Digitale Netzwerkanalyse dramatischer Texte

Peer Trilcke[1], Frank Fischer[2], Dario Kampkaspar[3]

[1] Georg-August-Universität Göttingen
[2] Göttingen Centre for Digital Humanities
[3] Herzog August Bibliothek Wolfenbüttel

1 Einleitung

Das Projekt 'Digitale Netzwerkanalyse dramatischer Texte' steht in der Tradition strukturanalytischer Ansätze in der Literaturwissenschaft (allgemein Titzmann 1977), die es einerseits im Sinne eines konsequent netzwerkanalytischen Relationismus (mit Rekurs auf die Social Network Analysis, siehe u. a. Wasserman/Faust 1998), andererseits unterstützt durch Verfahren der automatisierten Datenerhebung und -auswertung weiterentwickelt, um sie auf größere Textkorpora anzuwenden und so umfassende relationale Daten über Prozesse des literaturgeschichtlichen Strukturwandels gewinnen zu können.

Als theoretisches Fundament dient dabei eine netzwerkanalytische Konzeptualisierung dramatischer Interaktion (erste Ideen dazu prominent bei Moretti 2011; Kritik und literaturtheoretisch begründete Rekonzeptualisierung bei Trilcke 2013 – dort auch ein ausführlicher Forschungsüberblick), die – in Fortführung von Konzepten der dramatischen Konfiguration (Marcus 1973, Pfister 1977; problematisch hingegen, weil mit diffuser Konzeptualisierung: Pohlheim 1997) – zunächst bei einer rudimentären Operationalisierung ansetzt, nach der eine 'Interaktion' dann vorliegt, wenn zwei Figuren innerhalb einer durch die überlieferte Struktur des Textes vorgegebenen Subsegmentierungseinheit (in der Regel 'Szene' oder 'Auftritt') als Sprecher aufgeführt werden.

'Interaktion' wird in diesem Sinne – und zu Zwecken einer ersten Automatisierung – verstanden als 'szenische Kopräsenz zweier Sprecher'. Auf Grundlage der so definierten Relation werden im Rahmen des Projekts automatisiert netzwerkanalytische Daten erhoben, die sowohl global die 'Interaktions'-Netzwerke der Dramen (Density, Average Degree, Connectedness u. dergl.) als auch fokussiert einzelne Akteure charakterisieren (Degree sowie diverse weitere Centrality-Indices). Der erstellte Workflow ermöglicht auch die Datenerhebung auf Mesoebene (u. a. Identifizierung von Clustern) und beinhaltet darüber hinaus Visualisierungen der Netzwerkdaten, die wiederum zur Analyse des literaturgeschichtlichen Strukturwandels beitragen.

2 Wahl des Dramenkorpus

Für die automatisierte Analyse von Dramen war ein verlässliches und genügend großes Dramenkorpus vonnöten. Infrage kamen hier:

- Deutsches Textarchiv (DTA): 49 Dramen[1]
- Wikisource: 50 Dramen[2]
- Projekt Gutenberg-DE: 641 Dramen[3]
- Textgrid Repository: 690 Dramen[4]

[1] `http://www.deutschestextarchiv.de`
[2] `http://de.wikisource.org/wiki/Kategorie:Drama`
[3] `http://projekt.gutenberg.de`
[4] `http://www.textgridrep.de`

Das DTA hat zwar das qualifizierteste (TEI-)Markup, besteht aber bisher nur aus vergleichsweise wenigen Texten. Letzteres gilt auch für die Dramen im deutschsprachigen Zweig von Wikisource. Das Projekt Gutenberg-DE wiederum, das seit 2002 bei Spiegel Online gehostet wird, hat das Problem, dass es nicht mit brauchbarem Markup versehen ist, nur rudimentärem XHTML. Deshalb kam eigentlich nur das TextGrid Repository infrage, das sich aus den alten Zeno.org-Volltexten speist und basale TEI-Auszeichnungen aufweist.

Aus dem TextGrid-Gesamtkorpus wurden zunächst die in den Metadaten mit dem Genre 'drama' versehenen Texte extrahiert, insgesamt 690. Dazu gehören vor allem deutschsprachige Dramen von ca. 1500 bis 1930 sowie ferner u. a. Übersetzungen von einem Dutzend griechischer Tragödien und einiger Shakespeare-Dramen. Aus der Gesamtmenge lassen sich prinzipiell auch recht einfach zeitlich gestaffelte Teilkorpora erstellen, denn im TEI-Header stehen innerhalb von `<creation></creation>` rudimentäre Entstehungsdaten (Beispiele: `<date when="1802"/>`, aber auch weitläufige Eingrenzungen wie `<date notBefore="1738" notAfter="1758"/>`).

3 Erhebung der Netzwerkdaten

Als Zwischenschritt wurde für jede der 690 TEI-Dateien eine Relationsliste (CSV-Datei) erzeugt, die den gängigen Formaten der Speicherung netzwerkanalytischer Daten entspricht. Zur Extraktion der Sprecherdaten sind in der Regel zwei getrennte Schritte nötig: Das Erkennen der einzelnen Teile eines Theaterstückes und danach das Erkennen der einzelnen Sprecher.

Zur Erleichterung der nachstehenden Arbeiten teilt das Skript die vorliegenden Dateien auf: Für jede erkannte Ebene (die Datei selbst ist dabei auch eine) wird ein Unterverzeichnis angelegt, in dem wiederum TEI-Dateien mit den einzelnen Teilen stehen und in das auch die jeweiligen Registerdateien geschrieben werden. Anhand der aufgeteilten Dateien werden verschiedene Arten von Ausgaben erstellt. Zum einen ist dies ein kleinteiliges Register aller `<speaker>`-Tags, aber auch aller Auszeichnungen `<rs>` und `<person>`. Um eindeutige Referenzziele zu erhalten, werden ggf. ID-Nummern vergeben (dies erleichtert insbesondere auch spätere Eingriffe, wenn problematische Namen manuell korrigiert werden müssen). Zum andern werden die Kookkurrenzlisten erstellt. Im untersten Verzeichnis werden die Vorkommen aller Sprecherpaare in allen Dateien gezählt. In den darüberliegenden werden die Werte aller Unterverzeichnisse addiert.

Neben dem Erkennen der Struktur ist die korrekte Zuordnung von Namen die größte Herausforderung. Im Idealfall sind alle `<speaker>` mit einem Attribut `@who` versehen, über das eine normierte Form des Namens erreicht werden kann. Ist dies nicht der Fall (oder sind stattdessen die Tags `<rs>` oder `<person>` verwendet worden), muss das Skript den Textinhalt des Tags auswerten, wobei neben möglichen Verschreibungen (bei der Transkription oder in der Vorlage) auch syntaktische Änderungen auftreten können. So findet sich bei Lessing, *Nathan der Weise* V/1, neben Saladin "Ein Mameluck", der nach seinem ersten Auftreten als "Der Mameluck" geführt wird. Es folgt ein weiterer: "Ein zweiter Mameluck", danach "zweiter Mameluck". Ist es hier noch möglich, durch ein Berücksichtigen der Artikel mit einfachen Mitteln gute Ergebnisse zu erzielen, stellt sich dies bei Emilia Galotti etwas schwieriger dar, da teils nur "Odoardo", teils aber "Odoardo Galotti" erscheint. Auch Fälle mit mehreren Sprechern (z. B. "Alle") sind nicht ganz trivial zu bearbeiten.

Neben dem Versuch, diese Fälle automatisch zu klären, besteht in diesen Zweifelsfällen aber immer noch die Möglichkeit des manuellen Eingriffs, wozu die erstellten Indexdateien mit eindeutiger ID beitragen können. In einer weiteren Überarbeitung des Skriptes ist es vorgesehen, eine einfache graphische Oberfläche anzubieten, über die solche Zweifelsfälle bearbeitet werden können.

4 Datenauswertung und Visualisierung

Die Datenauswertung erfolgt über Python (3.4.x) mit dem igraph-Paket, das sowohl zum Visualisieren der Graphen als auch zum Berechnen der netzwerkanalytischen Daten genutzt wird.

Für eine erste Visualisierung des Datenbestands wurden die Graphdaten an eine spring-embedding-Methode übergeben (Fruchterman-Reingold), die versucht, affine Knoten näher beieinander anzuordnen und dadurch deutlich sichtbar zu clustern. Einen Eindruck des gesamten Korpus vermittelt

Abbildung 1, die 671 Dramen aus 2500 Jahren Dramengeschichte enthält, chronologisch links oben mit den Griechen beginnend und bis rechts unten ins zweite Viertel des 20. Jahrhunderts reichend:

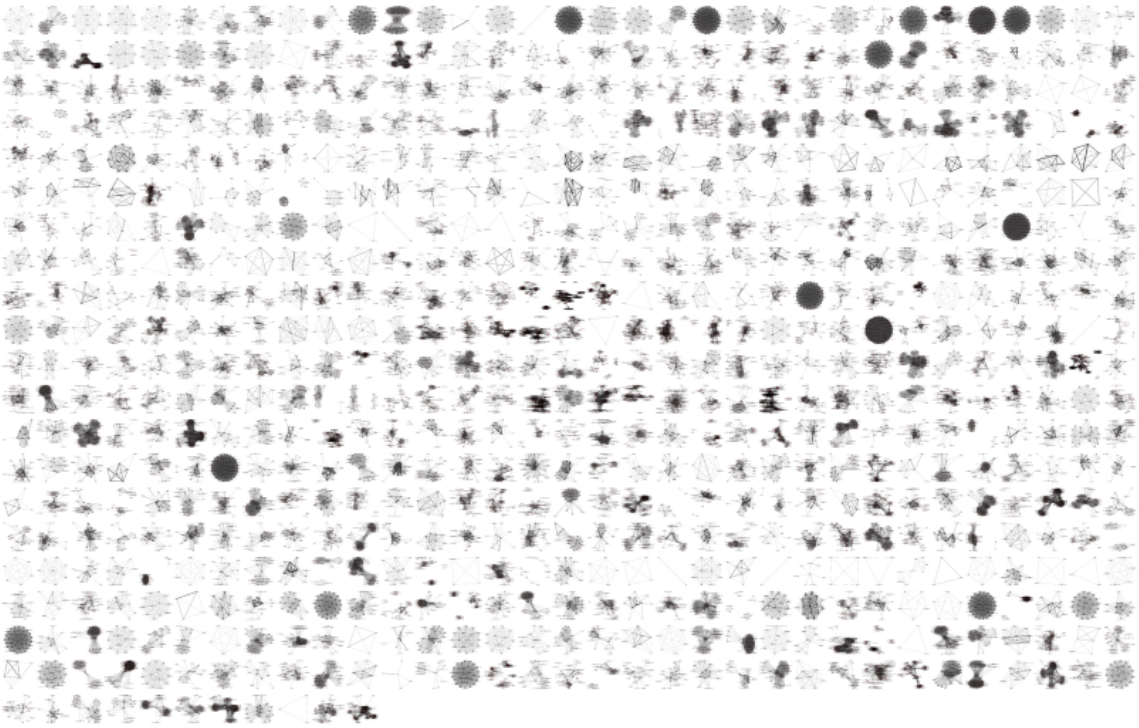

Abbildung 1: Netzwerkgraphen von 671 Dramen aus dem TextGrid Repository

Die visualisierten Graphen haben auch deutlich gemacht, dass die meisten berechneten CSV-Dateien wegen des teils nicht eindeutigen Markups zumindest kleine Fehler aufwiesen. Diese Erkenntnisse konnten zur Fehlerbehandlung an den vorhergehenden Schritt (Erhebung der Netzwerkdaten) zurückgegeben werden.

Erste strukturanalytische Berechnungen erfolgten auf Basis der 12 (vollendeten) Lessing-Dramen. Entsprechende Diagramme sind in Abbildung 2 zu finden.

5 Ausblick

Die erhobenen und bereinigten Netzwerkdaten sollen als Grundlage für alle statistischen Berechnungen dienen und auch öffentlich zur Verfügung gestellt werden. Im Mittelpunkt der Forschung steht nun die Implementierung zusätzlicher netzwerkanalytischer Berechnungstools (etwa zur Bestimmung der Betweenness Centrality, mit der die Wichtigkeit einzelner Figuren für das Netzwerk bestimmt werden kann). Darüber hinaus wird an der Qualifizierung der Netzwerkdaten gearbeitet (außer dem reinen Fakt, dass Figuren miteinander sprechen: Redeanteile quantifizieren, Bühnenpräsenz nicht sprechender Personen mit einbeziehen usw.) sowie an der Erstellung multiplexer Netzwerke, die nicht nur die oben definierten ‘Interaktions’-Relationen erfassen, sondern auch u. a. Verwandschafts- oder instrumentelle Beziehungen berücksichtigen.

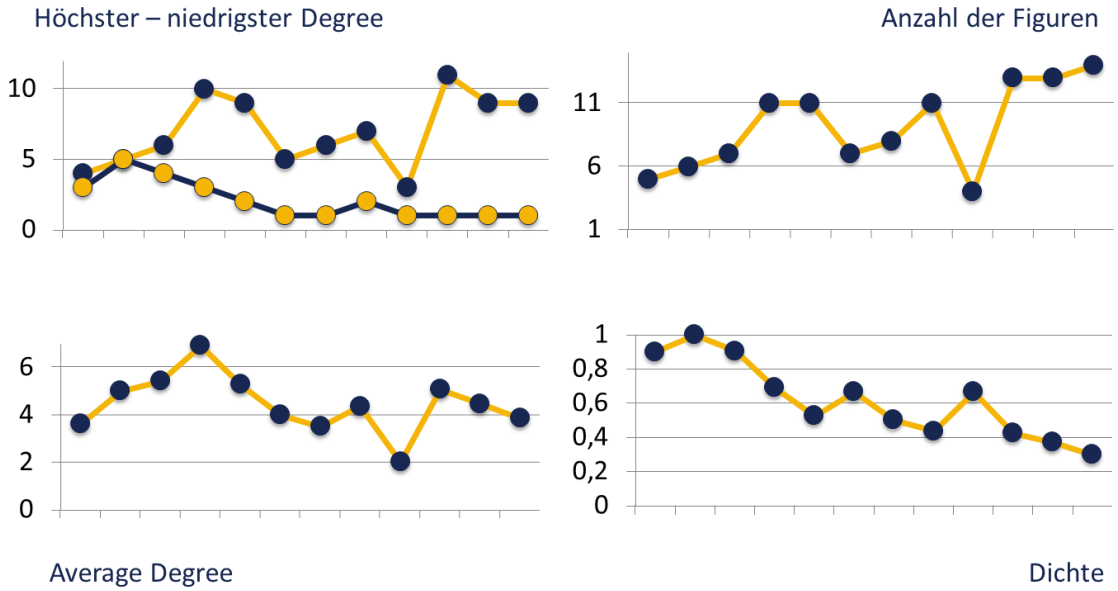

Abbildung 2: Beispielberechnungen anhand der Lessing-Dramen (x-Achse): Damon, 1747 – Der junge Gelehrte, 1747 – Der Misogyn, 1748 – Die alte Jungfer, 1748 – Der Freigeist, 1749 – Die Juden, 1749 – Der Schatz, 1750 – Miß Sara Sampson, 1755 – Philotas, 1759 – Minna von Barnhelm, 1767 – Emilia Galotti, 1772 – Nathan der Weise, 1779

Literatur

Marcus, Solomon. *Mathematische Poetik.* Frankfurt/M. 1973.

Moretti, Franco. *Network Theory, Plot Analysis.* Stanford Literary Lab Pamphlets 2 (1. 5. 2011). `http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf`.

Pfister, Manfred. *Das Drama. Theorie und Analyse.* München 1977 u. ö.

Pohlheim, Karl Konrad (Hg.). *Die dramatische Konfiguration.* Paderborn u. a. 1997.

Titzmann, Michael. *Strukturale Textanalyse. Theorie und Praxis der Interpretation.* München 1977 u. ö.

Trilcke, Peer. *Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft.* In: Philip Ajouri, Katja Mellmann u. Christoph Rauen (Hg.): Empirie in der Literaturwissenschaft. Münster 2013. S. 201–247.

Wasserman, Stanley; Katherine Faust. *Social Network Analysis. Methods and Applications.* Cambridge u. a. 1998.