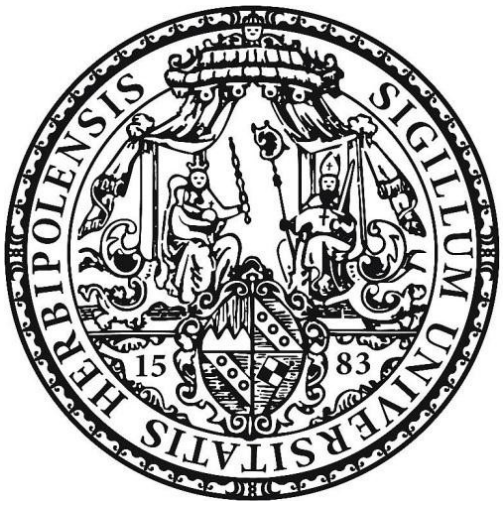
Julius-Maximilians-Universität Würzburg

Lehrstuhl für Computerphilologie
Studienfach: Digital Humanities

Masterarbeit



Figurennetzwerke als Ähnlichkeitsmaß

08.08.2016

Isabella Reger

Betreuer: Prof. Dr. Fotis Jannidis

Zweitgutachter: Dr. Christof Schöch

Inhaltsverzeichnis

1 Einleitung

Anfang 2011 veröffentlichte Hugh Craig einen Artikel, in dem er mit Hilfe quantitativer Methoden zeigen konnte, dass William Shakespeare im Vergleich zu seinen Zeitgenossen in seinen Werken keinen überdurchschnittlich hohen Wortschatz verwendet, obwohl dies eine gängige Annahme auf dem Gebiet der Shakespeare-Studien war (Craig 2011). Patrick Juola fand 2013 mittels statistischer Untersuchungen heraus, dass der unter dem Pseudonym Robert Galbraith erschienene Roman *The Cuckoo's Calling* mit hoher Wahrscheinlichkeit von Joanne K. Rowling stammt (Juola 2013). Kurze Zeit später bestätigte die Autorin der Harry Potter-Reihe dies öffentlich.

Diese bekannten Beispiele aus der Stilometrie zeigen, dass die quantitative Analyse von Literatur Erkenntnisse liefern kann, die mit klassischem Close Reading nicht möglich wären und die bestehende Forschungsansichten der Literaturwissenschaft revidieren können. In der jüngeren Vergangenheit ist daher auch die Zahl der Arbeiten, die quantitative Methoden auf literarische Texte - seien es Romane, Dramen oder Lyrik - anwenden, kontinuierlich gestiegen.

Neben der stilistischen Gestaltung können auch andere Aspekte von literarischen Werken mit computergestützten Methoden untersucht werden. Die inhaltlich-thematische Zusammensetzung eines Textes kann beispielsweise mit Hilfe von Topic Modeling repräsentiert und so unter anderem für die Betrachtung literarischer Gattungen genutzt werden. Derartige quantitative Ansätze basieren immer auf Modellen zur Operationalisierung bestimmter Eigenschaften literarischer Texte.

Ein weiterer Bereich in der Literaturwissenschaft, der mit Hilfe einer solchen Operationalisierung untersucht werden kann, ist die Figurenkonstellation in Romanen oder Dramen. Abgesehen vom Einzeltext ist die Untersuchung der Figurenkonstellation auch für größere Textsammlungen interessant: Mögliche Forschungsfragen könnten sein, ob es wiederkehrende Typen von Figurenkonstellationen gibt, die sich, unter Umständen in leichter Abwandlung, in verschiedenen Romanen finden lassen oder ob bestimmte Konstellationen typisch für bestimmte Gattungen sind.

Die Figurenkonstellation eines Romans lässt sich durch ein soziales Netzwerk repräsentieren, das automatisch aus dem Romantext extrahiert werden kann.

Die vorliegende Arbeit befasst sich anhand eines Korpus von deutschsprachigen Romanen aus dem 19. Jahrhundert mit der Frage, ob mit Hilfe solcher Figurennetzwerke automatisch festgestellt werden kann, inwieweit sich Romane im Hinblick auf ihre Figurenkonstellation ähnlich sind. Dazu werden Methoden der Sozialen Netzwerkanalyse verwendet, um Eigenschaften der Figurennetzwerke zu erheben, anhand derer dann Distanzen berechnet werden können, die die Ähnlichkeit zwischen Romanen angeben. Als Evaluationsgrundlage dient die menschliche Intuition solcher Ähnlichkeit, festgehalten in Form einer manuell erstellten Distanzmatrix.

Im Folgenden wird anhand eines Beispiels erläutert, welches Konzept von Ähnlichkeit zwischen Figurenkonstellationen dieser Arbeit zugrunde liegt. Der darauf folgende Abschnitt gibt einen Überblick über bestehende Forschung in den relevanten Themengebieten. Anschließend werden die Daten- und Evaluationsgrundlage, sowie die geplanten Experimente und Berechnungen vorgestellt. Die Abschnitte 5 und 6 erläutern die anhand von Figurennetzwerken durchgeführten Experimente und deren Kombination mit Topic Modeling. Danach findet eine Diskussion der Ergebnisse und Beobachtungen statt. Den Abschluss bildet ein Ausblick auf zukünftige nötige und mögliche Arbeiten.

2 Theoretischer und thematischer Hintergrund

Das Ziel dieser Arbeit ist, die Figurenkonstellation von Romanen mit computergestützten Methoden zu modellieren und davon ausgehend Ähnlichkeiten zwischen Romanen festzustellen. Daher werden diese beiden Begriffe – Figurenkonstellation und Ähnlichkeit – in diesem Abschnitt unter Einbeziehung eines Beispiels näher beleuchtet.

Das Reallexikon der Deutschen Literaturwissenschaft definiert Figurenkonstellation als „Ensemble aller in einem Drama oder Erzähltext vorkommenden fiktiven Personen“ (Weimar et al. 2010, S. 591). Zusätzlich zu ihren Eigenschaften und Merkmalen ist jede Einzelfigur über ihre Einbettung ins Figurenensemble definiert. Das macht die Figurenkonstellation zu einem wichtigen strukturgebenden Ordnungsprinzip in literarischen Texten. Zwischen den Figuren finden im Verlauf der Handlung die verschiedensten Interaktionen und Veränderungen statt; das Zusammenspiel aller Figuren zu einem bestimmten Zeitpunkt des Textes wird als Konfiguration bezeichnet. Das übergreifende Konzept der Figurenkonstellation

bleibt jedoch „im ganzen Text konstant“ (Weimar et al. 2010, S. 591), gewissermaßen in einem aggregierten Zustand über die einzelnen Konfigurationen.

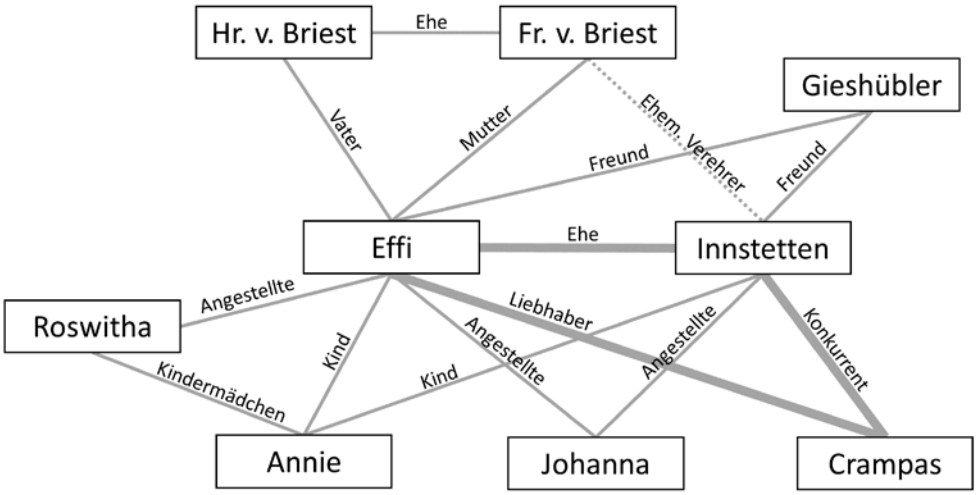
Alex Woloch (Woloch 2003) betrachtet den Raum innerhalb eines Romans, der einer Figur gewidmet wird (*character space*), als zentral für deren Charakterisierung. Mit ‚Raum‘ bezeichnet er hier die Aufmerksamkeit, quantifiziert über die Menge des Textes, die der Figur entgegengebracht wird. Dabei betont er, dass Romanfiguren insbesondere durch das Verhältnis der Figuren untereinander und ihre Einbettung in eine Gesamtstruktur (*character system*) für den Leser greifbar werden. Die Dominanz des Protagonisten wird erst durch seine Eingliederung in ein Ensemble weniger wichtiger Nebenfiguren realisiert. Obwohl Wolochs Studie im Folgenden eher auf die Charakterisierung von Figuren fokussiert ist, hebt er dennoch das Figurensystem als Ganzes als zentralen Aspekt von Romanen hervor.

Einem ähnlichen Konzept folgt auch Dieter Kafitz und versucht sich an einer auf der Figurenkonstellation basierenden Typologie des Romans (Kafitz 1978, S. 10–18). Er entwirft ein Spektrum, an dessen einen Ende er den „Roman mit herausgehobenem Helden“ sieht, bei dem die Nebenfiguren einem zentralen Helden untergeordnet sind, während am anderen Ende der „Vielheitsroman“ steht, in dem eine Vielzahl von Figuren nebeneinander gleichermaßen von Bedeutung ist und die Hierarchie in Haupt- und Nebenfiguren nur schwach ausgeprägt ist. Zwischen diesen beiden Polen finden sich verschiedene Abstufungen, wie der Roman des „Doppelhelden“, in dem die Relation zwischen zwei Hauptfiguren im Zentrum steht oder der Roman, in dem sich die Figurenkonstellation um eine Figurengruppe, beispielsweise eine Familie oder andere soziale Gemeinschaft, formiert. Kafitz sieht auch einen Zusammenhang zwischen den genannten Typen von Figurenkonstellationen und Gattungen: Laut seiner Darstellung findet sich die Konstellation des Doppelhelden oft in romantischen Romanen, während es sich bei „Romanen mit herausgehobenen Helden“ häufig um Entwicklungs- oder Bildungsromane handelt. Dies belegt er jedoch rein exemplarisch.

Diese Ausführungen von Kafitz zeigen, dass verschiedene Romane gewisse wiederkehrende Strukturen in ihrer Figurenkonstellation aufweisen können und unterstützen damit die Annahme, dass Romane anhand ihrer Figurenkonstellation als ähnlich (oder unähnlich) eingestuft werden können.

Im Folgenden soll nun anhand eines Beispiels verdeutlicht werden, wann man von einer Ähnlichkeit zwischen den Figurenkonstellationen von Romanen

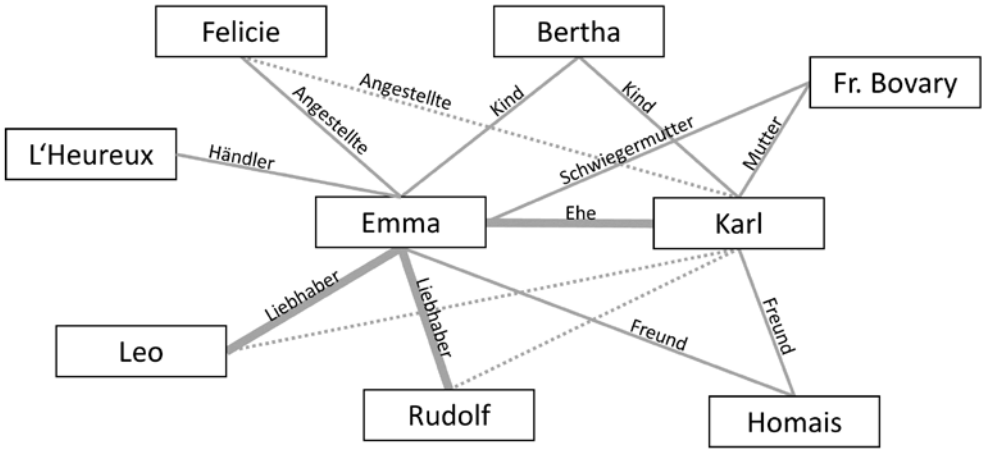
sprechen würde. Zwei Romane, die einige Gemeinsamkeiten aufweisen und in zahlreichen Studien miteinander in Verbindung gesetzt wurden, sind Theodor Fontanes *Effi Briest* und Gustave Flauberts *Madame Bovary* (Dethloff 2000; Degering 1978; Bonwit 1948). Tatsächlich lassen sich auch in den Figurenkonstellationen beider Romane (Abbildungen 1 und 2) viele Parallelen finden.



1 Figurenkonstellation zu Effi Briest

In beiden Texten steht eine junge Frau im Fokus, die in recht jungen Jahren einen deutlich älteren Mann heiratet, in dessen Haushalt in eine fremde Kleinstadt zieht und ihre Familie hinter sich lässt. Diese Ehe stellt jeweils die zentrale Relation des Romans dar, eingebettet in einen sozialen Kontext bestehend aus den Eltern bzw. Schwiegereltern, den Angestellten im Haushalt und Mitgliedern der Gemeinde. In beiden Fällen spielt hier der Apotheker des Ortes, Alonzo Gieshübler in *Effi Briest* und Monsieur Homais in *Madame Bovary*, eine Rolle, mit dem die Paare einen nachbarschaftlichen, freundschaftlichen Umgang pflegen. Es wird jedoch recht schnell klar, dass sich sowohl Effi, als auch Emma in ihrer Beziehung durch das hohe Arbeitspensum ihrer Ehemänner einsam und gelangweilt fühlen. Beide sehnen sich nach einem aufregenderen, stimulierenden gesellschaftlichen Leben, das ihre Wohnorte jedoch nicht zu bieten haben. Im Verlauf der Handlung bekommen beide Paare eine Tochter. Weder Effi, noch Emma sind jedoch in der Lage, ihr Kind bedingungslos zu lieben, weshalb beide Töchter hauptsächlich von Kinderfrauen aufgezogen werden. So nimmt die Entfremdung der Ehepartner ihren Lauf, bis die Frau sich zum heimlichen Ehebruch hinreißen lässt bzw. entscheidet.

Natürlich unterscheiden sich die beiden Romane in ihrem Handlungsverlauf und in den Details auch deutlich – legt man den Fokus aber auf das Figureninventar und die Beziehungen der Figuren untereinander, so sind zweifelsfrei deutliche Parallelen zwischen den Figurenkonstellationen beider Romane feststellbar.



2 Figurenkonstellation zu Madame Bovary

Abbildungen 1 und 2[1], sowie die vorhergehende Beschreibung, zeigen auch, dass sich der Eindruck der Figurenkonstellation eines Romans, sowie die Einschätzung von Ähnlichkeit zwischen den Figurenkonstellationen verschiedener Romane auf zwei Aspekte stützt: Zum einen auf die grundlegende Struktur, also die Art, wie die Figuren untereinander vernetzt sind und wie stark einzelne Figuren in das Gesamtbild eingebettet sind und zum anderen auf gewisse Eigenschaften von Figuren, wie Geschlecht, sozialer Status oder der Beruf, sowie die Art der Beziehungen und gemeinsame Motive zwischen Figuren. *Madame Bovary* und *Effi Briest* weisen in dieser Hinsicht, wie bereits beschrieben, einige deutliche Parallelen auf und können somit als gutes Beispiel für zwei Romane mit ähnlicher Figurenkonstellation gelten.

Die vorliegende Arbeit untersucht die Frage, ob es möglich ist, die Figurenkonstellation von Romanen mit computergestützten Methoden zu modellieren und

[1] Die Übersichtsgrafiken wurden von der Verfasserin nach Lektüre beider Romane erstellt. Dicke Linien deuten eine besonders starke Beziehung zwischen zwei Figuren an. Gestrichelte Linien stehen für Beziehungen, die zwar faktisch vorhanden, aber für den Roman von geringerer Bedeutung sind.

anhand dessen automatisch festzustellen, ob und welche Romane sich hinsichtlich der beschriebenen Aspekte ähnlich sind.

3 Methoden und Forschungsstand

In den folgenden Abschnitten werden die computergestützten Methoden, die in dieser Arbeit zum Einsatz kommen, eingeführt und bisherige Forschung auf den entsprechenden Gebieten thematisiert. Nach einem kurzen Überblick über die Grundlagen der Graphentheorie werden bestehende Ansätze zur Erstellung von Figurennetzwerken sowie darauf aufbauende Untersuchungen dargestellt. Anschließend wird die Latent Dirichlet Allocation als eine Methode des Topic Modeling vorgestellt und ebenfalls relevante Anwendungsbeispiele aus den Geisteswissenschaften und speziell zur Analyse von literarischen Texten aufgezeigt.

3.1 Figurennetzwerke

3.1.1 Grundlagen der Netzwerktheorie

Ein Netzwerk oder Graph besteht aus einer Menge von Objekten, den sogenannten Knoten, und bestimmten Verbindungen zwischen den Objekten, die als Kanten bezeichnet werden.[2] Je nach Art des Netzwerks kann es sich bei den Knoten um verschiedene Arten von Objekten handeln: In einem Computernetzwerk wäre beispielsweise jeder Rechner ein Knoten, während in einem sozialen Netzwerk Personen als Knoten repräsentiert werden. Als Kanten kämen entsprechend die Datenverbindungen zwischen Computern oder die Freundschafts- und Verwandtschaftsbeziehungen von Personen in Betracht. Die Kanten eines Netzwerks können gerichtet sein, falls eine Verbindung explizit von einem Knoten ausgeht und nur in diese Richtung durchlaufen werden kann, oder ungerichtet, falls dem nicht so ist. Außerdem kann ein Graph gewichtet sein: In diesem Fall tragen alle Kanten numerische Gewichte, die die Stärke der Verbindung angeben.

Solche Netzwerke können auf verschiedene Weisen repräsentiert werden. Gängige Methoden sind die Darstellung als Kantenliste, in der alle Verbindungen

[2] In Newman (2010) findet sich ab Seite 109 eine gute, umfassende Einführung in die konzeptionellen und mathematischen Grundlagen der Graphentheorie.

zwischen Knoten entsprechend ihrer Richtung und gegebenenfalls mit ihrem Gewicht aufgelistet sind, oder als Adjazenzmatrix, die für jeden Knoten eine Zeile und eine Spalte besitzt und eine 1 bzw. das Kantengewicht enthält, sofern zwischen zwei Knoten eine Verbindung existiert.

In Analogie zu einem sozialen Netzwerk kann auch die Figurenkonstellation eines Romans als Graph dargestellt werden. Dabei werden die Figuren durch Knoten und die Beziehungen zwischen Figuren durch Kanten repräsentiert. Je nach Modellierung der Relationen zwischen Figuren können die Kanten zudem gewichtet sein oder eine Richtung besitzen.

Mit entsprechenden computergestützten Methoden können derartige Netzwerke auch automatisch generiert werden. Im Folgenden werden zunächst bestehende Ansätze betrachtet, die Netzwerke aus Textdaten extrahieren, bevor diese Idee auf literarische Texte ausgeweitet wird. Anschließend werden verschiedene Möglichkeiten zur Modellierung von Interaktionen zwischen Figuren, sowie auf Basis von Figurennetzwerken durchgeführte Untersuchungen näher beleuchtet.

3.1.2 Netzwerke aus Textdaten

Die Idee, soziale Netzwerke aus textuellen Daten automatisch zu erstellen, existiert bereits seit längerem und wird in verschiedenen Domänen immer wieder aufgegriffen.

Culotta et al. (Culotta et al. 2004) analysieren das Email-Postfach eines Nutzers, um daraus dessen persönliches soziales Netzwerk zu extrahieren. Außerdem wenden sie Methoden der Sozialen Netzwerkanalyse an, um innerhalb dieses Netzwerks besonders wichtige Personen zu identifizieren oder Vorschläge zu generieren, welcher Kontakt sinnvoll bei einer bestimmten Email in Kopie (‚CC‘) gesetzt werden könnte.

Jing et al. (Jing et al. 2007) erstellen soziale Netzwerke anhand von gesprochener Sprache und verwenden als Datengrundlage Transkriptionen von aufgezeichneten Interviews mit Holocaust-Überlebenden. Eine besondere Herausforderung stellte hierbei dar, dass die einzelnen Sprecher zwar mit Bezeichnungen wie ‚Speaker1‘ und ‚Speaker2‘ gekennzeichnet waren, diese Begriffe jedoch nicht auf die tatsächlich während des Interviews anwesenden Personen aufgelöst wurden.

Auch Gruzd und Haythornthwaite (Gruzd und Haythornthwaite 2008) extrahieren soziale Netzwerke aus textuellen Daten, nämlich aus den Diskussionssträngen eines Online-Forums. Dabei wenden sie Methoden des Natural Language Processing an, um anhand syntaktischer und semantischer Eigenschaften die verschiedenen expliziten und impliziten Relationen zwischen den Nutzern des Forums, sowie die Stärke dieser Verbindungen, zu ermitteln.

3.1.3 Netzwerke aus literarischen Texten

Die bisher genannten Arbeiten haben eine Gemeinsamkeit, die als Vorteil bei der automatischen Erstellung von sozialen Netzwerken gewertet werden kann: Die zugrundeliegenden Daten weisen gewisse wiederkehrende, explizit ausgewiesene Strukturen auf. Interessiert man sich für die Analyse literarischer Texte, ist dies sehr viel seltener der Fall. In Dramen gibt es noch einige Strukturelemente, wie die Aufteilung in Akte und Szenen oder die Kennzeichnung der Sprecher, die sich zur automatischen Erstellung von sozialen Netzwerken nutzen lassen (Trilcke et al. 2015). Diese können sich jedoch von Werk zu Werk stark unterscheiden. Romane und andere Erzähltexte bieten jedoch kaum derartige Struktureigenschaften, sofern diese nicht aufwändig per Hand im Text kodiert wurden. Die automatische Generierung von Figurennetzwerken aus literarischen Prosatexten erfordert also einige Vorverarbeitungsschritte und gewisse Operationalisierungen, für die in der Forschungsgemeinde bereits verschiedene Vorschläge gemacht wurden.

Der erste Schritt besteht in der Regel darin, die im Text vorkommenden Figuren zu identifizieren. Hierfür werden meist bestehende Named Entity Recognition Systeme genutzt. Anschließend sollten die gefundenen Figurenreferenzen bestenfalls durch eine Koreferenzauflösung aufeinander bezogen werden. Da dies ohnehin ein schwieriger Task ist und bestehende Systeme in den meisten Fällen anhand von Zeitungskorpora entwickelt wurden, sind die Ergebnisse für Romantexte hier oft mäßig, sodass häufig auf stark vereinfachte Notlösungen zurückgegriffen oder auf Koreferenzauflösung verzichtet wird. Es ist jedoch nicht klar, wie stark sich das auf die resultierenden Figurennetzwerke auswirkt. Jannidis et al. und Krug et al. haben sowohl für die Erkennung der Figuren, als auch für die Koreferenzauflösung eigene Systeme entwickelt, die auf deutschsprachige Romane des 19. Jahrhunderts spezialisiert sind (Jannidis et al. 2015; Krug et al. 2015). Beide Systeme werden in dieser Arbeit genutzt und in Abschnitt 5.1 genauer erläutert.

Auch und vor allem bei der Modellierung von Interaktionen zwischen Figuren, also wann eine Kante zwischen zwei Figuren gezogen wird, können verschiedene Wege eingeschlagen werden. Elson et al. identifizieren Passagen direkter Rede in Romanen und ordnen jeder Instanz von direkter Rede eine Figur als Sprecher zu (Elson und McKeown 2010). Anhand der im Text vorkommenden Dialogstrukturen erstellen sie anschließend Figurennetzwerke aus britischen Romanen des 19. Jahrhunderts (Elson et al. 2010). Auch Celikyilmaz et al. extrahieren soziale Netzwerke auf Basis der Dialoge aus englischen Romanen des gleichen Zeitraums, nutzen dafür jedoch ein unüberwachtes Verfahren, das Actor-Topic-Model, für die Zuordnung von Sprechern zu jeweiligen direkten Reden (Celikyilmaz et al. 2010). Auch für deutsche Romane des 18. bis 20. Jahrhunderts haben Krug et al. (Krug et al. 2016a) ein regelbasiertes Verfahren entwickelt, das sowohl den Sprecher, als auch den Angesprochenen einer direkten Rede ermittelt. Diese Informationen können, ähnlich zu Elson et al., ebenfalls für die Erstellung von Figurennetzwerken herangezogen werden, sofern ein Roman einen ausreichenden Anteil an direkter Rede enthält.

Andere Ansätze stützen sich auf gemeinsames Vorkommen von Figuren im Text. Park et al. bestimmen den Abstand zwischen Figurenreferenzen im Text und definieren eine Distanzfunktion, um aus den gemessenen Abständen die Stärke der entsprechenden Relation zu berechnen (Park et al. 2013). Coll Ardanuy und Sporleder betrachten als Alternative zu Dialogstrukturen auch Netzwerke, bei denen eine Interaktion angenommen wird, sobald zwei oder mehr Figuren im gleichen abgeschlossenen Textabschnitt, beispielsweise einem Absatz, vorkommen (Coll Ardanuy und Sporleder 2014).

Die Arbeitsgruppe um Apoorv Agarwal verfolgt eine weitere Vorgehensweise, die sie zunächst in einer manuellen Studie vorschlagen (Agarwal et al. 2012) und anschließend in Form von SINNET als computergestützte Methode umsetzen (Agarwal et al. 2013a; Agarwal et al. 2013b). Dabei modellieren sie zwei Arten von Ereignissen zwischen Figuren: „Interactions“, bei denen beide Parteien sich der Interaktion bewusst sind, wie beispielsweise ein Gespräch, und „Observations“, die nur einer Figur bewusst sind, weil sie zum Beispiel über eine abwesende Figur spricht oder nachdenkt. Neben der Modellierung von zwei unterschiedlichen Interaktionstypen ist eine Besonderheit der Arbeit von Agarwal et al., dass sie anhand von Lewis Carrolls *Alice im Wunderland* ein automatisch generiertes Netzwerk und

ein manuell erstelltes Netzwerk im Detail miteinander vergleichen, um die Ergebnisse ihres Systems einschätzen zu können. In vielen anderen Arbeiten werden zwar Methoden zur Extraktion von Netzwerken aus Textdaten vorgestellt, die entstandenen Netzwerke als solches aber nicht weiter ausgewertet, sondern direkt zu weiterführenden Analysen herangezogen.

3.1.4 Analyse und Evaluation von Figurennetzwerken

In der Sozialen Netzwerkanalyse, wie sie in den Sozialwissenschaften oder auch der Psychologie angewandt wird, sowie auch in der Netzwerkanalyse seitens der Mathematik und der Informatik, gibt es zahlreiche Methoden und Metriken zur Beschreibung und Auswertung von Netzwerkdaten. Franco Moretti, als prominentes Beispiel, hat Konzepte wie Zentralität oder Dichte in die Domäne der Literaturwissenschaft übertragen: Unter Verwendung solcher Methoden analysiert er ein manuell erstelltes Figurennetzwerk aus Shakespeares *Hamlet* im Detail und aus klarer literaturwissenschaftlicher Perspektive (Moretti 2011). Peer Trilcke (Trilcke 2013) versucht, im Sinne einer Methodenentwicklung zur Analyse von Figurennetzwerken aus literarischen Texten, systematisch Konzepte aus der Sozialen Netzwerkanalyse auf die Literaturwissenschaft zu übertragen. Anhand eines Anwendungsbeispiels, in dem er Figurennetzwerke mehrerer deutscher Dramen vergleicht, macht er den Einsatz dieser Techniken in Bezug auf literaturwissenschaftliche Fragestellungen deutlich.

Nach der automatischen Erstellung von Figurennetzwerken aus literarischen Texten ist der nächste folgerichtige Schritt deren ebenfalls computergestützte Analyse auf Basis größerer Korpora. Elson et al. (Elson et al. 2010) nutzen die quantitative Netzwerkanalyse, um die literaturwissenschaftlich begründeten Hypothesen, dass die Struktur des Figurennetzwerks mit der Menge an Dialogen im Roman sowie dem Setting (urban oder ländlich) zusammenhängt, zu widerlegen. Krug et al. (Krug et al. 2016b) nutzen sowohl auf Dialogstrukturen als auch auf Kookkurrenzen basierende Figurennetzwerke, um die Hauptfiguren eines Romans zu identifizieren, und evaluieren ihren Ansatz mit Hilfe von automatisch ausgewerteten Inhaltszusammenfassungen. Coll Ardanuy und Sporleder (Coll Ardanuy und Sporleder 2014) extrahieren Features aus Figurennetzwerken, um anhand derer Romane nach Gattung und nach Autorschaft zu clustern. Insbesondere die Ergebnisse für

das Clustering nach Gattungen sind jedoch unzureichend, was sie auf die Problematik zurückführen, dass literarische Gattungen ein kaum klar abgrenzbares Konzept darstellen.

Dieser Überblick zeigt, dass vielfältige Fragestellungen mit Hilfe von Figurennetzwerken adressiert werden können, wobei in manchen Fällen die erwarteten Ergebnisse ausbleiben.

3.2 Topic Modeling

3.2.1 Latent Dirichlet Allocation

Ein weiteres Verfahren der quantitativen Textanalyse, das in dieser Arbeit zum Einsatz kommen soll, ist Topic Modeling. Dabei handelt es sich um einen Überbegriff für eine Reihe von Verfahren, die Muster von Kookkurrenzen in Daten, häufig Textdaten, aufdecken. Im Allgemeinen werden Texte dabei als sogenannte Bag-of-Words-Modelle repräsentiert, also nur anhand der vorkommenden Wörter, ungeachtet deren Reihenfolge und syntaktischer Struktur. Ein Ziel von Topic Modeling ist es, die versteckte inhaltliche Zusammensetzung eines Korpus ans Licht zu bringen.

Die bekannteste und am häufigsten genutzte Umsetzung dieser Technik ist die sogenannte Latent Dirichlet Allocation, kurz LDA (Blei et al. 2003). LDA ist ein generatives Modell, das davon ausgeht, dass sich jeder Text gemäß einer feststehenden Wahrscheinlichkeitsverteilung aus verschiedenen Topics zusammensetzt. Ein Topic ist dabei definiert als eine feste Menge von Wörtern, die innerhalb des Topics ebenfalls einer bestimmten Wahrscheinlichkeitsverteilung unterliegen. In einem Topic sammeln sich mit hoher Wahrscheinlichkeit Wörter, die aus einem gemeinsamen semantischen Feld, gewissermaßen einem Themengebiet, stammen. LDA nimmt an, dass ein Text in einem generativen Prozess entsteht, indem wiederholt entsprechend der jeweiligen Wahrscheinlichkeitsverteilungen verschiedene Topics und daraus wiederum Worte ausgewählt werden.

Um LDA zur Untersuchung eines Korpus anzuwenden, muss dieser Prozess gewissermaßen umgedreht werden. Ausgehend von den Originaltexten werden die zugrundeliegenden Topics und deren Verteilung über das Korpus berechnet, indem

Kookkurrenzen von Wörtern in den gleichen Dokumenten ausgewertet werden. Daher benötigt LDA als unüberwachtes Verfahren keine aufwändig annotierten Trainingsdaten. Die resultierenden Topics spiegeln die inhaltliche und thematische Zusammensetzung des betrachteten Korpus wider. In diesem interpretierbaren und für anschließende Experimente verwendbaren Ergebnis liegen die Nützlichkeit und der Charme von LDA (Blei 2012).

3.2.2 Anwendungsbeispiele

Aufgrund seiner Funktionsweise wird LDA häufig zur Modellierung und Untersuchung auch größerer Textkorpora eingesetzt, wodurch es als Methode auch für die Geisteswissenschaften von Interesse ist. Im Folgenden sollen typische Anwendungsbeispiele für Topic Modeling aufgezeigt werden und dargelegt werden, inwiefern Topic Modeling auch für die Arbeit mit literarischen Texten herangezogen werden kann.

Griffiths und Steyvers (Griffiths und Steyvers 2004) verwenden LDA, um in einer Datenbank von wissenschaftlichen Publikationen den Zusammenhang zwischen vergebenen Kategorien wie ‚Mathematik' oder ‚Psychologie' und den tatsächlich in den Abstracts vorkommenden Themen, repräsentiert durch Topics, zu untersuchen. Auch für diachrone Studien kann Topic Modeling eingesetzt werden: David Mimno (Mimno 2012) betrachtet eine Sammlung von mehreren Zeitschriften zu Altertumswissenschaften und verwendet LDA, um thematische Ähnlichkeiten zwischen verschiedenen Zeitschriften oder auch die Variabilität von Vokabular oder Themen in einem Journal über den Zeitverlauf zu untersuchen.

Im Hinblick auf die Analyse von literarischen Texten ist beispielsweise ein Blogartikel von Cameron Blevins häufig zitiert worden, in dem er das Tagebuch der Martha Ballard, einer Hebamme im 18. Jahrhundert, mit LDA modelliert und das Auftreten bestimmter Topics über den Textverlauf betrachtet (Blevins 2010). Matthew Jockers beschäftigte sich in Zusammenarbeit mit Mimno (Jockers und Mimno 2013) und auch in seinem Buch *Macroanalysis* (Jockers 2013) anhand von englischsprachigen Romanen ausführlich mit der Frage nach dem Zusammenhang zwischen dem Vorkommen bestimmter Themen, modelliert mit LDA, und Faktoren wie dem Geschlecht oder der Nationalität des Autors, dem Erscheinungsjahr oder der Gattung eines Textes. Dabei kann er zeigen, dass sich ein solcher Zusammenhang mit Topic Modeling durchaus nachweisen lässt: das Themengebiet ‚Mode der

Frau‘ ist beispielsweise bei weiblichen Autorinnen sehr viel stärker vertreten als bei Männern, während bei diesen das Thema ‚Feindschaft‘ sehr viel häufiger auftritt (Jockers 2013, S. 136–138). Ähnliche Ergebnisse hat Christof Schöch erhalten, der mit Hilfe von LDA anhand einer Sammlung französischer Kriminalliteratur zeigen konnte, dass verschiedene Autoren, Untergattungen und Zeiträume durch unterschiedliche Topic-Verteilungen charakterisiert sind (Schöch 2015). Ebenfalls mit Untergattungen von Romanen beschäftigt sich eine weitere Arbeit von Schöch et al. (Schöch et al. 2016), die Topics über den Textverlauf betrachtet und dabei feststellt, dass manche Themengebiete, wie beispielsweise „Schule“, eher am Anfang von Romanen auftreten, während gegen Ende häufiger abstrakte Topics, die beispielsweise auf bestimmte Wertvorstellungen hindeuten, gefunden werden. Hettinger et al. (Hettinger et al. 2016) verwenden Topic Modeling, um Untergattungen von deutschsprachigen Romanen des 19. Jahrhunderts automatisch zu klassifizieren. Eine Besonderheit dieser Arbeit ist, dass sie LDA, Stilometrie und Netzwerkanalyse kombiniert, um so verschiedene Aspekte literarischer Texte abzubilden.

4 Geplante Experimente

Die bis hierhin beschriebenen theoretischen Konzepte und Methoden sollen nun in computergestützten Experimenten auf die dieser Arbeit zugrundeliegende Fragestellung übertragen werden. Daher werden zunächst das verwendete Korpus, sowie die Erstellung der Evaluationsgrundlage, auf deren Basis die durchgeführten Untersuchungen ausgewertet werden, erläutert. Schließlich werden die geplanten Experimente kurz vorgestellt.

4.1 Korpus

Diese Arbeit befasst sich mit Romanen des 19. Jahrhunderts, die in deutscher Sprache verfasst sind. In dem entsprechend zusammengestellten Korpus befinden sich jedoch nicht nur Texte von deutschen Autoren, es können auch Romane von Autoren anderer Nationalitäten, wie beispielsweise Flaubert, enthalten sein, die in einer deutschen Übersetzung vorliegen. Die verwendeten Texte stammen aus der

Digitalen Bibliothek von TextGrid[3] und sind dementsprechend qualitativ hochwertige Digitalisate von Erstdrucken und Studienausgaben.

Das Korpus umfasst 35 Romane, deren Auswahl bestimmten Kriterien unterliegt. Diese recht niedrige Zahl ermöglicht es, einen guten Überblick über die Texte zu behalten sowie die einzelnen Arbeitsschritte detailliert nachvollziehen zu können. Auch die relativ zeitaufwändige händische Aufbereitung der Textdaten, die in Abschnitt 5.1 beschrieben wird, wäre mit einer größeren Textmenge im Rahmen dieser Arbeit nicht mehr zu bewältigen gewesen. Ein weiterer Grund für den überschaubaren Umfang des Korpus ist die Notwendigkeit, sich zur Erstellung der Evaluationsgrundlage, die im nächsten Abschnitt erläutert wird, eine gewisse Grundkenntnis zu jedem Roman anzueignen. Daher wurde als weiteres Auswahlkriterium bei der Korpuszusammenstellung geprüft, ob für den jeweiligen Text eine Zusammenfassung von ausreichender Länge vorliegt. Die Herkunft und Verwendung dieser Zusammenfassungen werden im folgenden Abschnitt beschrieben. Außerdem wurden Briefromane und in der ersten Person erzählte Romane ausgeschlossen, da diese Fälle zusätzliche Schwierigkeiten bei der Erstellung von Figurennetzwerken mit sich bringen können. Abgesehen von diesen Kriterien wurden die Texte für das Korpus willkürlich aus dem Bestand der Digitalen Bibliothek ausgewählt. Eine vollständige Übersicht über die im Korpus enthaltenen Romane befindet sich im Anhang.

4.2 Evaluationsgrundlage

Um computergestützte Experimente auswerten zu können, benötigt man in der Regel eine vorher festgelegte Evaluationsgrundlage. Da es sehr zeitaufwändig und auch nicht im Sinne einer quantitativen Analyse wäre, alle Texte des Korpus zu lesen, wurden Inhaltszusammenfassungen herangezogen, um sich einen Eindruck der Figurenkonstellation der Romane zu verschaffen. Die Zusammenfassungen stammen aus Kindlers Literatur Lexikon Online[4], sind von Experten verfasst worden und größtenteils relativ ausführlich. Zur Evaluation der folgenden Untersu-

[3] https://textgrid.de/digitale-bibliothek.

[4] Arnold, Heinz L. (Hg.) (2009). *Kindlers Literatur Lexikon*. 3. Aufl. Stuttgart/Weimar: Verlag J.B. Metzler. Online verfügbar unter kll-online.de, zuletzt geprüft am 02.08.2016.

chungen wurde die menschliche Intuition über die Ähnlichkeit zwischen den Figurenkonstellationen der im Korpus enthaltenen Romane in Form einer manuell erstellten Distanzmatrix festgehalten. Eine solche Matrix besitzt als Spalten und Zeilen jeweils alle n Romane des Korpus, sodass in den einzelnen Feldern die Abstände zwischen zwei Romanen als numerische Werte eingetragen werden können. Entsprechend ist eine Distanzmatrix symmetrisch und enthält in der Hauptdiagonalen nur Nullen. Aufgrund dieser Eigenschaften müssen nur $n \cdot (n-1)/2$ Felder ausgefüllt werden: Bei 35 Romanen bedeutet das dennoch 595 paarweise Vergleiche, was wiederum den überschaubaren Umfang des Korpus erklärt.

Nach der genauen und wiederholten Lektüre der jeweiligen Zusammenfassungen wurden die Romane entsprechend einer Skala von 0 (= identisch) bis 4 (= unähnlich) hinsichtlich ihrer Ähnlichkeit bewertet. Dabei lag der Fokus natürlich auf der Struktur der Figurenkonstellation und den vorkommenden Relationen und Konflikten zwischen Figuren. Dennoch handelt es sich dabei um eine zeitaufwändige und schwierige Aufgabe, da sich oft sowohl Parallelen als auch Unterschiede zwischen Romanen finden lassen.

Die so erstellte Distanzmatrix bietet die Möglichkeit einer direkten Auswertung der Experimente, im Gegensatz zu einer indirekten Evaluation durch die Klassifikation bestehender Kategorien wie beispielsweise Gattungen. Sie wurde manuell in Excel erstellt und kann unter Verwendung passender Packages gut in Python eingelesen werden.

4.3 Durchführung

Im folgenden praktischen Teil dieser Arbeit werden die Textdaten zunächst aufbereitet, um daraus mit Hilfe einer bestimmten Modellierung automatisch Figurennetzwerke zu extrahieren. Dabei soll auch genauer beleuchtet werden, wie gut ein automatisch erstelltes Netzwerk die Figurenkonstellation eines Romans widerspiegelt. Zu diesen Netzwerken werden unter Verwendung von Methoden aus der Sozialen Netzwerkanalyse verschiedene Eigenschaften erhoben. Auf Basis dieser Features können Distanzen zwischen den Netzwerken berechnet werden, die wiederum gegen die manuell erstellte Distanzmatrix evaluiert werden. Mit Hilfe von

Topic Modeling soll versucht werden, Informationen zu in den Romanen enthaltenen Motiven und Beziehungsarten zu erheben, um diese zusätzlich zur Distanzberechnung heranzuziehen bzw. direkt in die Figurennetzwerke einzubinden.

Zur Durchführung der Experimente wurde Python in Verbindung mit verschiedenen Packages verwendet und, wo erforderlich, weitere Open-Source-Software herangezogen.

5 Figurennetzwerke

5.1 Datenaufbereitung

Der erste Schritt auf dem Weg zu einer automatischen Erstellung von Figurennetzwerken ist die Erkennung aller Vorkommen von Figuren im Text. Es gibt zahlreiche Systeme zur Namenserkennung (Named Entity Recognition, kurz NER), wie zum Beispiel Stanford NER[5], die Personennamen in Texten finden. Da diese Werkzeuge üblicherweise auf Zeitungskorpora trainiert wurden, markieren sie jedoch nur Personennennungen mittels konkreter Namen. In literarischen Texten können Figuren jedoch oft auch durch sogenannte Appellative, beispielsweise Adelstitel wie ‚Baron' oder Berufsbezeichnungen wie ‚Gärtner', referenziert werden, die ein solches System nicht erkennen kann. Jannidis et al. (Jannidis et al. 2015) haben dafür ein System entwickelt, das auf einem manuell annotierten Goldstandard von Auszügen aus deutschen Romanen des 19. Jahrhunderts trainiert und somit optimal für diese erweiterte Definition von Named Entity Recognition geeignet ist. Um konkrete Namen und Appellative auch noch im Nachhinein voneinander unterscheiden zu können, verwendet das System die Tags ‚B-PER' und ‚I-PER' jeweils mit den Zusätzen ‚Core' bzw. ‚App', sowie ‚Pron' für pronominale Referenzen.

Anschließend sollte eine Koreferenzauflösung (Coreference Resolution, CR) erfolgen, also festgestellt werden, welche Referenzen sich auf die gleichen Figuren beziehen. In weiterführender Arbeit haben Krug et al. (Krug et al. 2015) auch hierfür ein Werkzeug entwickelt, das auf die Domäne literarischer Texte angepasst

[5] http://nlp.stanford.edu/software/CRF-NER.shtml.

ist und im Vergleich zu anderen State-of-the-Art Systemen zur Koreferenzauflösung wie CorZu[6], die typischerweise ebenfalls anhand von Zeitungsdaten entwickelt wurden, bessere Ergebnisse liefert.

Zusammen mit einigen weiteren Natural-Language-Processing-Komponenten wie Tokenisierung, Satz- und Absatzerkennung oder Part-of-Speech-Tagging sind die beiden genannten Werkzeuge in ein von Markus Krug erstelltes Kommandozeilenprogramm[7] integriert, das sich über Konfigurationsdateien an die Bedürfnisse des Benutzers anpassen lässt. Auf diese Weise wurden alle Texte des Korpus prozessiert.

Da Koreferenzauflösung jedoch, insbesondere auf Texten von Romanlänge, ein schwieriger Task ist, macht leider auch das hier verwendete, auf die Domäne angepasste System einige Fehler. Schwierigkeiten treten beispielsweise dann auf, wenn mehrere Figuren den gleichen Familiennamen oder sogar den gleichen Vornamen tragen, oder sich der Name oder Titel durch eine Hochzeit, Beförderung oder ähnliches ändert. Beispiele dafür sind die Namen ‚Briest' und ‚Innstetten' in *Effi Briest* oder zwei Figuren in Sudermanns *Frau Sorge*, die beide ‚Elsbeth' heißen. Auch wiederkehrende Adelstitel oder Berufsbezeichnungen können zu Problemen führen: Kommen in einem Roman mehrere verschiedene Figuren vor, die den Titel ‚Baron' tragen, so kann es passieren, dass sie als eine Figur zusammengefasst werden. Auch umgekehrt kann es vorkommen, dass zwei oder mehr Gruppen von Referenzen, die sich eigentlich auf die gleiche Figur beziehen, nicht kombiniert werden.

Um sicherzustellen, dass die verwendete Datengrundlage für diese Arbeit dennoch von möglichst hoher Qualität ist, wurde die Ausgabe der Koreferenzauflösung manuell nachkorrigiert. Natürlich kann nicht jede einzelne Referenz überprüft werden, es wurde jedoch für die am häufigsten vorkommenden Figuren überprüft, dass keine gravierenden Fehlzuweisungen bestehen bleiben. Dazu wurde ein ebenfalls von Markus Krug entwickelter Editor verwendet, der neben diverser anderer Funktionen einerseits einen Überblick über alle im Roman vorkommenden

---

6 http://www.cl.uzh.ch/de/research/completed-research/coreferenceresolution.html.

7 Das Programm ist zum aktuellen Zeitpunkt nicht veröffentlicht, wurde aber zur Verwendung im Rahmen dieser Arbeit zur Verfügung gestellt.

Figuren und andererseits eine Korrektur von Fehlern in der Koreferenzauflösung ermöglicht.

Da die verwendete Pipeline intern mit UIMA arbeitet und entsprechend das XMI-Format verwendet, wurden die Daten anschließend in ein anderes Format transformiert. Dabei handelt es sich um eine Tab-separierte, tabellarische Auflistung, bei der jede Zeile ein Wort und die dazugehörigen Informationen wie POS-Tag, NE-Tag und Koreferenz-ID in mehreren Spalten enthält (Abbildung 3). Neben der intuitiven und übersichtlichen Lesbarkeit für den Menschen, hat dieses Format den Vorteil, dass es sich recht komfortabel mit Skriptsprachen wie Python verarbeiten lässt.

SectionId	ParagraphId	SentenceId	TokenId	Begin	End	Token	Lemma	CPOS	POS	Morphology	DependencyHea	Dependencyf	NamedEntity	CorefId
null	0	7	162	2908	2910	1.	1.	ADJA	ADJA	masc|sg|	Kapitel	NK	O	-
null	0	7	163	2911	2918	Kapitel	Kapitel	N	NN	masc|sg|	Erstes	MO	O	-
0	1	7	164	3115	3121	Erstes	erst	ADJA	ADJA	masc|sg|	ROOT	--	O	-
0	1	7	165	3122	3129	Kapitel	Kapitel	N	NN	masc|sg|	Erstes	SB	O	-
0	2	7	166	3161	3163	In	in	APPR	APPR	||	fiel	MO	O	-
0	2	7	167	3164	3169	Front	Front	N	NN	||	In	NK	O	-
0	2	7	168	3170	3173	des	die	ART	ART	masc|sg|	Herrenhauses	NK	O	-
0	2	7	169	3174	3179	schon	schon	ADV	ADV	||	seit	MO	O	-
0	2	7	170	3180	3184	seit	seit	APPR	APPR	||	bewohnten	MO	O	-
0	2	7	171	3185	3193	Kurfürst	Kurfürst	N	NN	masc|sg|	seit	NK	B-PER_CORE	589
0	2	7	172	3194	3199	Georg	Georg	N	NE	masc|sg|	Wilhelm	PNC	I-PER_CORE	589
0	2	7	173	3200	3207	Wilhelm	Wilhelm	N	NE	masc|sg|	Kurfürst	NK	I-PER_CORE	589
0	2	7	174	3208	3211	von	von	APPR	APPR	||	bewohnten	SBP	O	-
0	2	7	175	3212	3215	der	die	ART	ART	fem|sg|	Familie	NK	O	-
0	2	7	176	3216	3223	Familie	Familie	N	NN	fem|sg|	von	NK	B-PER_CORE	67
0	2	7	177	3224	3227	von	von	APPR	APPR	||	Familie	PG	O	-
0	2	7	178	3228	3234	Briest	<unknown>	N	NN	||	von	NK	O	-
0	2	7	179	3235	3244	bewohnten	bewohnt	ADJA	ADJA	neut|sg|	Herrenhauses	NK	O	-
0	2	7	180	3245	3257	Herrenhauses	Herrenhaus	N	NN	neut|sg|	Front	AG	O	-

3 Ausschnitt aus tabellarischer Darstellung zu Effi Briest

5.2 Modellierung und Netzwerkerstellung

5.2.1 Erstellung von Interaktionslisten

Um auf Basis der aufbereiteten Textdaten Figurennetzwerke zu erstellen, ist noch eine konzeptionelle Entscheidung nötig: Auf welche Weise sollen Interaktionen zwischen Figuren modelliert werden, also wann existiert eine Kante zwischen zwei Figuren? Wie Abschnitt 3.1.3 gezeigt hat, gibt es dafür mehrere Ansätze. In dieser Arbeit wird eine Interaktion angenommen, sobald zwei Figuren gemeinsam in einem Absatz genannt werden. Die Informationen zu Absatzgrenzen sind aus dem tabellarischen Dateiformat durch eine fortlaufende Zählung ersichtlich. Diese Operationalisierung wurde unter anderem gewählt, da vorhergehende Arbeiten in Bezug auf die Identifikation der Hauptfiguren eines Romans keinen Vorteil in der Verwendung von komplizierteren, auf Dialogstrukturen beruhenden Figurennetzwerken erkennen konnten (Krug et al. 2016b). Da in Romanen zudem ein Großteil

der Informationen über Erzählerrede vermittelt wird und manche Romane nur wenig direkte Rede enthalten, sind Kookkurrenz-Netzwerke möglicherweise sogar besser geeignet, um diese Textsorte zu repräsentieren (Coll Ardanuy und Sporleder 2014). Charmant ist auch die relativ einfache Programmierbarkeit dieses Ansatzes. Im resultierenden Netzwerk wird also jede Figur durch einen Knoten und jedes gemeinsame Vorkommen zweier Figuren in einem Absatz durch eine Kante repräsentiert. Als Kantengewicht wird die Information verwendet, in wie vielen Absätzen dies der Fall ist. Darüber hinaus liefert diese Vorgehensweise jedoch keine weiteren Informationen zu den Kanten, wie beispielsweise die Art der Relation zwischen zwei Figuren.

Zum Erstellen der Interaktionsdaten wird die Tab-separierte Datei eines Romans zunächst unter Verwendung des Python-Pakets Pandas[8] als DataFrame eingelesen, um die `groupby`-Funktionen dieser Datenstruktur nutzen zu können. Mit deren Hilfe werden die Daten in eine interne Datenstruktur überführt, die aus verschachtelten Listen auf den Ebenen von Kapiteln, Absätzen und Wörtern besteht. Die einzelnen Wörter werden als Dictionaries repräsentiert, wobei nur die tatsächlich benötigten Informationen mitgeführt werden, nämlich das Wort an sich, das Named-Entity-Tag des Wortes und die Koreferenz-ID. Anschließend geht das Skript alle Absätze durch und ermittelt die enthaltenen Relationen, nämlich alle möglichen Zweierkombinationen zwischen allen im Absatz genannten Figuren. Außerdem wird gezählt, in wie vielen Absätzen die jeweiligen Relationen auftreten.

Für all diese Schritte wird die Koreferenz-ID herangezogen, da sie eine Figur eindeutig identifiziert. Im Hinblick auf eine spätere Visualisierung der Figurennetzwerke ist es allerdings wünschenswert, jede Figur zusätzlich durch einen Namen bezeichnen zu können. Jeder Koreferenz-ID ist jedoch eine ganze Reihe von Nennungen dieser Figur zugeordnet. Daher muss daraus ein Anzeigename ausgewählt werden, der für den Betrachter möglichst anschaulich deutlich macht, welche Figur sich hinter der jeweiligen ID verbirgt. Hier sind mehrere Varianten denkbar: Es könnte beispielsweise die kürzeste, als ‚Core' getaggte Referenz (also ein konkreter Name) gewählt werden. Bei wiederkehrenden Familiennamen kann dies jedoch dazu führen, dass mehrere Figuren den gleichen Anzeigenamen erhalten. In verschiedenen Versuchen stellte sich heraus, dass die sinnvollsten Anzeigenamen

[8] http://pandas.pydata.org.

entstehen, wenn diejenige Referenz gewählt wird, die am häufigsten vorkommt, wobei pronominale Referenzen natürlich ausgeschlossen werden müssen.

Da nun alle nötigen Informationen zum Aufbau eines Figurennetzwerks ermittelt wurden, können diese als Tabstopp-getrennte Liste gespeichert werden, wie Abbildung 4 zeigt. Diese Interaktionslisten enthalten für jede Figur sowohl die Koreferenz-ID, als auch den Anzeigenamen und die entsprechenden Kantengewichte zwischen zwei Figuren. Für eine bessere Übersichtlichkeit wurden die Knoten jeweils so angeordnet, dass der im Alphabet vorangehende Name zuerst genannt wird, und die gesamte Liste nach Kantengewichten absteigend sortiert.

Effi_569	Innstetten_779	286
Effi_569	Mama_880	135
Effi_569	Roswitha_860	95
Effi_569	Gieshübler_598	94
Crampas_445	Effi_569	85
Crampas_445	Innstetten_779	59
Innstetten_779	Mama_880	58
Gieshübler_598	Innstetten_779	56

4 Ausschnitt aus Interaktionsliste zu Effi Briest

Auf diese Weise muss die Extraktion der Daten nur einmal durchgeführt werden. Außerdem ermöglicht es die Zwischenspeicherung, die Daten in anderen Anwendungen, wie zum Beispiel zur Visualisierung in Gephi[9] oder zur Weiterverarbeitung in anderen Skriptsprachen wie R, zu nutzen. Auf Basis dieser Interaktionslisten können im Folgenden Figurennetzwerke erstellt und analysiert werden.

5.2.2 Netzwerkerstellung und -visualisierung

Zur Arbeit mit Netzwerken in Python wird die Bibliothek NetworkX[10] verwendet, die es erlaubt, mathematische Graphen zu repräsentieren, die Knoten und Kanten zu verwalten, sowie diesen Labels oder Eigenschaften wie Gewichte zuzuweisen. Außerdem stellt sie zahlreiche Funktionen für typische Berechnungen auf

[9] https://gephi.org.

[10] https://networkx.github.io.

Graphen oder zur Visualisierung bereit. Die Daten können direkt aus der Interaktionsliste eingelesen werden. Dabei muss lediglich angegeben werden, welche Kanteneigenschaft in der dritten Spalte steht und welchen Datentyp diese hat, da NetworkX auch weitere Spalten mit zusätzlichen Attributen erlaubt. Die entsprechenden Knoten und Kanten werden dann automatisch generiert.

Ein auf diese Weise entstandenes Figurennetzwerk kann je nach Roman aus mehreren Hundert Knoten bestehen, da auch unwichtige Figuren, die nur sehr selten auftreten, enthalten sind. Um Figurennetzwerke als Modell für die Figurenkonstellation von Romanen sinnvoll analysieren und visualisieren zu können und völlig überladene Darstellungen zu vermeiden, muss die Knotenzahl reduziert werden.

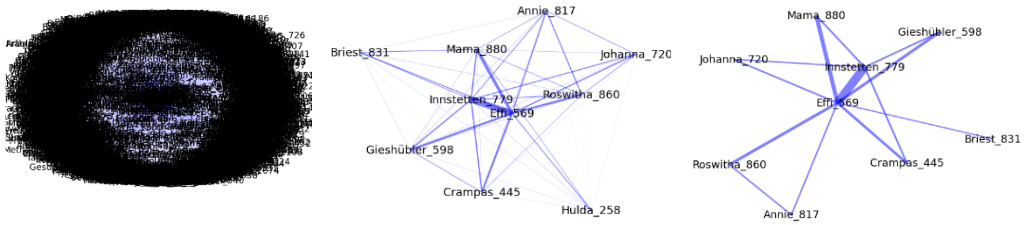
Zunächst wurden sehr generische Knoten wie ‚Menschen', ‚Leute' oder ‚Gesellschaft' mit Hilfe einer Liste herausgefiltert, ebenso wie Knoten mit Bezeichnern wie ‚Gott' oder ‚Person'. Diese werden von den Algorithmen zur Erkennung der Figurenreferenzen und Koreferenzauflösung wie alle anderen Figuren auch behandelt und erscheinen dann als relativ wichtig im Figurennetzwerk, obwohl sie inhaltlich wenig beitragen. Die Filterung wirkt diesem Effekt entgegen und wird zuerst durchgeführt, damit andere, inhaltlich sinnvollere Figuren ins Netzwerk „nachrücken" können.

Anschließend wurde ein sogenannter Knotenfilter eingeführt, bei dem die Knoten absteigend nach ihrem gewichteten Knotengrad (*weighted node degree*) sortiert werden und dann nur eine bestimmte Anzahl vom Beginn dieser Liste in die Visualisierung aufgenommen wird. Der gewichtete Knotengrad (auch *node strength* genannt) wird ermittelt, indem die Gewichte aller an dem Knoten anliegenden Kanten summiert werden (Costa et al. 2007, S. 174). Je mehr Kanten ein Knoten hat und je höher deren Gewicht ist, desto höher ist also dieser Wert und desto wichtiger der jeweilige Knoten. Beim tatsächlichen Sortieren in Python ist hier Vorsicht geboten, da NetworkX die Knoten und den dazu berechneten gewichteten Grad als Dictionary zurückgibt, in dem die Elemente folglich keine feste Reihenfolge besitzen. Es kann also passieren, dass bei einem zweiten Durchlauf des Skripts die Sortierreihenfolge bei Knoten mit gleichem Grad abweicht. Daher muss unbedingt sowohl nach Alphabet, als auch nach gewichtetem Knotengrad sortiert werden.

Nach Anwendung dieses Filters erhält man ein Netzwerk, das sehr stark verbunden ist, aber viele Kanten mit geringem Gewicht enthält. Daher liegt es nahe,

auch die Kanten auf die wichtigsten zu beschränken. Dazu wird ein bestimmter Prozentsatz der Kanten mit dem niedrigsten Gewicht herausgefiltert und alle isolierten Knoten, die danach ohne Kanten verbleiben, ebenfalls verworfen. Alternativen zu diesem Ansatz wären, nur Kanten ab einem bestimmten Mindestgewicht oder nur eine bestimmte Anzahl von Interaktionen mit hohem Gewicht vom Beginn der Interaktionsliste zu berücksichtigen. In Versuchen hat sich jedoch gezeigt, dass eine Begrenzung wie oben beschrieben auf die zehn wichtigsten Knoten und eine Filterung von 30% der schwächsten Kanten anschauliche Visualisierungen liefert, bei denen die wichtigen Hauptfiguren im Netzwerk verbleiben und dessen Grundstruktur gut hervortritt. Der empirische Eindruck, dass nach der Filterung vor allem Hauptfiguren im Figurennetzwerk enthalten sind, lässt sich weiter untermauern: Eine naheliegende Annahme ist, dass die Hauptfiguren in einem Roman auch am häufigsten vorkommen bzw. dass häufig vorkommende Figuren wichtig für die Figurenkonstellation sind. Betrachtet man die in den Netzwerken enthaltenen Figuren im Hinblick auf diese These, so fallen diese durchschnittlich unter die 15 häufigsten Figuren des Romans, wobei durchschnittlich nur 1,14 Knoten nicht aus den zehn häufigsten Figuren stammen. Diese Zahlen unterstützen den Eindruck, dass die Filterung ihren Zweck erfüllt.

Für die Darstellung wurde das ‚spring'-Layout aus NetworkX verwendet, welches den Fruchterman-Reingold-Algorithmus nutzt, um die Knoten auf der Bildfläche anzuordnen. Dabei werden anziehende und abstoßende Kräfte zwischen den Knoten angenommen und diese so platziert, dass sie den größtmöglichen Abstand zueinander haben und gleichmäßig verteilt sind. Außerdem wird versucht, Überschneidungen von Kanten soweit möglich zu vermeiden (Pfeffer 2010, S. 231). Abbildung 5 zeigt das Figurennetzwerk zu *Effi Briest* im ungefilterten Zustand, sowie nach Anwendung beider Filter. Bei den Zahlen hinter den Namen handelt es sich um die entsprechende Koreferenz-ID der Figur, die zur Unterscheidung dient, falls mehrere Figuren den gleichen Anzeigenamen haben.



5 Verschiedene Filterstufen zu Effi Briest: kein Filter (links), nur Knotenfilter (Mitte), Knoten- und Kantenfilter (rechts)

Die Grafik zeigt, dass die Filterung die Struktur des Netzwerks für den Betrachter viel deutlicher macht und für einen besseren Überblick sorgt. Bei der Filterung werden zwar Informationen entfernt, diese lassen sich jedoch auf Figuren zurückführen, die nur sehr selten im Roman vorkommen und daher für eine Repräsentation der Figurenkonstellation nicht relevant sind. Außerdem sorgt die Filterung dafür, dass das Beziehungsgeflecht zwischen den verbleibenden Figuren besser hervortritt. Dies verstärkt den Eindruck, dass diese Filterung auch für die automatische Analyse der Figurennetzwerke nur von Vorteil sein kann.

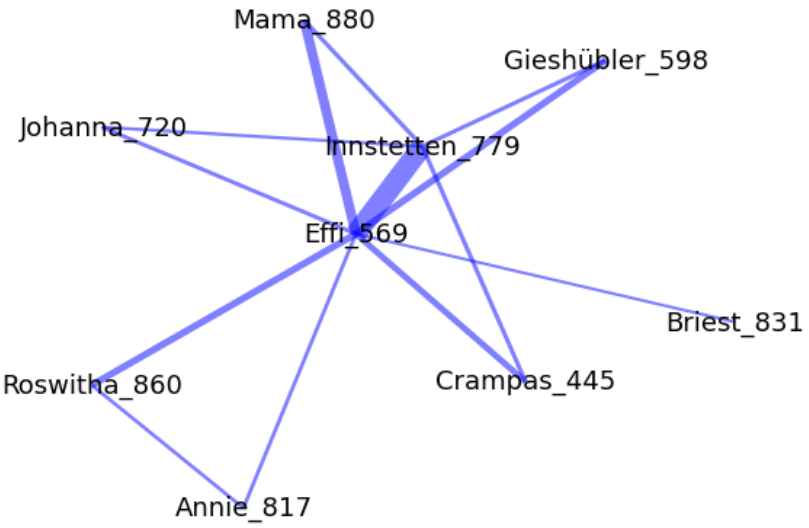
5.3 Vergleich der Netzwerke mit Zusammenfassungen

Im Folgenden werden die generierten Figurennetzwerke für *Effi Briest* und *Madame Bovary* im Detail betrachtet und mit der in Abschnitt 2 vorgestellten jeweiligen Figurenkonstellation in Bezug gesetzt, um zu sehen, wie gut ein solches Figurennetzwerk einen Roman repräsentieren kann. Außerdem werden Figurennetzwerke näher betrachtet, deren Visualisierung von der anhand der Zusammenfassung erwarteten Figurenkonstellation abweichen, um eventuelle Probleme oder Besonderheiten bei der Netzwerkerstellung zu identifizieren.

5.3.1 Figurennetzwerke zu Effi Briest und Madame Bovary

Betrachtet man das Netzwerk zu *Effi Briest* (Abbildung 6), so stellt man fest, dass alle in der schematischen Figurenkonstellation (Abbildung 1) dargestellten Figuren auch im Netzwerk zu finden sind. Zudem liegt der Fokus darauf, die Grundstruktur der Figurenkonstellation zu erfassen, die sich am Netzwerk recht deutlich erkennen lässt. Effi steht im Zentrum und ist der Knoten, an dem die meisten Kanten anliegen. Die zentrale Relation zwischen ihr und Innstetten ist stark ausgeprägt. Auch das Arrangement zwischen Effi als Mutter und Roswitha als Kindermädchen im Hinblick auf Tochter Annie lässt sich wiederfinden. Die Kante zwischen Effi

und Crampas fällt nicht ganz so stark aus, wie vielleicht zu erwarten wäre, allerdings ist ihre Beziehung im Roman auch nur von kurzer Dauer mit relativ wenig direkter Interaktion und wird viel in Andeutungen erzählt. Zwischen Innstetten und Annie, sowie ihm und Roswitha, sind keine Kanten realisiert. Dies könnte sich darauf zurückführen lassen, dass Roswitha hauptsächlich Effis Angestellte und Vertraute ist und weniger mit Innstetten interagiert. Annie ist zwar Innstettens Tochter, aber da der Roman auf die Protagonistin Effi fokussiert ist, die ihrer Tochter nur phasenweise Aufmerksamkeit schenkt und Innstetten viel auf Reisen ist, ist es nachvollziehbar, dass Innstetten und Annie seltener gemeinsam genannt werden. Obwohl ein Leser des Romans sich natürlich der Verwandtschaftsbeziehung zwischen den beiden Figuren bewusst ist, kann diese Information keineswegs trivial mit computergestützten Methoden erhoben werden und ist somit bei der Netzwerkerstellung nicht verfügbar.

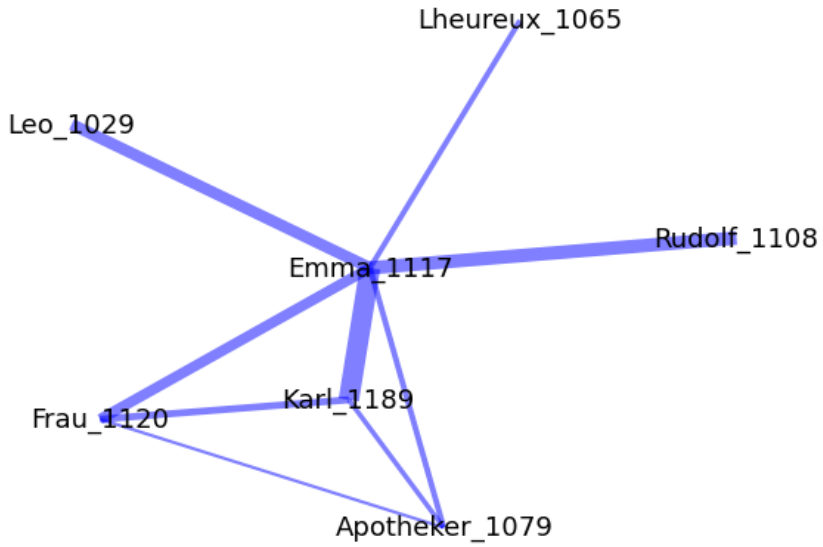


6 Figurennetzwerk zu Effi Briest

Am Figurennetzwerk zu *Madame Bovary*[11] lassen sich ähnliche Effekte beobachten (Abbildung 7). Auch hier steht Emma im Zentrum und ist durch eine starke Kante mit ihrem Ehemann Karl verbunden. Mit ihren beiden Liebhabern Leo und Rudolf verbinden sie relativ starke Relationen. Dienstmädchen Felicie und Tochter Bertha tauchen nicht im Netzwerk auf. Das ist erklärbar, da Felicie eher eine Nebenfigur ist und Bertha hauptsächlich bei einer Amme aufwächst, wodurch

[11] Bei dem Knoten „Frau_1120“ handelt es sich um die alte Frau Bovary, Karls Mutter.

die Interaktionen zwischen ihr und ihren Eltern eher gering bleiben. Außerdem hängt das „Fehlen“ von Figuren natürlich mit den angewendeten Filtern zusammen.



7 Figurennetzwerk zu Madame Bovary

In jedem Fall lässt sich bei Kenntnis der Romane sagen, dass keiner der Knoten und keine der Kanten, die in den Netzwerken enthalten sind, falsch oder nicht nachvollziehbar sind. Die hier beschriebenen Unterschiede zwischen den Figurennetzwerken und der manuell erstellten Übersicht über die Figurenkonstellation lassen sich im Wesentlichen auf zwei Punkte zurückführen. Zum ersten muss man sich bewusst sein, dass bei der manuellen Analyse der Figurenkonstellation Informationen über feststehende Beziehungen zwischen Figuren, wie zum Beispiel Verwandtschaft, einfließen, die der Leser während der Lektüre eines Romans erhält. Derartige grundlegende Relationen werden im Roman häufig nur einmal eingeführt und manifestieren sich, wie im Fall der Beziehung zwischen Innstetten und seiner Tochter Annie, nicht notwendigerweise in häufiger Interaktion zwischen den entsprechenden Figuren. Diese Interaktionen, modelliert über gemeinsames Vorkommen von Figuren, sind jedoch die einzigen Daten, die zur automatischen Erstellung von Figurennetzwerken genutzt werden können. Hier liegen also die Grenzen dessen, was ein solches Netzwerk darstellen kann. Entsprechend können die Kanten des Netzwerks auch nicht weiter in Kategorien wie Ehe, Freundschaft oder Familie unterteilt werden. Das Netzwerk erfasst lediglich, zwischen welchen Figuren eine Relation besteht und wie häufig diese im Roman vorkommt. Obwohl es

theoretisch denkbar wäre, Informationen über Familienbeziehungen und andere soziale Verhältnisse durch Textmining-Verfahren in Romanen zu ermitteln, ist dies jedoch keineswegs eine einfache Aufgabe. Zum Zeitpunkt dieser Thesis sind keine Arbeiten bekannt, in denen Kanten in Figurennetzwerken automatisch hinsichtlich verschiedener Beziehungsarten klassifiziert wurden.

Der zweite Punkt ist die Filterung des Netzwerks. Je nachdem, auf welche Werte die Knoten- und Kantenfilter gesetzt werden, kann es vorkommen, dass ein Leser bei der Betrachtung eines Figurennetzwerks manche Figuren als fehlend oder überflüssig empfindet. Andererseits sind die Filter notwendig, um ein übersichtliches Netzwerk darzustellen und dessen Struktur überblicken zu können. Leider ist keine Vorgehensweise bekannt, einen solchen automatischen Filter anders als durch empirisches Ausprobieren und Betrachtung der Visualisierungen für das komplette Korpus festzulegen. Auch eine individuelle Berechnung der Filter-Werte pro Roman könnte hilfreich sein. Andererseits entscheidet auch bei der manuellen Erstellung einer Figurenkonstellation der Leser, welche Figuren aufgenommen werden sollen und lässt andere weg, die ihm weniger wichtig erscheinen und wendet damit gewissermaßen einen Filter an.

5.3.2 Besonderheiten bei Figurennetzwerken

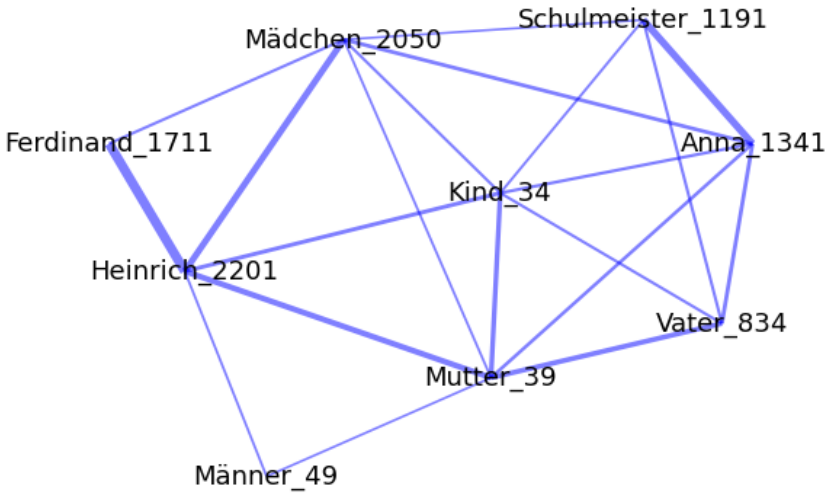
Neben diesen Beispielen wurde für alle Romane des Korpus betrachtet, ob es Netzwerke gibt, die wesentlich von der anhand der Zusammenfassung erwarteten Figurenkonstellation abweichen. Dabei fielen insbesondere zwei Texte auf.

Der erste ist *Der grüne Heinrich* von Gottfried Keller[12], in dem der Lebensweg des Künstlers Heinrich Lee geschildert wird, welcher von seinen Reisen und Begegnungen mit anderen Figuren geprägt ist[13]. Anhand der Zusammenfassung wäre ein tendenziell sternförmiges Netzwerk zu erwarten, dessen Zentrum Heinrich bildet, um den die anderen Figuren angeordnet sind. Tatsächlich ist das generierte Netzwerk zu diesem Roman jedoch relativ stark verbunden, wobei an allen Knoten eine ähnliche Zahl von Kanten anliegt und alle Kanten ein vergleichbares Gewicht haben (Abbildung 8).

---

[12] Erste Fassung von 1854.

[13] Alle in diesem Abschnitt enthaltenen Inhaltszusammenfassungen zu Romanen basieren auf den entsprechenden Artikeln aus Kindlers Literatur Lexikon Online.



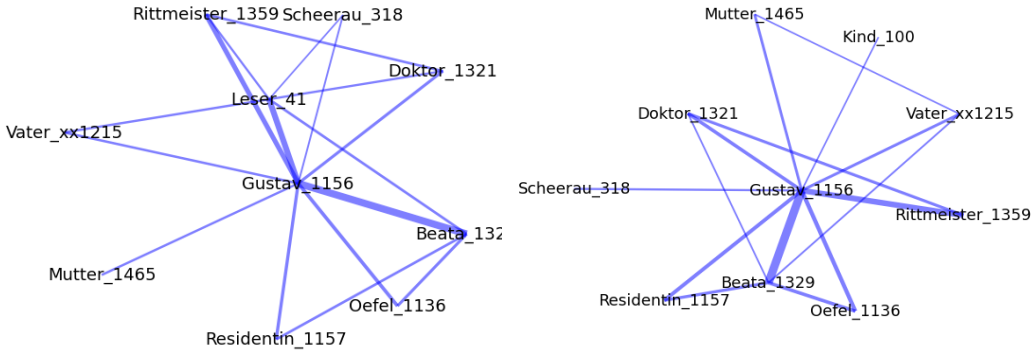
8 Figurennetzwerk zu Der grüne Heinrich

Bei näherer Betrachtung des Romans zeigt sich auch der Grund für diesen Effekt: Obwohl der Roman hauptsächlich heterodiegetisch erzählt ist, wird ein längerer Teil als Rückblick auf Heinrichs Kindheit und Jugendalter von ihm selbst aus der Ich-Perspektive geschildert. Dies führt zu besonders vielen Schwierigkeiten bei der Koreferenzauflösung, da die Figuren aus den beiden Teilen aufeinander bezogen werden müssen. Das ist insbesondere bei den vielen Referenzen auf Heinrich Lee in der ersten Person eine besondere Herausforderung, der der Algorithmus zur Koreferenzauflösung noch nicht gewachsen ist. Auch eine manuelle Korrektur wäre hier zu aufwändig gewesen. Es bleibt also festzuhalten, dass längere eingeschobene Passagen, die in einer anderen Erzählperspektive verfasst sind, möglicherweise zu Problemen bei der automatischen Erstellung von Figurennetzwerken führen können. Solche Einschübe könnten zum Beispiel längere Briefe, Tagebucheinträge oder Binnenerzählungen sein. Letztere sind ein Sonderfall, da sie andere Figuren enthalten, als in der Haupthandlung vorkommen. Für die Zukunft wäre es nützlich, solche Passagen automatisch zu erkennen, um sie dann gegebenenfalls anders prozessieren zu können.

Der zweite auffällige Fall ist Jean Pauls *Die unsichtbare Loge*. Auch hier wäre ein weitgehend sternförmiges Netzwerk zu erwarten, da der Roman die Lebensgeschichte von Gustav von Falkenberg erzählt, der zunächst abgeschieden von der Gesellschaft aufwächst und erzogen wird, und später bei Hofe lebt und einer

Geheimgesellschaft beitritt. Im generierten Figurennetzwerk (Abbildung 9) ist jedoch neben Gustav noch ein weiterer zentraler Knoten zu finden, der überraschenderweise die Bezeichnung ‚Leser' trägt.

Dies hängt mit der stilistischen Gestaltung des Romans zusammen: Im Text wird wiederholt der Leser direkt vom Erzähler angesprochen. Diese Vorkommen wurden vom NER-System fehlerhaft als Figur gekennzeichnet, obwohl - wenn überhaupt - die Auszeichnung als Pseudofigur[14] legitim wäre. Wenn ein solcher Fall jedoch nicht oder nur sehr selten in den NER-Trainingsdaten auftritt, kann die korrekte Auszeichnung auch nicht vom System erwartet werden. In der Konsequenz werden in der Koreferenzauflösung auch Pronomen und andere Referenzen fehlerhaft auf die Figur ‚Leser' aufgelöst. Auch dies ist ein Problem, das potentiell in anderen Romanen wiederkehren könnte. Im vorliegenden Fall wurden alle Vorkommnisse von ‚Leser' und die dazugehörigen Kanten aus der Interaktionsliste gestrichen, um eine bessere Repräsentation für diesen Roman zu erhalten.

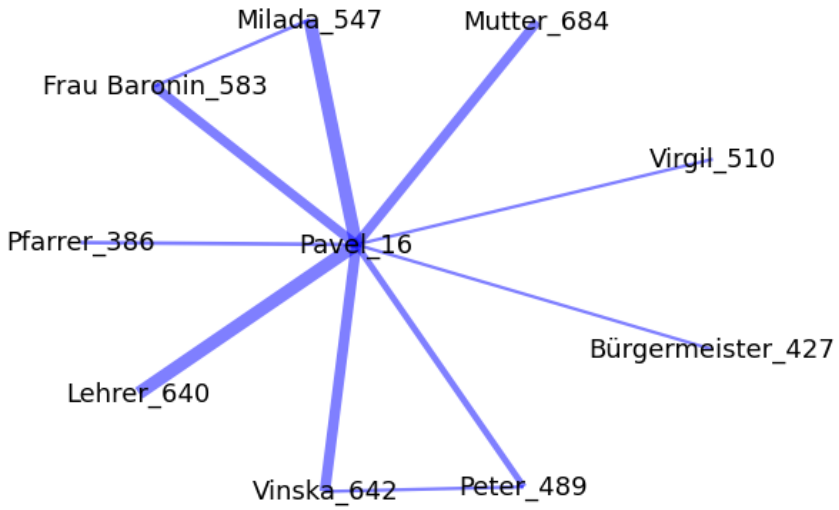


9 Figurennetzwerke zu Die unsichtbare Loge mit (links) und ohne ‚Leser' (rechts)

Darüber hinaus kann es vorkommen, dass zwar die Figuren wie anhand der Zusammenfassungen angenommen im Netzwerk enthalten sind, aber manche Kanten deutlich stärker oder auch deutlich schwächer ausgeprägt sind, als erwartet, und damit auch manche Figuren entsprechend zentraler oder weniger zentral erscheinen. Als Beispiel sei hier *Das Gemeindekind* von Marie von Ebner-Eschenbach genannt. Der Roman handelt von einem Geschwisterpaar, Milada und Pavel, die ohne Eltern aufwachsen. Während Milada von einer Gutsbesitzerin aufgenommen wird,

[14] Im Annotationsschema des NER-Systems von Jannidis et al. (2015) ist eine Pseudofigur eine Figur, die zwar im Text genannt wird, aber eigentlich keine Figur der erzählten Welt ist. Weitere Beispiele wären die Nennung von Schriftstellernamen oder mythologischen Figuren.

gerät Pavel auf die schiefe Bahn und findet erst Jahre später mit Hilfe seiner Schwester zu einem rechtschaffenen Leben zurück. Dies lässt vermuten, dass beide Geschwister gleichrangige Hauptfiguren des Romans und für die Figurenkonstellation von ähnlich hoher Bedeutung sind. Das generierte Figurennetzwerk ist allerdings klar sternförmig um Pavel aufgebaut und deutet darauf hin, dass die Handlung doch stärker auf Pavel konzentriert ist (Abbildung 10).



10 Figurennetzwerk zu Das Gemeindekind

Ebenso kann es vorkommen, dass Relationen im Netzwerk deutlich stärker ausgeprägt sind als erwartet oder dass Relationen im Figurennetzwerk enthalten sind, die in der Zusammenfassung nicht genannt wurden. Ob nun der durch die Zusammenfassung oder der durch das Netzwerk vermittelte Eindruck den Roman besser widerspiegelt, kann in solchen Fällen nur durch eine ausführliche Recherche oder schließlich die Lektüre des Romans geklärt werden.[15] Daher kann dieses Phänomen nicht als methodische Schwäche dieser Art der Netzwerkerstellung gewertet werden, sondern deutet vielmehr auf ein interessantes Verhältnis zwischen einem Roman und verschiedenen Zusammenfassungen desselben hin.

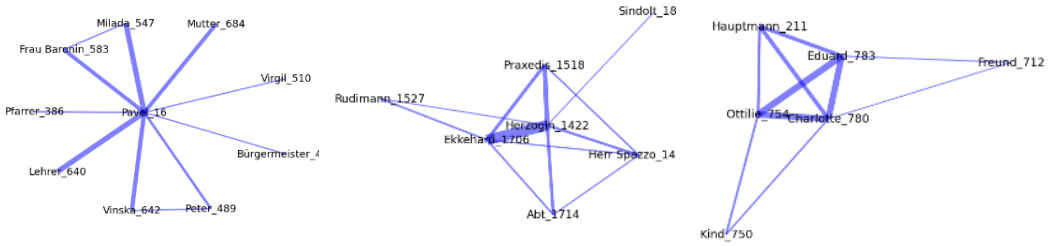
[15] Im Wikipedia-Artikel zu *Das Gemeindekind* steht Pavel deutlich stärker im Fokus und seine Entwicklung wird als Hauptstrang des Romans bezeichnet (https://de.wikipedia.org/wiki/Das_Gemeindekind). Das zeigt, dass verschiedene Zusammenfassungen durchaus unterschiedliche Eindrücke der Figurenkonstellation vermitteln können.

Insgesamt lässt sich feststellen, dass die automatisch generierten Figurennetzwerke bis auf vereinzelte Ausnahmen nicht nennenswert von der Figurenkonstellation, wie sie anhand der Zusammenfassungen eingeschätzt werden kann, abweichen. Am Beispiel von *Effi Briest* und *Madame Bovary* wurde gezeigt, dass die Figurennetzwerke die Figurenkonstellation der Romane hier sehr gut widerspiegeln. Es kann also davon ausgegangen werden, dass die vorgestellte Methode brauchbare modellhafte Repräsentationen der Figurenkonstellation von Romanen erzeugt.

5.4 Netzwerkfeatures

Um die generierten Figurennetzwerke automatisch analysieren und vergleichen zu können, müssen diese durch Merkmale, sogenannte Features, beschrieben werden, die dann als Ausgangspunkt für weitere Berechnungen dienen können. In der Graphentheorie und der Sozialen Netzwerkanalyse gibt es eine ganze Reihe von Netzwerkmaßen und -kennzahlen, die sich in verschiedene Typen unterteilen lassen: Manche Maße beziehen sich auf einzelne Knoten oder Kanten, wie zum Beispiel Zentralitätsmaße, die angeben, wie wichtig diese Bestandteile für das Netzwerk sind. Andere beschreiben einen Graphen als Ganzes und befassen sich mit der Verbundenheit der Knoten untereinander und der damit einhergehenden Gruppenbildung oder dem Informationsfluss im Netzwerk, je nach Anwendungsfall. Die meisten solcher Maße können Eigenschaften des Netzwerks, wie die Richtung oder die Gewichtung der Kanten, mit einbeziehen.

In diesem Abschnitt soll anhand dreier Visualisierungen von generierten Figurennetzwerken (Abbildung 11) dargelegt werden, welche strukturellen Eigenschaften der Netzwerke in dieser Arbeit von Interesse sind und mit Hilfe welcher Maße diese als Features modelliert werden.



11 Figurennetzwerke zu Das Gemeindekind (links), Ekkehard (Mitte) und Die Wahlverwandtschaften (rechts)

Der Fokus liegt hierbei darauf, ein Netzwerk als Ganzes zu beschreiben, um möglichst dessen Grundstruktur zu modellieren. Das erste Beispiel ist wiederum das Figurennetzwerk zu Marie von Ebner-Eschenbachs *Das Gemeindekind*. Dabei handelt es sich um ein fast vollkommen sternförmiges Netzwerk, bei dem ein Knoten ganz klar zentral und mit allen anderen verbunden ist. Außer zwei Kanten mit relativ geringem Gewicht sind unter den äußeren Knoten keine weiteren Relationen vorhanden. Insgesamt sind die Kantengewichte relativ gleichmäßig verteilt: Sie variieren natürlich, es gibt aber keine Kante, die ein deutlich höheres Gewicht als alle anderen hat. Dies ist dagegen im zweiten Beispiel der Fall. Im Figurennetzwerk zu Joseph Victor von Scheffels *Ekkehard* gibt es nicht eine einzelne Figur, die besonders zentral ist, sondern eine Kante mit besonders hohem Gewicht, die zwei wiederum recht stark verknüpfte Figuren verbindet. Im Unterschied zum ersten Beispiel gibt es nur einen Knoten, an dem nur eine einzige Kante anliegt. Das dritte Beispiel zeigt das Figurennetzwerk zu Johann Wolfgang von Goethes *Die Wahlverwandtschaften*. Hier sind fast alle Figuren untereinander verbunden. An jedem Knoten liegen mindestens zwei Kanten an. Fast alle Figuren sind ähnlich stark ins Netzwerk eingebunden, es gibt keinen Knoten, der besonders zentral erscheint.

Von dieser Beschreibung ausgehend, müssen also insbesondere zwei Netzwerkeigenschaften als Features modelliert werden: Die Existenz besonders zentraler Knoten oder Kanten und deren Bezug auf das Netzwerk als Ganzes sowie die Ausprägung der Verbundenheit der Knoten untereinander. Dafür bieten sich zum Beispiel Zentralitätsmaße an, mit denen nicht nur wichtige Knoten identifiziert, sondern auch Informationen über die Struktur des Netzwerks gewonnen werden können.

Ein einfaches, bekanntes und oft verwendetes Zentralitätsmaß ist die Gradzentralität (*degree centrality,* c_D) (Newman 2010, S. 169). Für einen Knoten v ist sie definiert als der Quotient zwischen der Anzahl aller an einem Knoten anliegenden Kanten $\deg(v)$ und der Anzahl aller anderen Knoten im Graphen, außer dem betrachteten Knoten selbst.[16]

[16] In den folgenden Formeln bezeichnet n die Anzahl aller Knoten, V die Menge aller Knoten, m die Anzahl aller Kanten und $w(i, j)$ das Gewicht der Kante zwischen den Knoten i und j.

$$c_D(v) = \frac{\deg(v)}{n-1}$$

Der mögliche Wertebereich der Gradzentralität liegt also zwischen 1, falls alle möglichen Kanten eines Knoten realisiert sind, und 0, falls ein Knoten gar keine Kanten hat.[17] Als erstes Feature wird der maximale im Netzwerk vorkommende Wert für dieses Maß verwendet. Dieser kann bei 1 liegen, falls es sich um ein sternförmiges Netzwerk mit einem einzelnen Knoten im Zentrum handelt, aber auch, falls es sich um ein stark verbundenes Netzwerk handelt. In einem komplett verbundenen Graphen hätten alle Knoten eine Gradzentralität von 1. Um diese Fälle voneinander abzugrenzen, wird der kleinste vorkommende Wert der Gradzentralität zusätzlich als Feature herangezogen, da dieser bei einem sternförmigen Netzwerk sehr klein ist und bei einem stark verbundenen Netzwerk auch größer sein kann. Außerdem wird die Varianz der Gradzentralitäten aller Knoten verwendet: Hier deutet ein kleiner Wert darauf hin, dass alle Knoten eine ähnliche Gradzentralität haben, was bedeutet, dass der Graph relativ stark verbunden sein muss.

Dieses Zentralitätsmaß lässt jedoch die Kantengewichte, die ebenfalls als Information in den Figurennetzwerken enthalten sind, außer Acht, die die Einbeziehung weiterer Feinheit erlauben. In Analogie zum Knotengrad wird in gewichteten Graphen häufig die „Stärke“ eines Knoten (*node strength*) berechnet, also die Summe der Gewichte aller an einem Knoten anliegenden Kanten (Costa et al. 2007, S. 9). Dividiert man diesen Wert durch die Summe aller Kantengewichte im ganzen Netzwerk, so erhält man ein recht anschauliches Zentralitätsmaß für gewichtete Graphen, das als gewichtete Gradzentralität (*weighted degree centrality,* c_{WD}) bezeichnet wird und dessen Wertebereich ebenfalls zwischen 0 und 1 liegt.

$$c_{WD}(v) = \frac{\sum_{i \in V,\, i \neq v} w(v,i)}{\sum_{i,j \in V,\, i \neq j} w(i,j)}$$

Diese Berechnung führt dazu, dass im Unterschied zur ungewichteten Gradzentralität Knoten, an denen mehr starke Kanten anliegen, als wichtiger betrachtet

[17] Durch die Filterung von isolierten Knoten kann dies bei den in dieser Arbeit betrachteten Figurennetzwerken jedoch nicht vorkommen.

werden als solche, an denen vor allem schwache Kanten anliegen. Als Features werden der höchste und der zweithöchste Wert dieses Zentralitätsmaßes verwendet, sowie die Differenz zwischen beiden. Ist der erste Wert hoch und die Differenz zum zweiten Wert ebenfalls, so besitzt das Netzwerk einen besonders zentralen Knoten. Falls der zweite Wert auch höher und die Differenz relativ klein ist, deutet dies auf ein stärker verbundenes Netzwerk hin.

Ein weiteres Maß, das zur Beschreibung der Graphstruktur beitragen kann, ist die sogenannte *Central Point Dominance* (Costa et al. 2007, S. 28).

$$CPD = \frac{1}{n-1} \sum_{i \in V} (B_{max} - B_i)$$

Dieses Maß wird mit Hilfe der Betweenness-Zentralität B berechnet und lässt sich als durchschnittliche Differenz zwischen dem zentralsten Knoten und allen anderen interpretieren. Die Betweenness-Zentralität eines Knotens gibt dabei an, wie oft der betrachtete Knoten auf einem kürzesten Pfad zwischen zwei anderen Knoten liegt. B_{max} ist die höchste im Netzwerk vorkommende Betweenness-Zentralität, während B_i für die Betweenness-Zentralität des Knoten i steht. Das Ergebnis der Central Point Dominance ist ein Wert von 1 für einen sternförmigen Graphen, bei dem alle Kanten in einem Knoten zusammenlaufen, und ein Wert von 0 für einen komplett verbundenen Graphen.

Gil et al. definieren in ihrer Studie zu Figurennetzwerken aus Dramen und Filmen ein Maß namens *Single Relationship Centrality*, welches misst, wie stark eine einzelne Relation im Fokus steht (Gil et al. 2011, S. 4). Berechnet wird es aus der Differenz der Gewichte der beiden stärksten Kanten, im Verhältnis zum gesamten Kantengewicht im Graphen.

$$SRC = \frac{max_{i,j,\, i \neq j}\big(w(i,j)\big) - \; next\_max_{i,j,\, i \neq j}\big(w(i,j)\big)}{\sum_{i,j \in V,\, i \neq j} w(i,j)}$$

Je höher dieser Wert, desto eindeutiger enthält das Netzwerk eine Kante, die gegenüber allen anderen besonders zentral ist und ein besonders hohes Kantengewicht hat. Analog verwenden Gil et al. ein Maß für die Dominanz einer einzelnen

Figur, das jedoch in dieser Arbeit durch die Verwendung der Differenz zwischen den beiden höchsten gewichteten Gradzentralitäten bereits abgedeckt ist und daher nicht zusätzlich genutzt wird, um Korrelationen zwischen den Features zu vermeiden.

Neben Maßen, die die Netzwerkstruktur über die Zentralität von Knoten oder Kanten beschreiben, werden solche herangezogen, die die Verbundenheit des Graphen messen. Eine sehr einleuchtende solche Kennzahl ist die Dichte (*density*, D) eines Graphen.

$$D = \frac{m}{\frac{1}{2}n(n-1)}$$

Dabei handelt es sich um das Verhältnis der Anzahl aller tatsächlich im Netzwerk realisierten Kanten zu der Anzahl aller theoretisch möglichen Kanten (Newman 2010, S. 134). Auch dieser Wert bewegt sich zwischen 1 bei einem komplett verbundenen Netzwerk und 0 bei einem hypothetischen Netzwerk, das nur aus isolierten Knoten besteht. Je höher also der Wert, desto stärker sind die Knoten untereinander verbunden.

Ein weiteres Maß für die Verbundenheit eines Netzwerks ist die Transitivität (*transitivity*, C). Sie gibt das Verhältnis der Zahl aller geschlossenen Dreiecke (N_Δ) zur Zahl aller „Triaden“ (N_3), also Pfaden der Länge 2, die nicht zu einem Dreieck geschlossen sind, an (Costa et al. 2007, S. 19). Der Faktor 3 ergibt sich daraus, dass jedes Dreieck eigentlich aus drei Triaden besteht und sorgt dafür, dass der Wertebereich der Transitivität im Intervall [0, 1] liegt.

$$C = \frac{3N_\Delta}{N_3}$$

Damit beschreibt sie die Verbundenheit eines Netzwerks zusätzlich unter einem anderen Blickwinkel als die Dichte. Beide Maße für die Verbundenheit eines Netzwerks lassen die Kantengewichte außer Acht. Dies ist beabsichtigt, da es bei der Analyse der Struktur der Figurennetzwerke zunächst primär darum geht, wie viele Kanten im Netzwerk realisiert sind und in welcher Form.

Feature	*Gemeinde-kind*	*Ekkehard*	*Wahlverwandt-schaften*
Max. Degree	1,0	1,0	1,0
Min. Degree	0,1111	0,1667	0,4
Degree Varianz	0,0735	0,0820	0,0587
Top 1 Weighted Degree	0,9159	0,6685	0,5412
Top 2 Weighted Degree	0,2073	0,5471	0,5011
Weighted Degree Differenz	0,7085	0,1214	0,0400
Central Point Dominance	0,8951	0,2222	0,2400
Single Relationship Centrality	0,0012	0,1719	0,0069
Density	0,2444	0,5714	0,6667
Transitivity	0,15	0,6316	0,6667

1 Feature-Werte für Beispielnetzwerke

Tabelle 1 zeigt die Zahlenwerte für die Features für die drei oben gezeigten Beispielnetzwerke. Hier wird deutlich, dass die Features die Netzwerke entsprechend der dargelegten Erwartungen beschreiben und dass die Werte sich je nach Struktur des Netzwerks unterscheiden. So sind beispielsweise die gewichtete Grad-Differenz und die Central Point Dominance für das sternförmige Netzwerk zu *Das Gemeindekind* deutlich höher als für die anderen beiden Netzwerke. Analog dazu sind die Dichte und die Transitivität bei *Ekkehard* und insbesondere bei den *Wahlverwandtschaften* stärker ausgeprägt. Ebenso zu beobachten, ist der im Vergleich zu den anderen Netzwerken hohe Wert für die Single Relationship Centrality bei *Ekkehard*, dessen Netzwerk zweifellos eine zentrale Kante aufweist.

Berechnet man die beschriebenen Features für die ungefilterten Netzwerke, so liegen die Werte bei allen Maßen in sehr schmalen Bereichen und unterscheiden sich von Roman zu Roman nur sehr viel schwächer als die hier gezeigten Werte für

gefilterte Netzwerke. Dies bestätigt nochmals, dass die Filterung nicht nur für die Visualisierung, sondern auch für die Analyse der Figurennetzwerke sinnvoll ist.

Die meisten hier genannten Maße können mit NetworkX direkt berechnet werden. Wo das nicht der Fall ist, wurden zusätzliche Funktionen implementiert. Die berechneten Features wurden zu einer Featurematrix zusammengefasst, in der jede Spalte den Featurevektor für einen Roman enthält. Diese Matrix wird zur weiteren Verwendung als csv-Datei gespeichert.

5.5 Berechnung von Distanzen

Nachdem nun alle Romane durch eine Featurematrix repräsentiert sind, können mit Hilfe eines Distanzmaßes paarweise Ähnlichkeiten zwischen den Romanen berechnet werden. Die Verwendung eines Distanzmaßes zur Berechnung von Ähnlichkeit impliziert, dass eine hohe Distanz für eine geringe Ähnlichkeit steht und umgekehrt eine niedrige Distanz für große Ähnlichkeit.

Zunächst muss jedoch die Featurematrix skaliert werden, da andernfalls Features mit hohen Werten und größeren Wertebereichen die Berechnung von Distanzen dominieren können. Daher werden alle Features mittels Min-Max-Scaling (Raschka und Olson 2015, S. 110) so umgerechnet, dass sie danach in einem gemeinsamen Intervall von [0,1] liegen. Die Anwendung von Min-Max-Scaling ist sinnvoll, da es sich bei den Features um Messwerte auf kontinuierlichen Skalen handelt und es mehr um die relative Größe der Werte zueinander geht, als um Abweichungen von einer Norm. Genutzt wird die Implementierung MinMaxScaler() aus der Python-Bibliothek Scikit-Learn[18].

Als Distanzmaß wird die Kosinusdistanz verwendet, wie sie im Paket Scipy[19] enthalten ist. Dabei wird der Kosinus des Winkels zwischen zwei Vektoren berechnet und von 1 abgezogen, um einen Distanzwert zu erhalten. Aus den Eigenschaften des Kosinus ergibt sich, dass die Werte der Kosinusdistanz im Intervall zwischen 1 (höchste Distanz) und 0 (keine Distanz) liegen können. Daraus resultiert eine berechnete Distanzmatrix, die als csv-Datei gespeichert wird und im Folgenden näher betrachtet werden soll.

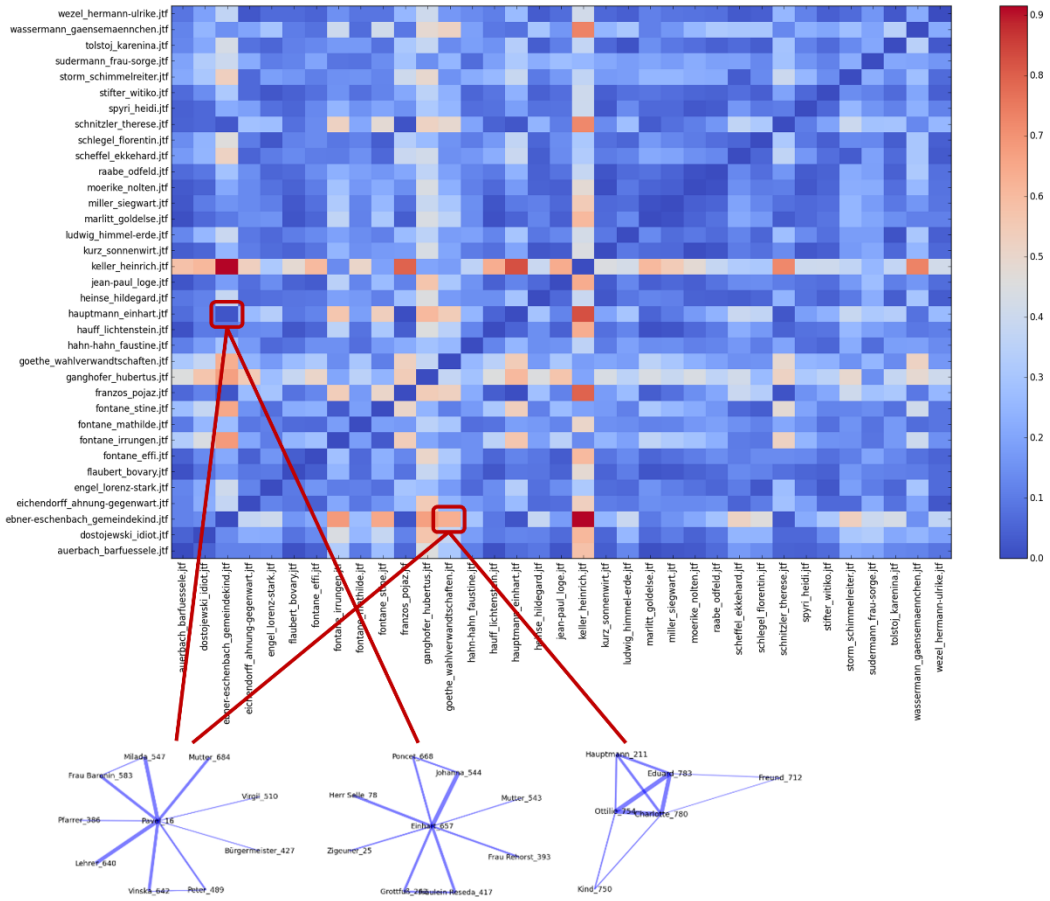
---

[18] http://scikit-learn.org/stable.

[19] https://www.scipy.org. Kosinusdistanz unter http://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cosine.html.

Ein genauerer Blick auf die berechneten Distanzen zeigt, dass der größte ermittelte Abstand bei 0,9157 liegt, während der kleinste Wert, abgesehen von der Diagonalen, auf der korrekterweise alle Werte 0 sind, bei 0,0064 liegt. Die in Bezug auf den Wertebereich der Kosinusdistanz relativ große Differenz zwischen Minimum und Maximum deutet darauf hin, dass eindeutig Unterschiede in den Daten messbar sind. Zwischen diesen Extremwerten können die Distanzen allerdings auch recht gleichmäßig verteilt sein.

Ein nützliches Hilfsmittel zur explorativen Untersuchung einer Distanzmatrix ist deren Visualisierung als Heatmap. Dabei werden die Zahlenwerte als Abstufungen von Farben kodiert, was dem Betrachter einen schnellen Überblick über die Daten ermöglicht. Anhand der Farben lässt sich schnell erkennen, zwischen welchen Romanen eine kleine bzw. große Distanz berechnet wurde. Abbildung 12 zeigt eine solche Visualisierung der berechneten Distanzmatrix. Für die Farbkodierung wurde eine von blau nach rot verlaufende Skala gewählt. Als Beispiel wurden je ein recht ähnliches und ein recht unähnliches Paar ausgesucht und die dazugehörigen Netzwerkvisualisierungen beigefügt. Das relativ dunkel blaue Feld zwischen Ebner-Eschenbachs *Gemeindekind* und Carl Hauptmanns *Einhart der Lächler* entspricht einer Kosinusdistanz von 0,0213, was anhand der Netzwerke sehr deutlich nachvollziehbar ist. Beide Netzwerke sind sternförmig mit einer zentralen Figur in der Mitte und besitzen je zwei Kanten zwischen äußeren Figuren. Deutlich rot eingefärbt ist mit 0,6349 dagegen die Distanz zwischen *Gemeindekind* und den *Wahlverwandtschaften*. Auch das entspricht den Erwartungen, da das Netzwerk zu den Wahlverwandtschaften im Gegensatz zum sternförmigen Netzwerk des *Gemeindekinds* auf vier stärker verbundene, gleichrangige Figuren fokussiert ist. Außerdem lässt sich erkennen, dass zueinander ähnliche Romane meist ähnliche Distanzen zu den restlichen Texten des Korpus aufweisen.

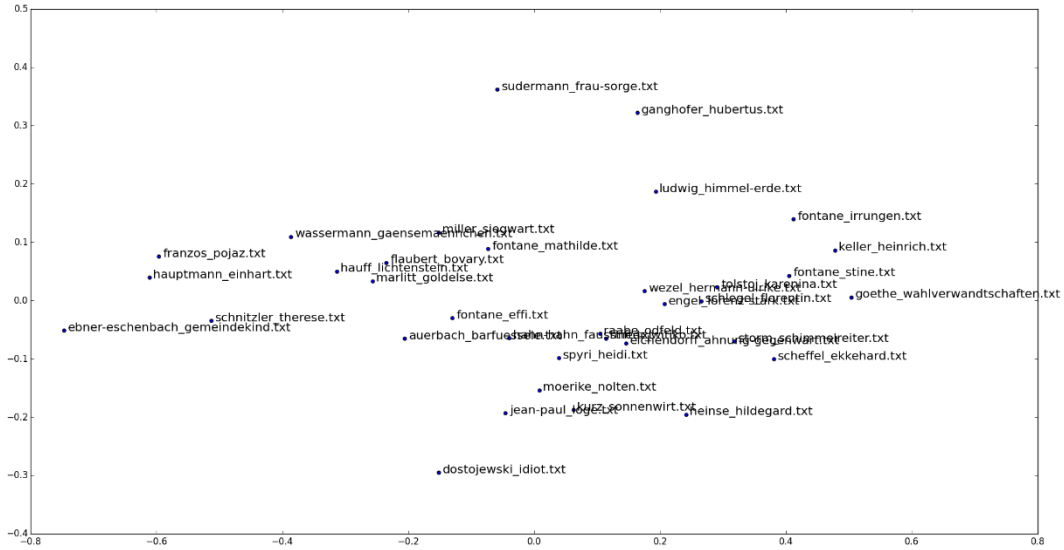


12 Heatmap der berechneten Distanzen mit Beispielnetzwerken[20]

Diese Visualisierung zeigt, dass die Berechnung von Distanzen auf Basis der Netzwerkfeatures Ergebnisse liefert, die bei Betrachtung der automatisch generierten Netzwerkvisualisierungen den Erwartungen entsprechen und durchaus nachvollziehbar sind. Die verwendeten Features scheinen also grundsätzlich geeignet zu sein, um strukturelle Eigenschaften von Figurennetzwerken zu modellieren.

Diese These lässt sich anhand einer Visualisierung auf Basis der Features selbst weiter untermauern. Berechnet man ausgehend von der skalierten Featurematrix eine Principal Component Analysis (PCA) zur Reduktion der Daten auf zwei Dimensionen und plottet diese als Scatterplot, so stellt man fest, dass auch diese Grafik sinnvoll interpretiert werden kann (Abbildung 13).

[20] Alle Abbildungen dieser Arbeit befinden sich zur genaueren Ansicht auch auf dem beiliegenden USB-Stick.



13 PCA-Scatterplot der Netzwerkfeatures

In der linken Hälfte der Grafik finden sich sternförmige Netzwerke, während auf der rechten Seite eher stärker verbundene zu finden sind. Am unteren Rand stehen einige Netzwerke, die eine zentrale, besonders starke Relation enthalten. *Effi Briest* und *Madame Bovary* liegen nicht allzu weit auseinander. Auch diese Betrachtung deutet darauf hin, dass die ausgewählten Features wie beabsichtigt bestimmte strukturelle Eigenschaften der Figurennetzwerke abbilden können.

5.6 Auswertung anhand der Evaluationsgrundlage

Um festzustellen, in wie weit mit der beschriebenen Vorgehensweise – automatische Netzwerkerstellung, Beschreibung durch Netzwerkfeatures und Berechnung von Distanzen – die zuvor als Evaluationsgrundlage festgehaltene Intuition über die Ähnlichkeit zwischen Romanen im Korpus abgebildet werden kann, müssen die manuell erstellte Distanzmatrix und die berechnete Distanzmatrix miteinander in Beziehung gesetzt werden.

Eine Möglichkeit zum Vergleich zweier Distanzmatrizen ist der Mantel Test, der 1967 von dem Biologen und Statistiker Nathan Mantel vorgeschlagen wurde (Mantel 1967). Der Test kann verwendet werden, um Distanzen zwischen den gleichen Objekten, die aus unterschiedlichen Quellen stammen, zu vergleichen, wie beispielsweise geographische und genetische Abstände zwischen Tierarten. Die Grundidee dieses Tests besteht darin, eine Korrelation zwischen beiden Matrizen zu berechnen, die zustande kommt, wenn jeweils an den gleichen Stellen hohe

bzw. niedrige Werte stehen. Ein Problem ist jedoch, dass paarweise Distanzen keine voneinander unabhängigen Daten sind. Mantels Lösungsansatz hierfür ist, eine der beiden Matrizen wiederholt zu permutieren und für jede Permutation die Korrelation zu berechnen. Dahinter steht die Annahme, dass die meisten dieser Permutationen keine Korrelation aufweisen, da dieser Zusammenhang durch das Permutieren einer der beiden Matrizen aufgehoben wird. Wenn wirklich eine Korrelation zwischen den beiden Matrizen besteht, dann sollte die tatsächliche Korrelation signifikant höher sein, als der Durchschnitt für alle Permutationen, was über den Z-Score angegeben werden kann. Der Linguist Jon W. Carr beschreibt die Funktionsweise des Tests und die gerade erläuterten Annahmen sehr anschaulich auf seinem Blog (Carr 2014) und hat zudem eine Python-Implementierung des Mantel Tests entwickelt, die er auf GitHub zur Verfügung stellt[21] und die in dieser Arbeit verwendet wird.

Carrs Funktion liefert die tatsächliche Korrelation r zwischen den zu vergleichenden Matrizen zurück, sowie den Signifikanzwert p und den Z-Score z. Wenn der p-Wert den festgelegten Schwellwert von 0,05 unterschreitet, deutet ein positiver Z-Score auf eine signifikant positive Korrelation hin und ein negativer Z-Score auf eine signifikant negative Korrelation. Der Wert r kann dabei als Stärke dieser Korrelation interpretiert werden. Da der Test zufällige Permutationen einer Matrix berechnet, können die Ergebnisse bei wiederholter Berechnung abweichen. Setzt man die Anzahl der zu betrachtenden Permutationen hoch genug, erhält man aber weitgehend stabile Ergebnisse, die sich erst in den hinteren Nachkommastellen unterscheiden. Je größer die betrachteten Matrizen und je höher die Zahl der Permutationen, desto rechenintensiver wird der Mantel Test jedoch auch.

Um die berechneten Werte besser einschätzen und interpretieren zu können, wurden vier Beispielmatrizen der Größe 10 erstellt und mittels des Mantel Tests verglichen. Matrix A wurde mit zufälligen Werten erstellt, Matrix B enthält an jeder Stelle den doppelten Wert von A, in Matrix C und D wurden zusätzlich die Werte für drei bzw. zehn Objektpaare verändert.

[21] http://jwcarr.github.io/MantelTest.

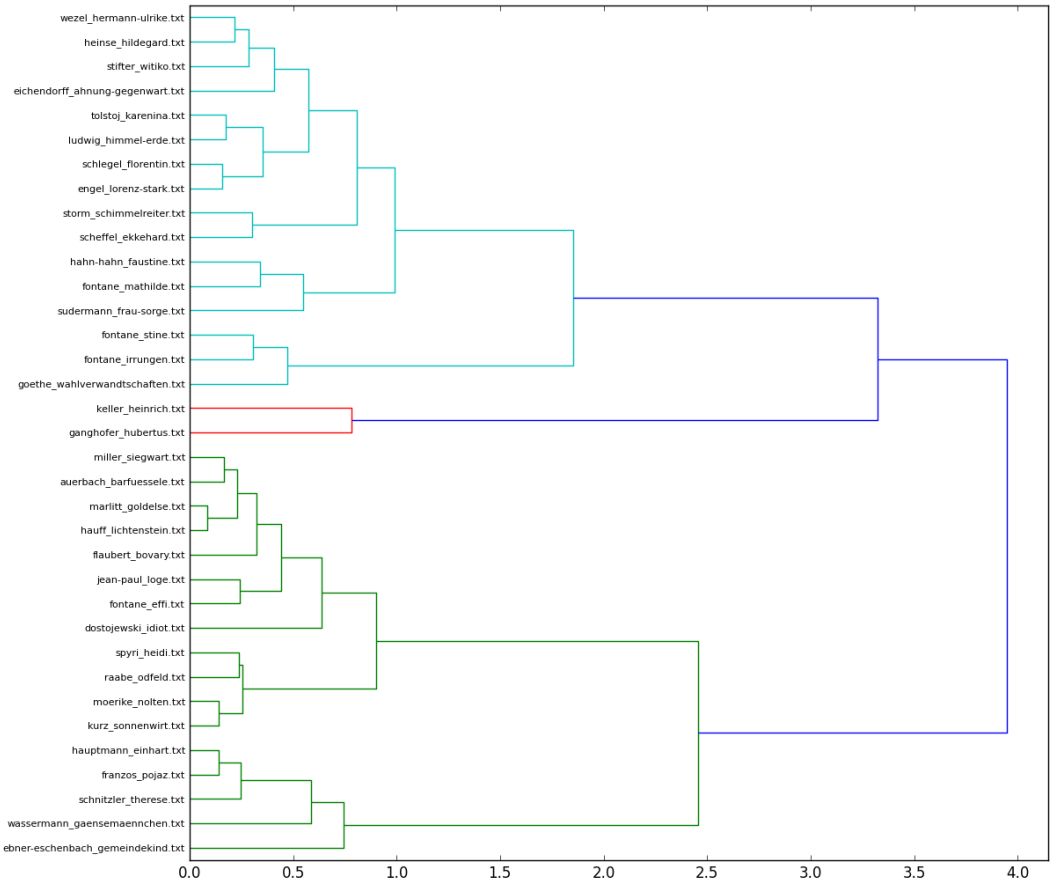
Vergleich	Beschreibung	r	p	z
A – B	Verdopplung	1,00	0,0001	6,55
A – C	3 Änderungen	0,90	0,0001	6,02
A – D	10 Änderungen	0,64	0,0001	4,29

2 Ergebnisse des Mantel Test für Beispielmatrizen

Die Tabelle zeigt, dass Matrix A zu allen anderen hoch signifikant positiv korreliert ist, was den Erwartungen entspricht. Werden jedoch mehr Abweichungen in die Matrizen eingebracht, so wird die Korrelation schrittweise schwächer.

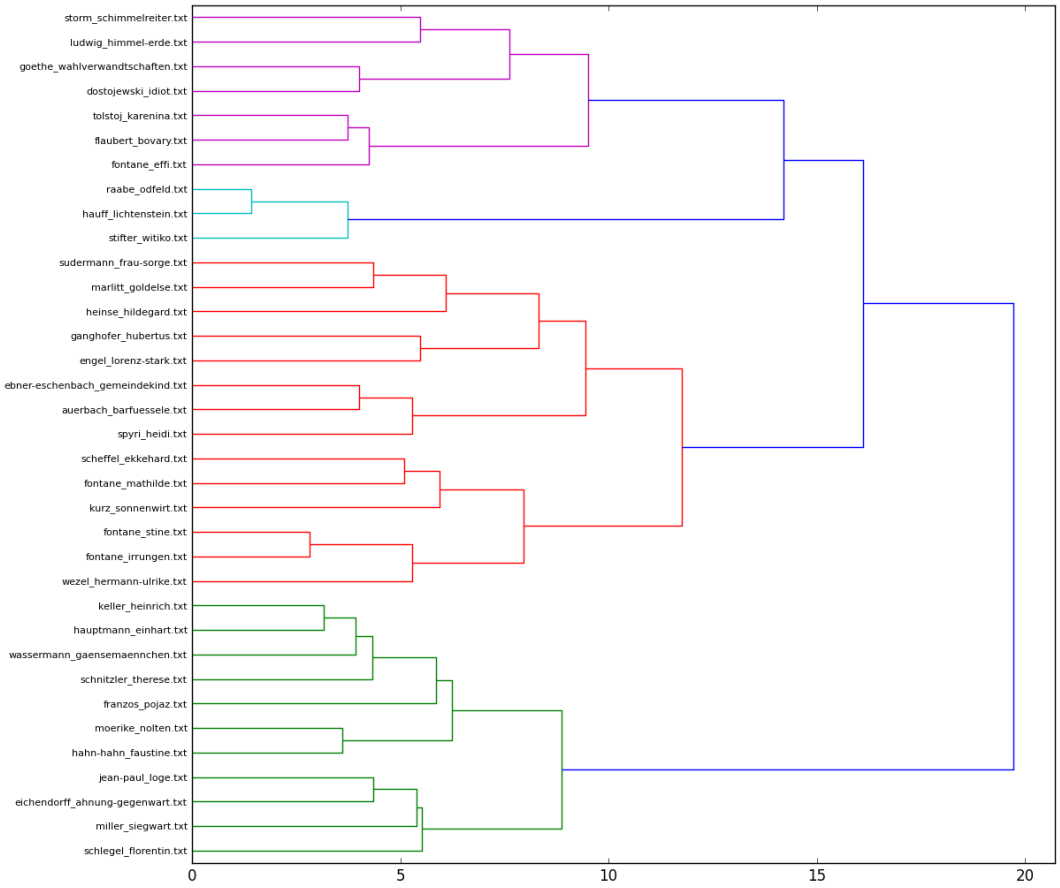
Nutzt man nun den Mantel Test, um die manuelle Distanzmatrix der Romane mit den anhand der Figurennetzwerke berechneten Distanzen zu vergleichen, erhält man diese Ergebnisse: r = -0,01, p = 0,89, z = -0,15. Negative Werte für r und z entstehen, wenn die verglichenen Matrizen negativ korreliert sind, also wenn Felder, die in der ersten Matrix hohe Zahlen enthalten, in der zweiten Matrix an diesen Stellen niedrige Zahlen enthalten und umgekehrt. Da die Werte für r und z in diesem Fall jedoch nahe an 0 liegen und der p-Wert sehr hoch ist, deuten sie leider in keiner Weise auf eine Korrelation hin. Ein mögliches Problem könnte darin liegen, dass die manuelle Distanzmatrix nur vier feste Werte für Distanzen enthalten kann (1, 2, 3, 4), während die berechneten Distanzen auf einer kontinuierlichen Skala zwischen 0 und 1 liegen. Um das zu umgehen, wurden die berechneten Distanzwerte ebenfalls auf die Kategorien 1 bis 4 abgebildet, indem alle Werte anhand der Quartile in vier Gruppen eingeteilt wurden. Lag der berechnete Wert zwischen 0 und dem ersten Quartil, so wurde eine 1 stattdessen eingesetzt und analog für die anderen Abschnitte. Daraufhin liefert der Mantel Test die Ergebnisse r = -0,02, p = 0,81 und z = -0,26. Auch das stellt leider keine Verbesserung dar.

Eine weitere Möglichkeit zur Analyse von Distanzmatrizen ist die Anwendung eines hierarchischen Clusterings und die anschließende Visualisierung mittels eines Dendrogramms. Dabei werden iterativ die Objekte bzw. Cluster mit dem geringsten Abstand zueinander zusammengefasst, sodass eine hierarchische Baumstruktur entsteht. Einer der verbreitetsten Algorithmen für hierarchisches Clustering ist die Ward-Methode, die darauf abzielt, die Varianz der Daten innerhalb eines Clusters zu minimieren (Ward 1963). Prozessiert man sowohl die manuelle, als auch die berechnete Distanzmatrix auf die gleiche Weise, so erhält man zwei Grafiken, die einem optischen Vergleich unterzogen werden können.



14 Dendrogramm der berechneten Distanzen

Betrachtet man das auf Basis der berechneten Distanzen erstellte Dendrogramm (Abbildung 14), so fällt zunächst auf, dass Ludwig Ganghofers *Schloß Hubertus* und Gottfried Kellers *Der Grüne Heinrich* als eigenes Cluster herausstechen. Zumindest bei letzterem Roman könnte dies auf die in 5.3.2 beschriebenen Probleme bei der Netzwerkerstellung zurückzuführen sein. Ansonsten lässt sich beobachten, dass das unterste grüne Cluster von Hauptmanns *Einhart der Lächler* bis zu Ebner-Eschenbachs *Gemeindekind* die am deutlichsten sternförmigen Netzwerke zusammenfasst. Bis auf *Das Gemeindekind* liegen diese auch im Dendrogramm zu den manuellen Distanzen (Abbildung 15) in einem Cluster, wenn auch nicht so deutlich von den anderen Romanen abgegrenzt.



15 Dendrogramm der manuellen Distanzen

Leider erschöpfen sich hier auch bereits die erkennbaren Gemeinsamkeiten, sodass eine weitere Interpretation kaum möglich ist und auf reiner Spekulation beruhen würde. Am Dendrogramm der manuellen Distanzen lassen sich allerdings Gruppierungen beobachten, die auf thematische Aspekte zurückgeführt werden können: Das türkise Cluster aus Raabe, Hauff und Stifter umfasst Romane, die sich mit der Thematik Krieg und Politik befassen. Das pinke Teilcluster darüber, bestehend aus *Anna Karenina*, *Madame Bovary* und *Effi Briest*, gruppiert Ehebruchsromane mit deutlichen Parallelen.

Um sich nicht auf einen optischen Vergleich verlassen zu müssen, können aus einem hierarchischen Clustering flache Cluster erzeugt werden, die jeden Roman zu einer bestimmten Gruppe zuordnen, indem das Dendrogramm sozusagen auf einer bestimmten Höhe abgeschnitten wird, um eine vorher festgelegte Zahl an Clustern zu erhalten. Da das Dendrogramm auf Basis der manuellen Distanzen vier recht deutliche Cluster zeigt, wurde die Clusteranzahl auf diesen Wert festgelegt. Mit Hilfe der `fcluster`-Funktion aus Scipy können aus beiden hierarchischen

Clusterings analog flache Gruppierungen erzeugt werden, die anschließend mit gängigen Maßen zur Auswertung der Performanz von Clustering-Algorithmen verglichen werden können. Ein solches Maß ist der *Accuracy Score*, der in scikit-learn implementiert ist. Dabei wird berechnet, für welchen Anteil der Romane die vorgegebene und die berechnete Clusterzuordnung miteinander übereinstimmen. Für den Vergleich zwischen dem Clustering anhand der manuellen Distanzen und dem Clustering anhand der berechneten Distanzen ergibt sich ein *Accuracy Score* von 0,34, was bedeutet, dass sich die beiden Clusterings sehr schlecht aufeinander abbilden lassen. Auch für andere Anzahlen von Clustern bleibt der Wert ähnlich schlecht oder schlechter.

Diese starken Abweichungen von der Evaluationsgrundlage sind im Hinblick auf die Tatsache, dass sowohl die auf Basis der Figurennetzwerke ermittelten Features als auch die berechneten Distanzen sinnvoll interpretiert werden können, durchaus erstaunlich. Dies führte zu der Idee, anhand einer simpleren Fragestellung zu untersuchen, ob die Netzwerkfeatures überhaupt grundsätzlich wie erwartet funktionieren. Daher wird nun die deutlich klarer definierte Frage untersucht, ob die Figurenkonstellation eines Romans um eine zentrale Hauptfigur angeordnet ist oder nicht. Dies wurde wiederum anhand der Zusammenfassungen festgehalten. Ausgehend von der berechneten Distanzmatrix wurde mit den bereits beschriebenen Methoden ein flaches Clustering mit zwei Clustern berechnet. Dieses wurde mit der manuellen Einteilung verglichen, wobei sich ein *Accuracy Score* von 0,74 ergibt. Dieser Wert liegt deutlich über der durchschnittlichen Performanz einer zufälligen Zuordnung und zeigt somit, dass die Features für diese Art der Klassifikation eindeutig relevante Informationen enthalten. Betrachtet man die Romane, die abweichend klassifiziert wurden, so handelt es sich dabei häufig um solche Fälle, in denen die anhand der Zusammenfassung erwartete Figurenkonstellation und das generierte Figurennetzwerk voneinander abweichen, ohne dass klar erkennbar ist, welche Darstellung den Roman besser widerspiegelt (vgl. Abschnitt 5.3.2). Das deutet darauf hin, dass die gewählten Features mindestens für die Unterscheidung zwischen sternförmigen und stärker verbundenen Netzwerke nachvollziehbare Ergebnisse liefern und nicht grundsätzlich schlecht funktionieren.

Klar ist: Obwohl die aus den Netzwerken generierten Features sinnvoll interpretierbar sind, wie in den Abschnitten 5.4 und 5.5 gezeigt wurde, passen die Evaluationsgrundlage und die berechneten Distanzen nicht zusammen. Dies stellte

sich sowohl bei der Verwendung eines Korrelationstests, als auch beim Clustering heraus. Ein möglicher Grund könnte sein, dass die Erstellung der Evaluationsgrundlage unterbewusst mehr als angenommen durch wiederkehrende Themen und Motive beeinflusst ist. Zwischenmenschliche Motive wie Ehe, gesellschaftlicher Stand oder Familienkonflikte können die Wahrnehmung der Figurenkonstellation deutlich formen, zumal wenn die Einschätzung lediglich auf der Lektüre von Zusammenfassungen basiert. Solche Informationen, die eher auf die verschiedenen Arten von Beziehungen zwischen Figuren abzielen, können Figurennetzwerke, wie sie in dieser Arbeit betrachtet werden, nicht abbilden.

6 Kombination mit Topic Modeling

Die vorhergehenden Experimente mit Figurennetzwerken haben zu der Annahme geführt, dass die menschliche Intuition von Ähnlichkeit in Bezug auf die Figurenkonstellation in Romanen neben der grundlegenden Struktur weitere Aspekte mit einbezieht. Im Folgenden wird versucht, mit Hilfe von Topic Modeling wichtige Themen zu modellieren, die im Korpus vorkommen. Die Idee dabei ist, dass auf diese Weise auch wiederkehrende zwischenmenschliche Motive erfasst werden können, die als Annäherung für die in den Romanen enthaltenen Beziehungsarten herangezogen werden können. So könnte eine weitere Dimension der Ähnlichkeit zwischen Figurenkonstellationen abgedeckt werden.

6.1 Preprocessing und Parameter

Zur Berechnung der Topics wird LDA verwendet, wie es in Abschnitt 3.2.1 beschrieben ist. Eine der bekanntesten und am weitesten verbreiteten Implementierungen dieser Methode ist Mallet[22]. Es ist in Java geschrieben, als Open-Source-Software frei verfügbar und lässt sich über die Kommandozeile bedienen. Aus Python heraus kann Mallet mit Hilfe des Moduls subprocess[23] angesteuert werden, welches es ermöglicht, Befehle auf der Kommandozeile durchzuführen und die

[22] http://mallet.cs.umass.edu/topics.php.

[23] https://docs.python.org/3/library/subprocess.html.

Rückgabewerte zu speichern. Es gibt natürlich auch direkt in Python implementierte Bibliotheken für Topic Modeling, wie beispielsweise Gensim[24]; Mallet hat sich jedoch für diese Arbeit, unter anderem im Hinblick auf die Laufzeit, als performanter erwiesen.

Als Input benötigt Mallet Plain-Text-Dateien. Da die Romane im zuvor beschriebenen tabellarischen Format (vgl. Abbildung 3) vorliegen, muss zunächst eine reine Textfassung daraus generiert werden. Davor werden die Daten jedoch einem recht umfangreichen Preprocessing unterzogen, wie es im Bereich des Topic Modeling üblich ist.

Typischerweise werden dabei die Texte gefiltert, sodass nur noch bestimmte Tokens in das Modell einfließen. Da vor allem thematische Informationen modelliert werden sollen, liegt es nahe, sich auf die Substantive zu beschränken, weil diese in einem Text den höchsten inhaltlichen Gehalt haben. Das tabellarische Format enthält bereits Part-of-Speech-Tags zu jedem Wort, sodass die gewünschten einfach ausgewählt werden können. Auf diese Weise werden, unter anderem, die sogenannten Funktionswörter wie zum Beispiel Artikel oder Pronomen ausgeschlossen, die aus geschlossenen Wortklassen stammen, meistens sehr häufig vorkommen und inhaltlich wenig beitragen.

Außerdem werden alle Figurenreferenzen ausgeschlossen, da diese spezifisch für einzelne Romane sind und somit keine hilfreichen Informationen beitragen können, wenn es darum geht, mehrere Texte zu vergleichen. Ähnliches gilt für die Hapax Legomena, also Wörter, die nur einmal im ganzen Korpus vorkommen. Diese werden aus dem gleichen Grund entfernt.

Analog zum sogenannten Culling, das beispielsweise in der Stilometrie eingesetzt wird, um Wörter herauszufiltern, die zu spezifisch für wenige Texte sind, wurden alle Wörter, die nur in einem einzigen Roman vorkommen, ignoriert (Eder et al. 2016). All diese Maßnahmen zielen auf die Vermeidung von Topics ab, die zu sehr auf einen Einzeltext fokussiert sind.

Nach den ersten Testläufen stellte sich heraus, dass sich einige Wörter in sehr vielen Topics wiederholten. Da es sich dabei vor allem um solche Wörter handelte, die im Gesamtkorpus sehr häufig vorkommen, wurden zusätzlich zu den anderen Preprocessing-Schritten die 50 häufigsten Wörter entfernt.

---

[24] https://radimrehurek.com/gensim.

Des Weiteren werden die lemmatisierten Formen der Substantive verwendet, die ebenfalls aus dem tabellarischen Format ersichtlich sind. Bei der Lemmatisierung werden der Plural und andere Flexionsformen auf eine gemeinsame Grundform zurückgeführt, zum Beispiel ‚Kinder' und ‚Kindes' auf die Form ‚Kind'. In Ausnahmefällen, in denen der Lemmatisierer keine Grundform finden konnte, wird das ursprüngliche Token beibehalten. Durch diesen Preprocessing-Schritt wird unerwünschte Variabilität im Vokabular reduziert, was zu konsistenteren Topics führt.

Ein weiterer Schritt, der oft unternommen wird, ist das Aufteilen von längeren Texten in kürzere Einheiten. LDA betrachtet jeden Text als Bag-of-Words, ohne die ursprüngliche Reihenfolge der Wörter zu berücksichtigen. Die Wahrscheinlichkeit, dass Wörter in einem ganzen Roman gemeinsam auftauchen, ist sehr viel höher und die damit verbundene Information sehr viel weniger aussagekräftig, als in einem kleineren Textabschnitt. Würde man die Romane im Ganzen als Input verwenden, so würden sehr breite und unspezifische Topics entstehen und andere Themen, die nur an einzelnen Stellen vorkommen, überlagert werden. Diese Problematik beschreibt auch Matthew Jockers in *Macronanalysis* und betont, dass die Segmentierung von Romantexten vor der Anwendung von Topic Modeling eindeutig sinnvoll ist und zu besseren Ergebnissen führt. Gleichzeitig weist er darauf hin, dass es keine allgemeingültige Richtlinie für die optimale Segmentierung gibt und diese in jeder Studie empirisch festgelegt werden muss (Jockers 2013, S. 134).

Zum Segmentieren der Romane gibt es mehrere Ansätze. Zum einen könnten Sinnabschnitte wie Kapitel oder Absätze verwendet werden. Diese sind jedoch nicht notwendigerweise in jedem beliebigen Roman enthalten und können in der Länge sehr stark variieren, sowohl innerhalb eines Textes als auch über ein Korpus hinweg. Eine andere Möglichkeit ist, die Texte in Abschnitte mit einer bestimmten Wortanzahl zu unterteilen. Welche Größe dabei für die Segmente gewählt werden sollte, kann nicht allgemein festgelegt werden. In der vorliegenden Studie wurde eine Segment-Länge von 300 Wörtern verwendet und außerdem darauf geachtet, Absatzgrenzen zu berücksichtigen, ähnlich wie bei Schöch et al. (Schöch et al. 2016). Hierbei wurden Segmente mit der festgelegten Wortanzahl betrachtet und dann jeweils die nächstliegende Absatzgrenze als tatsächlicher Trennpunkt herangezogen. Zudem wurde dafür Sorge getragen, dass das letzte Segment nicht zu klein wird, indem es gegebenenfalls zum vorletzten Segment dazu genommen wurde. Die

Einbeziehung von Absatzgrenzen dient dazu, den Sinnzusammenhang des Textes weitgehend zu bewahren.

Nach Durchführung des beschriebenen Preprocessing wurden die entstandenen Segmente als Textdateien gespeichert. Dafür wurden neue Dateinamen, bestehend aus dem ursprünglichen Namen und einer fortlaufenden Zählung, generiert, um die Segmente später wieder den jeweiligen Romanen zuordnen zu können.

Ein entscheidender Parameter, der bei LDA vom Benutzer festgelegt werden muss, ist die Anzahl der Topics, die das Modell ermitteln soll. Auch hier gibt es keine allgemeingültige Angabe: Die Wahl der Topic-Anzahl hängt von der Größe und Variabilität des Korpus, sowie der gewünschten Detailgenauigkeit der Topics ab (Jockers 2013, S. 128). Eine zu niedrige Topic-Anzahl resultiert in sehr allgemeinen Topics, die tendenziell in allen Texten vertreten sind, während eine zu hohe Anzahl zu schwerer interpretierbaren Topics führen kann. Nach mehreren Durchläufen wurde in dieser Arbeit eine Topic-Anzahl von 70 festgelegt.

Nach der Durchführung von LDA mit Mallet erhält man mehrere Ausgabedateien. Eine Textdatei enthält alle Topics mit der jeweiligen Gesamtwahrscheinlichkeit im Korpus, sowie den 20 Wörtern mit der höchsten Gewichtung für das jeweilige Topic, ohne jedoch die konkreten Wahrscheinlichkeiten für die einzelnen Wörter anzugeben. Diese Ausgabe ist eine gute Möglichkeit für den Benutzer, einen schnellen Überblick über die berechneten Topics zu erhalten. Die Informationen über alle in einem Topic enthaltenen Wörter mit den dazugehörigen Wahrscheinlichkeiten sind in einer separaten Datei aufgelistet. Auch das lässt sich gut für die Analyse der entstandenen Topics nutzen, beispielsweise für die im nächsten Abschnitt gezeigten Visualisierungen. Außerdem liefert Mallet die Topic-Verteilung über die einzelnen Dokumente in Form einer Matrix, bei der jede Zeile ein Dokument (also ein Romansegment) und jede Spalte ein Topic repräsentiert. In den Zellen stehen die Wahrscheinlichkeiten dafür, dass das jeweilige Topic im Dokument enthalten ist. Diese Datei bietet alle notwendigen Informationen für eine Analyse der Topics im Hinblick auf ihre Verteilung über die Dokumente.

6.2 Interpretation der entstandenen Topics

Bei der Einschätzung der Qualität der von einem LDA-Modell berechneten Topics lässt sich ein menschlicher Betrachter typischerweise von zwei Aspekten leiten: Wie gut passen die mit höchster Wahrscheinlichkeit in einem Topic enthaltenen Wörter zusammen und wie leicht fällt es, einen Überbegriff für das Topic zu finden? Diese beiden Dimensionen, auch als Kohärenz und Interpretierbarkeit bezeichnet, sind gerade bei der Arbeit mit literarischen Texten von besonderer Bedeutung, da sich Literaturwissenschaftler natürlich interpretierbare und nachvollziehbare Erkenntnisse über Romane erhoffen (Jockers 2013, S. 128). Beide Aspekte beeinflussen sich stark gegenseitig: je eindeutiger der Zusammenhang zwischen den Wörtern eines Topics, desto leichter fällt es einem Betrachter, das Topic mit einem Label zu benennen.

Obwohl natürlich im Optimalfall alle Topics klar interpretierbar sein sollten, bedeutet das Vorkommen von weniger kohärenten Topics keineswegs, dass das errechnete Modell schlecht oder unbrauchbar ist. Betrachtet man die 70 Topics, die für das in dieser Arbeit verwendete Korpus berechnet wurden, so zeigt sich, dass auch hier manche klarer und manche weniger klar interpretierbar sind.

Da Topics nach Wahrscheinlichkeit gewichtete Wortverteilungen sind, können sie gut als Wordclouds visualisiert werden. Die einzelnen Wörter werden dabei je nach ihrer Gewichtung in verschiedenen Schriftgrößen dargestellt: je häufiger ein Wort in einem Topic vertreten ist, desto größer wird es geschrieben. Diese Art der Visualisierung ermöglicht einen anschaulicheren Eindruck von der Zusammensetzung eines Topics, als die Betrachtung einer einfachen Wortliste, da die Wahrscheinlichkeitsverteilung der enthaltenen Wörter mit berücksichtigt wird. Zur Erstellung der Wordclouds wurde das gleichnamige Python-Package[25] verwendet und jeweils die 15 wichtigsten Wörter dargestellt.

Bei der Betrachtung dieser Wordclouds fällt auf, dass Topics auf verschiedene Aspekte hin interpretiert werden können. Manche Topics repräsentieren zentrale Probleme oder Ereignisse, die die Handlung eines Textes prägen können.

[25] http://amueller.github.io/word_cloud.

16 Topic 37: Krankheit/Tod

17 Topic 5: Geld(-sorgen)

Ein Beispiel dafür ist Topic 37 (Abbildung 16), das Wörter rund um ‚Krankheit' und ‚Tod' enthält und damit ein häufig in literarischen Texten vorkommendes Motiv beschreibt. Tatsächlich ist es in den meisten Romanen des Korpus vertreten. Auf der rechten Seite ist als weiteres Beispiel Topic 5 (Abbildung 17) mit dem Thema ‚Geld' zu sehen, das auch einige Wörter enthält, die Finanzprobleme und Geldsorgen andeuten. Dieses Topic findet sich zum Beispiel in *Madame Bovary*, da Emma gegen Ende des Romans immer häufiger auf Geldanlagen und Geschäfte mit dem Händler L'Heureux einlässt, die sie nicht vollständig überblickt und sich damit in eine finanzielle Notlage bringt.

Andere Topics beschreiben eher das Setting eines Romans. Topic 48 in Abbildung 18 umfasst Wörter, die sich auf das Thema ‚Kloster' beziehen. Romanen, in denen dieses Topic vertreten ist, spielen wahrscheinlich, zumindest stellenweise, in einer klösterlichen Umgebung, wie beispielsweise Scheffels *Ekkehard*, der von einem jungen Mönch handelt.

18 Topic 48: Kloster

19 Topic 43: Ländliches

Ähnlich spricht das Vorkommen von Topic 43 (Abbildung 19) dafür, dass ein Roman, oder Teile davon, in einer ländlichen Umgebung spielt, in der Landwirtschaft von Bedeutung ist. Im Hinblick auf die Figurenkonstellation deuten diese

Topics darauf hin, dass eine oder mehrere Romanfiguren mit den genannten Bereich in Bezug stehen, also beispielsweise Mönch oder Bauer sind.

Ähnlich lassen sich Topics betrachten, die bestimmte Interessensgebiete oder Lebensinhalte beschreiben. Auch hier liegt nahe, dass Figuren im entsprechenden Roman an diesen Bereichen interessiert sind und zugehörige Tätigkeiten ausüben.

20 Topic 11: Jagd

21 Topic 14: Musik

Beispiele hierfür sind Topic 11 (Abbildung 20), das sich mit dem Begriff ‚Jagd' beschreiben lässt und in Ganghofers *Schloß Hubertus* prominent vertreten ist, in dem ein alter Schlossherr sein Leben dieser Beschäftigung gewidmet hat. In Wilhelm Heises *Hildegard von Hohenthal* hat Topic 14 eine hohe Wahrscheinlichkeit, welches Wörter aus dem Themengebiet ‚Musik' umfasst (Abbildung 21).

Manche Topics bilden auch ganz klar zwischenmenschliche Beziehungen ab. Ein Beispiel dafür ist Topic 56 in Abbildung 22, welches Wörter rund um ‚Hochzeit' und ‚Ehe' enthält.

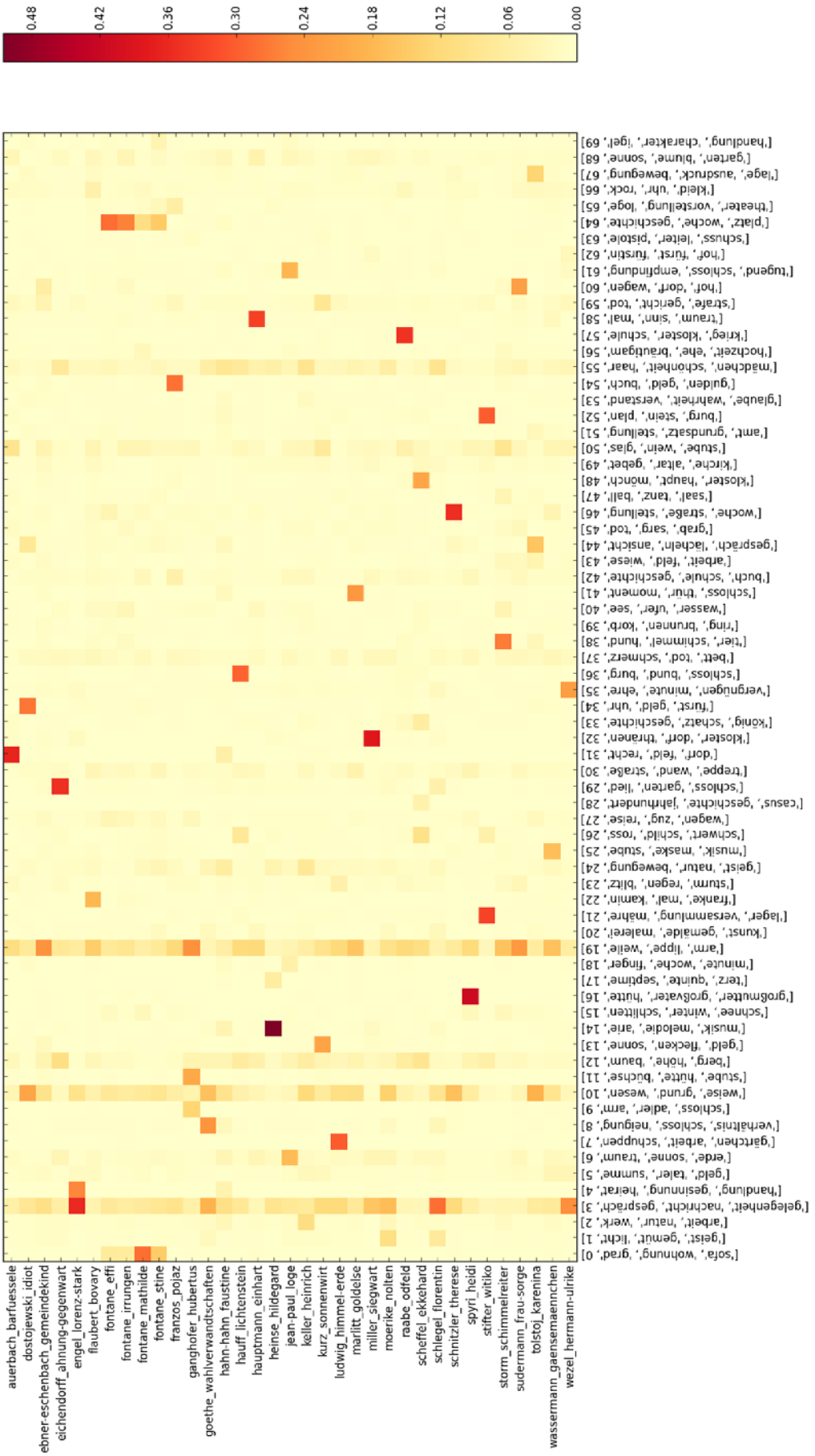
22 Topic 56: Hochzeit

In Romanen, in denen dieses Topic vertreten ist, findet also höchstwahrscheinlich eine Hochzeit statt oder das Thema wird angesprochen. Beispiele dafür

sind Theodor Fontanes *Mathilde Möhring* oder *Der Schimmelreiter* von Theodor Storm.

Außer den Topics liefert Mallet auch die Topic-Verteilung über die einzelnen Dokumente. Bei einem Dokument handelt es sich dabei um ein Romansegment. Im vorliegenden Anwendungsfall ist aber die Verteilung der Topics über die Romane als Ganzes von Interesse. Daher muss für jedes Topic die durchschnittliche Wahrscheinlichkeit über alle Segmente hinweg berechnet werden. Da die Topic-Verteilung als Matrix gespeichert ist, lassen sich dafür die `groupby`-Funktionen aus Pandas nutzen.

Um sich einen ersten Überblick zu verschaffen, wie die Topics über die Romane verteilt sind, bietet sich wiederum eine Heatmap zur Visualisierung an (Abbildung 23). Dabei werden auf der vertikalen Achse die Romane aufgetragen. Auf der horizontalen Achse befinden sich die Topics, wobei für jedes Topic die entsprechende Nummer sowie die ersten drei Wörter als Beschriftung angezeigt werden. Die Wahrscheinlichkeiten für die Topics werden durch eine Farbskala repräsentiert: je dunkler das Feld, desto höher die Wahrscheinlichkeit, dass ein Topics im entsprechenden Text vorkommt.



23 Heatmap der Topic-Verteilung

An der Heatmap lässt sich erkennen, dass die Verteilung über die Romane je nach Topic tatsächlich sehr unterschiedlich ausfallen kann. Manche Topics sind relativ gleichmäßig in fast allen Romanen vertreten. Andere hingegen stechen mit besonders hoher Wahrscheinlichkeit in einzelnen Texten heraus. Es gibt jedoch auch einige Topics, die manche Romane gemeinsam haben, während sie in anderen gar nicht auftreten, wie beispielsweise Topic 26, das sich um das Thema ‚Kampf' dreht. Ein weiteres Beispiel ist Topic 20 zum Thema ‚Kunst und Malerei', welches Romane verbindet, deren Hauptfigur Künstler ist oder die sich anderweitig mit dem Motiv ‚Kunst' auseinandersetzen.

Auch in den LDA-Ergebnissen scheinen also nützliche Informationen enthalten zu sein. Diese sollen im Folgenden zusätzlich zu den Netzwerk-Features zur Berechnung von Distanzen herangezogen werden, um die Evaluationsgrundlage eventuell besser annähern zu können.

6.3 Berechnung von Distanzen und Auswertung

Um die Informationen aus dem Topic Modeling in die Berechnung der Distanzen zwischen Romanen mit einfließen zu lassen, werden für jeden Roman die Wahrscheinlichkeiten für die einzelnen Topics als Features verwendet. Dafür wird, wie bereits oben beschrieben, die durchschnittliche Topic-Verteilung über alle Romansegmente ermittelt und in einer csv-Datei zwischengespeichert. Die Netzwerk-Features und die Topic-Features werden in einem Pandas-DataFrame zusammengeführt. Es sind also für jeden Text 70 weitere Features hinzugefügt worden. Analog zu der in Abschnitt 5.5 beschriebenen Vorgehensweise werden die Daten skaliert und die Kosinusdistanz berechnet.

Auf diese Weise erhält man wiederum eine Distanzmatrix und es gilt zu überprüfen, ob das erweiterte Feature-Set die Evaluationsgrundlage möglicherweise besser annähert. Der Mantel Test liefert die Werte $r = 0{,}18$, $p = 0{,}005$ und $z = 2{,}82$. Dies deutet auf eine signifikant positive Korrelation hin, auch wenn diese nur sehr schwach ausgeprägt ist. Die Idee, Topic Modeling als weitere Dimension mit einzubinden, scheint also im Sinne der Evaluationsgrundlage zu sein.

Obwohl das kombinierte Feature-Set sehr viel mehr Topic-Features als Netzwerk-Features enthält, tragen die Netzwerk-Features dennoch eindeutig zu den berechneten Distanzen bei und werden nicht von den Topic-Features überlagert.

Dies zeigt sich, wenn man die nur auf Basis der Netzwerk-Features berechnete Distanzmatrix mit der anhand der kombinierten Features ermittelten Distanzmatrix vergleicht. Die Werte r = 0,59, p = 0,0001 und z = 6,11 zeigen eine deutliche positive Korrelation.

Berechnet man die Distanzen jedoch nur auf Basis der Topic-Features, ohne die Netzwerk-Informationen zu berücksichtigen, und vergleicht diese mit der Evaluationsgrundlage, so liefert der Mantel Test r = 0,22, p = 0,0006 und z = 3,28. Diese Korrelation ist etwas stärker und hat gleichzeitig einen niedrigeren Signifikanzwert als für die Distanzen basierend auf den kombinierten Features. Das deutet darauf hin, dass die Informationen aus dem Topic Modeling eher in der Lage sind, die in der manuell erstellten Distanzmatrix festgehaltene Intuition über die Ähnlichkeit zwischen Romanen abzubilden, als die Netzwerkmaße. Angesichts der Tatsache, dass die Visualisierungen der Figurennetzwerke und die daraus ermittelten Kennzahlen doch sehr klar interpretierbar sind, ist dies erstaunlich und wirft die Frage auf, ob die Evaluationsgrundlage zu sehr von in den Romanen vorkommenden Themen und Motiven geprägt ist und bei der Erstellung die Figurenkonstellation unterbewusst in den Hintergrund getreten ist.

Im Gegensatz zu den Netzwerkfeatures, bei denen die berechnete Distanzmatrix anhand der Netzwerkvisualisierungen gut nachvollzogen werden konnte, ist eine solche Betrachtung der Distanzen hier kaum möglich. Obwohl der Mantel Test eine positive Korrelation anzeigt, ist diese nur sehr schwach und lässt keineswegs die Schlussfolgerung zu, dass Topic-Features die Ähnlichkeit zwischen den Figurenkonstellationen verschiedener Romane abbilden würden. Vielmehr drängen sich Zweifel daran auf, ob die Evaluationsgrundlage tatsächlich die Intuition von Ähnlichkeiten zwischen Romanen mit Fokus auf der Figurenkonstellation abbildet oder doch zu stark von anderen Faktoren beeinflusst ist.

Zudem wurden die beiden Featuresets lediglich durch einfaches Zusammenfügen miteinander kombiniert, sodass nun insgesamt 80 Features für jeden Roman betrachtet werden. Bei einem kleinen Korpus ist dies bereits extrem viel, sodass die Daten diesen höher-dimensionalen Raum nur noch spärlich abdecken (die Daten werden *sparse*). Im unendlich dimensionalen Raum wären alle Datenpunkte gleich weit voneinander entfernt. Das kann dazu führen, dass berechnete Distanzen nicht

mehr aussagekräftig sind. Dieses Phänomen wird auch als ‚Fluch der Dimensionalität' (*curse of dimensionality*) bezeichnet (Keogh und Mueen 2011) und könnte eventuell erklären, warum der Mantel Test hier eine Korrelation anzeigt.

Daher wird im Folgenden ein Ansatz vorgestellt, der es erlaubt, die Figurennetzwerke und die LDA-Topics auf differenziertere Weise zusammenzubringen.

6.4 Topics als Kanteneigenschaften

Dieser Abschnitt untersucht die Frage, ob und wie Figurennetzwerke mit Informationen aus einem Topic-Modell angereichert werden könnten. Die Idee dabei ist, jede Kante durch ein dazugehöriges Topic zu charakterisieren und so eventuell verschiedene wiederkehrende Kantentypen und Muster solcher Typen über die Romane hinweg identifizieren zu können. Auf diese Weise könnten die Informationen aus dem Topic-Modell direkter mit den Figurennetzwerken verknüpft werden, als nur durch die Verwendung der Topic-Verteilungen als Features.

Da die Kanten der Figurennetzwerke über gemeinsames Vorkommen von Figuren im gleichen Absatz modelliert sind, muss auch das Topic-Modell auf Absätzen beruhen, damit beide Informationen miteinander in Verbindung gesetzt werden können. Daher wurde erneut ein LDA-Modell berechnet, mit dem Unterschied, dass die Romantexte diesmal in die einzelnen Paragraphen segmentiert wurden. Aufgrund der variierenden Länge von Absätzen kann es vorkommen, dass für manche Paragraphen kein LDA-Dokument erstellt wurde, da diese durch ihre Kürze keine Substantive enthalten oder alle Substantive im Preprocessing herausgefiltert wurden. Die einzelnen Absatzsegmente wurden mit dem Dateinamen des Romans sowie der entsprechenden Absatznummer aus dem tabellarischen Format benannt, um sicherzustellen, dass die Informationen anschließend wieder mit den Figurennetzwerken zusammengeführt werden können. Abgesehen davon wurden das gleiche Preprocessing wie in Abschnitt 6.1 angewendet und ebenfalls 70 Topics berechnet.

Die entstandenen Topics sollen hier nicht im Detail beleuchtet werden. Es kann jedoch festgehalten werden, dass das Ergebnis vergleichbar zu den in Abschnitt 6.2 beschriebenen Topics auf Basis der Segmentierung nach Wortanzahl ist. Ein Großteil der Topics lässt sich analog wiedererkennen, natürlich jeweils mit einer anderen Topic-Nummer.

Bei der Erstellung der Interaktionslisten wurde nun für jeden Absatz das Topic mit der höchsten Wahrscheinlichkeit ermittelt und allen im Absatz vorkommenden Kanten zugeordnet. Sofern zu einem Absatz aufgrund leerer LDA-Dokumente keine Informationen über die Topic-Verteilung vorlagen, wurde der Platzhalter -1 vergeben. Auf diese Weise erhält man für jede Kante eine Liste von Topics, deren Länge dem Gewicht der Kante entspricht. Aus dieser Liste wurde das am häufigsten vorkommende Topic herausgesucht und als zusätzliches Kantenattribut neben dem Kantengewicht gespeichert. Als weitere Information wurde die Anzahl der Absätze, in denen das gewählte Topic am wahrscheinlichsten war, festgehalten (vgl. Abbildung 24).

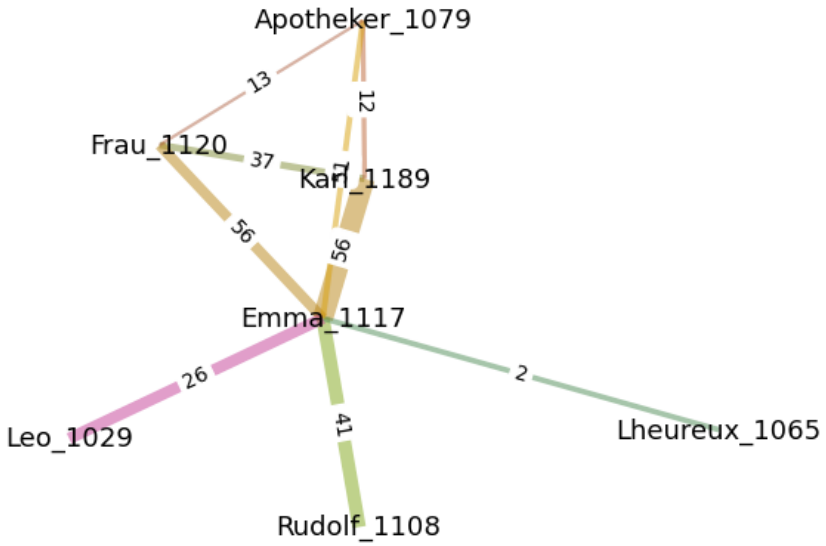
Diese Informationen können in der Form ‚gewähltes Topic_Anzahl' als weitere Spalte in die Interaktionslisten geschrieben werden, da NetworkX Mechanismen zum Einlesen mehrerer Kantenattribute bietet.

Emma_1117	Karl_1189	248	56_23
Emma_1117	Rudolf_1108	163	41_13
Emma_1117	Leo_1029	140	26_13
Emma_1117	Frau_1120	127	56_16
Frau_1120	Karl_1189	85	37_8

24 Ausschnitt aus Interaktionsliste zu Madame Bovary mit Topic-Informationen

Bei der Visualisierung dieser erweiterten Figurennetzwerke bietet es sich nun an, die Kanten je nach zugeordnetem Topic in verschiedenen Farben darzustellen. Dabei wurde mit Colormaps aus Matplotlib[26] gearbeitet und jede Topic-Nummer auf eine bestimmte Farbabstufung gemappt. Bei der Auswahl des Farbschemas aus dem limitierten Angebot an Colormaps wurde darauf geachtet, dass die Farben möglichst gut unterscheidbar sind. Außerdem wurde die jeweilige Topic-Nummer als Label an die Kanten geschrieben, damit die dazugehörigen Topics zur besseren Analyse nachgeschlagen werden können.

[26] http://matplotlib.org/users/colormaps.html.



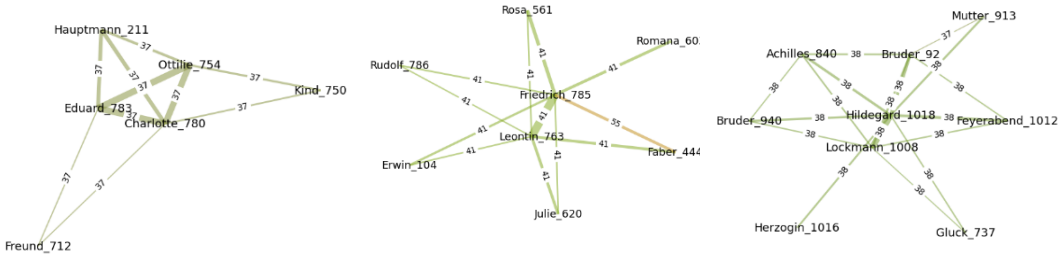
25 Figurennetzwerk zu Madame Bovary mit Topics als Kantenattributen

Betrachtet man beispielsweise das mit Topic-Informationen angereicherte Figurennetzwerk zu *Madame Bovary* (Abbildung 25), so stellt man fest, dass fast allen Kanten unterschiedliche Topics zugeordnet wurden. Zwei Kanten, zwischen Emma und Frau Bovary sowie zwischen Emma und Karl, haben das gleiche Topic erhalten. Da es sich bei den beiden Figuren um Emmas Ehemann und ihre Schwiegermutter und somit um recht unterschiedliche Arten von Beziehungen handelt, wären verschiedene Topics für die Kanten zu erwarten gewesen. Bei Kenntnis des Romans lässt sich andererseits sagen, dass Frau Bovary sehr häufig in Verbindung mit Emma und Karl auftritt, wenn sie beispielsweise für längere Besuche im Haus des Paars weilt. Das hat zur Folge, dass es zahlreiche gemeinsame Textstellen dieser drei Figuren gibt, was eine Erklärung für die übereinstimmenden Topic-Zuordnungen sein könnte.

Erfreulich ist das Topic für die Kante zwischen Emma und L‘Heureux: Topic 2 umfasst die Themen ‚Geld‘ und ‚Finanzielle Not‘. Das ist sehr passend, da die Figur L’Heureux ein nur scheinbar wohlgesonnener, skrupelloser und profitgieriger Händler ist, bei dem Emma häufig Dinge kauft, dabei Schulden macht und sich Geld leihen muss. Die Topic-Zuordnungen für die anderen Kanten lassen sich jedoch nicht so offensichtlich interpretieren.

Bei Betrachtung aller Visualisierungen lässt sich sagen, dass die Topics 37, 51 und 56 in sehr vielen Netzwerken vorkommen, wie das auch bei *Madame Bovary* der Fall ist. Das liegt daran, dass es sich dabei um die drei Topics mit der insgesamt

größten Wahrscheinlichkeit für das Korpus handelt. Außerdem kann es vorkommen, dass (fast) alle Kanten eines Netzwerks das gleiche Topic erhalten. Dies passiert vor allem dann, wenn ein einzelnes Topic für einen Roman eine besonders hohe Wahrscheinlichkeit aufweist. In einem solchen Fall bieten Topics als Kantenattribute kaum eine Möglichkeit zur Interpretation im Hinblick auf Relationen zwischen Figuren oder zur weiteren Analyse.



26 Beispielnetzwerke mit (fast) nur gleichen Kanten-Topics; Goethes Wahlverwandtschaften (links), Eichendorffs Ahnung und Gegenwart (Mitte), Heinses Hildegard von Hohenthal (rechts)

Vergleicht man die als Kantenattribute auftretenden Topics mit der Verteilung der Topics über die Romane als Ganzes, so stellt man fest, dass die für die Kanten ermittelten Topics im Großen und Ganzen die Topic-Verteilung des Romans widerspiegeln. Gibt es für den Roman ein dominierendes Topic, dessen Wahrscheinlichkeit deutlich höher ist, als bei den anderen Topics, erhalten höchstwahrscheinlich alle Kanten dieses Topic als Attribut. Gibt es mehrere Topics mit ähnlich hoher Wahrscheinlichkeit, werden diese als Kanteneigenschaften im Netzwerk zu finden sein. Folglich haben die Ergebnisse dieses Ansatzes im Vergleich zur Topic-Verteilung über die Romane als Ganzes leider keinen wesentlich höheren Informationsgehalt.

Außerdem fällt auf, dass die Zuordnung der Topics zu den Kanten oft nur auf einem kleinen Teil der Absätze, in denen die entsprechende Kante vorkommt, beruht. Ein Blick auf die vierte Spalte der in Abbildung 24 gezeigten Interaktionsliste zeigt, dass die Topic-Zuordnung beispielweise für die Kante zwischen Emma und Karl auf 23 von 248 Absätzen basiert. Auch bei anderen Kanten sind unter 10% der Absätze für die Auswahl entscheidend. In Anbetracht der Tatsache, dass für jeden Absatz 70 Topics theoretisch in Frage kommen, relativiert sich dieser Eindruck zwar etwas, es entstehen dennoch Zweifel, ob die Auswahl des Topics für eine Kante über die reine Häufigkeit eine sinnvolle Methode ist. Betrachtet man für die Kante zwischen Emma und Karl für alle 248 relevanten Absätze die Topics mit der höchsten Wahrscheinlichkeit und deren Häufigkeit, so stellt man fest, dass 50

verschiedene Topics vorkommen und die Häufigkeiten im oberen Bereich sehr nah beieinander liegen. Topic 56 kommt 23 Mal vor, Topic 37 an zweiter Stelle 22 Mal. Somit wird bei der einfachen Auswahl des häufigsten Topics eine nicht unbeträchtliche Datenmenge außer Acht gelassen. Eine denkbare Möglichkeit zur Abmilderung dieser Problematik ist, den Wahrscheinlichkeitswert zu speichern, wenn für jeden Absatz das Topic mit der höchsten Wahrscheinlichkeit ermittelt wird, und anschließend für jedes Topic die gesammelten Wahrscheinlichkeitswerte aufzusummieren. Auf diese Weise erhalten Topics, die zwar seltener sind, aber dafür höhere Wahrscheinlichkeiten aufweisen, insgesamt mehr Einfluss. Bei der Betrachtung der resultierenden Interaktionslisten und Visualisierungen zeigt sich jedoch, dass diese Vorgehensweise effektiv bis auf sehr wenige Ausnahmen die gleichen Ergebnisse liefert.

Eine alternative Herangehensweise wäre, für jede Kante die Häufigkeitsverteilung der zugeordneten Topics zu betrachten. Das würde jedoch eine Visualisierung und eine nachfolgende Analyse erheblich erschweren und die Entwicklung neuer Methoden erforderlich machen.

Der hier vorgestellte Ansatz zur Einbindung von Topics als Kantenattribute ist also durchaus problembehaftet. Aufgrund der Funktionsweise von Topic Modeling und der Netzwerkerstellung auf Basis von gemeinsamen Vorkommen von Figuren im gleichen Absatz liegt die Annahme nahe, dass die Idee besser funktioniert, wenn zwei Figuren durch eine klar abgegrenzte Relation verbunden sind, also häufig nur zu zweit und seltener in Verbindung mit anderen Figuren auftreten und sich ihr Umgang auf eine bestimmte Tätigkeit oder einen Lebensbereich fokussiert. Dies trifft beispielsweise auf die oben beschriebene Relation zwischen Emma und L‘Heureux zu. Bei Figuren mit intensiveren, vielschichtigeren Beziehungen ist diese Annahme eher problematisch, da ein einzelnes Topic dies kaum abbilden kann.

Zudem kann die beschriebene Vorgehensweise nur dann potentiell nützliche Informationen liefern, wenn unterschiedliche Relationen auch verschiedene Topics als Kantenattribute erhalten. Damit hängt sie stark vom berechneten LDA-Modell ab, welches wiederum vom Korpus abhängig ist. Ob die als Kantenattribute ermittelten Topics im Hinblick auf die Relation zwischen den zwei Figuren sinnvoll in-

terpretierbar sind, steht auf einem anderen Blatt. Klar ist allerdings, dass Netzwerke, in denen jede Kante das gleiche Topic hat, keinen Ansatzpunkt für weitere Untersuchungen bieten.

7 Fazit und Diskussion

Die in dieser Arbeit betrachtete Fragestellung beschäftigt sich mit der automatischen Erkennung von Ähnlichkeit zwischen den Figurenkonstellationen verschiedener Romane. Dazu wurden zunächst Figurennetzwerke automatisch aus den Romanen extrahiert und mit Hilfe von Visualisierungen gezeigt, dass diese die Figurenkonstellation weitgehend gut repräsentieren. Auf Basis der Figurennetzwerke wurden verschiedene Netzwerkmaße ermittelt und diese als Features zur Berechnung von Distanzen zwischen den Romanen verwendet. Obwohl diese Distanzen durchaus sinnvoll interpretiert werden konnten, konnte keine Verbindung zur manuell erstellten Evaluationsgrundlage nachgewiesen werden. Dies widerspricht den zuvor gehegten Erwartungen und lässt aufgrund der guten Interpretierbarkeit der Features und der berechneten Distanzen Zweifel an der Evaluationsgrundlage aufkommen. Ausgehend von der Annahme, dass die verschiedenen Arten von Beziehungen und bestimmte zwischenmenschliche Motive zwischen Figuren den Eindruck von Ähnlichkeit zwischen Figurenkonstellationen stark beeinflussen, wurde Topic Modeling verwendet, um derartige Informationen mit einzubinden. Durch die zusätzliche Verwendung der Topic-Verteilung für die Romane als Features konnte eine signifikant positive Korrelation zwischen den berechneten und den manuell ermittelten Distanzen festgestellt werden, die jedoch nur sehr schwach ausgeprägt ist. Das deutet dennoch darauf hin, dass die Evaluationsgrundlage unterbewusst doch relativ stark von den im Roman auftretenden Themen und Motiven beeinflusst ist und nährt somit weitere Zweifel daran, ob die Evaluationsgrundlage in ihrer vorliegenden Form geeignet ist, die in dieser Arbeit durchgeführten Experimente auszuwerten.

Tatsächlich gibt es einige Punkte, die im Zusammenhang mit der Erstellung der Evaluationsgrundlage als problematisch betrachtet werden können. Zum einen beruht die Einschätzung der Figurenkonstellation auf Zusammenfassungen. Obwohl diese Zusammenfassungen von Experten verfasst sind und aus einem qualitativ hochwertigen Literaturlexikon stammen, und bei der Zusammenstellung des

Korpus darauf geachtet wurde, dass die entsprechenden Zusammenfassungen von einer gewissen Länge sind, variieren sie dennoch in Umfang und Detailgehalt. Somit kann je nach Zusammenfassung ein besserer oder schlechterer Eindruck eines Romans und dessen Figurenkonstellation entstehen, wie das Beispiel von Marie von Ebner-Eschenbachs *Gemeindekind* in Abschnitt 5.3.2 zeigt. So können auch relevante Informationen verborgen bleiben, ohne dass es der Leser der Zusammenfassung merkt. Da es jedoch extrem zeitaufwändig und bei größeren Korpora gar nicht denkbar wäre, alle Romane komplett zu lesen, ist der Rückgriff auf Zusammenfassungen dennoch eine notwendige Alternative.

Zum anderen können Romane sehr vielseitig und unterschiedlich sein, wodurch es kaum möglich ist, feste Kriterien bei der Erstellung einer manuellen Distanzmatrix anzuwenden. Häufig lassen sich beim Vergleich zweier Romane sowohl ähnliche, als auch unähnliche Aspekte finden. Obwohl versucht wurde, sich auf den Vergleich der Figurenkonstellationen zu beschränken, kann der Eindruck von Ähnlichkeit trotzdem durch verschiedene Aspekte unterbewusst beeinflusst sein, beispielsweise durch vergleichbares Setting oder wiederkehrende zentrale Motive. Durch die Verwendung einer Skala für die Distanzen wurde zwar eine binäre Aufteilung in ‚ähnlich' und ‚unähnlich' vermieden, dennoch handelt es sich bei den manuell erstellten Distanzen um stark subjektiv geprägte Ad-hoc-Entscheidungen. Dies lässt vermuten, dass es durchaus auch zu nicht unerheblichen Abweichungen kommen würde, wenn eine zweite Person eine Distanzmatrix von Hand erstellen würde, da diese Aufgabe auch für Menschen schwierig ist.

Mögliche Alternativen, eine weniger stark subjektiv geprägte manuelle Distanzmatrix als direkte Evaluationsgrundlage zu erstellen, wären zum einen ein Crowdsourcing, in dem jeweils mehrere Personen die Ähnlichkeit der Figurenkonstellation von Romanen anhand der Zusammenfassungen bewerten. Einerseits könnte auf diese Weise das Inter-Annotatoren-Agreement berechnet werden, um die Schwierigkeit der Aufgabe besser einschätzen zu können. Außerdem könnten die durchschnittlichen Distanzen der verschiedenen Annotatoren verwendet werden, um subjektive Einflüsse zu vermindern. Andererseits ist ein solches Crowdsourcing natürlich zeitaufwändig und mit Kosten verbunden. Zudem müssten die Teilnehmer bei einer derartigen Aufgabe über eine gewisse literaturwissenschaftliche Kenntnis verfügen.

Ein weiterer Ansatz wäre eine umfassende Recherche in literaturwissenschaftlicher Sekundärliteratur, um festzustellen, welche Texte in der Forschung als ähnlich betrachtet werden. Dabei könnte eine Schwierigkeit darin liegen, Sekundärliteratur zu finden, die speziell auf den Bereich Figurenkonstellation fokussiert ist. Außerdem ist die klassische literaturwissenschaftliche Forschung stark am Einzelwerk interessiert und versucht eher, Besonderheiten von einem bestimmten Text in Abgrenzung zu anderen Texten herauszuarbeiten. Darüber hinaus ist literaturwissenschaftliche Fachliteratur größtenteils auf kanonisierte Werke beschränkt, sodass dieser Ansatz nicht mehr in Frage kommt, wenn das zu untersuchende Korpus auch unbekanntere Texte enthält.

Sofern die vorgestellten Untersuchungen auf größere Korpora übertragen werden sollen, sind beide Ansätze kritisch zu betrachten, da die Erstellung einer solchen Evaluationsgrundlage immer aufwändiger wird, je größer die Textsammlung ist und zudem zu allen Texten Zusammenfassungen bzw. Sekundärliteratur vorhanden sein müssten. Insgesamt lässt sich festhalten, dass für Experimente mit Figurennetzwerken, wie sie in dieser Arbeit betrachtet werden, eine alternative Art der Evaluation erforderlich ist, für die noch nicht der richtige Ansatz gefunden wurde. Hierbei wäre auch eine Form der indirekten Auswertung denkbar, bei der die Features und Methoden für eine Unterteilung von Romanen in nachprüfbare literaturwissenschaftliche Kategorien, wie beispielsweise Gattungen, oder zur Überprüfung literaturwissenschaftlicher Thesen herangezogen werden könnten.

Aufgrund der dargelegten Probleme ist es nicht möglich, anhand von Evaluationsergebnissen gezielt verschiedene methodische Details wie beispielsweise der Auswahl der Features oder des Distanzmaßes zu verändern. Dies gilt insbesondere für die Wahl der Parameter beim Topic Modeling, die folglich mittels einer näheren Betrachtung der entstehenden Topics festgelegt wurden.

Bei der Auswertung des Topic Modeling fiel auf, dass manche Topics für einzelne Texte eine besonders hohe Wahrscheinlichkeit aufwiesen, während sie in allen anderen Texten kaum vertreten waren. Ein solcher Effekt ist weniger hilfreich, wenn es darum geht, Ähnlichkeiten zwischen Romanen zu entdecken. Hier könnte der relativ kleine Umfang des Korpus eine Rolle spielen, da sich der Effekt in einer größeren Textmenge relativieren würde, weil dann die Wahrscheinlichkeit höher wäre, dass Topics in mehreren verschiedenen Texten auftreten.

Einige der entstandenen Topics bilden eine Reihe von Aspekten ab, die sich, wie in Abschnitt 6.2 beschrieben, hinsichtlich der Figuren eines Romans interpretieren lassen. Es gibt jedoch auch weitere Topics, die allgemeinere Konzepte repräsentieren und eher auf die Art und Weise hindeuten, wie Figuren, Orte oder Situationen in Romanen beschrieben werden. Diese Topics haben meist eine insgesamt recht hohe Wahrscheinlichkeit und kommen in den meisten Texten vor.

Im Hinblick auf den Versuch, Beziehungen zwischen Romanfiguren mittels Topic Modeling zu modellieren, lässt sich sagen, dass die Topics zu viele verschiedene Aspekte abbilden, um das leisten zu können. Zudem beschreibt die Wahrscheinlichkeitsverteilung der Topics über das Korpus eher die Romane als Ganzes. Gründe dafür könnten sein, dass feste Beziehungen zwischen Figuren, wie Verwandtschaft, Ehe oder Liebe, an wenigen Stellen eingeführt und dann nicht wiederholt direkt thematisiert werden, sondern über Verhalten, Handlungen und Rede der Figuren vermittelt werden. Andere Relationen, wie beispielsweise die Affäre zwischen Effi und Crampas, klingen nur an, wobei die letztliche Interpretation dem Leser überlassen wird. Solche Aspekte sind typisch für literarische Texte und generell eine Herausforderung für eine computergestützte Analyse.

Bezüglich Figurennetzwerken als solches kann festgehalten werden, dass diese die Figurenkonstellation von Romanen trotz der sehr einfachen Modellierung auf Basis von Figuren-Kookkurrenzen gut repräsentieren. In vereinzelten Fällen können Probleme entstehen, die stilistischen oder strukturellen Besonderheiten der entsprechenden Romane geschuldet sind. Bei der Evaluation stellte sich heraus, dass manche Netzwerke von der anhand der Zusammenfassung erwarteten Figurenkonstellation abweichen. Hier lässt sich jedoch nicht pauschal sagen, welche Darstellung den Roman besser repräsentiert. Dieser Umstand ist durchaus problematisch für die Idee, Figurennetzwerke anhand von Zusammenfassungen zu evaluieren, was sich sowohl bei der Evaluationsgrundlage als auch bei der Kategorisierung in Netzwerken mit und ohne einer eindeutig zentralen Hauptfigur zeigte. Da sich natürlich bereits geringe Unterschiede zwischen Figurennetzwerken in den extrahierten Features bemerkbar machen, decken die in dieser Arbeit generierten Figurennetzwerke den Merkmalsraum zwischen verschiedenen Extremfällen recht gleichmäßig ab. Da bereits in dem vorliegenden, relativ kleinen Korpus viele Zwischenstufen vorkommen, ist anzunehmen, dass dieser Effekt in größeren Korpora

noch stärker hervortritt. Dies könnte dafür sprechen, dass Autoren Figurenkonstellationen in ihren Werken recht frei gestalten, ohne sich zu stark an wiederkehrenden Typen zu orientieren, und somit Kafitz‘ Versuch einer Romantypologie anhand der Figurenkonstellation (vgl. Abschnitt 2) gewissermaßen relativieren. Hier müssten jedoch weitere Untersuchungen an umfangreicheren Korpora in Betracht gezogen werden.

8 Ausblick

Wie im vorhergehenden Abschnitt beschrieben, ist die Entwicklung einer alternativen Möglichkeit zur Evaluation von Experimenten mit Figurennetzwerken, wie sie auch in dieser Arbeit durchgeführt wurden, ein sehr zentraler Punkt für weitere Arbeiten in diesem Bereich. Eine solide Evaluationsgrundlage würde es ermöglichen, verschiedene Feature-Kombinationen und Distanzmaße zu vergleichen oder mit Hilfe von Learning-to-Rank-Verfahren eine passende Metrik anhand der Daten lernen zu lassen.

Darüber hinaus wäre eine Erkennung von eingeschobenen Passagen in Romanen, wie Briefen oder Binnenerzählungen, sowie von Vor- und Rückblenden, die möglicherweise in einer abweichenden Erzählperspektive verfasst sind, von großem Nutzen, sodass derartige strukturelle Besonderheiten bei der Erstellung von Figurennetzwerken berücksichtigt werden könnten. Ebenso wäre eine verbesserte Koreferenzauflösung für die Domäne literarischer Texte von Vorteil für die Erkennung der wichtigsten Romanfiguren und deren Darstellung als Figurennetzwerk. Die Einbindung von weiteren Figureneigenschaften wie dem Geschlecht, dem Alter oder dem gesellschaftlichen Stand und eine Erkennung der verschiedenen Typen von Beziehungen zwischen Figuren könnten zu einer umfassenderen Repräsentation der Figurenkonstellation beitragen und die Betrachtung neuer Fragestellungen ermöglichen.

Aufbauend auf die in dieser Arbeit betrachteten Netzwerke könnten dynamische Figurennetzwerke betrachtet werden, die die Entwicklung der Figurenkonstellation über den Verlauf eines Romans hinweg modellieren, erstellt und untersucht werden. Da die Figuren die Handlungsträger eines Romans sind, könnten dynamische Figurennetzwerke eventuell dazu beitragen, bestimmte Plot-Elemente in Romanen zu erfassen. Daran anknüpfend wäre es außerdem interessant, weitere

Aspekte, in denen Romane sich ähneln können, wie den Plot oder die stilistische Gestaltung mit computergestützten Methoden zu greifbar zu machen und somit mehrere Dimensionen von Ähnlichkeit zwischen literarischen Texten zusammenzubringen. In der Praxis könnten mit einem solchen System beispielsweise qualitative Leseempfehlungen generiert werden oder Romane abseits des literarischen Kanons erschlossen und mit ähnlichen kanonisierten Werken in Verbindung gebracht werden, wodurch sich neue Forschungsansätze ergeben könnten.

9 Literaturverzeichnis

Agarwal, Apoorv; Corvalan, Augusto; Jensen, Jacob; Rambow, Owen (2012). *Social Network Analysis of Alice in Wonderland.* Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. Montréal, Canada: Association for Computational Linguistics, S. 88–96.

Agarwal, Apoorv; Kotalwar, Anup; Rambow, Owen (2013a). *Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland.* Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, S. 1202–1208.

Agarwal, Apoorv; Kotalwar, Anup; Zheng, Jiehan; Rambow, Owen (2013b). *SINNET: Social Interaction Network Extractor from Text.* The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations. Nagoya, Japan: Asian Federation of Natural Language Processing, S. 33–36.

Arnold, Heinz L. (Hg.) (2009). *Kindlers Literatur Lexikon*. 3. Aufl. Stuttgart/Weimar: Verlag J.B. Metzler. Online verfügbar unter kll-online.de, zuletzt geprüft am 02.08.2016.

Blei, David M. (2012). *Introduction to Probabilistic Topic Models*. In: *Communications of the ACM* 55 (4), S. 77–84. DOI: 10.1145/2133806.2133826.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. (2003). *Latent Dirichlet Allocation*. In: *Journal of Machine Learning Research* (3), S. 993–1022.

Blevins, Cameron (2010). *Topic Modeling Martha Ballard's Diary*. Online verfügbar unter http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/, zuletzt geprüft am 21.07.2016.

Bonwit, Marianne (1948). *Effi Briest und Ihre Vorgängerinnen Emma Bovary und Nora Helmer*. In: *Monatshefte* 40 (8), S. 445–456. DOI: 10.2307/30164770.

Carr, Jon W. (2014). *A guide to the Mantel test for linguists*. Online verfügbar unter http://www.jonwcarr.net/blog/2014/9/19/a-guide-to-the-mantel-test-for-linguists, zuletzt geprüft am 21.07.2016.

Celikyilmaz, Asli; Hakkani-Tur, Dilek; He, Hua; Kondrak, Greg; Barbosa, Denilson (2010). *The Actor-Topic Model for Extracting Social Networks in Literary Narrative.* Proceedings of the NIPS 2010 Workshop Machine Learning for Social Computing. Whistler, Canada.

Coll Ardanuy, Mariona; Sporleder, Caroline (2014). *Structure-based Clustering of Novels.* Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL). Gothenburg, Sweden: Association for Computational Linguistics, S. 31–39.

Costa, Luciano. da F.; Rodrigues, Francisco. A.; Travieso, Gonzalo; Villas Boas, Paulino R. (2007). *Characterization of Complex Networks: A Survey of Measurements*. In: *Advances in Physics* 56 (1), S. 167–242. DOI: 10.1080/00018730601170527.

Craig, Hugh (2011). *Shakespeare's Vocabulary: Myth and Reality*. In: *Shakespeare Quarterly* 62 (1), S. 53–74. DOI: 10.1353/shq.2011.0002.

Culotta, Aron; Bekkerman, Ron; McCallum, Andrew (2004). *Extracting social networks and contact information from email and the Web.* First Conference on Email and Anti-Spam (CEAS).

Degering, Thomas (1978). *Das Verhältnis von Individuum und Gesellschaft in Fontanes "Effi Briest" und Flauberts "Madame Bovary"*. Bonn: Bouvier (Abhandlungen zur Kunst-, Musik- und Literaturwissenschaft, 274).

Dethloff, Uwe (2000). *Emma Bovary und Effi Briest. Überlegungen zur Entwicklung des Weiblichkeitsbildes in der Moderne*. In: Hanna Delf von Wolzogen (Hg.): Theodor Fontane - am Ende des Jahrhunderts. Würzburg: Königshausen & Neumann, S. 123–134.

Eder, Maciej; Rybicki, Jan; Kestemont, Mike (2016). *Stylometry with R: A Package for Computational Text Analysis*. In: *The R Journal* 2016 (1).

Elson, David; Dames, Nicholas; McKeown, Kathleen (2010). *Extracting Social Networks from Literary Fiction.* Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, S. 138–147.

Elson, David K.; McKeown, Kathleen (2010). *Automatic Attribution of Quoted Speech in Literary Narrative.* Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta.

Gil, Sebastian; Kuenzel, Laney; Suen, Caroline (2011). *Extraction and Analysis of Character Interaction Networks from Plays and Movies*. Online verfügbar unter http://web.stanford.edu/~cysuen/projects/GilKuenzelSuen-Character-InteractionNetworks.pdf, zuletzt geprüft am 21.07.2016.

Griffiths, Thomas L.; Steyvers, Mark (2004). *Finding scientific topics.* Proceedings of the National Academy of Sciences of the United States of America, 101 Suppl 1, S. 5228–5235.

Gruzd, Anatoliy; Haythornthwaite, Caroline (2008). *Automated Discovery and Analysis of Social Networks from Threaded Discussions.* International Network of Social Network Analysis Conference. St. Pete Beach, Florida.

Hettinger, Lena; Jannidis, Fotis; Reger, Isabella; Hotho, Andreas (2016). *Classification of Literary Subgenres.* DHd Tagung 2016. Leipzig.

Jannidis, Fotis; Krug, Markus; Reger, Isabella; Toepfer, Martin; Weimer, Lukas; Puppe, Frank (2015). *Automatische Erkennung von Figuren in deutschsprachigen Romanen.* DHd Tagung 2015. Graz.

Jing, Hongyan; Kambhatla, Nanda; Roukos, Salim (2007). *Extracting Social Networks and Biographical Facts From Conversational Speech Transcripts.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, S. 1040–1047.

Jockers, Matthew L. (2013). *Macroanalysis*. Urbana: University of Illinois Press (Topics in the digital humanities).

Jockers, Matthew L.; Mimno, David (2013). *Significant themes in 19th-century literature*. In: *Poetics* 41 (6), S. 750–769. DOI: 10.1016/j.poetic.2013.08.005.

Juola, Patrick (2013). *How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling*. In: *Scientific American*. Online verfügbar unter: http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/, zuletzt geprüft am 01.08.2016.

Kafitz, Dieter (1978). *Figurenkonstellation als Mittel der Wirklichkeitserfassung*. Kronberg/Ts.: Athenäum Verlag.

Keogh, Eamonn; Mueen, Abdullah (2011). *Curse of Dimensionality*. In: Claude Sammut und Geoffrey I. Webb (Hg.): Encyclopedia of machine learning. New York: Springer (Springer reference), S. 257–258.

Krug, Markus; Jannidis, Fotis; Reger, Isabella; Weimer, Lukas; Macharowsky, Luisa; Puppe, Frank (2016a). *Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts: Methoden zur Bestimmung des Sprechers und des Angesprochenen.* DHd Tagung 2016. Leipzig.

Krug, Markus; Jannidis, Fotis; Reger, Isabella; Weimer, Lukas; Macharowsky, Luisa; Puppe, Frank (2016b). *Comparison of Methods for the Identification of Main Characters in German Novels.* DH Conference 2016. Krakow.

Krug, Markus; Puppe, Frank; Jannidis, Fotis; Macharowsky, Luisa; Reger, Isabella; Weimar, Lukas (2015). *Rule-based Coreference Resolution in German Historic Novels.* Proceedings of the 4th Workshop on Computational Linguistics for Literature (CLfL). Denver, Colorado, USA: Association for Computational Linguistics, S. 98–104.

Mantel, Nathan (1967). *The Detection of Disease Clustering and a Generalized Regression Approach*. In: *Cancer Research* 27 (2), S. 209–220.

Mimno, David (2012). *Computational Historiography: Data Mining in a Century of Classics Journals*. In: *Journal on Computing and Cultural Heritage* 5 (1), S. 1–19. DOI: 10.1145/2160165.2160168.

Moretti, Franco (2011). *Network Theory, Plot Analysis (*Stanford Literary Lab Pamphlets, 2). Online verfügbar unter https://litlab.stanford.edu/LiteraryLab-Pamphlet2.pdf, zuletzt geprüft am 21.07.2016.

Newman, Mark E. J. (2010). *Networks. An Introduction*. Oxford: Oxford University Press.

Park, Gyeong-Mi; Kim, Sung-Hwan; Hwang, Hye-Ryeon; Cho, Hwan-Gue (2013). *Complex System Analysis of Social Networks Extracted from Literary Fictions*. In: *International Journal of Machine Learning and Computing (IJMLC)*, S. 107–111. DOI: 10.7763/IJMLC.2013.V3.282.

Pfeffer, Jürgen (2010). *Visualisierung sozialer Netzwerke*. In: Christian Stegbauer (Hg.): Netzwerkanalyse und Netzwerktheorie. Wiesbaden: VS Verlag, S. 227–238.

Raschka, Sebastian; Olson, Randal S. (2015). *Python machine learning*. Birmingham: Packt Publishing (Community experience distilled).

Schöch, Christof (2015). *Topic Modeling French Crime Fiction.* DH Conference 2015. Sydney.

Schöch, Christof; Henny, Ulrike; Calvo, José; Schlör, Daniel; Popp, Stefanie (2016). *Topic, Genre, Text Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880-1930).* DHd Tagung 2016. Leipzig.

Trilcke, Peer (2013). *Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft*. In: Philip Ajouri (Hg.): Empirie in der Literaturwissenschaft. Münster: Mentis-Verl. (Poetogenesis, 8), S. 201–247.

Trilcke, Peer; Fischer, Frank; Kampkaspar, Dario (2015). *Digitale Netzwerkanalyse dramatischer Texte.* DHd Tagung 2015. Graz.

Ward, Joe H. (1963). *Hierarchical Grouping to Optimize an Objective Function*. In: *Journal of the American Statistical Association* 58 (301), S. 236–244.

Weimar, Klaus; Fricke, Harald; Müller, Jan-Dirk (2010). *Reallexikon der deutschen Literaturwissenschaft*. Berlin: de Gruyter.

Woloch, Alex (2003). *The One vs. the Many*. Princeton: Princeton University Press.

10 Anhang

10.1 Korpus

Auflistung aller im Korpus enthaltenen Werke, wie in Abschnitt 4.1 beschrieben.

Titel	**Autor**	**Jahr**
Barfüßele	Auerbach, Berthold	1856
Der Idiot	Dostojewski, Fjodor Michailowitsch	1868
Das Gemeindekind	Ebner-Eschenbach, Marie von	1887
Ahnung und Gegenwart	Eichendorff, Joseph von	1815
Herr Lorenz Stark	Engel, Johann Jakob	1795
Madame Bovary	Flaubert, Gustave	1857
Effi Briest	Fontane, Theodor	1894
Irrungen, Wirrungen	Fontane, Theodor	1887
Mathilde Möhring	Fontane, Theodor	1906
Stine	Fontane, Theodor	1890
Der Pojaz	Franzos, Karl Emil	1905
Schloß Hubertus	Ganghofer, Ludwig	1895
Die Wahlverwandtschaften	Goethe, Johann Wolfgang	1809
Gräfin Faustine	Hahn-Hahn, Ida Gräfin von	1840
Lichtenstein	Hauff, Wilhelm	1826
Einhart der Lächler	Hauptmann, Carl	1915
Hildegard von Hohenthal	Heinse, Wilhelm	1795
Die unsichtbare Loge	Jean Paul	1793
Der grüne Heinrich [Erste Fassung]	Keller, Gottfried	1854
Der Sonnenwirt	Kurz, Hermann	1854
Zwischen Himmel und Erde	Ludwig, Otto	1856
Goldelse	Marlitt, Eugenie	1866
Siegwart. Eine Klostergeschichte	Miller, Johann Martin	1776

Maler Nolten	Mörike, Eduard	1832
Das Odfeld	Raabe, Wilhelm	1888
Ekkehard	Scheffel, Joseph Viktor von	1855
Florentin	Schlegel, Dorothea	1801
Therese	Schnitzler, Arthur	1928
Heidis Lehr- und Wanderjahre	Spyri, Johanna	1880
Witiko	Stifter, Adalbert	1865
Der Schimmelreiter	Storm, Theodor	1888
Frau Sorge	Sudermann, Hermann	1887
Anna Karenina	Tolstoj, Lev Nikolaevic	1878
Das Gänsemännchen	Wassermann, Jakob	1915
Hermann und Ulrike	Wezel, Johann Karl	1780

10.2 Verwendete NLP-Komponenten

Bei der Generierung des tabellarischen Formats, wie in Abschnitt 5.1 beschrieben, kamen folgende Natural-Language-Processing-Komponenten zum Einsatz.

Task	**Komponente**
Tokenisierung	OpenNLP[27] Tokenizer
Satzerkennung	OpenNLP Sentence Splitter
Paragraphenerkennung	Anhand einfacher Zeilenumbrüche
Part-of-Speech-Tagging	TreeTagger[28]
Morphologie	RFTagger[29], TIGER Morph[30], Mate Tools[31]
Lemmatisierung	TreeTagger
Chunking	TreeTagger
Dependency Parsing	Mate Tools

[27] https://opennlp.apache.org.

[28] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger.

[29] http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger.

[30] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.en.html.

[31] https://code.google.com/archive/p/mate-tools.

Named Entity Recognition	NER nach Jannidis et al. (2015)
Coreference Resolution	CR nach Krug et al. (2015)

10.3 Evaluationsgrundlage

Die verwendete Evaluationsgrundlage wie in 4.2 beschrieben, aus Platzgründen in mehrere Teile aufgeteilt. Außerdem ist die zugehörige Excel-Tabelle auf dem beiliegenden USB-Stick zu finden.

	auerbach_ba	dostojewski_	ebner-escher	eichendorff_	engel_lorenz	flaubert_bov	fontane_effi.j	fontane_irru
auerbach_barfuessele.jtf	0	4	1	3	3	4	4	3
dostojewski_idiot.jtf	4	0	4	4	4	2	4	3
ebner-eschenbach_gemeinde	1	4	0	3	3	4	4	3
eichendorff_ahnung-gegenw	3	4	3	0	4	3	3	3
engel_lorenz-stark.jtf	3	4	3	4	0	4	4	4
flaubert_bovary.jtf	4	2	4	3	4	0	1	3
fontane_effi.jtf	4	4	4	3	4	1	0	3
fontane_irrungen.jtf	3	3	3	3	4	3	3	0
fontane_mathilde.jtf	3	4	2	4	3	3	3	3
fontane_stine.jtf	3	4	3	4	4	3	3	1
franzos_pojaz.jtf	4	4	4	3	2	4	3	4
ganghofer_hubertus.jtf	3	4	4	4	2	4	4	4
goethe_wahlverwandtschaft	4	2	4	4	4	3	3	3
hahn-hahn_faustine.jtf	3	3	4	2	4	2	3	3
hauff_lichtenstein.jtf	4	4	4	4	4	4	4	4
hauptmann_einhart.jtf	4	4	4	2	3	3	4	4
heinse_hildegard.jtf	3	4	3	2	3	3	3	2
jean-paul_loge.jtf	4	4	4	2	4	4	3	4
keller_heinrich.jtf	3	4	4	3	4	3	4	4
kurz_sonnenwirt.jtf	3	4	2	3	2	4	4	2
ludwig_himmel-erde.jtf	4	4	3	4	3	3	3	3
marlitt_goldelse.jtf	2	4	4	3	3	4	4	4
miller_siegwart.jtf	3	4	4	2	3	3	3	4
moerike_nolten.jtf	4	3	4	2	3	3	3	4
raabe_odfeld.jtf	4	4	4	4	4	4	4	4
scheffel_ekkehard.jtf	4	4	3	3	3	4	4	2
schlegel_florentin.jtf	4	3	4	2	3	3	2	3
schnitzler_therese.jtf	3	4	4	3	4	3	3	4
spyri_heidi.jtf	2	4	2	3	3	4	4	4
stifter_witiko.jtf	4	4	4	4	4	4	4	4
storm_schimmelreiter.jtf	4	3	3	4	3	4	3	3
sudermann_frau-sorge.jtf	3	4	3	3	2	4	4	4
tolstoj_karenina.jtf	4	3	4	4	4	1	2	3
wassermann_gaensemaennc	4	4	4	3	3	3	3	3
wezel_hermann-ulrike.jtf	3	4	3	4	4	3	4	1

	fontane_mat	fontane_stin	franzos_poja	ganghofer_h	goethe_wahl	hahn-hahn_f	hauff_lichten	hauptmann_
auerbach_barfuessele.jtf	3	3	4	3	4	3	4	4
dostojewski_idiot.jtf	4	4	4	4	2	3	4	4
ebner-eschenbach_gemeinde	2	3	4	4	4	4	4	4
eichendorff_ahnung-gegenw	4	4	3	4	4	2	4	2
engel_lorenz-stark.jtf	3	4	2	2	4	4	4	3
flaubert_bovary.jtf	3	3	4	4	3	2	4	3
fontane_effi.jtf	3	3	3	4	3	3	4	4
fontane_irrungen.jtf	3	1	4	4	3	3	4	4
fontane_mathilde.jtf	0	3	3	4	4	3	4	3
fontane_stine.jtf	3	0	4	4	3	3	4	4
franzos_pojaz.jtf	3	4	0	4	4	2	4	2
ganghofer_hubertus.jtf	4	4	4	0	4	4	4	4
goethe_wahlverwandtschaft	4	3	4	4	0	3	4	4
hahn-hahn_faustine.jtf	3	3	2	4	3	0	4	2
hauff_lichtenstein.jtf	4	4	4	4	4	4	0	4
hauptmann_einhart.jtf	3	4	2	4	4	2	4	0
heinse_hildegard.jtf	3	2	3	4	3	3	4	2
jean-paul_loge.jtf	4	3	3	4	4	2	4	2
keller_heinrich.jtf	3	4	1	4	3	1	4	1
kurz_sonnenwirt.jtf	3	2	3	4	4	3	4	3
ludwig_himmel-erde.jtf	4	3	4	4	3	4	4	4
marlitt_goldelse.jtf	3	4	3	3	4	3	4	3
miller_siegwart.jtf	4	4	3	4	3	3	4	3
moerike_nolten.jtf	4	3	3	4	3	1	4	2
raabe_odfeld.jtf	4	4	4	4	4	4	1	4
scheffel_ekkehard.jtf	2	2	3	4	4	3	4	3
schlegel_florentin.jtf	4	3	3	4	3	3	4	3
schnitzler_therese.jtf	4	4	2	4	4	2	4	1
spyri_heidi.jtf	4	4	3	4	4	4	4	3
stifter_witiko.jtf	4	4	4	4	4	4	2	4
storm_schimmelreiter.jtf	3	3	2	4	3	4	4	4
sudermann_frau-sorge.jtf	4	4	3	3	4	3	4	3
tolstoj_karenina.jtf	4	3	3	4	3	3	4	3
wassermann_gaensemaennc	4	4	3	4	3	2	4	1
wezel_hermann-ulrike.jtf	3	2	4	3	4	3	4	4

	heinse_hilde	jean-paul_lo	keller_heinri	kurz_sonnen	ludwig_himn	marlitt_golde	miller_siegw	moerike_nolt
auerbach_barfuessele.jtf	3	4	3	3	4	2	3	4
dostojewski_idiot.jtf	4	4	4	4	4	4	4	3
ebner-eschenbach_gemeinde	3	4	4	2	3	4	4	4
eichendorff_ahnung-gegenw	2	2	3	3	4	3	2	2
engel_lorenz-stark.jtf	3	4	4	2	3	3	3	3
flaubert_bovary.jtf	3	4	3	4	3	4	3	3
fontane_effi.jtf	3	3	4	4	3	4	3	3
fontane_irrungen.jtf	2	4	4	2	3	4	4	4
fontane_mathilde.jtf	3	4	3	3	4	3	4	4
fontane_stine.jtf	2	3	4	2	3	4	4	3
franzos_pojaz.jtf	3	3	1	3	4	3	3	3
ganghofer_hubertus.jtf	4	4	4	4	4	3	4	4
goethe_wahlverwandtschaft	3	4	3	4	3	4	3	3
hahn-hahn_faustine.jtf	3	2	1	3	4	3	3	1
hauff_lichtenstein.jtf	4	4	4	4	4	4	4	4
hauptmann_einhart.jtf	2	2	1	3	4	3	3	2
heinse_hildegard.jtf	0	3	2	3	4	2	4	3
jean-paul_loge.jtf	3	0	2	2	4	3	3	2
keller_heinrich.jtf	2	2	0	3	4	3	3	2
kurz_sonnenwirt.jtf	3	2	3	0	3	3	3	3
ludwig_himmel-erde.jtf	4	4	4	3	0	4	3	3
marlitt_goldelse.jtf	2	3	3	3	4	0	3	3
miller_siegwart.jtf	4	3	3	3	3	3	0	2
moerike_nolten.jtf	3	2	2	3	3	3	2	0
raabe_odfeld.jtf	4	4	4	4	4	4	4	4
scheffel_ekkehard.jtf	3	3	3	3	4	3	3	3
schlegel_florentin.jtf	3	3	3	3	4	3	3	3
schnitzler_therese.jtf	2	3	2	3	3	4	3	2
spyri_heidi.jtf	3	4	3	3	4	4	4	4
stifter_witiko.jtf	4	4	4	4	4	4	4	4
storm_schimmelreiter.jtf	4	4	4	2	3	4	3	4
sudermann_frau-sorge.jtf	2	3	3	3	3	2	3	3
tolstoj_karenina.jtf	3	4	4	4	4	4	4	4
wassermann_gaensemaennc	2	3	1	3	3	3	3	2
wezel_hermann-ulrike.jtf	3	4	4	3	4	2	4	4

	raabe_odfeld	scheffel_ekke	schlegel_flor	schnitzler_th	spyri_heidi.jt	stifter_witik	storm_schim	sudermann_f	tolstoj_karen	wassermann	wezel_herma
auerbach_barfuessele.jtf	4	4	4	3	2	4	4	3	4	4	3
dostojewski_idiot.jtf	4	4	3	4	4	4	3	4	3	4	4
ebner-eschenbach_gemeinde	4	3	4	4	2	4	3	3	4	4	3
eichendorff_ahnung-gegenw	4	3	2	3	3	4	4	3	4	3	4
engel_lorenz-stark.jtf	4	3	3	4	3	4	3	2	4	3	4
flaubert_bovary.jtf	4	4	3	3	4	4	4	4	1	3	3
fontane_effi.jtf	4	4	2	3	4	4	3	4	2	3	4
fontane_irrungen.jtf	4	2	3	4	4	4	3	4	3	3	1
fontane_mathilde.jtf	4	2	4	4	4	4	3	4	4	4	3
fontane_stine.jtf	4	2	3	4	4	4	3	4	3	4	2
franzos_pojaz.jtf	4	3	3	2	3	4	2	3	3	3	4
ganghofer_hubertus.jtf	4	4	4	4	4	4	4	3	4	4	3
goethe_wahlverwandtschaft	4	4	3	4	4	4	3	4	3	3	4
hahn-hahn_faustine.jtf	4	3	3	2	4	4	4	3	3	2	3
hauff_lichtenstein.jtf	1	4	4	4	4	2	4	4	4	4	4
hauptmann_einhart.jtf	4	3	3	1	3	4	4	3	3	1	4
heinse_hildegard.jtf	4	3	3	2	3	4	4	2	3	2	3
jean-paul_loge.jtf	4	3	3	3	4	4	4	3	4	3	4
keller_heinrich.jtf	4	3	3	2	3	4	4	3	4	1	4
kurz_sonnenwirt.jtf	4	3	3	3	3	4	2	3	4	3	3
ludwig_himmel-erde.jtf	4	4	4	3	4	4	3	3	4	3	4
marlitt_goldelse.jtf	4	3	3	4	4	4	4	2	4	3	2
miller_siegwart.jtf	4	3	3	3	4	4	3	3	4	3	4
moerike_nolten.jtf	4	3	3	2	4	4	4	3	4	2	4
raabe_odfeld.jtf	0	4	4	4	4	2	4	4	4	4	4
scheffel_ekkehard.jtf	4	0	2	3	4	3	3	3	4	3	3
schlegel_florentin.jtf	4	2	0	3	3	4	4	3	3	3	4
schnitzler_therese.jtf	4	3	3	0	4	4	4	3	3	2	4
spyri_heidi.jtf	4	4	3	4	0	4	4	3	4	3	4
stifter_witiko.jtf	2	3	4	4	4	0	4	3	4	4	4
storm_schimmelreiter.jtf	4	3	4	4	4	4	0	3	4	4	4
sudermann_frau-sorge.jtf	4	3	3	3	3	3	3	0	4	2	3
tolstoj_karenina.jtf	4	4	3	3	4	4	4	4	0	4	4
wassermann_gaensemaennc	4	3	3	2	3	4	4	2	4	0	4
wezel_hermann-ulrike.jtf	4	3	4	4	4	4	4	3	4	4	0

11 Eigenständigkeitserklärung

Ich versichere, dass ich die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sämtliche wörtlichen oder sinngemäßen Übernahmen und Zitate sind kenntlich gemacht und nachgewiesen.

Ferner versichere ich, dass das Thema dieser Arbeit nicht identisch ist mit dem Thema einer von mir bereits für eine andere Prüfung eingereichten Arbeit.

Ich erkläre weiterhin, dass ich die Arbeit nicht bereits an einer anderen Hochschule als Prüfungsleistung eingereicht habe.

Datum, Unterschrift