TNA ACTIVITY REPORT

PROJECT TITLE (**Formulaic language in Latin funerary epigraphy**)

Author: **Laura Soffiantini**

Current position: **FWO aspirant (PhD)**

Affiliation: **KU Leuven**

Host institution: **Center of Digital Humanities (CDH), Ghent University**

Mentor(s): **Julie M. Birkholz**

Period of stay: **August 22 – October 31**

Introduction: my research question

The aim of my project "Formulaic language in Latin funerary epigraphy" is to analyze formulaic sequences in Latin tomb inscriptions to **(OBJ1)** explore the **combinations** of the formulae in the texts. By relying upon an existing dataset that we developed in previous steps, I was especially interested in applying semantic network analysis to investigate how and to what extent formulae are related to each other. This permitted us to innovatively **(OBJ2)** reconstruct **global properties** of the formulaic language used in Latin funerary inscriptions that are still virtually unexplored.

Funerary texts are the most copious inscriptions that survived our time. Written on tombstones to commemorate the loss of a beloved one, it was estimated that between 170,000 and 190,000 Latin funerary texts came down to us. Widespread throughout the Roman territories, tomb inscriptions shared a set of highly standardized formulae. The same stock expressions (e.g., 'sit tibi terra levis', "the ground be light to you" or 'hic situs est', "he lies here") are repetitively used in a fixed form occurring in thousands of inscriptions. Despite the apparent uniformity of the texts, a closer look reveals some significant **variations** in the use of formulae with regional and chronological differences.

With my research, I aim at studying how formulae are combined in the texts by revealing (**OBJ1**) which formulae occur together and which are more strictly connected. In this way, by focusing on the relational aspects of the formulaic language, my scope is to (**OBJ2**) reconstruct the patterns in the use of formulaic sequences. More broadly, my project aims at contributing to our understanding of three research questions: (a) to what extent the meaning of the text depends on the combination of formulae used in it; (b) how the variations in a formula affect the other formulae in the text, (c) in which way formulae circulated in the past and were chosen in text composition.

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Methodological plan

To analyze the combinations of formulae and explore their relations, we generated a network of co-occurring formulae (**Step 1**). The nodes of the networks represent the formulae. Two nodes are connected by a link (edge) if the formulae co-occur in the same text. Then, by performing semantic network analysis, we quantified different metrics of the network both on a local and global level (**Step 2**). In this way, we were able to investigate the connections of the formulae (**OBJ1**) and, more broadly, to extract the global properties of the formulaic language in the texts (**OBJ2**).

Step 1: Construction of the network
The first step consisted of the generation of a network of formulae. To do this, we relied upon an existing dataset of 323,058 funerary formulae. Each formula in the dataset was linked to the corresponding text via IDs.

WP1: **Determine the window size**. To generate the links between the formulae, it was necessary to determine a window size, that is the range of words within which a connection was established. In our project, we used a window size of one text. In other words, if two formulae are present in the same text, an edge is created between them.
WP2: **Create the link list**. The link list of co-occurring formulae contains the pairs of formulae to import into the graph. To create it, we generated a tuple of each pair of formulae in the same text (that is, with the same text ID).
WP3: **Generate the graph**. The link list was imported to the Python library NetworkX to create the network graph.
WP4: **Weight the edges**. The value of the strength of each link was determined by the frequency of the co-occurrence. The weight of the link is equal to the number of times the formulae co-occur in the texts.

Step 2: Network analysis
The second objective consisted of the analysis of the generated network. Different metrics were extracted to investigate the relational properties of the formulae.

WP1: **Extract the descriptive statistics**. Once the network was generated some structural characteristics were examined including density, average shortest path length, and average clustering coefficient and the correlations between them.
WP2: **Perform centrality measures**. The total degree and betweenness centrality were extracted for each node. Moreover, the degree distribution of the network was analyzed and interpreted.
WP3: **Sparsify the network.** The structure of the network was simplified (sparsified) by removing the most central nodes. This step, which is commonly used in semantic network approaches, permitted us to delete the words that potentially could overshadow the network structure.
WP4: **Perform community detection.** The application of the Louvain and the Leiden modularity algorithm permitted us to detect the communities of formulae. The clusters contain formulae that tend to appear together in the same sentence.

My research visit at CDH

- **Development of the workflow** (August 25th-September 1st, CDH). I was introduced to researchers at the CDH and got familiarized with the department, their research focuses, and the ongoing project. Together, we discussed how to apply network analysis to the dataset of formulae in Latin. We discussed the pro and cons of the use of different tools for semantic network analysis (NetworkX, ConText, AutoMat) to find convenient solutions. Specific bibliography was provided to me, and we discussed together the possibility of applying machine learning techniques to the corpus of texts (see Future Work).
- **Crash course "Historical Social Network Analysis"** (September 29th, CDH). During the crash course held by prof. Birkholz, we discussed the basics of networks, with particular attention to (1) how to structure information for network analysis starting from historical data (more specifically, from periodicals) and (2) how to model data in relational terms. The course was based on the equivalent of the material of three lectures from the Bachelor course 'Historical Data Visualization'. The course provided me with a comprehensive overview of social network theories, network structures, and network measures. Moreover, I was introduced to the use of the program UCINET. During the demo session, we generated a network, implemented network measurements, and visualized it. Then, the discussion focused on the advantages and limitations of unimodal or **bimodal networks**, and **path analysis** for the network of formulae. Additional bibliography was provided at the end of the course.
- **Weekly meetings.** In joint meetings with my mentor and my current supervisor, we discussed how to structure the network of formulae and how to improve it. In particular, the discussion focused on (1) how to interpret the measurements of networks of concepts—that is, networks mapping the relations between words—and (2) how to visualize them. Particular attention was paid to discussing which kind of relationships we considered as a starting point for understanding the meaning of the network measurements. A comprehensive review of how classical network measurements are translated to networks of concepts is still lacking in the secondary literature and we discussed the possibilities of future investigations in this direction (see Future Work). Moreover, we discussed together my progress on my paper on methods of formulaic extraction and I received relevant feedback and suggestions.
- **Workshop "From Words to Networks: Semantic Network Analysis"** (October 24th, online). The introductory workshop was held by prof. Jana Diesner (School of Information Science, University of Illinois at Urbana-Champaign) and her team and focused on the application of network analysis to textual data with particular attention on the use of natural language text data as source information about the networks. The workshops introduced me to fundamental theories, concepts, and methods for constructing a network of words and provided me with a practical overview of basic techniques for structuring networks based on texts. Then, the discussion focused on the (1) impact of relation extraction methods on the reliability of the resulting network and the (2) use of cognitive/mental models to represent (shared) knowledge and information (which will be particularly relevant also for my current Ph.D. project). Then, I was provided with a demo session on the use of the program ConText. Finally, we discussed (3) how network measures are currently interpreted in semantic networks and the possible directions for further investigations to provide better

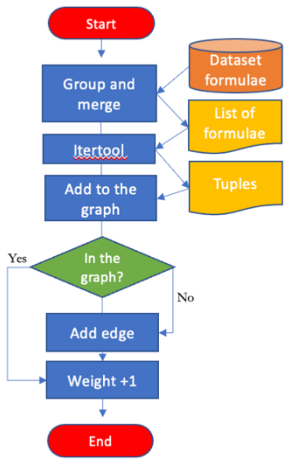
standardization and specific metrics for networks of words. Additional bibliography was provided at the end of the course.

- **Final presentation (October 25th, Ghent University Library)**. During the presentation hosted by the Library of Ghent University (45 min + discussion), I had the possibility to present the results of my research (dissemination). My talk focused on the methodologies we developed to study formulaic language in Latin inscriptions for both the automatic detection of formulae in the texts and the analysis of the combinations of formulae. The audience of the presentation included different members of the University including students of the course of Ancient History, Ph.D. students and professors and members of the department of DH, Ancient History, and Latin Linguistics.

Outcomes

The outcome of my research was a **network** of formulae (Table 1). The network is unimodal, undirected, and weighted. It contains 168 nodes connected by 3,866 ties. The sum of links, that is the sum of the values of the weighted edges, is 536,055. The **density** of the network is 0.275, measuring the proportion between the number of existing edges and the number of all the possible edges. The value is relatively low (< 1) showing that the network is sparsely connected and that the attested combinations of formulae are quite limited if compared with all the possible ones. Moreover, comparing the average shortest path length and the clustering coefficient, we discovered that the network is a **small-world network**. This implies that most nodes can be reached from every node by a small number of steps. In other words, even if two formulae never co-occur in the same text, they share at least one close formula which co-occurs with both of them.

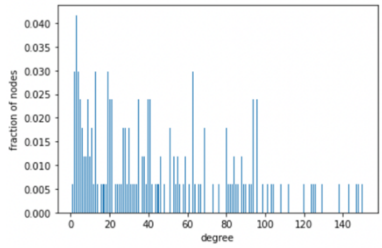
Semantic Network Analysis (SNA)



Undirected and **weighted** network. **Nodes**: 168; **Edges**: 3,866; **Density**: 0.275; **Sum of links**: 536,055; **Average degree**: 23; **Average clustering coefficient**: 0.826.

Formula	Degree	Normalized degree
vixit	150	0.8
annos	148	0.8
Dis Manibus	147	0.8
fecit	143	0.8
annis	138	0.8
bene merenti	129	0.7
Dis Manibus sacrum	126	0.7
sibi	125	0.7
fecerunt	124	0.7
annorum	120	0.7

The centrality degree is the number of edges touching a node The degree is normalized by dividing by the maximum possible degree in the graph.



The plot shows the degree distribution in the network. By counting how many nodes have each degree, the degree distribution of a degree (k) is the fraction of the nodes with degree k and the total number of nodes.

Table 1. Descriptive statistics of the semantic network.

The extraction of the **total degree** of the nodes permitted us to identify the most central formulae. As we can see in Table 1, the formula 'vixit' has the highest degree of 150. This means that the 'vixit' co-occurs at least one time with other 150 formulae out of a total of 168. Other central formulae are 'annos', 'Dis Manibus', 'fecit', 'annis' which co-occur in many different combinations. Then, we analyzed the **degree distribution** of the network which shows how the edges are distributed over the nodes (Table 1). The measurement revealed that most of the formulae have a low degree, whereas only a small number of nodes have a high degree (hubs). On average each node has 23 edges, but many formulae have a degree less than 20. This means that while a few formulae co-occur with almost all the other ones, most of the formulae are attested in a limited number of combinations. Furthermore, by comparing the degree of the nodes and their **clustering coefficient**, we also observed that the neighbors of the formulae with low degrees tend to be highly connected. In other words, the formulae with a small number of combinations tend to be used with formulae that are strictly connected to each other. Finally, after the sparsification of the network, we detected the communities of formulae. The same process was applied to the ego networks of single formulae. This permitted us to focus on a formula and to explore in detail its connections. In the example (Table 2), we can see the results of the analysis of the formula 'viva'. By **community detection**, we identified three clusters. The groups correspond to the formulae that most frequently appear together and with the formula 'viva'. The first community, for instance, corresponds to a **pattern** of formulae that is attested in the tomb inscription *CIL* 3.7408 (Table 2). The pattern focuses on the role of the woman in making the tomb set up ('fecit', 'posuit', "she did") for herself ('sibi') and for her 'well-deserving' relatives ('bene merenti') while she was alive ('viva').

'She did it (for herself)"

CIL III 7408 (Thracia, undated):
[D]is Manibus // L(ucio) Titovio L(uci) l(iberto) Diadu/meno Flavia Vera / coniugi **bene merenti** / et **sibi** et suis **viva fecit**.

'viva': frequency: 691; **degree centrality**: 93; **closeness centrality**: 0.68.

Ego-network

fecit	0.25	annos	0.18	annorum	0.01
Dis Manibus	0.02	vixit	0.01	de suo	0.009
sibi	0.02	Dis Manibus sacrum	0.01	posuit	0.009
bene merenti	0.01	hic situs est	0.006	sit tibi terra levis	0.005
annis	0.01	hic sita est	0.006	ossa	0.003
fecerunt	0.01	militavit	0.005	faciendum curaverunt	0.002
dulcissimo	0.008	pia	0.004	defuncto	0.001

By community detection, 3 clusters were identified. The table shows the nodes with the highest betweenness centrality in each community.

Bibliography

Edmondson, J. "Family relations in Roman Lusitania: Social change in a Roman province?" In M. George (eds.), *The Roman Family in the Empire: Rome, Italy, and Beyond*, 183–229. Oxford. 2015.
Mueller, I. "Single Women in the Roman Funerary Inscriptions." *ZPE* 175 (2010), 295–303.
Saller, R., Shaw, B. "Tombstones and Roman Family Relations in the Principate: Civilians, Soldiers and Slaves." *JRS* 74 (1984), 124–56.

Table 2. Results of the community detection of the ego-network of the formula 'viva'.

Future work

In conclusion, the application of semantic network analysis to the Latin texts showed how the use of computational methods can expand the analysis of ancient inscriptions by revealing unexplored features in the texts. One of the biggest challenges of our work was the lack of standard metrics for the analysis of networks of words. From this point of view, a review of the measurement currently in use will contribute to defining their standardization. Furthermore, an ongoing Trismegistos project (KU Leuven) is investigating the use of abbreviations in Latin inscriptions (e.g., 'Dis Manibus Caius' written 'DMC'). Including the abbreviations in the network of formulae will permit us to (b) explore to what extent formulaic words are abbreviated, which formulae are abbreviated in the same cluster, are more broadly which is the relationship between formulaic language and abbreviations in Latin inscriptions.