**TRILCKE — Network Analysis**

<2 figures (captions embedded in figures)>

**Digital Network Analysis of Dramatic Texts**

Trilcke, P., Fischer, F. and Kampkaspar, D.

The project ‘Digital Network Analysis of Dramatic Texts’ follows in the tradition of structural analysis approaches in the literary studies (Titzmann, 1977) and is aimed at advancing them in recourse to established methods like Social Network Analysis (Wasserman and Faust, 1998). The project also enhances these elder approaches through automated data acquisition and analysis in order to handle much larger corpora and to be able to generate comprehensive relational data to analyse structural transformations in the history of literature.

As theoretical foundation we used a network-analytic conceptualisation of dramatic interaction (first ideas in Moretti, 2011; critical reading and theoretical reconceptualisation in Trilcke, 2013; also a detailed research summary). In continuation of concepts of dramatic configuration (Marcus, 1973; Pfister, 1977; Pohlheim, 1997), we resorted to a basic operationalisation according to which an ‘interaction’ takes place if two characters are listed as speakers within a given segment of a text (usually a ‘scene’).

For our first automation purposes, we considered ‘interaction’ as ‘scenic co-presence of two speakers’. Based on this concept of relations, network data is extracted automatically—both the global networks of ‘interactions’ of the plays (density, average degree, connectedness, etc.) and data that characterises individual actors (degree and various other centrality indices). The workflow we created also allows the collection of data on a meso-level (e.g., identification of clusters) and includes visualisations of the extracted network data, which in turn contributes to the analysis of the structural transformation in the history of literature.

# Corpora of Dramatic Texts

For the automated analysis of theatre plays, a reliable and sufficiently large (German-language) corpus was needed. The following corpora were reviewed:

• Deutsches Textarchiv / German Text Archive (DTA): 54 dramatic texts.

• Wikisource: 50 dramatic texts.

• Projekt Gutenberg-DE: 641 dramatic texts.

• TextGrid Repository: 690 dramatic texts.

The DTA corpus has the best quality of TEI markup, but so far only incorporates few texts. The latter also applies to the German-language branch of Wikisource. The Gutenberg-DE archive is problematic due to the poor markup it provides (just some basic XHTML). Thus, only the TextGrid Repository (containing basic TEI markup) was really applicable.

First, we extracted all dramatic texts from the repository, 690 texts altogether that are marked as ‘drama’ in the ‘genre’-field of the metadata. Most of them are German plays from about 1500 to 1930, plus a dozen translations of Greek tragedies and some Shakespeare plays.

# Acquisition of Network Data

As an intermediate step, we created a list of relations between all the persons appearing in a play for each of the 690 TEI files and wrote them into a CSV file, one of the standard formats for the storage of network data. To extract the speaker data, usually two separate steps are required: The recognition of the individual segments of a play and the recognition of individual speakers.

To facilitate the following work, the script first splits the files. For each level recognised in the document tree, a subdirectory is created containing all the individual parts of the TEI files, along with the respective index files. Different kinds of outputs are generated this way: for one, a detailed register of all <speaker> tags, but also all the <rs> and <person> tags. To obtain unambiguous reference targets, ID numbers are assigned (also facilitating later interventions, if erroneously assigned names must be manually corrected). Furthermore, co-occurrence lists are created. In the bottom directory, the occurrence of all speaker pairs in all files are counted. In the upper directories, the values of all subdirectories are added.

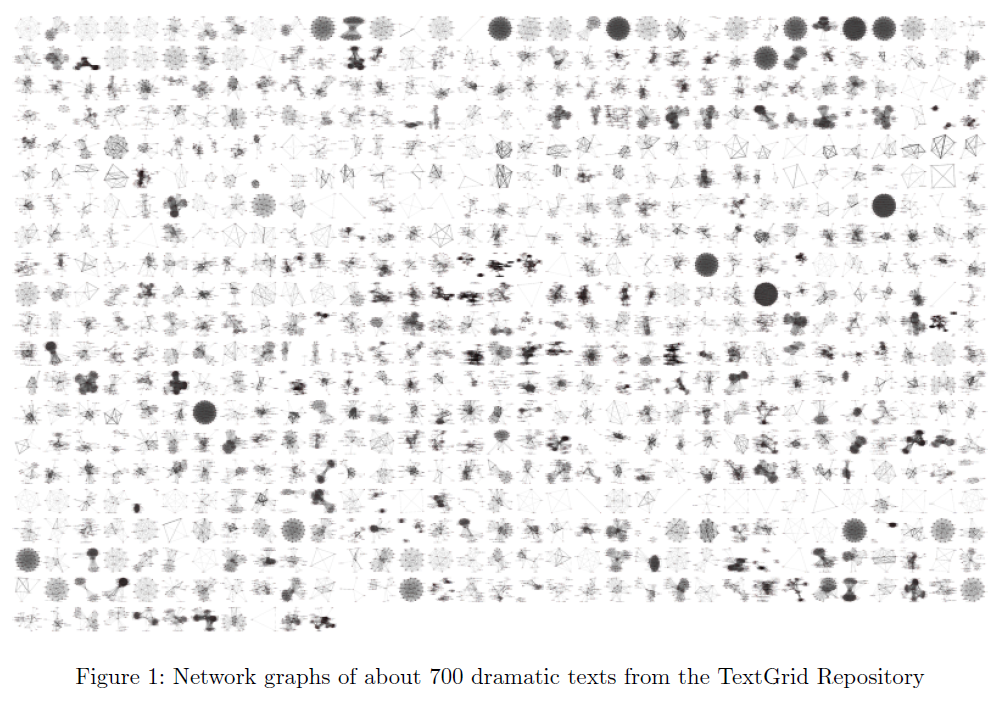
In addition to the recognition of the structure, the correct assignment of names is one of the bigger challenges. Ideally, all <speaker> tags would contain a @who attribute to provide a unique ID for each character. If this is not the case (or if <rs> or <person> tags were used instead), the script has to analyse the textual content of the tag. Possible misspellings (due to the transcription or the original) and grammatical changes have to be considered. For instance, in Lessing, *Nathan the Wise* V/1, we encounter different sorts of Mamalukes, be it ‘A Mamaluke’, ‘The Mamaluke’, ‘A second Mamaluke’, or just ‘Second Mamaluke’. In this case, some linguistic knowledge about adjectives is enough to rule out mistakes when assembling the lists of relations. But there are more complicated cases when automatising this process, e.g., if multiple characters speak up at once (‘All’).

In addition to trying to clarify these cases automatically, there is still the possibility of manual intervention in cases of doubt. Here, the generated index files with unique IDs will be of help. For an upcoming version of the script, it is intended to provide a simple GUI to allow easy editing in such cases.

# Data Analysis and Visualisation

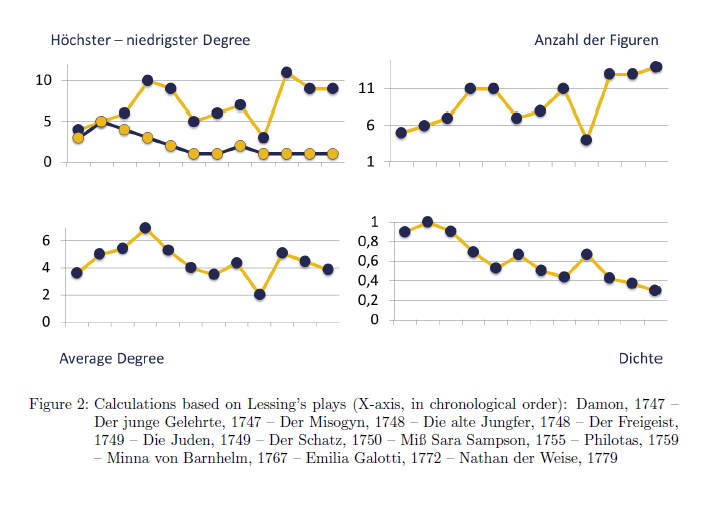
The data analysis was done by using the igraph package via Python (3.4.x). It was used for both the visualisation of the graphs and the calculation of the network data.

For a first visualization, we fed the graph data into a spring-embedding method (Fruchterman-Reingold), which tries to arrange related nodes closer together (clustering). A first impression of the entire corpus is provided in Figure 1. It comprises nearly 700 plays from 2,500 years of theatre history, starting chronologically at the top left with the ancient Greeks ranging to the bottom right and the second quarter of the 20th century.



The visualised graphs (a ‘distant reading’ of its own) also suggested that most of the calculated CSV had at least minor flaws in them due to ambiguous markup. These findings contributed to the error handling in the previous step (Acquisition of Network Data).

Some basic network calculations were done on the basis of the 12 (completed) theatre plays by Gotthold Ephraim Lessing. Corresponding diagrams are show in Figure 2.



# Conclusions

The extracted (and adjusted) network data will serve as a basis for further statistical calculations and also be made publicly available. Our research focus will now shift to implementing further calculation tools for the network analysis of theatre plays (e.g., to calculate the betweenness centrality to determine the importance of individual characters in a network). We will also work on enhancing the network data (quantify speech units, include stage presence of non-speaking persons, etc.) and try to build multiplex networks that not only capture the ‘interactions’, but also consider parental or instrumental relations.

References

**Marcus, S.** (1973). *Mathematische Poetik.* Athenäum-Verlag, Frankfurt.

**Moretti, F.** (2011). Network Theory, Plot Analysis. Stanford Literary Lab Pamphlets 2, http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.

**Pfister, M.** (1977). *Das Drama. Theorie und Analyse.* Fink, Munich.

**Pohlheim, K. K.** (ed.). (1997). *Die dramatische Konfiguration.* Schöningh, Paderborn.

**Titzmann, M.** (1977). *Strukturale Textanalyse. Theorie und Praxis der Interpretation.* Fink, Munich.

**Trilcke, P.** (2013). Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In Ajouri, P., Mellmann, K. and Rauen, C. (eds), *Empirie in der Literaturwissenschaft.* Münster: Mentis, pp. 201–47.

**Wasserman, S. and Faust, K.** (1998). *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge.