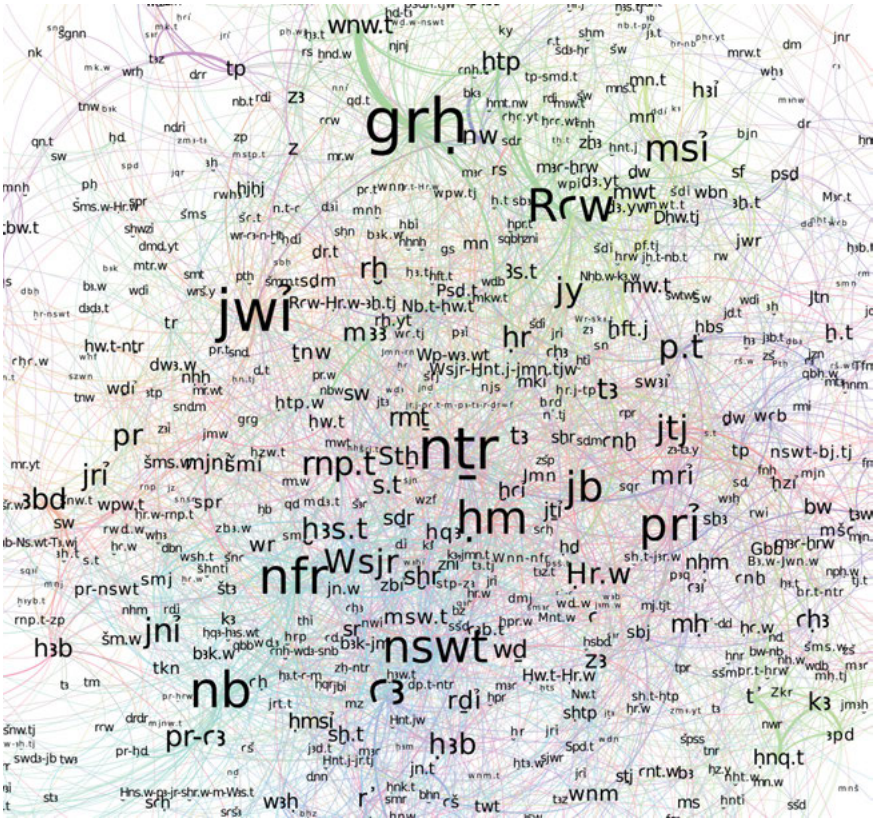

Fig. 2: Excerpt of the entire network constructed on the basis of all 678 references of *hrw*

Nodes	Count	Degree	Translation
jri (*hrw*) (29030)	35	22	to spend (the day)
pr (60220)	16	14	house
ṯ𓄿i (174260)	6	10	to take
rmṯ (94530)	12	10	human being, man
ḥmsi (105780)	8	10	to sit
p𓄿-hrw (58940)	3	9	this day
b𓄿k.w (53820)	4	8	work
spr (132830)	11	8	to reach
mjni (70060)	12	7	to arrive at, to reach

Table 1: Community structure around *jri* (*hrw*)

Tab. 1 shows an excerpt of the second largest community, which consists of 8.31% of the total nodes. The dominating word is *jri (hrw)* with a meaning of 'to spend the day with sth.' (WB I, 109.24 f.). Most of the following terms are verbs, which directly refer to a certain action, for instance, *to take*, *to sit* or *to reach*. Together with the terms *work* and *man/human being* this would suggest that this community shows the concept of *hrw* as moment or time of activity. Other words with a lower degree (< 8) are, for example, *sšr* with the meaning of 'corn' (TLA no. 851818; WB IV, 541.11–542.1) or *wrš*, another word translated with 'to spend the day with sth./to be awake' (TLA no. 48130; WB I, 335.10–18). Fig. 3 shows the community and its structure in detail. *jri (hrw)* is in a central position of the sub-network while all other community members are grouped around it. The words with the highest degree (see tab. 1) are the ones which hold the community together. In this way, this subgraph is rather coherent in its structure.

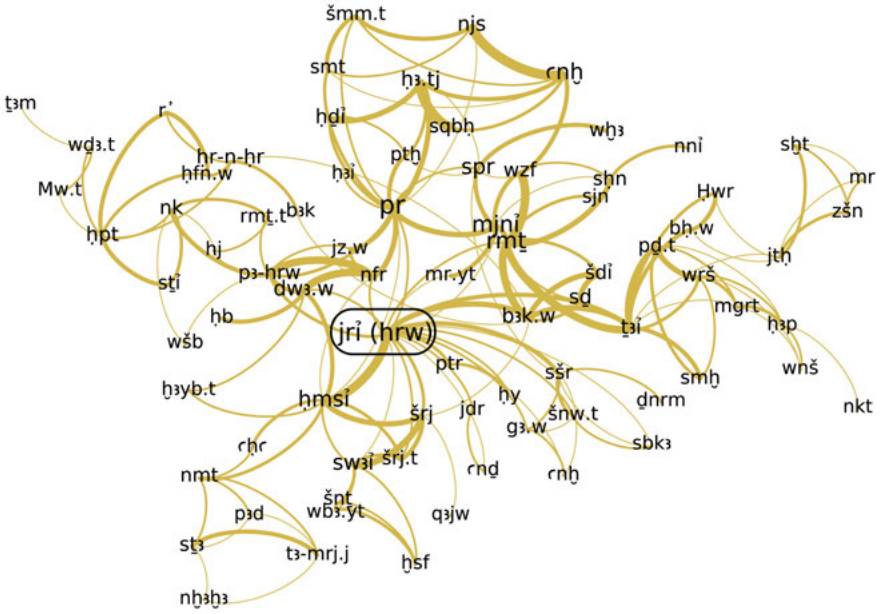

Fig. 3: Visualization of the community structure around *jri* (*hrw*)

Nodes	Count	Degree	Translation
nṯr (90260)	35	21	god
mri (72470)	16	16	to love
ʿḥ3 (39930)	15	16	battle
sʿḥ3 (129190)	6	11	to erect; to set up
nḥm (86430)	12	11	to take away, to rescue
mḥ (73290)	10	11	to fill
mšʿ (76300)	7	11	army, workforce
sbtj (132260)	3	10	wall, rampart
mnfȝ.t (70670)	2	10	infantry, soldiers

Table 2: Community structure around *nṯr*

The community listed in tab. 2 consists of 7.58% of all nodes. It is obvious that the words mentioned here (except for the first two) belong to a topic of fighting and battle. That means day in this community (and context) is characterized by war. At first sight, it seems that *god* and *love* do not fit into a war community, but by looking into the text sources (here it is inevitable to go back to close reading), one finds examples in which divine intervention is mentioned in a context of a battle.[14] It is interesting that some words with a lower degree (< 8) refer to a very different context. *psḏ* 'to shine' (TLA no. 62420; WB I, 556.14–558.3), *ȝḫ.t* 'horizon' (TLA no. 227; WB I, 17.12–23), *wbn* 'to shine/to rise' (TLA no. 45050; WB I, 292.9–294.3) and *Jtn* 'Aten' (TLA no. 850317; WB I, 145.4–7) belong to the context of *hrw* as part of the solar circle. But this special connection of the words just occurs in one sentence from the 'great hymn to the Aten'[15] dating back to the time of king Akhenaton. They are linked to the community by their common connection with terms and entities like *nṯr* (TLA no. 90260; WB II, 358.1–360.14), *Rʿw-Ḥr.w-ȝḫ.tj* 'Re-Herakhty' (TLA no. 70002) and *Mȝʿ.t* 'Maat' (TLA no. 66630; WB II, 20.10–13). Close reading reveals that this particular sub-group can be regarded as an artefact and should be excluded for interpretation. Otherwise, the community shows a thematically coherent conceptualization of day in the context of war/battle (see fig. 4).

14 For example pNew York MMA 35.9.21, col. 27.14–27.15.

15 The reference is from Tell el-Amarna, tomb 25 (Eje), entrance, west wall, PM IV, 228.

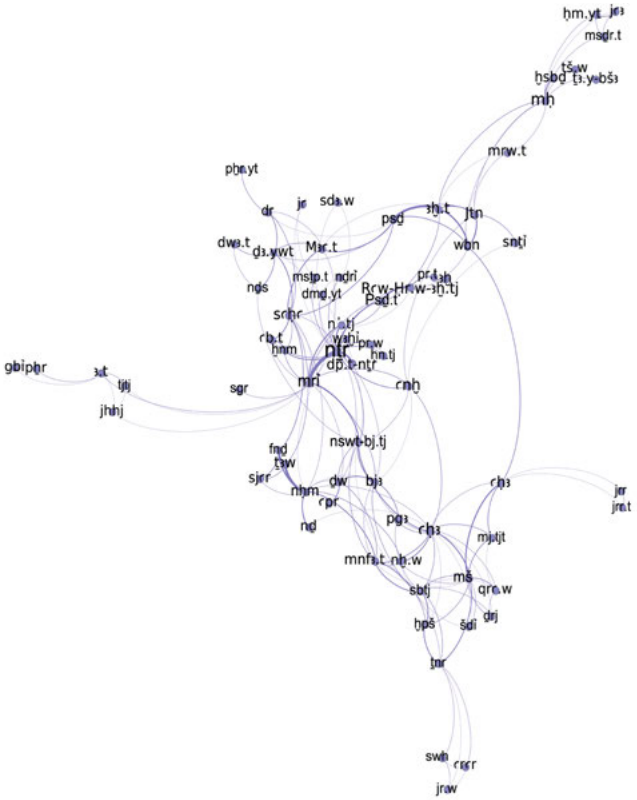

Fig. 4: Visualization of the community structure around *nṯr*

Nodes	Count	Degree	Translation
grḥ (167920)	60	34	night
Rᶜw (400015)	38	26	Re
wnw.t (46420)	21	21	hour
ḥtp (111230)	14	14	to be pleased, to set (of sun)
ḥr (107510)	19	11	face
ᶜḥᶜ.yt (40380)	4	9	midday
tp-ꜣbd (852477)	4	8	beginning of month
sḫm (142180)	2	7	might, power
psḏn.tjw (850569)	5	7	new-moon day

Table 3: Community structure around *grḥ*

The community consisting of 7.17% of all nodes is characterized by the word *night* (tab. 3). In addition, some group members like *hour*, *midday*, *beginning of the month*, etc. refer to time and date designations. The second predominating term is the divine entity Re. This gives a focus on the religious field of the connection between day and night. If you look at the network visualization (fig. 5), the direct link between *grḥ* and *Rʿw* is very thin. They are indirectly connected by the links with *ḥtp* 'to be pleased' (TLA no. 111230; WB III, 188.2–192.10), *ʿnḫ.t* 'the west (lit. place of life)' (TLA no. 38780; WB I, 205.16), and *mkw.t* 'protection' (TLA no. 76880; WB II, 160.22–161.4). As a result, the community can be interpreted as consisting of two aspects of *hrw*, its relation to other time specifications and its religious connotation. However, they are not strictly separable, but interrelated through shared references.

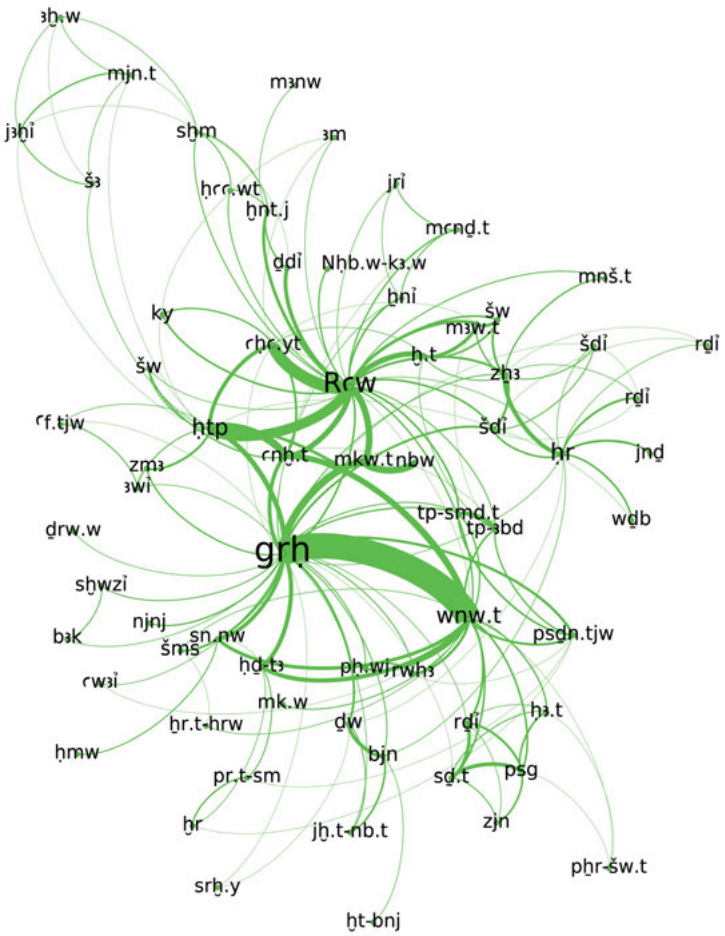

Fig. 5: Visualization of the community structure around *grḥ*

Nodes	Count	Degree	Translation
ꜣ (450158)	24	18	great, large
sn (136230)	10	17	brother
wꜣḥ (43010)	12	14	to put
wnm (46710)	8	12	to eat
ḥm.t (104730)	4	12	wife
ꜥš (40940)	6	11	fir tree
ꜥšꜣ (41011)	4	10	numerous
Bꜣtꜣ (850439)	4	10	Bata
sḫ.t (141480)	10	9	field
jn.t (26780)	4	9	valley

Table 4: Community structure around *ꜣ*

The next community (tab. 4) is one of the smallest and consists of only 6.44% of all nodes. Their relation can be ascribed to a specific context they have in common: they occur in sentences after expressions like *ḫr jr m-ḫt hrw.w qn.w – After many days (sth. happens)*. In this context *day* has no extensive meaning, because it occurs not in its function of a literal day, but as a stylistic expression to mark that a new text section begins (see Erman 1933: 286; Junge 1999: 276–277). One can find this expression, for example, in *The tale of the two brothers*[16] or *The tale of the doomed prince*[17]. This finding highlights the fact that especially smaller communities might not only stem from conceptual complexes, but also can express stylistic functions. Depending on the research question, they might or might not be relevant for further interpretation.

In comparison to the general network constructed on the basis of all references, the one generated from the Old Kingdom texts will be presented below. Here just 84 unique references for *hrw* are to be found. Most of them stem from pyramid text spells and funerary tomb inscriptions, only a very few are from letters or administrative documents. This suggests that in the network religious and funeral aspects will be more visible than in the aforementioned network from all references.

It consists of 305 nodes and 1225 edges. After applying the procedure of dropping unconnected nodes (see above), 209 nodes and 649 edges remain. Because it is considerably smaller than the one constructed from all references, the network is already legible as a whole (see fig. 6). The nodes with the highest degree are *grḥ*, *pri*, *nṯr*, *Wsjr*, *Ḥrw*, *tp* (TLA no. 170880) and *Rꜥw*. Comparison with the full network shows that *Wsjr*

16 pBM EA 10183 (pD'Orbiney), recto, 2,7.
17 pHarris 500, verso, 7,5.

and *Ḥrw* did not play such a central role there. This supports the assumption that the communities will be characterized mostly by religious concepts due to the pyramid text spells. In the following, we will discuss one new community and the one around *grḥ* that was already present.

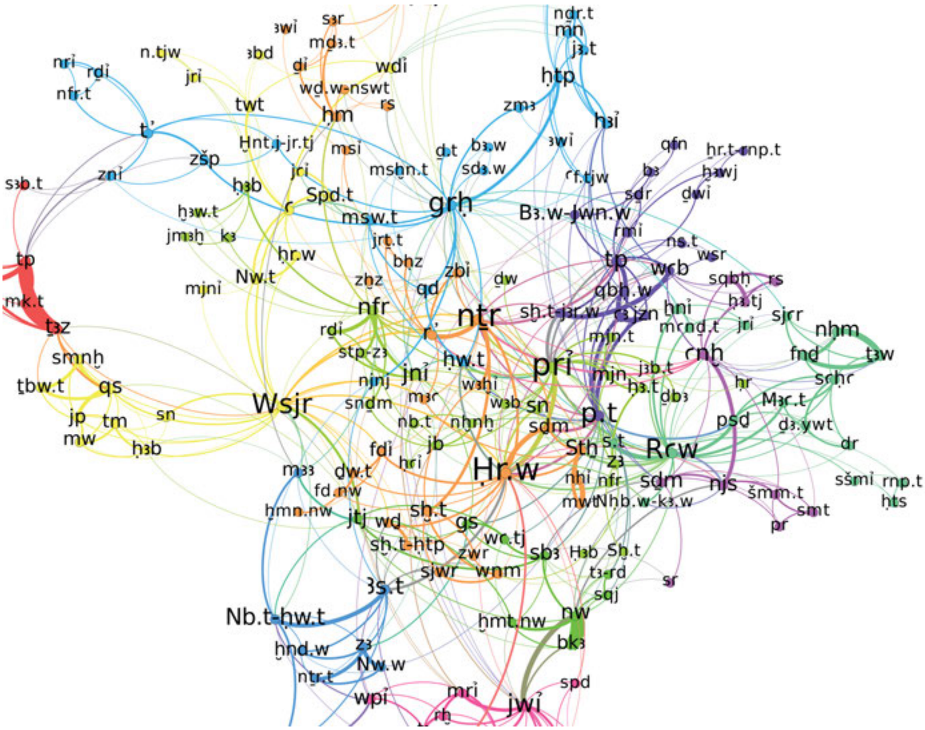

Fig. 6: Excerpt of the network constructed on the basis of 84 references of *hrw* from the Old Kingdom.

Node	Count	Degree	Translation
Wsjr (49460)	8	11	Osiris
ꜥ (34320)	4	9	arm
jp (24070)	2	8	to count, to asses
qs (162200)	3	8	bone
ḫ3b (850109)	2	6	to fish
tm (172000)	1	6	to be complete
mw (69000)	1	6	water
sn (136230)	2	5	brother
twt (170470)	2	5	statue, image
smnḫ (135360)	2	4	to make distinguished

Table 5: Old Kingdom community structure around *Wsjr*

This community around the divine entity *Wsjr* – the Egyptian god usually identified as the god of the underworld, afterlife and dead, but also as the god of regeneration and resurrection – was not visible in the full network. *Wsjr* connects all words and is the only link between two *sub-structures* (see fig. 7). This could explain why the community as a whole is hardly interpretable. The *sub-community* in the right upper corner (fig. 7) goes back to a spell of the pyramid texts (PT 536, §1297a–§1297c) in which Osiris and the words *jp*, *qs*, *ḫ3b*, *tm*, *mw*, and *sn* occur (cf. tab. 5). Here, a day is characterized by the concepts of *fishing*, *counting of bones*, and *making endure the soles (of the foot)*. Some of the links of the lower left corner of fig. 7 can be traced back to PT 505 §1089a–§190f in which the *day of the landing of Osiris* is mentioned. This refers to the day of Osiris' death. This community highlights the importance of going back to the text in order to understand the origin of certain connections. Additionally, it shows how references to gods tend to influence the community structure. Generally speaking, a day in this community is conceptualized by (the death/dying of) Osiris. This is a result of the time and genre of the pyramid texts.

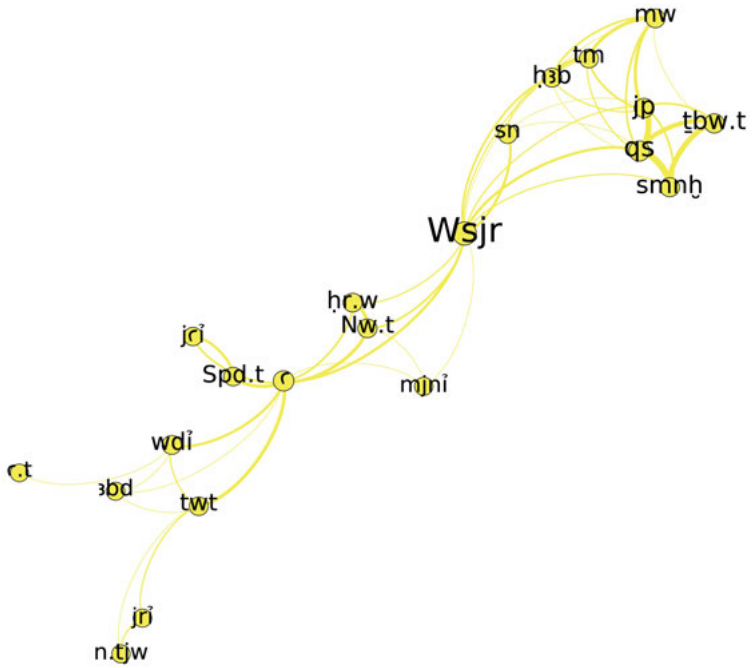

Fig. 7: Visualization of the Old Kingdom community structure around *Wsjr*

Nodes	Count	Degree	Translation
grḥ (167920)	11	16	night
ḥtp (111230)	4	11	to be pleased, to rest
t (168810)	2	7	bread
j3ḫi (20870)	1	5	to be flooded
mjn.t (68370)	1	5	stagnant water (?)
3wi (50)	1	4	to be extended
š3 (151110)	1	4	marsh, meadow
3ḫ.w (253)	1	4	(magical) power, mastery
h3i (97350)	2	4	to descend, to fall, to strike
zšp (144830)	1	4	to polish, to smooth

Table 6: Old Kingdom community structure around *grḥ*

This community (tab. 6.) is the largest one with 13.88% of all nodes. The term *night* is dominant, but in comparison to the community of *night* from the full network, only *ḥtp* is left. In the references from the Old Kingdom the term is not used with the meaning of *to set (of the sun)*, but rather represents concepts from the pyramid text spells, in which is said that a night should be merciful (PT 44, §34b) or a night is satisfying (PT 406, §708a–708c). Looking at the counts of the words in the community, one can see that most nodes are just mentioned once, making it difficult to interpret the remaining words from the community (see fig. 8). This example demonstrates clearly that a certain number of references is urgently needed to obtain an interpretable result.

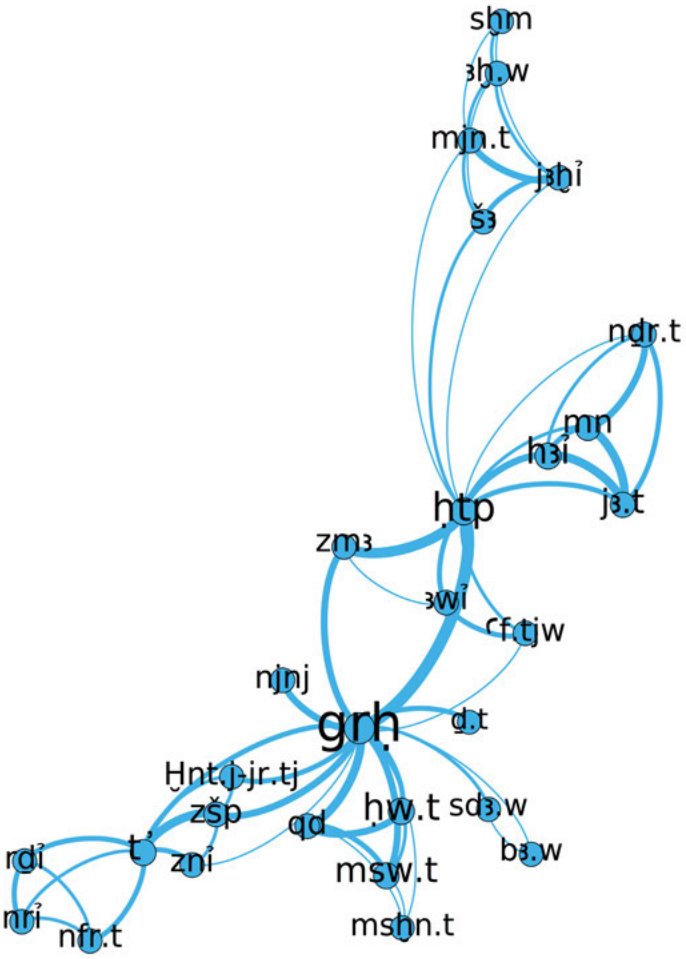

Fig. 8: Visualization of the Old Kingdom community structure around *grḥ*

In the following, the network constructed from the New Kingdom references (1550–1070 BC) will be presented. There are 349 references in total – this actually represents half of all unique references of *hrw* in the TLA database. The network consists of 833 nodes and 4328 edges; after removing unconnected nodes, 581 nodes and 2574 edges remain. If you look at the network of the New Kingdom (fig. 9 shows an excerpt), one can see in comparison to the one from the Old Kingdom that the influence of *Ḥrw* and *nṯr* fades into the background, while new context fields like the ones around *rnp.t* and *msi* arise. Only *grḥ* remains one of the most central nodes in all networks. In the following, we will discuss the community around *rnp.t* and again the one around *grḥ* in the New Kingdom in order to highlight significant differences.

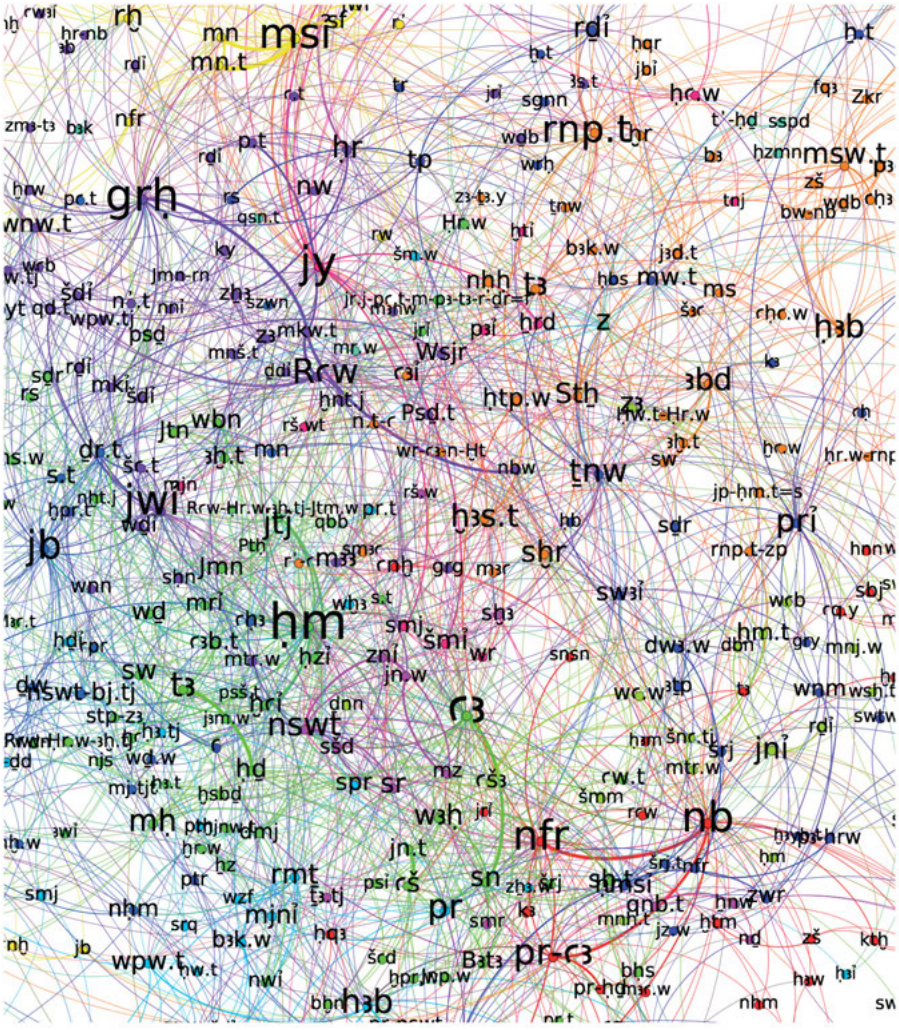

Fig. 9: Excerpt of the network constructed on the basis of 349 references of *hrw* from the Old Kingdom

Nodes	Count	Degree	Translation
rnp.t (94920)	20	24	year
ḫ3b (103300)	9	22	festival
t3 (400096)	10	16	earth
msw.t (75070)	10	15	birth
3bd (93)	14	13	month
ḥtp.w (111260)	7	9	peace, contentment
p3q.t (59330)	3	9	fine linen
r3 (92560)	2	8	mouth
ḫ3s.t (114300)	8	7	foreign land, desert
3.t (5)	3	7	moment, instant

Table 7: New Kingdom community structure around *rnp.t*

Tab. 7 shows the largest community, consisting of 13.79% of all nodes. It extends around the words *rnp.t* 'year' and *ḫ3b* 'festival' and characterizes *ḥrw* as a feast day, when earth is at peace and festivals take place. This community, too, exists in the network constructed on the basis of all references (where it consists of 7.68% of all nodes). Since the New Kingdom references account for a large share of all references, it is plausible that basic structures that can be found in this sub-network also show in the full network. However, it is absent in the Old Kingdom network.

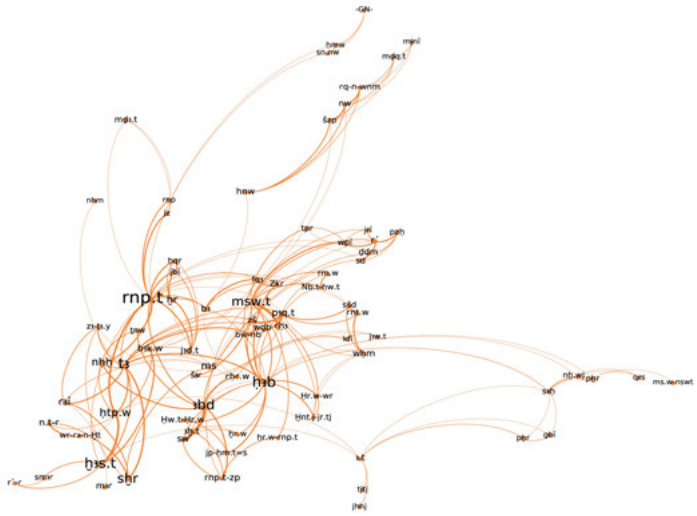

Fig. 10: Visualization of the New Kingdom community structure around *rnp.t*

Nodes	Count	Degree	Translation
grḥ (167920)	23	24	night
jwi (21930)	23	14	to come
Rᶜw (400015)	13	14	Re
ḥr (107510)	10	13	face
nw (80840)	6	7	moment
Psḏ.t (62500)	4	7	divine ennead
Ḏḥw.tj (185290)	2	7	Thot

Table 8: New Kingdom community structure around *grḥ*

Tab. 8 shows the community that is built around *night*. In comparison to the Old Kingdom network, the link to the god Re becomes visible and at the same time central. Therefore, it is present in the full network as well. Very interesting is that the relation to *ḥtp* does not exist any more. Nevertheless, a new node *jwi* 'to come' (TLA no. 21930, WB I, 44.6–45.6) appears (cf. fig. 11). Although *jwi* has the same count as *grḥ* its degree is much lower. Because *hrw* and *grḥ* are very close to each other in the texts, the weight between the nodes becomes higher.

By comparing all three sub-networks (fig. 5, fig. 8 and fig. 11) in which *grḥ* is the central node, the different composition of each becomes visible. This result indicates shifts in the conceptualization of the relation between day and night and can serve as a starting point for further analysis.

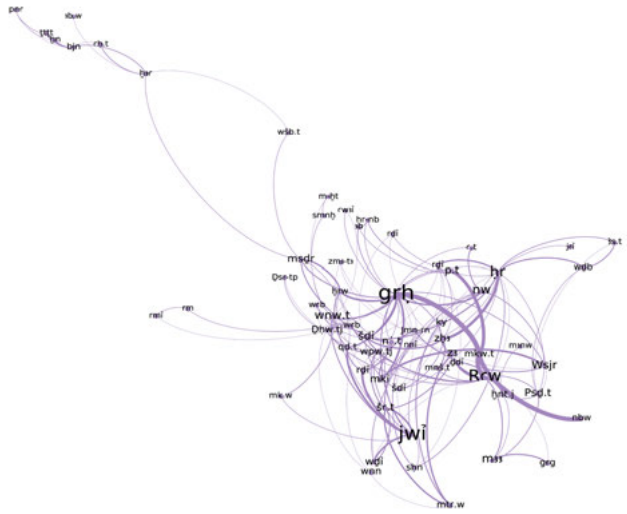

Fig. 11: Visualization of the New Kingdom community structure around *grḥ*

6 Conclusions: why network analysis for tracing concepts and categorizations?

The exemplary analyses of *hrw* in the TLA corpus show how semantic network analysis enables new approaches in the study of textual corpora. The representation of word co-occurrences as a network and computational analysis of this network allows the detection of semantic structures. In this way, it is possible to see at a glance indicators for different conceptualizations, like the example above showed: day connected to feasts, war and certain activities; day connected to entities like Osiris or Re; day as a stylistic expression to mark the beginning of a new text section, and the most frequent: day as part of a categorization pattern with other time specifications like night, midday, month or year. From an emic perspective, an ancient Egyptian day was closely connected to the solar cycle and the resulting change between light and dark (Hornung 1956: 26 f.). Thus, the expression *hrw* means, roughly speaking, either a *whole* day (corresponds approximately to 24 hours) or a time unit for a literal day (corresponds to the time when sun is shining), also referred to as daytime as the direct opposite of night (see Hornung 1956: 26 f.; Spalinger 1992: 147; Hoffmann 1996: 184; Westendorf 2003: 89–91). Through diachronic comparison, the relation to night becomes visible over the time. While this in itself is neither new nor surprising, semantic network analysis allows changes to be traced in the connotations of this relation through different time periods.

The networks of themselves should not be seen as a final result or as an answer to research questions, but as a tool to get from distant to close reading. Analysis does not stop on the level of the network. Interesting relationships between words, visible by distant reading, can be reconstructed through close reading. Because each edge in the graph can be traced back to specific source passages in the text, the network can be used as an alternative index. Instead of just listing words, it also allows word co-occurrences to be looked up. This kind of relationality can be seen as the main advantage. It shows words in their context, allowing relations to be studied, instead of isolated lexical items. In addition, the visualization of these relations allows relevant structures to be spotted more easily than traditional methods of corpus analysis do.

At the same time, there are also limitations. The method unfolds its full potential only if a sufficient number of references are available in the corpus. Only then do structural patterns begin to emerge that go beyond individual texts. Moreover, computational methods tend to make their results seem objective and unbiased. It requires a certain level of familiarity with the underlying methods to understand the implicit assumptions hidden in the algorithms and potential sources of bias. We thus see semantic network analysis as a heuristic device for tracing concepts. However, it should not be used without going back to the original source to understand the origin of a certain word connection in the corpus.

References

Atteveldt, Wouter van. 2008. *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*. Amsterdam: Vrije Universiteit. http://vanatteveldt.com/p/vanatteveldt_semanticnetworkanalysis.pdf. [Accessed 1/2017]

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In *International AAAI Conference on Weblogs and Social Media*, 361f. ICWSM. The AAAI Press. http://aaai.org/ocs/index.php/ICWSM/09/paper/view/154. [Accessed 1/2017]

Biemann, Chris. 2006. "Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems." In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, 73–80. http://dl.acm.org/citation.cfm?id=1654774. [Accessed 1/2017]

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10). doi:10.1088/1742-5468/2008/10/P10008.

Borgatti, Stephen P., and Martin G. Everett. 2006. "A Graph-Theoretic Perspective on Centrality." *Social Networks* 28 (4): 466–84. doi:10.1016/j.socnet.2005.11.005.

Bubenhofer, Noah. 2009. *Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin: de Gruyter.

Bullinaria, John A., and Joseph P. Levy. 2007. "Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study." *Behavior Research Methods* 39 (3): 510–26. doi:10.3758/BF03193020.

Dorow, Beate, and Dominic Widdows. 2003. "Discovering Corpus-Specific Word Senses." In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, 79–82. EACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1067737.1067753.

Elwert, Frederik. 2014. *TCFnetworks*. https://github.com/SeNeReKo/TCFnetworks (23.01.2017).

Elwert, Frederik. 2016. "Network Analysis between Distant Reading and Close Reading." In *DHLU 2013. Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age*, edited by Lars Wieneke, Catherine Jones, Marten Düring, Florentina Armaselu, and René Leboutte. CEUR Workshop Proceedings 1681. Luxembourg: University of Luxembourg. http://ceur-ws.org/Vol-1681/Elwert_Network_analysis.pdf. [Accessed 1/2017]

Erman, Adolf. [2]1933. *Neuägyptische Grammatik*. Leipzig: Wilhelm von Engelmann.

Freeman, Linton C. 1978. "Centrality in Social Networks Conceptual Clarification." *Social Networks* 1 (3): 215–39. doi:10.1016/0378-8733(78)90021-7.

Ganesan, Kavita. 2014. "All About Stop Words for Text Mining and Information Retrieval." *Text Mining, Analytics & More*. http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html. [Accessed 1/2017]

Hoffmann, Friedhelm. 1996. "Der Literarische Demotische Papyrus Wien D6920–22." *Studien zur Altägyptischen Kultur* 23: 167–200.

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, edited by Georg Gafus and Stefan Wimmer. (Ägypten und Altes Testament 80, 235–245). Wiesbaden: Harrassowitz.

Hornung, Erik. 1956. *Nacht und Finsternis im Weltbild der Alten Ägypter*. Tübingen: Universität Tübingen.

Junge, Friedrich. [2]1999. *Neuägyptisch. Einführung in die Grammatik*. Wiesbaden: Harrassowitz.

Knopp, Johannes, Johanna Völker, and Simone Paolo Ponzetto. 2013. "Topic Modeling for Word Sense Induction." In *Language Processing and Knowledge in the Web*, edited by Iryna Gurevych, Chris Biemann, and Torsten Zesch, 97–103. (Lecture Notes in Computer Science 8105). Berlin and Heidelberg: Springer. http://link.springer.com/chapter/10.1007/978-3-642-40722-2_10. [Accessed 1/2017]

Lietz, Haiko. 2007. "Mit neuen Methoden zu neuen Aussagen: Semantische Netzwerkanalyse am Beispiel der Europäischen Verfassung." http://www.haikolietz.de/docs/verfassung.pdf. [Accessed 1/2017]

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge (Mass.)/London: MIT press.

Melville, Herman. 1851. *Moby-Dick, Or, the Whale*. New York: Harper & Brothers; London: Richard Bentley.

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

Paranyushkin, Dmitry. 2011. *Identifying the Pathways for Meaning Circulation Using Text Network Analysis*. Berlin: Nodus Labs. http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis. [Accessed 1/2017]

PM = Porter, Bertha, Moss, Rosalind L. B. 1960–1981. *Topographical Bibliography of Ancient Egyptian Hieroglyphic Texts, Reliefs and Paintings*, 7 vols. Oxford: Oxford University Press.

Rieger, Burghard B. 1988. "Bedeutungsanalyse und Dispositionsstrukturen. Zum Problem einer empirischen Komponente der Situationssemantik." In *Angewandte Linguistik und Computer. Kongreßbeiträge zur 18. Jahrestagung der Gesellschaft für Angewandte Linguistik (GAL)*, 44–47. (Forum Angewandte Linguistik, Band 16). Tübingen: Narr.

Rieger, Burghard B. 1989. *Unscharfe Semantik: Die Empirische Analyse, Quantitative Beschreibung, Formale Repräsentation Und Prozedurale Modellierung Vager Wortbedeutungen in Texten,* edited by Bernd Spillner. Frankfurt am Main u. a.: Lang.

Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. "Big Data: Big Gaps of Knowledge in the Field of Internet Science." *International Journal of Internet Science* 7 (1): 1–5. http://iscience.deusto.es/wp-content/uploads/2012/08/ijis7_1_editorial.pdf. [Accessed 1/2017]

Spalinger, Anthony J. 1992. "Night into Day." *Zeitschrift für Ägyptische Sprache und Altertumskunde* 119 (2): 144–157.

Strauss, Anselm L., and Juliet M. Corbin. 1996. *Grounded Theory: Grundlagen qualitativer Sozialforschung*. Weinheim: Psychologie Verlags Union.

Teubert, Wolfgang. 2005. "My Version of Corpus Linguistics." *International Journal of Corpus Linguistics* 10 (1): 1–13. doi:10.1075/ijcl.10.1.01teu.

WB = Erman, Adolf, and Hermann Grapow. 1926–1963. *Wörterbuch der Ägyptischen Sprache*, 5 vols. Berlin: Akademie-Verlag.

Westendorf, Wolfhart. 2003. "'Der kleine oder vollendete Tag' im Ägyptischen Weltreich." *Göttinger Miszellen* 193: 87–94.

Wittgenstein, Ludwig. (1953) 2008. *Philosophische Untersuchungen*. Edited by Joachim Schulte. Frankfurt am Main: Suhrkamp.

Figures

Fig. 1–11: Frederik Elwert / Simone Gerhards