DISS. ETH NO. 28703

NOVELLA NETWORKS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES (Dr. sc. ETH Zurich)

presented by

SIMON PÄPCKE

MSc ETH Mathematics, ETH Zurich

born on 19.09.1990

accepted on the recommendation of

Prof. Dr. Ulrik Brandes
Prof. Dr. Thomas Weitin
Prof. Dr. Nils Reiter

2022

Abstract

This thesis deals with network science approaches to literary corpus analysis. In an at-scale approach, it uses intermediate resolution levels to combine the advantages of both hermeneutical and statistical tools in a mixed-methods approach to release synergies.

The *Deutsche Novellenschatz*, a collection of 19th century German language novellas, will be the starting point of the analyses. The texts where edited and published by Paul Heyse and Hermann Kurz between 1871 and 1876 in 24 volumes. The corpus was digitized by Weitin in 2016 and is part of the *Deutsche Textarchiv* (German text archive). Investigating structural properties of this corpus will lead to analyses that help to understand whether these properties resonate with the intentions of the editors in respect of quality and delimitation for the novella genre. A key tool in the at-scale approach will be network analysis to consider different levels of examination from single texts to the entire corpus.

More concrete, the thesis can be split into three parts. Each part takes a certain viewpoint to disentangle the object of study. First, we try to validate the underlying novella theory by a reading experiment. The editors claim that a novella is characterized by the fact that it can be summarized within a few sentences. The presented study shows that this can only hold under certain circumstances. While it is possible that the claim was made as a general principal, we converted their recommended procedure into a hypothesis. In that way, we tested the summarizability of Hieronymus Lorm's novella *Ein adeliges Fräulein*. As a result, we observed that only frequent readers agree on content and ordering of the summaries. In the second research project, we analyze the entire corpus with stylometric distance measures. The goal of this is twofold. On the one hand, it makes it possible to gain new insights into the corpus. On the other hand, we could use the *Novellenschatz* as a prime example of a closed corpus. As such, it is used to gain a deeper understanding about the implications and effects towards the application of the underlying methods. Based on this, we closer analyzed the resulting subgroups to deliver meaningful results on the subject matter.

The final part of the thesis is dedicated to the creation and examination of character networks. Therefore, we construct these networks in a semi-automated approach from the novellas. These networks are special in manifold ways. First, we do not only consider character names but also enrich our data set by also taking into account synonyms. In this way, we can assure that each character co-occurrence has an underlying contextual interaction. Moreover, it makes it less likely to miss out character interactions. This is important since we do use this information to create two-mode networks. Here, the links are present between characters and paragraphs to generate a dynamic plot of the interactions. These networks are then studied for structural (dis-)similarities. In particular, we test again for the intention of the editors on how the strict form of the novella gives rise to a specific character constellation network. For that reason, we employ a measure that uses the spectra of the adjacency matrix to convert structural similarities in networks into distances. As a result, we show that the distances for the novellas are small in general. However, on a relative scale, we can see differences between the network types. Similar to the second project, we

analyze and classify the resulting clusters and subgroups.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Anwendung von Netzwerkwissenschaften in der literarischen Korpusanalyse. Dabei werden in einem skalierten Vorgehen verschiedene Auflösungsstufen genutzt, um hermeneutische und statistische Methoden im Rahmen eines *mixed-methods* Ansatzes bestmöglich zu verbinden und Synergien nutzbar zu machen.
Als Ausgangspunkt dient dabei der *Deutsche Novellenschatz*, eine Sammlung von deutschsprachigen Novellen des 19. Jahrhunderts, die von Paul Heyse und Hermann Kurz zwischen 1871 und 1876 in 24 Bänden veröffentlicht wurde. Diese wurden 2017 von Weitin digitalisiert und sind Teil des *Deutschen Textarchivs*. Es wird geprüft, ob sich die im Vorwort des *Novellenschatzes* von den Herausgebern angegebenen Qualitäts- und Abgrenzungsabsichten bezüglich dessen, was gemäss ihrer Einschätzung eine gute Novelle ausmacht, validieren lassen, indem diese in den strukturellen Eigenschaften des Korpus widerhallen. Dabei wird der Netzwerkanalyse eine zentrale Rolle zukommen, um Gemeinsamkeiten und Unterschiede zwischen und innerhalb von Texten zu bestimmen.
Konkret lässt sich die Dissertation in drei Bereiche unterteilen. Diese nehmen jeweils einen konkreten Blickwinkel ein, um den Untersuchungsgegenstand zu erfassen. Zunächst soll die vorliegende Novellentheorie anhand eines Leseexperiments bestätigt werden. Die vorliegende Studie zeigt auf, dass die Behauptung von Heyse und Kurz, eine Novelle sei dadurch gekennzeichnet, dass sie sich in wenigen Sätzen zusammenfassen lasse, nur unter gewissen Voraussetzungen als haltbar angesehen werden kann. Während es sich hierbei vielleicht nur um ein allgemeines Prinzip gehandelt haben mag, haben wir die von ihnen mitgegebene Handlungsempfehlung in eine Hypothese umformuliert. Auf diese Weise wurde die Zusammenfassbarkeit der Novelle *Ein adeliges Fräulein* von Hieronymus Lorm untersucht. Es stellt sich heraus, dass insbesondere häufige Leser:innen eine gemeinsame Vorstellung von Inhalt und Reihenfolge beim Zusammenfassen des Textes haben.
In einem zweiten Projekt wird der gesamte Korpus mit stylometrischen Distanzmassen untersucht. Dabei werden zwei Ziele gleichzeitig verfolgt. Zum einen werden neue Einblicke in den Korpus gewonnen, andererseits wird der *Novellenschatz* als Paradebeispiel eines in sich abgeschlossenen Korpus angesehen. Als solcher wird er genutzt, um ein tieferes Verständnis im Hinblick auf die Anwendung der genutzten Methoden zu bekommen. Darauf aufbauend werden sich hieraus ergebende Untergruppen von Texten im Korpus näher untersucht und erläutert.
Der letzte Teil der Arbeit widmet sich dem Erstellen und Untersuchen einer grossen Kollektion von Figurennetzwerken. Dabei werden die Figurennetzwerke aus den einzelnen Novellen in einem semiautomatischen Verfahren erstellt. Diese sind in verschiedenerlei Hinsicht besonders. Zum einen werden für die verschiedenen Figuren nicht nur ihre Namen, sondern auch ihre Synonyme verwendet. Damit wird sichergestellt, dass es wahrscheinlich ist, dass den Verbindungen in den Netzwerken tatsächliche Interaktionen zugrunde liegen. Ausserdem werden so möglichst wenige Interaktionen übersehen. Dies ist zentral, da zum anderen die Netzwerke als sogenannte *two-mode* Netzwerke eingelesen werden. Hierbei bilden die Figuren einerseits und die Textparagrafen andererseits die Knoten des Netzw-

erkes, sodass ein dynamisches Bild der Interaktionen entsteht. Die so erstellten Netzwerke werden wiederum auf strukturelle Ähnlichkeiten und Unterschiede untersucht. Insbesondere wird anhand der Position der Herausgeber des *Novellenschatzes* überprüft, inwieweit die strikte Form der Novelle die Art der gefundenen Figurennetzwerke einschränkt. Dabei wird ein Mass verwendet, welches strukturelle Ähnlichkeiten in Netzwerken anhand der Spektra ihrer Adjazenzmatrix in Distanzen umwandelt. Hierbei ergibt sich ein Bild, welches sowohl aufzeigt, dass die untersuchten Novellennetzwerke global gesehen kleine Abstände aufweisen, jedoch innerhalb dieser sich grössere Unterschiede bezüglich der relativen Distanzen aufzeigen lassen. Dieses führt zu Clustern und Untergruppen, die analog zum zweiten Projekt analysiert und klassifiziert werden.

Acknowledgments

While a large part of this theses deals with character networks in novellas it took a complete character network around me to write it.

First, I want to thank my supervisor Ulrik Brandes. He gave me the chance to work on this interdisciplinary project and has supported me with great advice and helpful suggestions.

Likewise, I am grateful to Thomas Weitin. After each visit in Darmstadt I came back with a new enthusiasm for my work. It helped me a lot in understanding the relevance of my subject and gave me a different perspective on my own work. Moreover, I want to thank the complete group in Darmstadt, especially Katharina Herget and Anastasia Glawion, for great scientific and non-scientific discussions.

I also thank Nils Reiter for agreeing to review this thesis.

I am thankful to the whole Social Networks Lab at ETH for insightful discussions, their time to listen to a lot of trial presentations over the years, and all the joy that you brought into the office. Thanks to my office mates over the years Julian Müller, Felix Schönenberger, Alvaro Uzaheta, Meher Pindiprolu, and Annina Stahl for all the discussions, ideas and support; as well as to Denise Weber who has always been a helping hand and brought the social part in the Social Networks group.

I would like to thank my loving parents and my brother for everything. A special thanks goes to my wife Diana who is always by my side and supported me especially in the hard last months of writing this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advantages in computer text analysis made it possible to study larger collections of texts, i.e., corpora, simultaneously and in a comparative manner. Many studies first prepare a corpus that is a representative sample of possible texts with respect to some common property (e.g., Trilcke et al., 2013 look at dramatic texts from Greek tragedies to 20th-century plays in different languages, Nalisnick and Baird, 2013 consider specifically the plays of Shakespeare, while Diederich et al., 2003 used texts from the Berliner Zeitung published between 12/1998 and 02/1999). Hence, the research itself is part of some formative process that canonizes texts for its own purpose. However, the results of these studies still depend on the underlying corpus. This can be delicate if the analysis is directly affected by the choice of the elements in the corpus as for example in authorship attribution where the removal of texts with non-allocatable authors improve the overall performance.

In contrast, the center of all analyses in this thesis is a corpus of 86 German language novellas from the 19th century edited and published by Paul Heyse and Hermann Kurz as a series called *Deutscher Novellenschatz*. Therefore, the corpus generation is out of the researcher's hand which gives rise to the possibility of employing literary theory proposed by the editors of the corpus themselves to derive verifiable research questions. The size of the corpus gives way for a dissection on an intermediate level where the strengths of both classical hermeneutical and computational methods can interlink to disentangle the connections within the corpus and answer research questions that are based on the way the corpus was claimed to be compiled.

1.1 Theoretical background

This work wants to contribute to the thriving field of digital humanities. Therefore, a short introduction to the field as well as to the key concepts that are used to investigate our corpus is given. These concepts are corpus analysis, stylometry, and literary network science.

1.1.1 Digital Humanities

Digital humanities is a relatively young scientific discipline that uses methods from computer science to answer research questions in the humanities. The term *digital humanities* itself first came up in 2001 (Berry, 2017, p. 10) but one can argue that the idea of using computers in literary studies and other fields in the humanities dates back to the early years of computer science or even before that. Many nowadays scholars (e.g. Schreibman et al., 2004; Jannidis et al., 2017; Berry, 2017) see the founding myth[1] of the digital humanities in 1949 when the Jesuit Roberto Busa wrote his dissertation on the use of the presence term[2] in the work of Thomas Aquinas and convinced IBM to use their equipment (resulting in the *Index Thomisticus* by Busa, 1973). Even though this is most likely the first use of software in a humanities context, Le Deuff (2018, p. 25) looks on the approach of Busa as an idea already employed by Jean Hautfuney (c. 1320) to index the multi-volume work *Speculum historiale* by Vincent de Beauvais (Paulmier-Foucart, 1980) that was adapted to the technical capabilities of the first half of the 20th century. In the same way, Le Deuff identified several genealogies for different fields of the digital humanities from antiquity up to the present day (Le Deuff, 2018, p. 26). And also Busa himself acknowledges that "although some say I am the pioneer of the computers in the humanities, such a title needs a good deal of nuancing ... isn't it true that all new ideas arise out of a milieu when ripe, rather than from any one individual?" (Busa, 1980, p. 84).

Aside from these earlier tendencies it is still quite noticeable that digital humanities has, unlike other disciplines, a relatively clear starting point. Moreover, there is no doubt that the fast development of digital technology in recent years has led to a rapid growth of applications of digital methods in the humanities. Nowadays, there is a very broad and vibrant field of research including all fields in the humanities. Yet, this raises the question whether *the* digital humanities can be seen as a unified research field or if digital humanities are an area of several research fields. In the "Digital Humanities Manifesto 2.0", the scholars argue that "Digital Humanities is not a unified field but an array of convergent practices that explore a universe" (Schnapp et al., 2008). However, as Liu (2016) points out the authors of the manifesto themselves refer to the digital humanities as a collective noun and therefore taking a singular verb (*is*). He further argues to treat digital humanities as a singular field which "need not imply consensus; it signals only that members of the field agree to participate in a common conversation" (Liu, 2016). We will make use of this notion in the following course of the thesis.

A development of the field can also be seen by the wide variety of scientific journals covering it.[3] A long standing journal in the field was known as *Literary and Linguistic Computing (LLC)*.[4] But, as the field evolved, the title became too narrow so that the own-

[1] "Gründungsmythos" (Jannidis et al., 2017)

[2] e.g., we can look up the lemma `sum` to find the inflected words `sum`, `es`, `fui`, `fuisti`, `essem`, `esses`, `fore`, `futurus`, etc. (Busa, 1980; Winter, 1999).

[3] A survey of journals can be found here `https://digitalhumanities.berkeley.edu/resources/digital-humanities-journals` (Digital Humanities at Berkeley, 2016) and here `https://guides.library.harvard.edu/c.php?g=310256&p=2071428` (Dressen, 2022).

[4] They even claim to be the "longest standing journal in the field" (Vanhoutte, 2015). However, in 1966

ers of the journal, the *Association of Digital Humanities Organizations* and the *European Association for Digital Humanities*, decided to change it to *Digital Scholarship in the Humanities (DSH)* to reflect "all digital scholarship undertaken in the Humanities in its widest meaning" (Vanhoutte, 2015). Today, there are both, journals that are targeted to a specific area within the field (e.g., *Journal of the Text Encoding Initiative*, *Journal of Cultural Analytics*, ...) and journals covering a wide range of digital humanities topics (e.g., *Digital Humanities Quarterly*, *Digital Studies / Le champ numérique*, ...).

Current projects reach from machine learning backed musical composition[5] to historical transportation networks in the Roman World.[6] Another wide range of projects aim to build up a searchable information database in art,[7] history,[8] and literature.[9] Beside these database approaches, a different branch in digital literary science is trying to go beyond pure gathering and digitization of texts and use quantitative tools like natural language processing (NLP), machine learning and network science to analyze large text corpora at once.

With the sudden plurality of new technological approaches, some scholars tried to unleash any technical method available to a text corpus of their choice. This led to the first major crises of the digital humanities with critics scathing that digital humanities could not uncover anything that was not already known by traditional hermeneutics.[10] The critics resulted in the common understanding that like the traditional humanities, a profound research question needs to set the limits and demands of any digital method used.

Moreover, it has become increasingly popular to use a mixed methods approach that makes use of both, the digital and the analogue world where "the specific affordances of each form should be understood and used together" (Berry, 2017).

"the American scholar Joseph Raben founded the first professional journal in the field, Computers and the Humanities (*CHum*)." (McCarty, 2003, p. 3) while LLC was founded in 1986 based on the *ALLC Bulletin* issued by the Association for Literary and Linguistic Computing.

[5] Human-Machine Interactive Composition Using Machine Learning at Berkeley DH `https://digitalhumanities.berkeley.edu/projects/human-machine-interactive-composition-using-machine-learning`.

[6] ORBIS at Stanford `https://digitalhumanities.stanford.edu/projects/orbis`.

[7] e.g., digitale diathek `http://digitale-diathek.net/`.

[8] Digital Manuscript Library Cologne (CEEC) `http://www.ceec.uni-koeln.de/`.

[9] e.g., Project Gutenberg `https://www.gutenberg.org/` or Projekt Gutenberg-DE `https://www.projekt-gutenberg.org/`.

[10] See for example Kathryn Schulz' criticism of Moretti's network analysis of *Hamlet* in the New York Times: "I mostly vacillated between two reactions: 'Huh?' and 'Duh!' - sometimes in response to a single sentence. For example, Moretti [...] means the protagonist is the character with the smallest average degree of separation from the others, 'the center of the network.' So guess who's the protagonist of Hamlet? Right: Hamlet. Duh?" (Schulz, 2011). A possible response is given by other scholars of the field, i.e., "Canons exist [...]. If we need to read books in order to extract information [...], we're going to spend most of our time dealing with a relatively small set of texts. [...] This is a canon. [...] When we try to do better [...], we do so both with a deep set of ingrained cultural biases that are largely invisible to us and (because we've read so little) in ignorance of their alternatives. We need to do less close reading and more of anything and everything else that might help us extract information from and about texts as indicators of larger cultural issues" (Wilkens, 2012). Other types of criticism were collected by Gold (2012).

1.1.2 Corpus analysis

As we have pointed out in the previous section, one of the oldest branches in the digital humanities is the field of linguistic computing.[11] Within this branch, corpus linguistics and corpus analysis are central elements. And similar to the digital humanities itself, there is a debate in the research community whether corpus linguistics is a research tool or a discipline in its own (While Gries, 2009 see it as a tool while Kuebler and Zinsmeister, 2015 argue that it is both).

A possible definition is that corpus analysis is the "study of language in use through corpora" (Bennett, 2010). The first major compiled corpus of various genres was the so called *Brown Corpus* (Kucera and Francis, 1967). Even though state-of-the-art text corpora in corpus linguistics do consist of a much greater number of words[12] Kucera and Francis (1967) introduced many types of analysis that are still widely used. Among others, they sorted the words in descending order of frequency[13] and gave "three numbers [...] which indicate its range of occurrence: its frequency in the Corpus, and the number of different GENRES and SAMPLES" (Maverick, 1969).

A typical corpus linguistic approach is described by Wallis and Nelson (2001) as the *3A perspective*. The three A's stand for *Annotation*, *Abstraction* and *Analysis*.

> "*Annotation* consists of collection, transcription standardisation, segmentation, processing (POS-tagging, parsing) and often further hand-correction and adding material to text.
> *Abstraction* consists of the process of choosing a research topic, defining parameters and variables and then extracting a sample, or 'abstract model', from the corpus.
> *Analysis* consists of processes which generate and test hypotheses to search for interesting patterns in this model" (Wallis, 2007).

Corpus linguistics usually tries to answer questions about the patterns associated with lexical or grammatical features using quantitative (e.g., collocations of words) and qualitative analyses (why do these patterns exist) (Biber et al., 1998) with the desire to "subject quantitative results to qualitative scrutiny" (Leech et al., 2009). While the results of corpus linguistics usually become more significant as the size of the corpus increases, this does not hold for other research questions that are based in the (digital) humanities. Hodel (2013) suggests using small corpora in the humanities (and not Google's n-gram viewer) to analyze those with specified methods. A similar demand for a variety in methods is given by Chouquer (2002).

While the research questions in this thesis go beyond the ones from corpus linguistics, we will use the *3A perspective* as one of the guiding principles in tackling some of these

[11] Scheinfeldt (2014) states that the "digital humanities family tree has two main trunks, one literary and one historical, that developed largely independently into the 1990s."

[12] e.g., the Corpus of Contemporary American English (Davies, 2010) with 1 billion words compared to the 1 million words in the Brown Corpus

[13] Showing that indeed Zipf's law, a claim that there exists an approximately proportional relation between the nth most frequent word and $\frac{1}{n}$ holds for the corpus (Zipf, 1935, 1949).

questions. More concretely, word frequencies, POS-tagging and annotations of characters will be used in Chapters 3 and 5.

1.1.3 Stylometry

The branch in computational linguistics that approaches corpus analysis via "quantitative assessment of linguistic features" (Lagutina et al., 2019) is called stylometry. Typical applications of stylometry include authorship attribution (defining the author of a given text based on a corpus of texts, whose authorship is known), authorship verification (binary authorship attribution), authorship profiling (deriving information about gender, age, psychological characteristics [BIG 5], etc.), style change detection (change of style over time or within a document), and classification of written texts (by genre, sentiment etc.) (Lagutina et al., 2019). Nonetheless, authorship attribution is by far the most pursued research topic.[14] This might be due to the simple and typical machine learning approach of performing a training-test split of the underlying corpus to verify predicted outcomes against true values and compare the studied method with a certain benchmark.

Any task performed in stylometry usually employs a classification step. The methods used for this can be divided into the broader categories machine-learning (ML) algorithms, minimum-distance techniques, and probabilistic models (Neal et al., 2017). Especially in the tasks described above, where scholars compare the proposed technique against existing methods, a race for the best performing algorithm usually splits the research community into those favoring a ML viewpoint and those preferring 'classical' statistical tools like Bayesian classifiers.[15] A widely used method in stylometry is the Delta distance introduced by Burrows (2002). While originally used as an intertextual distance measure it has since been reinterpreted as either a nearest-neighbor approach (and hence a ML algorithm) or "an approximation to ranking candidates by probability according to an estimated Laplace distribution" (Argamon, 2008). This reassessment led to a variety of similar distance measures that keep their straightforward interpretability while emphasizing certain feature components (e.g., Smith and Aldridge, 2011, Jannidis and Lauer, 2014, Jannidis et al., 2015, Evert et al., 2015; Büttner et al., 2017, and Posadas-Durán et al., 2017). The development of open source software like Signature[16] , JGAAP[17], and the R-package stylo[18] as well as language specific tools like Zemberek-NLP for Turkish[19], LexTo for the Thai language[20], and ICTCLAS for Chinese[21] made the application of many stylometric ideas easily accessible to other scientists and led to numerous analyses on several text corpora.

[14] In the survey by Lagutina et al. (2019) 24 out of 41 studies deal with authorship attribution and only four with non-authorship classification tasks.

[15] Depending on the data set and interpretation, a direct comparison leads to conflicting results on which method is superior(Kjell, 1994; Zhao and Zobel, 2005).

[16] https://www.philocomp.net/texts/signature.htm

[17] Java Graphical Authorship Attribution Program: http://evllabs.github.io/JGAAP/

[18] https://github.com/computationalstylistics/stylo (Eder et al., 2016)

[19] https://github.com/ahmetaa/zemberek-nlp

[20] http://www.sansarn.com/lexto/

[21] http://ictclas.nlpir.org/ (Zhang et al., 2003a)

Neal et al. (2017) group the features used in stylometric studies in lexical (character and word-based), syntactic (organization of sentences), semantic (meaning of words), structural (document organization), and domain-specific (content-specific information) categories as well as additional features like topic models. Moreover, they state that "there may exists a relationship between semantic features and topic and authorship, while syntactic features may be related to the length of the document" (Neal et al., 2017).[22] This brings them to the conclusion that an "optimal set of features that efficiently captures these correlations has yet to be discovered." While this hunt for the best combination of features might be fruitful in some contexts we want to argue that it can be much more helpful to understand the influence of the use of certain features in detail to better understand and interpret the results and conclusions that are drawn from analyses that are based on them.

1.2 Deutscher Novellenschatz – A 19th century German novella corpus

As stated above the corpus of *Deutscher Novellenschatz* is a collection of German language novellas that have been published in 24 volumes between 1871 and 1876 by Paul Heyse and Hermann Kurz. With the rise of mass literature production, a main goal was to establish a canon by assembling a paradigmatic sample of the novella style. The texts where originally published between 1811 (Kleist's *Die Verlobung in St. Domingo* as the oldest novella) and 1875 (Horner's *Der Säugling*) and span the epochs of German Romanticism, Biedermeier, Young Germany, Vormärz, and Poetic Realism. Main topics that can be seen as subgenres are wedding or marriage live, village live (Dorfgeschichten), art novellas, as well as crime and justice stories. As expected for a corpus compiled in the 19th century by two male authors, its gender distribution is highly skewed. While 74 of the authors are male there are only twelve[23] female authors.

The large success of the series led to two subsequent collections, the *Novellenschatz des Auslands* and *Neuer Deutscher Novellenschatz*. The later was edited by Heyse together with Ludwig Laistner after the sudden death of Kurz. As opposed to the original *Novellenschatz*, nearly all novellas from the *Neuer Deutscher Novellenschatz* are from the epoch of Realism.[24]

Each text was published with a short introduction by the editors. Moreover, a larger introduction was given to the *Novellenschatz* itself. These introductions served several larger purposes. In the introduction to the corpus, Heyse and Kurz develop a theory that can be converted into a definition of what a novella is to them.

[22] As a hint that this might exist he cites Sapkota et al. (2013).

[23] Adelheid Reinbold is using the male pseudonym Franz Berthold and the gender of the anonymous writer Fräulein Wolf is unknown but assumed to be female.

[24] An exception is Schiller's *Verbrecher aus verlorener Ehre.* A novella already mentioned in the introduction to the *Novellenschatz* but then ignored because it was seen as something from a pre period ("Vorperiode") while the beginning of German novella writing should be marked by Goethe's *Unterhaltungen deutscher Ausgewanderten.*

A classical definition of a novella was earlier given by Goethe who called a central element of a novella a past egregious incident[25] (conversation between Goethe and Eckermann on January 29, 1827, cited after Trunz, 1960, p.726). In 1881 Storm called the novella the sister of the drama[26] (Storm, 1920) to highlight that it has the strictest closed form of all prose texts. In a similar manner, the introduction to the *Novellenschatz* establishes the concept of a strong silhouette[27] to distinguish a novella from a novel. By a strong silhouette they mean the concentration on a guiding theme and a simple test is the possibility to summarize the text in a few short sentences. A second important feature of a novella is said to be the motif[28] that is a representative for the problem within the novella. They display their ideas on Boccaccio's so-called falcon novella[29] (Decamerone V,9) as a prime example, which is why the approach became famous as falcon theory (Heyse and Kurz, 1871). The importance of this genre marker is also discussed in the editor's correspondence letters.

Moreover, the short prefaces to each novella were twofold. On the one hand, they gave a short biographical overview of the novella's author. On the other hand, they contextualized and assessed the text within their own framework. In this assessment we can generally observe a more negative tone for novellas from romantic or female authors.[30]

The *Novellenschatz* was digitized by Weitin and made available in "Deutsches Textarchiv" (DTA).[31] The accessibility of the text lead to various previous work on the corpus. Weitin (2016, 2017) looked at text distances by word frequency to observe that indeed the novella of Heyse *Der Weinhüter von Meran* had the lowest distance from the corpus average and Kurz's *Die beiden Tubus* is central to a certain subgroup of novellas. As an extension, Weitin (2019) considered the entropy of the underlying values to investigate the quality of stylometric differences. Moreover, Weitin and Herget (2017) used topic modelling to produce what they call 'falcon topics', a number of topics that contain words characteristic for a single novella. Jannidis (2017) analyzed certain aspects of the corpus by identifying stylistic role models[32], comparing simple character networks to those of novels, and determine a certain female writing style for those novellas included in the *Novellenschatz.* Finally, Mellmann (2017) explored the corpus by quantitatively counting locations and tense.

[25] "sich ereignete unerhörte Begebenheit"

[26] "Schwester des Dramas"

[27] "starke Silhouette"

[28] "Dingsymbol"

[29] "Falkennovelle"

[30] e.g., introduction to Bretano's *Geschichte vom braven Kasperl und dem schönen Annerl*: "Der Volkston zeigt mitunter etwas Gemachtes und wird gelegentlich von einer nicht sehr volksmäßigen Sprache unterbrochen. Das zuckende Richtschwert sodann und der abgehauene Kopf, der dem Kind in das Röckchen beißt, sind aus der Dämmerung, wohin sie gehören, zu sehr in das Licht des Tages gerückt: denn der Aberglaube kann nur dann poetisch wirken, wenn er durch dritte oder vierte Hand überliefert wird; hier aber ist es eine als glaubwürdig vorgeführte Person, die uns Selbsterlebtes erzählt und hiemit Glauben an die Thatsachen so wie an deren Folgen beansprucht." and in the introduction to Dincklage's *Der Striethast* "Wenn es der Dichterin gelänge, ihre zuweilen aus Rand und Band gehende Phantasie zu zügeln [...], wäre sie ihrem Talent nach dazu berufen, [...] ihre Heimath so erfolgreich zu vertreten, wie es Fritz Reuter in der 'Stromtid' mit seinem Mecklenburg gethan hat."

[31] `https://www.deutschestextarchiv.de/search/metadata?corpus=novellenschatz`

[32] He sees Eichendorff's *Die Glücksritter* as such a role model.

1.3 Research theme

1.3.1 Distant reading vs. close reading

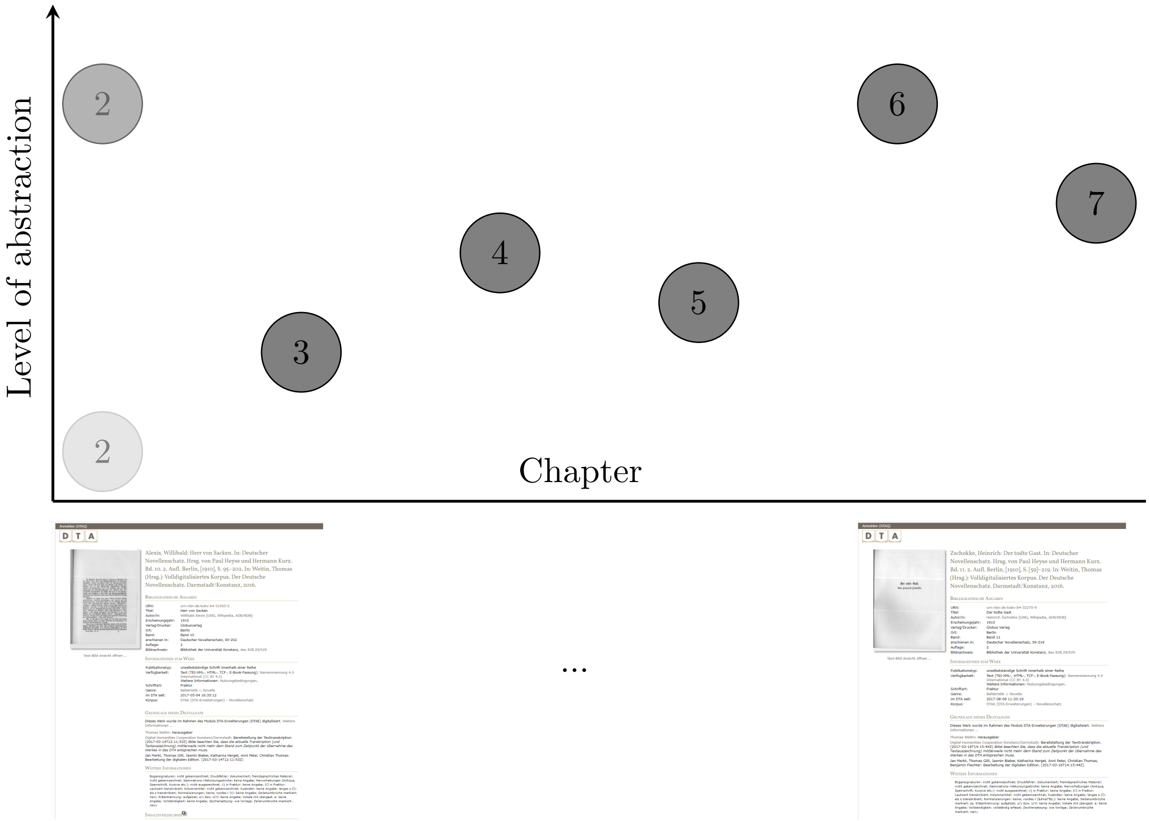
The term *distant reading* was introduced by Moretti (2000) and can be defined as the study of a large collection of literature via computational analysis of literary data to verify hypotheses for corpora that include non-canonical texts ("the great unread"). He admitted in an interview that the term was a last-minute addition to the essay "Conjectures on World Literature" (Hackler and Kirsten, 2016) but established it in 2013 in his book "Distant Reading" (Moretti, 2013). The new expression in the literary discourse naturally raises the question of how we can understand its antonym *close reading*. While Moretti himself refers to close reading as the "most complex possible analysis of [a *single* text] as such" (Hackler and Kirsten, 2016, p. 6), many others consider close reading to be the hermeneutic analysis of words, sentences, and texts in a subjective but well undergirded perception. In a broader sense, the distinction between close and distant reading is sometimes just seen as the difference between approaching a text via reading versus via statistical tools. However, this is a quite extreme point of view that expects the two methods being opposed to each other and does not allow for any combination or finer nuances. In contrast, we argue for a perspective that give space for intermediate resolution level.

1.3.2 Reading@scale

In the discussion about the correct relation between close and distant reading Mueller introduced the term "Scalable Reading" (Mueller, 2012; Müller, 2013). In contrast to his notation, we want to use the term at-scale to highlight the adequacy of the different domains covered. As stated above, the *Deutsche Novellenschatz* is a medium-sized corpus that can be accessed via statistical tools while still small enough to at the same time clarify results via close reading. Hence, it is perfectly suited for an at-scale approach to get insight from both, close readings and corpus data analysis. Instead of trying to find distant reading evidence for arguments won by close reading, one can fine-tune the methods to use the appropriate level of abstraction to answer a certain research question. This will lead to an effective compromise between precision and scope. Knowledge gained from close reading is detailed but lacks comparability. In contrast, pure word counting leads to statistical results that lack any contextual understanding.

1.4 Thesis structure

This thesis will use the at-scale approach to answer research questions that arise from the corpus itself. To achieve this goal, we use network analysis as a key method. The different resolution levels of the analysis contribute to experimental research in literary analysis, the theoretical foundations of network science in the humanities, and to network science itself. In Figure 1.1 we displayed how the six main chapters of this thesis are integrated within the at-scale approach.



Anmelden (DTAQ)

D T A

Alexis, Willibald: Herr von Sacken. In: Deutscher Novellenschatz. Hrsg. von Paul Heyse und Hermann Kurz. Bd. 10. 2. Aufl. Berlin, [1910], S. 95–202. In: Weitin, Thomas (Hrsg.): Volldigitalisiertes Korpus. Der Deutsche Novellenschatz. Darmstadt/Konstanz, 2016.

Text-Bild-Ansicht öffnen …

Bibliographische Angaben

URN:	urn:nbn:de:kobv:b4-31505-0
Titel:	Herr von Sacken
Autor/in:	Willibald Alexis (GND, Wikipedia, ADB/NDB)
Erscheinungsjahr:	1910
Verlag/Drucker:	Globusverlag
Ort:	Berlin
Band:	Band 10
erschienen in:	Deutscher Novellenschatz, 95-202
Auflage:	2
Bildnachweis:	Bibliothek der Universität Konstanz, deu 838.29/h29

Informationen zum Werk

Publikationstyp:	unselbstständige Schrift innerhalb einer Reihe
Verfügbarkeit:	Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung): Namensnennung 4.0 International (CC BY 4.0) Weitere Informationen: Nutzungsbedingungen.
Schriftart:	Fraktur
Genre:	Belletristik :: Novelle
im DTA seit:	2017-05-04 16:35:12
Korpus:	DTAE (DTA-Erweiterungen) – Novellenschatz

Grundlage dieses Digitalisats

Dieses Werk wurde im Rahmen des Moduls DTA-Erweiterungen (DTAE) digitalisiert. Weitere Informationen …

Thomas Weitin: Herausgeber

Digital Humanities Cooperation Konstanz/Darmstadt: Bereitstellung der Texttranskription. (2017-03-14T12:11:53Z) *Bitte beachten Sie, dass die aktuelle Transkription (und Textauszeichnung) mittlerweile nicht mehr dem Stand zum Zeitpunkt der Übernahme des Werkes in das DTA entsprechen muss.*

Jan Merkt, Thomas Gilli, Jasmin Bieber, Katharina Herget, Anni Peter, Christian Thomas: Bearbeitung der digitalen Edition. (2017-03-14T12:11:53Z)

Weitere Informationen

Bogensignaturen: nicht gekennzeichnet; Druckfehler: dokumentiert; fremdsprachliches Material: nicht gekennzeichnet; Geminations-/Abkürzungsstriche: keine Angabe; Hervorhebungen (Antiqua, Sperrschrift, Kursive etc.): nicht ausgezeichnet; i/j in Fraktur: keine Angabe; I/J in Fraktur: Lautwert transkribiert; Kolumnentitel: nicht gekennzeichnet; Kustoden: keine Angabe; langes s (ſ): als s transkribiert; Normalisierungen: keine; rundes r (ꝛ): keine Angabe; Seitenumbrüche markiert: nein; Silbentrennung: aufgelöst; u/v bzw. U/V: keine Angabe; Vokale mit übergest. e: keine Angabe; Vollständigkeit: keine Angabe; Zeichensetzung: wie Vorlage; Zeilenumbrüche markiert: nein;

Inhaltsverzeichnis

…

Anmelden (DTAQ)

D T A

Zschokke, Heinrich: Der todte Gast. In: Deutscher Novellenschatz. Hrsg. von Paul Heyse und Hermann Kurz. Bd. 11. 2. Aufl. Berlin, [1910], S. [59]–219. In: Weitin, Thomas (Hrsg.): Volldigitalisiertes Korpus. Der Deutsche Novellenschatz. Darmstadt/Konstanz, 2016.

Der todte Gaſt.
Von Heinrich Zſchokke.

Text-Bild-Ansicht öffnen …

Bibliographische Angaben

URN:	urn:nbn:de:kobv:b4-32270-9
Titel:	Der todte Gast
Autor/in:	Heinrich Zschokke (GND, Wikipedia, ADB/NDB)
Erscheinungsjahr:	1910
Verlag/Drucker:	Globus Verlag
Ort:	Berlin
Band:	Band 11
erschienen in:	Deutscher Novellenschatz, 59-219
Auflage:	2
Bildnachweis:	Bibliothek der Universität Konstanz, deu 838.29/h29

Informationen zum Werk

Publikationstyp:	unselbstständige Schrift innerhalb einer Reihe
Verfügbarkeit:	Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung): Namensnennung 4.0 International (CC BY 4.0) Weitere Informationen: Nutzungsbedingungen.
Schriftart:	Fraktur
Genre:	Belletristik :: Novelle
im DTA seit:	2017-08-09 11:30:19
Korpus:	DTAE (DTA-Erweiterungen) – Novellenschatz

Grundlage dieses Digitalisats

Dieses Werk wurde im Rahmen des Moduls DTA-Erweiterungen (DTAE) digitalisiert. Weitere Informationen …

Thomas Weitin: Herausgeber

Digital Humanities Cooperation Konstanz/Darmstadt: Bereitstellung der Texttranskription. (2017-03-16T14:15:44Z) *Bitte beachten Sie, dass die aktuelle Transkription (und Textauszeichnung) mittlerweile nicht mehr dem Stand zum Zeitpunkt der Übernahme des Werkes in das DTA entsprechen muss.*

Jan Merkt, Thomas Gilli, Jasmin Bieber, Katharina Herget, Anni Peter, Christian Thomas, Benjamin Fiechter: Bearbeitung der digitalen Edition. (2017-03-16T14:15:44Z)

Weitere Informationen

Bogensignaturen: nicht gekennzeichnet; Druckfehler: dokumentiert; fremdsprachliches Material: nicht gekennzeichnet; Geminations-/Abkürzungsstriche: keine Angabe; Hervorhebungen (Antiqua, Sperrschrift, Kursive etc.): nicht ausgezeichnet; i/j in Fraktur: keine Angabe; I/J in Fraktur: Lautwert transkribiert; Kolumnentitel: nicht gekennzeichnet; Kustoden: keine Angabe; langes s (ſ): als s transkribiert; Normalisierungen: keine; rundes r (&#xa75b;): keine Angabe; Seitenumbrüche markiert: ja; Silbentrennung: aufgelöst; u/v bzw. U/V: keine Angabe; Vokale mit übergest. e: keine Angabe; Vollständigkeit: vollständig erfasst; Zeichensetzung: wie Vorlage; Zeilenumbrüche markiert: nein;

Fig. 1.1: We plot the position of each chapter by their distance from the actual texts of the *Novellenschatz* that are displayed in their digitized form (as screenshots from the DTA websites). The further up, the further away an analysis is from the underlying text.

In Chapter 2 we tested the claim made by the editors that a novella is summarizable in a few sentences. More concretely, we want to answer the following research question.

Research Question 1. *Can we empirically prove the property of the novella as defined by Heyse as an effect on readers?*

In an experiment we had 86 participants read and summarize the text *Ein adeliges Fräulein* by Lorm which is claimed to be a novella because of its presence in the *Novellenschatz*. Later, we tested whether these summaries are consistent in content and chronology which might be expected by the arguments of Heyse and Kurz on how they compiled the corpus. Therefore, the level of abstraction in Chapter 2 works on two scales. The experiment itself is a close reading task and operates on the closest level you can get to the text. In contrast, the analysis takes place on a quite abstract level because it totally discards the underlying text and only focuses on the participants answers. An article version of this chapter is in preparation. The data collection and experiment realization were done by Katharina Herget, Anastasia Glawion, Judith Brottrager, and Thomas Weitin.

In Chapter 3 and 4 we look at stylometric similarities within the corpus. As such, both chapters are using word similarities and are therefore very close to the texts of the underlying corpus. More detailed, Chapter 3 demonstrates the use of different stylometric features on a fine-grain scale by studying the effect of these features on selected novellas. In addition, Chapter 4 looks at the effect of the interplay of these features on the whole corpus. Both chapters together want to answer the following question.

Research Question 2. *Can stylometric features and similarity networks that are derived from them give insight into both, the position of the novellas in the corpus and the applicability of these features in an attribution context other than authorship?*

Moreover, we want to contribute to a better theoretical understanding of the use and misuse of stylometric features in general by applying them outside of their typical field of authorship attribution. Overcoming the possibility of comparing their results to a ground truth, this gives way to address a qualitative level of the analysis. In fact, the research question will be answered by looking at the two subparts of the it separately. We will answer the second part of the research question in Chapter 3 by looking at the more detailed question.

Research Question 2a. *How does the process of determining similarity influences the notion of similarity itself and how can it lead to an informed construction of a similarity measure tailored to the specific needs of the researcher?*

In Chapter 4 we want to use our gained knowledge to apply specific similarity measures to demonstrate how these measures help in identifying subgroups in the corpus. We therefore ask:

Research Question 2b. *Which groups can be found or deliberately deconstructed by the application of a tailored similarity measure.*

This was a joined work with Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes and a modified version is submitted to the DSH Journal and accepted for publication.

The third part of the thesis consists of the Chapters 5, 6 and 7. While previous parts considered the network between texts, this part will also focus on networks within the novellas. In Chapter 5 we extract high-quality character co-occurrence networks. However, in contrast to the common approaches used in the literature, we instead visualize them as two-mode networks with links between characters and paragraphs. In this way we can use these networks in a very flexible way and make them later available to the research community[33]

Chapter 6 has again the highest degree of abstraction. It considers the network of networks or more precisely a network ensemble based on the character networks extracted from the corpus in Chapter 5. By comparing those networks with a distance measure, it tries to answer the following question:

[33] Among others, a possible use case could be the comparison to real world networks, as in (Volker and Smeets, 2020).

Research Question 3. *Does the restricted form of a novella give preference to a specific character constellation?*

Different stages of this work have been presented at the 4th European Conference on Social Networks (EUSN 2019) in Zurich, the "Historical Networks – Réseaux Historiques – Historische Netzwerke" conference (HNR+ResHist 2021), and at the 2nd International Conference of the European Association for Digital Humanties (EADH 2021). A predecessor version is published in the book of abstracts of EADH 2021.[34] The result of this analysis will show that this holds partially true when contrasted with a comparative corpus of novels. Furthermore, the approach reveals an inner structure of the novellas in the corpus. This structure will then be studied in detailed in Chapter 7.

Besides the different views this thesis takes on the *Novellenschatz* corpus, a common theme in each of the projects lies in the type of analysis. We will display the decisions in use of techniques that are taken along the way side by side to possible alternatives and explain why and how the answer to the research questions can profit from the chosen option. Hence, all choices are well-founded, and we can largely bypass the risk to be guided by the availability of methods instead of profound theory. In addition, all research questions have common underlying motivations. On the one hand, they will give new insights in the theoretical foundations of network science in text analysis. On the other hand, the use of them will not be an end in itself. It will rather give a clearer understanding of the corpus generation process in a time of arising literary mass production as well as the idea of how close or far the selected texts come to the claimed ideal of a characteristic novella.

This work will be primary addressed to the digital humanities community. Above the described methodological contribution to the field the research project aims to popularize a philosophy that is shaped not only by the qualitative vs. quantitative divide but also by the more granular difference of structured and unstructured data that looks out to bridge between statistical and hermeneutical comprehension. While this work uses, adapts and where necessary refines methods that are more prevalent in the social sciences, the benefit for scholars in the social sciences is limited to few domains. However, we do not want to leave out the importance of well-founded data sets that can be applied in various contexts. Therefore, the data set in Chapter 5 is constructed in a way that it can satisfy the needs of the social science community.

This dissertation was embedded in a larger research project "reading@scale" funded by the *Volkswagen Foundation* in the program "Digital Methods in the Humanities".

[34] `https://eadh2021.culintec.de/P_PCKE_Simon_Character_Networks_in_a_Collection_of_19th_Cent.html`

Chapter 2

Silhouette experiment

2.1 Introduction

With the falcon theory, the editors of the *Deutsche Novellenschatz* offer a definition for a novella in general and a quality check specifically for the novellas in their collection. An excellent novella – they say - shows a *strong silhouette* (Heyse and Kurz, 1871, p. XIX), which means it has one strong theme. Even more, they offer an operationalization to test this by stating that this silhouette can be detected by summarizing its content in just a few lines. On that basis, we aimed to test Heyse's hypothesis, in which a novellas potential for density is understood as a quality criterion of a novella by presenting probands a text from the *Deutsche Novellenschatz* that they had to read and summarize in a few sentences. Inspired by Roland Bathes' concept of an action sequence (Barthes, 1971) we decided to use a strictly structured experiment design.

The founding research question is:

Research Question 1. *Can we empirically prove the property of the novella as defined by Heyse as an effect on readers?*

Accordingly, the hypothesis of this experiment is

Hypothesis. X is a novella (by acceptance in the *Novellenschatz*) and can therefore be summarized by readers in a few lines (effect).

2.1.1 Description of the data

For the experiment we needed to choose the novella X. While in theory any novella should be suitable to test our hypothesis (as they are all part of the *Novellenschatz*) we had the following practical limitations in our choice. First, it needed to be read und summarized in approximately 60 minutes (the average reading speed is 100-250 words/min Hunziker, 2006). This broke down the novella selection process to the following three novellas: Hieronymus Lorm's *Ein adeliges Fräulein* (9568 words), Theodor Storm's *Eine Malerarbeit* (9411 words), or Moritz Hartmann's *Das Schloß im Gebirge* (8208 words). After a pretest, we settled on

Lorm's *Ein adeliges Fräulein* as the text for the experiment. While all three novellas are relatively short and compact, the story of Lorm has the further advantage that its clear motif (a painting) is easily recognizable. It combines the novella-typical frame story and internal plot and is therefore particularly suitable for our research topic. After reading the participants answers it seems difficult to clearly identify a common structure and chronology. If one would comprise the answers in some kind of reference summary that is taken from participants answers, it could look like this:

> *"Der Kurator eines Fürsten begibt sich zur Weihnachtszeit zur Freifrau von Börte, um ihr ein Gemälde für seinen Fürsten abzukaufen."* (participant 28)
> *"Von der Wirtin des Wirtshauses, in das er einkehrt, erfährt er die Geschichte der Frau von Börte.*
> *Sie schien glücklich verlobt mit dem bürgerlichen Landschaftsmaler Thurn zu sein, brach jedoch plötzlich jeden Kontakt mit ihm ab."* (participant 17)
> *"Als er von der Freifrau während eines Spaziergang für einen Stadtmenschen, welcher sich verirrt hat, gehalten wird, wird er in Ihr Haus eingeladen.*
> *Er hat die Möglichkeit ein Blick auf das besondere Bild zu erhaschen und erfährt bei seinem Besuch mehr über die Geschichte der Freifrau und über das Bild."* (participant 33)
> *"Ihre tragische Liebesgeschichte mit einem Kunstkenner soll durch eine Zwangsheirat durch ihren Vater mit einem anderen beendet werden, doch durch ihre unfreiwillige Zustimmung, fühlt sie sich selbst nach dem Gedankenwechsel des Vaters des Ehebruchs und Verrats schuldig und beendet ihre Beziehung.*
> *Sie sagt, dass sie ihn nie wiedersehen kann, verschweigt ihm aber den Grund.*
> *Da das Bild von ihm war, kann sie es nicht verkaufen, da es keinen Eigentümer hat."* (participant 61)

A translation into English could be

> *Around Christmas time, the curator of a prince repairs to the baroness von Börte to buy a painting for his prince from her.*
> *The story of Madame von Börte is told to him by the innkeeper of a pub.*
> *She seemed to be happily engaged with the landscape painter Thurn but suddenly broke tie with him.*
> *While taking a walk, he meets the baroness, who considers him as a lost city slicker and invites him into her home.*
> *He has the possibility to catch a glimpse of the special painting. During his visit he learns more about the baroness and the painting.*
> *Her father wants to end her tragic love story with an art connoisseur should by forced marriage with another man. Even after her father changes his mind, she feels guilty for her involuntary approval and ends the relationship.*
> *She tells Thurn that she cannot see him again but keeps quite about the reasons.*
> *Because she has the painting from him, she must not sell it, since it has no owner.*

However, one should note that this is not claimed to be the 'correct' answer but can be viewed as the least common denominator.

2.1.2 Empirical Setting.

The experiment took place on May 20, 2019, at the TU Darmstadt. A total of 85 subjects participated. The mean age of the sample was 24.1 years (SD = 6.2 years). All but three participants where in the range of 18 to 31 years with the outliers being 41, 50, and 62 years. Since it might be relevant to the way participants write their answers, we included questions about first language(s) and reading frequency. 65 of the participants had German as their mother tongue (15 of these had a second mother tongue), other mother tongues where Chinese (10), Arab (2), Turkish (2), Dutch (1), Italian (1), Romanian (1), Spanish (1), Tigrinya (1), and Urdu (1). 20 study participants self-identified as high frequent readers while 15 identified as low frequent readers (all others identified as medium frequent readers). Media types that have been mentioned included social media (76), news (69), fiction books (57), non-fiction books (47), scientific papers (3), graphical novels (3), and forums or blogs (2). The gender distribution was 41 male and 44 female participants. Finally, no participant answered that they previously knew the novella.[1] Besides the socio-demographic data the questionnaire consisted of four elements. The first task was to read the given text carefully. As a second question we asked for the most memorable element of the text.[2] One objective was to test the text's impact on the reader. One could have even hoped to identify the novellas motif. However, there were only four answers mentioning the painting. Instead, frequent answers contained the themes tragic love story (18), christmas / wintertime (13), nobility and commons as classes in society (11), and nature (6) as well as general comments about orthography (12), emotions that rose in the probands (7), and the epoch of Romanticism (4).

The second task was to summarize the plot in full sentences.[3] This was the main element of our survey, in order to support a sequence-like description of the plot; we have provided six text fields here, each of which was suitable for a short sentence. In addition, some blank spaces have been left in case the available space would not be sufficient. The last question was about text understanding to examine the participants attention for the text.[4]

The experiment was approved by the ethics commission of TU Darmstadt.

2.2 Analysis

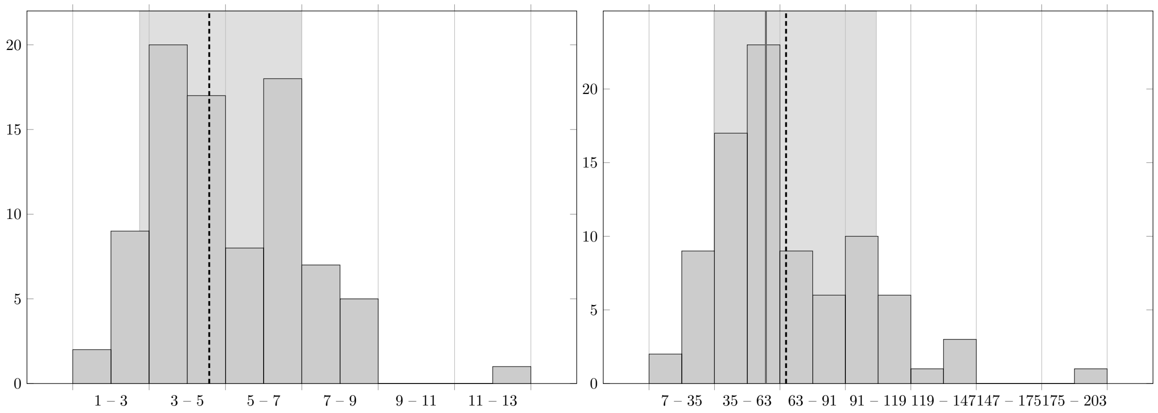
The analysis of the received summaries is divided into a number of steps. With the idea of looking at action sequences we can justify our first step of splitting the summaries by sentence. To employ some form of a distance measure that should capture the sentences

[1] On the question, do you know the novella, 81 answered "no", one person "I don't know", and 3 provided no answer.

[2] The precise wording in German was "Was kommt Ihnen vom Inhalt des Textes als erstes in den Sinn?"

[3] "Fassen Sie das Geschehen in ganzen Sätzen zusammen."

[4] "Was ist die wesentliche Aussage des Textes?"



(a) Sentences. Median is four sentences, mean (dashed line) is 4.586 sentences.

(b) Words. Median (gray line) is 57 words, mean (dashed line) is 65.621 words.

Fig. 2.1: Histograms for the number of sentences and words used by the participants to write the summary. The gray shadowed domain marks the area in which 75% of the summaries fall.

closeness in meaning we converted each sentence into a vector based on the words that occur. After utilizing such a suitable measure, we used a clustering methods to show dissimilarities within the answers and plotted an embedded two-dimensional version of the distance matrix for visualization purposes. In a final step we compare how similar the experiment participants answers are by computing edit distances based on the cluster profile of an answer.

Sentence splitting. In a first step, we are splitting the summaries sentence by sentence. The empirical distribution of the number of sentences used can be seen in Figure 2.1a. We observe that already in this early step of the analysis there is no agreement between the probands on how many words or sentences are needed to build the required action sequence. Instead, they use between one and twelve sentences to summarize the text. If we instead look at the number of words (Figure 2.1b) the result is not getting any clearer (Answers range from 7 to 203 words per summary). However, this is not at all unexpected since expression and language should vary between individuals. Therefore, we need to find a way of making the summaries comparable even if they differ in length of words and sentences.

Word occurrences. A first step towards this direction is done by the removal of stopwords as well as all other words that do not transport meaning. After a pretest it turned out that significant plot description is condensed in the used nouns and verbs. Thus, we decided to only consider words with those part-of-speech (POS) tags.[5] This is directly in

[5] The POS tags where generated with `spacy`, a python software package for NLP (Honnibal and Montani, 2017). It is available via `spacy.io`.

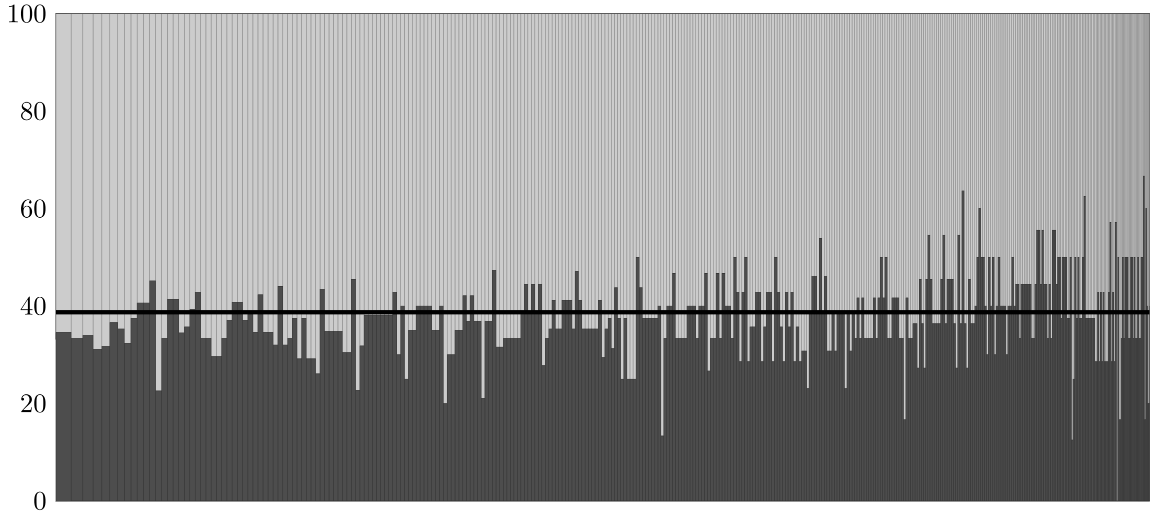


Fig. 2.2: Each thin bar represents a sentence. The area of the bar is proportional to the sentence length (in descending order from left to right). The darker area covers the percentage of words that are considered for the subsequent analysis. The black line shows the average proportion of words kept (≈ 39%).

line with Barthes concept of action sequences (Barthes, 1971).

In Figure 2.2 we can see the proportion of words that are used per sentence. It can be observed that the removal of words is equidistributed throughout the sentences with no connection to sentence length except for higher volatility for shorter sentences.[6] In a second step we combine words that are very similar via Levenshtein distance to account for spelling errors and lemmas of the same word.[7] This shrinks our vocabulary further down from originally 1577 to 759 words.

In addition we can observe in Figure 2.3 in how many sentences the most frequent words appear. However, this might be misleading because we instead want to know how discriminative a word is for the description of different sections of the novalla.

To get a first hint what we want to do, we look at the distribution of some sample words over the summaries sentences in Figure 2.4. The following observations can be made. While the theme Christmas (words starting with *Weihnacht* or *Weinacht* [sic!]) is highly present in the first sentences, we find the theme village (words starting with *Dorf*) and pub (words starting with *Wirt*, e.g. *Wirthshaus*) in the middle and the topics father (*Vater*) and marriage (*heirat*) at the end of the summaries. The more general noble titles for the eponymous hero (*Freifrau* and *Baronin*) are more equally distributed. One possibility to get a general quantification of this phenomenon is derived from the *tf-idf*-score (Salton and Buckley, 1988).

Therefore, we first convert our leftover sentences into sparse vectors that have the length

[6] This is expected. A sentence of length n can take $n + 1$ possible values. Hence, shorter sentences have

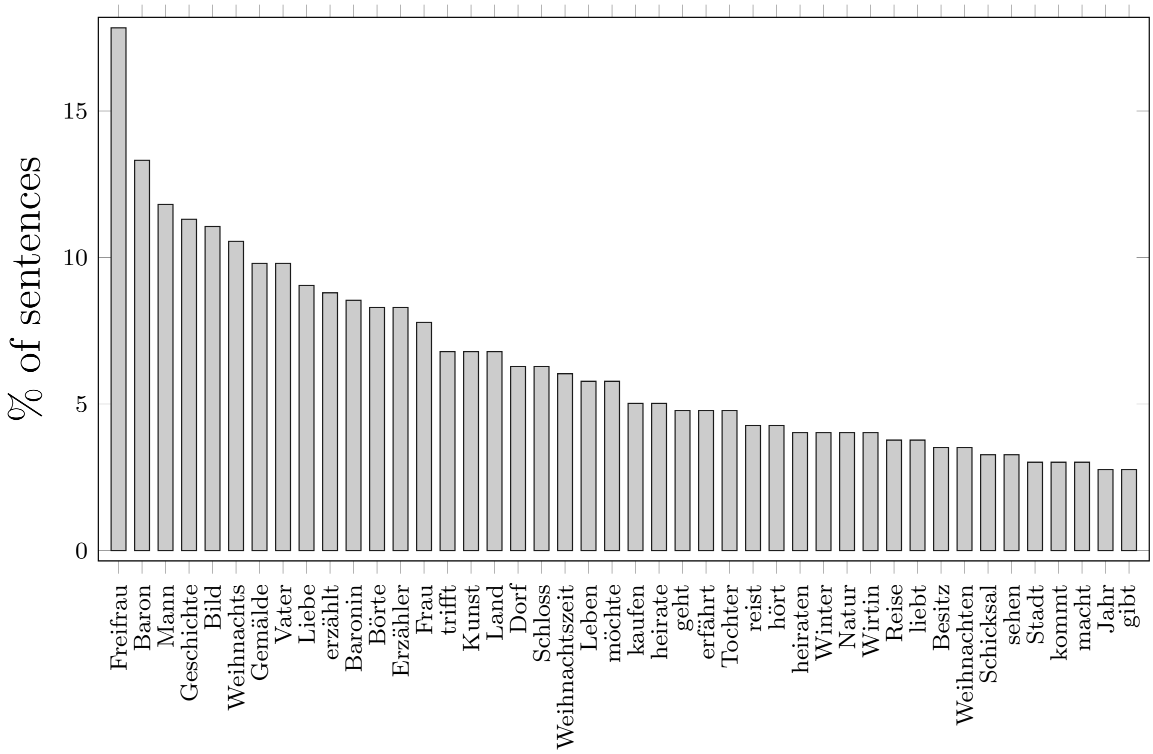


Fig. 2.3: Distribution of the words that appear at least ten times.

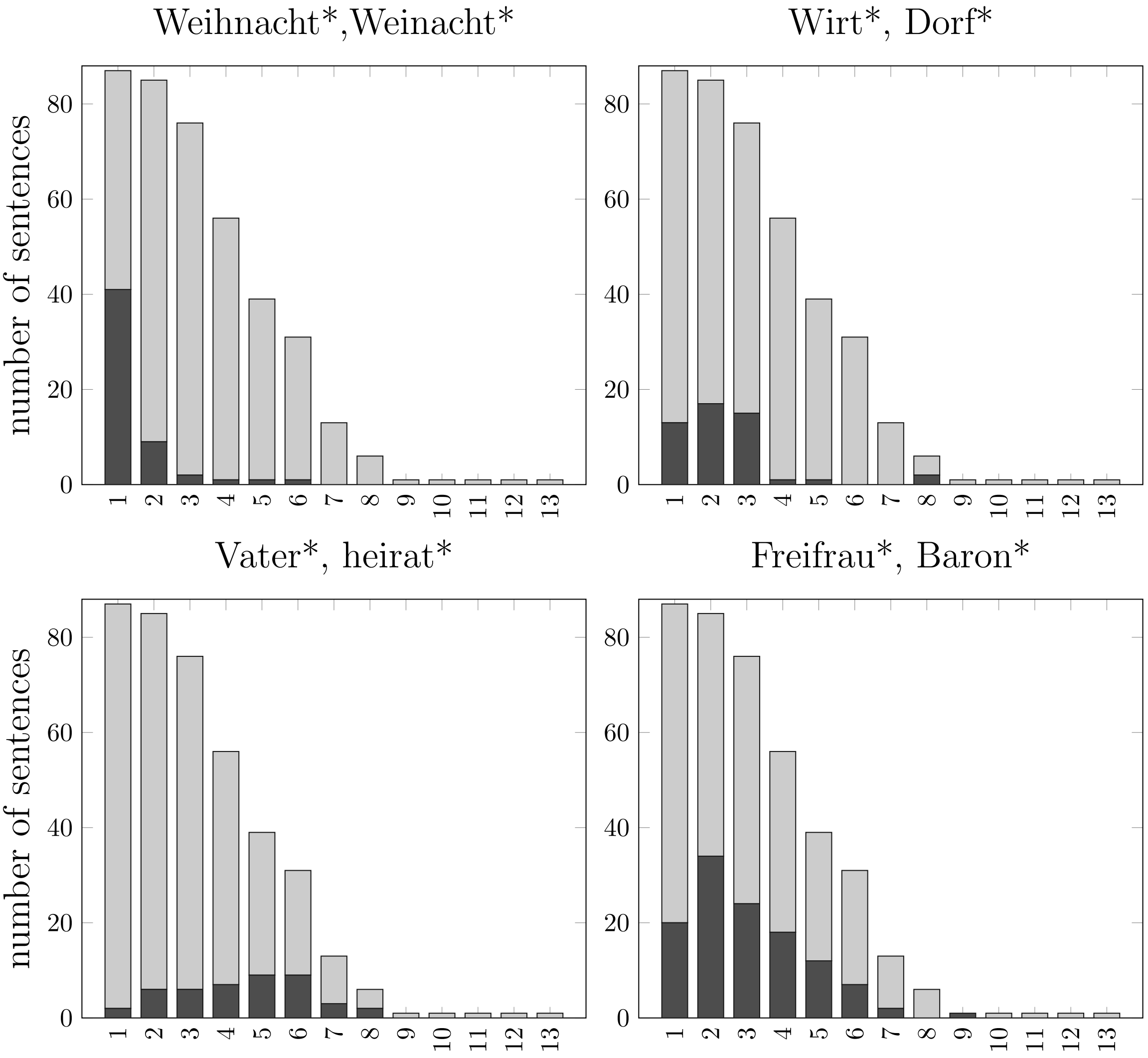


Fig. 2.4: The x axis indicates the ordering number n of a sentence within a summary. The height of each bar is given by the number of summaries with at least n sentences. The darker part of each bar illustrates the share of sentences using the designated words (where a star means that it any characters can follow).

(dimension) of the vocabulary (759) and the only non-zero entries of the vector appear if a word is present in the sentence. Starting from these sparse vectors, we build a vector for each word containing its *tf-idf*-score according to the order of the sentence in which it is used. Concrete, we consider all first / second / third / ... sentences together as singular documents D and compute the *tf-idf*-score based on this for each term t in the summaries. The *tf-idf*-score is a term from information retrieval. It is obtained by multiplying relative frequencies with the logarithmically scaled inverse of the share of documents in which a term appears,

$$\mathit{tf\text{-}idf}(t) = \mathit{tf}(t, D) \cdot \log\left(\frac{N}{\sum_{t \in D} 1}\right).$$

Here, $\mathit{tf}(t, D)$ (the *term frequency*) is the relative frequency of the term t in document D, i.e., the number of times term t appears among all words[8] in all first / second / third / ... sentences. where $|D|$ is the number of words of a document D in the corpus. Moreover, $N = 13$ is is the total number of documents, hence the maximum number of sentences that a summary contains. This weighting, and minor variants thereof, are referred to as *inverse document frequency* (Salton and Buckley, 1988; Manning et al., 2008). It serves to boost terms that appear in fewer documents and thus are potentially more informative to discriminate documents. Note that the weight of terms appearing in all documents equals zero and that the weighting by inverse document frequency is constant, and thus irrelevant, when all n terms appear in equal numbers of documents.

We then allocate the computed value to the word vector of each sentence where the word appears. In this way, words that are specific for a certain sentence get a higher value if they occur in that sentence than the same word occurring outside the expected sentence. In the same way terms that are used throughout the summary get lower values in general.

Example . *The words considered are Weihnachtszeit and Freifrau. If we compute their tf-idf-scores we get the following.*

$$\begin{aligned} \mathit{Weihnachtszeit} &\ [0.060, 0.003, 0.004, 0, 0, 0, 0, 0, 0, 0, 0, 0] \\ \mathit{Freifrau} &\ [0.012, 0.023, 0.018, 0.018, 0.012, 0.016, 0, 0.081, 0, 0, 0, 0] \end{aligned}$$

Let us now consider three comprised "sentences".

s_1 =*Kurator Fürsten Gemälde begibt Weihnachtszeit*
Freifrau Börte abzukaufen

s_2 =*Frau Freifrau Diskussionen Börte*

s_3 =*verlassen Frost Schnee aufzunehmen liebt Weihnachtszeit Stadt*

For s_1 we have a word vector that has entry 0.060 *for the word Weihnachts-zeit and* 0.012 *for Freifrau since it is a first sentence of a participants summary. In contrast, s_2 is a second sentence within a summary and therefore has the value* 0.023 *for the Freifrau entry. Finally, s_3 is a third sentence of a summary and has* 0.004 *as its Weihnachtszeit entry.*

fewer values.

[7] A list of words replaced can be found in Appendix A1

[8] Where all words refer to those that are leftover after preprocessing.

Sentence distances. A common measure to compare such vectors is the cosine similarity. However, the pure use of cosine similarity is too strict because we do not care so much about the actual wording and want to find sentences with similar content. This can be considered by using soft cosine similarity that maintain a small distance between sentences using similar words. This generalization of cosine similarity has been proposed by Sidorov et al. (2014). They suggest using an additional term-similarity matrix $S = (s_{ij}) \in \mathbb{R}_{\geq 0}^{n \times n}$ to weight the contributions of all pairs of entries x_i, y_j. The *soft cosine* similarity measure

$$\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} x_i y_j}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} x_i x_j} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} y_i y_j}}$$

reduces to cosine similarity for $S = I_n$, i.e.,

$$s_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

where terms are similar only to themselves. To compute the similarity matrix we relied on spacy's similarity scores.[9] These similarity scores are between 0 and 1 and give small but non-zero values when the similarity is not obvious to humans. Because many of these non-zero values do add up when looking at a vocabulary of 759 words and therefore influencing the results, we decided to only consider the 5% that are the most similar to each word and set all other similarities to 0. Because of the symmetry of S this can be more than 5% of the matrix entries.

As an example, we want to compare two summaries that are shown in Table 2.1. We used background coloring of the sentences to propose a possible clustering that a human reader could suggest. The corresponding soft cosine similarity matrix is shown in Table 2.2.

First, we can observe that the relatively short sentences **A 1**, **A 2**, and **B 3** are actually having only small similarity to all other sentences. Their highest similarity scores do not exceed values of 0.4. In contrast, sentence **B 2** is having a low similarity on average, while having high similarity with the sentence *A 3*. However, this represents distinctiveness of the sentence while the sentences **A 1**, **A 2**, and **B 3** are quite generic. Second, the fact that part of the grey shaded numbers lie on a diagonal can lead to the conclusion that there might be some detectable common understanding of content and chronology.

Clustering. To gain deeper knowledge whether this holds true on a global scale, we try to cluster our results. First, we use the dimensionality reduction technique t-sne (Hinton and Roweis, 2002; Maaten and Hinton, 2008) to get an embedding of the similarity matrix in a two dimensional space. This is shown in Figure 2.5a. We see a strong cohesiveness for the first sentence (top left corner). Many of the second and third sentences are in the bottom left corner. However, the separation is much less visible then for the group of first

[9] A list of 30 out-of-vocabulary words can be found in Appendix A1. Besides the name *Börte* these are mostly long composite words and misspellings.

Participant A	**Participant B**
Ein Mann reist jedes Jahr zur Weihnachtszeit aufs Land.	Ein Bürgerlicher erinnert sich an eine Winterreise und seine dortige Begegnung mit einer Baronin von der er ein Bild erwerben möchte.
In einem Jahr mit dem Vorhaben, ein besonderes Bild von einer Freifrau von Börte zu beschaffen.	In einem Wirtshaus klärt er erst genauere Informationen zu dem Fräulein.
Im Gebirgsdorf nahe ihres Herrenhauses erfährt er von der Wirtin die Lebensgeschichte der Freifrau.	Niemand weiß genau was ihr Schicksal ist.
Er macht sich selber auf den Weg zu ihr und sie erzählt ihm die tragische Geschichte ihrer Jugend.	Bei einem zufälligen Treffen finden sie sich sympathisch und er erfährt ihre Geschichte.
Sie wohnte mit ihrem Vater alleine in dem Haus und lernte über geschäftliche Beziehungen einen jungen Mann kennen und lieben.	Trotz ihrer Liebe willigte ihrem Vater zu Liebe einer anderen Hochzeit ein und danach findet sie, dass sie kein Glück mehr empfinden darf, da sie ehrlos ist.
Auf Verlangen ihres Vaters, da der Mann nicht ihrem Stand entspricht, will sie ihn nie wiedersehen und verbringt ihr Leben nur damit, sich an der Natur zu erfreuen.	

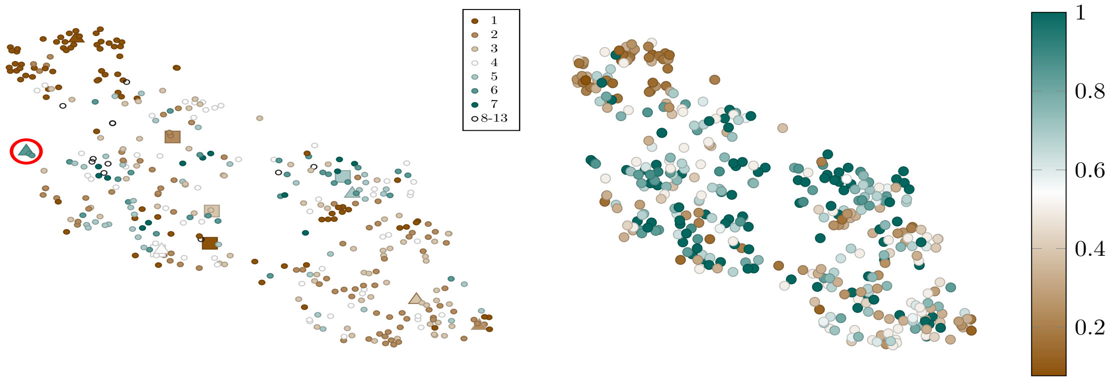
Table 2.1: Two summaries of Lorm's novella *Ein adeliges Fräulein* with a possible human selected clustering.

	A 1	**A 2**	**A 3**	**A 4**	**A 5**	**A 6**
B 1	0.068505	0.227016	0.306183	0.484832	0.199082	0.113300
B 2	0.052517	0.129623	0.515282	0.025771	0.234336	0.040373
B 3	0.301810	0.159635	0.131499	0.358703	0.260174	0.216767
B 4	0.033738	0.128108	0.168511	0.730363	0.199355	0.043729
B 5	0.267364	0.000000	0.178595	0.047258	0.588535	0.411057

Table 2.2: Soft cosine similarities. For each row and column, the highest entry is colored.

sentences. This is not necessary a bad sign since what some persons write in one sentence needs two sentences for others. This shift propagates to the next sentences such that it is even harder to find groups there. Nonetheless, there are some very dense groups of very similar sentences as the three sentences in the middle left (red circle).

A possibility to avoid this problem is given by coloring the nodes according to their position within a summary. However, as seen in Figure 2.5b this does not significantly improve the situation. Instead we use a clustering algorithm to find and further investigate



(a) Number of the sentence indicated by color. The sentences of **A** are marked as triangles, those of **B** as squares.

(b) Position of the sentence indicated by color. Brown (=0) marks the first, turquoise (=1) the last sentence.

Fig. 2.5: t-SNE plot of the participants answer sentences.

groups in the answers that might not be bounded to the pure position of a sentence in a summary.[10] In Figure 2.6 we compare the effects of two techniques, namely density based clustering (Ester et al., 1996, DBSCAN) and average linkage hierarchical clustering (Sokal and Michener, 1958). To consider possible differences between the two methods, we plotted the dendrogram resulting from the hierarchical clustering in a two-folded way. In Figure 2.6a the color coding is based on a truncation by a distance of 0.75 . In contrast, in Figure 2.6b the color coding represent the labels as they occur in the DBSCAN clustering. First, we see that the small clusters in the hierarchical method are mostly considered as noise by the DBSCAN (in fact we also color-coded clusters of size ≤ 4 in gray to visualize the similarities). Second, we can observe that the large cluster on top of Figure 2.6b[11] is split into two subclusters[12] that are connected at a distance of ≈ 0.78. Finally, we see that DBSCAN splits a larger cluster on the bottom[13] into two clusters.[14] This split would have been performed for a distance threshold of ≈ 0.65.

While having a refinement and a union of two clusters at once when looking at a specific threshold for hierarchical clustering we can further investigate where this split is happening. We can see in Figure 2.7a that the union combines two clusters that are mostly covering the beginning of the summaries. In Table 2.3 we observe the differences between these two clusters. Most prominent, we see a different focus of the participants when summarizing the opening of the plot. While those in cluster 1 (based on the DBSCAN clustering numbering)

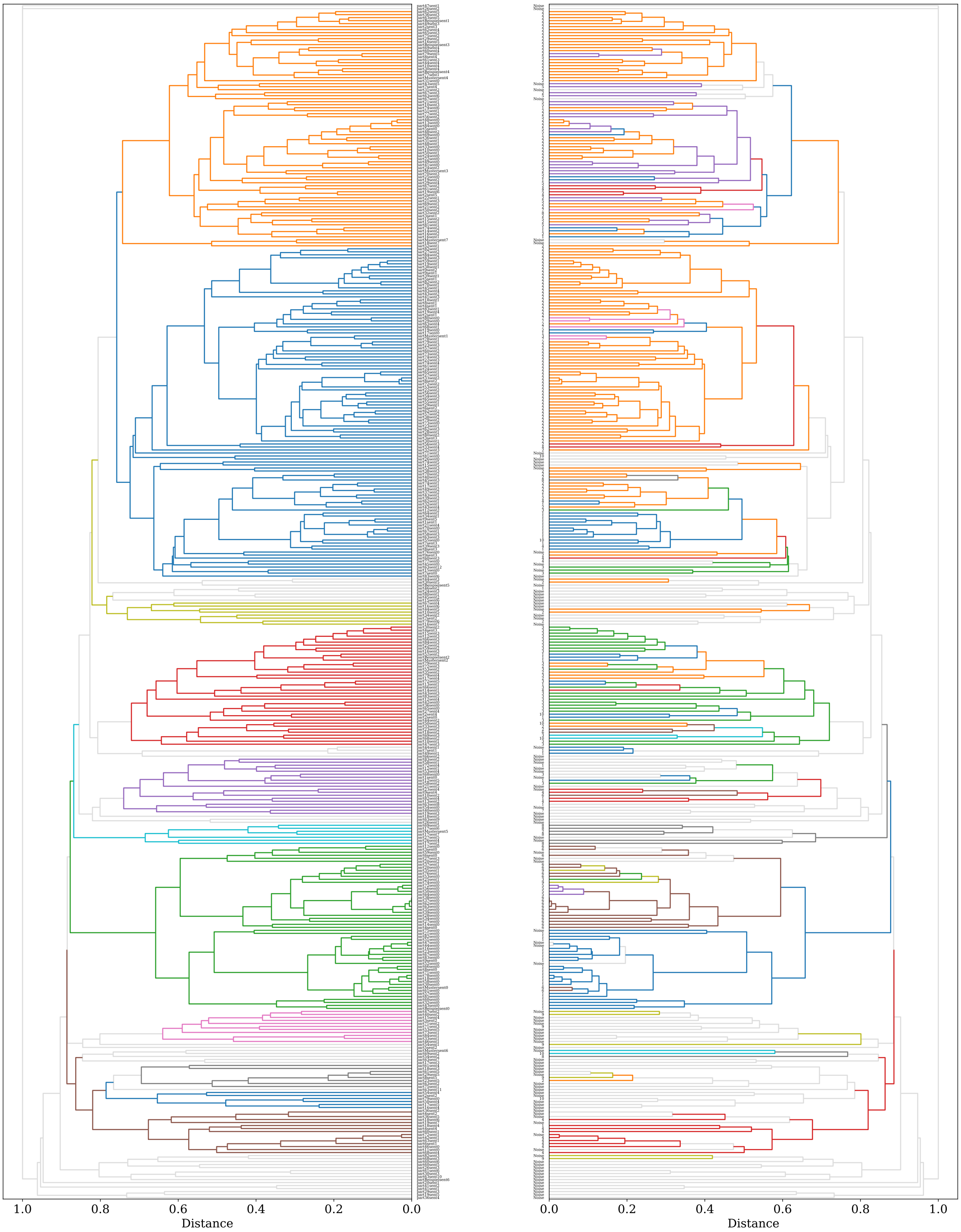
[10] In this case, we look whether there is some form of agreement in content while we discard ordering.

[11] The orange cluster with cluster number 2.

[12] The orange and the blue cluster.

[13] The green cluster.

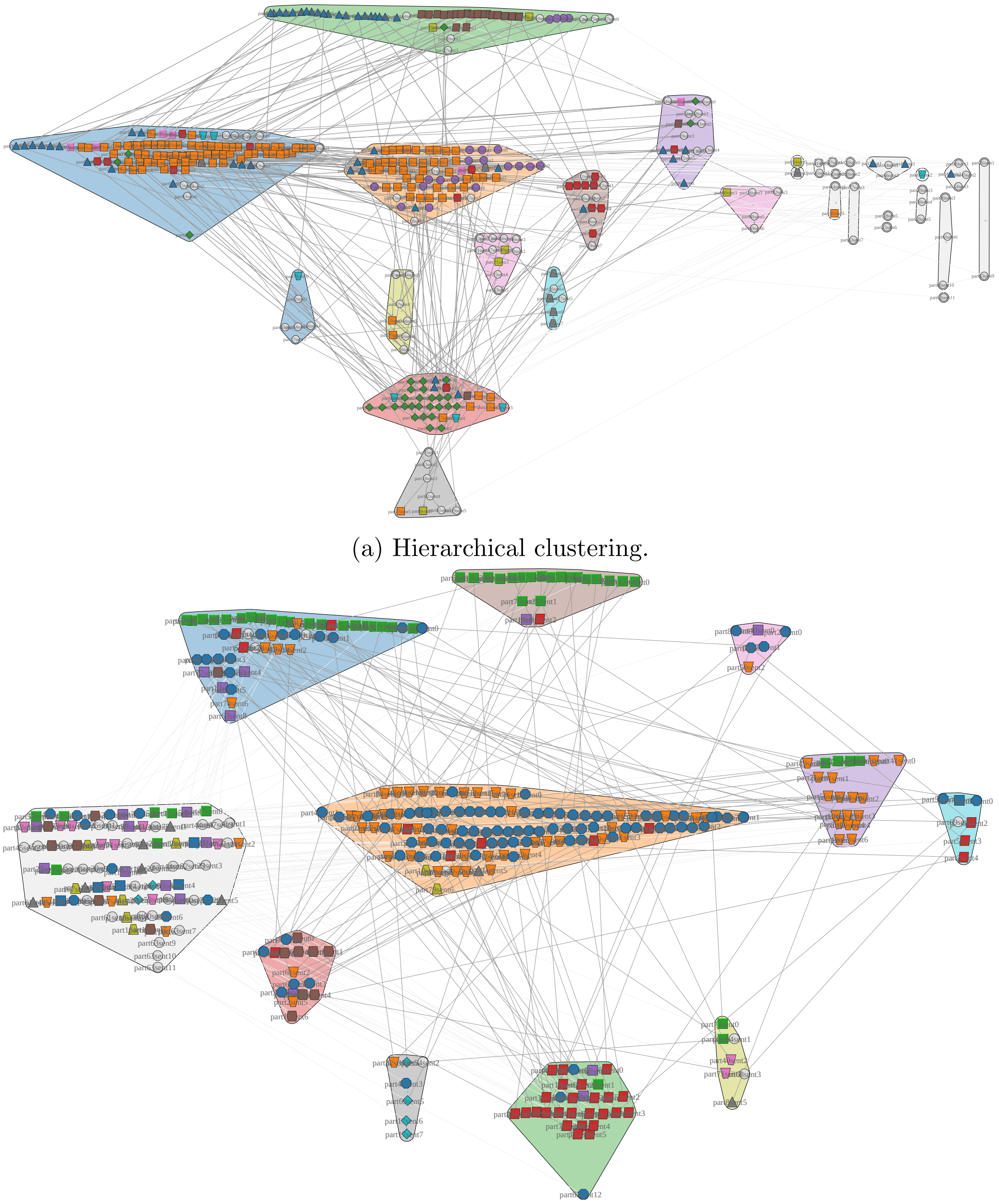
[14] The brown (cluster number 6) and the blue cluster (1).



(a) Hierarchical clustering.

(b) DBSCAN clustering.

Fig. 2.6: The dendrogram of the hierarchical clustering with colors indicating two different clustering algorithms. We use gray for noise.



(a) Hierarchical clustering.

(b) DBSCAN clustering.

Fig. 2.7: The clusters that are formed are plotted as follows. Each summary consists of a series of nodes (representing sentences) connected by directed links (first to last sentence of a summary). Then the nodes are clustered and displayed as group nodes in a way that we can see how many first (top line per cluster), second (second line per cluster), etc. sentences are found in each cluster. Thus, we can do a possible ordering of the clusters by avoiding upward pointing ties. This is indicated by the position of the cluster. Color and shape of each node represents the alternative clustering (i.e., DBSCAN for hierarchical and vice versa). The color is the same as in Figure 2.6.

focus on the narrator[15] the sentences in cluster 6 emphasize the setting of the scene[16]. The opposite is happening in the middle of the story. Here, we observe that the cluster build by DBSCAN is split into two when considering hierarchical clustering. The differences can be seen in Table 2.4. While the DBSCAN cluster takes into account a large part of the story-line, the hierarchical clustering differentiates between the middle part (where the story of Freifrau von Börthe is told by the innkeeper[17]) and the last part of the story (where the father forces his daughter into a marriage). While some of the clustering in the hierarchical tree is considered noise for the DBSCAN we have two clusters that have some alignment. Namely, these are the DBSCAN clusters 3 and 4 labeled in Figure 2.6b and its corresponding clusters for the hierarchical clustering. The alignment is given by a Jaccard index of 0.42 and 0.57 respectively when discarding noise. Here, the Jaccard index is obtained by calculating the ratio between the cases that are classified in the same group by the two methods and this number plus the case that mismatch.

We can look at the most common words of these clusters that are mostly describing the end of the story in Table 2.5. In total these differences can lead to some deviation in the result. However, both methods have their difficulties of finding cohesive groups.

Participants distances. As noted before it appears to be difficult to identify a singular and unique way to summarize the novella. Instead, we want to study the network of dissimilarity of the participants answers. This gives us the possibility to categorize the types of answers that we have received. First, we mark all answers that have a high percentage of their sentences clustered as noise.[18] We expect that these answers themselves are problematic to include in the further analysis.[19]

We now take the clusters and place them in the order as they were given in the answers. This means, we represent each summary by a sequence whose elements are the clusters per sentence. Each of this leaves us with a *word*[20] and hence, we can compute an edit distance (Levenshtein, 1965a,b) on the answer profile. This distance matrix itself can then be interpreted as a network. Here, we start from the complete graph where the nodes are the summaries and the links are weighted by edit distance. In a second step we reduce the number of links to only display connections that are close to each other. A detailed description of the visualization can be found in Chapter 3.2.5. In Figures 2.8 and 2.9 we see how these answers form cohesive groups.

[15] "Erzähler","Mann","Geschichte"

[16] "Weihnachtszeit", "Bild", "Weihnachten"

[17] "Wirtin"

[18] Explicitly, we mark those sentences that have 20%/40% noise in the hierarchical clustering algorithm/DBSCAN algorithm.

[19] The whole list of outliers is given in Appendix A1. By close reading, we can observe that in fact, many of these sentences do not have any human interpretable connection to the investigated novella. This might be either due to the lack of language (however, only few where written by speakers of non-German mother tongue) or motivation to seriously participate in the task.

[20] In the sense of computer science as a finite sequence of characters that are taken from a finite set of symbols, called alphabet.

DBSCAN cluster 1		DBSCAN cluster 6		hierarchical cluster	
Geschichte	19	Weihnachtszeit	9	Weihnachtszeit	23
Land	16	Bild	4	Land	22
Erzähler	14	Weihnachten	4	Weihnachten	11
Dorf	13	Fürsten	3	Mann	10
erzählt	11	Winter	3	reist	10
Mann	9			Stadt	9
Weihnachtszeit	9			Winter	8
Bild	8			Erzähler	8
reist	7			Bild	7
Gemälde	7			Dorf	6
erfährt	6			Fürsten	5
Stadt	6			kaufen	5
möchte	5			möchte	4
Freifrau	5			Gemälde	4
finden	4			erzählt	4
Winter	4			Jahr	4
Ich-Erzähler	4			macht	4
geht	4			Freifrau	4
Weihnachten	4			Natur	4
gehört	4			geht	4
trifft	4			begibt	3
Text	3			Kurator	3
Schloss	3			kommt	3
Kunstwerk	3			Frau	3
Frau	3			berichtet	3
Baronin	3			Text	3
erzählen	3				
Vater	3				
Jahr	3				
macht	3				
Tochter	3				
Fräulein	3				
Wirtshaus	3				
Lebensgeschichte	3				

Table 2.3: Frequencies of the most common words in two clusters of DBSCAN (1, 6) and the corresponding cluster of hierarchical clustering. They mostly cover the beginning of the story.

DBSCAN cluster		hierarchical cluster 1		hierarchical cluster 2	
Freifrau	56	Freifrau	50	Mann	28
Baronin	26	Geschichte	30	Vater	26
Vater	25	Baronin	29	Frau	17
Börte	20	Börte	20	Tochter	13
Mann	18	Wirtin	13	möchte	11
Tochter	14	Bild	12	erzählt	10
Frau	13	erzählt	11	Gemälde	9
Wirtin	13	erfährt	10	wollte	8
Baron	12	Gemälde	8	Baron	8
Gemälde	12	Liebe	7	Börte	8
Geschichte	12	Frau	7	heiraten	7
erzählt	10	Fräulein	7	Hochzeit	6
möchte	10	trifft	6	Mädchen	6
Liebe	9	gehört	6	Liebe	6
Schloss	8	Schloss	6	Geschichte	6
trifft	8	Baron	6	Dorf	6
Bild	7	Erzähler	6	Freifrau	5
heiraten	6	geht	6	Schicksal	4
erfährt	6	lebt	5	Dame	4
Hochzeit	5	Lebensgeschichte	5	Manne	4
geht	5	Vater	5	sehen	3
Spaziergang	5	Schicksal	5	Kunstexperte	3
lebt	5	handelt	5	kann	3
wollte	5	Leben	5	zwangsverheiraten	3
Erzähler	4	Liebesgeschichte	5	Protagonist	3
Verwandte	4	macht	4	Ich-Erzähler	3
Bräutigam	4	möchte	4	reist	3
Geliebten	4	Begegnung	4	Dorfbewohner	3
Manne	4	kommt	4	Schloss	3
Dorf	4	Dorf	4	Geliebten	3
betrogen	3	hört	3	Haus	3
Informationen	3	Einsamkeit	3	Verwandte	3
Lebensgeschichte	3	Besitz	3		
Protagonist	3	besitzt	3		
besitzt	3	Reise	3		
Leben	3	Bilder	3		
machen	3	Bräutigam	3		
Begegnung	3	erklärt	3		
kommt	3	Wanderer	3		
Ich-Erzähler	3	Geschehen	3		
Liebesgeschichte	3	Informationen	3		
Mädchen	3	Wirtshaus	3		
Haus	3	erzählen	3		
handelt	3	Autor	3		
		Erzählungen	3		
		Kunstwerk	3		

Table 2.4: Frequency of the most common words in two hierarchical clusters and the corresponding DBSCAN cluster. They mostly cover the middle part of the story.

cluster 3		**hierarchical 5**		**cluster 4**		**hierarchical 8**	
trifft	14	trifft	19	Bild	6	erzählt	5
Bild	5	geht	6	Geschichte	5	berichtet	5
geht	5	Schloss	5	Erzähler	4	Erzähler	3
Geschichte	4	Spaziergang	5	berichtet	4		
erfährt	4	Bild	5	Bilder	3		
Liebe	3	verirrt	5	erzählt	3		
Reise	3	erfährt	5				
		Freifrau	5				
		Geschichte	5				
		Wanderung	4				
		Reise	3				
		Erzähler	3				
		Dorf	3				
		lädt	3				

Table 2.5: Frequency of the most common words in the DBSCAN clusters 3, 4, and the corresponding hierarchical clusters 5 and 8. They mostly cover the intra diegetic story at the end.

2.3 Results

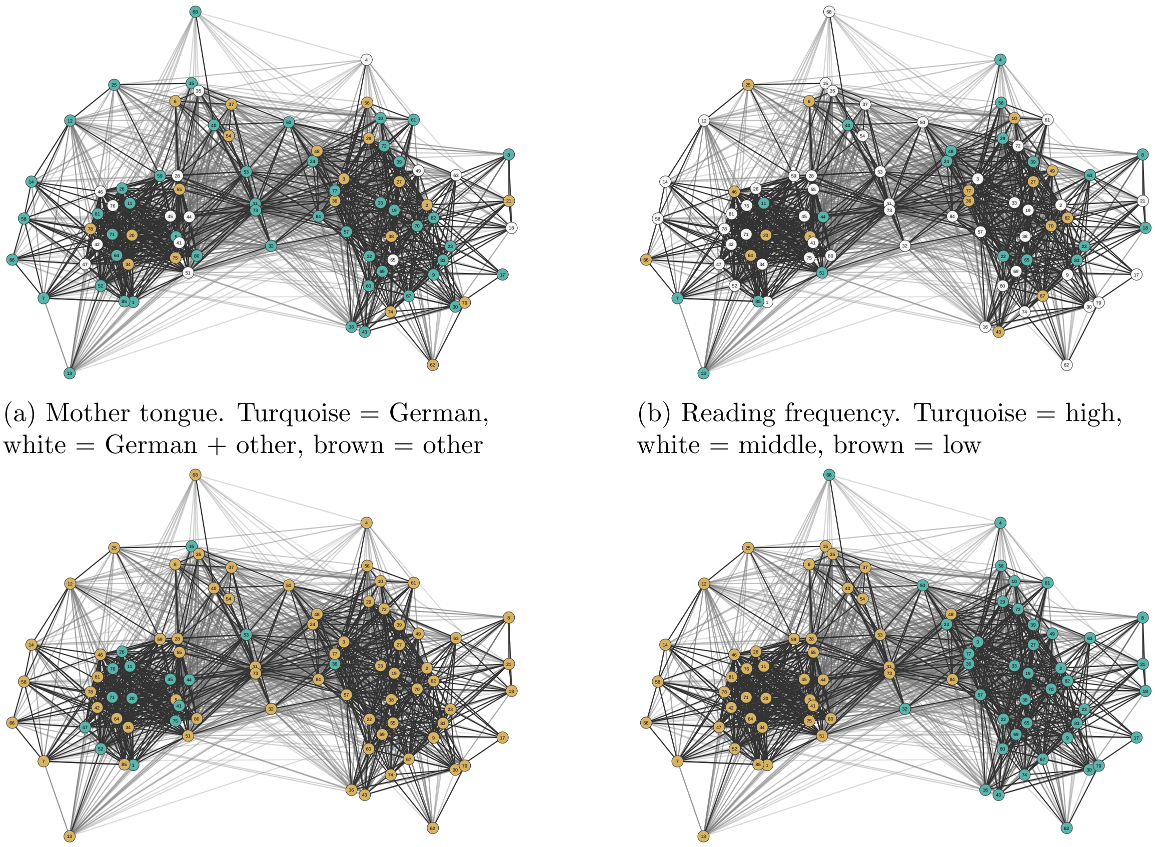
No participant used more than twelve sentences to sum up the novella. However, the used clustering algorithms identified more than twelve distinct clusters. Hence, it does not really make sense to directly connect the clusters to the sentences to find a general agreement on how to summarize this novella from the *Novellenschatz*. Moreover, we already saw in the analysis of our example summaries that while there might be some agreement on cluster level, there seems to be no clear general way of ordering the participants answers. Instead it appears that the answers (asides from containing some noise) can be grouped (e.g. as in Figure 2.8d and 2.9) and we try to identify building patterns of these groups. Consequently, we look at the Louvain clusters that result from the edit distances to investigate common properties of these groups.

We can use statistics to get the distribution of frequent readers and German mother tongue speakers in our sample. From the sample, we can compute a confidence interval in which the correct mean number (μ) of frequent readers and native speakers is located with a probability of 95%. This is done by the following computations. With

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$$



(a) Mother tongue. Turquoise = German, white = German + other, brown = other

(b) Reading frequency. Turquoise = high, white = middle, brown = low

(c) outliers in turquoise

(d) color represents Louvain clustering

Fig. 2.8: Plot of word edit distance with the coloring indicating mother tongue, reading frequency, and Louvain clustering respectively. The plot is based on the 5-out-of-10 Simmelian backbone algorithm (Nick et al., 2013).

we get

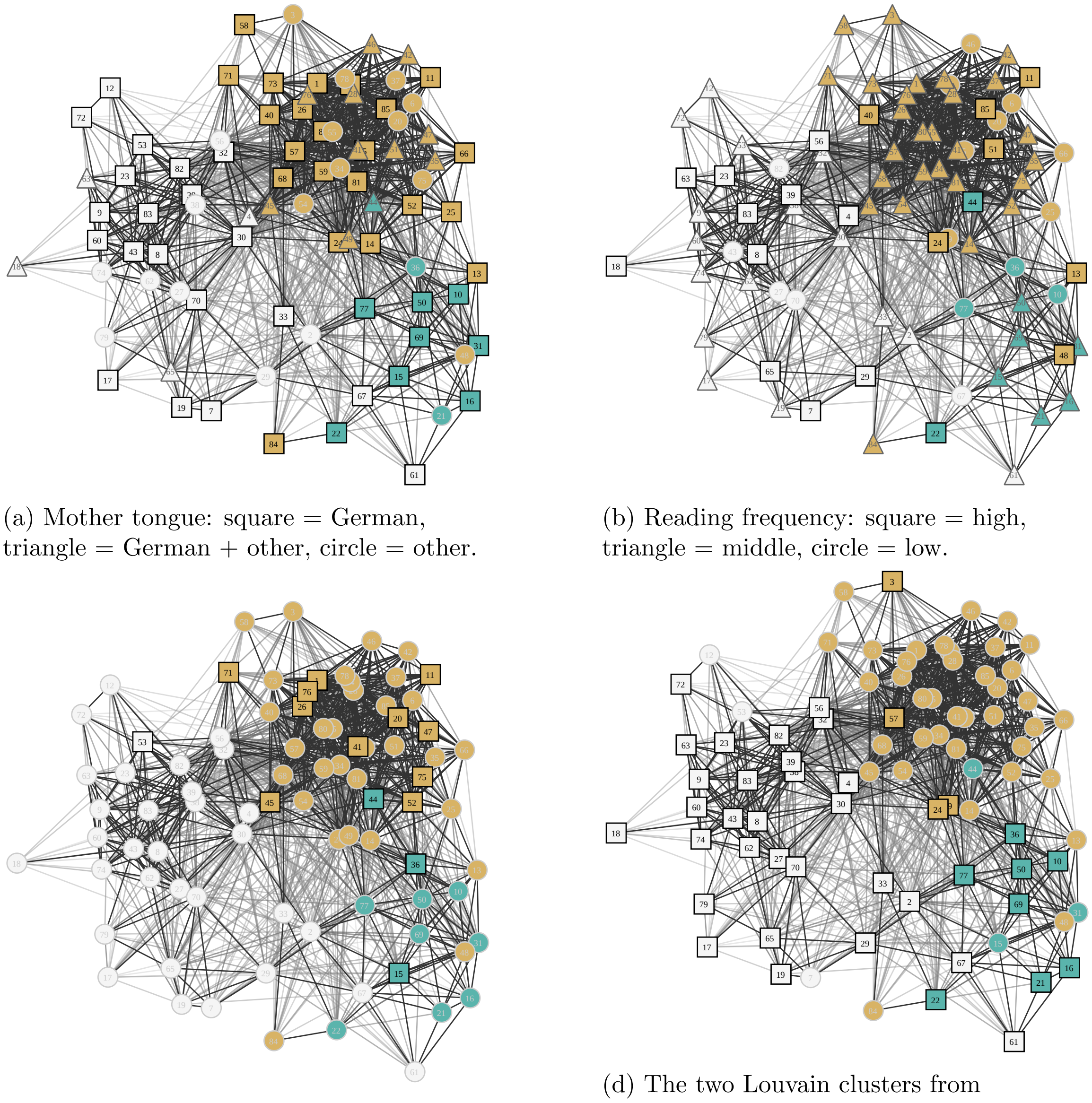
$$\mu \in C_{95\%} = \left[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right].$$

In our case we have $n = 85$, 16 times low reading frequency (encoded as $X_i = 0$), 49 times medium reading frequency ($X_i = 1$) and 20 times high reading frequency ($X_i = 2$). This leads to

$$\bar{x}_{\text{rf}} = \frac{89}{85} \approx 1.0471, \quad s^2_{\text{rf}} = \frac{16 * 89^2 + 49 * 4^2 + 20 * 81^2}{85^2 * 84} = \frac{258740}{606900} \approx 0.4263.$$

Hence,

$$\mu_{\text{rf}} \in C_{95\%,\text{rf}} = [0.9082, 1.1859].$$



(a) Mother tongue: square = German, triangle = German + other, circle = other.

(b) Reading frequency: square = high, triangle = middle, circle = low.

(c) Outliers: square = yes, circle = no.

(d) The two Louvain clusters from DBSCAN.

Fig. 2.9: Word edit distance based on hierarchical clustering. The plots are produced in the same way as in Figure 2.8. The color represents the Louvain clustering. We number the clusters the following way: white = Louvain 1, turquoise = Louvain 2, brown = Louvain 3. In (d) we see the similarity to the word edit distance based on DBSCAN. One can observe that the hierarchical clustering further refines the clustering given by DBSCAN.

In the same way we can get the intervals for mother tongue and outliers. With the distribution described in Chapter 2.1.2 we get the following intervals.

$$\bar{x}_{\text{mt}} = \frac{115}{85} \approx 1.3529, \quad s^2_{\text{mt}} = \frac{429250}{606900} \approx 0.7073, \quad C_{95\%,\text{mt}} = [1.1742, 1.5317]$$

and

$$\bar{x}_{\text{out}} = \frac{15}{85} \approx 0.1765, \quad s^2_{\text{out}} = \frac{346375}{606900} \approx 0.5707, \quad C_{95\%,\text{out}} = [0.0159, 0.3371]\,.$$

If we now consider the Louvain clustering of the edit distance in Figure 2.9d we can see how many mother tongue speakers and how many frequent and infrequent readers are in each group. This is summarized in Table 2.6.

		Louvain 1	**Louvain 2**	**Louvain 3**
reading frequency	**0**	5	3	8
	1	16	6	27
	2	11	2	7
mother tongue	**0**	8	2	10
	1	4	1	10
	2	20	8	22
outlier	**0**	31	8	31
	1	1	3	11

Table 2.6: Reading frequency, language use, and outliers by Louvain clusters.

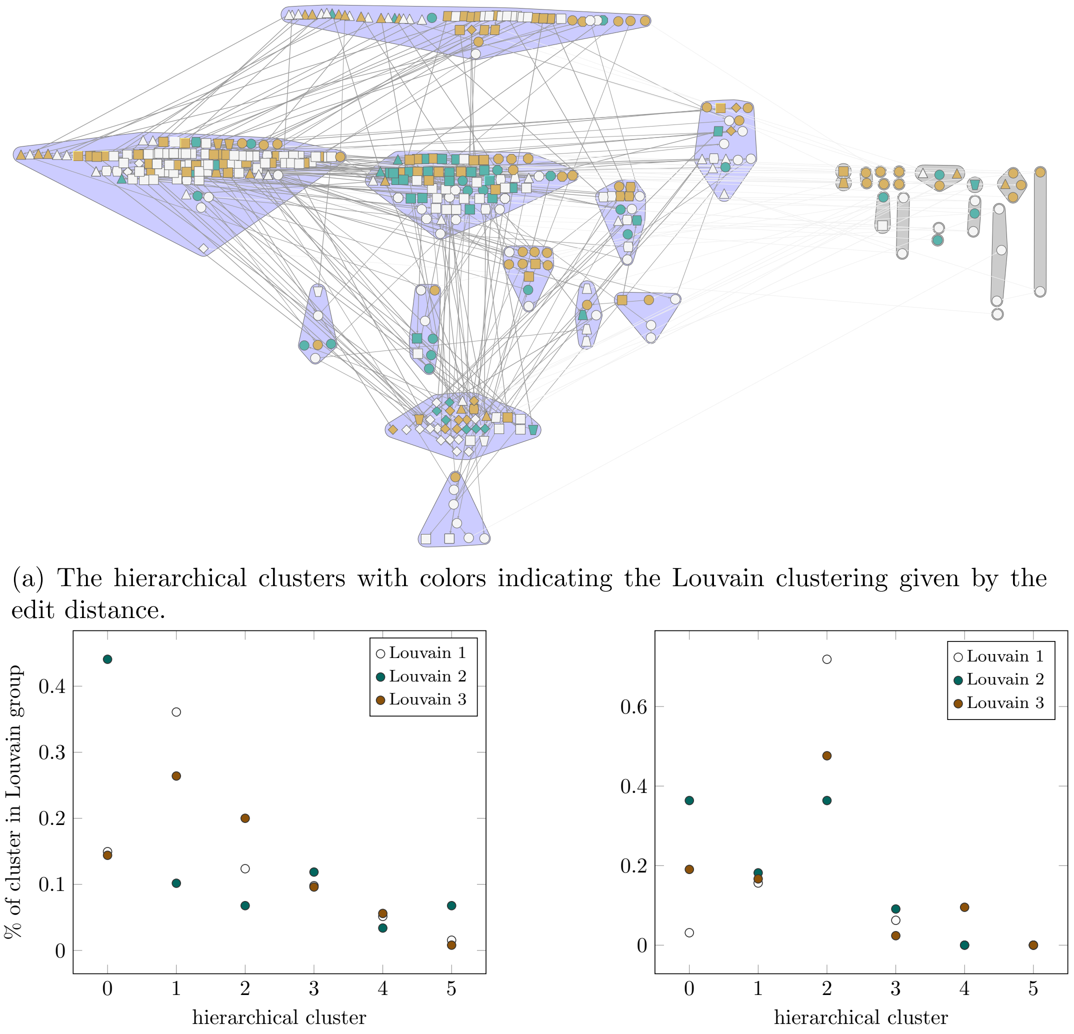
With this we can compute $\bar{x}^i$ for Louvain $i = 1, 2, 3$. They are found in Table 2.7. While

$\bar{x}^i$	Louvain 1	Louvain 2	Louvain 3
reading frequency	0.9091	0.9762	1.1875
mother tongue	1.375	1.5455	1.2857
outlier	0.03125	0.2727	0.2619

Table 2.7: The statistical mean for the values within the groups.

$\bar{x}^2_{\text{rf}}$ and $\bar{x}^3_{\text{rf}} \in C_{95\%,\text{rf}}$, we get $\bar{x}^1_{\text{rf}} \notin C_{95\%,\text{rf}}$. Hence, we can argue the following. While there does not seem to be a general alignment between the participants on how to summarize the novella, there appears to be a group (we call it Louvain one) that distinguishes itself by the way they describe the storyline. Moreover, in this group there is an over-representation of frequent readers. Furthermore, we can look on the number of outliers within the different groups. Here, we can see, that while not statistically significant, there is only one outlier in this group while there are eleven outliers in the 42 samples of Louvain three and three in the eleven samples of Louvain two.

Now, it can be considered how the differences in the Louvain cluster arose. In Figure 2.10 there are different plots that describe the distribution of sentences by Louvain clusters. In Figure 2.10b we observe that Louvain two (turquoise) is characterized by a clear overrepresentation of the central hierarchical cluster[21] (cluster 0). In contrast Louvain one and Louvain three are generally more aligned. However, if we look at Figure 2.10c we see that the differences between Louvain one and Louvain three do arise from ordering instead of content. Louvain one uses nearly exclusively words from the top[22] hierarchical cluster (cluster 2) while Louvain three also uses words from the middle cluster (cluster 0).



(a) The hierarchical clusters with colors indicating the Louvain clustering given by the edit distance.

(b) All sentences.

(c) First sentences.

Fig. 2.10: The distribution of largest hierarchical clusters (containing at least 5% of the sentences within one of the Louvain groups) for the three Louvain groups.

[21] Where we mean *central* based on Figure 2.10b.

[22] Again from Figure 2.10a

2.4 Conclusion

Against our hypothesis, it was not possible to observe a clear agreement on ordering and content. First, one needs to point out that we have not primarily exposed literary scholars but a sample of subjects from different backgrounds (although mostly students) to our questionnaire. In contrast, Heyse's theory was clearly focused on experienced if not professional readers even though it was not stated in such a way. Nonetheless, a pretest performed by German literary students does not make us confident that the result would significantly differ if we would have restricted our participant set to those familiar with 19th century German novellas and novels. Second, while the selection of the novella should not matter, it is very likely that Heyse and Kurz deviated heavily from their own theory. Already the choice of Goethe's *Die neue Melusine* instead of the more obvious choice *Die Novelle* shows that the selection process was driven by more than a fixed format. Hence, it is possible that other novellas (that where not within our reading speed constrains) where more suitable to verify our hypothesis. Nonetheless, it was possible to identify three subgroups that largely agreed in how to summarize the novella. Moreover, one of the groups that had a large agreement contained disproportionately many frequent readers and only few outliers. The answers of this group could also be verified by expert readers to be more meaningful summaries of the underlying text.

Despite giving mixed results, this experiment can be seen as a first example of how literary theory can lead to research questions that can be studied in an uncommon setting. In fact, we used methods that are typical in the social sciences but very infrequent in the humanities. The analysis itself is again different of what we would see in either of them. Instead, the digital humanities perspective on a quantitative and qualitative scale gives us insights into the way, the novella is perceived by its readers. Further ideas on future research are discussed in Chapter 8.2.

Chapter 3

Stylometric Similarity

A distant-reading task in literary corpus analysis is the grouping of stylometrically similar texts (see Chapter 1.1.3). Since there are many ways to define writing style, the result not only depends on the clustering method but even more so on the measure of similarity. With correct authorship attribution as a benchmark much research has addressed the utility of methods for measuring similarity. In contrast, we look at the *Novellenschatz* to demonstrate that one may be interested in very different meaningful groups of texts simultaneously, and that these can be recovered from stylometric clustering if the measure is chosen accordingly. As can be expected, different measures do better at recovering groups associated with, for instance, subgenre, author gender, or narrative perspective. As a consequence, it is suggested that corpus analyses should not be based on what is currently considered the most refined measure of stylometric similarity, but rather break down the decisions that yield a specific measure and provide substantively justified arguments for them.

3.1 Introduction

Authorship attribution from stylometric similarity rests on the assumption that there is an immutable signal that authors emit involuntarily. This signal is often claimed to manifest itself in the use of function words. The utility derived from an author invariant is that it yields a higher similarity between texts from the same author than between texts from different authors, so that authorship can be recovered from clustering similar texts.

Since stylometric similarity can be measured in different ways it appears a fair question to ask which measure is most suitable? Indeed, scholars have championed various measures with strong support articulated, for example, for Burrows's Delta (Burrows, 2002) and Zeta (Burrows, 2007; Schöch et al., 2018). The arguments are largely derived from empirical findings about relative performance, however, and therefore do not necessarily generalize.

Moreover, the same measures have also been used for stylometric analyses of corpora not aimed at authorship but, for instance, genre discrimination (Schöch, 2014). Although individual analyses led to convincing results, it is implausible that signals underlying different categorizations should be comparable and that one notion of stylometric similarity should serve multiple classification purposes.

The purpose of the Chapters 3 and 4 is therefore twofold. In Chapter 3, we want to break down the decisions made during the construction of a similarity measure, interspersed with smaller-scale examples taken from the *Novellenschatz*. In Chapter 4, we want to demonstrate that, indeed, specifically designed similarity measures can identify a variety of meaningful groups of texts in a literary corpus, beyond authorship. This is done by exploring groupings of novellas obtained from different similarity measures.

To keep the discussion concentrated, it is restricted to stylometric similarity in terms of word occurrences. Focusing on the *Deutscher Novellenschatz* as a single corpus is particularly suited for that objective as it is not a representative sample of novellas but the result of a historical process during which the editors aimed for a canonical collection and were acutely aware of compositional effects. Using different stylometric similarity measures, recent studies already found the two novellas of the editors to be central in different clusters of a similarity network (Weitin, 2016), and a group of novellas that appear to have been influenced stylistically by Eichendorff's very last novella, *Die Glücksritter* (Jannidis, 2017).

3.2 Stylometric similarity

As already stated in Chapter 1.1.3, stylometry uses quantification to study writing style. Here we focus specifically on the frequencies of common words as a means to group the novellas in the *Novellenschatz* by similarity.

Document similarity in terms of common words is an established concept in information retrieval (Baeza-Yates and Ribeiro-Neto, 2011). It is often considered in the abstract, with evaluation performed on generic document collections. Here however, the interest is not in the discovery of a specific group of document clusters but in demonstrating differences among results obtained by different methods.

It is worth noting that authorship attribution is a distinct problem for which it is indeed meaningful to compare methods in terms of their ability to identify texts of the same authors with high accuracy. There is less reason to assume that a method designed specifically to discriminate between authors would also serve well to identify stylometric differences across epochs, genres, and other grouping criteria. In fact, this would not even be desirable because of interaction effects between author invariants and other signals.

Instead of proposing a particular notion of (dis)similarity, we therefore break down the process of determining similarity into a number of generic steps. The hope is to thus inform the construction of similarity measures tailored to specific interests.

In the following, we denote with $\mathcal{C} = \{D_1, \ldots, D_N\}$ a collection (the corpus) of documents (the texts). To determine pairwise similarities, each document is characterized by an n-dimensional feature vector $t(D) \in \mathbb{R}^n$ that will be derived from the occurrences of words in document D. A distance $\delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ in the corresponding vector space defines then their dissimilarity.

3.2.1 Features

We will characterize texts by features that originate from word frequencies. The large set of stylometric features (Stamatatos, 2009) thus excluded from the discussion includes n-grams, co-occurrences, word or sentence length distributions, and sentence complexity.

Even in this restricted setting, several choices have to be made. Unlike suggested by the quest for ever more accurate authorship attribution, their relative merit may change with context. A choice may lead to superior analysis in one respect but fail to do so in another. It is thus worthwhile to consider multiple options and evaluate their consequences.

Canonicalization. The first major decision is the granularity at which lexical items are distinguished and whether different meanings of the related items are considered. The resulting classes of lexical items that are treated as equivalent constitute the feature variables $t_1, \dots, t_n$. We refer to them as *terms*, even when they represent entire classes of character strings because these are treated as equivalent manifestations of the same unit of observation.

For instance, we may use all-lowercase to avoid distinguishing occurrences of words at the beginning or in the middle of a sentence. On the other hand, we may still want to retain the distinction between the capitalized version of the German word ‘liebe’ (dear, lovely) and the noun ‘Liebe’ (love).

> Gnädiger Herr, antwortete die Frau mit neuer Betrübniß, meine Liebe trägt die Schuld von alle dem Unglück [...]
>
> Arnim, *Der tolle Invalide auf dem Fort Ratonneau*
>
> Warum wollen Sie so rasch fort von hier, liebe Emma?
>
> Grimm, *Das Kind*

Similarly, stemming, lemmatization, and even more inclusive abstractions are often used to eliminate undesired distinctions (Figure 3.1).

Multiple lexically different references to a named entity such as ‘Meran’ can be counted towards that same entity, or not. Consequently, the number of occurrences of a particular place (Meran), of a class of places (city), or of any kind of place (locations) yield different features and thus lead to different similarities later on (Figure 3.2).

The degree to which syntactic, semantic, and contextual information is used to aggregate or disaggregate the words of a document into classes that define term features t_i, $i = 1, \dots, n$, will have a profound impact on which texts are found to be similar.

Counting occurrences. With the units of observation decided upon, we turn to determining their associated values. A straightforward measurement considers raw counts: each time a member of an equivalence class (say, all words associated with the infinitive ‘to be’) is encountered, it contributes one unit towards the value of the associated variable.

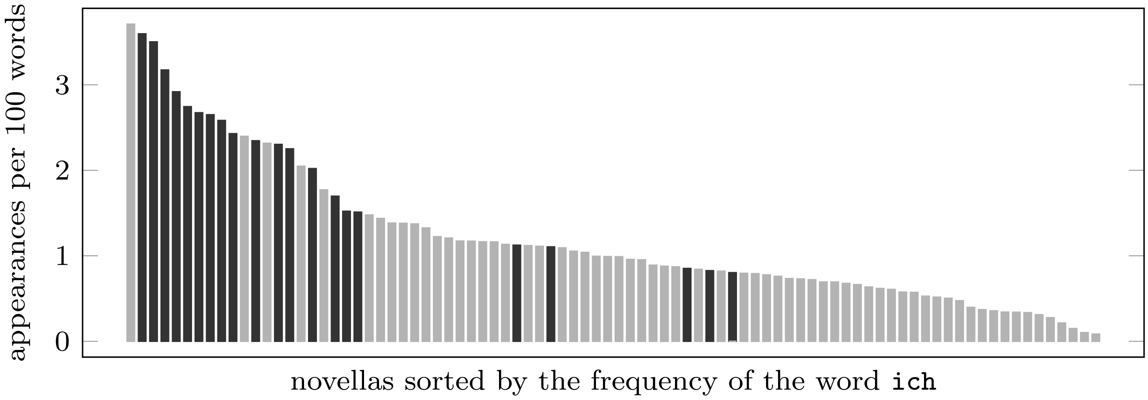
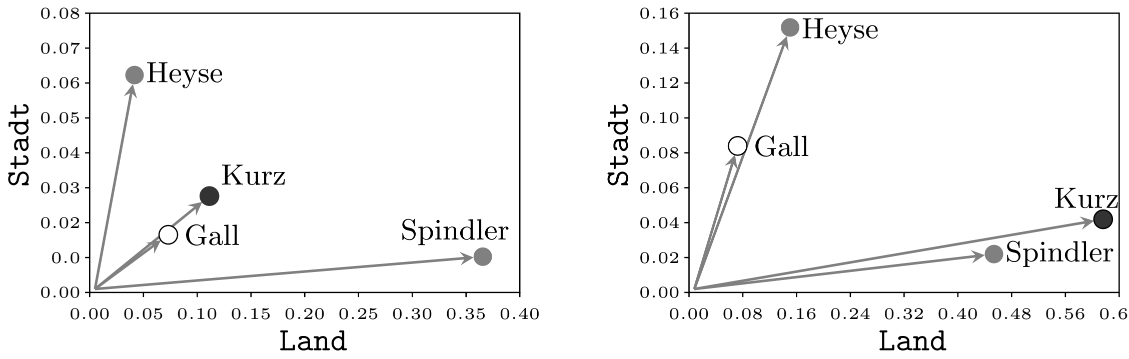


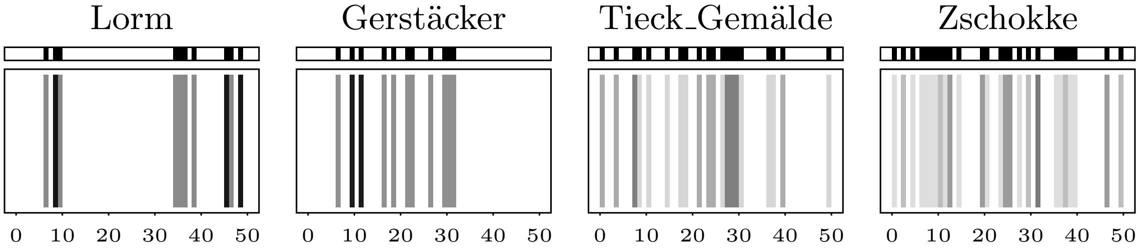
Fig. 3.1: Relative frequencies of the word `ich` in the novellas with first-person narratives marked in black. Without changing pronouns to a generic term, we expect a strong connection of first-person narratives as an artefact of the method.



(a) Term frequency vectors for the words `Stadt` (urban setting) and `Land` (rural setting) in four novellas.

(b) The same vectors after replacing named locations by the appropriate generic term `Stadt` or `Land`.

Fig. 3.2: By using a dictionary to map named locations to classes identifiers of locations we note a change of document frequency vectors of Kurz's *Die beiden Tubus* and Gall's *Eine fromme Lüge*. Up to scaling, there is a strong similarity between the vectors of Kurz and Gall on the left. However, merging particular places into generic classes leads to a situation in which the novellas of Gall and Kurz are rather dissimilar and much closer to the texts of Heyse and Spindler, respectively.
In this example, we have mapped `Basel, Wien, Bordeaux, Laibach, Verona, Meran, Innsbruck, Venedig, Berlin` and `Bremen` to `Stadt`, whereas `A...berg, Appenzell, Aarlberg, Bernerland, Y...burg, Sch...ingen, Tirol, Burgland, Etschtal, Küchelberg, Vitschgau, Algrund` and `Trautmannsdorf` to `Land`.



(a) 13.60/1000 words (b) 13.67/1000 words (c) 13.25/1000 words (d) 13.64/1000 words

Fig. 3.3: Appearance profile of the word `Vater` (father). Each novella is split into 50 equal slices and we denote a gray bar if the term appears in a slice and a white bar otherwise where darker color indicate a higher accumulated appearance in the slice. In Lorm's *Ein adeliges Fräulein* and Gerstäcker's *Germelshausen* we can identify the present story within a story by the accumulated occurrence of the term that is not present in the texts of Tieck and Zschokke.

Raw counts do not discriminate between occurrences of a particular word that are highly concentrated in one section or spread over the entire text. In variations, a text may therefore be chunked into segments, say of equal length, and each segment is quantified separately, or one occurrence is counted if and only if the raw count or relative frequency in a segment surpasses a threshold (Figure 3.3).

While the later analysis is highlighting stylometric and content-wise differences within a text, the former is suited for highlighting stylometric differences between texts. To keep the discussion focused, we will concentrate on the differences between texts. In the following we will use $t_i(D_k)$ to denote the count of the ith term (equivalence class of words) obtained for the kth document (corpus text).

Filtering. For common levels of granularity the number of terms in a typical document collection is rather large. It usually includes many elements that occur only rarely or are not informative for other reasons. Function words, for instance, are considered essential elements in stylometric analysis for authorship attribution but irrelevant in topic analysis. Terms that appear only occasionally or in few documents may be informative or constitute noise. Weitin (2019) used entropy analysis to illustrate this point.

Filtering serves to obtain a dimensionality n, of the feature vector that is much smaller than the total number of terms actually present. It is often based on a dictionary (e.g., blacklisting of stop words) or numbers of occurrences (e.g., top-n most frequent words) (Figure 3.4).

Culling is a variant of the frequency-based approach in which a word must, in addition to being frequent, appear in a minimum number of documents (Figure 3.5).

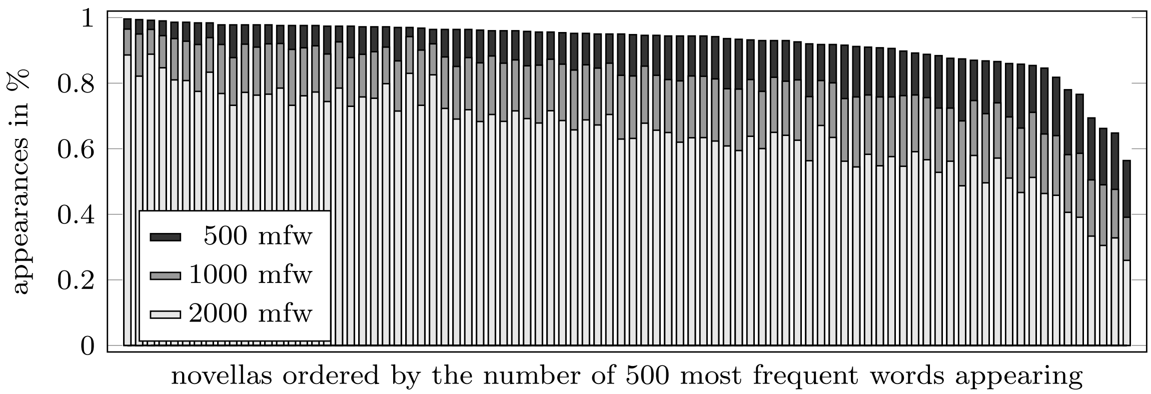
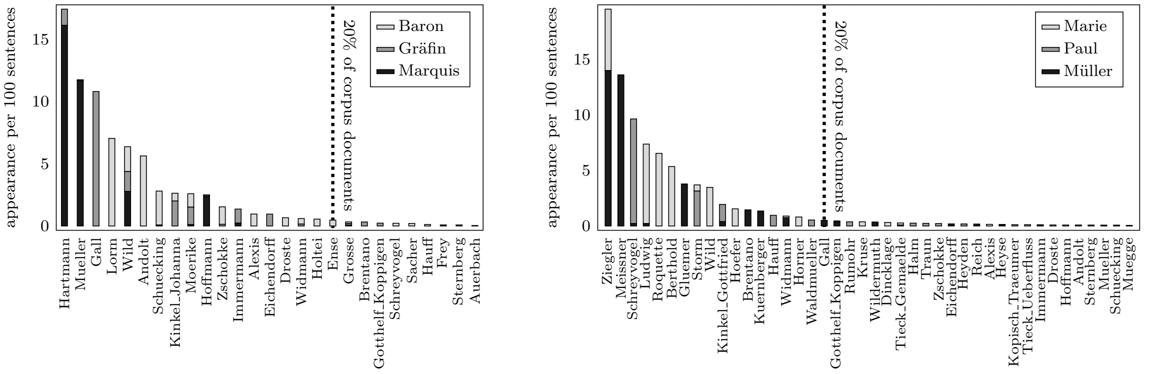


Fig. 3.4: While nearly every word in the 100 most frequent words occurs in every text, differences are more pronounced in the tail of the most frequent word vector. Novellas containing many words in the tail are affected by an increased dimensionality.



(a) `Baron`, `Gräfin` and `Marquis`

(b) `Marie`, `Müller` and `Paul`

Fig. 3.5: Names make up 27 of the 34 words removed by 20% culling. Three of the other seven are `Baron`, `Gräfin`, and `Marquis`. These titles are present each in 15 or 16 of the 86 novellas and therefore close to the cutoff, which they pass if combined. A similar situation arises for the three most common names. While there may be substantive reasons such as a genre signal to combine noble titles, this is not expected to be equally meaningful for character names.

3.2.2 Normalization

After determining features by deciding what to count, the comparison of texts by their feature vectors requires adjustments in order to take into account that frequencies may have different baselines in different texts.

A straightforward quantity to control for is the length of the text from which a feature is derived. A common normalization is therefore the share of occurrences counted toward a particular feature,

$$\mathrm{tf}(i, D) = \frac{t_i(D)}{|D|} \ ,$$

where $|D|$ is the text length of a document D in the corpus. The resulting quantity is called *relative term frequency.* Although length-normalization yields a distribution of occurrences, we may need to establish a corpus baseline to identify the special role of a text in a corpus. Therefore, we will use the *tf-idf*-score that has been introduced in Chapter 2.2. In our setting it is given by

$$\text{tf-idf}(i, D) = \mathrm{tf}(i, D) \cdot \log \frac{N}{N(i)} \ ,$$

where $N = |\mathcal{C}| = 86$ denotes the number of documents in the corpus and $N(i) = |\{k = 1, \ldots, N : t_i(D_k) > 0\}|$ the number of documents in which the term i appears.

A different weighting is obtained by comparing relative frequencies across documents. Assuming a normal distribution we let $\mu(\mathrm{tf}(i))$ denote the expected relative frequencies of term i across all documents, i.e., the average over the entire corpus, and $\sigma(\mathrm{tf}(i))$ its standard deviation. Then,

$$z_i(D) = \frac{\mathrm{tf}(i, D) - \mu(\mathrm{tf}(i))}{\sigma(\mathrm{tf}(i))}$$

defines the *z-score* of the ith term in document D. It is thus positive (negative), if the relative frequency of $\mathrm{tf}(i, D)$ in D is higher (lower) than expected across the corpus, where differences are scaled by their standard deviation, and thus made comparable. Since z-scores vary greatly for rare terms, they are generally used for the most frequent terms only.

It may seem so far that normalization is largely with respect to a document or the corpus. For relative frequencies of all terms in any document $D \in \mathcal{C}$ we have, of course, $\sum_i \mathrm{tf}(i, D) = \sum_i t_i(D)/|D| = 1$, i.e., the relative frequencies of all terms sum to one in each document. Depending on the distances used later on, normalization may be more suitable if with respect to a different norm. With Euclidean distance, for instance, we obtain a vector of unit length for D from the division of each entry by its vector length $\|t(D)\|_2 = \sqrt{\sum_{i=1}^{n} t_i(D)^2}$. In comparison to relative frequencies, the use of Euclidean unit-length normalizations puts relatively more weight on larger deviations. In addition to the above within-vector normalizations, Büttner et al. (2017) list two variants that use thresholding. The first introduces a lower and an upper bound and clips frequencies outside of this interval to its boundaries. In a stronger discretization, the second variant replaces each entry by -1, 0, or 1 depending on whether the term is infrequent, standard, or frequent relative to the document or the entire corpus.

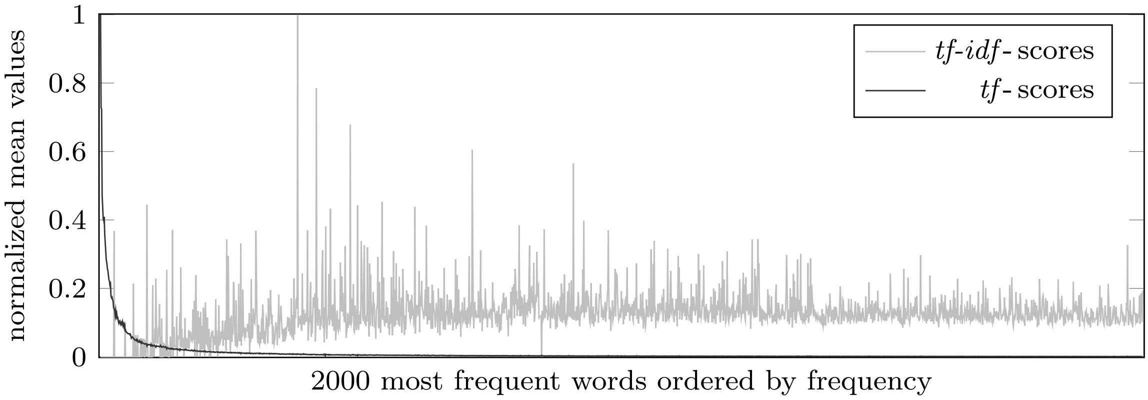


Fig. 3.6: We plot the mean *tf*- and *tf-idf*- score over all novellas. We observe the well-established Zipf's law (Zipf, 1935, 1949) for the *tf*- scores. Hence, any analysis that uses absolute differences of this score will be highly skewed towards the very frequent words while the *tf-idf*- score is reaching its peak at the medium frequent words. Meanwhile, the z - score does not discriminate between frequent and infrequent words. In fact, its mean is zero by definition.

3.2.3 Dissimilarity

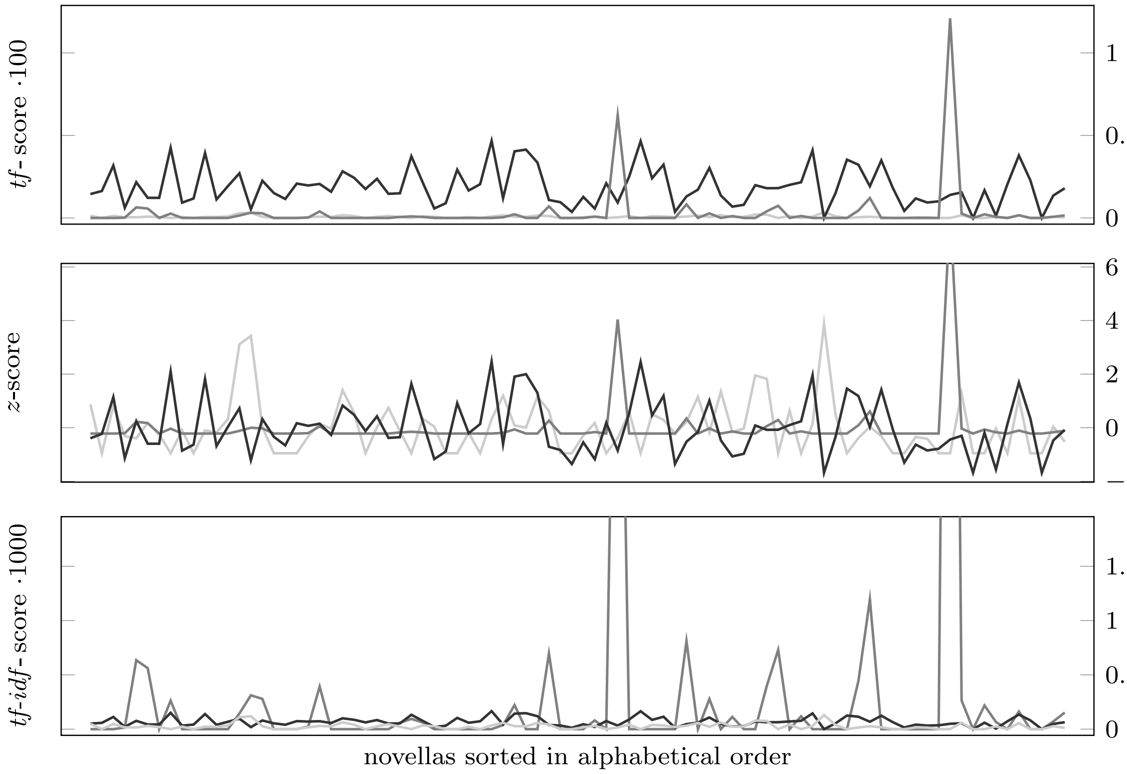
The basic question for which we are trying to find quantitative answers is whether two texts are similar with respect to the prevalence of words. We have already discussed that such comparison requires us to be precise about the term features that words are aggregated into, and the way we normalize frequencies across a document, the corpus, or with respect to each other. However, we also have to be clear about how to compare the frequencies of each term and how to aggregate their individual differences. It is a task-specific question, for instance, whether large differences with respect to some specific terms imply the same level of dissimilarity as many small differences across the board.

Options to assess such trade-offs have been discussed extensively (Burrows, 2002; Argamon, 2008; Smith and Aldridge, 2011; Sidorov et al., 2014) although generally in attempts to establish the superiority of one measure over others. We briefly review the most commonly used concepts.

The distance of two numbers $a, b \in \mathbb{R}$ is computed as $|a - b|$. In mathematics, a norm is an extension of the absolute value that can, among other purposes, be used to translate the concept of a distance to multidimensional objects. In stylometric analysis we can measure the distance of two feature vectors $x, y \in \mathbb{R}^n$ that result from some selection of terms and a normalization of their occurrence counts (e.g. *tf*-scores or z-scores) by

$$\delta_p(x, y) = \frac{1}{n} \cdot \|x - y\|_p^p = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^p \ ,$$

where $\|x - y\|_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$ is the so-called p-norm for $p \geq 1$. For $p = 1$, we



	μ (σ)	*tf*- score ·100	*z*-score	*tf-idf*- score ·1000
—	`ihre`	0.1891 (0.1126)	0.0 (1.0)	0.0672 (0.0400)
—	`Pfarrer`	0.0317 (0.1460)	0.0 (1.0)	0.3139 (1.4434)
—	`eigenes`	0.0069 (0.0072)	0.0 (1.0)	0.0259 (0.0271)

Fig. 3.7: Comparison of the *tf-idf*- score, the *tf*- score and the *z*-score of `ihre` (63rd mfw), `Pfarrer` (378th mfw), and `eigenes` (1247th mfw). While the *tf-idf*- score is low for the frequent and the infrequent words `ihre` and `eigenes` it is high for the medium frequent word `Pfarrer` which is very distinct for certain texts. In comparison, the term frequency of `ihre` is much higher for almost all novellas than for the other terms. With the *z*-score, these differences are balanced out by design.

obtain the average absolute difference over all feature values also known as the Manhattan distance. Burrows's Delta (Burrows, 2002) is the application of δ_1 to feature vectors that consist of the z-scores for the n most frequent words.

For $p = 2$ we get the distance in Euclidean space, and δ_2 is also referred to as Quadratic Delta in the present context (Argamon, 2008). In fact, the larger p, the more emphasis is put on entries with large differences, and, as p approaches infinity, only the maximum difference matters, $\|x - y\|_\infty = \max_{i=1}^n |x_i - y_i|$. Rather common choices are $p = 1, 1.4, 2, 4$ (Büttner et al., 2017).

The feature vectors are called vectors since they can be visualized as arrows in the n dimensional space. The δ_p measures above have in common that they focus on the distance of the endpoints of those vectors. A different approach is often used in information retrieval to find documents similar to a query text. With

$$\mathrm{sim}_{cos}(x, y) = \frac{1}{\|x\|_2 \cdot \|y\|_2} \cdot \sum_{i=1}^{n} x_i \cdot y_i \ ,$$

we can compute the cosine between the angle of the two vectors and hence, this is called *cosine similarity.* The reverse, $\delta_{cos}(x, y) = 1 - \mathrm{sim}_{cos}(x, y)$, is called *cosine distance.* It places the focus on the similarly skewed distributions of weighted term frequencies, rather than individual frequencies. If the feature vectors are already normalized, δ_{cos} is essentially the same as δ_2 since $\|x\|_2 = 1 = \|y\|_2$ implies $\|x - y\|_2^2 = 2\delta_{cos}(x, y)$.[1]

Independent of the particular term occurrence-based construction, we can compute the feature vectors $x(D_1), ..., x(D_N)$ with documents $D_1, ..., D_N$ in the corpus $\mathcal{C}$. With any dissimilarity measure δ defined on them, we obtain a document-dissimilarity matrix $\Delta = (\delta_{k\ell}) \in \mathbb{R}_{\geq 0}^{N \times N}$ with entries $\delta_{k\ell} = \delta(t(D_k), t(D_\ell))$. This matrix summarizes the relationships between the texts in the corpus with respect to the frequencies of words from which they are composed.

3.2.4 Clustering, Scatterplots, and Networks

To structure a corpus into groups of texts that are similar within, and dissimilar across groups, a matrix of dissimilarities is constructed as outlined in the previous section and

[1] A generalization of cosine similarity has been proposed in Sidorov et al. (2014). They suggest using an additional term-similarity matrix $S = (s_{ij}) \in \mathbb{R}_{\geq 0}^{n \times n}$ to weight the contributions of all pairs of entries x_i, y_j. This *soft cosine* similarity measure

$$\frac{\sum_{i=1}^n \sum_{j=1}^n s_{ij} x_i y_j}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n s_{ij} x_i x_j} \sqrt{\sum_{i=1}^n \sum_{j=1}^n s_{ij} y_i y_j}}$$

reduces to cosine similarity for $S = I_n$, i.e.,

$$s_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j, \end{cases}$$

where terms are similar only to themselves. Similarities can be motivated syntactically or semantically, and thus allow for a more nuanced aggregation than the combination of multiple words into a single term feature.

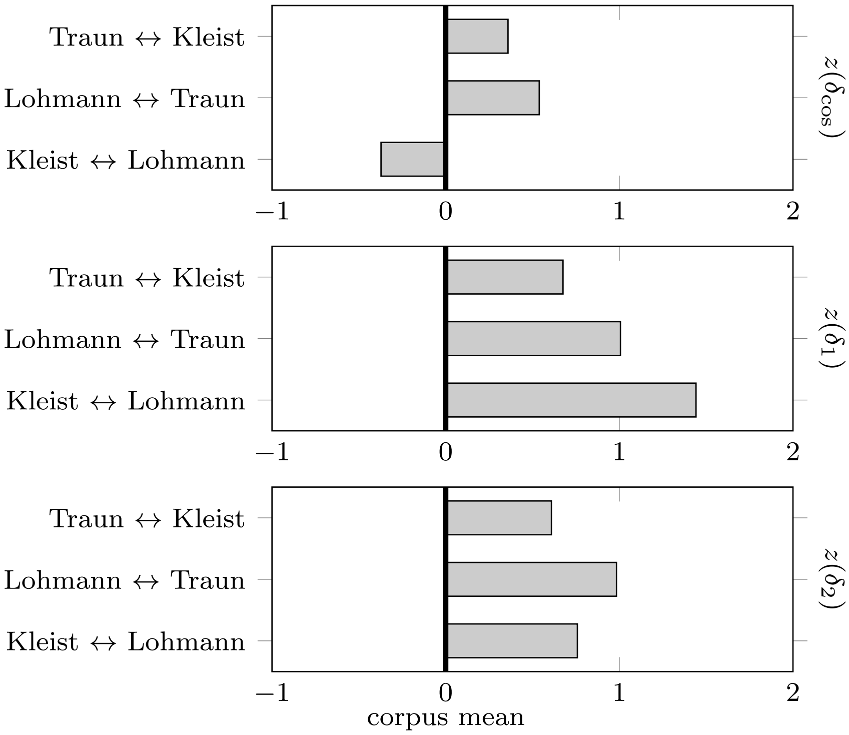


Fig. 3.8: We compare pairwise distances of the 500 mfw for the three novellas of *Kleist*, *Lohmann* and *Traun*. To make distances comparable we show the discrepancy to the mean in units of standard deviation (i.e., z-scores of distances). For cosine distance, Kleist and Lohmann are closest whereas Lohmann and Traun are far apart. For Burrows's Delta the novellas of Kleist and Lohmann are far apart and Traun and Kleist have the smallest distance. Finally, for quadratic delta, the distance of Lohmann and Traun is larger than the one of Kleist and Lohmann.

then subjected to a clustering method. As seen in Chapter 2, each clustering method strikes a different balance between the number of groups, group sizes, intra-group similarity, and inter-group dissimilarity. Beside the already used and described hierarchical clustering and DBSCAN there is an abundance of research on clustering methods for various contexts (Estivill-Castro, 2002; Berkhin, 2006).

Since we do not want the effects of different notions of word occurrence-based dissimilarity measures to be confounded by the selection of a particular clustering method, we refrain from applying one at all. Instead, we represent the dissimilarity matrices such that groupings likely to be stable across multiple clustering methods are recognizable.

To explore and present similarity-based clusterings it is typical to depict results obtained from dimensionality reduction methods such as a principal component analysis (PCA) (Pearson FRS, 1901; Hotelling, 1933), multidimensional scaling (MDS) (Torgerson, 1952, 1958; Kruskal, 1964), or t-SNE (Hinton and Roweis, 2002; Maaten and Hinton, 2008) as scatterplots.

However, pairwise dissimilarities are necessarily distorted when projecting them into only two or three dimensions. Whether this affects the recognizability of clusters depends on the context. Hence, to represent dissimilarity matrices we choose networks.

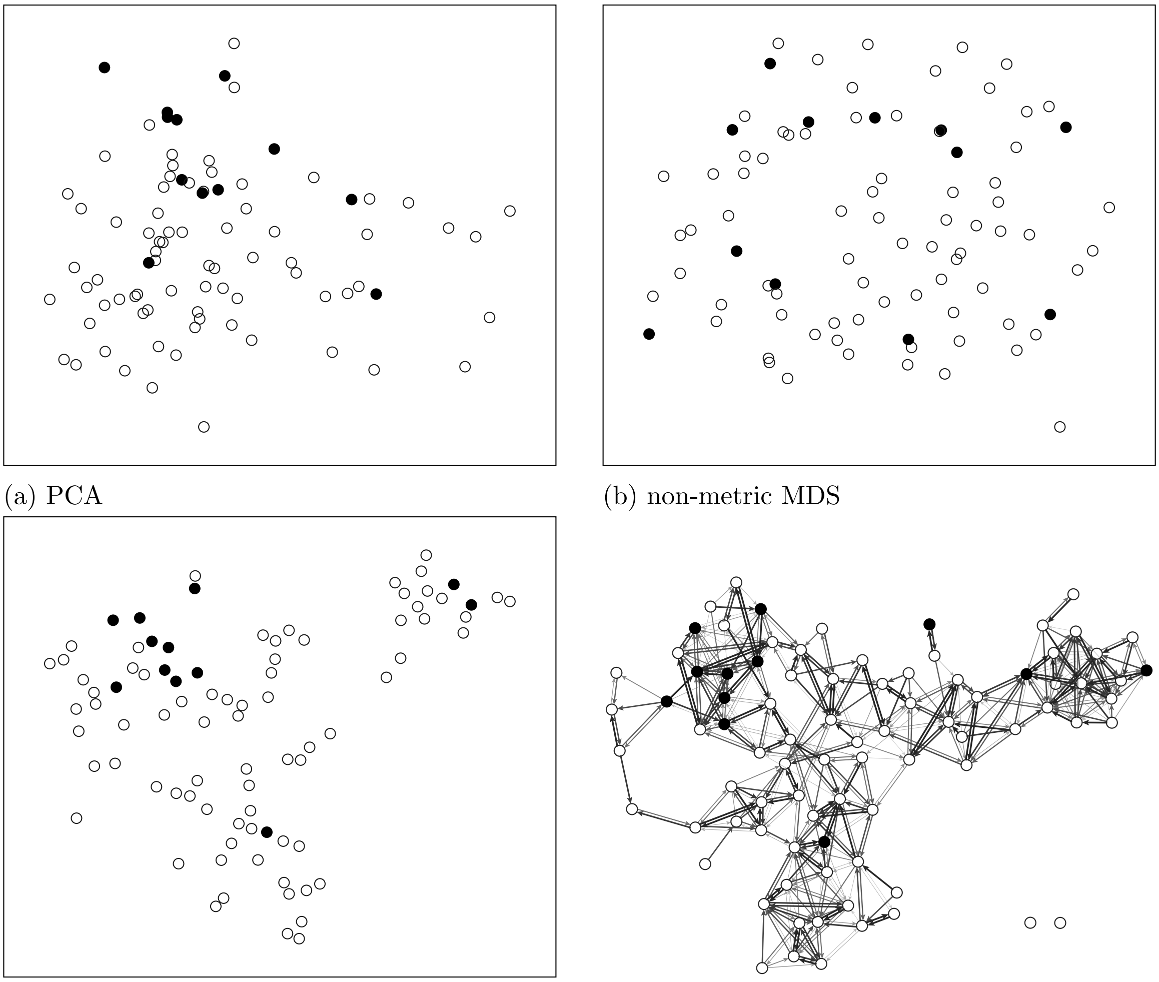
Consider as an example the boxed scatterplots in Figures 3.9a, 3.9b, and 3.9c. All three represent the same δ_2-distance matrix between feature vectors containing (case-sensitive) term frequencies of the 100 most frequent words in the *Novellenschatz* corpus. Figure 3.9a thus reproduces Figure 6 from Jannidis (2017) where the group of female authors is found to cluster in the upper left quadrant. This is of interest because PCA maximizes variance one dimension at a time. Female authors are located in a group that is somewhat recognizable in the primary (horizontal) dimension and clearly distinguished along the secondary (vertical) dimension.

The other two scatterplots are obtained from other common dimensionality-reduction methods. In the scatterplot obtained from MDS, axes are not relevant. Positions are determined instead to minimize the percentage error in the representation of dissimilarities by distances in the scatterplot. The MDS scatterplot suggests that female authors are rather peripheral but do not cluster.

PCA and MDS generally provide an overview of similarities and dissimilarities in terms of spatial distance, but may fail to represent clusters well because of overplotting in low-dimensional display space. The more recent non-linear dimensionality-reduction technique, t-SNE that has also been used in Figure 2.5a and Figure 2.5b yields a more pronounced clustering overall, and for all but three female authors in particular.

All three methods are designed to minimize misrepresentation of input distances in low-dimensional output space, but their underlying objectives represent different trade-offs between large and small misrepresentations of large and small distances. The focus on nearby nodes in the formulation of t-SNE, for instance, is apparent in the result.

Even for apparent groupings, uncertainty remains. Each dimensionality-reduction method represents a unique compromise, and it is difficult to assess whether what appears to be a strong clustering is really a coincidence resulting from that compromise.



(a) PCA

(b) non-metric MDS

(c) *t*-SNE

(d) 5-out-of-10 backbone

Fig. 3.9: Three two-dimensional scatterplots and a network backbone representing the (dis)similarities between texts in the *Novellenschatz* based on the 100 most frequent words. Black dots highlight novellas written by female authors. Note that the backbone is a structure with no prescribed geometry; the layout has been determined for clarity but should be considered flexible.

3.2.5 Backbone networks

In an attempt to represent dissimilarities without introducing projection artifacts, we opt for a representation that is more qualitative and retains more degrees of freedom.

A (dis)similarity matrix can be viewed as a complete weighted network, and one way to reduce it to only the strongest similarities is by thresholding. However, a certain level of similarity that may be high for one text can be comparatively low for another, for instance, because the latter is part of a group of mutually similar texts. An alternative, based on relative rather than absolute similarity, are nearest-neighbor networks in which relationships with the most similar other texts are retained.

We here use a restricted variant of Simmelian backbones (Nick et al., 2013), a filtering technique that uses locally adaptive thresholds considering also the vicinity around a pair of texts. It is designed to keep only those pairwise relationships that are relatively strong and sufficiently reinforced by joint similarity to others.

For each novella, the other novellas can be ranked by their similarity to the first, and we consider a fixed number, say ten, of the most similar ones, independent of their absolute similarity score. The networks resulting from this first step are nearest-neighbor networks. They are weighted by similarity rank and also directed, because a novella may be among the most similar of another without the reverse being the case as well.

In a second step, we remove all those relationships where the similarity may be relatively high but not indicative of joint group membership. The relationship of a text with a neighboring text is retained only if it has, among its ten most similar others, at least five of the neighbors of the first text. This results in a tendency to maintain relationships with texts that are not relatively more similar to other groups.

The backbone network in Figure 3.9d corroborates the groupings from the PCA and t-SNE scatterplots but also indicates that at least one novella with a female author is less strongly linked to the main cluster of female authors. Emmy von Dincklage's *Der Striethast* is located above but close to eight other female-authored novellas in the upper left of the t-SNE scatterplot. The backbone network suggests that this is an artifact of the projection because it does not retain any link to that group. Dincklage's novella is a pendant vertex in the top middle of the backbone network and shows strong similarity only to one novella of a male author (Sternberg's *Scholastika*).

The layout of a network is not given but generally determined such that densely connected groups are placed closer together, and unconnected parts farther apart to allow for an interpretation similar to scatterplots. Still, no deeper meaning should be read into the layouts, as the analytic information is in the local structure of links rather than relative positioning of nodes. We use standard network visualization tools and some manual editing to make the structure visible where the layout algorithms do not and to reduce differences in layouts across similar networks. Note that, in principle, we could have used scatterplot positions for layout; and the links would still add information on where the actual similarities are. Despite being close in t-SNE coordinates, Dincklage's novella would still be recognized as more dissimilar from the novellas of other female authors because the backbone contains no link between hers and theirs.

3.3 Conclusion

In the section above we displayed how decisions on a fine granular scale influence the building of a similarity score at large. In the next chapter We will discuss in detail how these micro level changes have an impact on the similarity scores at large. Nonetheless, we already want to give a first hint how the decision points influence the analysis.

We have shown that it is possible to have control over the similarity measure building process and its visualization and interpretation. Therefore, it is feasible to use adapted measures to tackle research questions that arise from a corpus that is not focused on authorship attribution. Moreover, one can use similarity measures to excavate cohesive groups that where preliminary invisible to the researcher. In this case, the understanding of the underlying mechanisms that shaped the similarity measure can lead to insights (or hints that can then be tested by other hermeneutical methods) about the located cliques.

Chapter 4

Demonstrations

4.1 Introduction

In the last chapter we have seen possible decision points when applying stylometric similarity measures to a text corpora. Now we want to demonstrate the effect of some of these decisions on the *Novellenschatz*. This list is by no means complete or even representative. Instead, the purpose is twofold. First, it gives an idea of how a possible application of the gained insight could look like. Second, the applications that are displayed here give a direct insight into the underlying corpus itself and reveals unknown facts about the collection.

4.2 Similarity networks

A full quantitative analysis of the sensitivity of stylometric similarities with respect to changes in the measures and their parametrization is beyond the scope of any contribution. There are simply too many degrees of freedom, and any distinction is muted or amplified by the specific corpus studied. We instead content ourselves with raising awareness for the non-negligible consequences that choices of features and similarity measures have on corpus analysis.

All examples in this section consist of the 86 novellas of the *Novellenschatz* and are based on the frequencies of words that have been converted to lower-caps only. No stemming or stop-word filtering was applied, but we did filter words that appear in only few novellas (20% culling). From the 500 most frequent words thus obtained for the corpus, we generate three feature vectors for each novella. One consists of relative frequencies of words (*tf*-scores) and in the other two they are weighted by their prevalence in the documents (*tf-idf*-scores), and replaced by their normalized deviation from the expectation (z-scores).

To determine the (dis)similarity of novellas, we compare their 500-dimensional feature vectors using absolute differences between their entries (δ_1-distance) as well as Euclidean distance (δ_2-distance) and cosine similarity ($\delta_{\cos}$-distance) in the feature space. This yields nine combinations of feature vectors and dissimilarities.

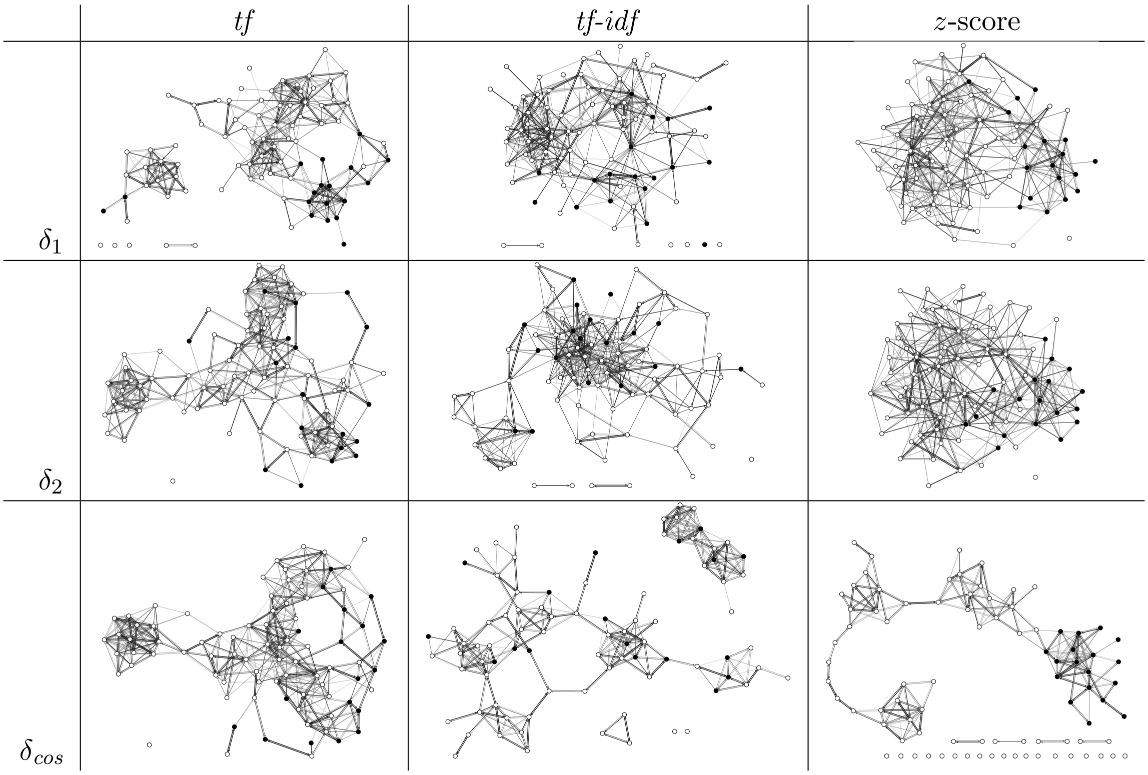
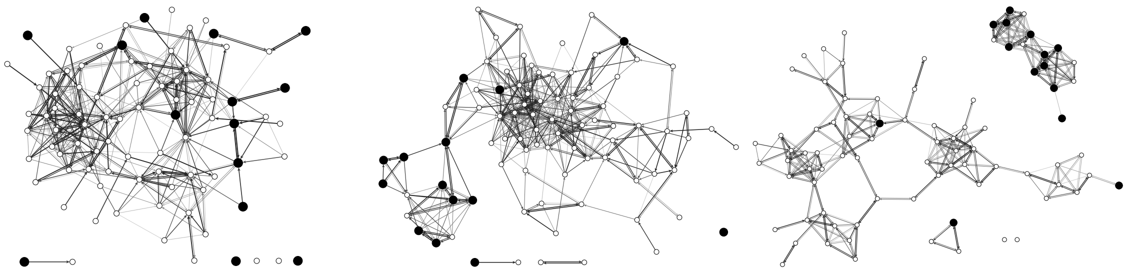


Fig. 4.1: Backbone networks based on three different distance measures and three different feature vectors from the 500 mfw. An edge points from one novella to another, if five out of the ten novellas closest to the first are also among the ten closest to the second. Edge thickness indicates the rank of the neighbor among the closest novellas. For comparison, we highlighted an apparent group from the lower right (cosine similarity of *z*-scores) in all backbones; this mystery group is discussed at the end of the section.



(a) δ_1 (b) δ_2 (c) $\delta_{\cos}$

Fig. 4.2: Backbones based on *tf-idf*-scores where highlighted nodes represent *Adelsnovellen* (novellas involving nobility).[1]They tend to cluster if larger deviations in single entries are emphasized, which δ_1 does not. According to Figure 4.6 (middle), the most indicative words for this subgenre are `Graf` (noble title) and `Fräulein` (young lady).

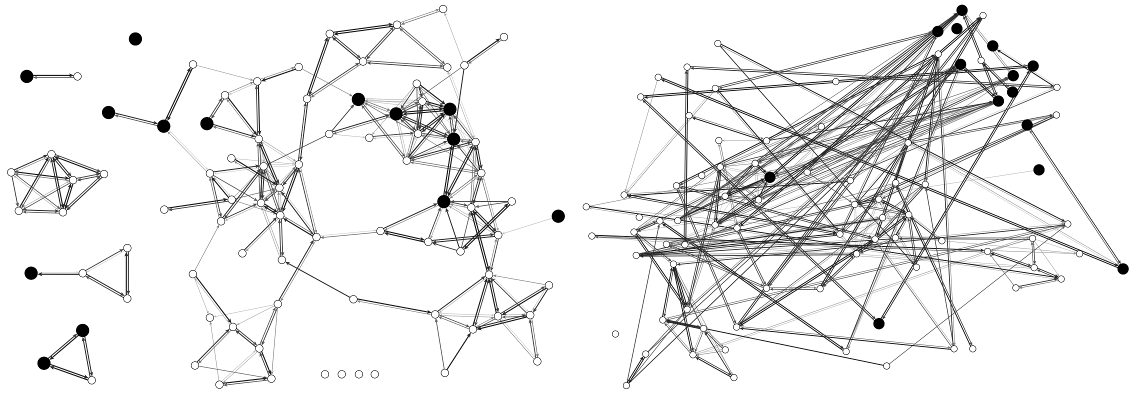
To understand grouping tendencies that most clustering methods will exhibit, a Simmelian backbone is determined for each of these nine dissimilarity matrices. For each novella, we create a ranking of the ten most similar other novellas. A link is created from one novella to another, if at least five of the ten novellas closest to it are also in the top ten of the other. We also require that the neighbor is among the top ten itself, so that the relation need not be symmetric.

The resulting backbones are shown in Figure 4.1. The layouts have been adjusted to ease recognition of similar substructures. While all combinations of feature vectors and distances suggest that there are groups of relatively more similar novellas, substantial differences seem to exist. Without knowing what links them together we have chosen one apparent group in the lower right and highlighted the corresponding novellas in all nine backbones. We will return to this group at the end of the section, but want to discuss first how known groups can be found in some backbones but not others. This serves to demonstrate that, for groups that are not defined by authorship, the choices made during the construction of similarity measures have a strong impact on the potential for uncovering groups using clustering.

In Figure 4.2, the subgroup of *Adelsnovellen* (novellas involving nobility) is clearly identifiable, if clustering is based on the presence of associated words of high discriminatory power. With 40% culling, however, the group dissolves because the indicative noble titles are no longer part of the feature vector (see Figure 3.5). Without prior knowledge of the significance of certain words for a subgenre, focusing on more widely appearing words bears the risk of losing the possibility to identify a relatively small group characterized by them.

We can observe very similar consequences with first-person narration. The pronounced cluster on the left of Figure 4.4 becomes part of a larger group if we use generic pronouns and lemmatization instead. Note that this may very well be the desired outcome, for instance, if groups of novellas are sought while controlling for the perspective in which they are narrated. Depending on analytic interest, the presence of certain pronouns among the

[1] In Appendix A2 we can find a list of meta data including the novellas involving nobility.



(a) 40% culling

(b) same with layout from Figure 4.2c

Fig. 4.3: Backbones based on cosine similarity of *tf-idf*-scores with 40% culling where highlighted nodes represent *Adelsnovellen*. The layout on the right is the same as in Figure 4.2c and thus illustrates the large differences between 20% and 40% culling where, e.g., `Graf` is no longer a feature.

most frequent words may thus be a signal or a confounding factor.

This is further emphasized in Figure 4.5, where we determined main characters as those with maximum degree in the co-occurrence network, i.e., characters who are referred to in the same paragraph with the largest number of others. Novellas with a female main character tend to cluster, and these groups are very similar to those obtained when selecting novellas in which `sie` (she) or `die` (female definite article) is the most frequent word.

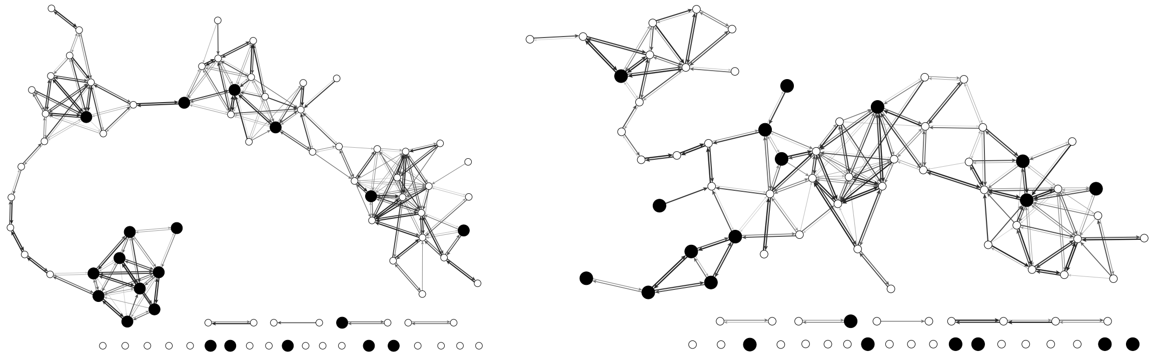
The examples in Figures 4.6 and 4.7 demonstrate that sometimes very few words explain a clustering obtained after a series of steps obfuscating their significance. Cohesive groups in Figure 4.6 can be discriminated with a short list of articles, pronouns, and a conjunction, if features are based on term frequencies.

Amplifying the frequency of otherwise unusual words with *tf-idf*-scores, on the other hand, leads to clusters determined by the cast of characters as shown in Figure 4.7. The words `Pfarrer`, `Fräulein`, `Bruder`, `Vetter`, and `Graf` are the six nouns with the highest average *tf-idf*-scores across all novellas, and also have large variance.

Of course, the presence or absence of certain groups rarely depends on one single, or even a small class, of features. A cluster that is apparent in one backbone and relatively stable across multiple analyses was already highlighted in Figure 4.1. It consists of 14 novellas written by *Arnim, Halm, Heyden, Gottfried and Johanna Kinkel, Kleist, Kruse, Kugler, Kurz, Lorm, Müller, Raabe, Riehl, Rumohr, Schücking, Sternberg, Waldmüller*, and *Wallner*.

Searching for a reason why these novellas are considered similar across a number of operationalizations, we find that 13 of the 40 verbs most underrepresented in this subgroup[2]

[2] The verbs with the least average z-score in this group are: `sind`, `ist`, **`sagte`**, `bin`, `ging`, `thun`, `muß`, `bist`, **`gesagt`**, `fuhr`, `kannst`, `kommt`, **`glaube`**, `will`, **`rief`**, `geht`, **`denken`**, `hast`, `kommen`, **`hörte`**, `sehe`, `thut`, **`gehört`**, **`sagen`**, `kann`, **`weißt`**, **`dachte`**, **`weiß`**, **`wissen`**, `lachte`, `willst`, `sah`, `gehen`, `hat`, `helfen`, `gethan`,



(a) pronouns as separate features

(b) same token for all pronouns

Fig. 4.4: Backbone networks based on cosine similarity of z-scores. One cluster consists entirely of first-person narratives (dark nodes) and is largely due to an overrepresentation of the word `ich` (first-person pronoun). Combining pronouns into one generic token has a substantial effect on the cluster structure.

are associated with direct speech. Storm (1881) later referred to the novella as "Schwester des Dramas" (drama's sister), which prompts us to expect a prevalence of direct speech. A straightforward test for the presence of direct speech is challenging because of the variety of markers used; instead of by quotation marks, direct speech is often indicated by starting a new line, hyphens, or no syntactical element at all.

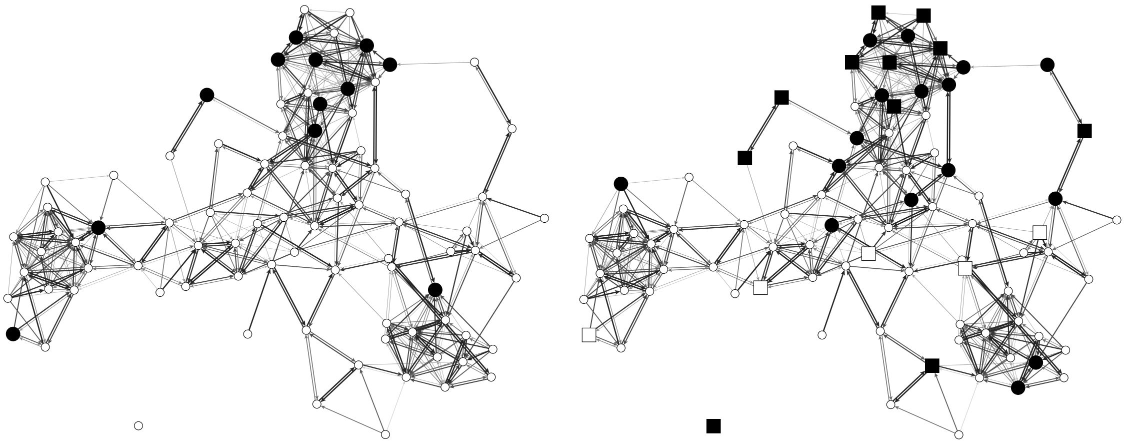
A second factor that appears to contribute to the discrimination of this group is the more frequent use of past tense.While the classification of tense is notoriously difficult for German. We used tags from the German-language model of `spacy` and used a simple algorithm (Alg. 1) to decide for past tense. As shown in Figure 4.8, all but two are above the median.

4.3 Methodological consequences

We have broken down the process of grouping texts by stylometric features, and more specifically the similarity of word use. Using the corpus of *Deutscher Novellenschatz* as a case study, we have demonstrated that the choices made in the grouping process have substantial impact on the groups found, and that groups are sometimes determined by seemingly trivial factors which may or may not be desirable for the research question at hand.

As a consequence, discussions of the suitability of similarity measures need to take into account the context in which the measures are applied, how the data is prepared, and which kind of signal is discriminative. Author signals, genre characteristics, narrative perspective, plot elements, and many other aspects confound the definition of stylometric similarity. The

`reden`, `essen`, `saßen`, `wird`.



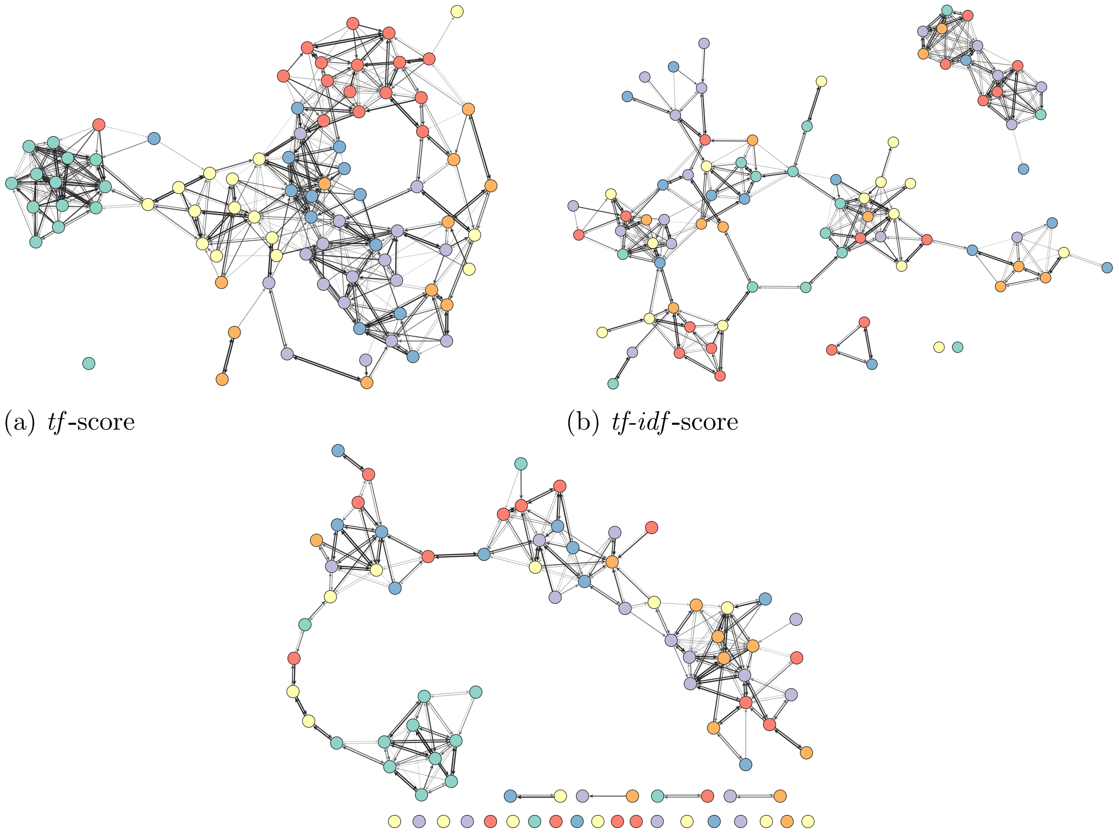
(a) female author

(b) highest *tf*-score for `sie` or `die`

Fig. 4.5: Backbone network based on δ_2 of *tf*-scores. Female authors tend to cluster (dark nodes on the left) as do novellas (dark nodes on the right) in which the most frequent word is either `sie` (she) or `die` (female definite article). These are not the same as the novellas with a female main character (squares).

Algorithm 1 get past tense

```
Input word.tag and word.text from spacy and tense
if word.text $\in$ {war, warst, waren, wart} then
    tense $\leftarrow$ past
else if word.tag $\in$ {VAFIN, VMFIN, VVFIN} then
    if last two letters = te or = ten then
        tense $\leftarrow$ past
    else
        tense $\leftarrow$ present
    end if
else if word.tag $\in$ {VAPP, VMPP, VVPP} then
    tense $\leftarrow$ past
else if word.tag $\in$ {VAIMP, VAINF, VMINF, VVIMP, VVINF, VVINF, VVIZU} then
    tense $\leftarrow$ present
end if
```

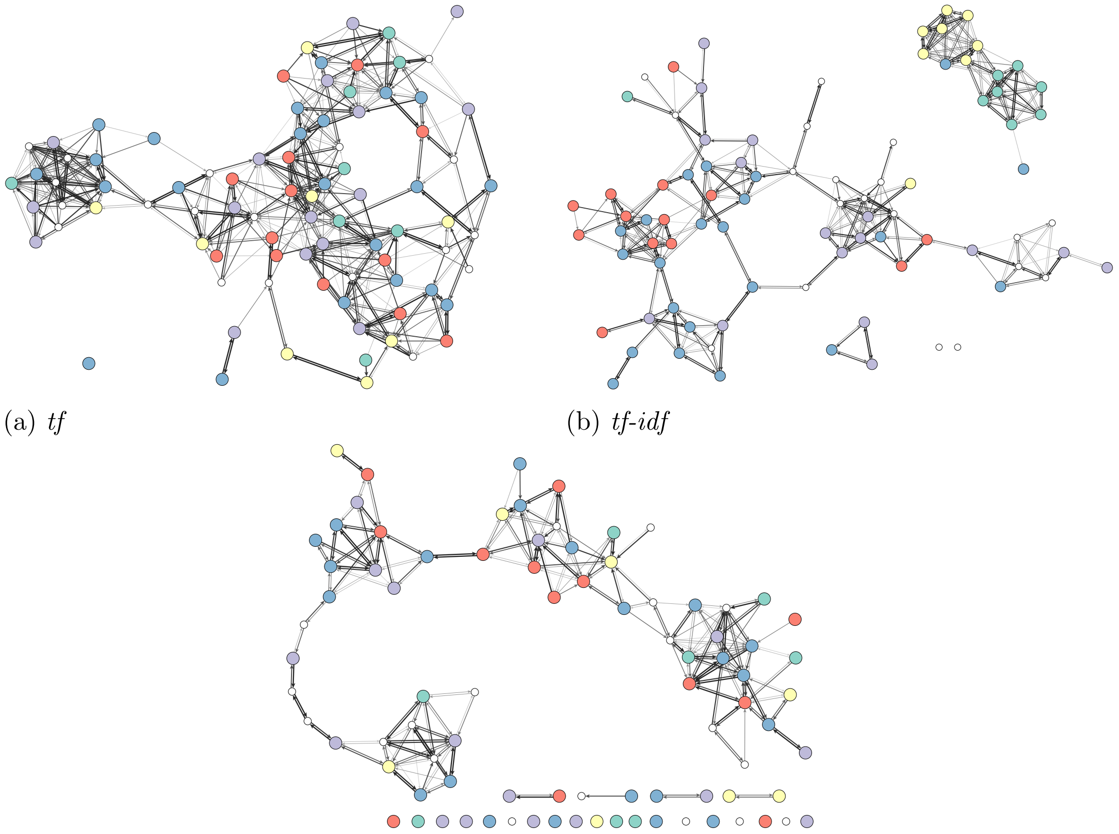


(a) *tf*-score

(b) *tf*-*idf*-score

(c) *z*-score

Fig. 4.6: Backbones for cosine similarity of the 500 mfw. Each novella is colored according to the word with highest *tf*-score among `er` ●, `der` ●, `sie` ●, `die` ●, `ich` ●, and `und` ●. For the *tf*-features, but not the others, most of the clustering is already explained by this one word.



(a) *tf*

(b) *tf-idf*

(c) z-score

Fig. 4.7: Backbones for cosine similarity with novellas colored according to the word with highest *tf-idf*-score among `Pfarrer` ●, `Fräulein` ○, `Bruder` ●, `Vetter` ●, and `Graf` ●. The unmarked novellas contain neither of these words. For the *tf-idf*-features, but not the others, most of the clustering is already explained by this one word.

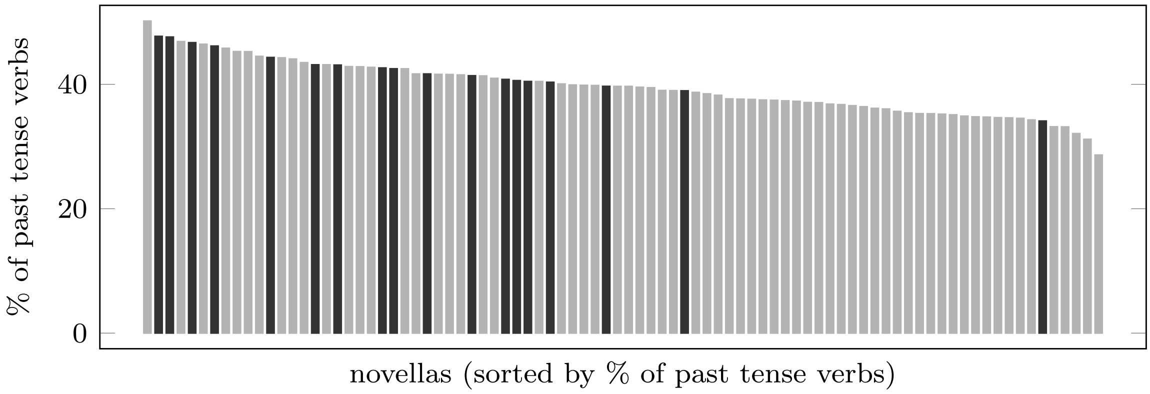


Fig. 4.8: With the exception of the novella by *Heyden*, those in the mystery group from Figure 4.1 (marked black) have an comparatively high use of past tense verbs.

task of amplifying desired aspects and controlling for others is thus both a theoretical and an empirical one, and it should be taken lightly.

If a resulting grouping matches the expectation, but the employed stylometric similarity does not operationalize the hypothesized mechanism, this still does not confirm a grouping hypothesis. Confidence in the meaningfulness of a clustering is gained only through justified decisions, not from decisiveness of results or their robustness with respect to different parameters.

For a given corpus and an idea about distinctive qualities of the texts in it, the task is to exploit the individual steps in the clustering process to let it be governed by these qualities. If individual authors can be distinguished by their use of function words, but male-female differences are suspected to manifest themselves in the use of certain adjectives, different stylometric similarity measure should be used for these tasks. We have pointed out several options for adaptation, from tokenization to baseline distributions, but the list of course goes on.

Our discussion was focused on the construction of similarities, and did not include the influence of specific clustering methods. There is a qualitative difference between these two aspects, and there is no point in accurate clustering if the similarity measure is inappropriate. This motivated the use of backbone networks because they point to relatively cohesive groups any clustering method will tend to preserve.

Through the presentation of several examples on a medium-size literary corpus, the *Deutscher Novellenschatz*, we attempted to make the practical relevance of these general methodical considerations more tangible. While plagiarism and authorship attribution may be major use cases for stylometric similarity analyses, these examples show that there is great potential for the identification of other clusters of texts, not related to authorship. Each kind of clustering will require its own combination of methods, because what works for authorship attribution may not apply to gender differentiation, and what works for genre classification may not apply to narration styles. Rather than contributing a solution, it seems, we are thus creating more problems.

Chapter 5

Novella Character Networks

5.1 Introduction

The foundations of network science lie in the social sciences[1] and as such some of the most natural objects of study are interpersonal connections like friendship, work, or club membership ties. A large subfield of network science is therefore dedicated to social networks and social network analysis (SNA). Early research includes the study of friendship networks in school children (Moreno, 1934). Aside from the developments in SNA, narrative theory uses character constellations to investigate plays and prose (Wilpert, 1959). Hence, the natural question arises whether we can operationalize the generation of character constellation networks to observe and investigate ties for characters in literary fiction via SNA. However, the methods used need to be customized to the specific nature of persons in texts.

First, the standard data collection methods like surveys (Freeman et al., 1987), organigrams (Luhmann, 1964), or observation by the researcher (Malinowski, 1922) are not suitable for characters in texts. Instead, we need to employ techniques to locate persons and their interpersonal connections in prose texts. Second, it is very frequent in literature that characters are relatives and share the same last name, are called by their nicknames, are given craft names by the author, or are even inanimate objects (Elsner, 2012). Third, a key characteristic of (good) literature is the use of synonyms, pronoun references and non-mentioning of characters that are present in a scene. Therefore, a careful and well-founded decision must be taken regarding the level of automation and to which granularity personal references should be extracted when investigating prose texts. Usually, a higher degree of automation is prone to a larger amount of misallocation while a lower degree of automation increases the workload and therefore limits the quantity of analyzed text. In addition, manual annotation could be prone to subjective perception by the researcher. Finally, many works of literary fiction are deliberately composed and therefore *closed worlds*. As stated by Labatut and Bost (2020), “real-world networks are the result of microscopic processes, whereas fictional ones are caused by a macroscopic process (Rochat and Kaplan, 2014)”.

As described in Chapter 1.2 a main argument for choosing the *Deutsche Novellenschatz*

[1] Especially social psychology, sociology, and anthropology as Brandes et al. (2013) point out.

as our subject matter is the possibility of taking an intermediate approach that is in line with the theory of "scalable reading" (Müller, 2013; Weitin, 2017). By using a semi-automated approach, we thus will accommodate for the corpus size to best possible combine the advantages of closed reading (interpretability of results and precision with respect to correct synonym allocation) with those of distant reading (speed and precision with respect to otherwise missed character allocations). Moreover, this will reduce the risk of overfitting the data.

5.2 Theory and existing methods

Character network extraction has some history in the digital humanities and a number of works deal with character networks in literature. Labatut and Bost (2020) identify three motives for researchers to look at those networks. First, a *narrative analysis* can give a simplification of the plot to get a "distant reading" perspective (Moretti, 2011). Second, the *network science* approach looks at the network itself as the object of study to apply network analysis tools on them (Prado et al., 2016; Chakraborty et al., 2021) or compare them with other kinds of networks (e.g., real-world social networks as in Alberich et al., 2002). A third line of research comes from *machine learning*. Here, the focus is on the automation of the network extraction process. Moreover, generated networks (usually large network ensembles) are used to perform tasks based on network comparison like genre classification (Coll Ardanuy and Sporleder, 2014; Gil et al., 2011).

The typical network extraction procedure is illustrated in Figure 5.1 and consists of three steps that involve a number of decisions producing various (and possibly extremely different) types of networks. Typically, the first move when creating a character network is the character identification. Here, the multiple decision points include level of automation as well as the types of character representations to consider.

One strategy is using a pretrained named entity recognizer (*NER*) to find all mentions of names in a text. This is for example done in Hettinger et al. (2015). This approach usually has two disadvantages, the inaccuracy of named entity recognition systems in the first place (especially when trained on a 21st century newspaper data set and applied to a 19th century data set of literary prose), and the above already discussed phenomenon that names are often only mentioned once per scene. More elaborate ways include the enhancing of NER by text patterns that particularly identify noble titles[2] (Trovati and Brady, 2014), the removal of infrequent names (Elsner, 2012; Sack, 2012), the search for characters as subjects of action verbs (Coll Ardanuy and Sporleder, 2014), and the search for characters by relations of possession (Karsdorp et al., 2012). Other ideas of alias resolution are proposed by Elson et al. (2010), Elsner (2012), Vala et al. (2015), and Jannidis et al. (2016).

As an alternative, many researchers annotate (usually a single) work by hand such that (depending on the granularity) proper nouns, pronouns, synonyms and other nominals, as well as speakers of direct speech are correctly allocated to the characters in the text (see

[2] "King",...

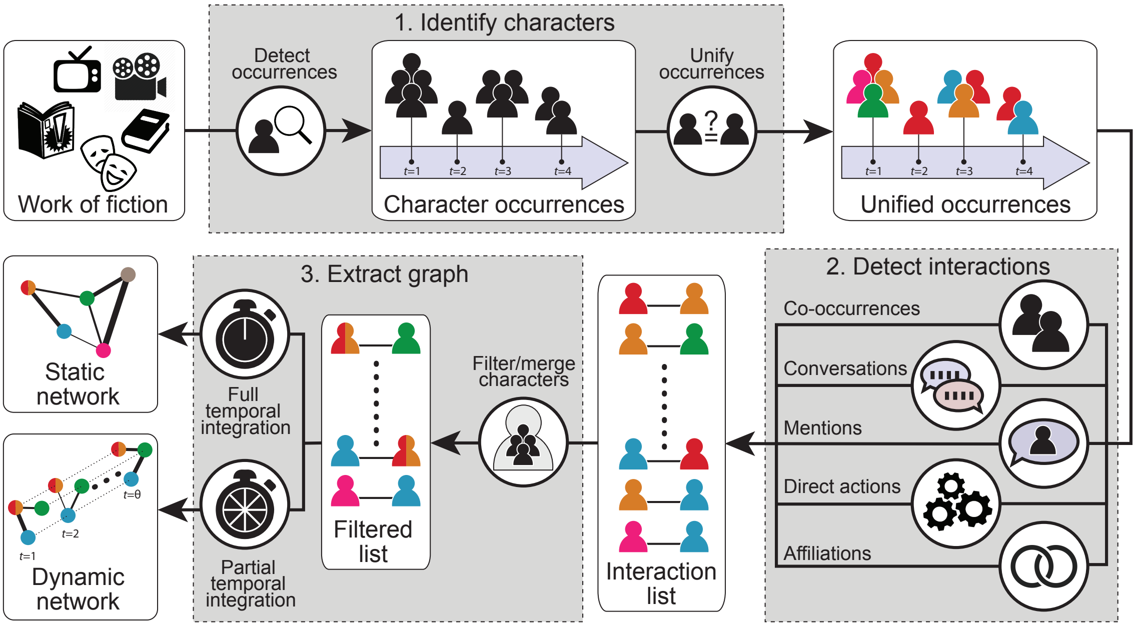


Fig. 5.1: Pipeline to describe the generic character network extraction process. The figure is taken from Labatut and Bost (2020) and available at `10.6084/m9.figshare.7993040` under CC-BY license.

e.g., Kydros and Anastasiadis, 2015), but even though this seems to be the most accurate way this approach is – aside from the required amount of time – problematic in several ways. It can make the investigator overconfident in the results while there was a certain amount of disambiguation intended by the author. Moreover, it does not solve problems like mentions of non-present characters in direct speech among others. Therefore, one should be aware that the selection of characters and their connections will always be a choice that should take into consideration the purpose of the data collection.

Other approaches (that could also be seen as semi-automated) include character indexes (Alberich et al., 2002; Sack, 2012; Rochat and Kaplan, 2014; Krug et al., 2017) and crowdsourcing (Rochat and Triclot, 2017). In particular, crowdsourcing can be a good workaround if one is looking for similar accuracy to manual annotation without having to restrain the investigation to just a few works.

After fixing the set of characters and a possible unification (alias resolution) the second major step is the choice of interaction manner. Those can include conversations between characters (Zhang et al., 2003b; He et al., 2013; Nalisnick and Baird, 2013; Waumans et al., 2015), direct actions such as “thinking about” or “fighting” (Franzosi, 2008; Bossaert and Meidert, 2013; Srivastava et al., 2016), and affiliations (Krug et al., 2017). However, the by far most common way of interaction detection is simple co-occurrence. The popularity of this strategy is mainly due to the simplicity of the method. Classical points of criticism are the undirectedness of inferred links or the ignorance of sentiment (Kwon and Shim, 2017) as well as the susceptibility of the method for producing false positives since textual co-occurrence does not need to imply actual interaction of the characters (Prado et al., 2016; Min and Park, 2016). This is mostly due to the fact that co-occurrence networks usually include the above-mentioned network types as subgraphs. However, one can counter this criticism by acknowledging that other network types usually have the opposite problem of producing false negatives (missing actual linkage of the considered interaction type). In fact, Coll Adanay and Sporleder (2015) argue that those false positives are rare.

Co-occurrence is implemented in several ways which are typically adapted dependent on the underlying work of fiction. Those include sentence-based (Lee and Yeung, 2012; Park et al., 2013), page-based (Rochat, 2014), paragraph-based (Elson et al., 2010; Elsner, 2012; Jannidis, 2017), scene-based (Rieck and Leitte, 2016; Gil et al., 2011, for plays), or chapter-based co-occurrences (Min and Park, 2016; Fischer et al., 2017) as well as sliding window approaches (Hutchinson et al., 2012; Grener et al., 2017).

The resulting link list (which could be interpreted as event data with an event being represented by time, sender and receiver) can then be used to extract the final network. These networks can then either be static (Moretti, 2011; Kydros and Anastasiadis, 2015) or dynamic (Prado et al., 2016; Fischer et al., 2017). Prado et al. (2016) points out that the amount of temporal integration should be carefully adjusted to the underlying work. Typical dynamic networks are actually a list of static networks on the chapter level.

5.3 Our approach

We want to extract an ensemble of character networks based on our corpus of novellas that can be used in a variety of studies. Therefore, we made some preliminary analyses to choose the appropriate character list. We will highlight the considerations that led to the final novella character networks on Droste-Hülshoff's *Die Judenbuche*. Effectively, the approach will be the same for all 86 novellas of the *Novellenschatz*.

First, we want to look at the effect of including the synonyms of characters. In Table 5.1 we can observe that the ordering and the distribution itself changes a lot when including synonyms. This is due to the fact that the names-only approach does not work well on all characters that are usually called by their family attributes like sister (*Schwester*), mother (*Mutter*), or uncle (*Oheim*). Moreover, unnamed characters like *Herr von S.* are not detected by the names-only approach. While this may be fixed in some special cases an approach including synonyms is much more flexible in handling these instances.

character	**only names**	**with synonyms**
Friedrich Mergel	33.75	15.86
Margareth Semler	5.04	12.79
Brandes	6.05	8.21
Herr von S.	0.00	7.28
Simon Semler	11.84	6.55
Johannes	12.85	6.50
Franz Semler	3.78	4.88
Aaron	5.03	4.62
Wilm Hülsmeyer	8.06	2.50
Franz Ebel	3.78	0.99
Frau Hülsmeyer	0.50	0.83

Table 5.1: Comparison between the relative number of appearances (in %) of the most frequent characters in Droste-Hülshoff's *Die Judenbuche* with and without synonym resolution.

If we consider synonyms there are different ways of how to handle name disambiguation[3]. In the first approach that included the synonyms, the correct alias resolution was ignored and each synonym that was used for several characters was counted to each of those characters. This might lead to an undesired overrepresentation of uncommon characters. In addition, dead characters could suddenly reappear. However, the removal of such words can lead to undesired behavior itself as observed for the character *Franz Semler* who is often called *Franz* or *Ohm*, both words that are used for different characters. In Table 5.2 we compare the effect of removing all synonyms that cannot be unambiguously assignment to the assignment of such words to all characters in question. One can observe that the way

[3] Generic synonyms that are used for multiple characters throughout the novella.

character	**with double mentions**	**without name disambiguation**
Friedrich Mergel	15.86	19.86
Margareth Semler	12.79	9.56
Brandes	8.21	10.56
Herr von S.	7.28	9.37
Simon Semler	6.55	5.96
Johannes	6.50	8.50
Franz Semler	4.88	0.32
Aaron	4.62	5.96
Franz Ebel	0.99	0.00
Wilm Hülsmeyer	2.50	3.18
Frau Hülsmeyer	0.83	0.24

Table 5.2: Table as above (in %) with and without name disambiguation (*Ohm*, *Frau*, *Hausfrau*, *Braut*, *Franz*, *Kerl*, *Wittwer*, *Bruder*, *Bräutigam*).

of coping with name disambiguation drastically alters the result.

Therefore, it is desirable to find a way of assigning double mentions to the right entity. While there are different approaches in the literature to handle name disambiguation (e.g., Mann and Yarowsky, 2003; Han et al., 2004; Huang et al., 2006; Fan et al., 2011; Khabsa et al., 2015) we decided to implement a simple algorithm to decide to which person the synonym in question should be assigned. First, we always store last names together with the article such that the number of name disambiguation is reduced beforehand. In our example of Droste-Hülshoff's *Die Judenbuche*, this was applied to *die Hülsmeyer* and *der Hülsmeyer*, who are correctly identified as *Frau Hülsmeyer* and *Wilm Hülsmeyer.* Furthermore, we make sure that only the longest version of a name is recognized as such. Thereby, *Franz* cannot be recognized as *Franz Semler* if the text is directly mentioning *Franz Ebel.* Finally, we assigned a synonym in question to the person that has the closer word distance[4] in the novella. In this way we could correctly classify 80% of the synonyms with name disambiguation for a test corpus of 5 novellas.[5]

While the usage of double mentions comes with some advantages and disadvantages, we want to argue that, depending on the examined research question, synonyms do add valuable information to character networks. Therefore, we choose to add them to our data set. Moreover, we assigned each synonym a type such that the data set can later be filtered

[4] We compute the word distance as the number of words between the synonym in question and all other synonyms used for the possible entities the synonym could belong to.

[5] 136 out of 171. Apart from Droste-Hülshoff's *Die Judenbuche* the test corpus contained the novellas of Eichendorff (*Die Glücksritter*), Gerstäcker (*Germelshausen*), Goldammer (*Eine Hochzeitsnacht*), and Hartmann (*Das Schloss im Gebirge*). Typical misclassifications include words like *Bruder* in direct speech where it is usually a salutation for the respondent (also called a *mention* in Figure 5.1) and not the name of the speaker.

by those types (e.g., going back to just considering names).

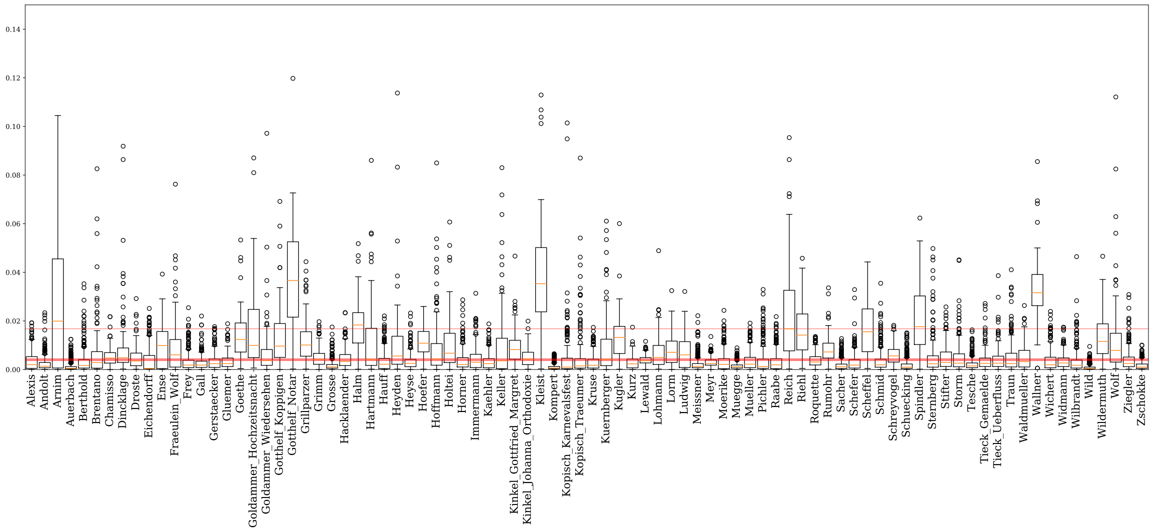


Fig. 5.2: Boxplots for the distribution of paragraph lengths relative to text length in all novellas. The thicker red line indicates the mean relative paragraph length and 95% of the paragraphs have a relative length below the thinner red line.

Contrary to the pipeline described in Section 5.2 we are aiming for a data set that can be used in many different circumstances. That is why we want to argue that the pipeline misses a step between 2 and 3. Namely, if we detect a mentioning of a character it happens on some text unit (sentence, paragraph, page etc.). Therefore, we create a two-mode network with characters as our source and text unit as our target value.

Because we are aiming to mimic a spacial connection (i.e., being in the same scene, room, conversation, etc.) by measuring co-occurrences, we decided to choose the paragraphs as our text unit. While the condensed format of a novella does not allow us to use chapter-based connections[6], the use of sentence-based links seems to be too restricting. Moreover, paragraphs can be seen as the smallest contexts in which sentences are embedded. Zadrozny and Jensen (1991, 1993) call them "units of thought". In Figure 5.2 we observe that the paragraph length is usually quite stable across novellas. Exceptions are Arnim's *Der tolle Invalide auf dem Fort Ratonneau*, Gotthelf's *Der Notar in der Falle*, Halm's *Die Marzipan Lise*, Kleist's *Die Verlobung in St. Domingo*, Spindler's *Die Engel Ehe*, and Wallner's *Der arme Josy* with a median above the thinner red line. The results do not change if we consider the absolute paragraph length or some damped weighting (e.g., by logarithmic text length) with a sole exception. The second novella of Gotthelf *Kurt von Koppigen* has very long paragraphs but is also quite long.

In this way we can store the synonyms and types as edge attributes. Moreover, the interaction list and extracted graph can be reinterpreted as the one mode representation

[6] In fact, only 38 out of 86 novellas are split into more than one chapter.

of such a two-mode network. However, this step will lose some information and should therefore only be performed when necessary. In the following sections we will explain the exact procedure to create the bipartite networks that is schematically described in Figure 5.3.

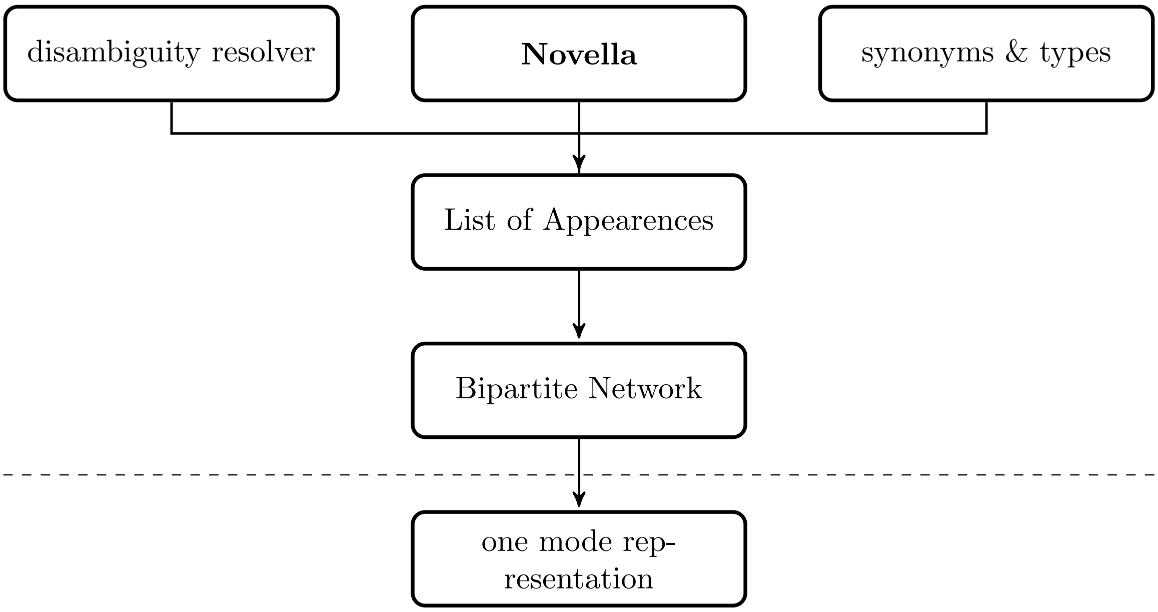


Fig. 5.3: Road map for the novella network extraction.

5.3.1 Annotating the data

The first step is the annotation of the novellas. Due to the size of the corpus and the accuracy we aspire to reach we decided for a semi-automated procedure that includes a crowdsourcing component similar to Rochat and Triclot (2017). Therefore, a *Citizen Science Project*[7] was created to set the collective reading task. The project is called "Koreferenzen in Novellen"[8] (*co-references in novellas*) and can be found under `https://lab.citizenscience.ch/de/project/478`. The project participants were supposed to answer the following questions (see Figures A.1 and A.2 in the appendix for a view of the website):

- Read the novella. Name up to 9 characters as well as their synonyms and aliases.[9]

[7] Citizen Science is "a web-based tool that allows researchers, students, and all members of the public to create and run Citizen Science Projects [where] volunteer contributors ('citizens') and scientists work together to produce scientific knowledge" (see `https://lab.citizenscience.ch/en/about`).

[8] The German project description should make sure that some proficiency of German language is given by the project participants.

[9] The precise wording in German was "Lies die Novelle. Nenne bis zu 9 Figuren und alle ihre Synonyme und Aliasse"

- Are there any more characters in the novella?[10]

Whenever the answer to the second questions was *yes* they were presented with a new text field and the task to enter again up to 9 different characters. The limitation of characters in a single text field were given for readability and verifiability purposes. The answers were then stored in a table and enhanced by the synonym type (See Table 5.3 and 5.4).

type	**description**	**examples**
name	proper names	Maria, Johannes, Aaron, Samuel, Christian Jacob Heyliger
family	family and marriage relations	Vater, Mutter, Ohm, Wittwer, Braut
title	(noble) titles	Herr, Frau, Baron, Euer Gnaden, König
job	job titles and functions	Kutscher, Gerichtsschreiber, Bauer, Pfarrer, Hausfrau
pseudonym	alternative names given by the author	Hausbesitzer, Kräuter-Ev, Hexe Wanderer, Langschläferin
salutation	general names of characters e.g., in direct speech	Bursche, Knabe, Kerl, Freund, Alte

Table 5.3: We allocated the synonyms to six categories. While the categories *name*, *family*, *title*, and *job* are self-explanatory we also decided to separate the categories *pseudonym* and *salutation*. While pseudonyms relate to specific character descriptions, salutations are of a more general form. Therefore, we used 10% culling, i.e., all synonyms that appear in at least 10% of the 86 novellas of the corpus and do not fall into the other four categories above are considered salutations.

5.3.2 Bipartite network extraction

Given the text of a novella and the list of synonyms for all characters present in the novellas, we extract a link list for each novella containing the referenced figure as the source value, the paragraph of appearance as the target value, as well as the used synonym and synonym type as link attributes (see Table 5.5).

The visualization was then done in the following way.

Nodes. First, we plotted the paragraph nodes in their order of appearance in one horizontal line in the center of the graph. Second, we plotted the character nodes either above or below this line. The position of the characters was determined by its average positional appearance in the text. I.e., a character appearing in the left of the graph is usually more

[10] "Gibt es noch weitere Figuren in der Novelle?"

name	synonyms	type
Friedrich Mergel	Friedrich, Friedrich Mergel, Sohn, Fritzchen, Junge, ...	name, name, family, pseudonym, salutation, ...
Hermann Mergel	Hermann, Hermann Mergel, Grundeigenthümer, Wittwer, ...	name, name, job, family, ...
Ulysses	Ulysses	name
Margareth Semler	Margareth, Margareth Semler, Braut, Frau, Mutter, Schwester, Hexe	name, name, family, title, family, family, pseudonym
Maria	Maria	name
Franz Semler	Franz, Franz Semler, Ohm, Bruder, Oheim, Wittwer	name, name, family, family, family, family
Frau Hülsmeyer	Nachbarin, Hausfrau, Frau Hülsmeyer, die Hülsmeyer	salutation, job, name, name
Aaron	Aaron, Jude	name, pseudonym
⋮	⋮	⋮

Table 5.4: Part of the synonym list for Droste-Hülshoff's *Die Judenbuche.*

name	paragraph	synonym	type
Friedrich Mergel	3	Sohn	family
Herrmann Mergel	3	Grundeigenthümer	job
Ulysses	3	Ulysses	name
Friedrich Mergel	6	Knaben	salutation
Margareth Semler	6	Braut	family
Brandes	6	Förster	job
Braut	6	Braut	family
Bräutigam	6	Mann	salutation
Friedrich Mergel	7	Friedrich	name
Herrmann Mergel	7	Vater	family
Margareth Semler	7	Braut	family
Braut	7	Braut	family
Friedrich Mergel	8	Friedrich	name
Herrmann Mergel	8	Wittwer	family
Margareth Semler	8	Margareth	name
⋮	⋮	⋮	⋮

Table 5.5: Head of the link list (first 15 entries) of Droste's *Die Judenbuche.*

present in the beginning then at the end and vice versa. This might lead to a set of overlapping nodes in the middle for those characters that appear throughout the story. Therefore, minor manual perturbations have been performed to make underlying nodes visible. Finally, the node area size is dependent on its out degree, and we decided to only display the main characters based on that value (i.e., the number of paragraphs a character is present).

Links. Each link was given a color that is based on the degree of the source node. In this way we can both, easily connect the links to the characters as well as highlighting the "importance" of the corresponding character (the darker the link, the higher the characters frequency). The link line style was depended on the link type. Here, a solid line represents a name, a dashed line a family connection, a rounded dashed line a job description, a square dotted line a title, and a dotted line pseudonym or salutation.

Three examples of those networks are seen in Figures 5.4 – 5.6. The novella of Droste is a good example for the representation of the different types. While *Friedrich Mergel* and *Johannes* are usually called by their names, the character *Margareth Semler* is mostly characterized by their family relations (*Mutter, Braut, Schwester*). Moreover, the character *Herr von S.* is called by his title (*Gutsherr, Baron, Gnaden*) while the character *Brandes* is called by his profession (*Oberförster, Förster*).

Figure 5.7 lists all network visualizations while enlarged versions for each novella can be found in Appendix A3. In general, the visualizations can lead to first observations about the investigated networks that are being thoroughly studied in Chapters 6 and 7.

5.4 Data management

As stated earlier, the digitized texts of the *Novellenschatz* are published by Weitin and publicly available via `https://www.deutschestextarchiv.de/search/metadata?corpus=novellenschatz`. In a similar way, we use the Open Science Framework (OSF)[11], a data storage to publish our networks and the synonym lists resulting from the citizen science task that precede them. The data is stored as graphml and txt files and can be found at `https://osf.io/m3yz2/`. The 86 synonym lists and corresponding bipartite networks are accompanied by an explanatory pdf file.

[11] `osf.io`

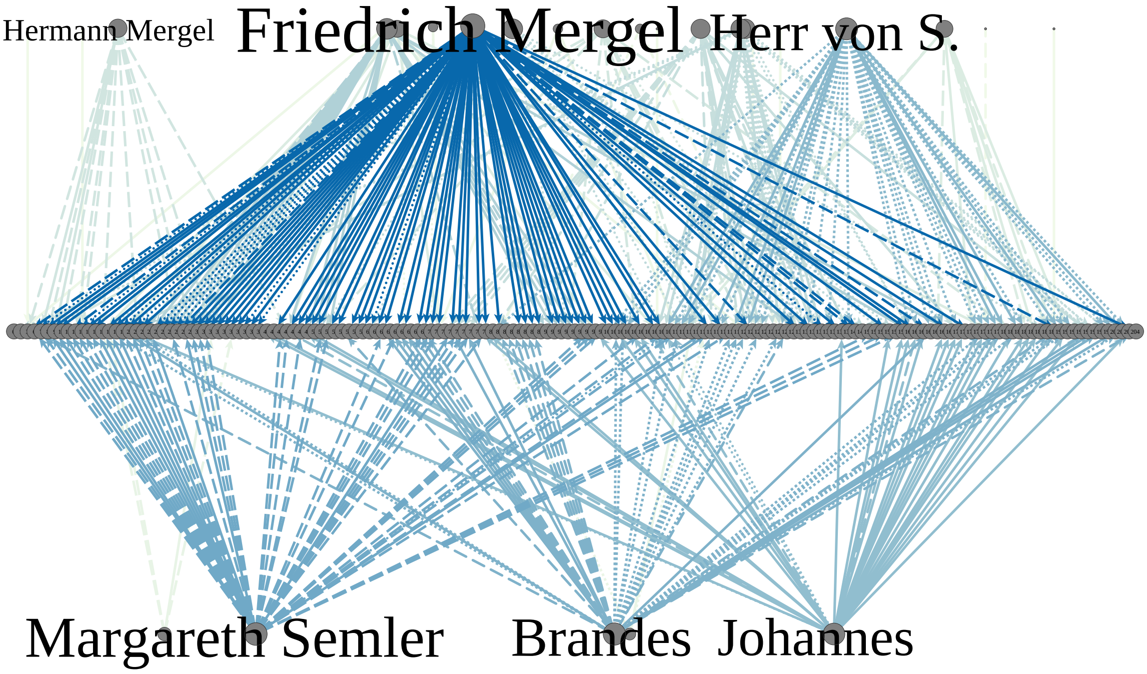


Fig. 5.4: Bipartite network of Droste-Hülshoff's *Die Judenbuche*. One can clearly identify the main characters of the novella. Moreover, one can observe that while Friedrich Mergel is present throughout most parts of the text other characters exclusively appear in specific sections of the novella. e.g., readers of the novella directly identify the parts where *Simon Semler* and *Friedrich Mergel* co-occur or recognize that *Herr von S.*, *Johannes*, and *Förster Brandes* do appear mostly in the second half of the text while *Herrmann Mergel* is only present in the beginning (he dies).

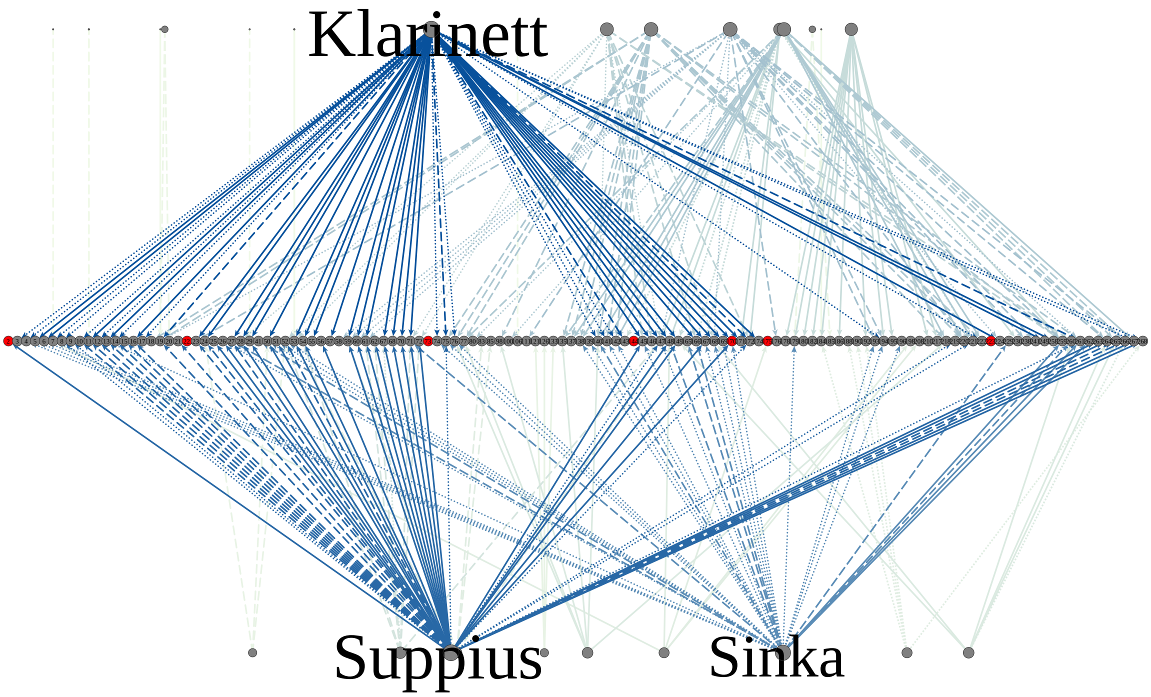


Fig. 5.5: Bipartite network of Eichendorff's *Die Glücksritter*. The official chapters of the novella can already be induced from the link pattern. To make it visible the new start of a chapter has been marked with a red node. We see that the two main characters *Klarinett* and *Suppius* are not present in the novellas chapters three and five.

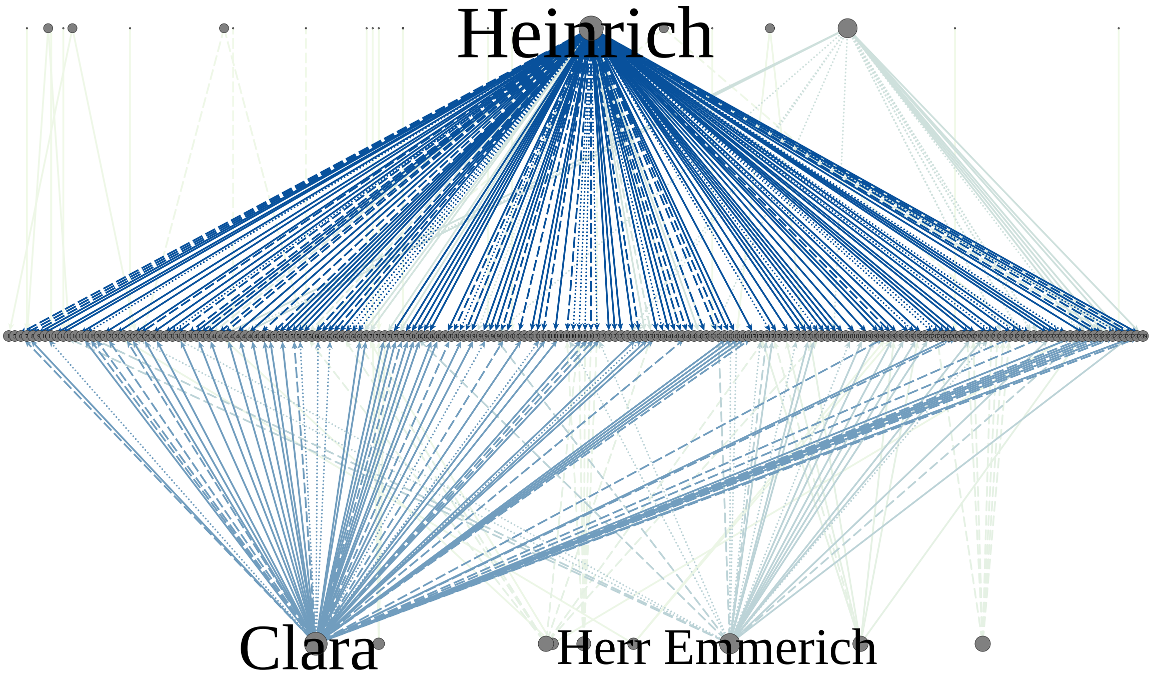


Fig. 5.6: Bipartite network of Tieck's *Des Lebens überfluss.* It can be easily recognized that a couple is in the center of the narrative and all other figures only accompany their presence.

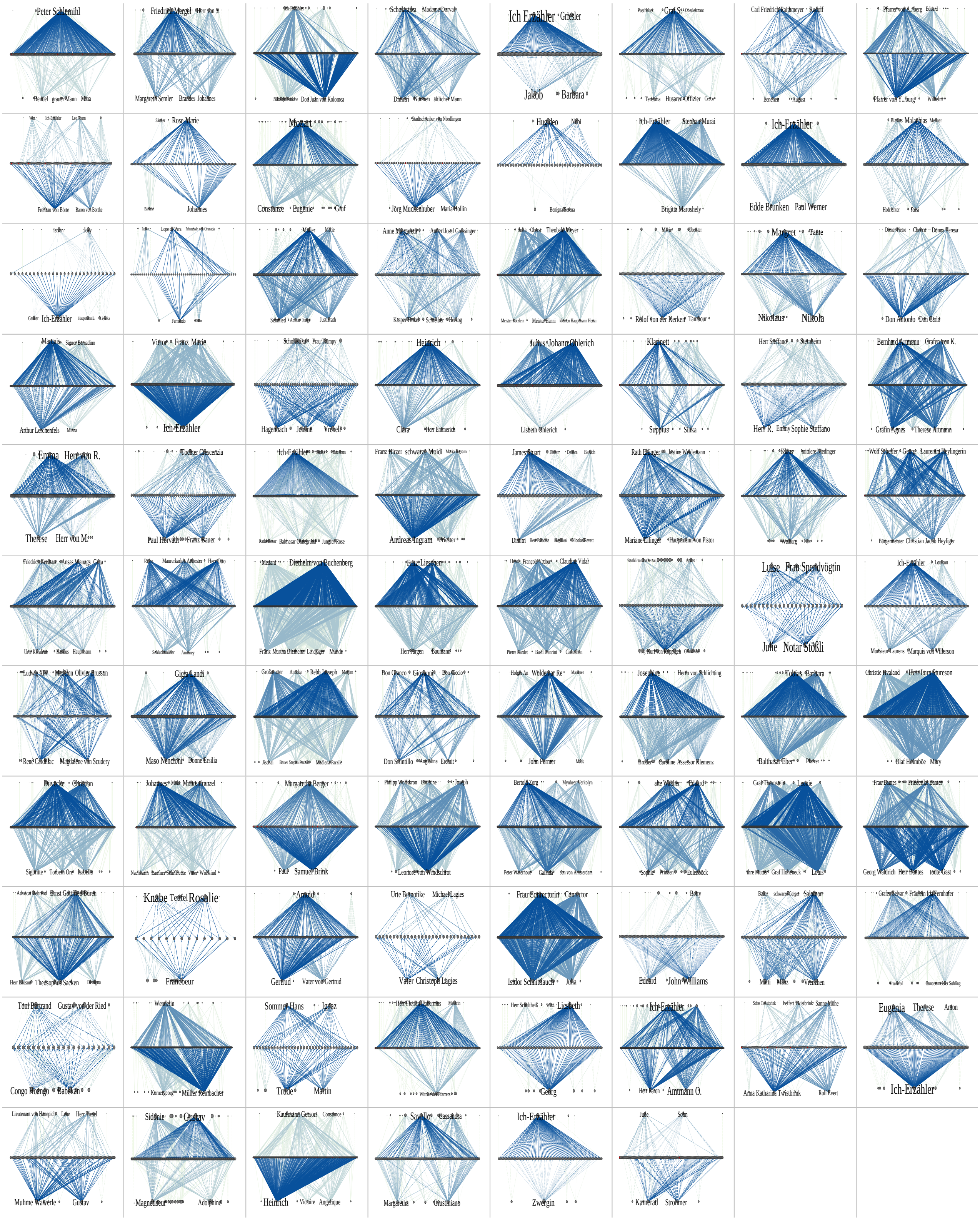


Fig. 5.7: All bipartite networks extracted from the Novellenschatz.

Chapter 6

Character Network Ensembles

6.1 Introduction

Many of the studies described in Chapter 5.2 deal with a singular work or look at individual outcomes for the networks of a larger corpus. In the present work, we want to contribute to this research by analyzing and comparing a corpus as an entire network ensemble. Hence, we use the networks themselves as our objects of study to understand their relationship within the corpus.

As described in Chapter 1.2 the *Deutsche Novellenschatz* was composed with the editors' intention to be a paradigmatic sample of the novella style and a strict formal criterion was given to distinguish novellas from novels. Therefore, the editors established the phrase strong silhouette ("starke Silhouette") and claimed it to be their guiding principle in the selection of texts. As such, a text does not demand to have a certain text length to be rated as a novella but instead needs to stay focused on a single topic that then can be executed thoroughly.[1] Hence, we hypothesize that the novellas in the corpus have a similar character constellation network which further motivates our research question whether the restricted form of a novella gives preference to a specific character constellation. To test our claim, we use a network distance measure and display the result using a dimensionality reduction technique. Therefore, we will consider the one-mode representations of the constructed two-mode networks.

6.2 Character networks

While the two-mode networks have been introduced in Chapter 5 we here will explicitly describe the building process of the one-mode representations for these networks. They are

[1] In the introduction they differentiate a novella from a novel as following: "Wenn der Roman ein Cultur- und Gesellschaftsbild im Großen [...] entfaltet, bei dem [...] ein concentrisches Sichumschlingen verschiedener Lebenskreise recht eigentlich abgesehen ist, so hat die Novelle in einem einzigen Kreise einen einzelnen Conflict [...] darzustellen und die Beziehungen der darin handelden Menschen zu dem großen Ganzen des Weltlebens nur in andeutender Abbriviatur durchschimmern zu lassen. [...]" (Heyse and Kurz, 1871)

constructed by connecting character nodes with a link whenever they point to the same paragraph and are hence the classical co-occurrence networks. Moreover, we combine the bundle of duplicated links between two characters to a single link with the multiplicity as a weight. The weighted degree of a node stays almost the same as in the two-mode network with exception for those links that point to paragraphs with just one character present because we will not consider loops.[2] In Table 6.1 we can see how these two-mode networks translate into one-mode networks for three example networks. If we compare the two relatively short novellas *Der Stern der Schönheit* by Wolf and *Auf Wiedersehen!* by Goldammer, the one-mode networks do look quite different. The two-mode network representation reveals that this difference is mostly due to the story within a story present in Wolf's novella.[3] Therefore, the characters appearing in the novella of Wolf form nearly disjunct sets with only few links between them.

6.3 Network ensembles

Network ensembles are collections of networks. As an object of study, one typically wants to find common structure among those networks. Possible ways of comparing networks include the differentiation of their global properties (Nagel (2011) calls them network indices), e.g., number of nodes, edge density, centralization, clustering coefficient (Watts and Strogatz, 1998), etc. Another possibility is the comparison of their inner structural properties, i.e., graph similarity. To do so there are several possibilities. First, one could consider graph edit distances. These are similarly defined as the word edit distances in Chapter 2.2. An overview over existing methods is given by Bunke (2000). Other approaches compare networks by statistical similarity, model similarity (Faust and Skvoretz, 2002) or consider certain subgraphs within the networks (e.g., Deshpande et al., 2005). Finally, spectral graph distances are another way of comparing networks. They totally discard the underlying networks and instead study the eigenvalues of the adjacency matrices of the networks. This is possible since these spectra, i.e., the sets of eigenvalues, do encode valuable structural information about the network. This is an extensive research field in graph theory. Nagel (2011, Ch. 7) provides a comprehensive review of spectral graph distances. Some of the properties incorporated in the spectrum of an adjacency matrix are the following:

1. Graph similarity measured by edit distance can be related to changes in the spectra.

2. The resulting changes can be interpreted as shifts of the involved eigenvalues and thus a measurement of this "eigenvalue movement" relates well to edit distance (Nagel, 2011, p. 84).

3. The skewness of the eigenvalue distribution is related to the number of triangles.

[2] links to the node itself that appear when a paragraph in the two-mode network has degree one.

[3] It could be even seen as meta diegetic because of the story that *Fernando* is sharing with *Lope de Vera* about how he imagined the *Stern der Schönheit / Prinzessin von Granada* to be.

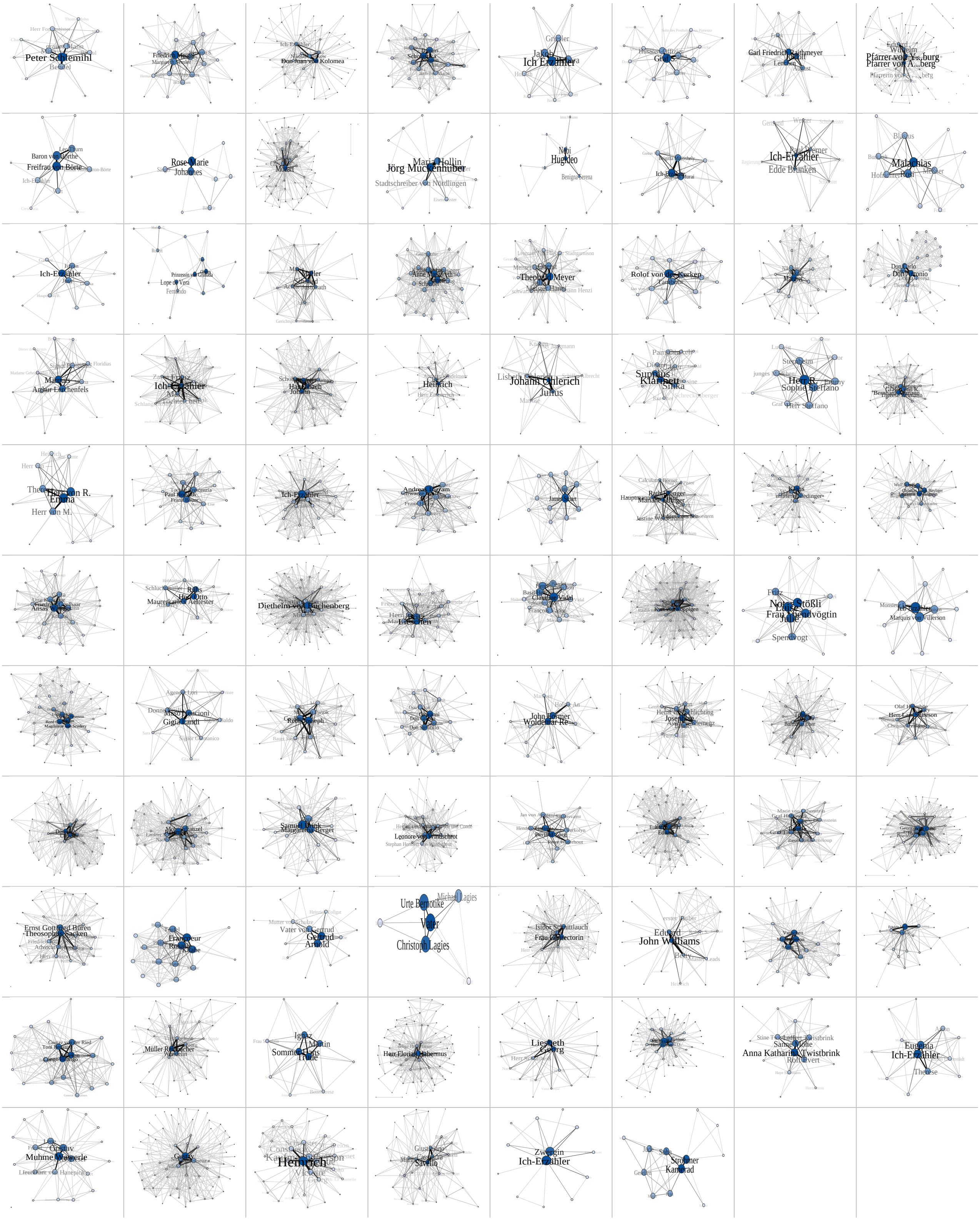


Fig. 6.1: All one-mode representations of character networks extracted from the *Novellenschatz*.

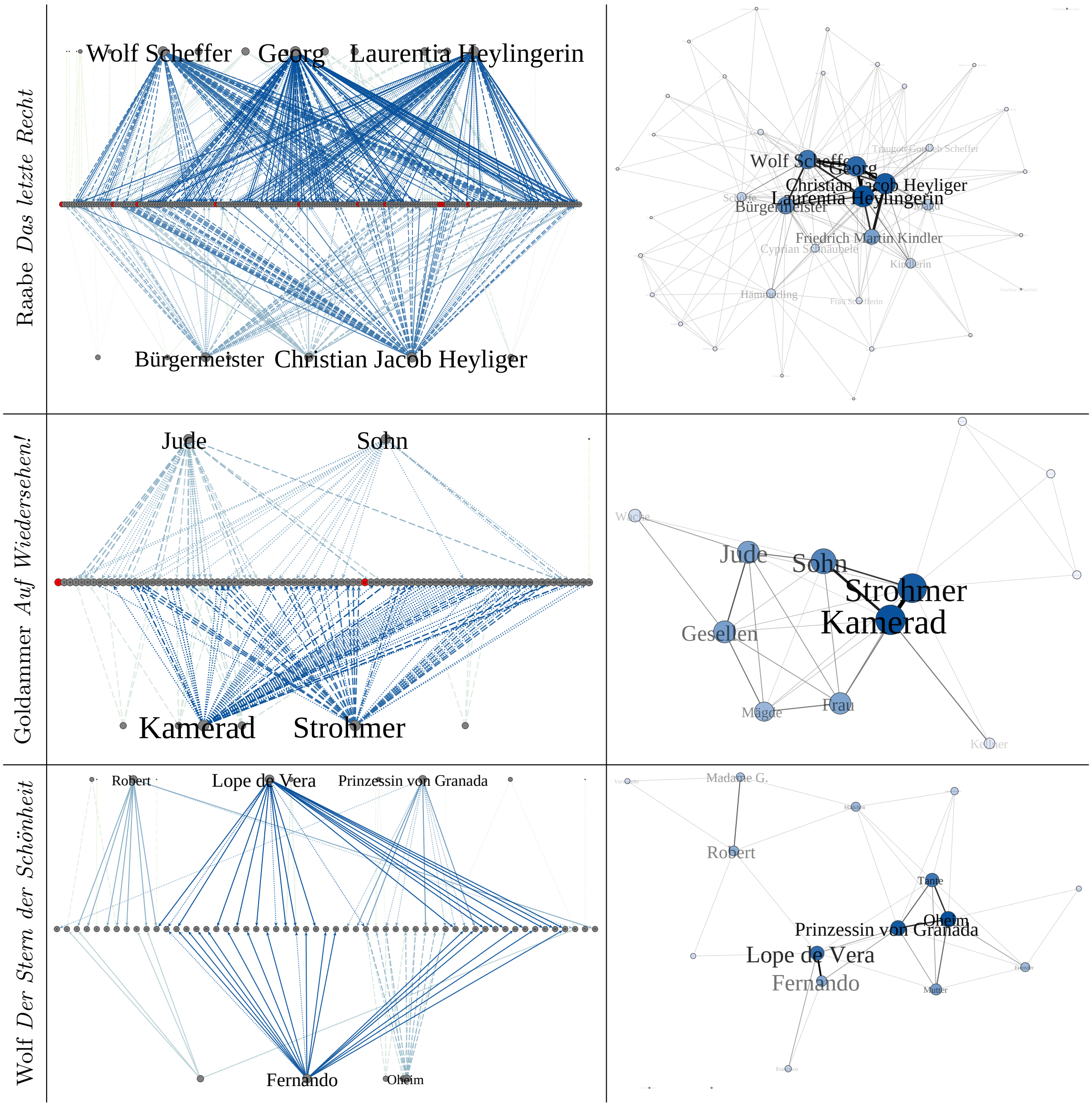


Table 6.1: Three examples of one-mode networks resulting from different two-mode networks. Node color, node area and node label size indicate node degree for the one-mode networks.

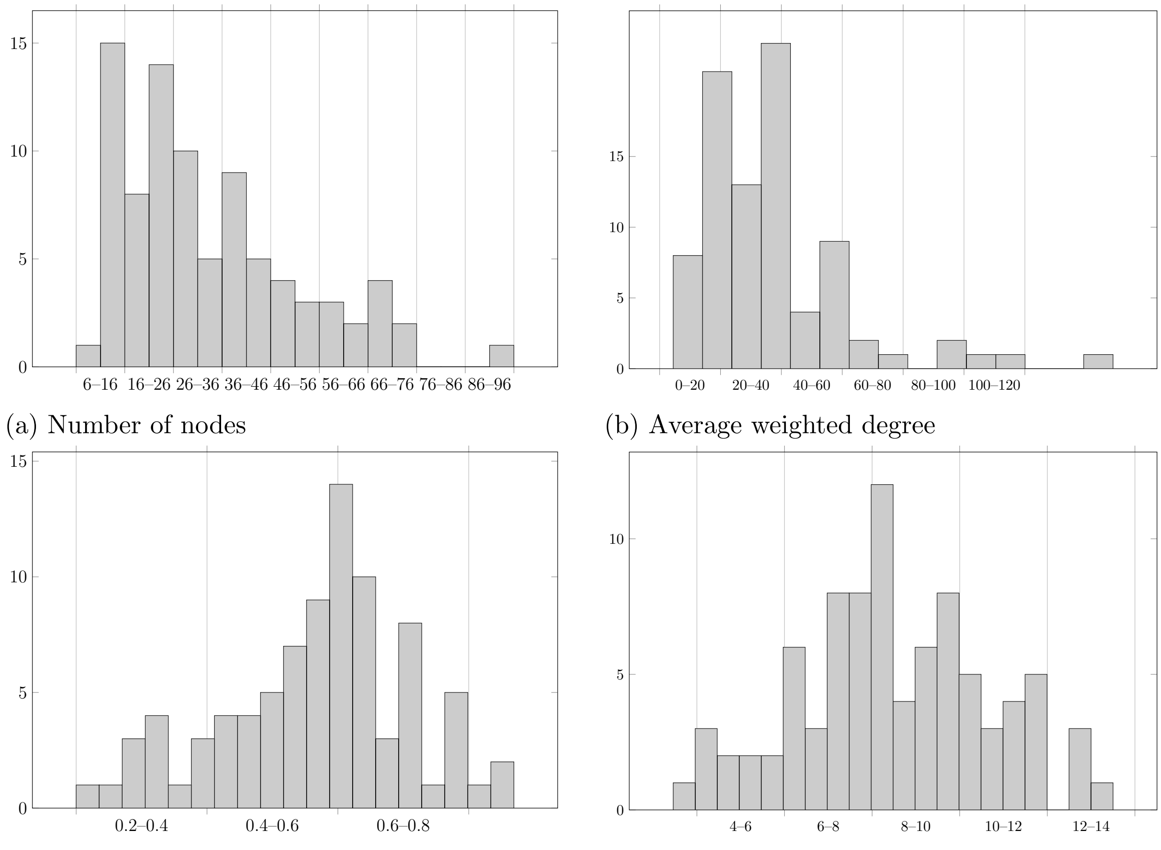
4. One can relate the eigenvalue distribution to the degree distribution and the density over the moments of the eigenvalue distribution.

5. Under certain conditions the exponential distribution of degrees results in an exponential distribution of the largest eigenvalues (Chung et al., 2003).

To answer our research question whether there is a predominant character constellation network type we will therefore use a twofold approach. In a first step, we will look at the similarity of some network statistics. In a second step, we compute a specific spectral distance that is based on spectrum transformation costs.

6.4 Network similarities

6.4.1 Network statistics



(a) Number of nodes

(b) Average weighted degree

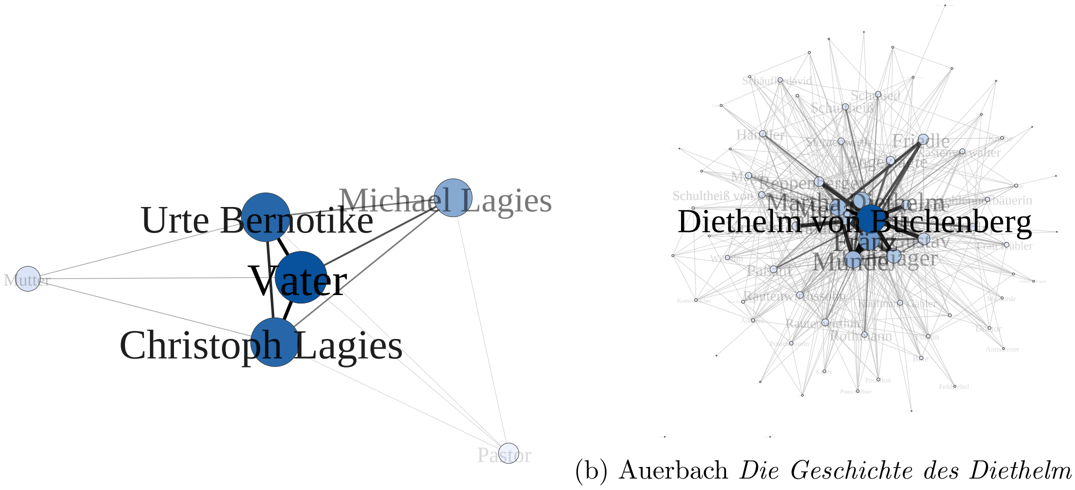
(c) Centralization

(d) Entropy

Fig. 6.2: Histograms for four network statistics of the one-mode networks.

If we want to compare the character networks of the novellas, we can get first insights when considering certain network statistics. In Figure 6.2 we plotted histograms of four network statistics that can explain certain aspects of our research question.

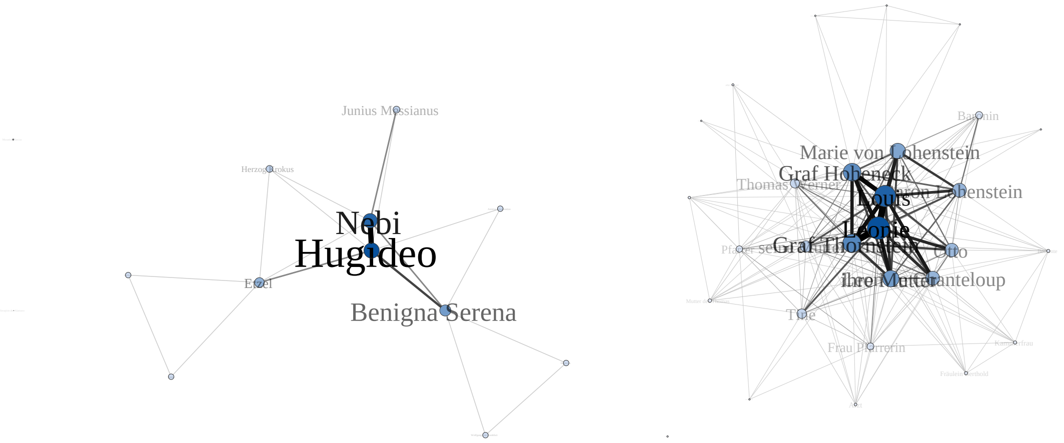
First, the number of nodes and hence characters are plotted. These range from 6 characters to 96 characters and are displayed in Figure 6.3. Second, we plot the average weighted degree, i.e., the average number of ties for each node with each tie counted by its multiplicity. Here, the numbers range from 4.46 to 149.45 (Figure 6.4).



(a) Goldammer *Eine Hochzeitsnacht*

(b) Auerbach *Die Geschichte des Diethelm von Buchenberg*

Fig. 6.3: The smallest and largest number of nodes.



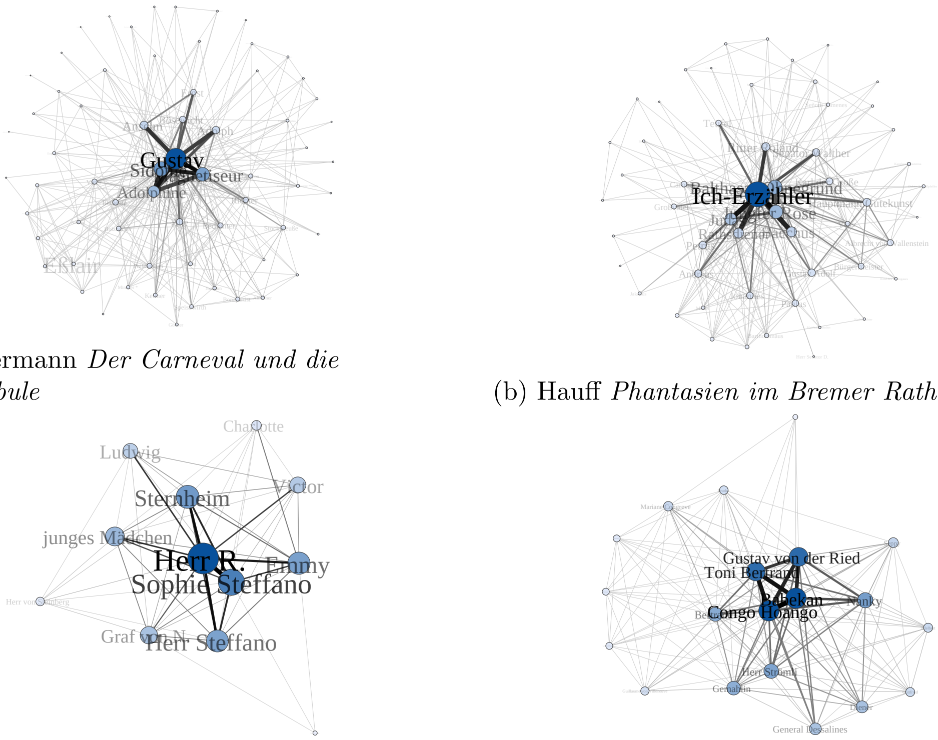
(a) Scheffel *Hugideo*

(b) Wild *Eure Wege sind nicht meine Wege*

Fig. 6.4: The smallest and largest average weighted degree.

The third histogram shows the graph centralization w.r.t. degree (gdc) as defined by Freeman (1978) and indicates how 'star'-like a network is. Jannidis (2017) showed that the novellas from the *Novellenschatz* do have a higher centralization then a comparative novel corpus. This might be a first hint that the restricted form indeed gives rise for a specific character constellation. The novellas with the lowest centralization are Kleist's *Die Verlobung von St. Domingo* ($C_{deg} = 0.275$), Fräulein Wolf's *Gemüth und Selbstsucht*

($C_{deg} = 0.236$), and Goldammer's *Eine Hochzeitsnacht* ($C_{deg} = 0.200$). The one with the highest centralization are Immermann's *Der Carneval und die Somnambule* ($C_{deg} = 0.869$) and Hauff's *Phantasien im Bremer Rathskeller* ($C_{deg} = 0.834$). Their 'star'-like structure and the opposing nearly full graph are plotted in Figure 6.5.



(a) Immermann *Der Carneval und die Somnambule*

(b) Hauff *Phantasien im Bremer Rathskeller*

(c) Fräulein Wolf *Gemüth und Selbstsucht*

(d) Kleist *Die Verlobung von St. Domingo*

Fig. 6.5: High (top) and low (bottom) centralized networks.

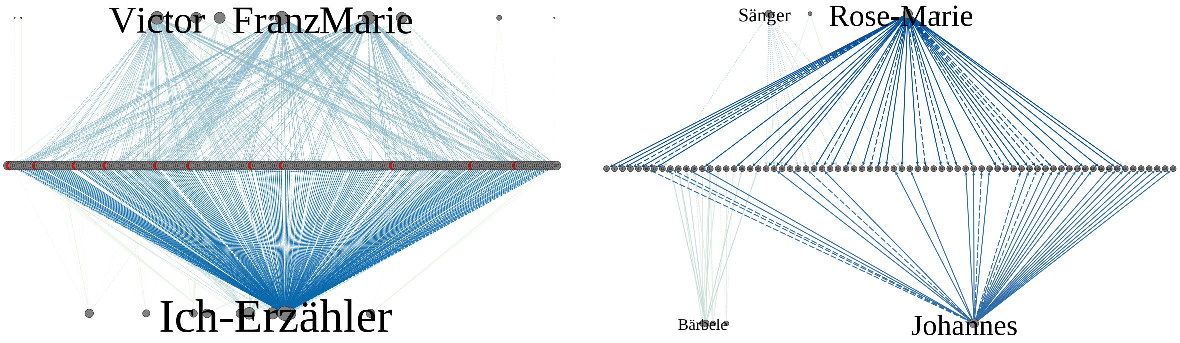
In Figure 6.2d we show a histogram on the entropy of the characters within a novella, i.e., whether they are equally distributed over the text or highly concentrated in certain paragraphs. Here, the entropy is given by

$$H(D_i) = -\sum_{i=1}^{m} p_i \log p_i \ ,$$

where

$$p_i = \frac{n(s_i)}{\max_s n(s)}$$

is the number of characters $n(s_i)$ in paragraph s_i divided by the maximum number of characters in one paragraph of the novella. Here, we find that Lewald's *Die Tante* has the highest entropy (and is hence, the story with the most equidistributed characters) and Ludwig's *Das Gericht im Walde* has the lowest entropy (i.e., a high concentration of characters). The corresponding two-mode networks can be found in Figure 6.6.



(a) Roquette *Die Schlangenkönigin*

(b) Ludwig *Das Gericht im Walde*

Fig. 6.6: Two-mode networks for the novellas with the highest and lowest entropy.

6.4.2 Spectral graph distance

Signature transformation cost. To compare the networks in the ensemble we regard each network as an instance in a metric space. Therefore, we use a specific spectral graph distance (Nagel, 2011, Ch. 6 and 7), a similarity measure that compares networks by considering the signature transformation costs. Its invariance under network size and automorphisms make it applicable for our purpose of comparing 86 different networks of diverse size with unrelated characters.

More formally, the distance between two novellas is given as

$$\mathrm{dist}(D_1, D_2) = \min_{F \in \mathcal{F}} \sum_{e_i, e_j} ||\lambda_i - \lambda_j|| \cdot F(e_i, e_j), \qquad e_i = (\lambda_i, \frac{m_i}{n})$$

where all λs are eigenvalues of the corresponding adjacency matrix from the character network, m_i the multiplicity of λ_i and $n = \sum m_i$.

Implementation. To compute the signature transformation cost, one can convert the question into the search for a perfect matching of minimum weight. We have implemented the methods described by Nagel (2011) for the undirected and directed version of the approach as a python package. As described by Nagel (2011), the undirected approach is far more efficient than the directed one. This is because the directed approach needs to find a minimum weight full matching. However, this needs a runtime of $\mathcal{O}(\mathrm{lcm}(n, m)^2 \log(\mathrm{lcm}(n, m))$ for a bipartite graph of size $2 \cdot \mathrm{lcm}(n, m)$ where n and m are the number of nodes in the two networks (Karp, 1980).

Application. The computed similarity measure is stored as an adjacency matrix containing the pairwise distances of each of the 86 networks. For visualization, we use metric MDS to translate the pairwise distances into two-dimensional data points (Figure 6.7). The result can be reviewed for structural properties.

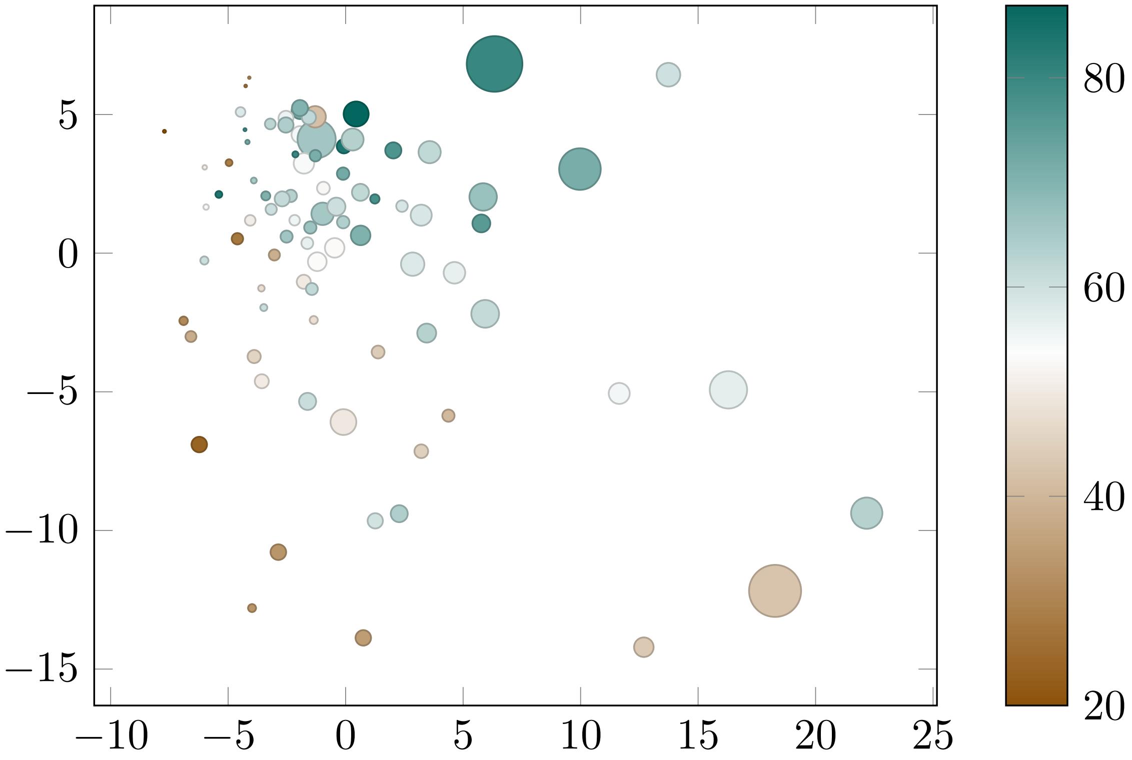


Fig. 6.7: Each point represents a novella as a character one-mode network. The points are plotted with metric MDS using the SMACOF algorithm (Kruskal, 1964). The area of each circle is proportional to the text length and the color indicates the graph degree centralization (in %).

6.5 Findings

We hypothesized that the novellas would have a common character constellation. In contrast, we observed in Figure 6.2 that the different novellas do have a variety of present characters, diverse network densities, as well as some differences in centralization and entropy. However, in Figure 6.8 we can observe that at least two features are highly correlated to text length. Nonetheless, these two values are also competing. The fewer characters there are on a large amount of text, the more likely we have a high average weighted degree. The two longest novellas of Auerbach and Wild do lead in one of the two categories but lag in the other.

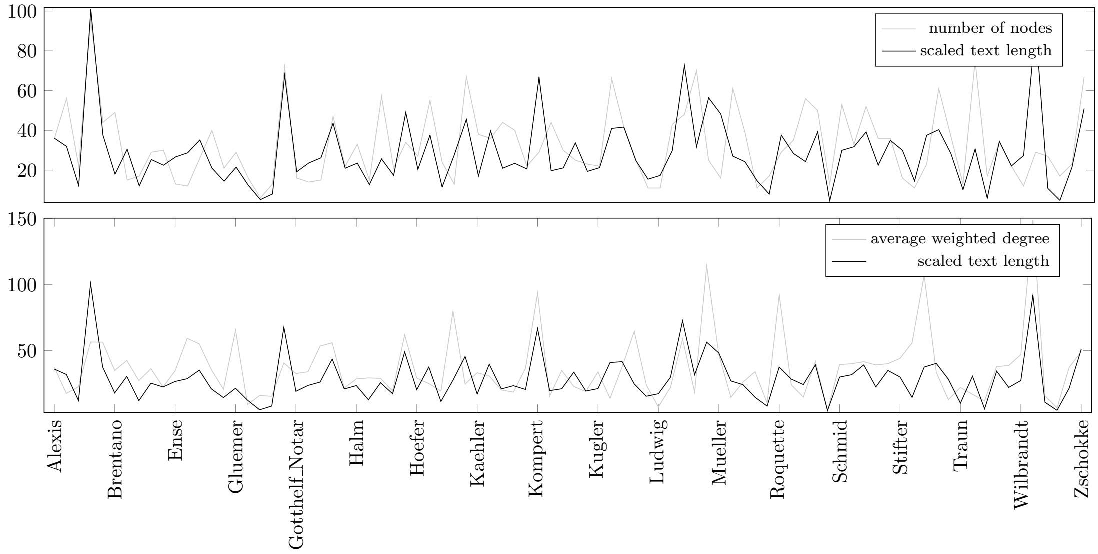


Fig. 6.8: Comparison of network statistics and text length

Instead, we used a spectral graph distance and indeed observed a short distance for a large share of the corpus elements. Moreover, we want to emphasize the positions of two exemplary novellas. Auerbach's *Die Geschichte des Diethelm zu Buchenberg* is the longest novella in the corpus and is seen as a novel instead of a novella by many literary scholars. This could be a possible explanation for its outlier position at the top of the plot in Figure 6.7. In addition, for the co-occurrence network Heyse's *Der Weinhüter von Meran* is the novella closest to the centroid, an interesting finding, especially if we recall that Heyse is one of the editors of the *Novellenschatz* and was called the virtuoso of the average[4] (Jeziorkowski, 1987) by other scholars.

6.6 Conclusion

We used a spectral graph distance measure to analyze character constellations in a corpus of novellas, and found that, outliers notwithstanding, high similarity overall. However, this can only truly be verified by holding the texts against a contrasting corpus (see Chapter 8.2. Additionally, the approach can be understood as a guiding principle for the more general comparison of network ensembles in the digital humanities beyond considering the descriptive statistics. We showed that the method yields interpretable results for our weighted networks. As other applications one could imagine to cluster character networks from movies to investigate whether there is a genre specific character constellation, compare networks in diverse archeological settings with each other (e.g., trade networks of different cultures), or analyze the style of correspondence networks for authors.

[4] "Virtuose des Durchschnitts"

Chapter 7

Classification of Character Networks

In Chapter 6 we observed that despite the fact that the criterion for being included as a novella in the *Novellenschatz* is rather strict this can still lead to a variety of different character networks. We now want to investigate and classify the possible network structures in detail. In Figure 7.1 we see a different visualization of the spectral graph distance. Instead of using MDS we consider again the Simmelian backbone network that is created with the same parameters as in Figure 4.1. We performed a Louvain modularity clustering on the complete distance matrix (as in Chapter 2) to better visualize the clusters formed by these networks.

We will analyze the networks that have been created in Chapter 5 at different scales. Hence, this can be seen as another example of the at-scale-approach as a guiding principle for our analysis. On the macro level our research approach is guided by the question of whether it is possible to use the novella networks to separate disjunct character sets. This could potentially lead to the separation of extra diegetic from intra diegetic narrative or expose relocations of the plot. On the micro level we want to know what archetypal network types for the novella ensemble exist and how those networks can be categorized, characterized, and analyzed. Therefore, we will stick to the approach from Chapter 3. We start each paragraph with some examples from the corpus. Then, we will describe a generalized procedure that (best possibly) reconstructs the group and name all other novellas that fall into this category.

First, we want to further understand how the specific network ensemble in Figure 7.1 was formed. Since the network distance is based on the weighted network, the various types of main character constellations should have an influence on the position in the network ensemble. A different way of characterizing each network is by considering its degree sequence. In Figure 7.2 we see the result of applying a k-means clustering (MacQueen, 1967; Lloyd, 1982) on the sequences to find five clusters. Interestingly, the novellas of Auerbach and Gotthelf were so different, that they fall into a category of their own. By observing the different node shapes in Figure 7.1 we see that only parts of the group building process visi-

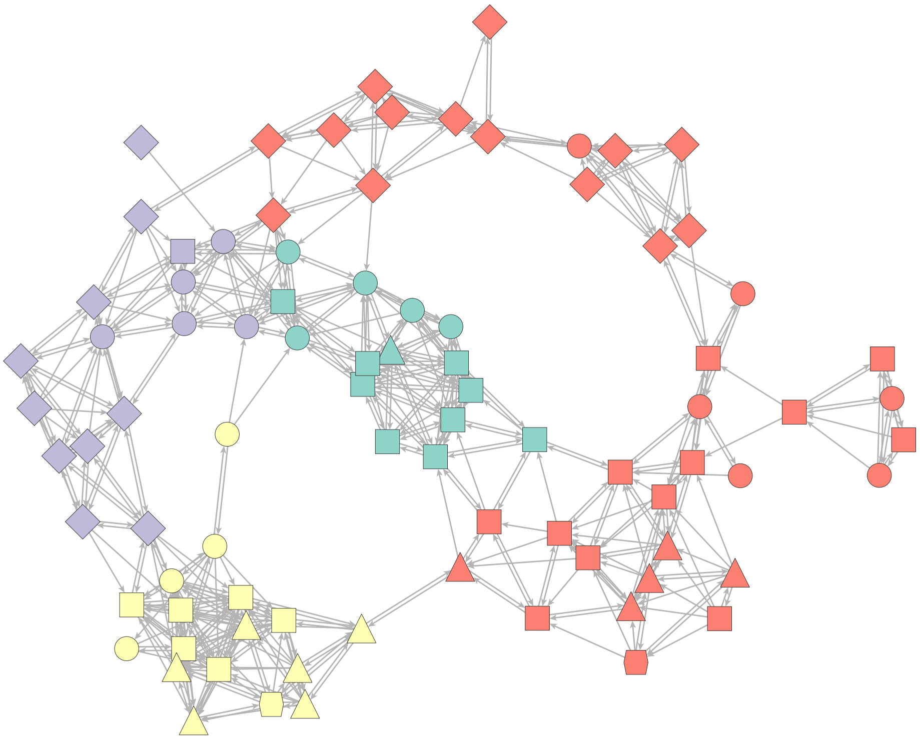


Fig. 7.1: Simmelian backbone network of the ensemble of one-mode co-occurrence networks based on the distance matrix given by the spectral graph distance. The color is based on Louvain modularity clustering while shape indicate the respective k-means cluster of the degree sequences (see Figure 7.2).

ble in the backbone visualization can be explained by the degree sequence. If we recall that the spectrum incorporates many more properties, this is highly expected. In the following we will discuss certain subgroups in further detail.

7.1 Macro level structure

7.1.1 Story within a story

A typical element of a novella is the story within a story. It is common that a protagonist becomes the narrator of a longer section of the text in which they either tell a story about themself or about someone else. One possibility is that the novella is part of a larger novella cycle. In that case, a loose extra diegetic narrative is used to combine different novellas with each other. Here, the extra diegetic narrative is usually extremely short compared to the intra diegetic narrative (e.g., Storm *Eine Malerarbeit*). Alternatively, the intra diegetic narrative is told at some point in the story (e.g., Lorm *Ein adeliges Fräulein*). In that case it usually contains the key information to dissolve the plot. Recognizing these second type

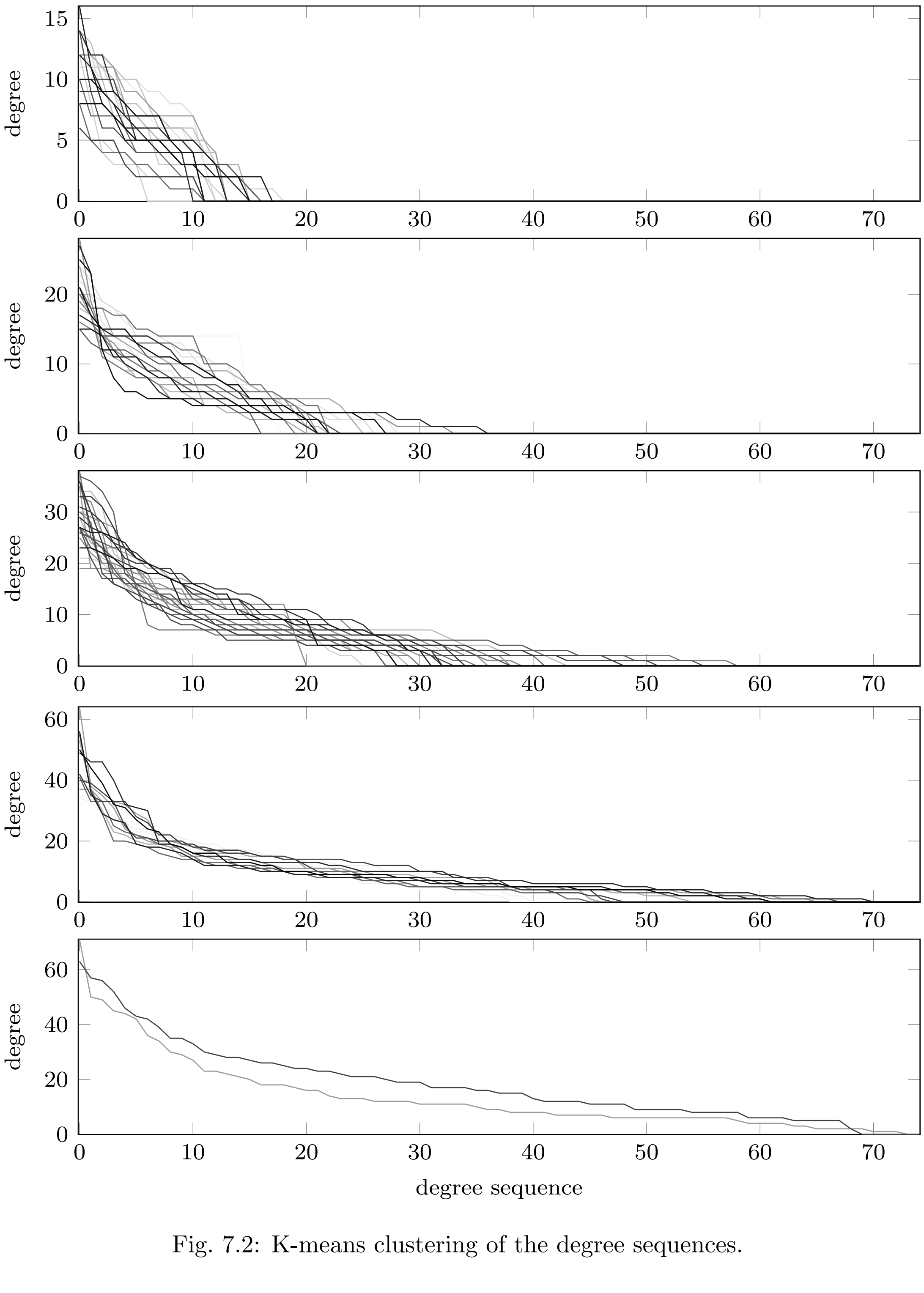


Fig. 7.2: K-means clustering of the degree sequences.

of stories within a story is harder but also more fruitful.

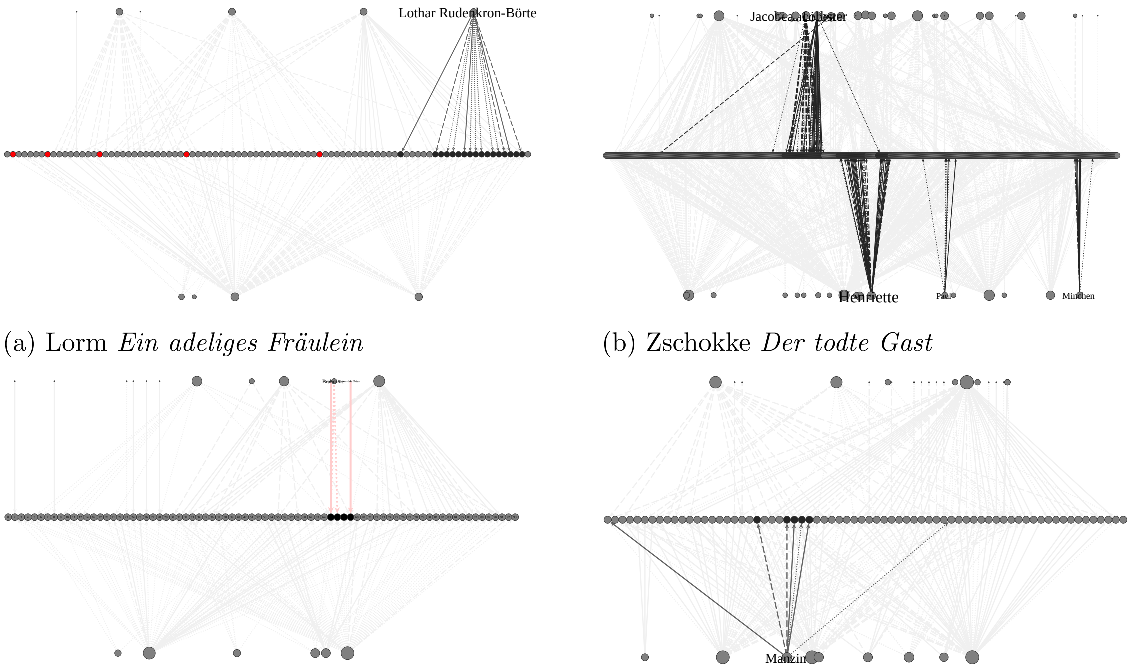
The most straightforward way would be by looking for keywords in the text that point to a narrative being told. However, we want to find those stories based on the underlying networks. Our approach therefore spots highly dense character appearances of otherwise unpresent figures (recall Figure 3.3). A high density of a character or a group of characters in a certain region of a story that did not appear in the same frequency outside of this region is a clear marker for such an intra diegetic narrative. If on the one hand this is happening for characters at the beginning and end it usually points to the first case and therefore to an extra diegetic narrative. If on the other hand this appears somewhere in between it usually points to an intra diegetic narrative. To identify such characters in our novellas, we use an automated procedure (Alg. 2).

Algorithm 2 get dense character appearance

```
Input List of characters with at least 5 appearances, link list from Chapter 5
for character in character list do
    for link in link list do
        if link contains character then
            start ← number of section
            skip to next iteration
        end if
    end for
    for link in link list[last:first:-1] do
        if link contains character then
            end ← number of section
            skip to next iteration
        end if
    end for
    if $\frac{deg_{char}}{\text{end}-\text{start}} > 0.5$ and $\frac{deg_{char}}{\text{number of paragraphs}} < 0.5$ then return Character
    end if
end for
```

We now want to consider the example networks of four novellas by Heyden, Keller, Lorm, and Zschokke. In Figure 7.3 we highlight four sets of dense character groups. However, on the one hand, only three of them could be found by Alg. 2. The intra diegetic story in Heyden's *Der graue John* could not be retrieved because it was to short in comparison with the relatively long paragraph length. Therefore, we fail to find the characters that are relevant. On the other hand, a character that matches the criterion is `Manzin` in Keller's *Romeo und Julia auf dem Dorfe*. While her presence points to a specific period in the narrative[1] it is not part of an intra diegetic story.

[1] When the destitute couple `Manz` moves to the town to work in a pub.



(a) Lorm *Ein adeliges Fräulein*

(b) Zschokke *Der todte Gast*

(c) Heyden *Der graue John*

(d) Keller *Romeo und Julia auf dem Dorfe*

Fig. 7.3: Plot of four two-mode networks. In (a), (b) and (d) we highlight the characters that are retrieved by Alg. 2.

7.1.2 Connecting characters

As we have observed above, it is possible that some characters are only present in a subpart of the story. While this might hint to an intra diegetic narrative it is very much likely that many characters do only play a role in certain chapters. However, other characters are presents throughout the entire plot and therefore connect these different characters. There is a possibility in grouping the characters. Indeed, if we apply a Louvain clustering to the novellas, we can assign different "roles"[2] to the characters.

We want to study six novellas in detail for this purpose (see Figure 7.4). In Brentano's *Geschichte vom braven Kasperl und dem schönen Annerl* we observe that the eponymous characters `Annerl` and `Kasper` fall into the role of a connector between `Anne Margareth` and the group of `Josef Grossinger` and the duke (`Herzog`). For Droste's novella we can highlight two groups that are connected by two different characters. The couple `Hermann Mergel` and `Margareth Semler` and `Förster Brandes` together with `Herr von S.`. They are connected via `Friedrich Mergel` and `Johannes`. However, they do not play the same role. While `Friedrich Mergel` is the connector in the first part, `Johannes` takes over in the last part of the novella. For Kopisch's *Ein Karnevalsfest auf Ischia* we can see the connection between `Don Antonio` and his servant `Pietro`. In Johanna Kinkel's and Rumohr's novallas we see clear separations between characters `Frau Werl` and `Concertmeister Sohling` respectively between `Margaretha` and `Giustiniano` with his wife `Cassandra`. Finally, for Sternberg's *Scholastika* the groups again consist of multiple characters.

7.2 Micro level structure

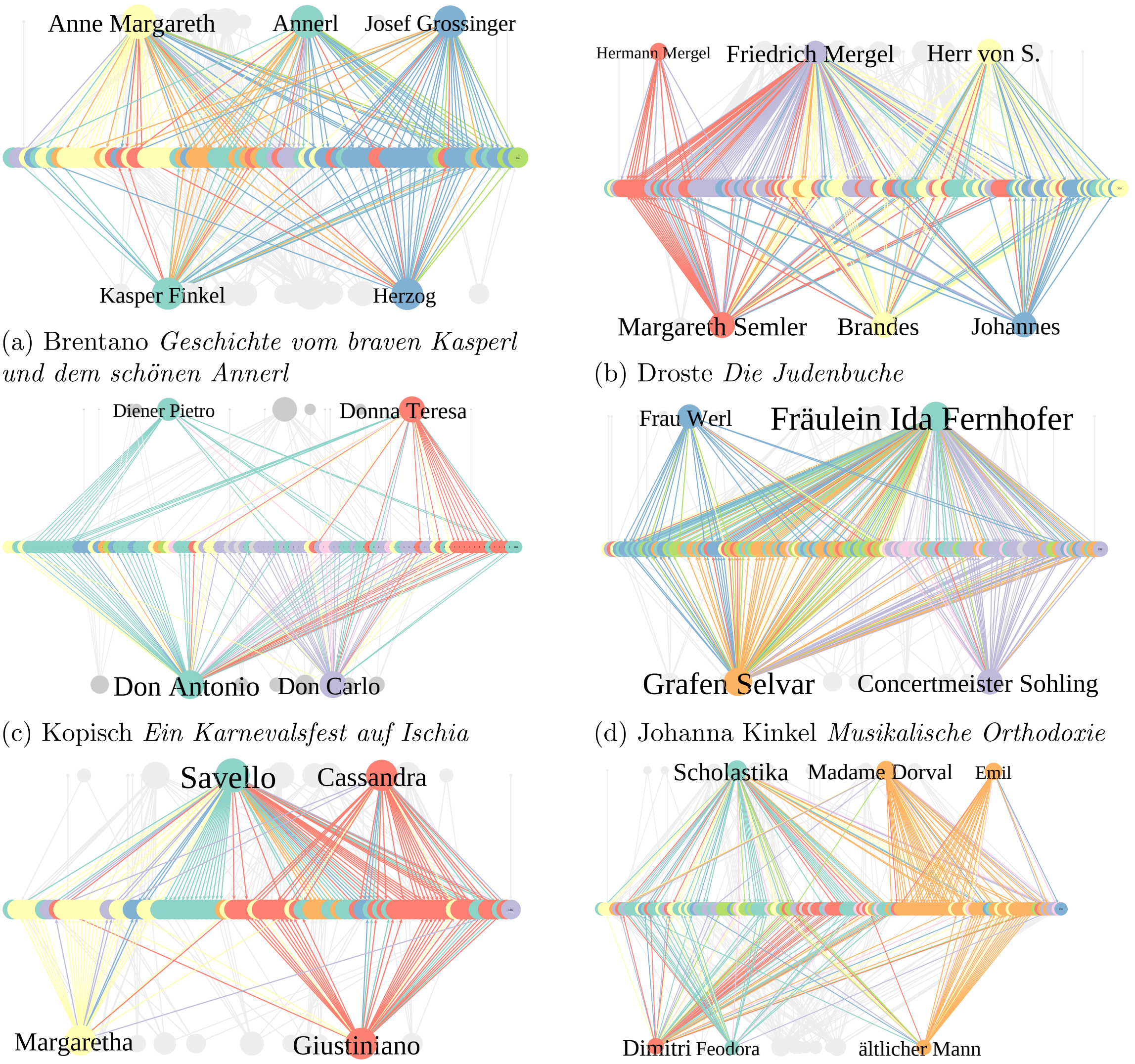
7.2.1 Star networks

Star networks are characterized by a central hero whose interactions defines the story line. By construction, first-person narratives are very likely to fall in this category since they could also be seen as ego networks. In practice, we want to separate those ego networks from the other star-like networks. Examples of such first-person networks are Chamisso's *Peter Schlemihls wundersame Geschichte* or Hauff's *Phantasien im Bremer Rathskeller.* Other novellas that can be characterized like this are Auerbach's *Die Geschichte des Diethelm zu Buchenbach* and Waldmüller's *Es ist nicht gut, dass der Mensch allein sei.*

In Figure 7.5 we see that their one-mode networks are highly centered around the main character. As already discussed in Chapter 6 this can be incorporated in the graph degree centralization. However, other novellas could also have a high centralization score. As an example, we want to point out Goethe's *Die neue Melusine.* While the novella is highly centralized ($gcd(D) = 0.833$) it is based on two central characters.

Instead, we want to consider a large difference between the highest (d_0) and the second highest weighted degree node (d_1) as the most distinctive attribute for this group. Here,

[2] Using an actual role analysis would be preferable. However, defining a useful role equivalence is complicated on the two-mode structure.



(a) Brentano *Geschichte vom braven Kasperl und dem schönen Annerl*

(b) Droste *Die Judenbuche*

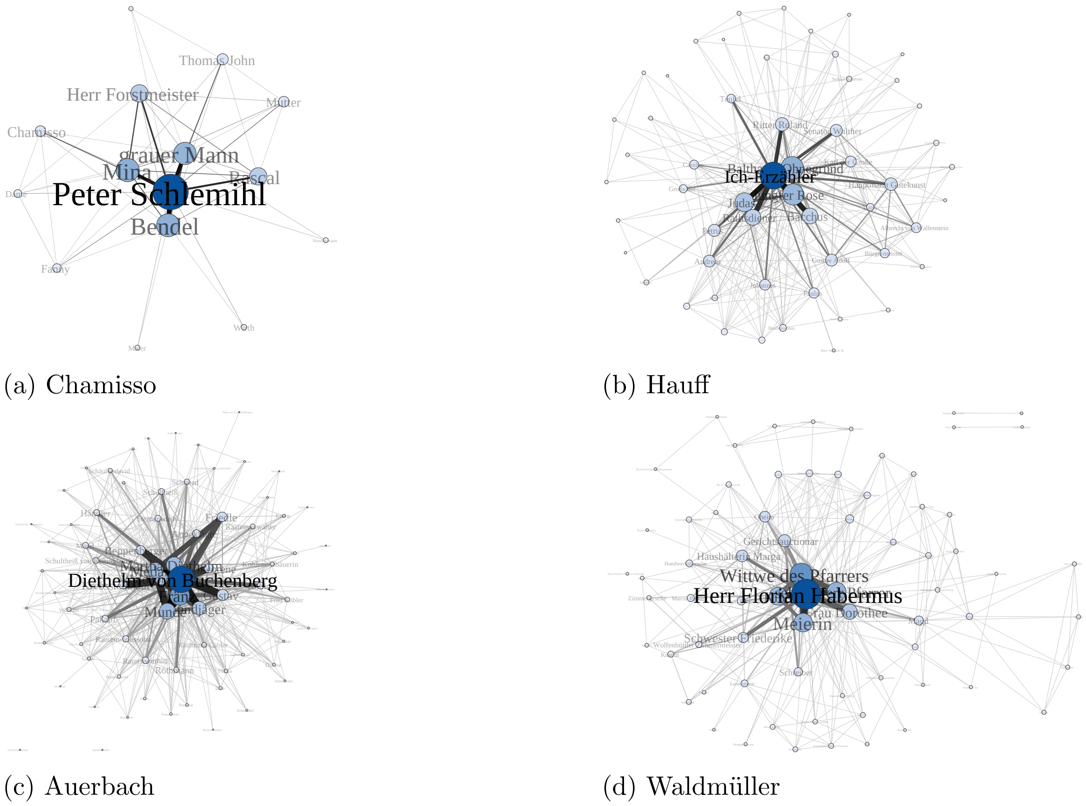
(c) Kopisch *Ein Karnevalsfest auf Ischia*

(d) Johanna Kinkel *Musikalische Orthodoxie*

(e) Rumohr *Der letzte Savello*

(f) Sternberg *Scholastika*

Fig. 7.4: Plot of six two-mode networks. We apply Louvain clustering on the nodes. This makes it possible to identify the role and position of certain characters that are highlighted.



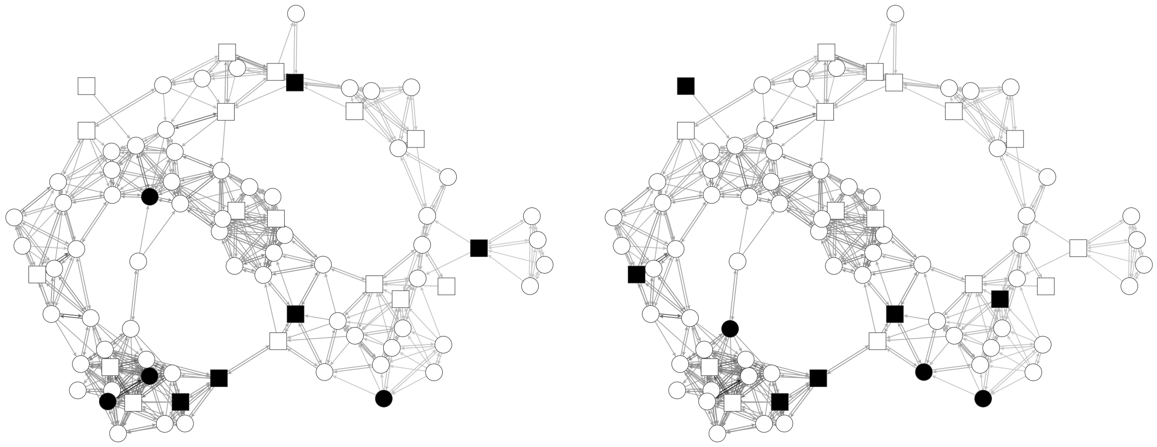
(a) Chamisso

(b) Hauff

(c) Auerbach

(d) Waldmüller

Fig. 7.5: Display of certain characters in two-mode networks. We highlight certain characters. Color represents the Louvain cluster.



(a) High decline in degree sequence

(b) group degree centralization

Fig. 7.6: Same network as in 7.1 with star networks highlighted in black. The sqares indicate first-person narratives.

Goethe's fairy tale does not pass a meaningful threshold, since the two main characters have weighted degrees of 112 and 96.

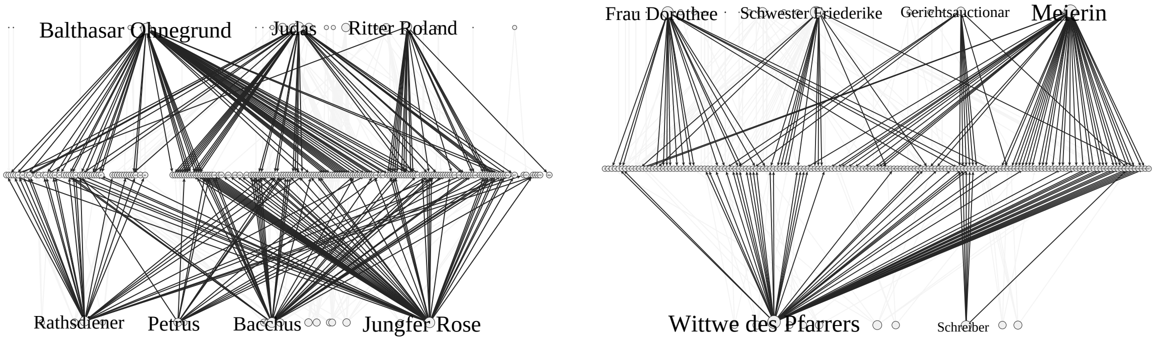
In Figure 7.6a all novellas that have a decline between the first two entries of the degree sequence of more than 40% are marked in black.[3] Beside the above-mentioned texts, these are the first-person narratives of Kähler, Roquette, and Immermann as well as the novellas of Hackländer and Mörike. In contrast, next to Goethe, also Wildermuth, Wallner, Meißner, and Schreyvogel have a high centralization without being centered around a singular figure.

For the star networks, it is interesting to investigate the two-mode network without the main character. Then we see who the protagonist meets up with and it usually gives much more information about the story line. Even though we cannot reconstruct the story (because the key motifs[4] are usually missing) we can see the characters that are involved and those that the main figure meets. In Figure 7.7 we can view the networks of Hauff and Waldmüller. While the characters of Hauff's *Phantasien im Bremer Rathskeller* do appear and reappear over time in complex connections[5], the characters of Waldmüller's *Es ist nicht gut, dass der Mensch allein sei* appear in nicely ordered sequences. Hence, it is possible to retrieve some of the story line. If we recall that in Chapter 2 we pursued the question whether the novellas have a strong silhouette, the silhouette can here be identified in the two-mode graph itself.

[3] $\frac{d_0-d_1}{d_0} > 0.4$.

[4] "Dingsymbol"

[5] In fact, apart from the `Rathsdiener` and `Ritter Roland` they are all in the same room and join and leave the conversation as they please.



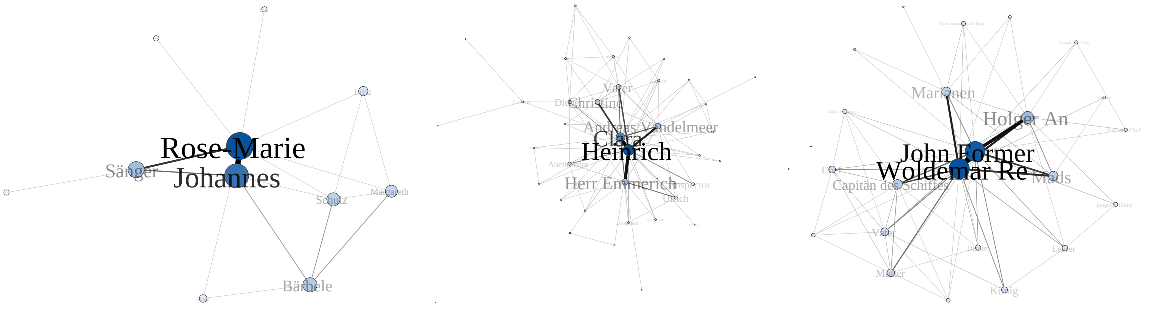
(a) Hauff without the narrator

(b) Waldmüller without Florian Habermus

Fig. 7.7: Two-mode networks of Hauff and Waldmüller without the main characters. We highlighted the most common characters and their links to the paragraphs where they are present.

7.2.2 Love and friendship

A second reoccurring theme in the *Novellenschatz* is the story of a couple or a friendship. In these novellas there is usually one strong edge in the one-mode network. Prime examples include Tieck's *Des Lebens Überfluß*, Ludwig's *Das Gericht im Walde*, and Kruse's *Nordische Freundschaft* that already bears the connection in its name (Figure 7.8). If we assume that the two main characters are present throughout the story it is again of interest to study the network without them. By highlighting the node degree of the paragraphs, we can carve out periods where the couple is alone. Here, the three novellas differ. While the two main characters in Ludwig's *Das Gericht im Walde* (The court in the forest) actually leave into the woods and abandon all other characters this does not hold for the other novellas. The two characters in Tieck's novella also experience a phase of loneliness (they even burn the stairs that lead to their room). However, at the end they reunite with other characters from earlier on. For the novella of Kruse there is no part where the main characters are alone.

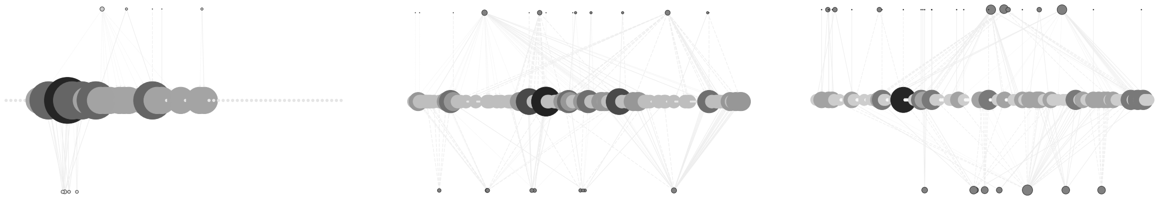


(a) Ludwig

(b) Tieck überfluß

(c) Kruse

Fig. 7.8: Three networks that have a couple in focus.



(a) Ludwig (b) Tieck *Des Lebens überfluß* (c) Kruse

Fig. 7.9: The two-mode networks of Ludwig, Tieck, and Kruse without their two main characters. The degree of the paragraph nodes is mapped on the node size and node color.

If we mimic the computation from Chapter 7.2.1 but replace the largest and second largest degree with the largest and second largest edge multiplicity we can extract such novellas. If we take a threshold of 0.55 we find eight novellas that fall into this category. Beside the novellas from Figure 7.8 and the above mentioned *Die neue Melusine* from Goethe, these are Johanna Kinkel's *Musikalische Orthodoxie*, Riehl's *Jörg Muckenbuber*, Scheffel's *Hugideo*, and Wildermuth's *Streit in der Liebe und Liebe im Streit.* In Figure 7.10 we observe that these texts are indeed close within the network ensemble. However, if we consider the Louvain clustering of Figure 7.1 they belong to different clusters. But this can be explained since the texts from the two clusters significantly differ by text length.

An outlier in the corpus in general is the text of Goethe. However, this is expected and has already been clear to the editors, who placed the fairy tale as a role model despite its uncommon choice.[6]

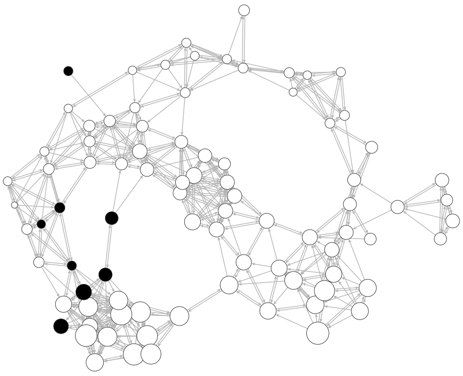
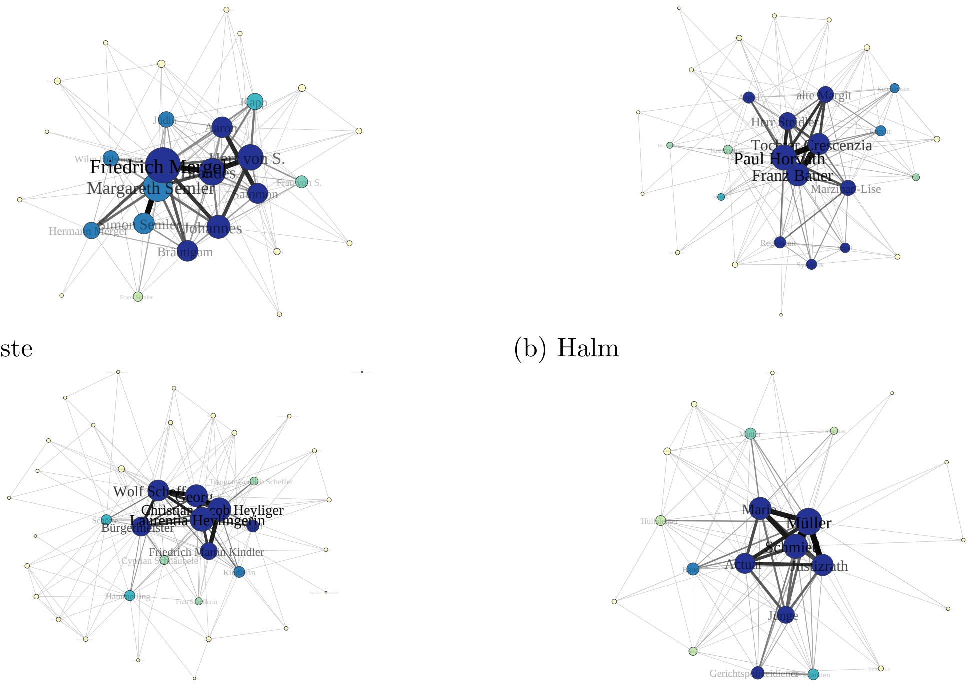


Fig. 7.10: Same network as in 7.1 with love and friendship networks highlighted in black. The node size is proportional to text length.

[6] A straightforward choice would have been Goethe's text "Novelle".

7.2.3 Justice novellas

If we approach the texts based on the topic they deal with, there are few surprises. 46 of 86 novellas are classical love and marriage themes. The other 40 novellas spann categories like village live (8), nobility (6), and art (6). However, there are also six novellas that center around justice. Namely Droste's *Die Judenbuche*, Halm's *Die Marzipan-Lise*, Raabe's *Das letzte Recht*, Ziegler's *Saat und Ernte*, Meißner's *Der Müller vom Höft*, and Riehl's *Jörg Muckenbuber*. As mentioned above, the text of Riehl is also very centered around two characters and therefore a little different from the other texts from this category. Interestingly, the first four mentioned justice novellas are all in the central cluster (green) of Figure 7.1.[7] They are joint by the texts of Kugler, Rumohr, Eichendorff, Höfer, Brentano, Spindler, Keller, Sternberg, Wichert, Alexis, and *Der Träumer* by Kopisch.



(a) Droste

(b) Halm

(c) Raabe

(d) Ziegler

Fig. 7.11: Display of the four justice novellas. The node color is based on the k-core number.[8]

In Figure 7.11 we observe that these networks are characterized by a relatively large core (typically 5-6 novellas) and an inner and outer periphery. Like in a court situation

[7] One could bring forward that justice novellas try to moderate between the other texts but due to the arbitrariness of the visualization one should be careful with such claims.

[8] We computed the unweighted k-core (Batagelj and Zaveršnik, 2003) but discarded links with multiplicity one and two.

the novellas argue for different points of view and the multitude of main characters usually present the different sides.

The only other justice novella that does not fall into the above category is the one of Meißner. Instead, it has outgoing ties to Schäfer, Schmid, Gall, and Kähler. However, it is by far not obvious why it should not fall into the cluster of the other since its network shares a lot of the properties discussed above.

7.2.4 Long tail networks

If we consider the bottom left (yellow) Louvain cluster of Figure 7.1 we observe that it contains many novellas that have a large amount of characters that are only mentioned once (Figure 7.12). Among these characters, a large group ($> 30\%$) consists of biblical, historical or literary figures. While is no correlation between the epoche or the time of first publication of a text and the use of those figures, there seems to be some variation in the names used. In both epochs, biblical and literary references where equally common. However, the use of references to characters from the ancient world increased over time. Not as pronounced, we also observe an increase of mentions of political and historical figures. One interesting fact is, that the novella *Die beiden Tubus* of the editor Kurz is among those that has a "long tail". It has 39 characters that are mentioned only once and is therefore second place in this statistic only beaten by Waldmüller's text (44 single mentioned characters).

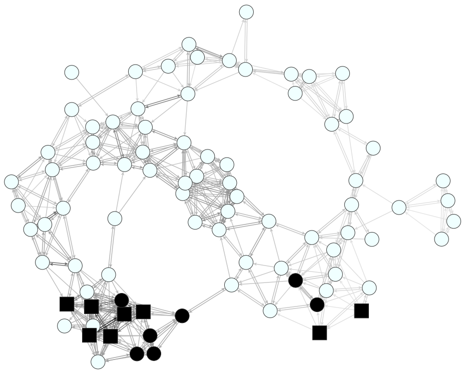


Fig. 7.12: Same network as in 7.1 with networks highlighted in black that have at least 20 characters that appear only once. The shape of these novellas marks their year of publication (before 1848 = circle, after 1848 = square).

7.2.5 Hidden persons

A last group that we want to discuss are novellas that have a common way of plot resolution. As examples, we want to point to Kähler's *Die drei Schwestern*, Stifter's *Brigitta*, and

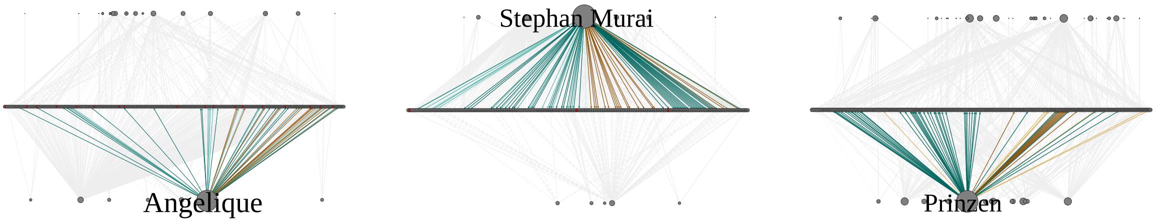
[8] Biblical references in before 1848 $\frac{6}{116}$ vs $\frac{12}{255}$ after 1848. Literary references before 1848: $\frac{11}{116}$ vs $\frac{24}{255}$ after 1848

biblical	**historical/political**		**greek or roman**		**literary**	
Paulus	*Vicekönig von Aegypten*	Kaiserin von Rußland	*Amor*	Empedokles	*Don Miguel*	Laura
Lea	*Livius*	Tuchmacher	*Prometheus*	Carluccio	*Horaz*	Beatrice
Rachel	*Lea*	Carls X.	*alten Juno*	Redner Stomachus	*Luther*	Jungfer Gretchen
Moses	*Hannibal*	Gustav Adolph	*Apollo*	Antoniello	*Tieck*	Vittoria
Laban	*Gouverneur von Martinique*	König von Syrien	Homer	Lunardo	*Wallenstein*	Paul Gerhardt
Jesus Maria	*heiligen Aloysius*	Heinrich II.	Sophokles	Raffaele	*Don Quixote*	Linné
Joseph	*heiligen Antonius*	Friedrichs von Braunschweig	Aristoteles	König von Thule	*Dürer*	Goethe
fromme Sirach	Kaiser in Rußland	Napoleon	Hekuba	Juno	*Maria zu Dresden*	Schiller
Salomo	Herzog von Enghein		Andromache		*Van Dyk*	Faust
Eva	Friedrichs des Großen		Magdalene		*Van der Werst*	Arlekin
Adam	Rousseau		Ariadne		*Rembrandt*	Don Carlos
Jesus Christus	Kaiser Napoleon		Antigone		Jean Paul	Magister Hölderlin
Goliath	Kaiser von Rußland		Minerva		Münchhausen	Magister Schelling
Judas Jschariot	Magister Hegel		Bacchus		Diderots Jaques	Grete
Propheten Ezechiel	Kaiser Joseph		Saturn		Tieck	Goethe
Adam	Königin Carolina		Periander		Götz von Berlichingen	Schiller
Juda	Kaiser Franz		Thales		Dulcinea	
Prophet Elisa	Papst		Anacharsis		Don Quixote	
Abraham	Herzog von Danzig und Wrede		Chilom		Magdalena	

Table 7.1: Biblical and historical figures in the novella.

Tieck's *Die Gemälde* where there are characters that hide their true identity behind a pseudonym to stay undiscovered in executing some plan.

With the two-mode networks there is a possibility to operationalize the discovery of such hidden persons. First we split each novella into n equally sized parts.[9] Then we mark for each of the 5 most frequent characters (we only care for main characters) the synonym type that is used the most within each segment as well as its frequency. Whenever there is a change of the most frequent type between two segments this is a hint for a hidden character. However, not each change needs to be considered. Instead, we look for those changes that reveal names and titles at the end of the novella or use a pseudonym in the middle part. And indeed, the three characters `Angelique`, `Stephan Murai`, and the prince are among the discovered characters. In Figure 7.13 we highlighted these characters in the two-mode networks and their two different pseudonyms.



(a) Kähler

(b) Stifter

(c) Tieck *Die Gemälde*

Fig. 7.13: The two-mode networks of Kähler, Stifter, and Tieck. Link color indicates synonym type. We chose the colors such that one can observe the two identities, a person pretends to be.

While this approach is only a hint towards hidden characters it cannot assure that an uncovered character was actually hidden. However, we can use the same procedure to

[9] Same number of words. Different values for n where tested but $n = 5$ appears to be a useful choice.

discover something else. If we look at all names that change their synonym type during the end of the novella to `family` we can conclude that the relationship status of that person has changed.[10] In Table 7.3 we see a list of these characters.

novella	**name**	**1**	**2**	**3**	**4**	**5**
Chamisso	Mina	family		family		pseudonym
Fräulein Wolf	Emmy	pseudonym	family		name	name
Gotthelf *Notar*	Notar Stößli		job	job	name	name
Halm	Franz Bauer	pseudonym				name
Hauff	Balthasar Ohnegrund	job				name
Heyden	Eduard	pseudonym	pseudonym		name	
Heyden	Betty			family		name
Hoffmann	Olivier Brusson	pseudonym				name
Holtei	Gustav	name	name	name	name	job
Immermann	Adolphine	family	family		family	name
Kähler	Angelique	family	family	family		name
Kopisch *Träumer*	Eremit		salutation	pseudonym		name
Müller	Arthur Lerchenfels	job	job	job	title	
Raabe	Laurentia Heylingerin	pseudonym				name
Rumohr	Cassandra	family				name
Schefer	Christian	title	pseudonym	title	title	title
Schreyvogel	Max Spohr		pseudonym			name
Schücking	Philipp Wolfskron	job		job	title	name
Stifter	Stephan Murai		title	name	title	title
Tieck *Gemälde*	Prinzen	pseudonym	pseudonym	pseudonym	title	
Wallner	Istvan	salutation	salutation	salutation	pseudonym	salutation

Table 7.2: Possible hidden figures. Black values indicate that there was either no presence of the character or no synonym type dominated the others (was present more than 50% of all character mentions within this section).

novella	**name**	**1**	**2**	**3**	**4**	**5**
Dincklage	Sanne Möhe	pseudonym	name	name	name	family
Gall	Gräfin Agnes	title	title		family	family
Große	Frau Conrectorin	name	name	name	family	
Höfer	Mutter	pseudonym	pseudonym		family	family
Holtei	Muhme Wawerle	name	name		name	family
Holtei	Herr Tiesel	name	name	name	name	family
Lohmann	Rath Ellinger	pseudonym	pseudonym	pseudonym		family
Tieck *überfluss*	Clara	name	name	name	name	family
Wichert	Urte Karalene	pseudonym	pseudonym	pseudonym	pseudonym	family

Table 7.3: Table of characters that change their family relation.

[10] Not limited to marriage but in any sense of relationship including mother-/fatherhood, being a sister in law, etc.

Chapter 8

Discussion and Conclusion

The thesis used network science to investigate a corpus of novellas. We started with a reading experiment to test a hypothesis that comes from literary theory. In the second part, we study the stylometric similarity in the corpus. Finally, we studied networks in novellas and again developed our research questions from the claims of the editors. This led to further classifications of the network types that can occur in 19th century novellas. In this chapter we will list the main contributions of the thesis as well as some limitations. Moreover, we will give ideas on future research.

8.1 Contributions and Limitations

8.1.1 Contributions

While each chapter highlights a different aspect there are a few overarching themes that are covered in each part. In the following, we want to point out three of these aspects. In detail, we emphasize our methodological viewpoint, the use of backbone networks, and the compilation of a new data set.

Methodological viewpoint. All studies were inspired by the specific properties of the *Novellenschatz* itself. The use of a pre-compiled corpus made it possible to derive our research questions from the corpus building process. Here, a focus was on the text genre novella especially when compared to the more common novel genre. In particular, we did not primary form our research questions based on the available methods. Instead, we analyzed the effects of a variety of methods in question to choose the best possible solutions for each research goal. While this is done explicitly in Chapter 3 it is also an overall attitude that is implicitly done in all other chapters as well.[1]

[1] i.e., even if it is not stated everywhere, there were a lot of alternatives considered for each choice of method that did not make it into the thesis.

Novella networks. The study of a variety of text corpora has become increasingly popular in nlp research and digital humanities. While nlp usually deals with present day texts like newspaper articles, social media posts or legal documents, digital humanities research and in particular computational literary studies worked with novels, plays, poems, and other forms of written text. In each case the methods used need to be adapted to the specific object of study. This thesis expanded the catalogue of feasible measures to the novella genre. Network science was thereby a key tool. More concrete, the Simmelian backbone algorithm was used in all projects for various network visualizations of diverse distance measures and proved productive in investigating our data sets and answer our research questions.

New data set. The creation, provision, and description of a new data set is another contribution of the thesis. The novelty of the provided network data is manifold.

First, larger collections of high-quality character network collections are rare. This is due to the fact that quality and quantity usually compete. While other network types like infrastructure networks, citation networks, or chemical networks usually have (more or less) trivial choices for the node and edge building process and are hence easily scalable[2] this is not true for social networks or those of literary figures. For such networks high quality network building processes come with the cost of close investigations of the object of study, i.e., questionnaires for real life social networks and close reading for literary character networks. While advances in NLP research leads to improvements in entity recognition and co-reference resolution the state-of-the-art tools are not yet capable to compete with human reading skills.[3] The semi-automated approach (that automated the link generation and disambiguation resolution but hand-coded the node creation) made it possible to make use of the medium sized corpus to create a data set that is large enough for network ensemble analyses (it forms a reasonably sized network of networks) without trusting of-the-shelf algorithms for the underlying networks which usually come as a black box.

Second, the consideration of two-mode networks instead of directly looking at the resulting one-mode networks is unusual. While being a logical step that has the advantage that we can directly infer the network evolution and dynamics we have not found similar procedures in the literature. Even though, we only made use of these dynamics in very limited ways (e.g., in Chapter 7.1) we could think of further analyses based on the data that are discussed in Chapter 8.2.

Finally, the visualization of the two-mode networks already embedded and displayed a lot of information over the story line. One can use these networks as starting points for automated plot summaries.

[2] Meaning the network (and network ensemble) size is not bounded by practical considerations. However, it might be limited by computational power.

[3] This is especially true for the co-reference resolution task and even more, when a text corpus is used that does not use present-day orthography.

8.1.2 Limitations

There are numerous limitations to the studies and analyses. Here, we will discuss four major points in detail. Namely, we will consider the text and proband selection, possible feature combinations for stylometric differences, the measurement of character relations, and the way we studied the two-mode networks. Besides making aware of the named problems, we will also discuss why we did not directly tackle them in our work.

Text and proband selection. While many of the analyses in this thesis dealt with the whole corpus, the first study focused on a singular work. Due to the nature of the experiment and the feasibility of the reading task we could not choose freely from the complete corpus but needed to concentrate on a small number of possible novellas. Within these novellas a small pretest was made to choose the selected text of Lorm. However, we did not compare our results to texts within or outside of the *Novellenschatz.* While the comparison to a text outside of the *Novellenschatz* can be seen as future research, the comparison with other texts from the present corpus could have helped to select a text where Heyse's claim of a strong silhouette is highly pronounced. However, we decided to see the claim as a selection criterion for inclusion in the *Novellenschatz* such that it should not matter which novella we chose.

Moreover, the call for experiment participants had only a few limitations (age minimum of 18 and German level C1[4]). This encoded several problems. On the one hand, the way the research question was formulated, it was necessary to give the experiment participants the possibility to write their summaries in free text form with only a little interference by the researcher (e.g., by providing a limited amount of lines). On the other hand, this led to a huge range of quality in the answers. With quality we do not mean the closeness to the expected result, but instead the use of language in general. A larger sample of the probands had major problems in text understanding because of the unfamiliarity of 19th century literature. We could possibly argue that the claims made by Heyse were definitely targeted to an experienced audience.

Possible feature combinations. In Chapter 4 we have demonstrated the usefulness of stylometric similarity measures for several examples. While we based our choices on informed decisions which combinations are useful and which are not and considered many other combinations that were not included in the analysis (e.g., because they led to uninterpretable results), we did not consider all possible combinations of features. If we would have done this, it would have multiplied the number of combinations by each decision point. Therefore, this easily would lead into a feasibility problem because of the magnitude of investigated networks. However, it would also suggest spurious accuracy because it might still be that we missed out a feature or as a simple example δ_p can take any $p \in [1, \infty)$, hence there are infinitely many possibilities. Nonetheless, it is quite likely that we missed

[4] Common European Framework of Reference for Languages `https://www.coe.int/en/web/common-european-framework-reference-languages`.

an interesting combination that is worth studying.

Measurement of character relations. We have used a semi-automated procedure to measure character relations. Here, there are several problems with the annotation process of nodes and links. As already described in Chapter 5, character disambiguation cannot be resolved completely (but at least in around 80% of the time). Second, the disregard of pronouns will lose a lot of precision in the correct character allocation.

The largest issue is the correct allocation of the narrator in first-person narratives. In such a novella we associated the word I ("ich") with the narrator. However, whenever the term I shows up in direct speech it is usually referring to the current speaker instead of the narrator. As already discussed in Chapter 4 the documents do not use quotation marks or any other unified form of marking direct speech (possible markers of direct speech reach from starting a new line, using a dash (–), or three points (...) to simply including direct speech in the text without any indication) it would need very elaborate methods to allocate sentences to a speaker. One could think of rule-based systems (regex) combined with methods that study syntax, tone or the use of second person pronouns.

Other accuracy issues that have not yet been discussed in detail include the use of co-occurrence within a paragraph. While this is useful for the waste majority of texts it is questionable when the paragraphs are not used as the smallest thematic unit. On the one hand, Arnim's *Der tolle Invalide auf dem Fort Ratonneau*, Spindler's *Die Engel-Ehe* and Gotthelf's *Kurt von Koppigen* use extremely lengthy paragraphs such that all main characters usually co-appear in each paragraph even if they do not interact. On the other hand, Schücking's *Die Schwester*, Wild's *Eure Wege sind nicht meine Wege*, and Sacher-Masoch's *Don Juan von Kolomea* use new paragraphs as a marker for direct speech. Hence, it might be useful to consider more appropriate units for these texts.

While all these can lead to a more accurate character recognition, there is a much more profound and complicated limitation to any analyses of the present type, i.e., the definition of an interaction in the first place. While we do measure co-occurrence on different scales as a proxy for an interaction, we are not able to identify the underlying process or the way, characters are in fact related. In Chapter 5 we name a few possible relation types that all come with their specific pros and cons.

Sentiment analysis. When investigating the novella networks in Chapter 7.2.3 we observed that there is a larger core of main characters. However, close reading hinted us that these characters actually build fractions. Hence, it would be insightful to use signed networks for further analyses. One could think of different procedures to construct such signed networks. Straightforward approaches include the following. First, one can compute the overall sentiment of a paragraph (e.g., by of-the-shelf sentiment analysis tools[5]) and assign this sentiment score to each of the individual characters present in the paragraph. A second approach would look only on subsets of these like the sentiment of the verb that

[5] `Spacy` has the German sentiment analysis corpus `SentiWS` (Goldhahn et al., 2012) incorporated.

connects two different characters (in the sense of an action sequence described by Franzosi 2008).

Possible applications would have been an analysis of the sentiment dynamics similar to Nalisnick and Baird (2013). Other applications include the use of dynamic models (Traag et al., 2013) to infer whether the novellas reach for social balance in the sense of social balance theory (Heider, 1946, 1958; Cartwright and Harary, 1956).

8.2 Future research

The thesis opens the possibility to various follow-up research questions. As an example, we want to point out four directions that one can pursuit.

Reading experiment. Compare the effect on readers of different texts from within and outside of the *Novellenschatz* and its subsequent corpora.

As an extension that can give further insights to the given research question one could hold the answers given by the participants against answers to the same question, summarizing a different underlying text from that epoch of similar length and style that has *not* make it into the *Novellenschatz*. In this way, it might be possible to spot differences on the summarizability of the given texts. Then, one could claim that one text is a novella by Heyse's theory and therefore rightfully belongs to the *Novellenschatz* while the other is not (even though it claims to be). However, we want to point out that Heyse's criterion might be a necessary but not sufficient condition and hence it does not need to be true that other texts are not summarizable in a similar way.

Text similarity. Our research highlighted different pitfalls and considerations that need to be made when working with text similarity for non-authorship attribution purposes. However, we used text similarity as a synonym for stylometric similarity. In reality, there are many more levels on which a text could be similar. Other possible directions to define similarity on a text bases are syntactic similarity and semantic similarity. State-of-the-art architectures like google's BERT (Devlin and Chang, 2018; Devlin et al., 2019) can compute similarities that take into account relationships between the words of a sentence.

Relationship types in fictional networks. In Chapter 7 we have introduced a first use case for the *types* attribute. However, one could think of a bundle of other studies that make use of it. First, one can observe the change of narrative perspective. In Figure 8.1 we have highlighted the synonyms I ("ich") and "Gustav" in brown and turquoise for the main character (other synonyms are in dark gray). In the last part of the novella (from paragraph 183 to 217) the name is suddenly the predominant synonym. Before, it was a classical first-person narrative.

Second, it might be valuable to look at the number of synonyms used as a text immanent feature. Like the type/token ratio (TEMPLIN, 1957) to measure lexical diversity, we could

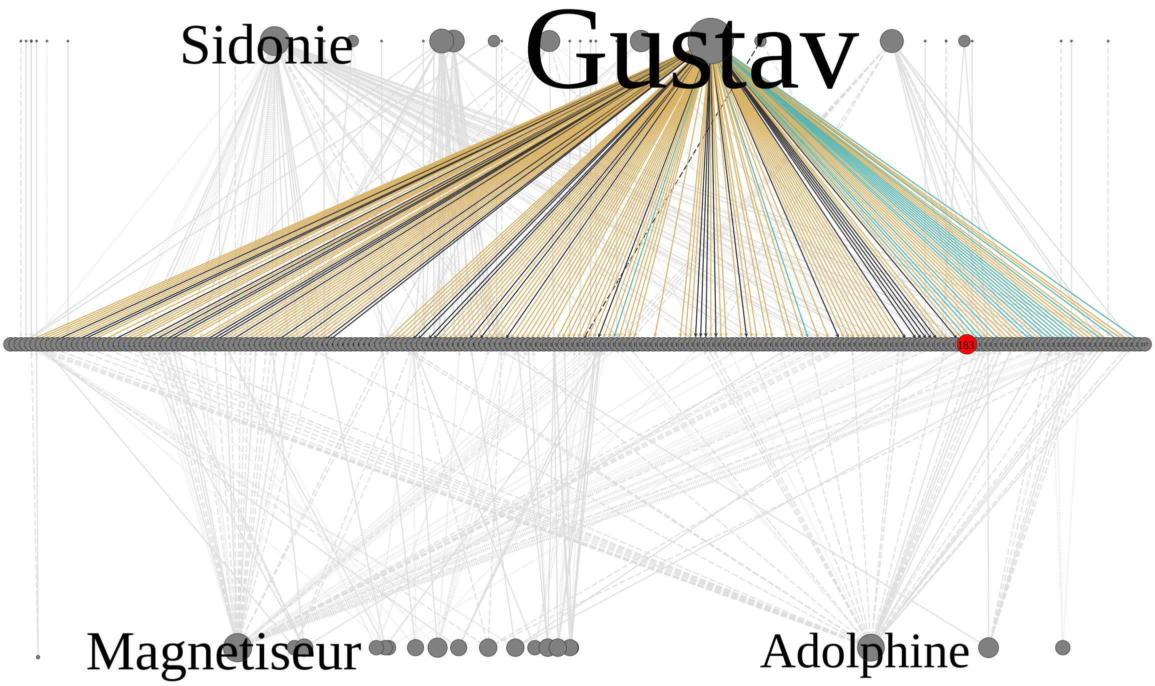


Fig. 8.1: Two-mode network of Immermann's *Der Carneval und die Somnambule.* The links of the main character Gustav are colored w.r.t. the synonyms.

use the synonyms/character mentions ratio (SCMR) to narrow down this diversity to the way authors write about their characters. This could be computed as

$$\mathrm{SCMR}(D) = \frac{\#ofsynonymsinnovella}{\#\text{ number of character mentions excluding I (\texttt{ich})}}.$$

As a possible application one could ask whether a richer use of synonyms relates to a higher degree of canonization. Brottrager et al. (forthcoming) developed a canonization score to measure the valuation by contemporary and nowadays literary scholars. If we consider a subsample of the *Novellenschatz*[6] we can observe a weak correlation between the two values (Figure 8.2. It might be of interest to study these connections in detail.

A third application is the use of the data as a valuable training set for co-reference resolution tasks[7] in machine learning. While not being the largest data set it at least consists of over 8300 synonyms that must be classified to more than 2800 characters. To do so, there are 41844 character mentions in the text.

[6] The subset consisted of 19 novellas that were also part of Brottrager et al.'s corpus.

[7] A good overview can be found in Sukthanker et al. (2020)

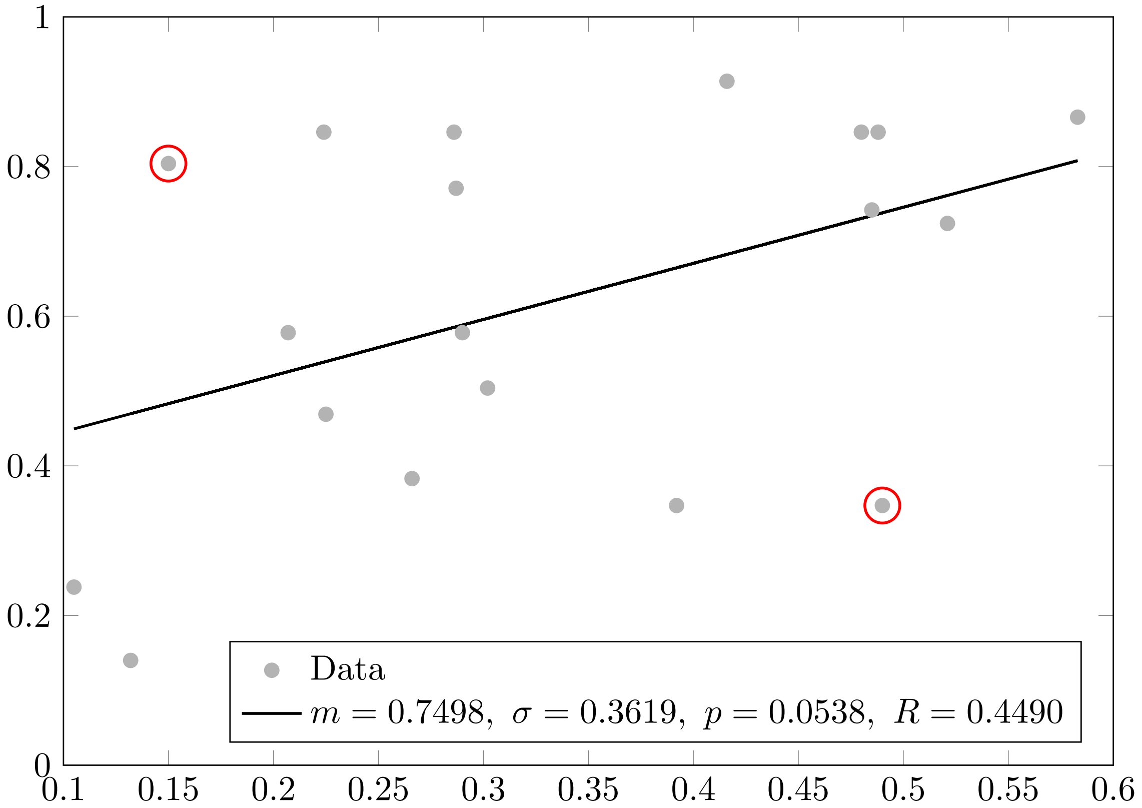


Fig. 8.2: Plot of canonization scores vs SCMR. Two outliers (red circles) are Tieck's *Des Lebens Überfluß* with a high level of canonization but relatively low number of synonyms and the relatively unknown novella *Der Drache* by Kürnberger with a high number of synonyms. Without these outliers the correlation is much stronger ($m = 1.1243,\ \sigma = 0.3357,\ R = 0.6541,\ p = 0.0044$). Also Tieck's second novella in the *Novellenschatz* has a low SCMR score (0.2130) such that it might be worthy to consider this a stylistic choice.

Comparison to novels. In Chapter 6 we have analyzed the inner structure of the network ensemble. However, to make a more convincing argument that the novellas indeed do have similar character constellation networks one could compare them to other texts from the same time. However, this again would be a lot of work because for a fair comparison, the character networks from the contrasting corpus need to be prepared with the same granularity. Instead, in an earlier stage of the work,[8] we automatically extracted the characters from the text.[9] In this case the network creation process could be replicated for different corpora and hence, we compared the documents from the *Novellenschatz* with those of the *Neuer Deutscher Novellenschatz* as well as a corpus of German language novels from 1655 to 1881.

While we lose precision when using automated methods we will consider static networks of quite long texts (when compared to most of the novellas from the *Novellenschatz*). Thus, even if we miss a bunch of links, we can assume that most of the connections in those texts will be captured eventually such that the network as a whole is still representative for the text and as a result comparable to our novella networks.

In Figure 8.3 we can indeed observe, that the novellas from the *Novellenschatz* are densely connected while the novels are further spread around the space. Moreover, one can observe that the novellas from the subsequent collection *Neuer Deutscher Novellenschatz* are between the two corpora. They are not as far spread then novels, but they do not hold as strong to the orignial paradigma of a highly similar character network.

Event models. The detailed two-mode networks make it possible to assign each character appearence a "point in time". Therefore, we can interpret the networks as dynamic event networks. Different approaches on modelling such networks are given by Butts (2008), Stadtfeld (2012), and Stadtfeld and Block (2017). Among others, one could investigate the following research questions. Can we use dynamic event models to show that there are family ties build in certain networks? Is it the case that some link attributes are more likely than others?

[8] That has been presented at the EADH Conference 2021 and is available via `https://eadh2021.culintec.de/P_PCKE_Simon_Character_Networks_in_a_Collection_of_19th_Cent.html`

[9] The networks are derived from the texts by a manually improved named entity recognition. More concretely, we take into account all entities that are marked as persons by either `spacy`'s german model parser[10] or Akers part-of-speech tagger (available via `http://staffwww.dcs.shef.ac.uk/people/A.Aker/activityNLPProjects.html`) or are present in a list of German noble titles and remove all names that only occur once in the text. We then manually remove false positives and do a simple matching for same names with wrong lemmatization (e.g., *Rosalie* vs. *Rosalien*).

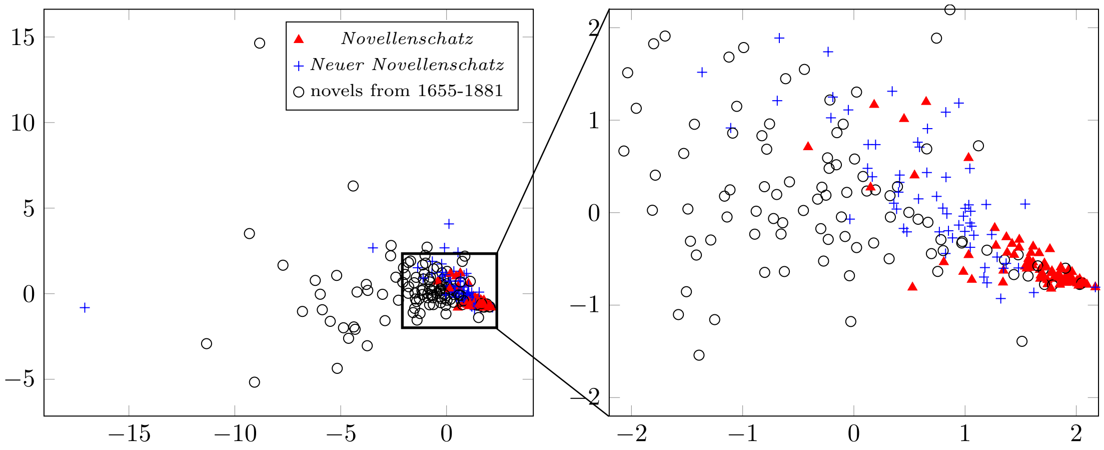


Fig. 8.3: MDS plot for network distances based on spectral transformation costs in a comparison with novels. The right side is an magnification of the densest area.

8.3 Conclusion

If we ask people what their definition of a novella is, they usually believe it is a relatively short prose text. While there have been many definitions given by literary scholars over the last centuries, none of these could ever be sharp. Instead, each meaningful definition leaves space for interpretation such that the assignment of a work to the novella genre needs to be done either by the author themself or by recipients of the text.

This thesis took the process one step further. It's starting point was a corpus of texts that have been edited by well-known literary scholars and authors of the 19th century. By naming their collection *Deutscher Novellenschatz* (and motivating their approach in a larger preface) Heyse and Kurz made clear that each document in the compilation should match *their* criterion of a novella. With the use of tools from the digital humanities and network science the thesis studied these novellas. The methods made it possible to observe the effect of the editor's definition to the scope and boundaries of the readers perception, the texts stylometric variety, and the novellas character networks. Moreover, the thesis showed that one of the reasons behind the relatively large variety within these categories could have resulted from the interpretation of the underlying criterion itself. By conducting an experiment to test a literary theory one observed that cohesiveness in summarization of a novella cannot be achieved on a global scale.

Finally, we want to point out one more detail. Starting with Moretti's book *Graphs, Maps, Trees* (Moretti and Piazza, 2007) the study of larger text corpora has proven to lead to fruitful results in the digital humanities research. However, right from the beginning there was the danger and criticism that the existing large amount of arbitrariness in literary corpus studies should not be underestimated. By narrowing our research to a specific corpus of a specific and less studied text sort (the novella) this thesis aimed for a better understanding of the tools at hand. Here, a good contextual knowledge is necessary for both, the algorithms and parameterizations used (digital) and the textual basis they are applied on. Only then, conclusions drawn from computational literary studies can persist.

Bibliography

Alberich, R., Miro-Julia, J., and Rosselló, F. (2002). Marvel Universe looks almost like a real social network. *arXiv*.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, 23(2):131–147.

Baeza-Yates, R. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval: the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England, Addison-Wesley.

Barthes, R. (1971). Action Sequences. In Joseph Strelka, editor, *Patterns of Literary Style*, pages 5–14. Pennsylvania State University Press.

Batagelj, V. and Zaveršnik, M. (2003). An O(m) Algorithm for Cores Decomposition of Networks. *CoRR*, cs.DS/0310049.

Bennett, G. R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* University of Michigan Press, Ann Arbor.

Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping multidimensional data: Recent advances in clustering*, pages 25–71.

Berry, D. M. (2017). *Digital humanities: knowledge and critique in a digital age.* Polity Press, Cambridge, England.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge Approaches to Linguistics. Cambridge University Press, Cambridge.

Bossaert, G. and Meidert, N. (2013). "We Are Only as Strong as We Are United, as Weak as We Are Divided" a Dynamic Analysis of the Peer Support Networks in the Harry Potter Books. *Open Journal of Applied Sciences*, 3(2):174–185.

Brandes, U., Robins, G., McCranie, A., and Wasserman, S. (2013). What is network science? *Network Science*, 1(1):1–15.

Bunke, H. (2000). Graph matching: Theoretical foundations, algorithms, and applications. In *In Proceedings of Vision Interface 2000, Montreal*, pages 82–88.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, 22(1):27–47.

Busa, R. (1973). *Index Thomisticus.* Frommann-Holzboog.

Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, 14(2):83–90.

Büttner, A., Dimpel, F. M., Evert, S., Jannidis, F., Pielström, S., Proisl, T., Reger, I., Schöch, C., and Vitt, T. (2017). "Delta" in der stilometrischen Autorschaftsattribution. *Zeitschrift für digitale Geisteswissenschaften.*

Butts, C. T. (2008). A Relational Event Framework for Social Action. *Sociological Methodology*, 3:155–200.

Cartwright, D. and Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277.

Chakraborty, S., Muhuri, S., and Das, D. (2021). Detection of constant member and overlapping community from dynamic literary network. *Social Network Analysis and Mining*, 11(1):77.

Chouquer, G. (2002). Alain Guerreau, L'avenir d'un passé incertain. Quelle histoire du Moyen Âge au XXIe siècle ? Paris, Le Seuil, 2001. *Études rurales*, 161-162. Online since 17.06.2003, Accessed: 09.02.2022. `http://journals.openedition.org/etudesrurales/106`.

Chung, F., Lu, L., and Vu, V. (2003). Eigenvalues of Random Power law Graphs. *Annals of Combinatorics*, 7(1):21–33.

Coll Adanay, M. and Sporleder, C. (2015). Clustering of Novels Represented as Social Networks. In *Linguistic Issues in Language Technology, Volume 12, 2015 - Literature Lifts up Computational Linguistics*. CSLI Publications.

Coll Ardanuy, M. and Sporleder, C. (2014). Structure-based Clustering of Novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 31–39, Gothenburg, Sweden. Association for Computational Linguistics.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

Deshpande, M., Kuramochi, M., Wale, N., and Karypis, G. (2005). Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering*, 17(8):1036–1050.

Devlin, J. and Chang, M.-W. (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. `http://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html`. Accessed: 27.06.2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 19(1):109–123.

Digital Humanities at Berkeley (2016). Digital humanities journals | Digital Humanities. `https://digitalhumanities.berkeley.edu/resources/digital-humanities-journals`. Accessed: 19.01.2022.

Dressen, A. (2022). Research Guides: Rinascimento: Digital Humanities journals + blogs. `https://guides.library.harvard.edu/c.php?g=310256&p=2071428`. Accessed: 19.01.2022.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1).

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644, Avignon, France. Association for Computational Linguistics.

Elson, D. K., Dames, N., and McKeown, K. R. (2010). Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 138–147, USA. Association for Computational Linguistics.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD '96, page 226–231. AAAI Press.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, 4(1):65–75.

Evert, S., Proisl, T., Vitt, T., Schöch, C., Jannidis, F., and Pielström, S. (2015). Towards a better understanding of burrows's delta in literary authorship attribution. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 79–88, Denver, Colorado, USA. Association for Computational Linguistics.

Fan, X., Wang, J., Pu, X., Zhou, L., and Lv, B. (2011). On graph-based name disambiguation. *J. Data and Information Quality*, 2(2).

Faust, K. and Skvoretz, J. (2002). 8. Comparing Networks across Space and Time, Size and Species. *Sociological Methodology*, 32(1):267–299.

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., and Trilcke, P. (2017). Network dynamics, plot analysis: Approaching the progressive structuration of literary texts. In *Digital Humanities 2017 (Montréal, 8–11 August 2017). Book of Abstracts.* McGill University.

Franzosi, R. (2008). Content analysis: Objective, systematic, and quantitative description of content. *Content analysis*, 1:26.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Freeman, L. C., Romney, A. K., and Freeman, S. C. (1987). Cognitive Structure and Informant Accuracy. *American Anthropologist*, 89(2):310–325.

Gil, S., Kuenzel, L., and Suen, C. (2011). Extraction and Analysis of Character Interaction Networks From Plays and Movies. Technical Report, Stanford University.

Gold, M. K. (2012). *Debates in the Digital Humanities.* University of Minnesota Press.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Grener, A., Luczak-Roesch, M., Fenton, E., and Goldfinch, T. (2017). Towards a Computational Literary Science: A Computational Approach to Dickens' Dynamic Character Networks. (v1.0). Zenodo.

Gries, S. T. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, 3(5):1225–1241.

Hackler, R. M. and Kirsten, G. (2016). Distant Reading, Computational Criticism, and Social Critique: an Interview with Franco Moretti. `https://doi.org/10.13095/uzh.fsw.fb.144`.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '04, pages 296–305, New York, NY, USA. Association for Computing Machinery.

He, H., Barbosa, D., and Kondrak, G. (2013). Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Wiley, New York.

Hettinger, L., Becker, M., Reger, I., Jannidis, F., and Hotho, A. (2015). Genre Classification on German Novels. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 249–253.

Heyse, P. and Kurz, H. (1871). Einleitung. In: Deutscher Novellenschatz. Hrsg. von Paul Heyse und Hermann Kurz. Bd. 1. München. In Weitin, T., editor, *Volldigitalisiertes Korpus. Der Deutsche Novellenschatz. Darmstadt/Konstanz, 2016*, Deutsches Textarchiv, page V–XXIV. `https://www.deutschestextarchiv.de/heysekurz_einleitung_1871`. Accessed: 02.03.2022.

Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Hodel, T. (2013). Das kleine Digitale: Ein Plädoyer für Kleinkorpora und gegen Grossprojekte wie Googles Ngram-Viewer. *Nach Feierabend. Zürcher Jahrbuch für Wissensgeschichte*, 9:103–119.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Huang, J., Ertekin, S., and Giles, C. L. (2006). Efficient Name Disambiguation for Large-Scale Databases. In Fürnkranz, J., Scheffer, T., and Spiliopoulou, M., editors, *Knowledge Discovery in Databases: PKDD 2006*, Lecture Notes in Computer Science, pages 536–544, Berlin, Heidelberg. Springer.

Hunziker, H. W. (2006). *Im Auge des Lesers : vom Buchstabieren zur Lesefreude : foveale und periphere Wahrnehmung*. Transmedia, Zürich, originalausg. edition.

Hutchinson, S., Datla, V., and Louwerse, M. (2012). Social Networks are Encoded in Language. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Jannidis, F. (2017). Perspektiven quantitativer Untersuchungen des Novellenschatzes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 47(1):7–27.

Jannidis, F., Kohle, H., and Rehbein, M. (2017). *Digital Humanities: Eine Einführung.* J.B. Metzler, Stuttgart.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In Erlin, M. and Tatlock, L., editors, *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, Studies in German Literature Linguistics and Culture, page 29–54. Camden House, Rochester.

Jannidis, F., Pielström, S., Schöch, C., and Vitt, T. (2015). Improving burrows' delta. an empirical evaluation of text distance measures. In *Digital Humanities Conference*, volume 11.

Jannidis, F., Reger, I., Krug, M., Weimer, L., Macharowsky, L., and Puppe, F. (2016). Comparison of methods for the identification of main characters in german novels. In Eder, M. and Rybicki, J., editors, *Digital Humanities 2016, DH 2016, Conference Abstracts, Jagiellonian University & Pedagogical University, Krakow, Poland, July 11-16, 2016*, pages 578–582. Alliance of Digital Humanities Organizations (ADHO).

Jeziorkowski, K. (1987). *Eine Iphigenie rauchend: Aufsätze und Feuilletons zur deutschen Tradition.* Number 1365 in Suhrkamp Taschenbuch. Suhrkamp, Frankfurt am Main.

Karp, R. M. (1980). An algorithm to solve the m × n assignment problem in expected time O(mn log n). *Networks*, 10(2):143–152.

Karsdorp, F., van Kranenburg, P., Meder, T., and van den Bosch, A. (2012). Casting a Spell: Identification and Ranking of Actors in Folktales. In Mambrini, F., Passarotti, M., and Sporleder, C., editors, *Proceedings of the Second Workshop on Annotation of Corpora for Research in het Humanities (ACRH-2)*, pages 39–50. Edicoes Cilibri, Lisbon.

Khabsa, M., Treeratpituk, P., and Giles, C. L. (2015). Online Person Name Disambiguation with Constraints. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, pages 37–46, New York, NY, USA. Association for Computing Machinery.

Kjell, B. (1994). Authorship attribution of text samples using neural networks and Bayesian classifiers. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1660–1664.

Krug, M., Reger, I., Jannidis, F., Weimer, L., Madarász, N., and Puppe, F. (2017). Overcoming Data Sparsity for Relation Detection in German Novels. In *ADHO 2017-Montréal.*

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Kucera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Kuebler, S. and Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic; Annotated edition, London; New York.

Kwon, H.-C. and Shim, K.-H. (2017). An Improved Method of Character Network Analysis for Literary Criticism: A Case Study of <Hamlet>. *International Journal of Contents*, 13(3):43–48.

Kydros, D. and Anastasiadis, A. (2015). Social network analysis in literature. The case of The Great Eastern by A. Embirikos. In *Proceedings of the 5th European Congress of Modern Greek Studies of the European Society of Modern Greek Studies*, volume 4, pages 681–702.

Labatut, V. and Bost, X. (2020). Extraction and Analysis of Fictional Character Networks: A Survey. *ACM Computing Surveys*, 52(5):1–40.

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., and Demidov, P. (2019). A Survey on Stylometric Text Features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195, Helsinki, Finland. IEEE.

Le Deuff, O. (2018). *Digital Humanities: History and Development*, volume 4 of *Intellectual technologies set*. ISTE, London UK.

Lee, J. and Yeung, C. Y. (2012). Extracting Networks of People and Places from Literary Texts. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 209–218, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Leech, G., Hundt, M., Mair, C., and Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Studies in English Language. Cambridge University Press.

Levenshtein, V. I. (1965a). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848. Original in Russian – translation in Soviet Physics Doklady 10(8):707-710, 1966.

Levenshtein, V. I. (1965b). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17.

Liu, A. (2016). Is Digital Humanities a Field? An Answer from the Point of View of Language. *Journal of Siberian Federal University. Humanities & Social Sciences*, 9(7):1546–1552.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Luhmann, N. (1964). *Funktionen und Folgen formaler Organisation*, volume 20 of *Schriftenreihe der Hochschule Speyer*. Duncker & Humblot, Berlin.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 281–298, Los Angeles, LA, USA. University of California Press.

Malinowski, B. (1922). *Argonauts of the Western Pacific: an account of native enterprise and adventure in the Archipelagoes of Melanesian New Guinea.* G. Routledge & Sons; E.P. Dutton., London.

Mann, G. S. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 33–40, USA. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*, chapter Scoring, term weighting, and the vector space model, pages 100–123. Cambridge University Press.

Maverick, G. V. (1969). Computational Analysis of Present-Day American English. Henry Kučera , W. Nelson Francis. *International Journal of American Linguistics*, 35(1):71–75.

McCarty, W. (2003). Humanities computing. *Encyclopedia of library and information science*, 2:1224–35.

Mellmann, K. (2017). Zeit- und Ortsangaben im Deutschen Novellenschatz. *Zeitschrift für Literaturwissenschaft und Linguistik*, 47(1):49–66.

Min, S. and Park, J. (2016). Network Science and Narratives: Basic Model and Application to Victor Hugo's Les Misérables. In Cherifi, H., Gonçalves, B., Menezes, R., and Sinatra, R., editors, *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016*, Studies in Computational Intelligence, pages 257–265. Springer International Publishing, Cham.

Moreno, J. L. (1934). *Who shall survive?: Foundations of sociometry, group psychotherapy, and sociodrama.* Nervous and Mental Disease Publishing Co., Washington, D.C.

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1:54–67.

Moretti, F. (2011). Network theory, plot analysis. *New Left Review*, 68:80–102.

Moretti, F. (2013). *Distant Reading.* Verso, London.

Moretti, F. and Piazza, A. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History.* Verso Books.

Mueller, M. (2012). Scalable Reading. `https://scalablereading.northwestern.edu/?page_id=22`.

Müller, M. (2013). Morgenstern's Spectacles or the Importance of Not-Reading. `https://scalablereading.northwestern.edu/2013/01/21/morgensterns-spectacles-or-the-importance-of-not-reading/`. Accessed: 25.01.2022.

Nagel, U. (2011). *Analysis of Network Ensembles.* PhD thesis, Konstanz University, Konstanz.

Nalisnick, E. T. and Baird, H. S. (2013). Character-to-Character Sentiment Analysis in Shakespeare's Plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, Sofia, Bulgaria. Association for Computational Linguistics.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6):1–36.

Nick, B., Lee, C., Cunningham, P., and Brandes, U. (2013). Simmelian backbones: Amplifying hidden homophily in facebook networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 525—532, New York, NY, USA. Association for Computing Machinery.

Park, G.-M., Kim, S.-H., Hwang, H.-R., and Cho, H.-G. (2013). Complex System Analysis of Social Networks Extracted from Literary Fictions. *International Journal of Machine Learning and Computing*, 3(1):107–111.

Paulmier-Foucart, M. (1980). Jean Hautfuney, Tabula super Speculum historiale fratris Vincentii - A-L. *Spicae, Cahiers de l'Atelier Vincent de Beauvais, Nouvelle série*, 2:p. 19–263.

Pearson FRS, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., and Chanona-Hernández, L. (2017). Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3):627–639.

Prado, S. D., Dahmen, S. R., Bazzan, A. L. C., Carron, P. M., and Kenna, R. (2016). Temporal network analysis of literary texts. *Advances in Complex Systems*, 19:1–19.

Rieck, B. and Leitte, H. (2016). 'Shall I compare thee to a network?': Visualizing the Topological Structure of Shakespeare's Plays. In *Proceedings of the 1st Workshop on Visualization for the Digital Humanities.*

Rochat, Y. (2014). *Character networks and centrality.* PhD thesis, Université de Lausanne, Faculté des sciences sociales et politiques.

Rochat, Y. and Kaplan, F. (2014). Analyse des réseaux de personnages dans Les Confessions de Jean-Jacques Rousseau. *Les Cahiers du numérique*, 10(3):109–133.

Rochat, Y. and Triclot, M. (2017). Les réseaux de personnages de science-fiction : échantillons de lectures intermédiaires. *ReS Futurae. Revue d'études sur la science-fiction*, 10.

Sack, G. (2012). Character networks for narrative generation. In *Intelligent Narrative Technologies: Papers from the 2012 AIIDE Workshop, AAAI Technical Report WS-12-14*, pages 38–43.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Sapkota, U., Solorio, T., Montes-y Gómez, M., and Rosso, P. (2013). The Use of Orthogonal Similarity Relations in the Prediction of Authorship. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 463–475, Berlin, Heidelberg. Springer.

Scheinfeldt, T. (2014). The Dividends of Difference: Recognizing Digital Humanities' Diverse Family Tree/s. *Found History.* `https://foundhistory.org/2014/04/the-dividends-of-difference-recognizing-digital-humanities-diverse-family-trees/`.

Schnapp, J., Presner, T., Lunenfeld, P., and Drucker, J. (2008). The Digital Humanities Manifesto 2. *Humanities Blast Engaged Digital Humanities Scholarship.* `humanitiesblast.com/manifesto/Manifesto_V2.pdf.` . Accessed: 03.04.2022.

Schöch, C. (2014). Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In Schöch, C. and Schneide, L., editors, *Literaturwissenschaft im digitalen Medienwandel*, Beihefte zu Philologie im Netz, 7, pages 130–157. PhiN.

Schöch, C., Schlör, D., Zehe, A., Gebhard, H., Becker, M., and Hotho, A. (2018). Burrows' zeta: Exploring and evaluating variants and parameters. In Palau, J. G. and Russell, I. G., editors, *Digital Humanities*, pages 274–277.

Schreibman, S., Siemens, R. G., and Unsworth, J. (2004). *A companion to digital humanities.* Blackwell Pub., Malden, MA.

Schulz, K. (2011). What is Distant Reading? *The New York Times*. `https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html`. Accessed: 01.06.2022.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., and Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504.

Smith, P. W. H. and Aldridge, W. (2011). Improving authorship attribution: optimizing Burrows' Delta method. *Journal of Quantitative Linguistics*, 18(1):63–88.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438.

Srivastava, S., Chaturvedi, S., and Mitchell, T. (2016). Inferring interpersonal relations in narrative summaries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2807–2813. AAAI Press.

Stadtfeld, C. (2012). *Events in social networks: a stochastic actor-oriented framework for dynamic event processes in social networks.* PhD Thesis, KIT Scientific Publishing.

Stadtfeld, C. and Block, P. (2017). Interactions, Actors, and Time: Dynamic Network Actor Models for Relational Events. *Sociological Science*, 4:318–352.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Storm, T. (1920). Eine zurückgezogene Vorrede aus dem Jahre 1881. In Köstner, A., editor, *Theodor Storm, Sämtliche Werke*, volume 8, pages 122–23, Leipzig.

Sukthanker, R., Poria, S., Cambria, E., and Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

TEMPLIN, M. C. (1957). *Certain Language Skills in Children: Their Development and Interrelationships*, volume 26. University of Minnesota Press, Minneapolis, MN, US, ned - new edition edition.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419.

Torgerson, W. S. (1958). *Theory and methods of scaling.* Theory and methods of scaling. Wiley, Oxford, England.

Traag, V. A., Van Dooren, P., and De Leenheer, P. (2013). Dynamical models explaining social balance and evolution of cooperation. *PLOS ONE*, 8(4):1–7.

Trilcke, P., Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., Meiners, H.-L., and Skorinkin, D. (2013). dlina - Digitally-Driven Literary Network Analyis (of Dramatic Texts). `https://dlina.github.io/`. Accessed: 01.03.2022.

Trovati, M. and Brady, J. (2014). Towards an Automated Approach to Extract and Compare Fictional Networks: An Initial Evaluation. In *2014 25th International Workshop on Database and Expert Systems Applications*, pages 246–250.

Trunz, E. (1960). *Goethes Werke*, volume VI. Wegner Verlag, Hamburg.

Vala, H., Jurgens, D., Piper, A., and Ruths, D. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Vanhoutte, E. (2015). The Journal is dead, long live The Journal! `https://academic.oup.com/dsh/pages/dsh_name_change`. Accessed: 10.01.2022.

Volker, B. and Smeets, R. (2020). Imagined social structures: Mirrors or alternatives? A comparison between networks of characters in contemporary Dutch literature and networks of the population in the Netherlands. *Poetics*, 79:101379.

Wallis, S. and Nelson, G. (2001). Knowledge Discovery in Grammatically Analysed Corpora. *Data Min. Knowl. Discov.*, 5:305–335.

Wallis, S. A. (2007). Annotation, Retrieval and Experimentation. *Studies in Variation, Contacts and Change in English*, 1.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Waumans, M. C., Nicodème, T., and Bersini, H. (2015). Topology analysis of social networks extracted from literature. *PLOS ONE*, 10(6):1–30.

Weitin, T. (2016). Selektion und Distinktion. Paul Heyses und Hermann Kurz' Deutscher Novellenschatz als Archiv, Literaturgeschichte und Korpus. *Archiv/Fiktionen. Verfahren des Archivierens in Literatur und Kultur des langen 19. Jahrhunderts*, pages 385–408.

Weitin, T. (2017). Scalable Reading. *Zeitschrift für Literaturwissenschaft und Linguistik*, 47(1):1–6.

Weitin, T. (2019). Burrows's Delta und Z-Score-Differenz im Netzwerkvergleich. Analysen zum Deutschen Novellenschatz von Paul Heyse und Hermann Kurz (1871-1876). In Jannidis, F., editor, *Digitale Literaturwissenschaft. Beiträge des DFG-Symposiums 2017*. J.B. Metzler, Stuttgart.

Weitin, T. and Herget, K. (2017). Falkentopics: über einige Probleme beim Topic Modeling literarischer Texte. *Zeitschrift für Literaturwissenschaft und Linguistik*, 47(1):29–48.

Wilkens, M. (2012). Canons, Close Reading, and the Evolution of Method. In Gold, M. K., editor, *Debates in the Digital Humanities*, pages 249–258. University of Minnesota Press.

Wilpert, G. v. (1959). *Sachwörterbuch der Literatur*, volume 231 of *Kröners Taschenausgabe*. Kröner, Stuttgart.

Winter, T. (1999). Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance. *Faculty Publications, Classics and Religious Studies Department*, 70.

Zadrozny, W. and Jensen, K. (1991). Semantics of Paragraphs. *Computational Linguistics*, 17(2):171–210.

Zadrozny, W. and Jensen, K. (1993). The Paragraph as a Semantic Unit. In Jensen, K., Heidorn, G. E., and Richardson, S. D., editors, *Natural Language Processing: The PLNLP Approach*, volume 196 of *The Kluwer International Series in Engineering and Computer Science*, pages 285–301. Springer US, Boston, MA.

Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., and Liu, Q. (2003a). HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, volume 17 of *SIGHAN '03*, pages 184–187, USA. Association for Computational Linguistics.

Zhang, J. Y., Black, A. W., and Sproat, R. (2003b). Identifying speakers in children's stories for speech synthesis. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*. ISCA.

Zhao, Y. and Zobel, J. (2005). Effective and Scalable Authorship Attribution Using Function Words. In Lee, G. G., Yamada, A., Meng, H., and Myaeng, S. H., editors, *Information Retrieval Technology*, Lecture Notes in Computer Science, pages 174–189, Berlin, Heidelberg. Springer.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort.* Addison–Wesley, Cambridge, MA.

Zipf, G. K. (1935). *The Psychobiology of Language.* Houghton-Mifflin, New York, NY, USA.

Appendices

A1 Silhouette Experiment

List of out-of-vocabulary words in spacy:

> Börte, Dorfmensch, Dorfwirtin, Freiherrin, Galeriedirektor, Gedankenwechsel, Geisteskrank, Geschichterzähler, Glücklichkeit, Heimattal, Kankheit, Kinderheit, Kunstexperte, Kunstgespräch, Kunstkurator, Kunstliebhabers, Kunsttechniken, Lebensglücks, Liebestragödie, Lückenhafte, Millionärssohnes, Naturliebe, Reisegeschichte, Unterzieht, Unverkäuflichkeit, Verstieß, Wachstumserfahrung, Weihnachts-Ausflug, Weihnachtsausflug, Wirtsfrau, Zwangsverheirateten, reingebeten, tränig, umhüpfen, vorgeworden, zwangsverheiraten

word	**replacement**	**word**	**replacement**
Adligen	Adlige	Weinachten	Weihnachten
Alters	Alter	Weihnachtsausflugs	Weihnachtsausflug
Barons	Baron	Weihnachtstage	Weihnachtstag
Besuchs	Besuch	begegnet	begegnen
Bildern	Bilder	begegnete	begegnet
Bildes	Bilder	gekommen	bekommen
Bräutigams	Bräutigam	gemerkt	bemerkt
hörte	Börte	besuchte	besucht
Dingen	Dinge	betrachtet	betrachten
Dorfbewohnern	Dorfbewohner	endete	endet
Drängens	Drängen	erinnert	erinnere
Familien	Familie	verstehen	erstehen
Freirau	Freifrau	erzählte	erzählt
Freiherrn	Freiherrin	findet	finden
Fräuleins	Fräulein	folgte	folgt
Gefühlen	Gefühle	hängt	fängt
geliebten	Geliebten	fühlt	fühle
Gemäldes	Gemälde	gehen	geben
Gerüchten	Gerüchte	gelangen	gegangen
Geschichten	Geschichte	sehen	gehen
Gespräche	Gespräch	gelingen	gelangen
Gesprächen	Gespräche	handelte	handelt
Hauses	Hause	heiraten	heirate
Hintergedanken	Hintergedanke	herausstellte	herausstellt
Ich-Erzählers	Ich-Erzähler	können	könne
Jahren	Jahre	nennen	kennen
wahren	Jahren	lernte	lernt
kommt	Kommt	lieben	liebe
Kunstexperten	Kunstexperte	liebt	liebe
Kunstwerke	Kunstwerk	machte	macht
Kunstwerks	Kunstwerke	möchte	machte
Kunstwerken	Kunstwerke	passierte	passiert
Künstlers	Künstler	platzte	platzt
Lebens	Leben	steht	sieht
geben	Leben	zieht	sieht
liebe	Liebe	sollen	solle
Mannes	Manne	sollte	solle
Mädchens	Mädchen	wollen	sollen
Reichtums	Reichtum	wollte	sollte
solle	Rolle	stellte	stellt
machen	Sachen	verheiratet	verheiraten
Schicksale	Schicksal	verlieren	verlieben
Schicksals	Schicksale	verliebte	verliebt
Schluss	Schloss	verliert	verliebt
Sohnes	Sohne	verloren	verloben
Vaters	Vater	verwoben	verloben
erhalten	Verhalten	versprochen	versprechen
verliebt	Verliebt	versuchte	versucht
Verwandten	Verwandte	willigte	willigt
Vorurteilen	Vorurteile	wohnte	wohnt
Weihnachtens	Weihnachten		

Table A.1: Words with Levenshtein distance > 0.6. The left word has been changed into the right one possibly multiple times (as in the case of Gesprächen → Gespräche → Gespräch).

A2 Text similarity

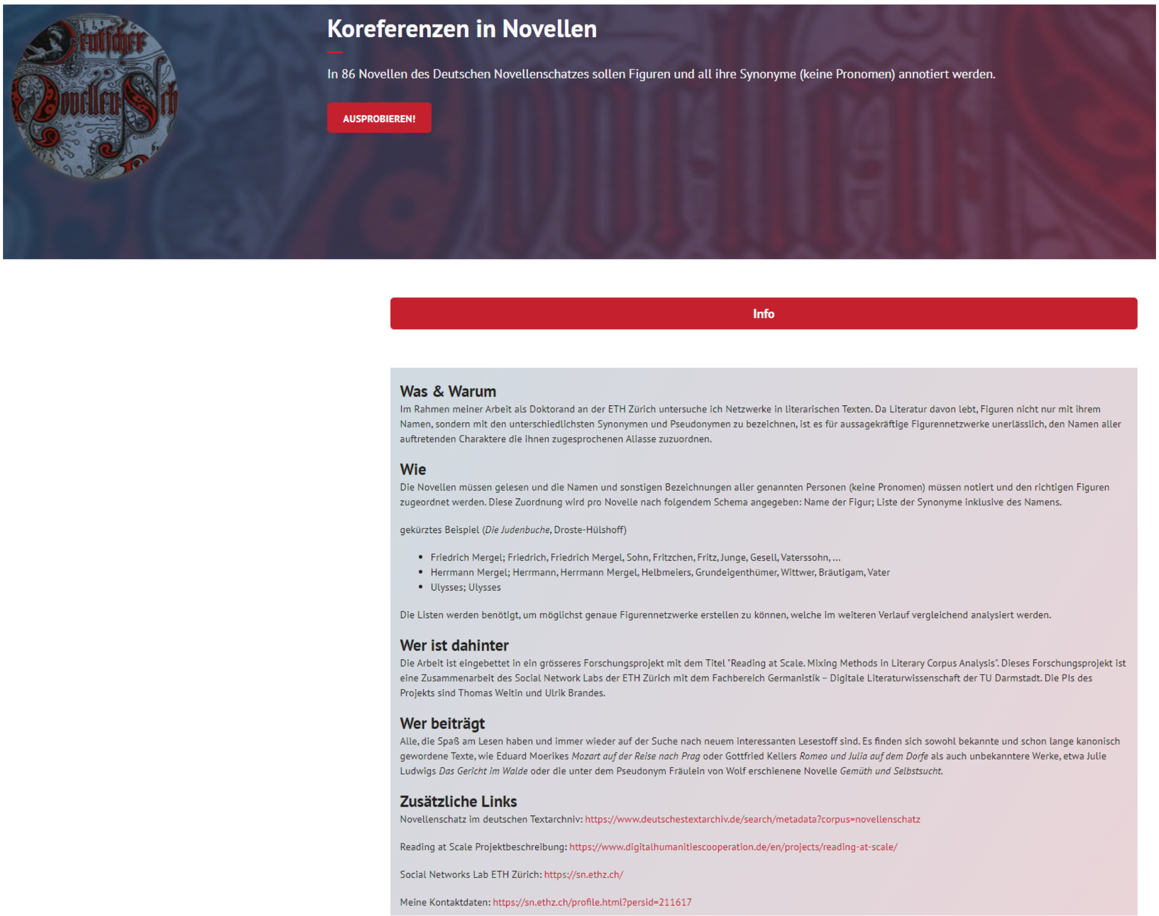
author	**title**	**first-person narrative**	**gender**	***Adelsnovelle***
Alexis, W.	Herr von Sacken		m	x
Andolt, E.	Eine Nacht	x	m	x
Arnim, A. v.	Der tolle Invalide auf dem Fort Ratonneau		m	
Auerbach, B.	Die Geschichte des Diethelm von Buchenberg		m	
Berthold, F.	Irrwisch-Fritze		f	
Brentano, C.	Geschichte vom braven Kasperl und dem schoenen Annerl	x	m	x
Chamisso, A. v.	Peter Schlemihls wundersame Geschichte	x	m	
Dincklage, E. v.	Der Striethast		f	
Droste-Hüllshof, A. v.	Die Judenbuche		f	
Eichendorff, J. v.	Die Glücksritter		m	x
Ense, K. A. V. v.	Reiz und Liebe	x	m	
Fräulein v. Wolf	Gemüth und Selbstsucht		f	
Frey, J.	Das erfüllte Versprechen		m	x
Gall, L. v.	Eine fromme Lüge		f	x
Gerstäcker, F.	Germelshausen		m	
Glümer, C. v.	Reich zu reich und arm zu arm		f	
Goethe, J. W. v.	Die neue Melusine	x	m	
Goldammer, L.	Eine Hochzeitsnacht		m	
Goldammer, L.	Auf Wiedersehen!		m	
Gotthelf, J.	Kurt von Koppigen		m	
Gotthelf, J.	Der Notar in der Falle		m	
Grillparzer, F.	Der arme Spielmann	x	m	
Grimm, H.	Das Kind		m	
Gross, J.	Vetter Isidor		m	x
Hackländer, F. W.	Zwei Nächte		m	x
Halm, F.	Die Marzipan-Lise		m	

author	**title**	**first-person narrative**	**gender**	***Adelsnovelle***
Hartmann, M.	Das Schloß im Gebirge	x	m	
Hauff, W.	Phantasien im Bremer Ratskeller	x	m	
Heyden, F. v.	Der graue John		m	
Heyse, P.	Der Weinhüter von Meran		m	
Höfer, E.	Rolof, der Rekrut		m	
Hoffmann, E. T. A.	Das Fräulein von Scuderi		m	
Holtei, K. v.	s Muhme-Leutnant-Saloppel		m	
Horner, H.	Der Säugling		m	
Immermann, K.	Der Carneval und die Somnambule	x	m	x
Kähler, L. A.	Die drei Schwestern	x	m	
Keller, G.	Romeo und Julia auf dem Dorfe		m	
Kinkel, G.	Margret		m	
Kinkel, J.	Musikalische Orthodoxie		f	x
Kleist, H. v.	Die Verlobung von St. Domingo		m	
Kompert, L.	Eine Verlorene		m	
Kopisch, A.	Ein Karnevalsfest auf Ischia		m	
Kopisch, A.	Der Träumer		m	
Kruse, L.	Nordische Freundschaft		m	
Kürnberger, F.	Der Drache		m	
Kugler, F.	Die Incantada		m	
Kurz, H.	Die beiden Tubus		m	
Lewald, F.	Die Tante	x	f	
Lohmann, F.	Die Entscheidung bei Hochkirch		f	
Lorm, H.	Ein adeliges Fräulein	x	m	x
Ludwig, J.	Das Gericht im Walde		f	
Meißner, A.	Der Müller vom Höft		m	
Meyr, M.	Der Sieg des Schwachen		m	
Mörike, E.	Mozart auf der Reise nach Prag		m	x
Mügge, T.	Am Malanger Fjord		m	
Müller, W.	Debora		m	

author	**title**	**first-person narrative**	**gender**	***Adelsnovelle***
Pichler, A.	Der Flüchtling	x	m	
Raabe, W.	Das letzte Recht		m	
Reich, M.	Mammon im Gebirge		m	
Riehl, W. H.	Jörg Muckenhuber		m	
Roquette, O.	Die Schlangenkönigin	x	m	
Rumohr, K. F.	Der letzte Savello		m	
Sacher-Masoch, L.	Don Juan von Kolomea		m	
Schefer, L.	Die Düvecke oder die Leiden einer Königin		m	
Scheffel, J. V. v.	Hugideo		m	
Schmid, H.	Mohrenfranzel	x	m	
Schreyvogel, J.	Samuel Brinks letzte Liebesgeschichte	x	m	
Schücking, L.	Die Schwester		m	x
Spindler, K.	Die Engel-Ehe		m	
Sternberg, A. v.	Scholastika		m	
Stifter, A.	Brigitta	x	m	
Storm, T.	Eine Malerarbeit	x	m	
Tesche, W.	Der Enten-Piet		m	
Tieck, L.	Die Gemälde		m	
Tieck, L.	Des Lebens Überfluß		m	
Traun, J. v. d.	Der Gebirgspfarrer		m	
Waldmüller, R.	Es ist nicht gut, daß der Mensch allein sei		m	
Wallner, F.	Der arme Josy	x	m	
Wichert, E.	Ansas und Grita	x	m	
Widmann, A.	Die katholische Mühle	x	m	
Wilbrandt, A.	Johann Ohlerich		m	
Wild, H.	Eure Wege sind nicht meine Wege		f	x
Wildermuth, O.	Streit in der Liebe und Liebe im Streit		f	
Wolf, A.	Der Stern der Schönheit		m	
Ziegler, F. W.	Saat und Ernte		m	
Zschokke, H.	Der tote Gast		m	

Table A.2: Used metadata in the analysis of the novellas.

A3 Novella Character Networks



Koreferenzen in Novellen

In 86 Novellen des Deutschen Novellenschatzes sollen Figuren und all ihre Synonyme (keine Pronomen) annotiert werden.

Was & Warum

Im Rahmen meiner Arbeit als Doktorand an der ETH Zürich untersuche ich Netzwerke in literarischen Texten. Da Literatur davon lebt, Figuren nicht nur mit ihrem Namen, sondern mit den unterschiedlichsten Synonymen und Pseudonymen zu bezeichnen, ist es für aussagekräftige Figurennetzwerke unerlässlich, den Namen aller auftretenden Charaktere die ihnen zugesprochenen Aliasse zuzuordnen.

Wie

Die Novellen müssen gelesen und die Namen und sonstigen Bezeichnungen aller genannten Personen (keine Pronomen) müssen notiert und den richtigen Figuren zugeordnet werden. Diese Zuordnung wird pro Novelle nach folgendem Schema angegeben: Name der Figur; Liste der Synonyme inklusive des Namens.

gekürztes Beispiel (*Die Judenbuche*, Droste-Hülshoff)

- Friedrich Mergel; Friedrich, Friedrich Mergel, Sohn, Fritzchen, Fritz, Junge, Gesell, Vaterssohn, ...
- Herrmann Mergel; Herrmann, Herrmann Mergel, Helbmeiers, Grundeigenthümer, Wittwer, Bräutigam, Vater
- Ulysses; Ulysses

Die Listen werden benötigt, um möglichst genaue Figurennetzwerke erstellen zu können, welche im weiteren Verlauf vergleichend analysiert werden.

Wer ist dahinter

Die Arbeit ist eingebettet in ein grösseres Forschungsprojekt mit dem Titel "Reading at Scale. Mixing Methods in Literary Corpus Analysis". Dieses Forschungsprojekt ist eine Zusammenarbeit des Social Network Labs der ETH Zürich mit dem Fachbereich Germanistik – Digitale Literaturwissenschaft der TU Darmstadt. Die PIs des Projekts sind Thomas Weitin und Ulrik Brandes.

Wer beiträgt

Alle, die Spaß am Lesen haben und immer wieder auf der Suche nach neuem interessanten Lesestoff sind. Es finden sich sowohl bekannte und schon lange kanonisch gewordene Texte, wie Eduard Moerikes *Mozart auf der Reise nach Prag* oder Gottfried Kellers *Romeo und Julia auf dem Dorfe* als auch unbekanntere Werke, etwa Julie Ludwigs *Das Gericht im Walde* oder die unter dem Pseudonym Fräulein von Wolf erschienene Novelle *Gemüth und Selbstsucht*.

Zusätzliche Links

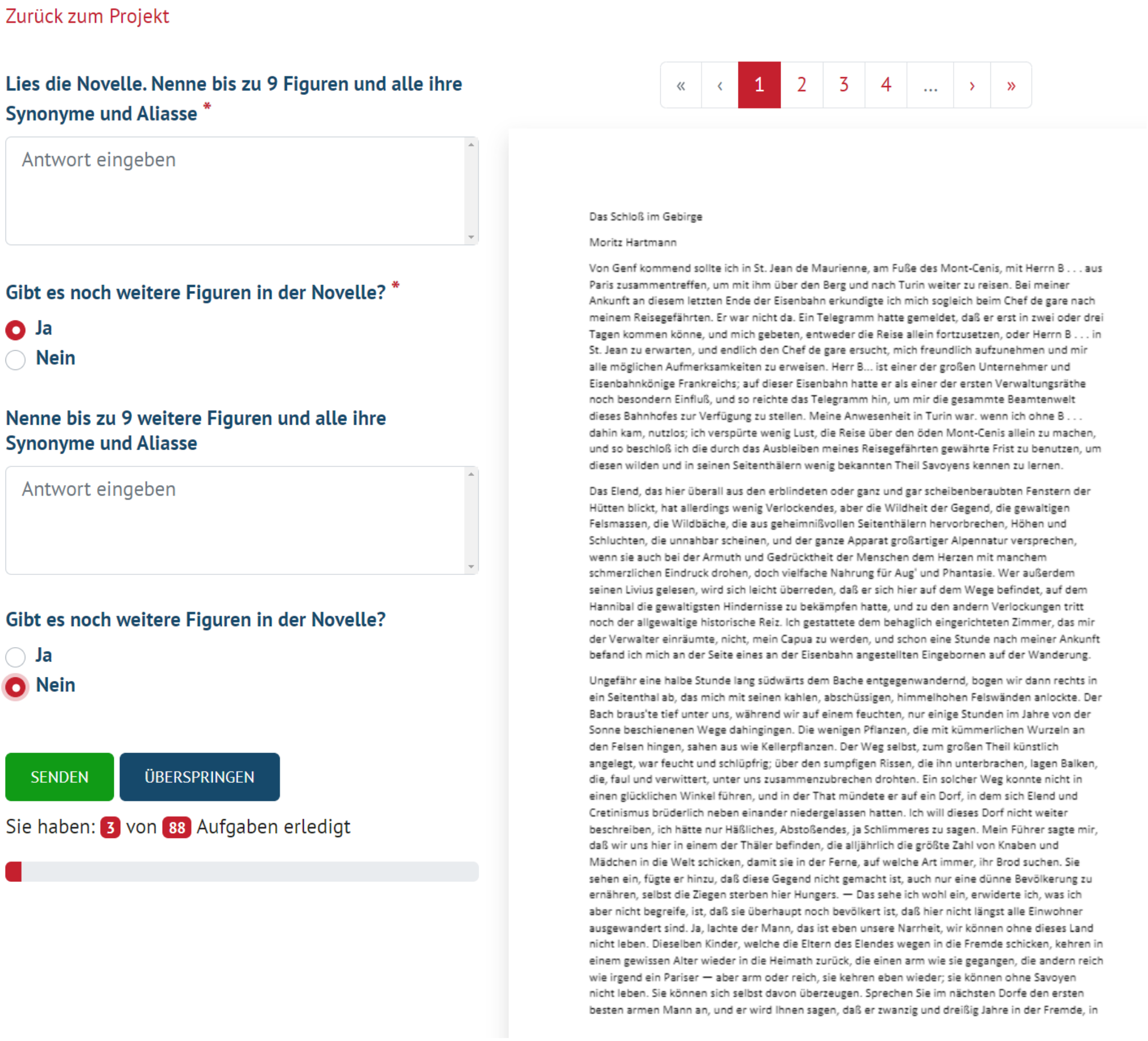
Novellenschatz im deutschen Textarchniv: https://www.deutschestextarchiv.de/search/metadata?corpus=novellenschatz

Reading at Scale Projektbeschreibung: https://www.digitalhumanitiescooperation.de/en/projects/reading-at-scale/

Social Networks Lab ETH Zürich: https://sn.ethz.ch/

Meine Kontaktdaten: https://sn.ethz.ch/profile.html?persid=211617

Fig. A.1: Frontpage of the Citizen Science Project website.



Lies die Novelle. Nenne bis zu 9 Figuren und alle ihre Synonyme und Aliasse *

Antwort eingeben

Gibt es noch weitere Figuren in der Novelle? *

● Ja
○ Nein

Nenne bis zu 9 weitere Figuren und alle ihre Synonyme und Aliasse

Antwort eingeben

Gibt es noch weitere Figuren in der Novelle?

○ Ja
● Nein

Sie haben: 3 von 88 Aufgaben erledigt

Das Schloß im Gebirge

Moritz Hartmann

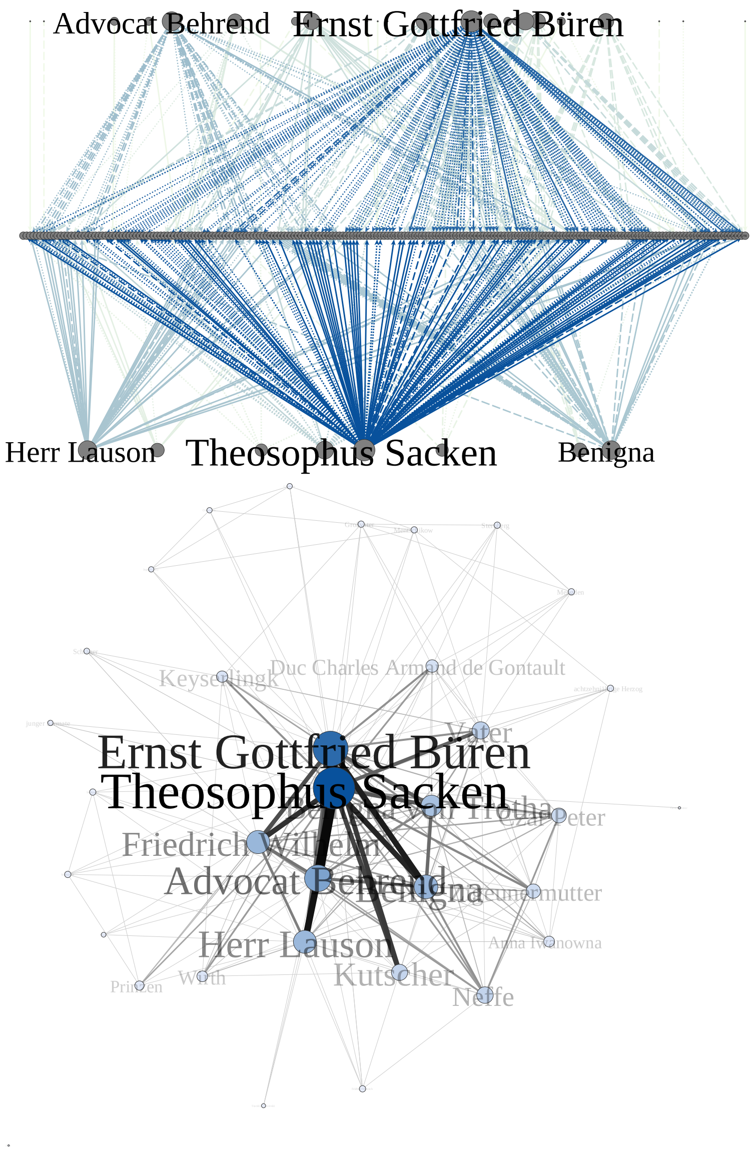
Von Genf kommend sollte ich in St. Jean de Maurienne, am Fuße des Mont-Cenis, mit Herrn B . . . aus Paris zusammentreffen, um mit ihm über den Berg und nach Turin weiter zu reisen. Bei meiner Ankunft an diesem letzten Ende der Eisenbahn erkundigte ich mich sogleich beim Chef de gare nach meinem Reisegefährten. Er war nicht da. Ein Telegramm hatte gemeldet, daß er erst in zwei oder drei Tagen kommen könne, und mich gebeten, entweder die Reise allein fortzusetzen, oder Herrn B . . . in St. Jean zu erwarten, und endlich den Chef de gare ersucht, mich freundlich aufzunehmen und mir alle möglichen Aufmerksamkeiten zu erweisen. Herr B… ist einer der großen Unternehmer und Eisenbahnkönige Frankreichs; auf dieser Eisenbahn hatte er als einer der ersten Verwaltungsräthe noch besondern Einfluß, und so reichte das Telegramm hin, um mir die gesammte Beamtenwelt dieses Bahnhofes zur Verfügung zu stellen. Meine Anwesenheit in Turin war. wenn ich ohne B . . . dahin kam, nutzlos; ich verspürte wenig Lust, die Reise über den öden Mont-Cenis allein zu machen, und so beschloß ich die durch das Ausbleiben meines Reisegefährten gewährte Frist zu benutzen, um diesen wilden und in seinen Seitenthälern wenig bekannten Theil Savoyens kennen zu lernen.

Das Elend, das hier überall aus den erblindeten oder ganz und gar scheibenberaubten Fenstern der Hütten blickt, hat allerdings wenig Verlockendes, aber die Wildheit der Gegend, die gewaltigen Felsmassen, die Wildbäche, die aus geheimnißvollen Seitenthälern hervorbrechen, Höhen und Schluchten, die unnahbar scheinen, und der ganze Apparat großartiger Alpennatur versprechen, wenn sie auch bei der Armuth und Gedrücktheit der Menschen dem Herzen mit manchem schmerzlichen Eindruck drohen, doch vielfache Nahrung für Aug' und Phantasie. Wer außerdem seinen Livius gelesen, wird sich leicht überreden, daß er sich hier auf dem Wege befindet, auf dem Hannibal die gewaltigsten Hindernisse zu bekämpfen hatte, und zu den andern Verlockungen tritt noch der allgewaltige historische Reiz. Ich gestattete dem behaglich eingerichteten Zimmer, das mir der Verwalter einräumte, nicht, mein Capua zu werden, und schon eine Stunde nach meiner Ankunft befand ich mich an der Seite eines an der Eisenbahn angestellten Eingebornen auf der Wanderung.

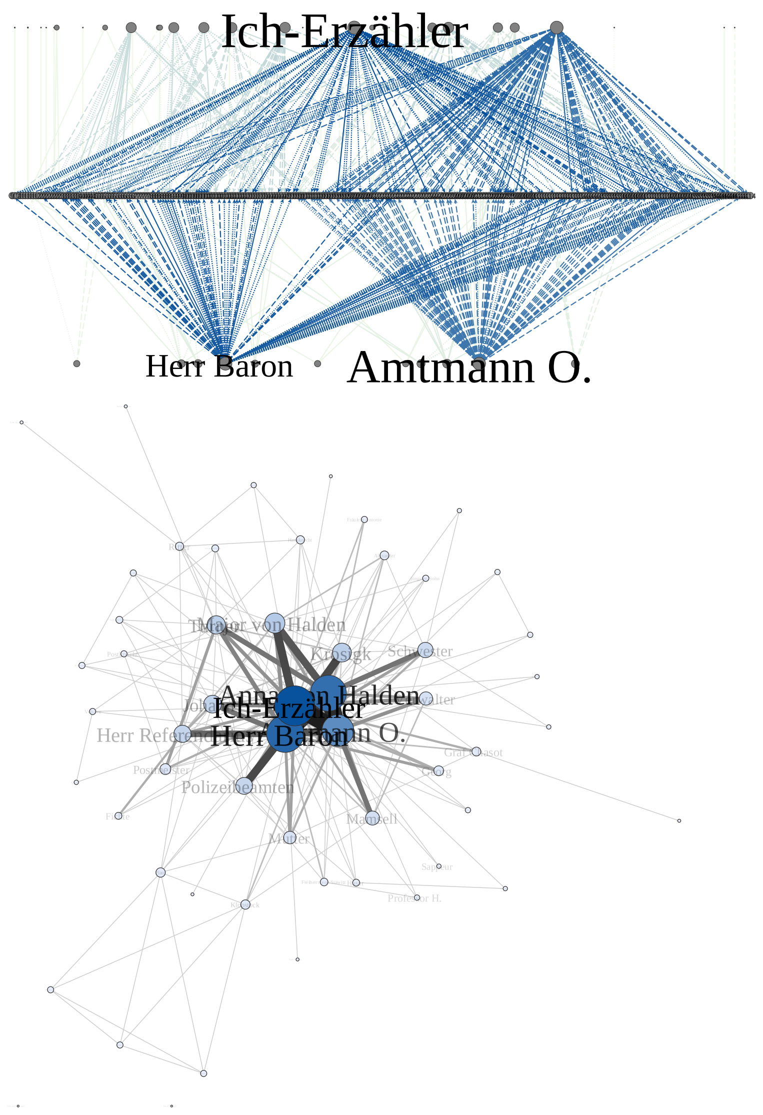
Ungefähr eine halbe Stunde lang südwärts dem Bache entgegenwandernd, bogen wir dann rechts in ein Seitenthal ab, das mich mit seinen kahlen, abschüssigen, himmelhohen Felswänden anlockte. Der Bach braus'te tief unter uns, während wir auf einem feuchten, nur einige Stunden im Jahre von der Sonne beschienenen Wege dahingingen. Die wenigen Pflanzen, die mit kümmerlichen Wurzeln an den Felsen hingen, sahen aus wie Kellerpflanzen. Der Weg selbst, zum großen Theil künstlich angelegt, war feucht und schlüpfrig; über den sumpfigen Rissen, die ihn unterbrachen, lagen Balken, die, faul und verwittert, unter uns zusammenzubrechen drohten. Ein solcher Weg konnte nicht in einen glücklichen Winkel führen, und in der That mündete er auf ein Dorf, in dem sich Elend und Cretinismus brüderlich neben einander niedergelassen hatten. Ich will dieses Dorf nicht weiter beschreiben, ich hätte nur Häßliches, Abstoßendes, ja Schlimmeres zu sagen. Mein Führer sagte mir, daß wir uns hier in einem der Thäler befinden, die alljährlich die größte Zahl von Knaben und Mädchen in die Welt schicken, damit sie in der Ferne, auf welche Art immer, ihr Brod suchen. Sie sehen ein, fügte er hinzu, daß diese Gegend nicht gemacht ist, auch nur eine dünne Bevölkerung zu ernähren, selbst die Ziegen sterben hier Hungers. — Das sehe ich wohl ein, erwiderte ich, was ich aber nicht begreife, ist, daß sie überhaupt noch bevölkert ist, daß hier nicht längst alle Einwohner ausgewandert sind. Ja, lachte der Mann, das ist eben unsere Narrheit, wir können ohne dieses Land nicht leben. Dieselben Kinder, welche die Eltern des Elendes wegen in die Fremde schicken, kehren in einem gewissen Alter wieder in die Heimath zurück, die einen arm wie sie gegangen, die andern reich wie irgend ein Pariser — aber arm oder reich, sie kehren eben wieder; sie können ohne Savoyen nicht leben. Sie können sich selbst davon überzeugen. Sprechen Sie im nächsten Dorfe den ersten besten armen Mann an, und er wird Ihnen sagen, daß er zwanzig und dreißig Jahre in der Fremde, in

Fig. A.2: Project page with questions on the left and text of the novella on the right.

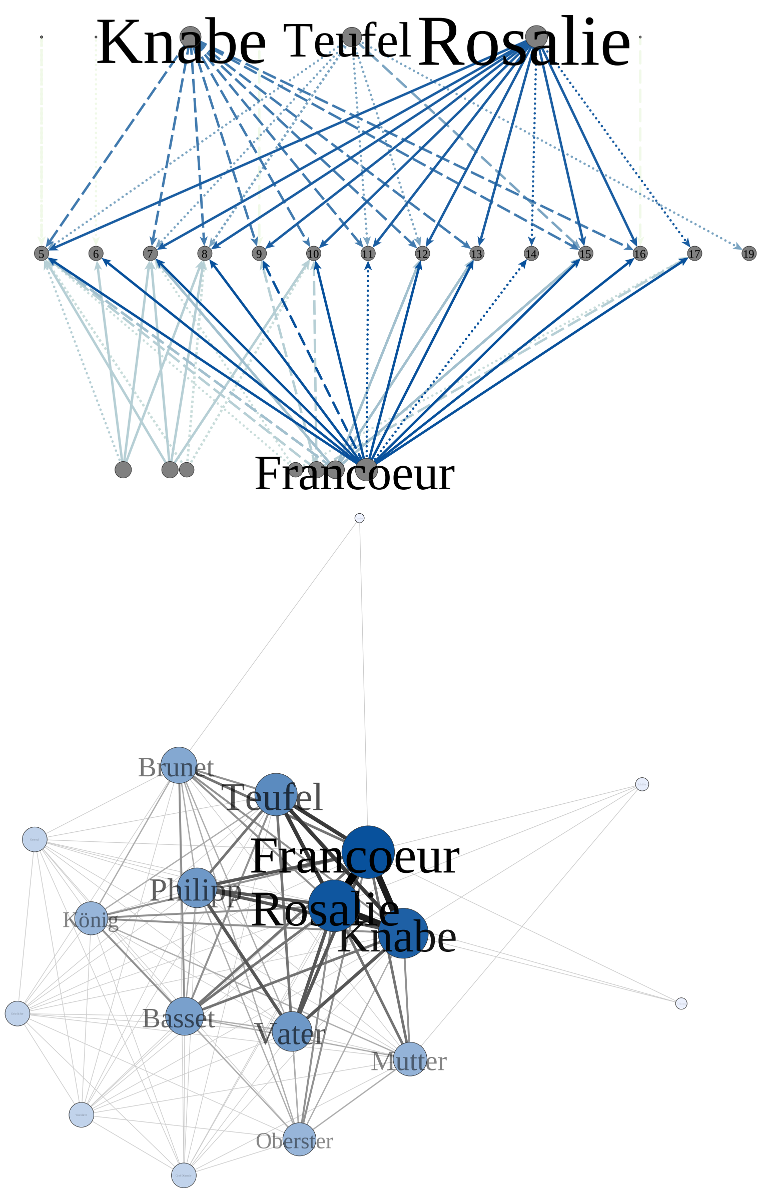
Alexis
Herr von Sacken



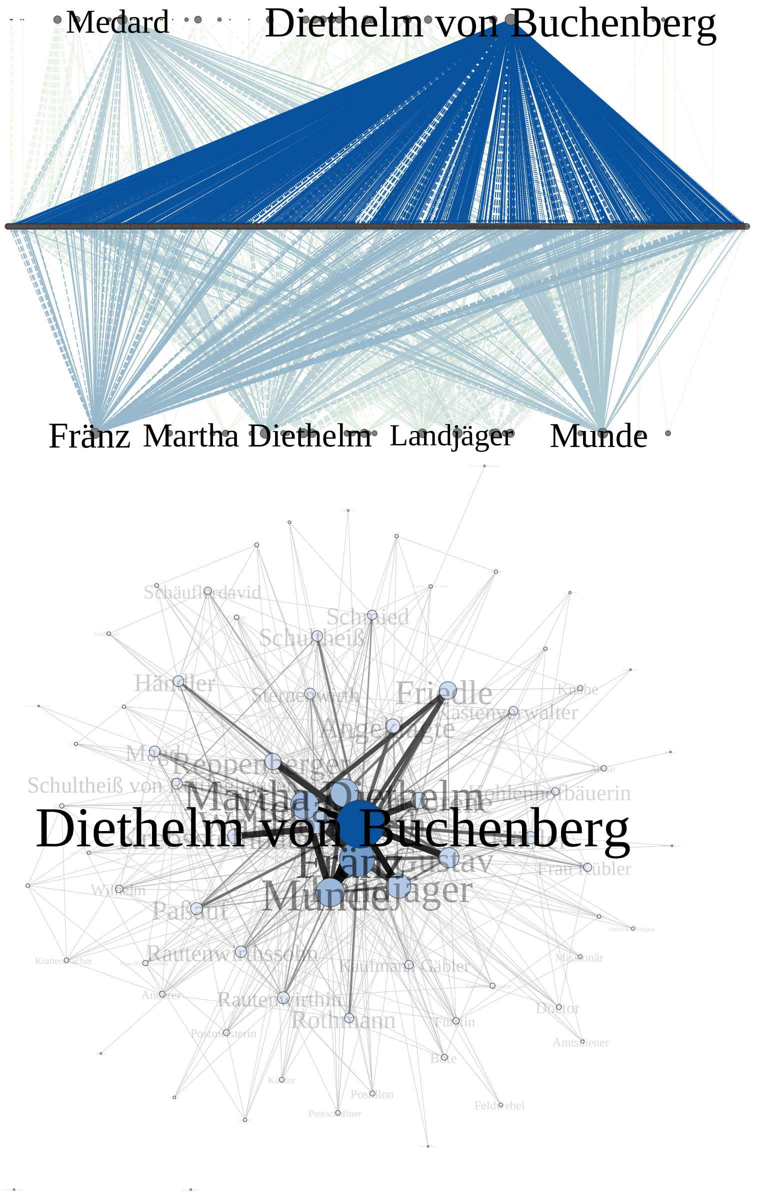
Andolt
Eine Nacht



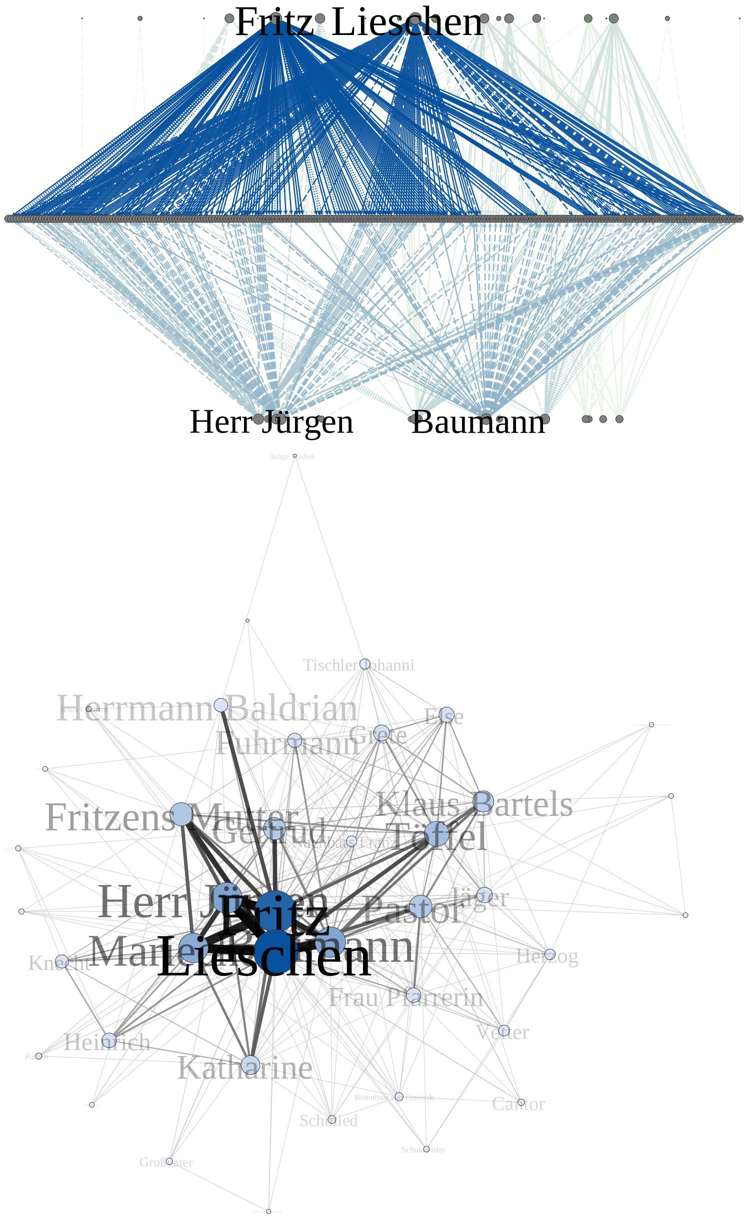
Arnim
Der tolle Invalide auf dem Fort Ratonneau



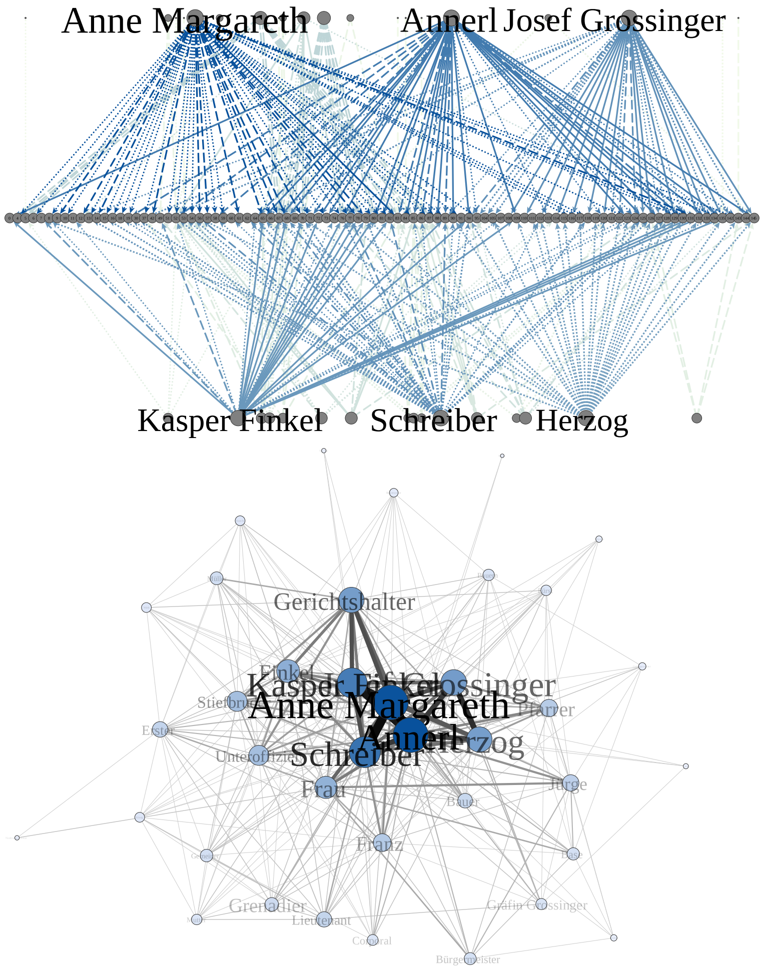
Auerbach
Die Geschichte des Diethelm von Buchenberg



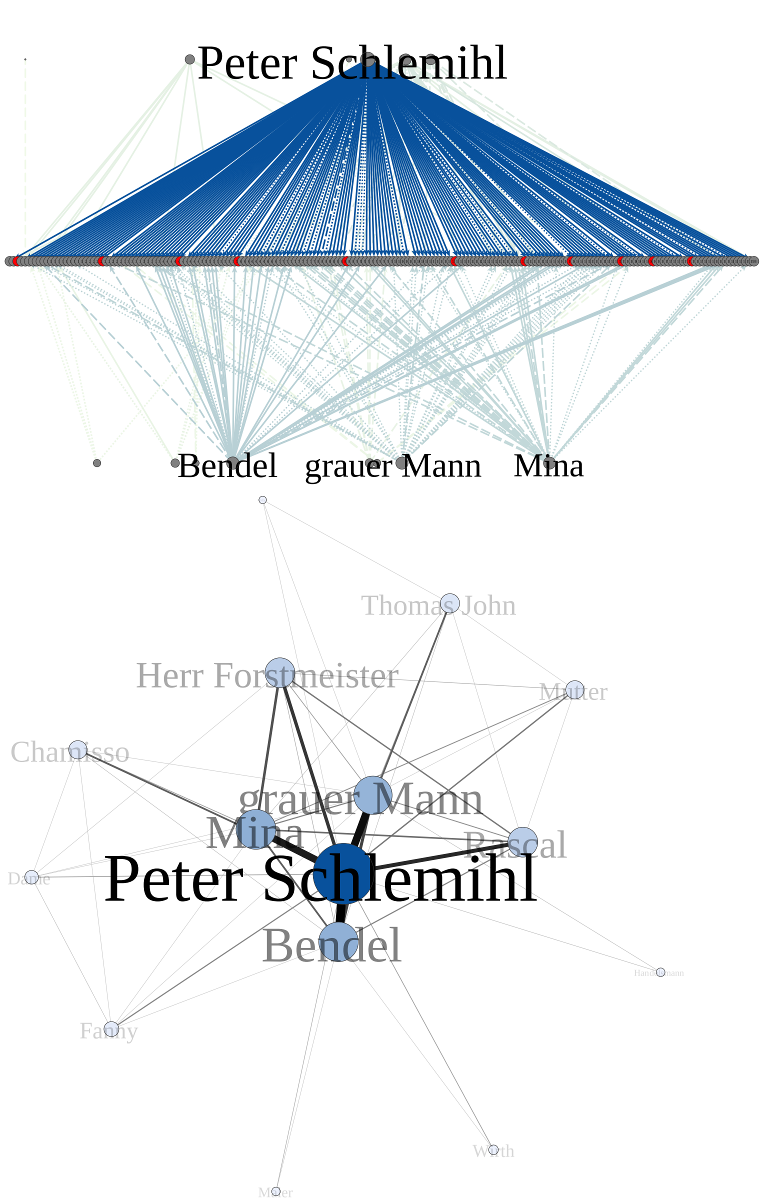
Berthold
Irrwisch-Fritze



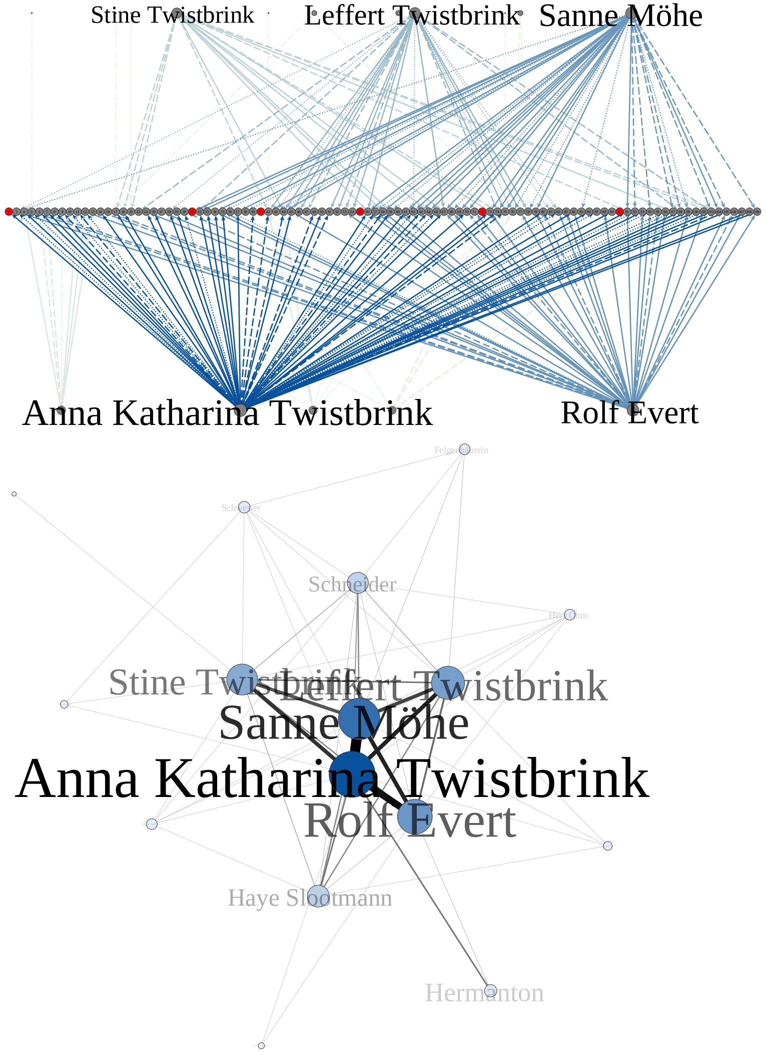
Brentano
Geschichte vom braven Kasperl und dem schönen Annerl



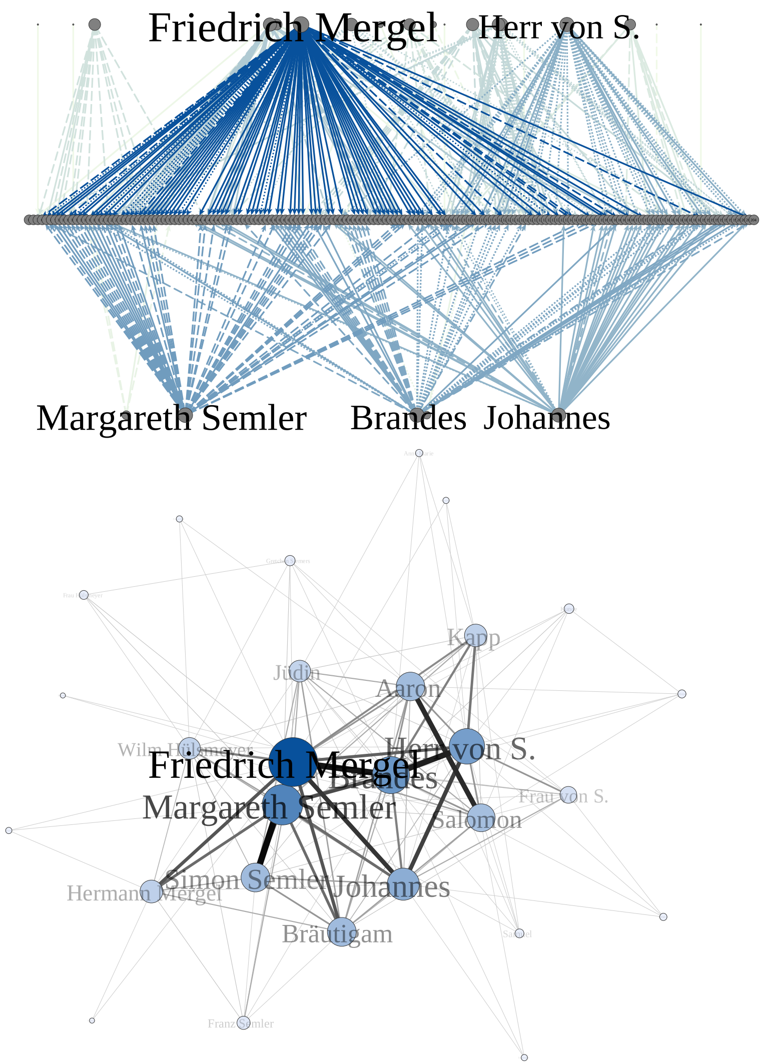
Chamisso
Peter Schlemihl's wundersame Geschichte



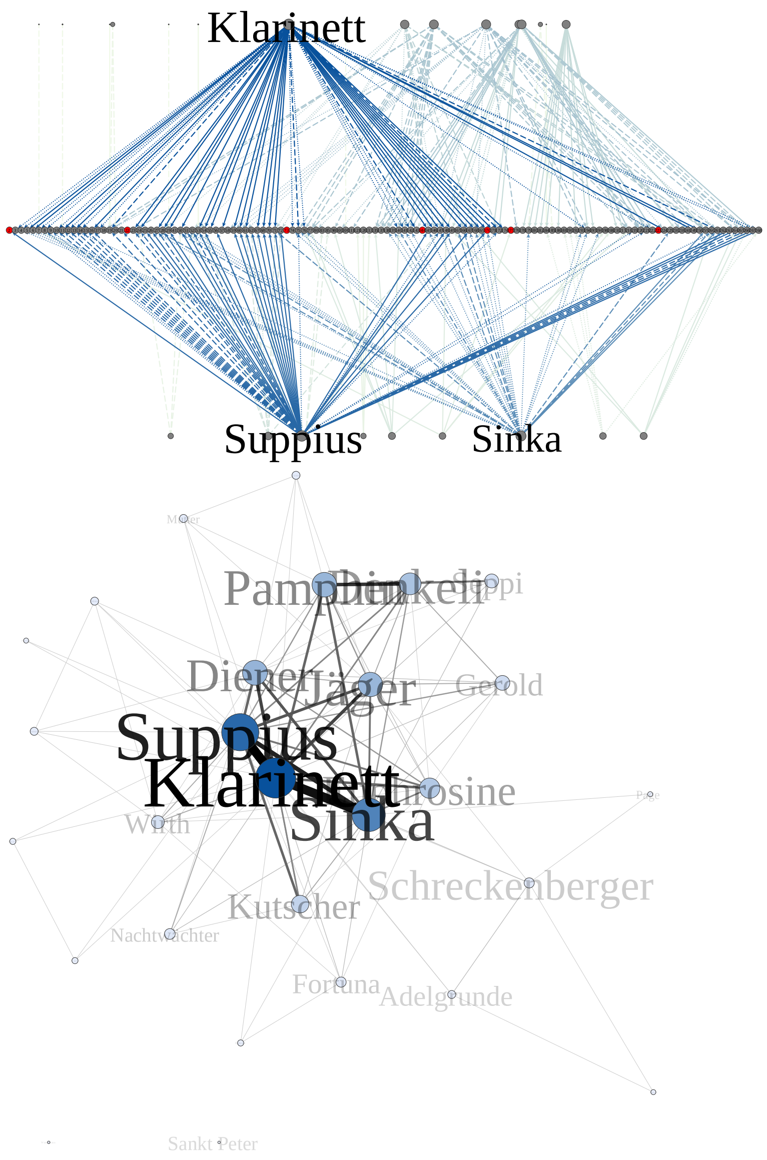
Dincklage
Der Striethast



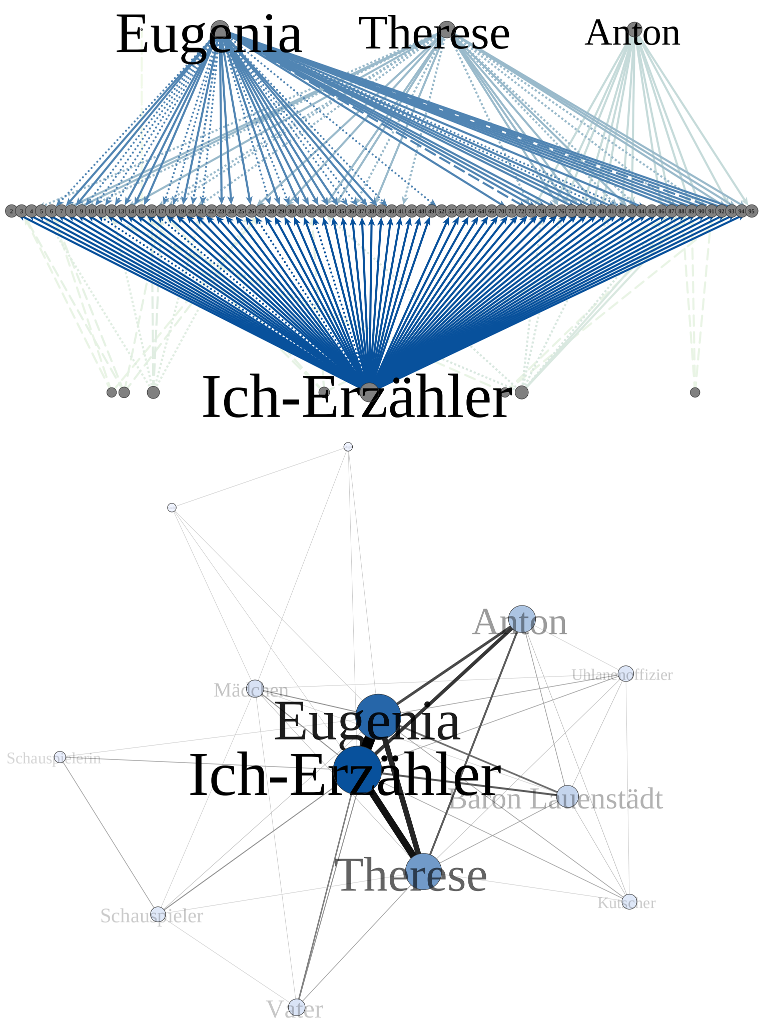
Droste
Die Judenbuche



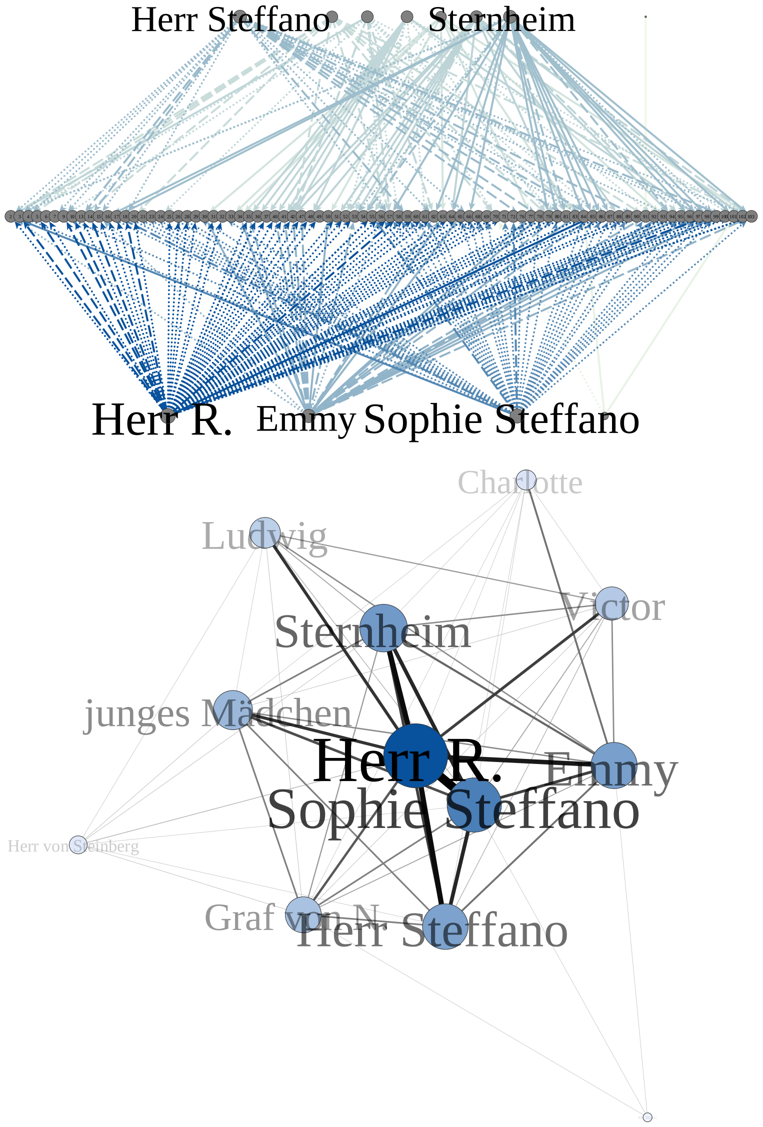
Eichendorff
Die Glücksritter



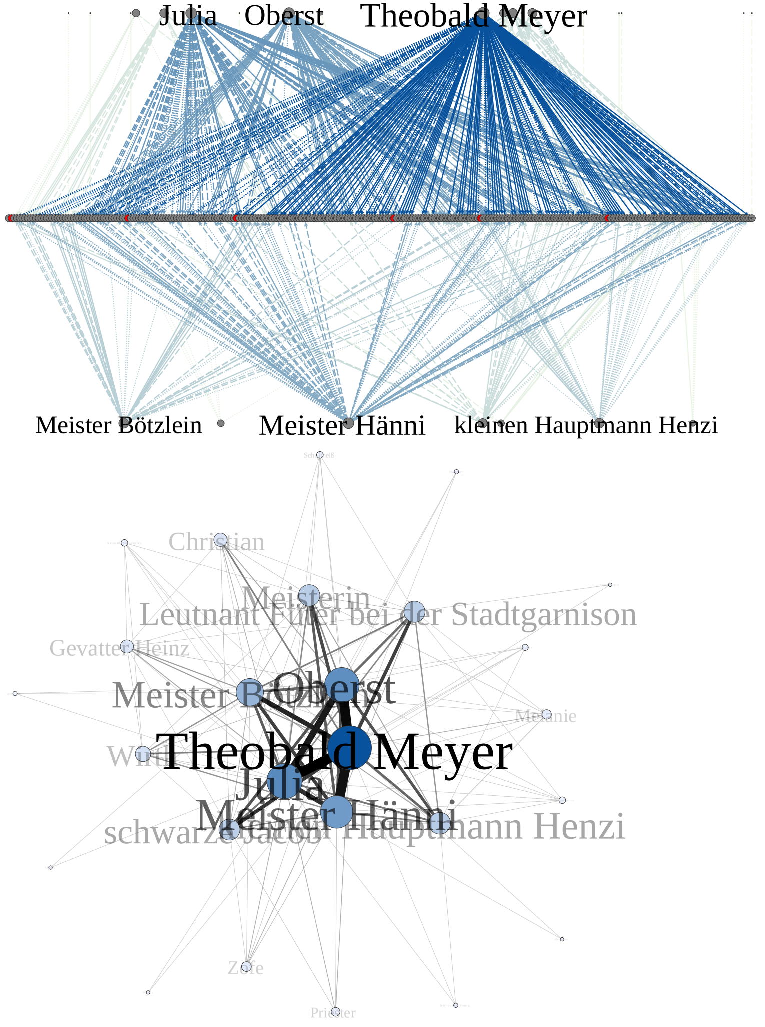
Ense
Reiz und Liebe



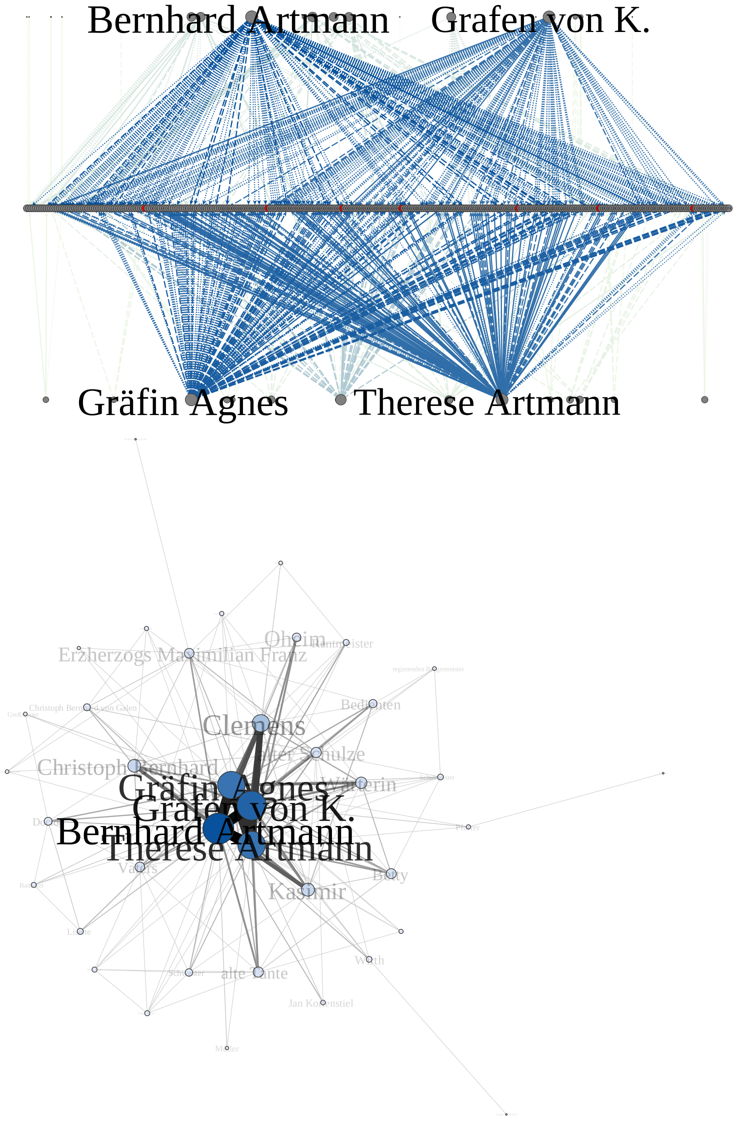
Fräulein Wolf
Gemüth und Selbstsucht



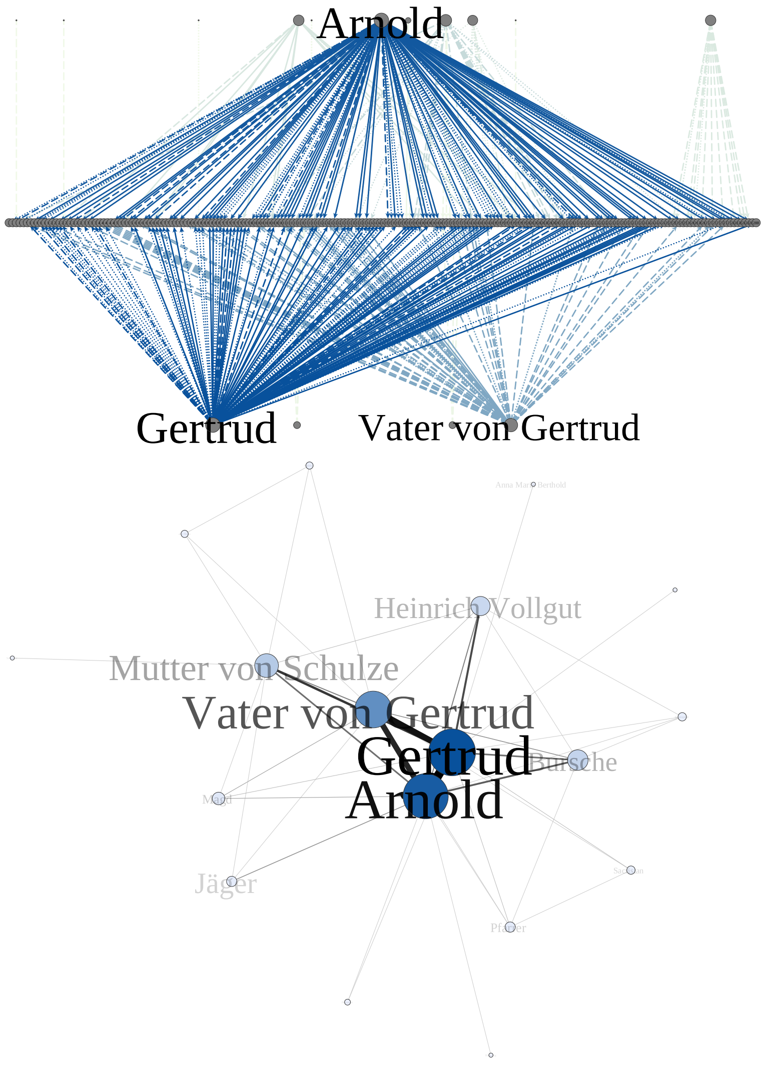
Frey
Das erfüllte Versprechen



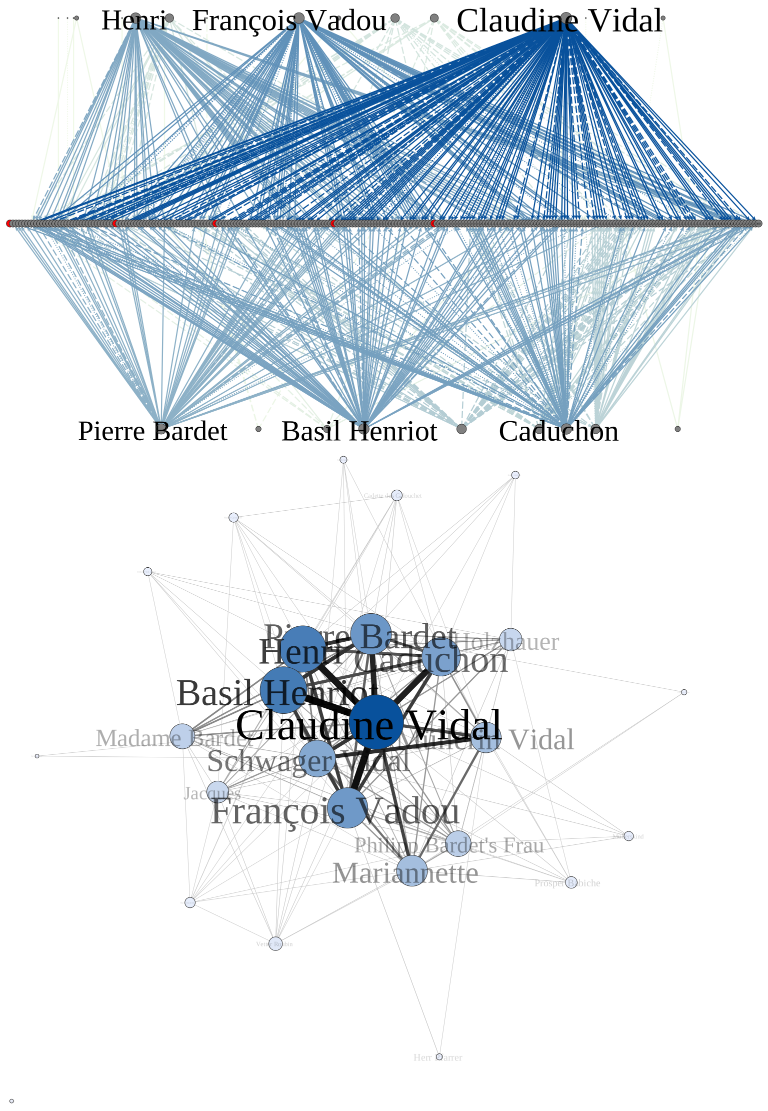
Gall
Eine fromme Lüge

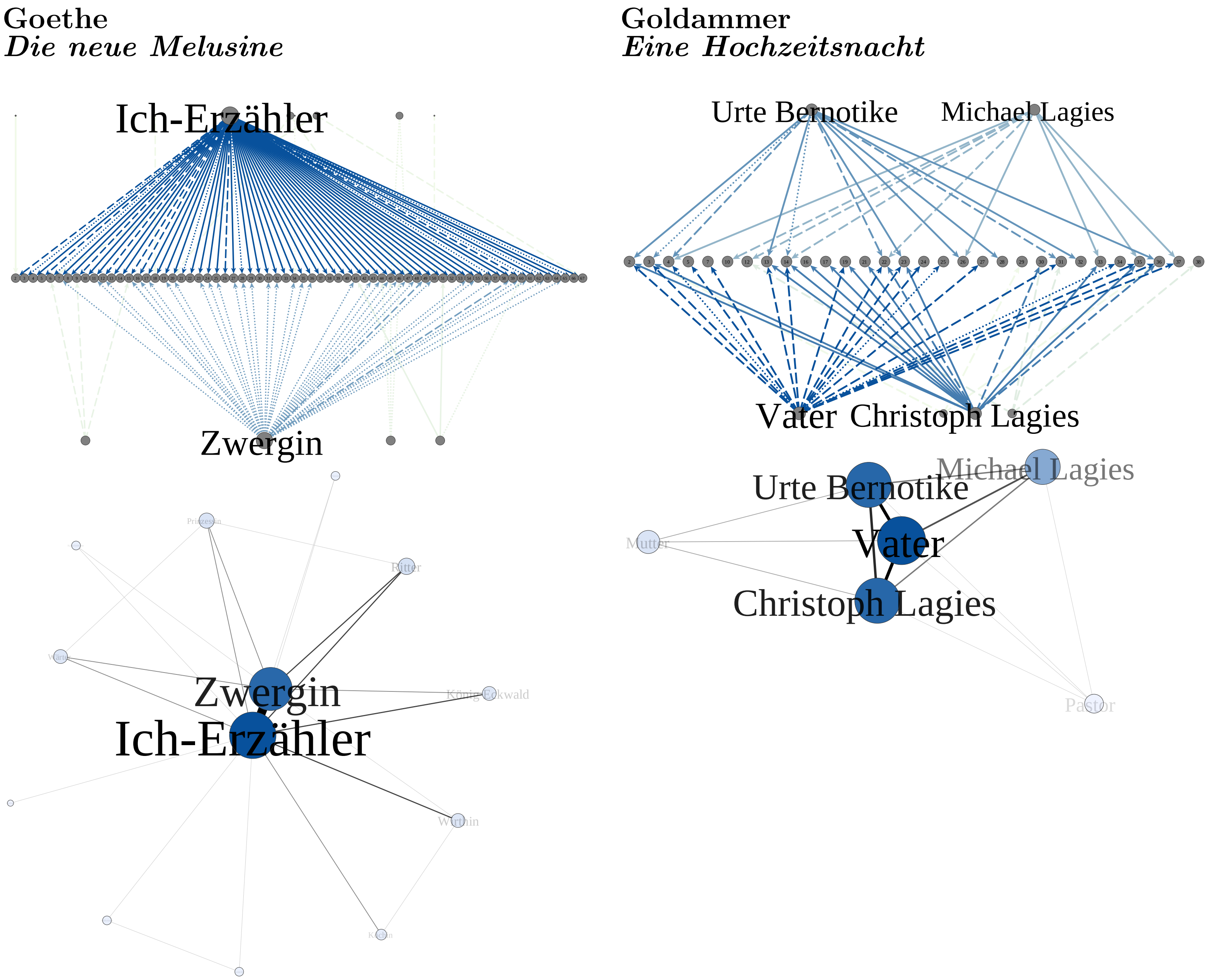


Gerstäcker
Germelshausen

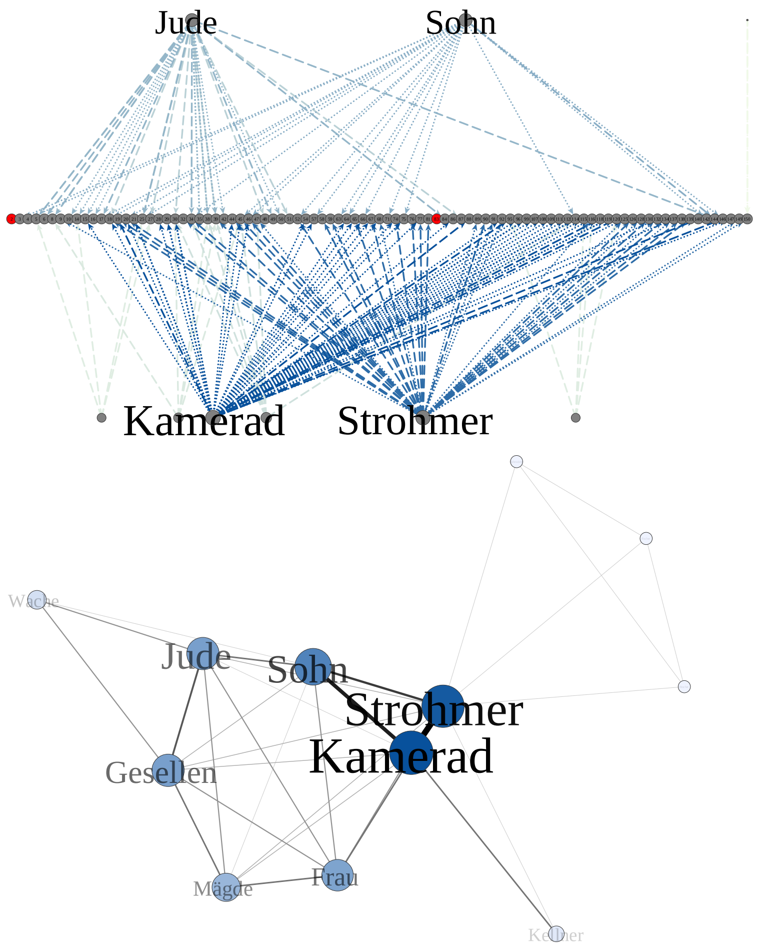


Glümer
Reich zu reich und arm zu arm

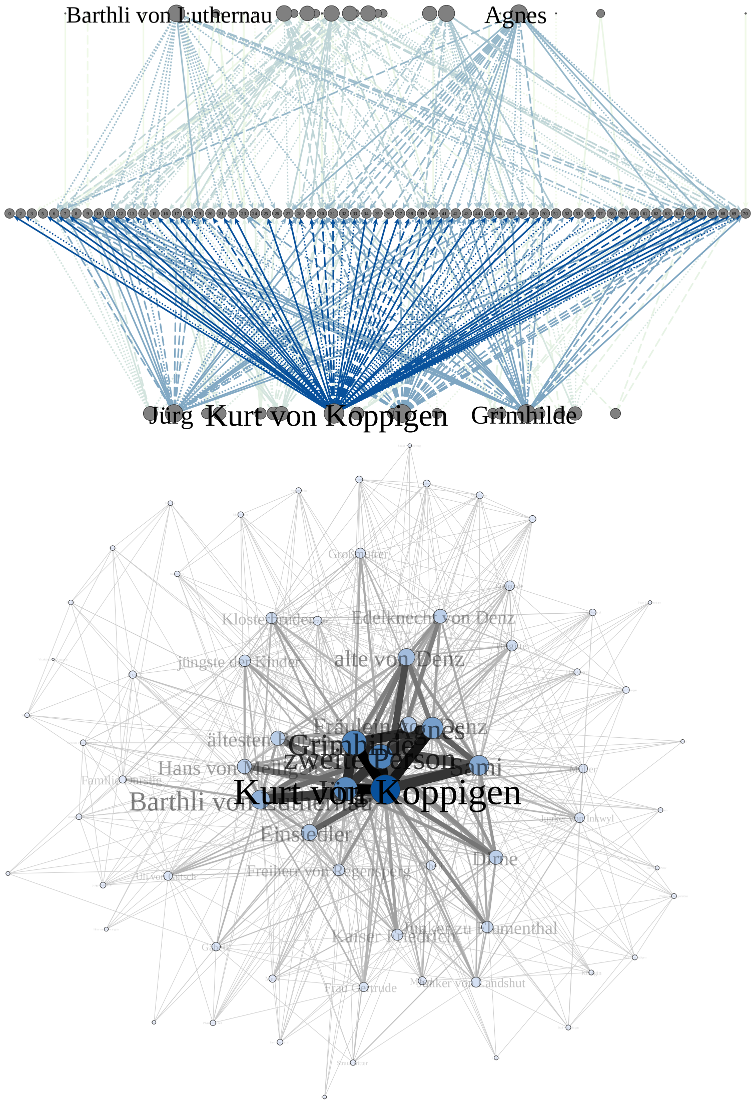



Goethe
Die neue Melusine
Ich-Erzähler
Zwergin
Prinzessin
Ritter
Wärter
Zwergin
König Eckwald
Ich-Erzähler
Wirthin
Köchin
Goldammer
Eine Hochzeitsnacht
Urte Bernotike
Michael Lagies
Vater
Christoph Lagies
Urte Bernotike
Michael Lagies
Mutter
Vater
Christoph Lagies
Pastor

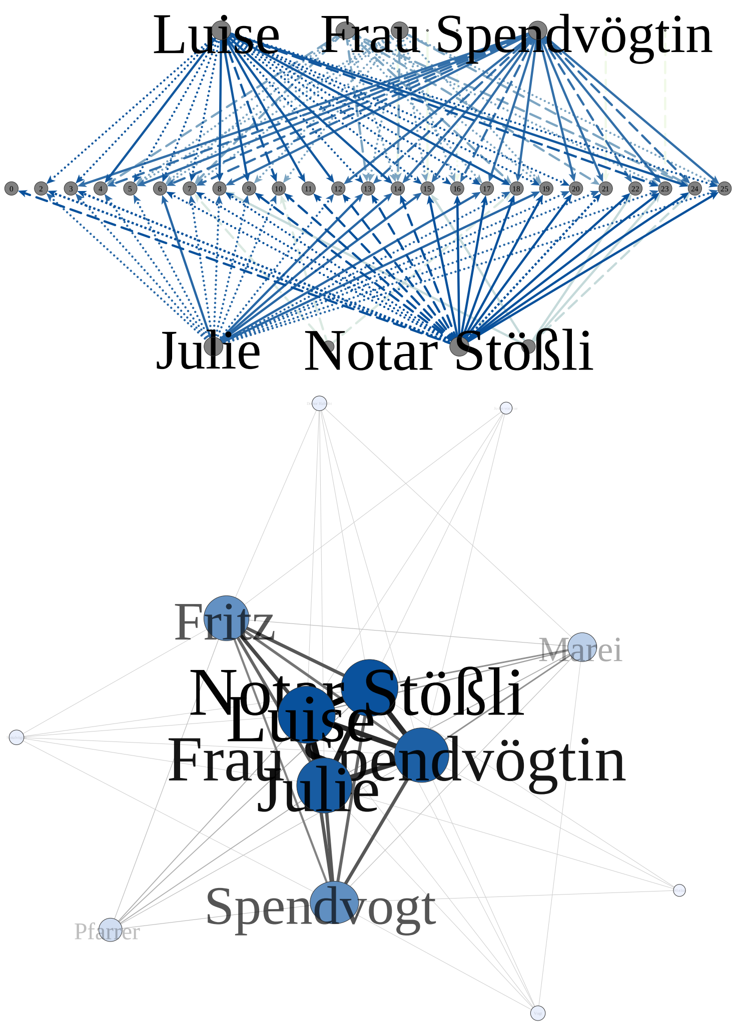
Goldammer
Auf Wiedersehen!



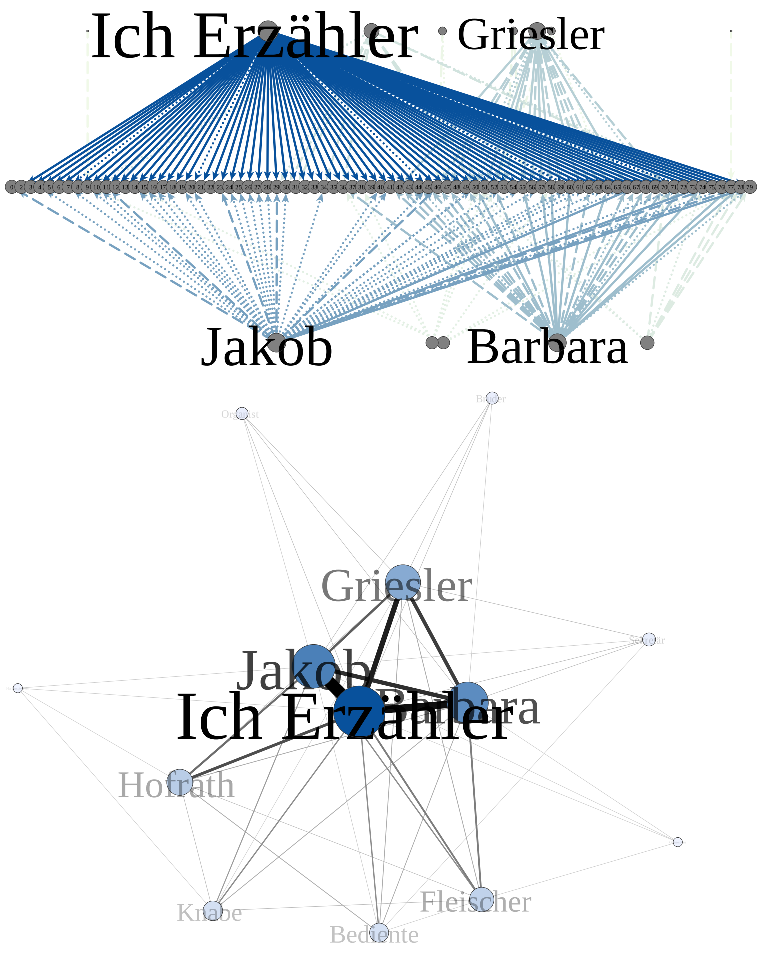
Gotthelf
Kurt von Koppigen



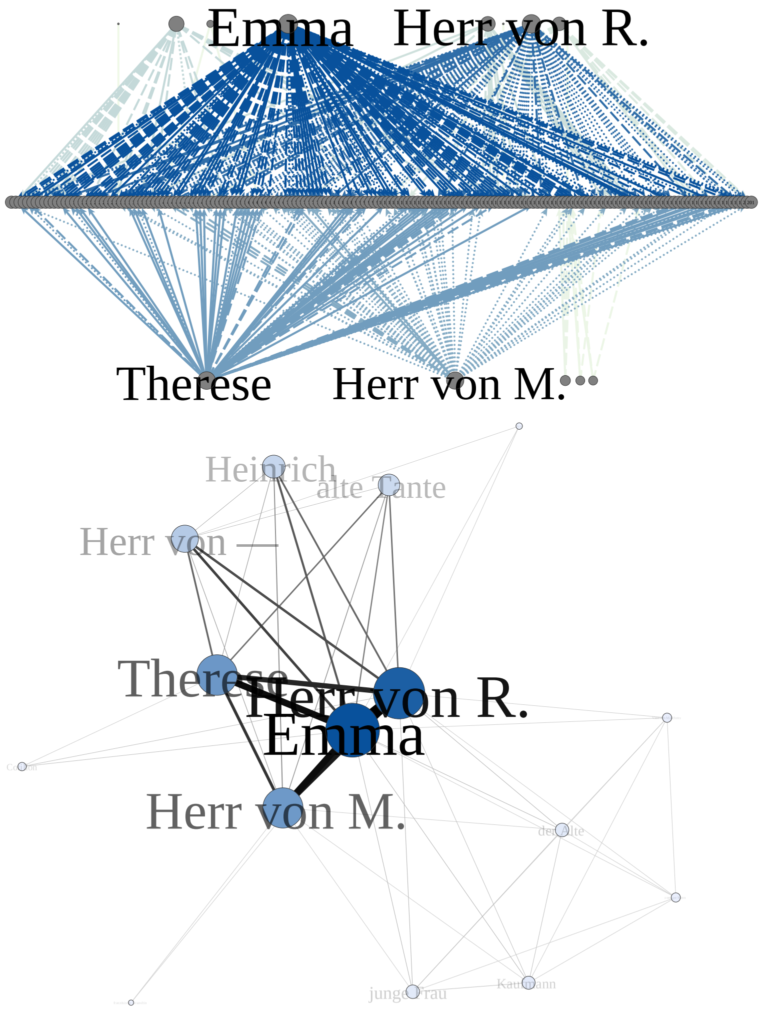
Gotthelf
Der Notar in der Falle



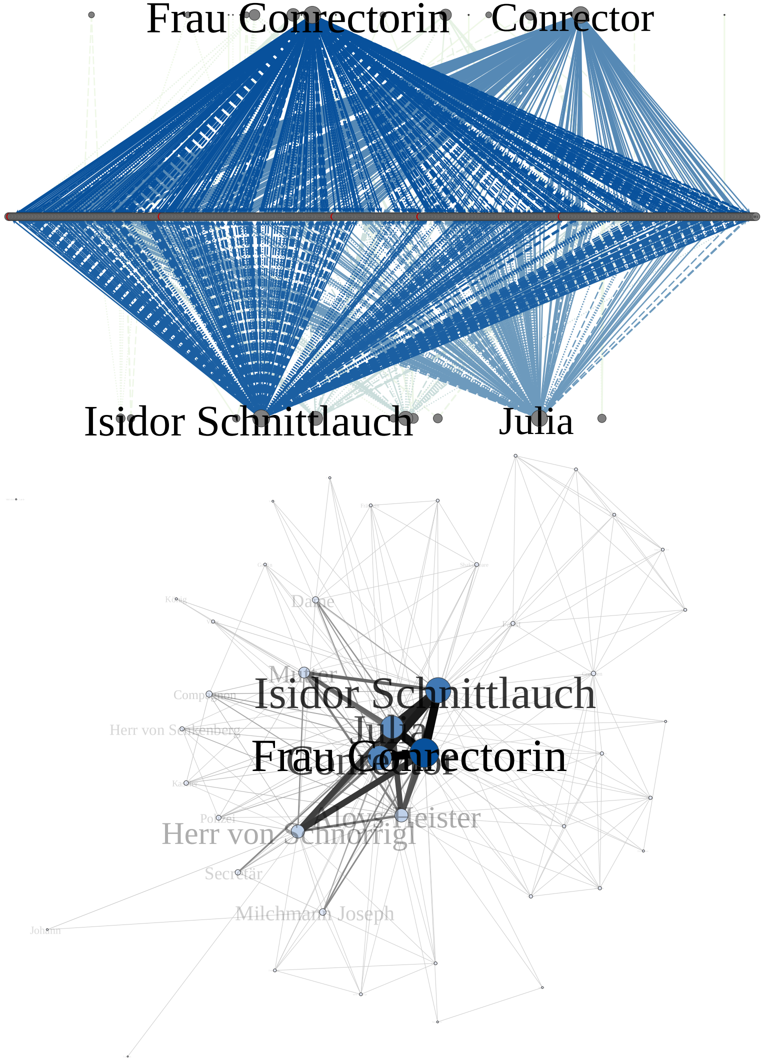
Grillparzer
Der arme Spielmann



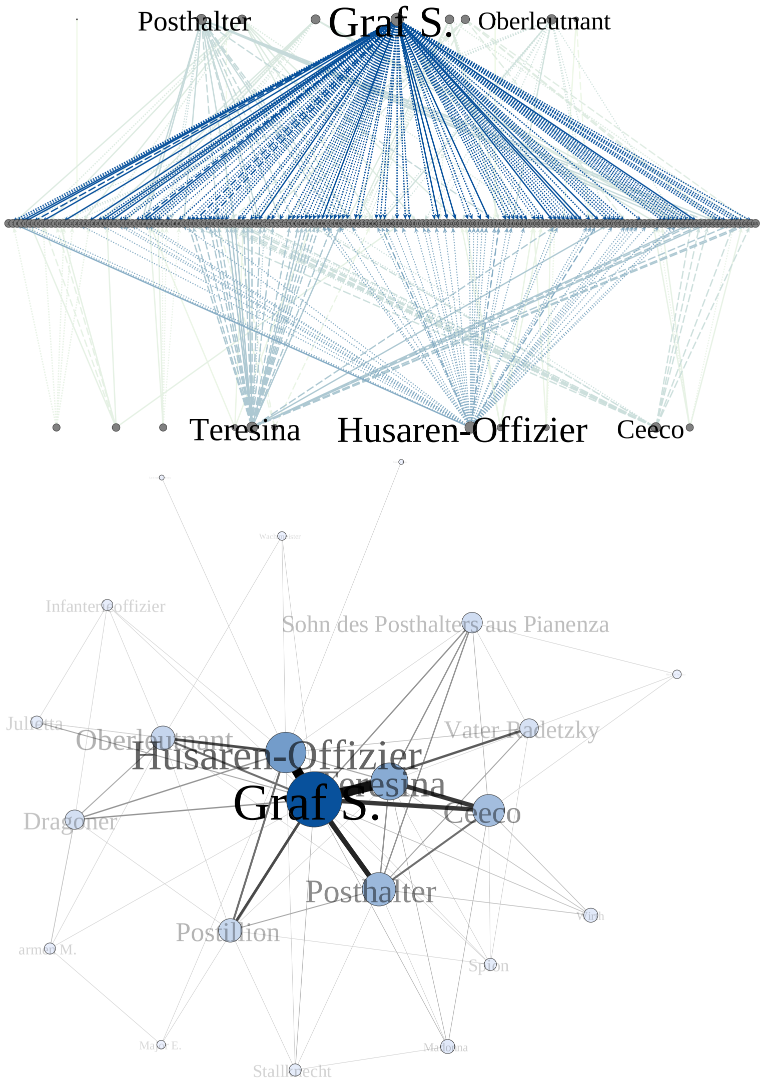
Grimm
Das Kind



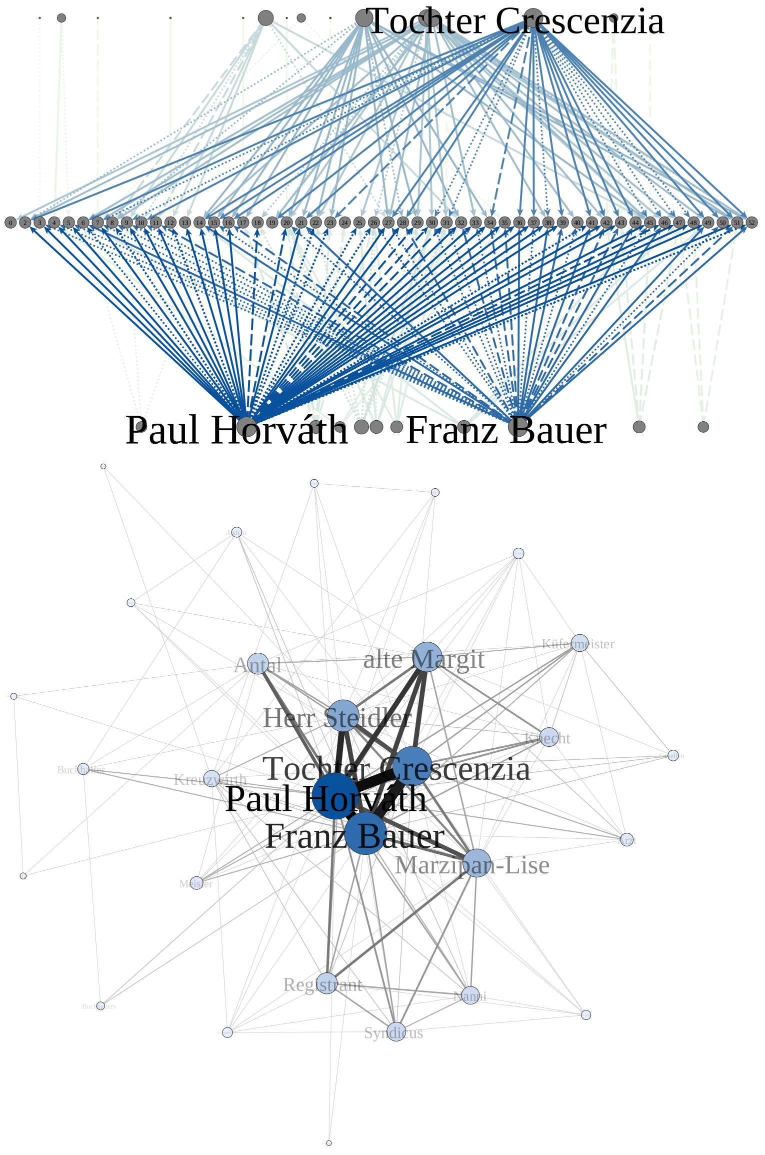
Grosse
Vetter Isidor



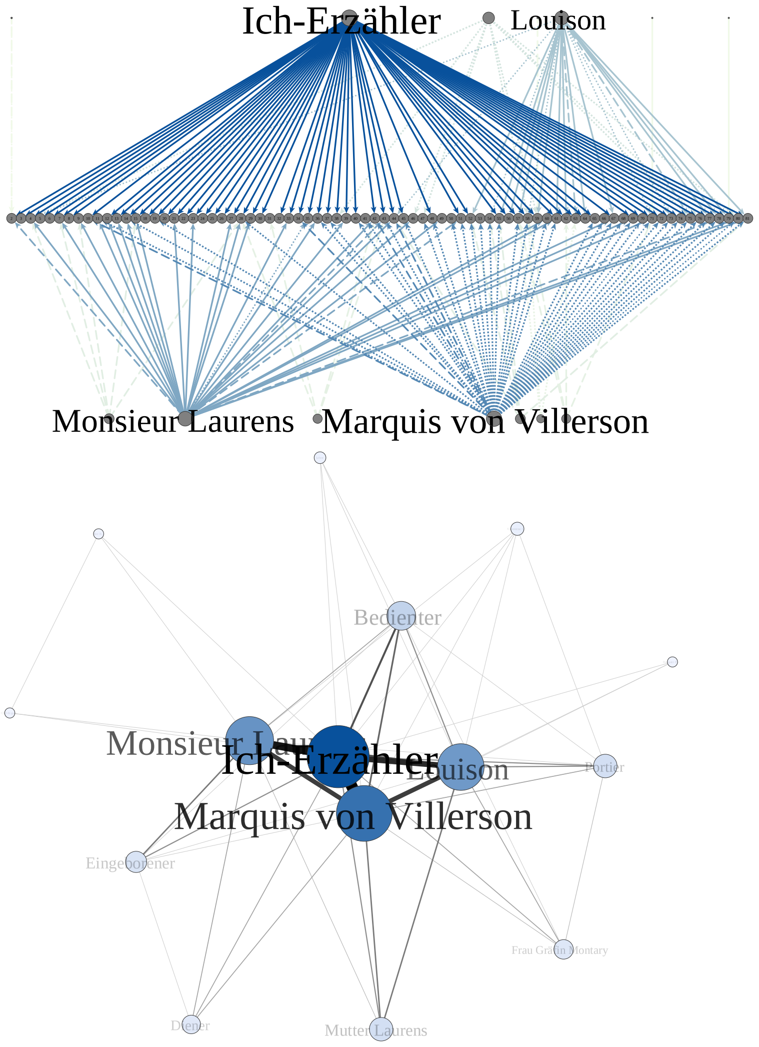
Hackländer
Zwei Nächte



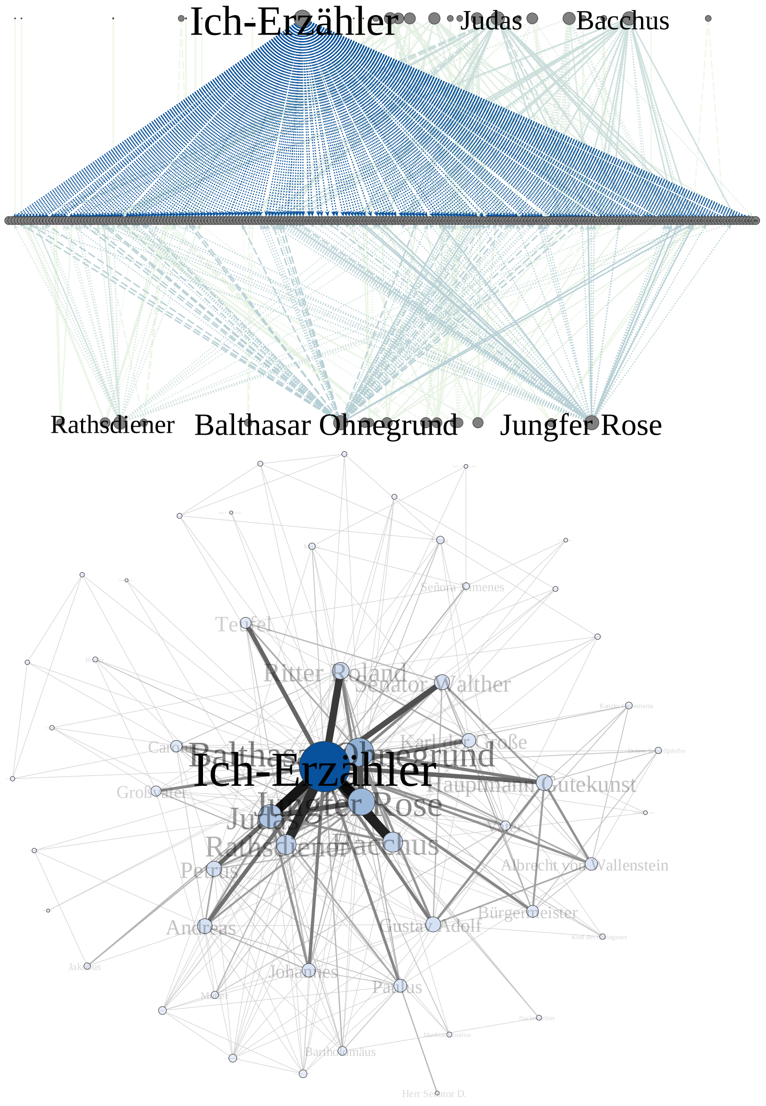
Halm
Die Marzipan-Lise



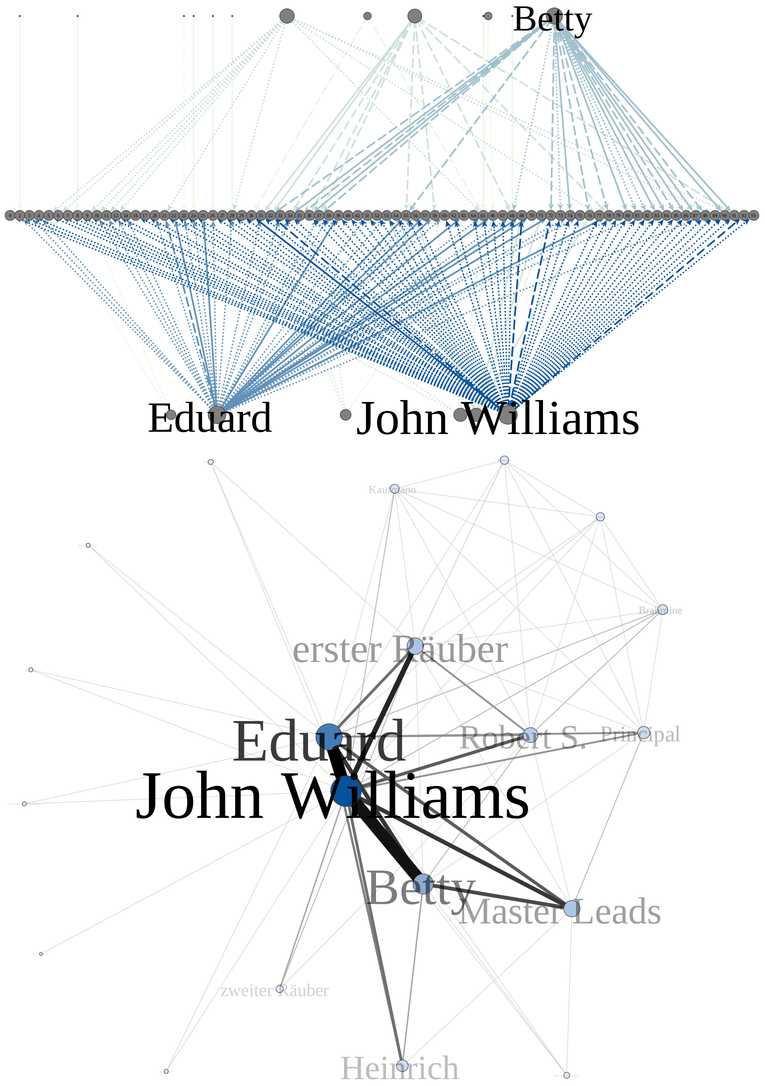
Hartmann
Das Schloß im Gebirge



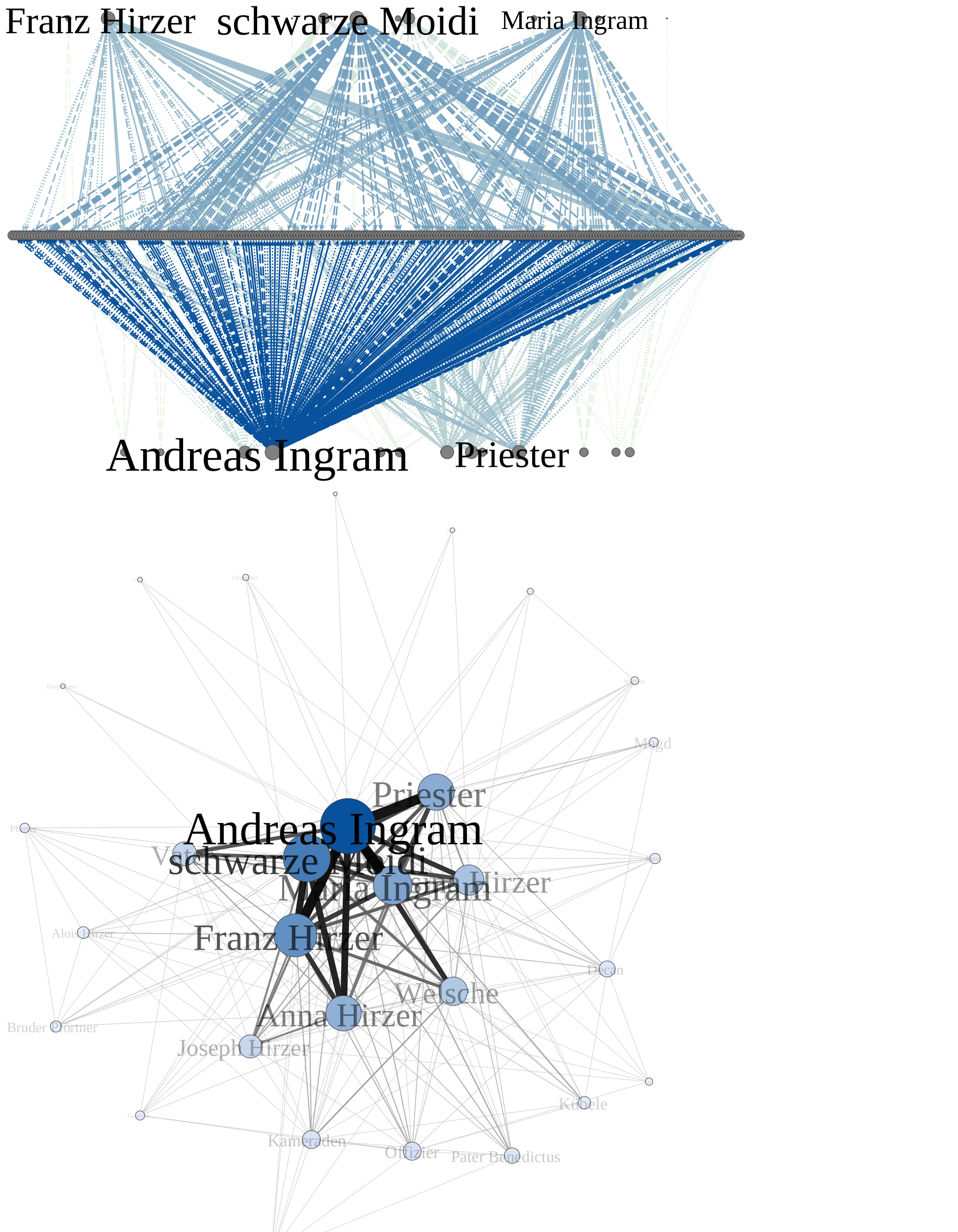
Hauff
Phantasien im Bremer Ratskeller



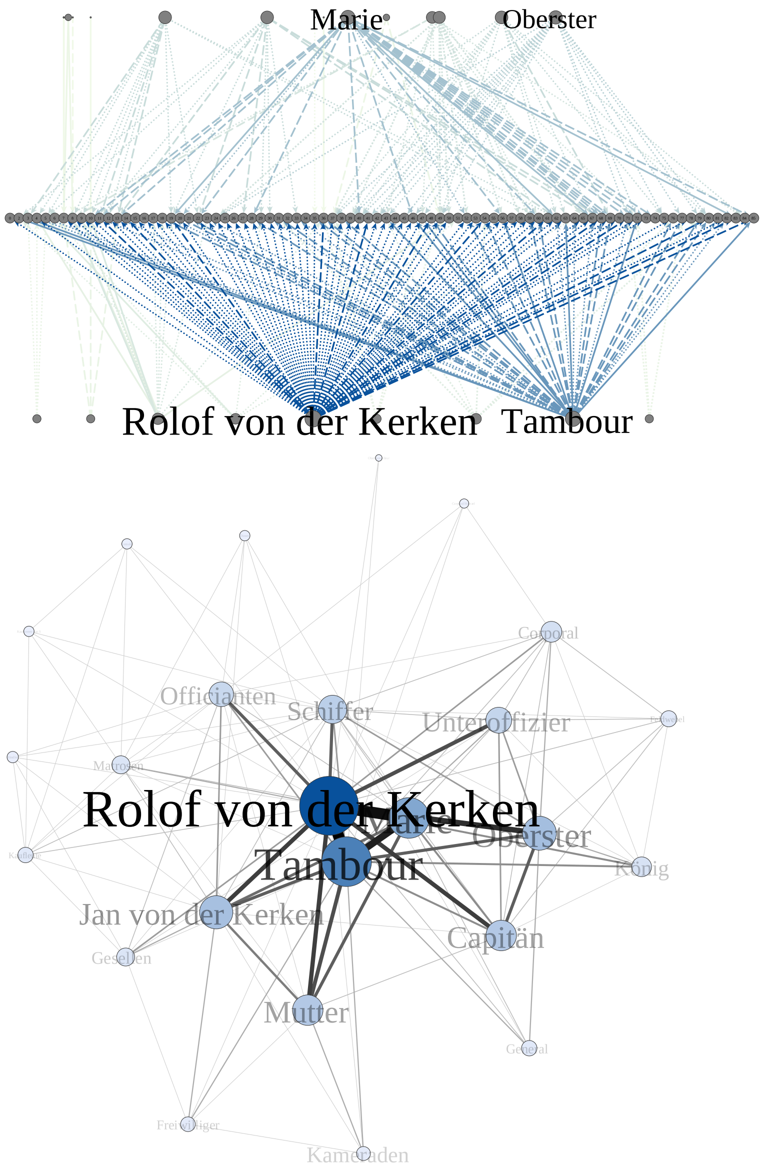
Heyden
Der graue John



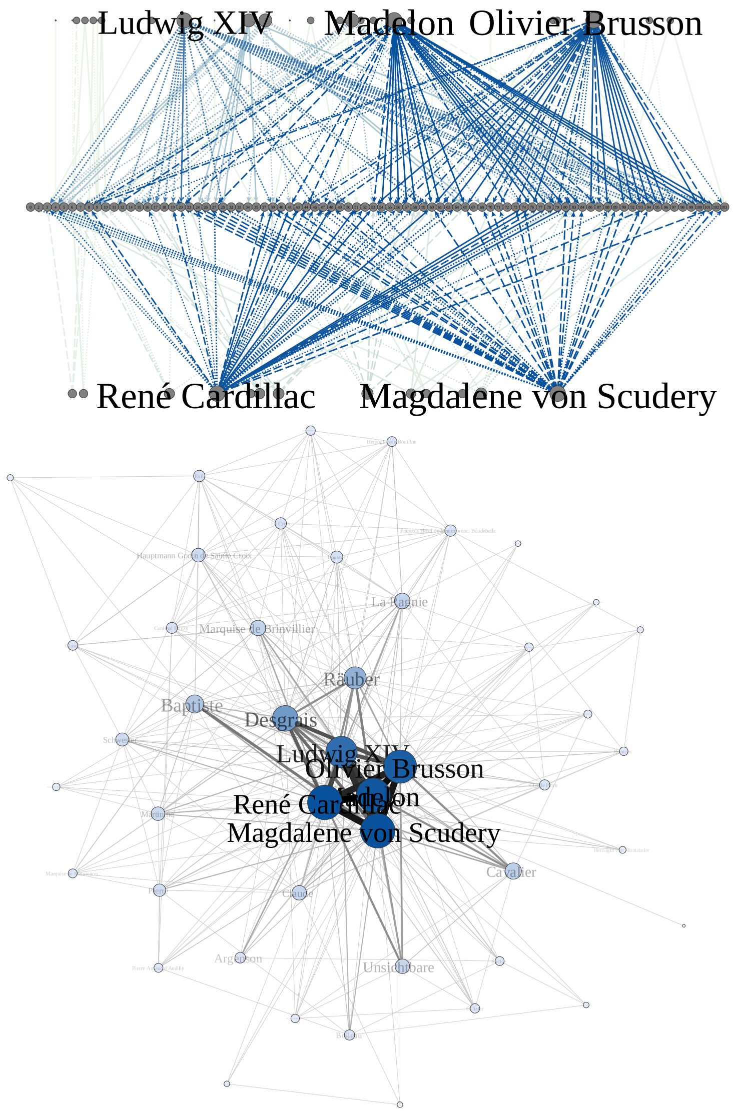
Heyse
Der Weinhüter von Meran



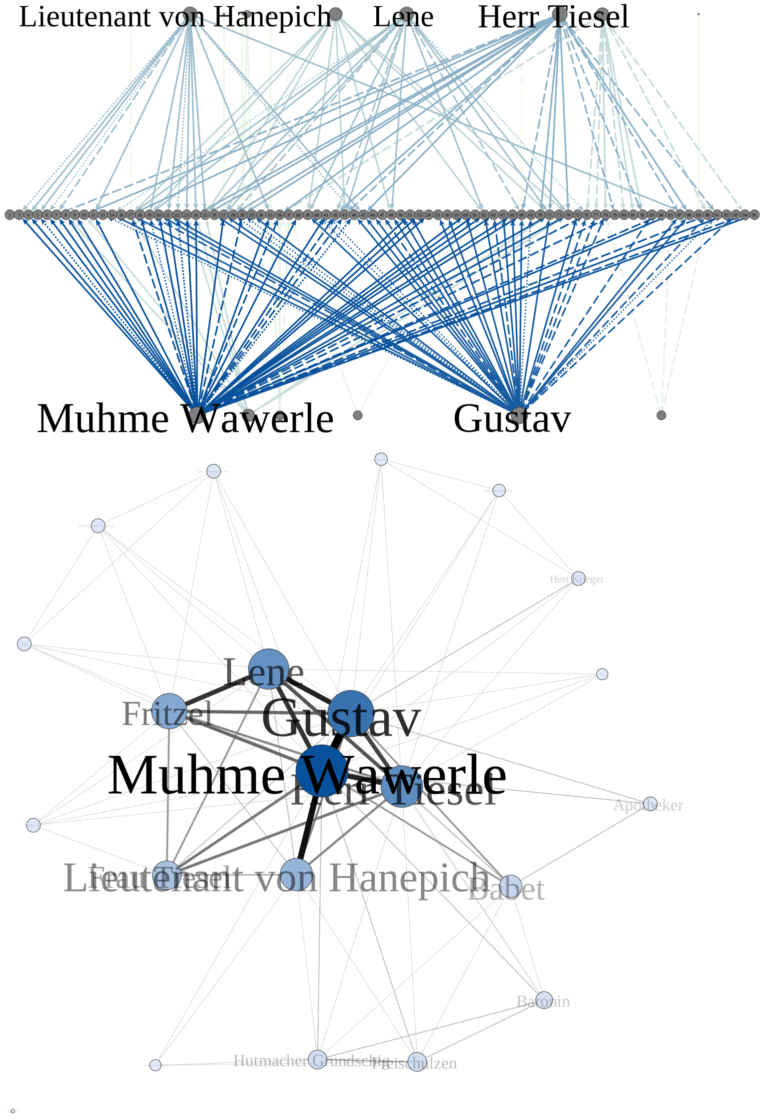
Höfer
Rolof, der Rekrut



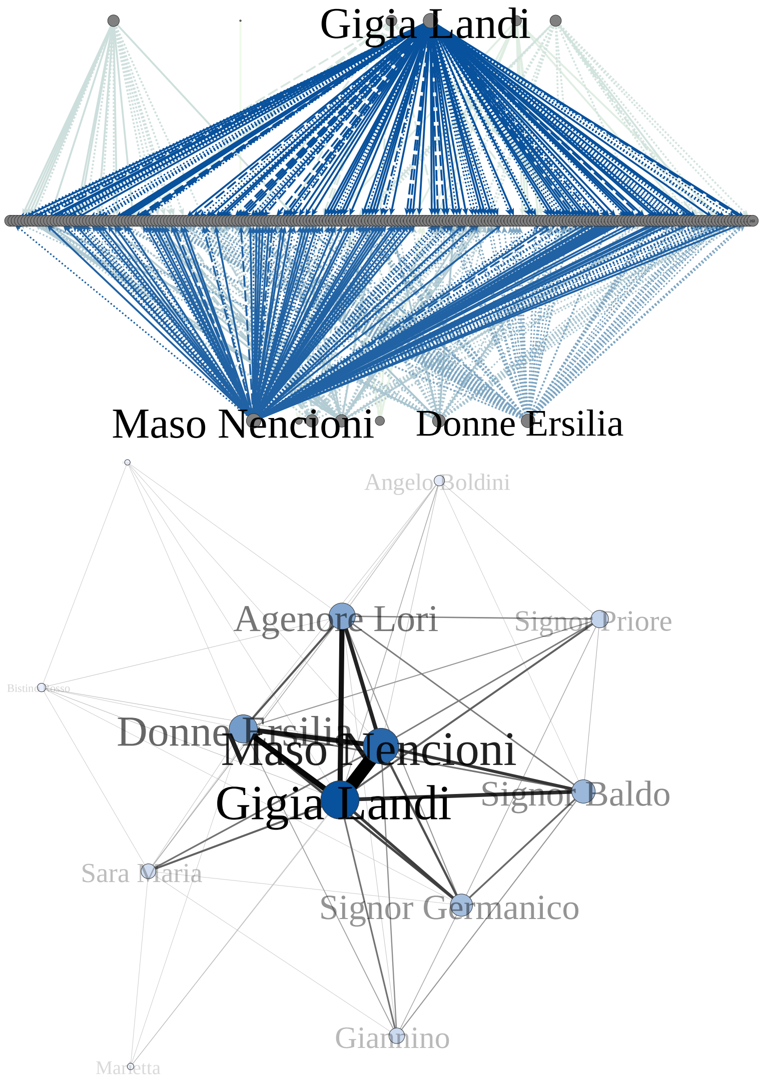
Hoffmann
Das Fräulein von Scuderi



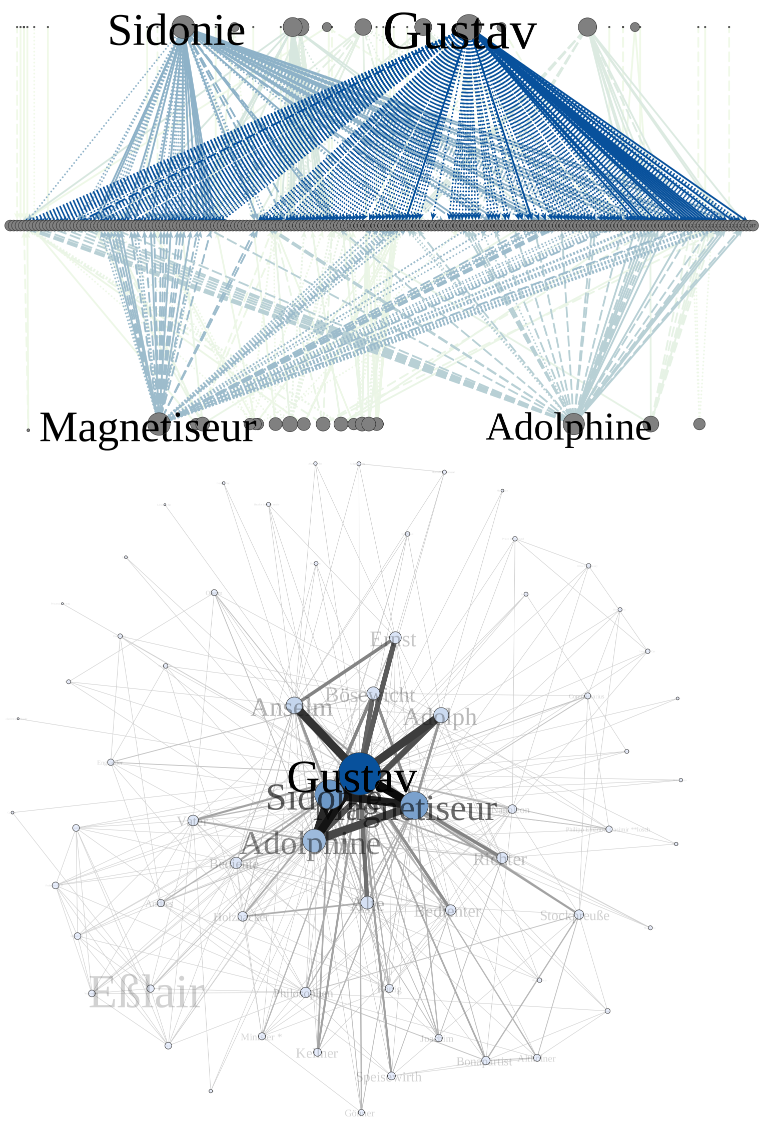
Holtei
s Muhme-Leutnant-Saloppel



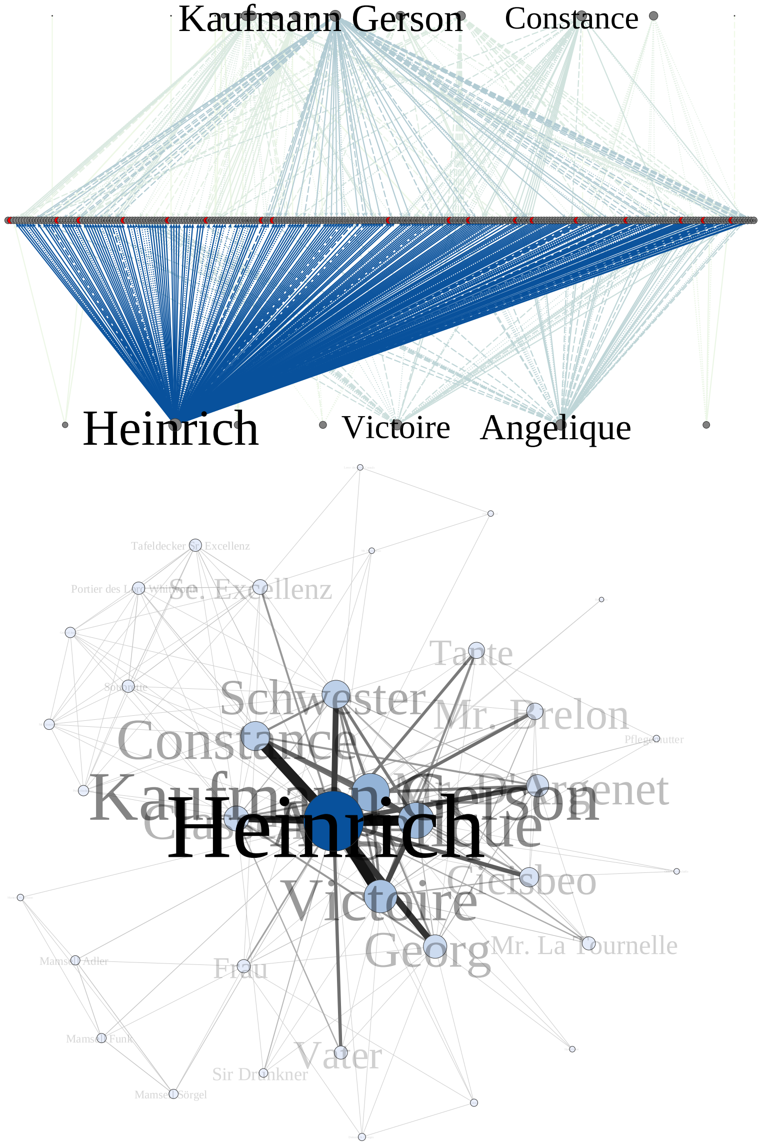
Horner
Der Säugling



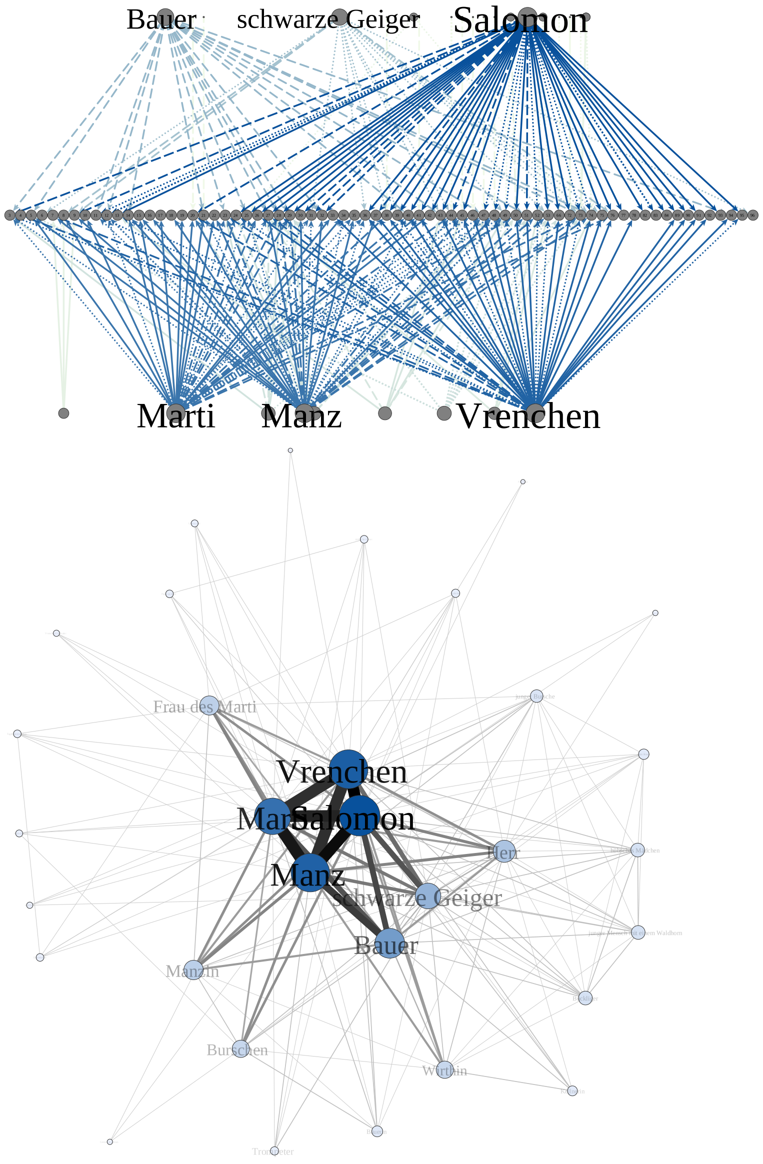
Immermann
Der Carneval und die Somnambule



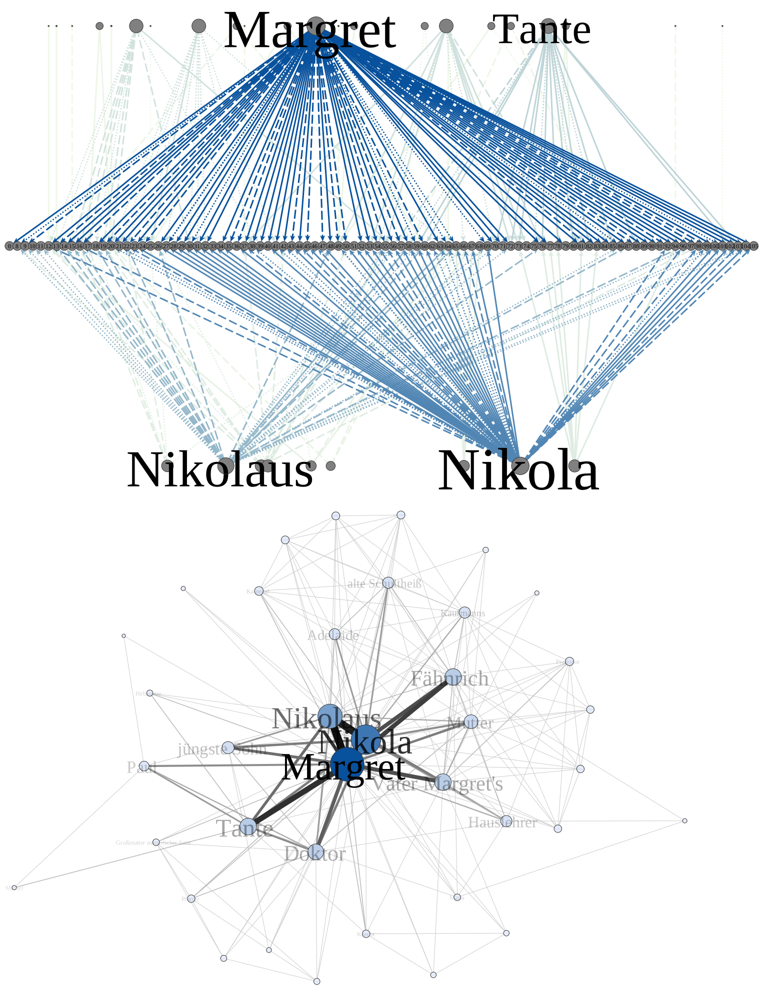
Kähler
Die drei Schwestern



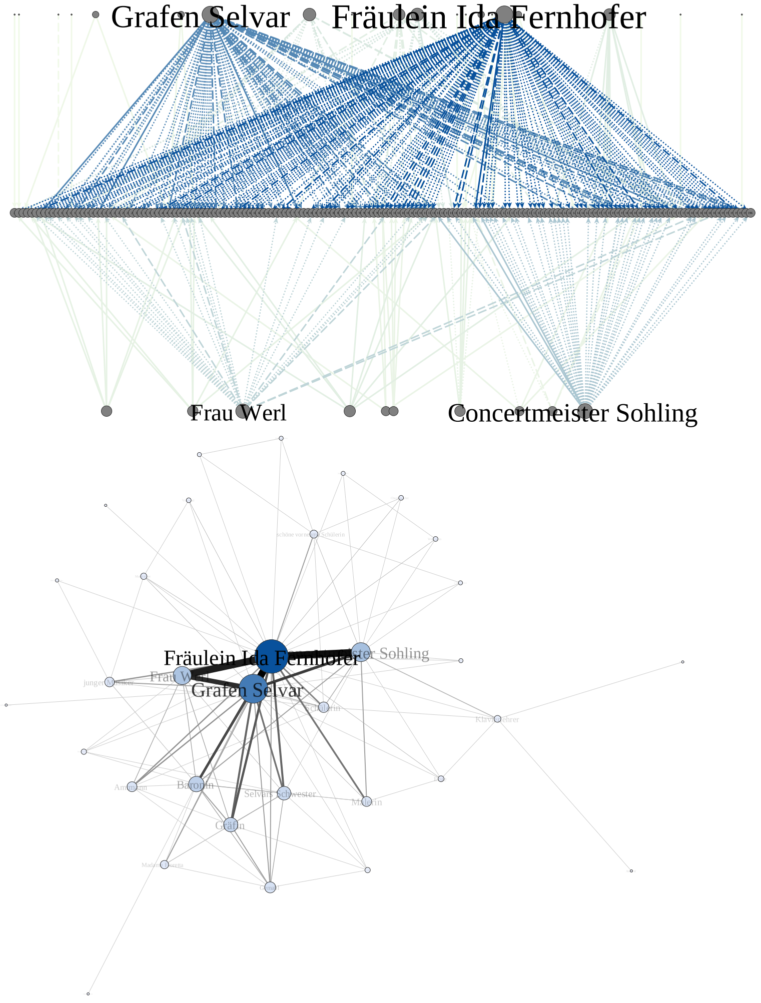
Keller
Romeo und Julia auf dem Dorfe



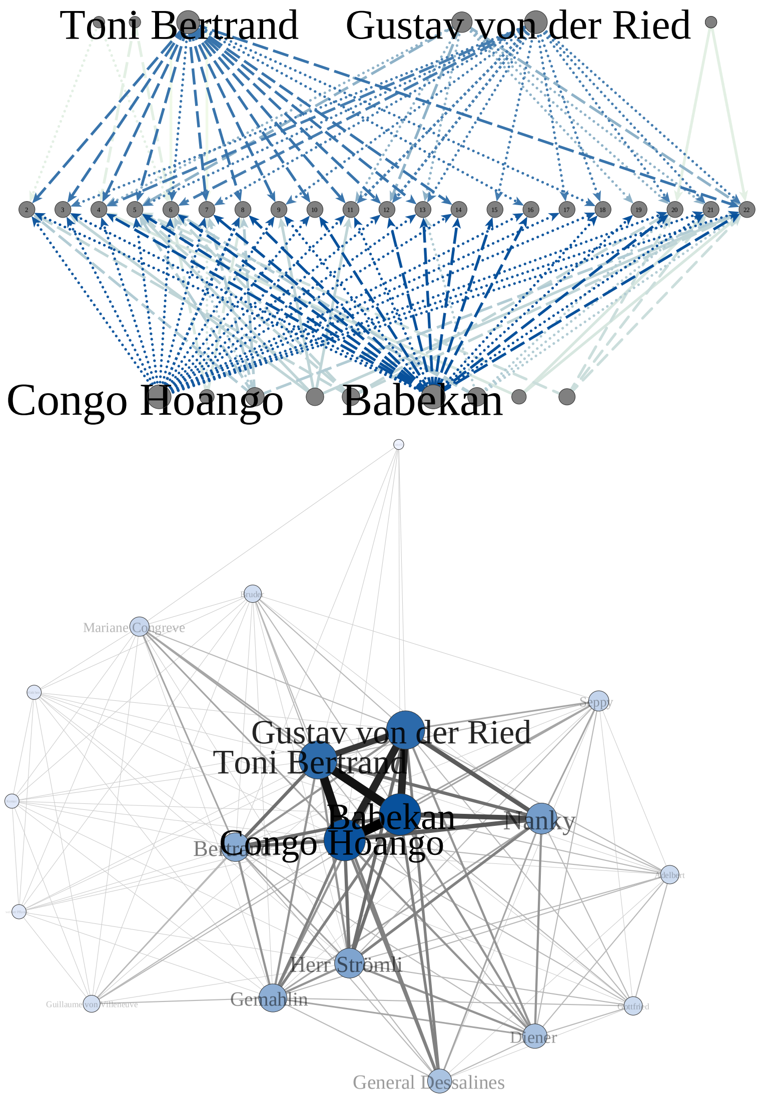
Kinkel G.
Margret



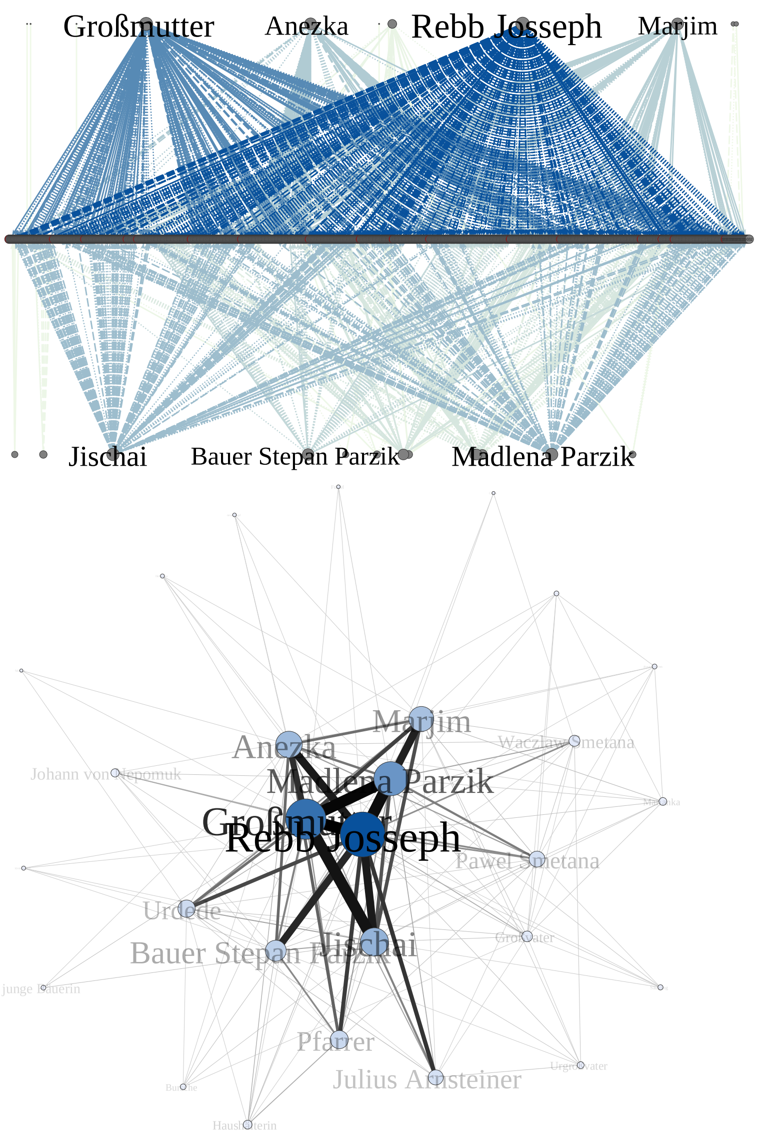
Kinkel J.
Musikalische Orthodoxie



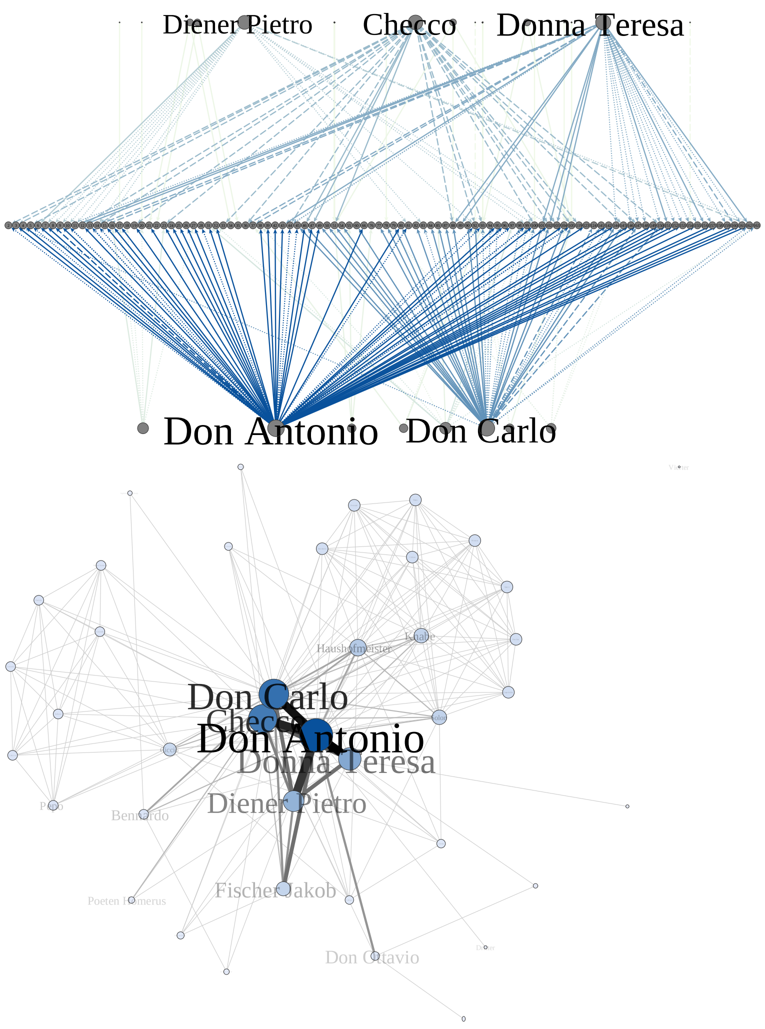
Kleist
Die Verlobung von St. Domingo



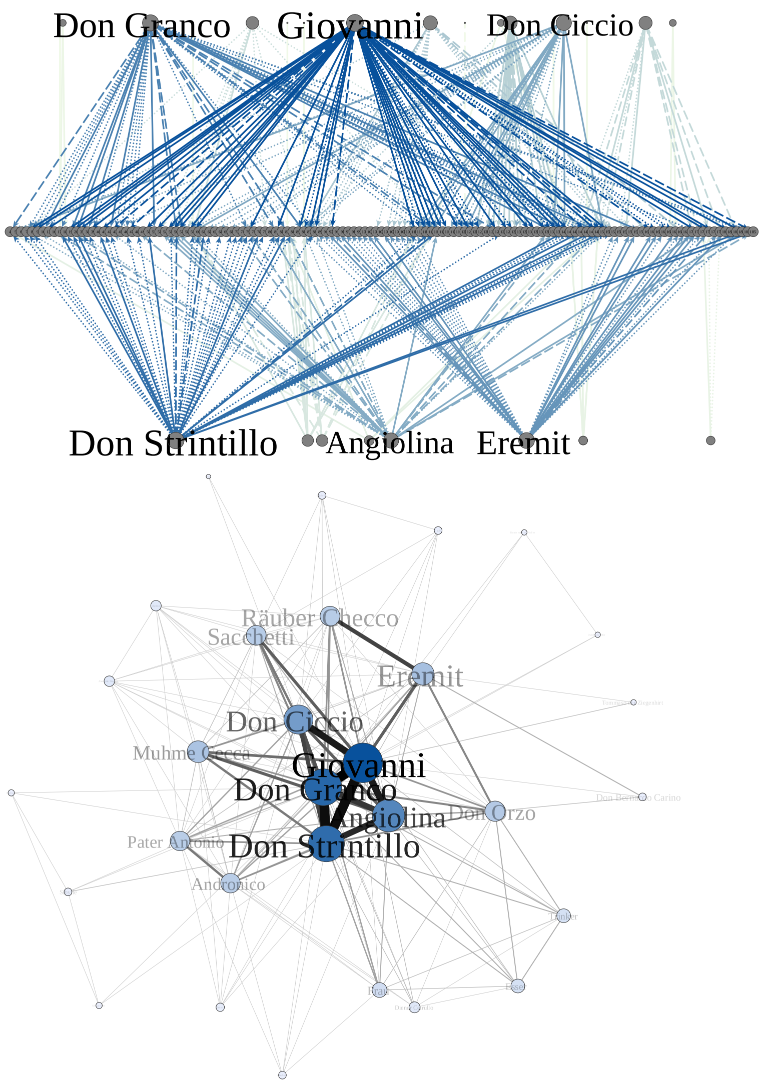
Kompert
Eine Verlorene



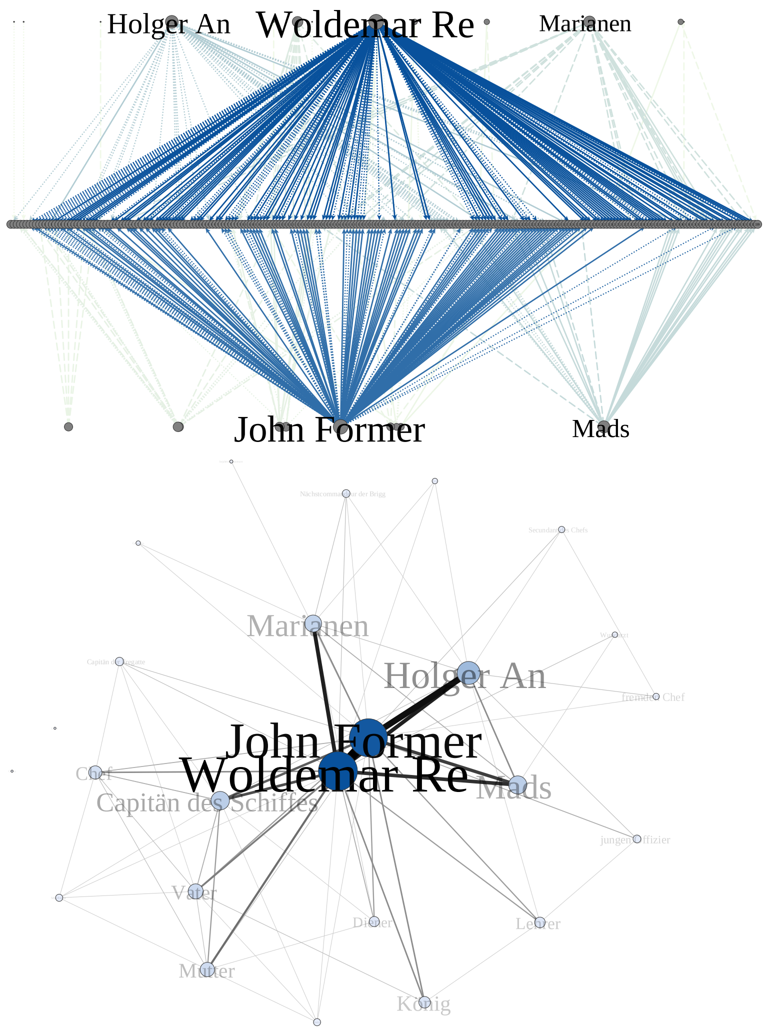
Kopisch
Ein Karnevalsfest auf Ischia



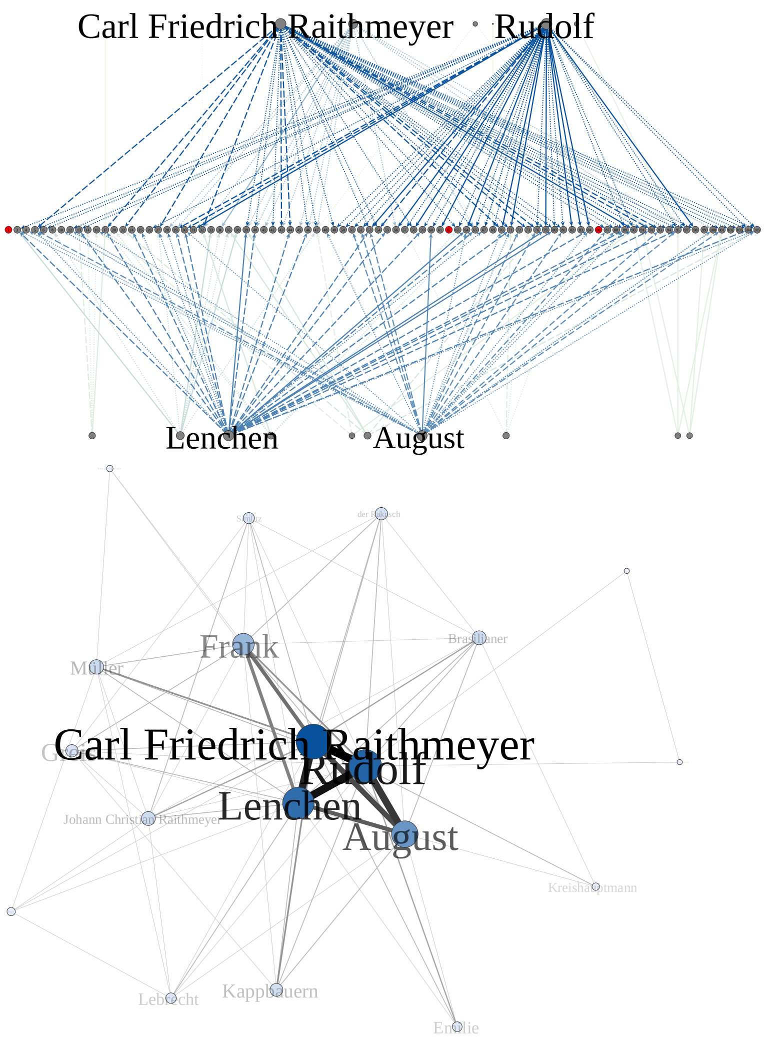
Kopisch
Der Träumer



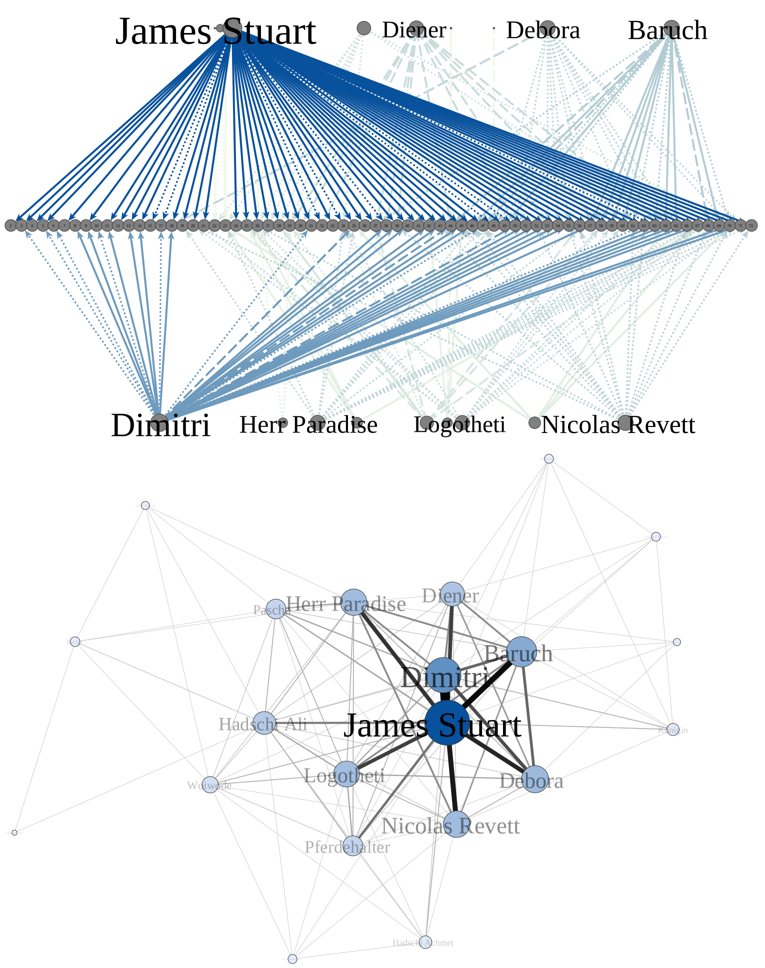
Kruse
Nordische Freundschaft



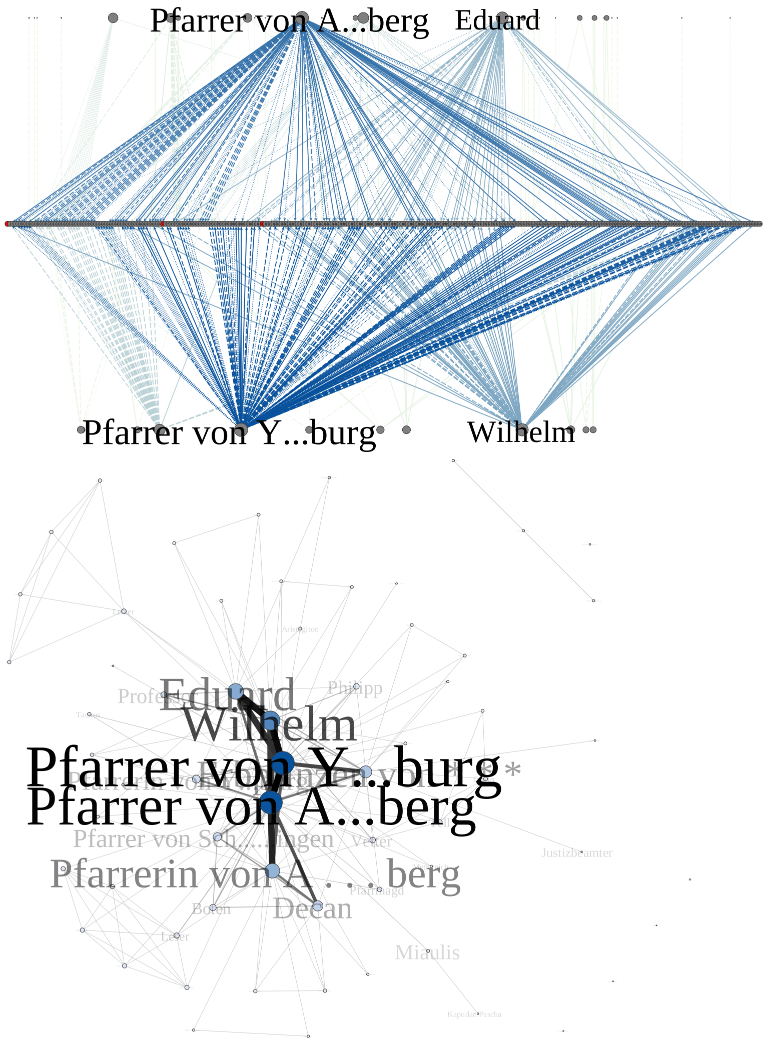
Kürnberger
Der Drache



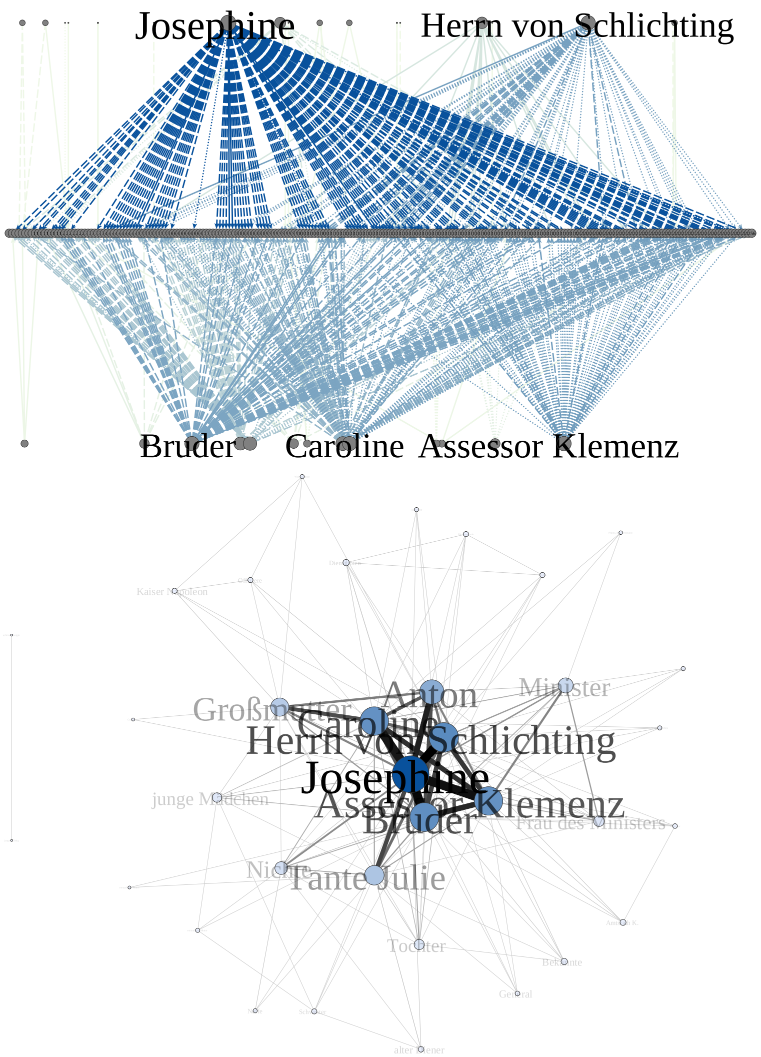
Kugler
Die Incantada



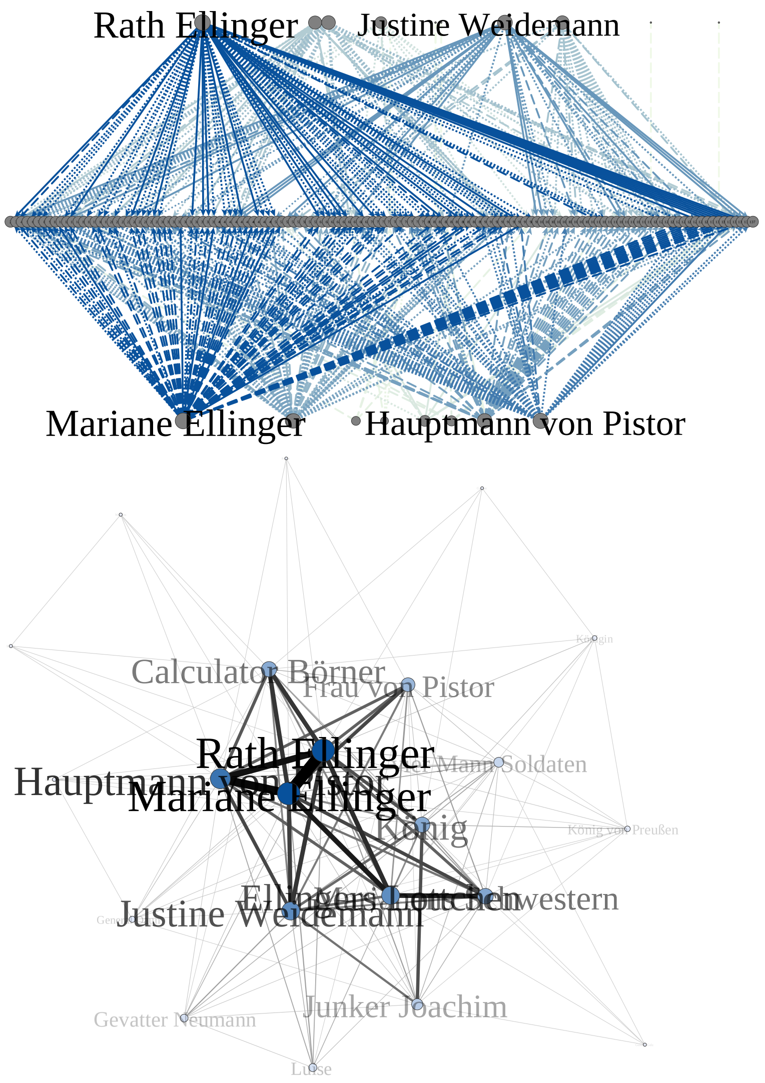
Kurz
Die beiden Tubus



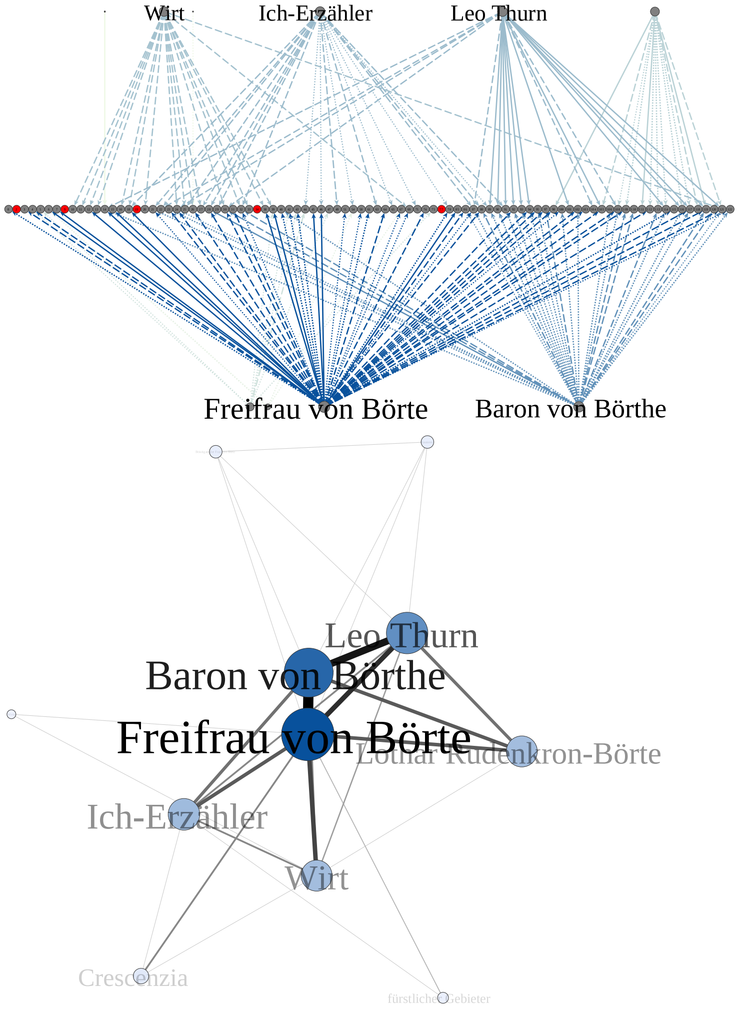
Lewald
Die Tante



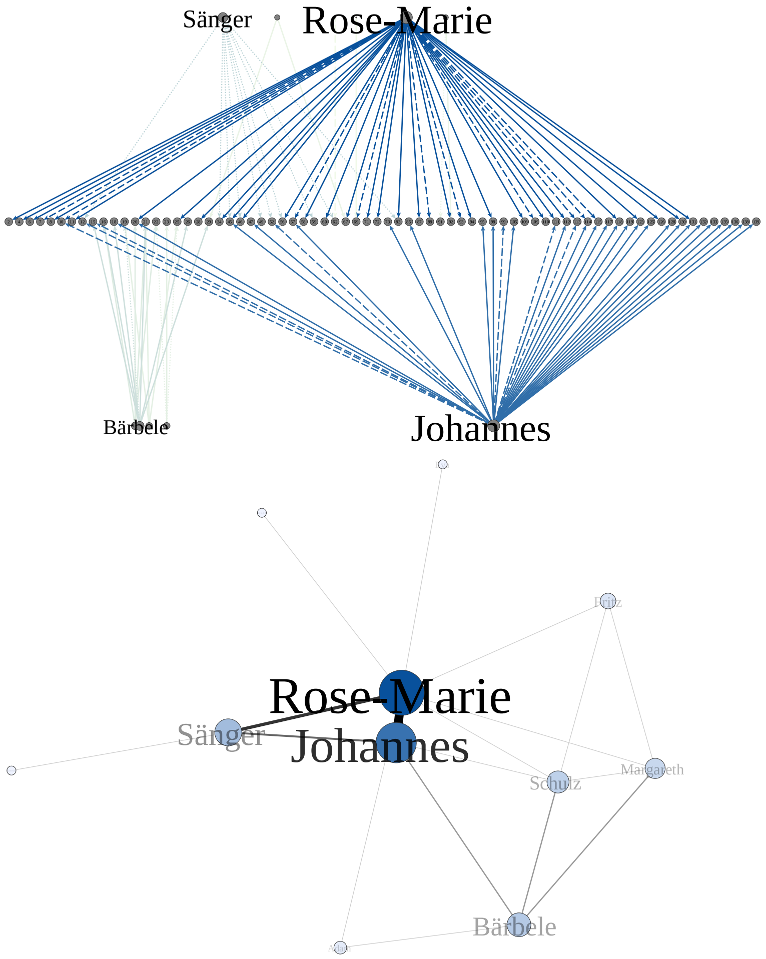
Lohmann
Die Entscheidung bei Hochkirch



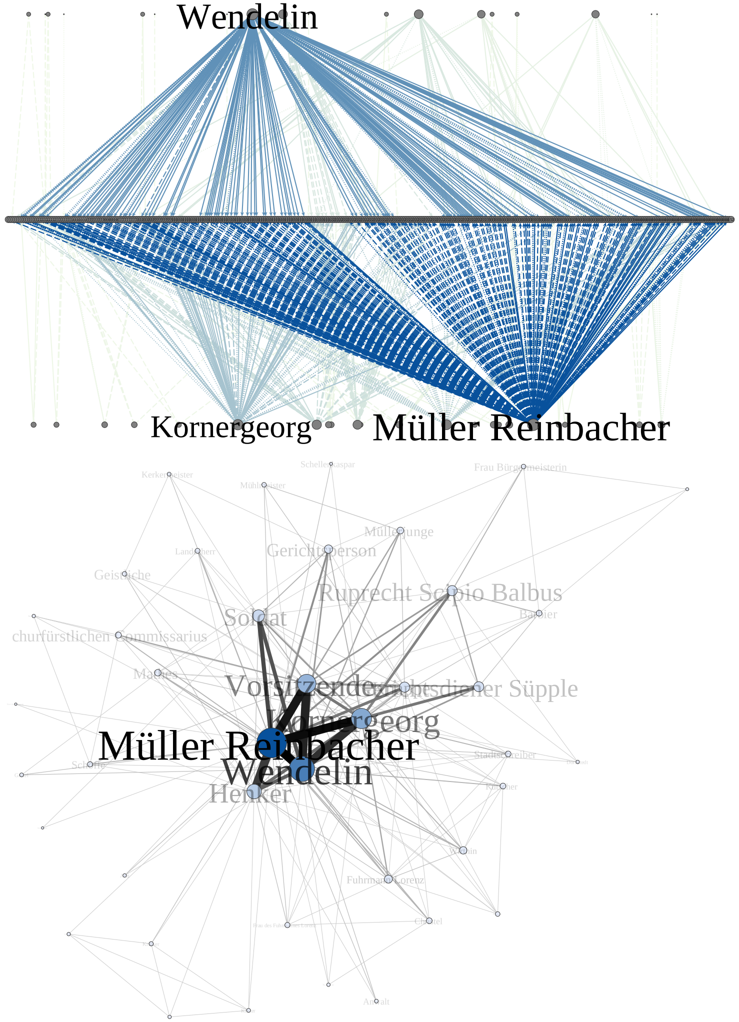
Lorm
Ein adeliges Fräulein

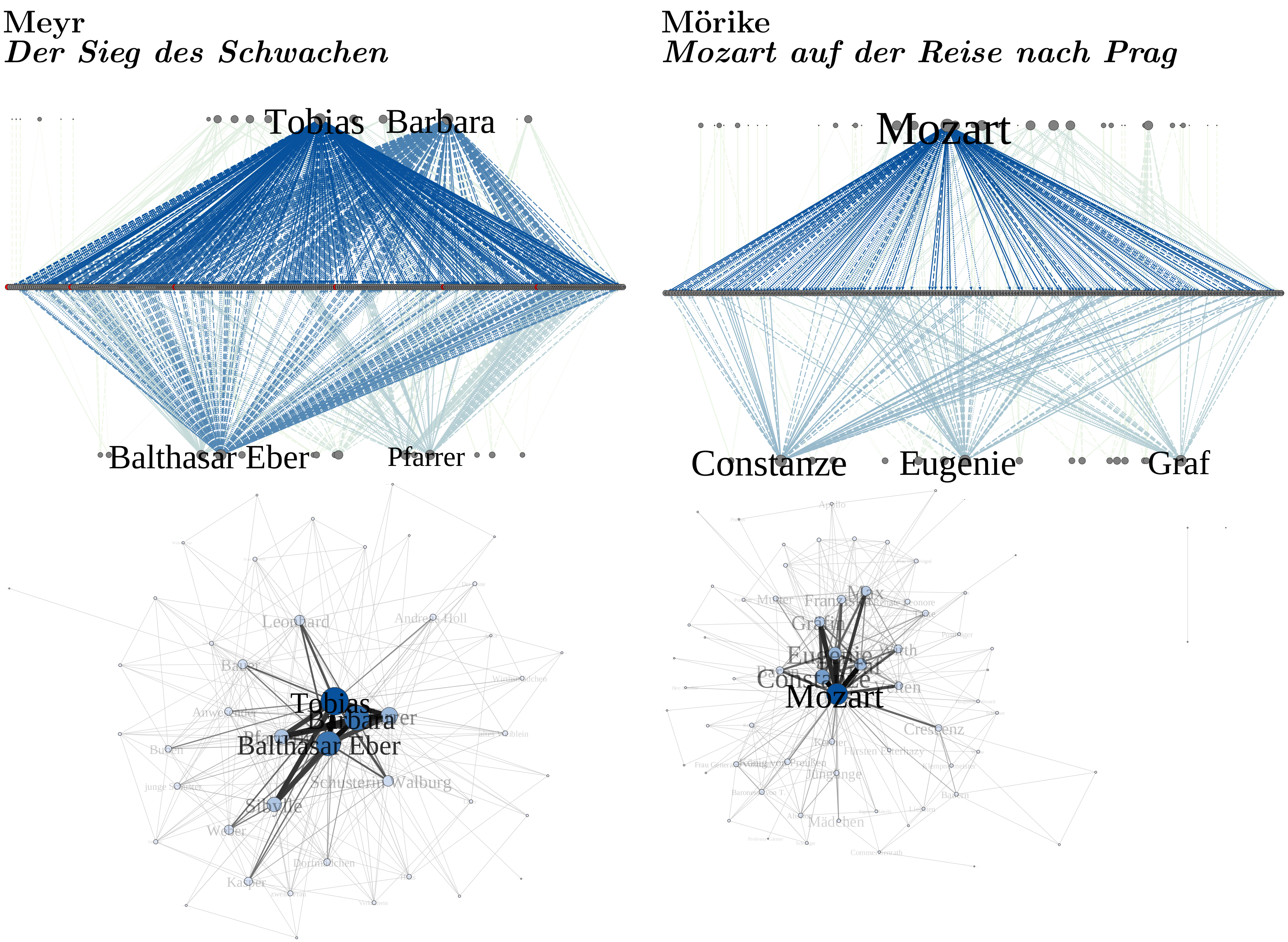


Ludwig
Das Gericht im Walde



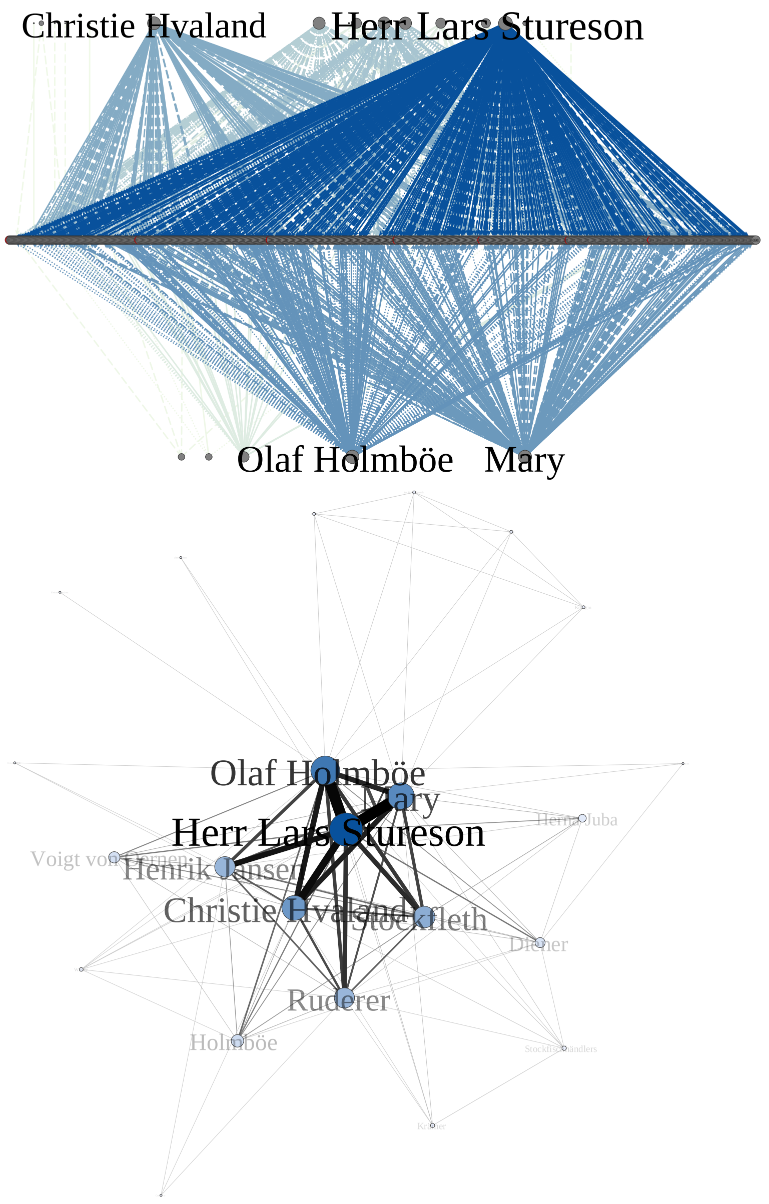
Meißner
Der Müller vom Höft



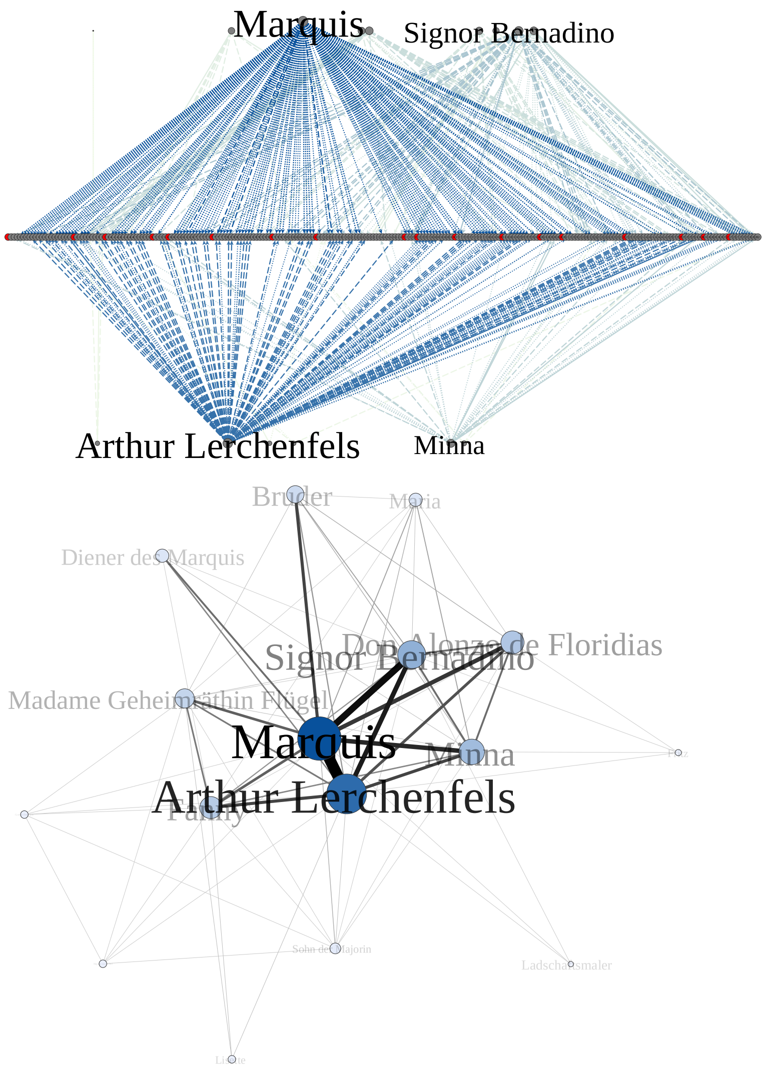

Meyr
Der Sieg des Schwachen

Mörike
Mozart auf der Reise nach Prag

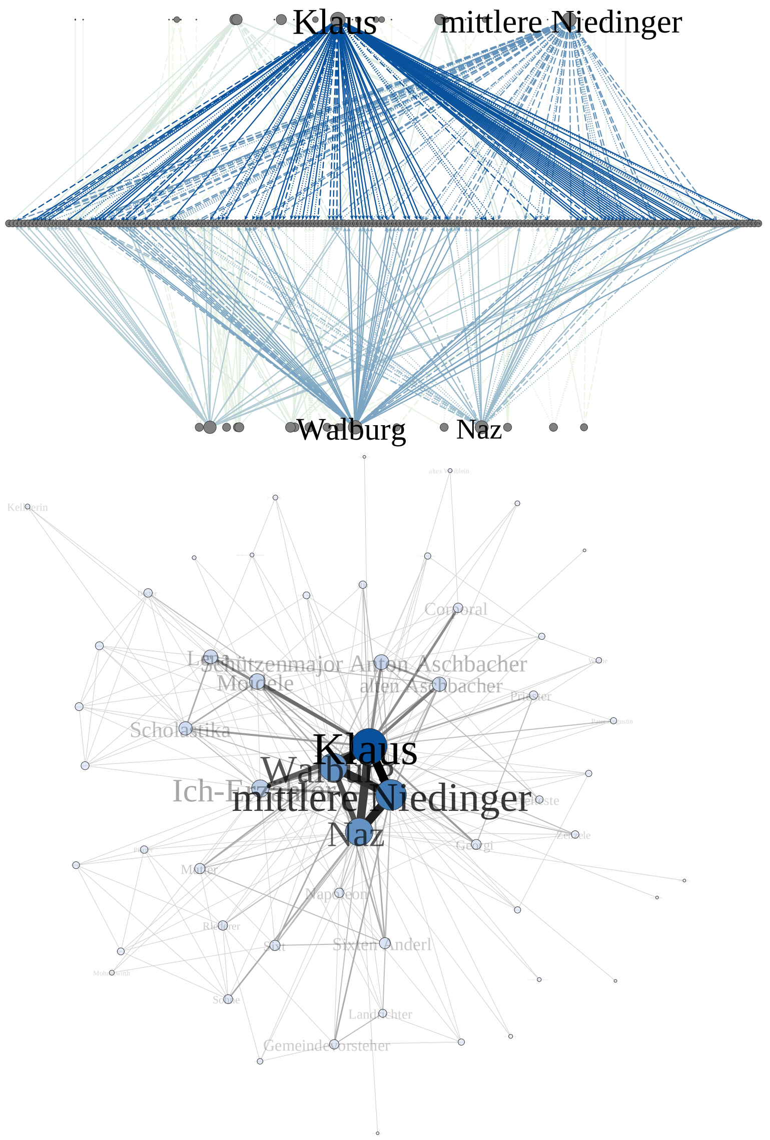
Mügge
Am Malanger Fjord



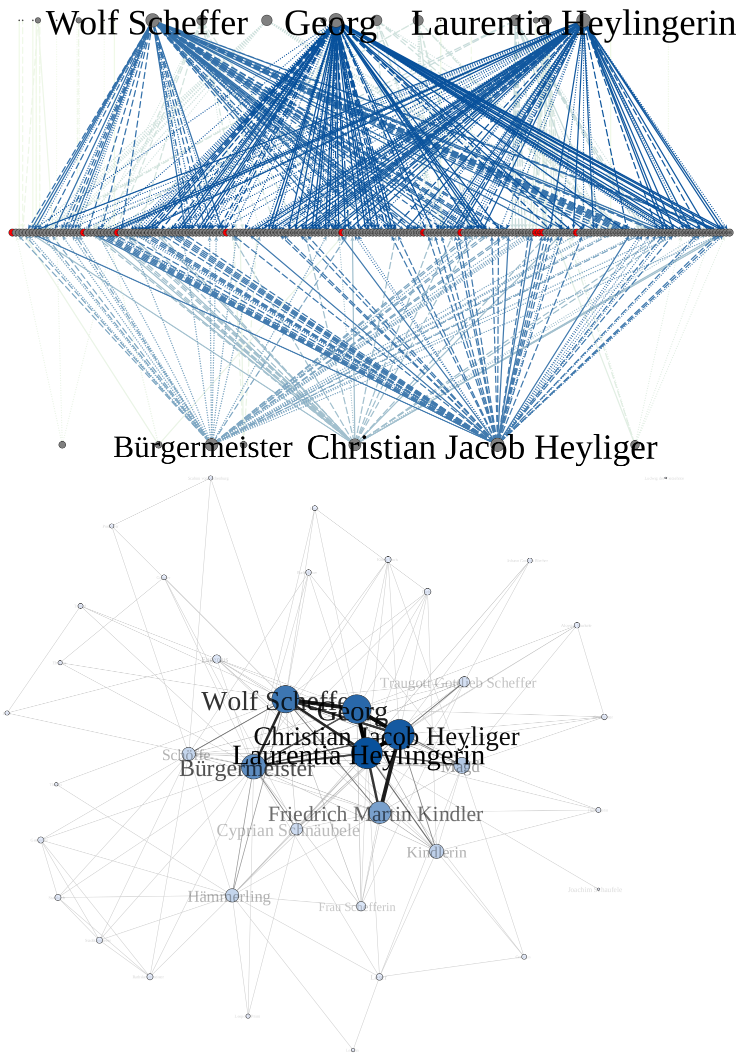
Müller
Debora



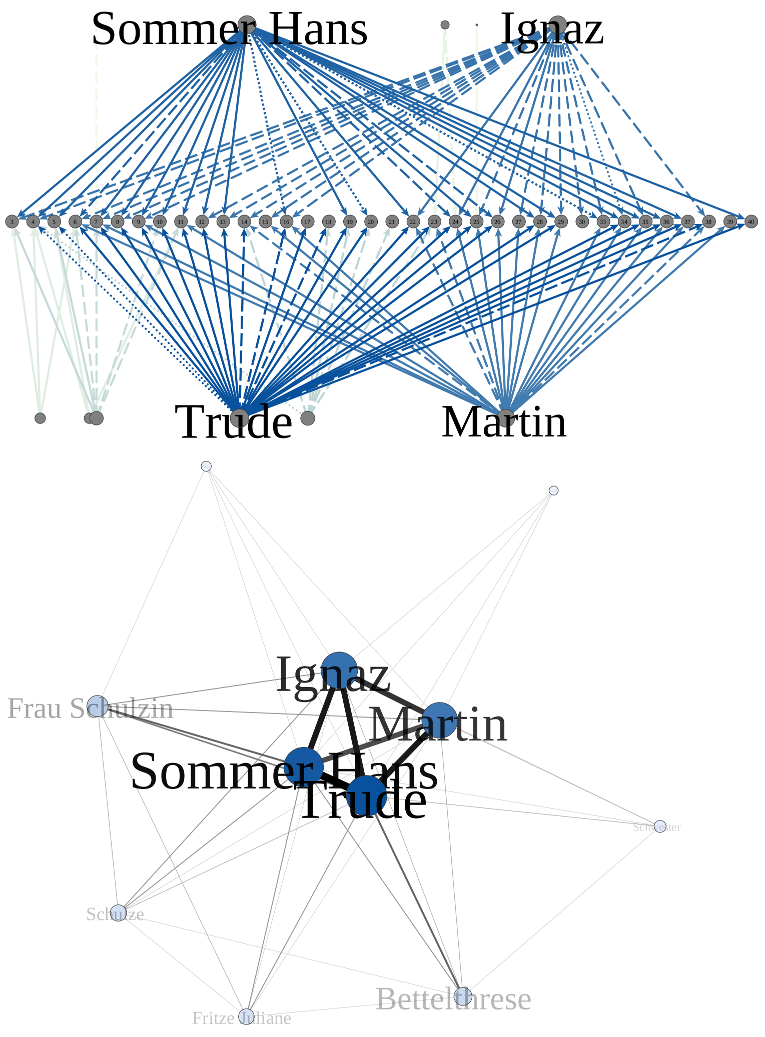
Pichler
Der Flüchtling



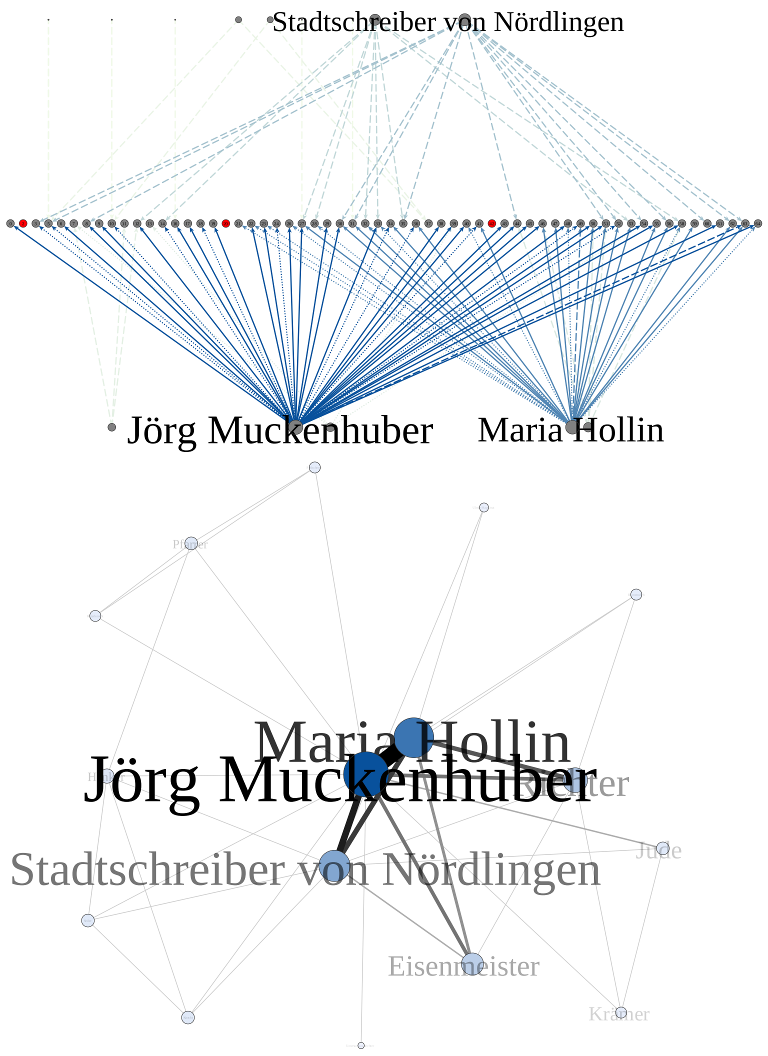
Raabe
Das letzte Recht



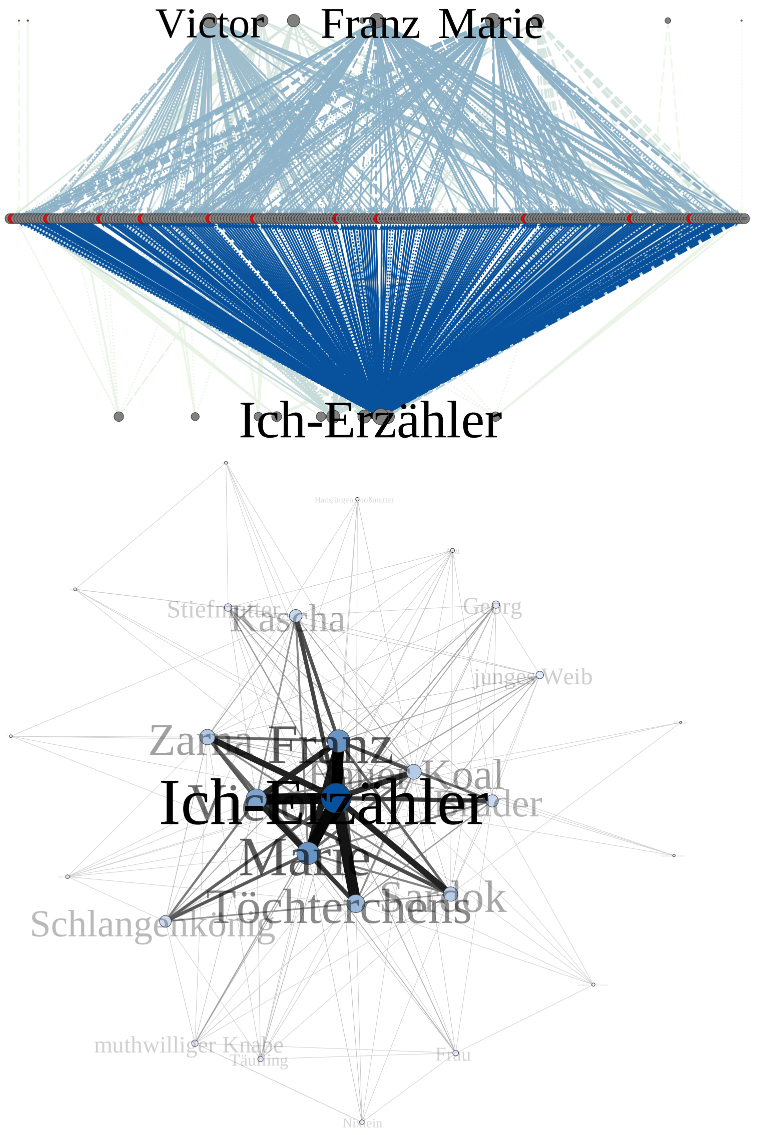
Reich
Mammon im Gebirge



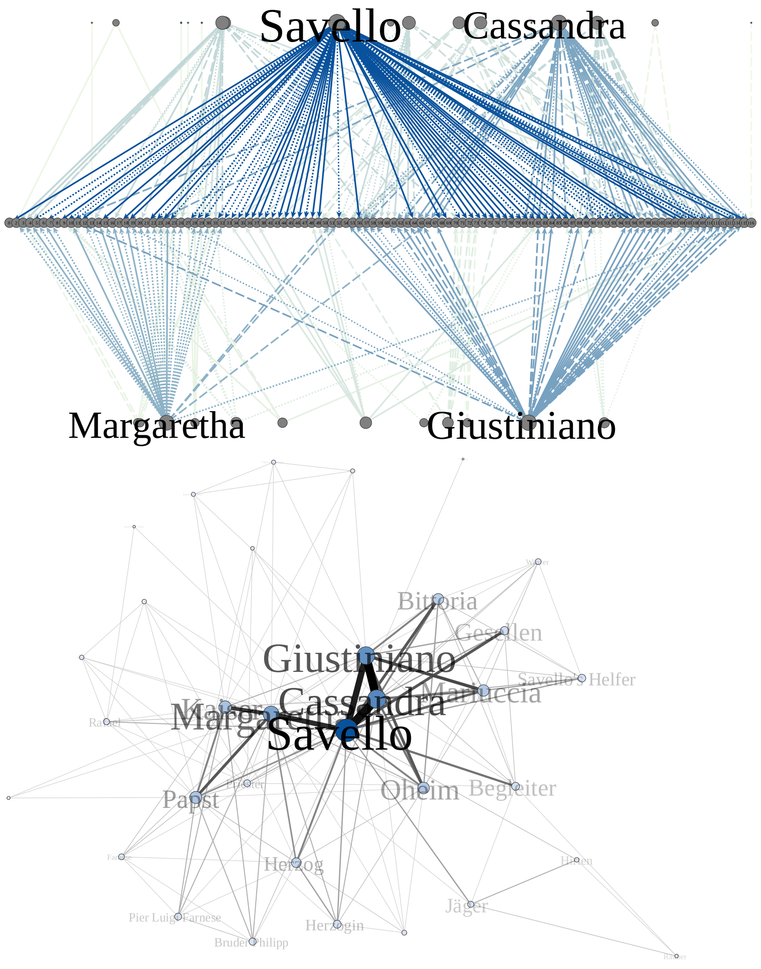
Riehl
Jörg Muckenhuber



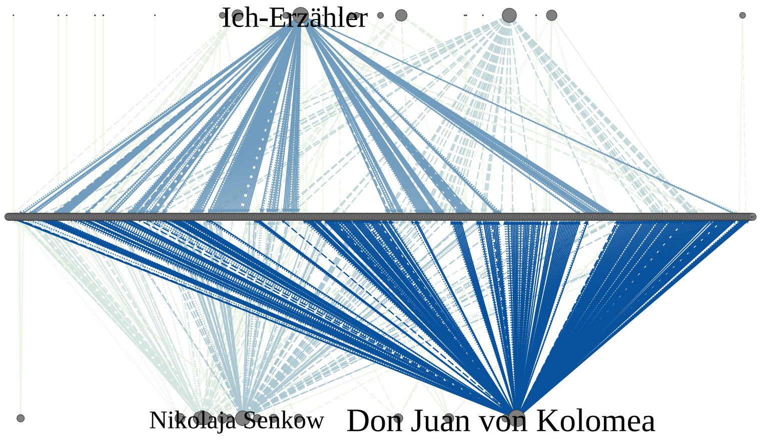
Roquette
Die Schlangenkönigin

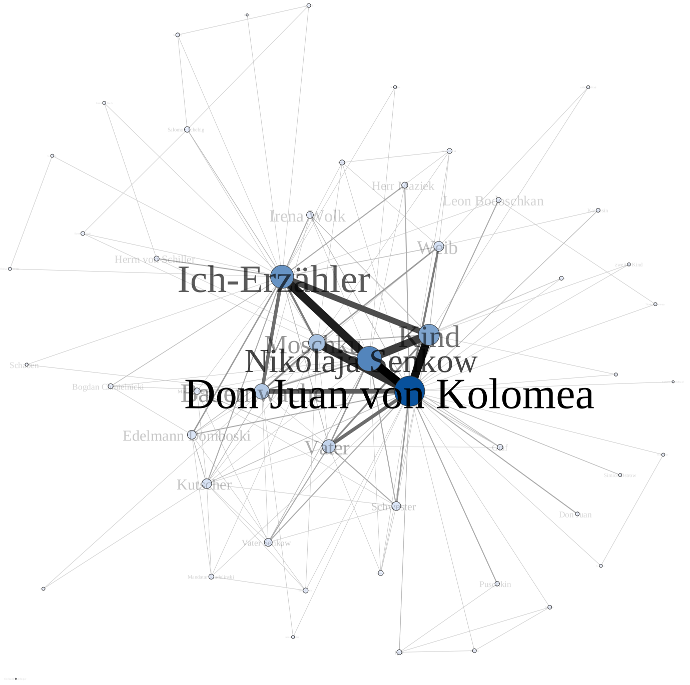


Rumohr
Der letzte Savello

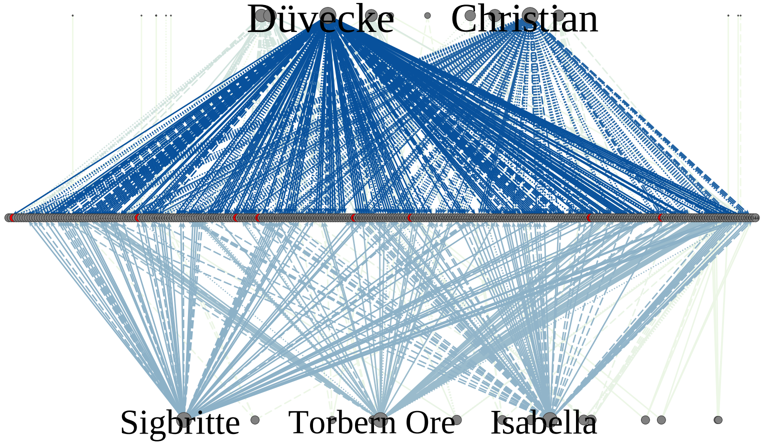


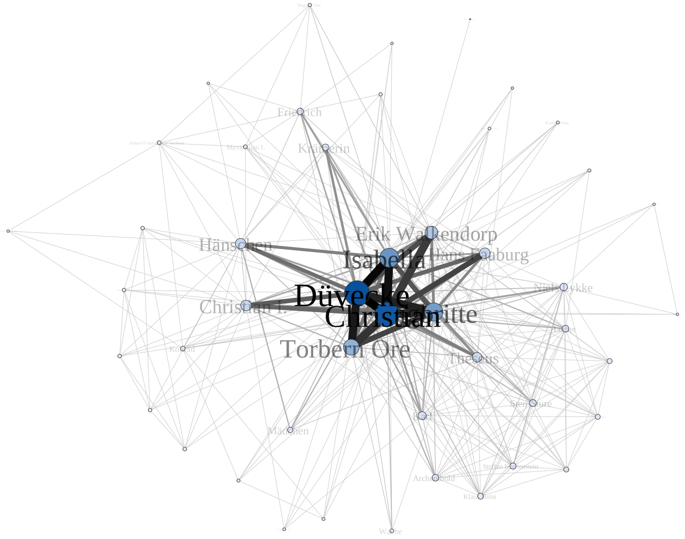
Sacher-Masoch
Don Juan von Kolomea



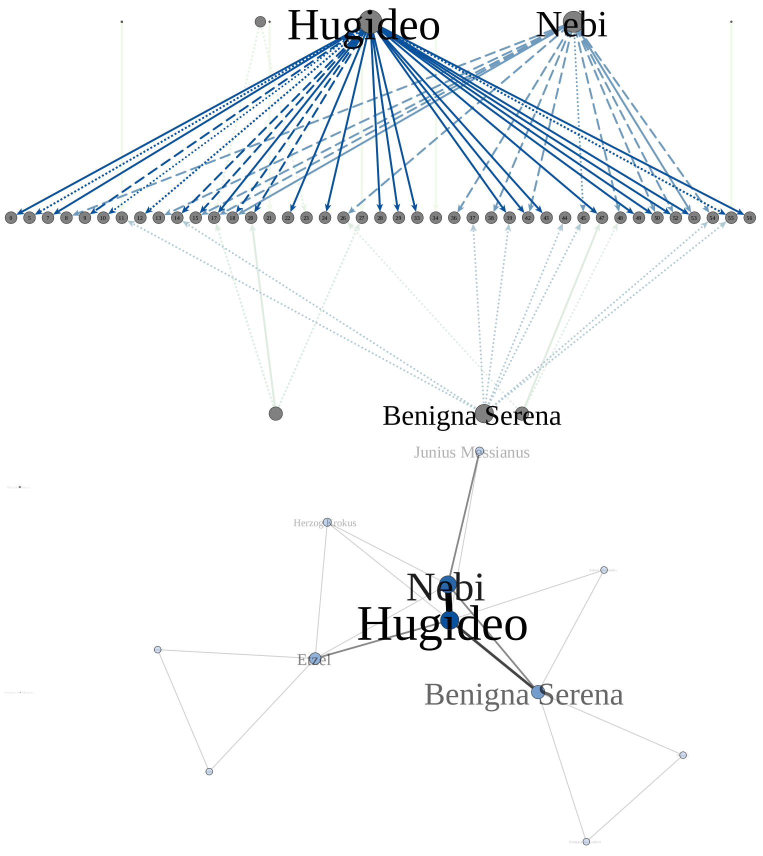


Schefer
Die Düvecke, oder die Leiden einer Königin

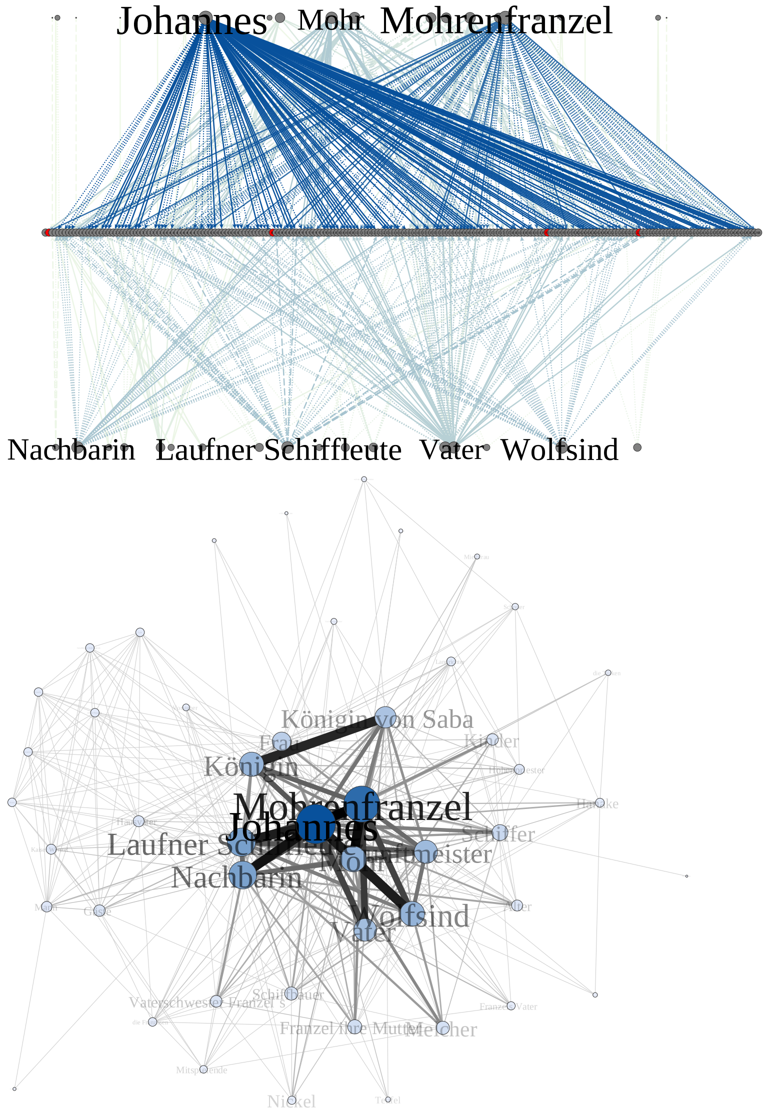




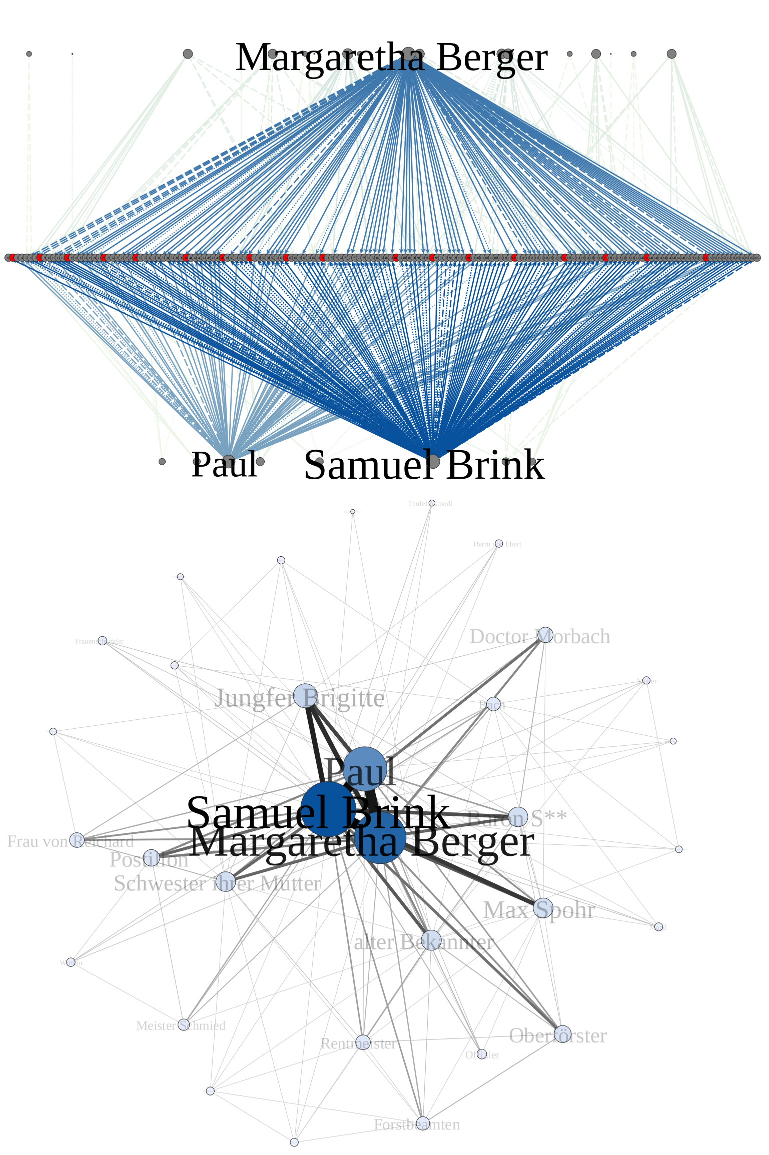
Scheffel
Hugideo



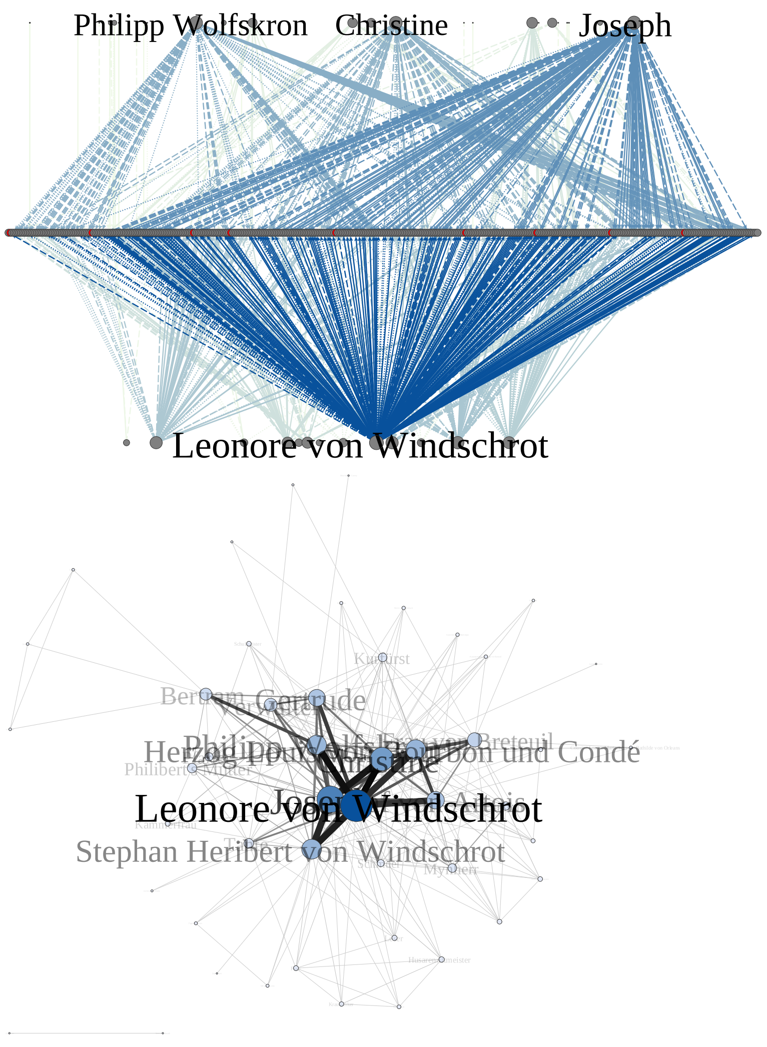
Schmid
Mohrenfranzel

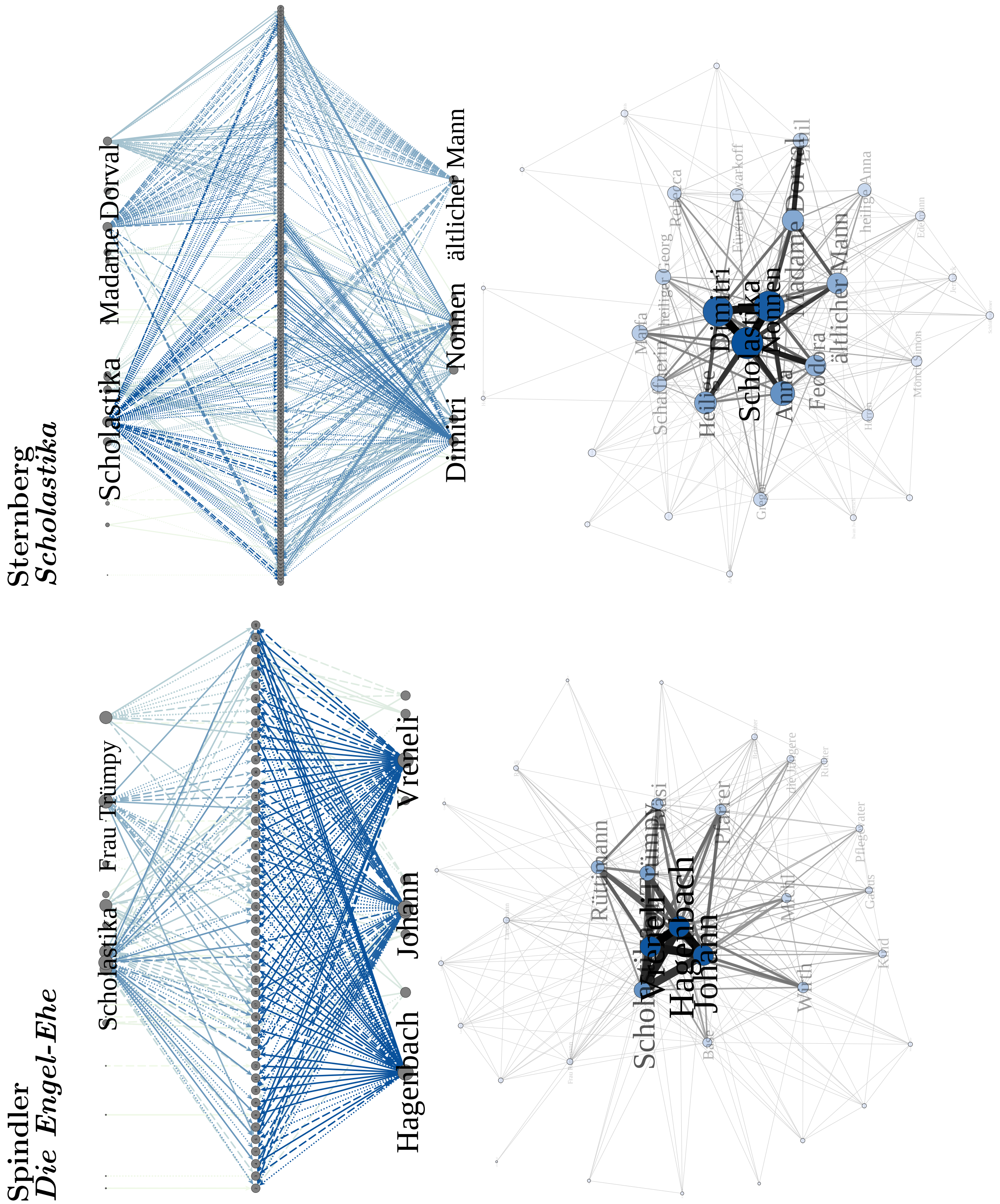


Schreyvogel
Samuel Brinks letzte Liebesgeschichte



Schücking
Die Schwester

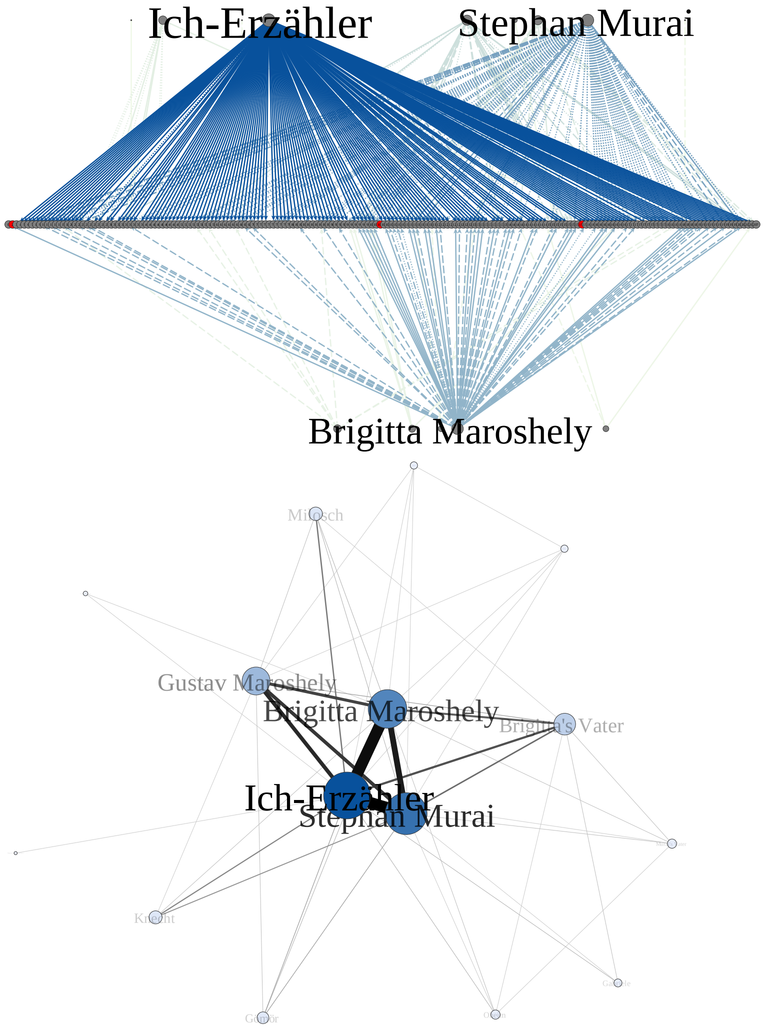




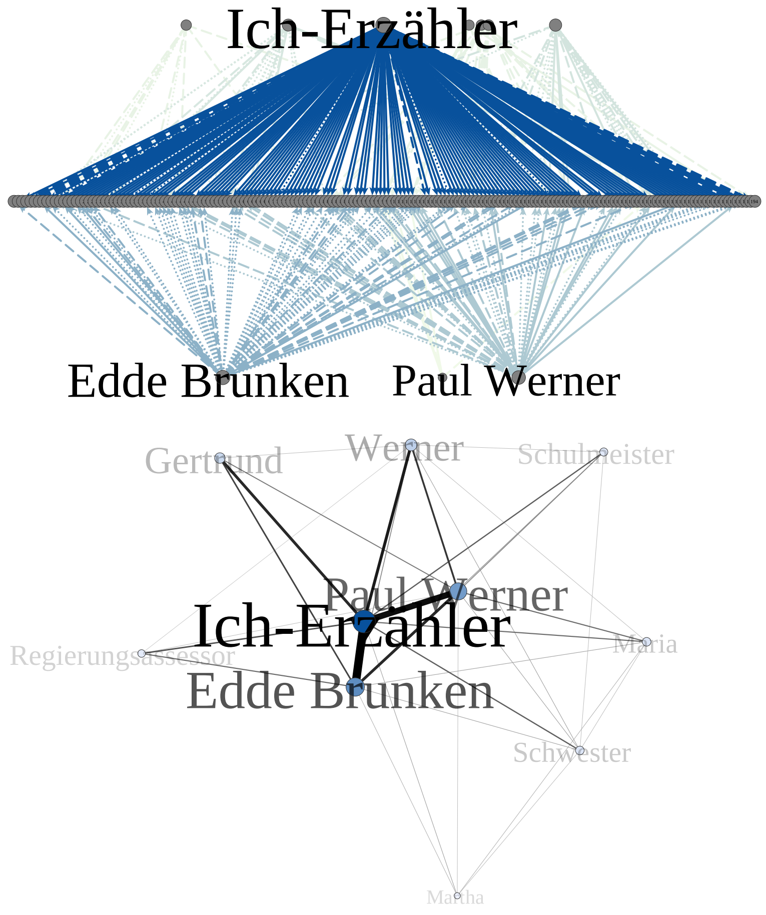
Sternberg *Scholastika*

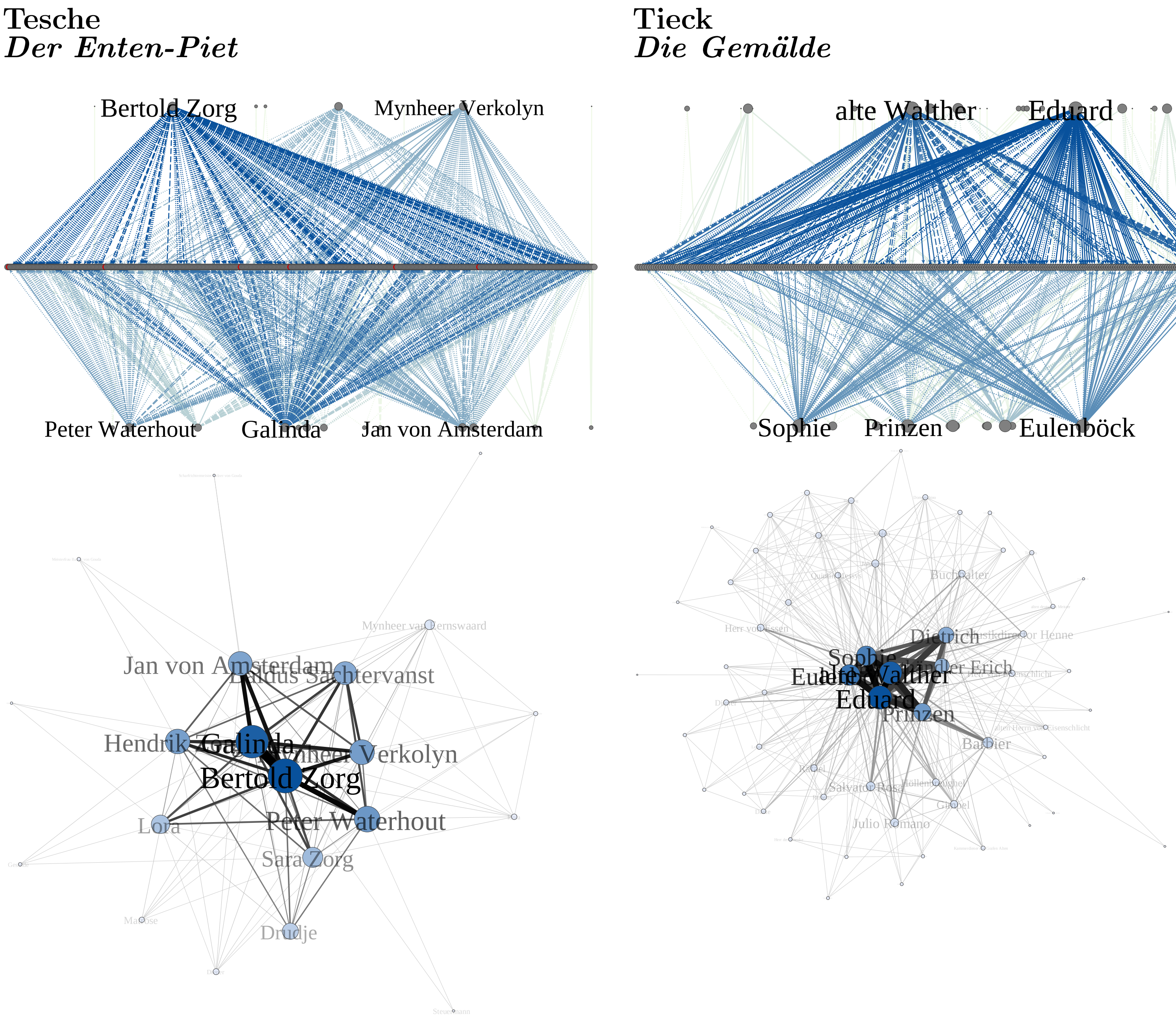
Spindler *Die Engel-Ehe*

Stifter
Brigitta

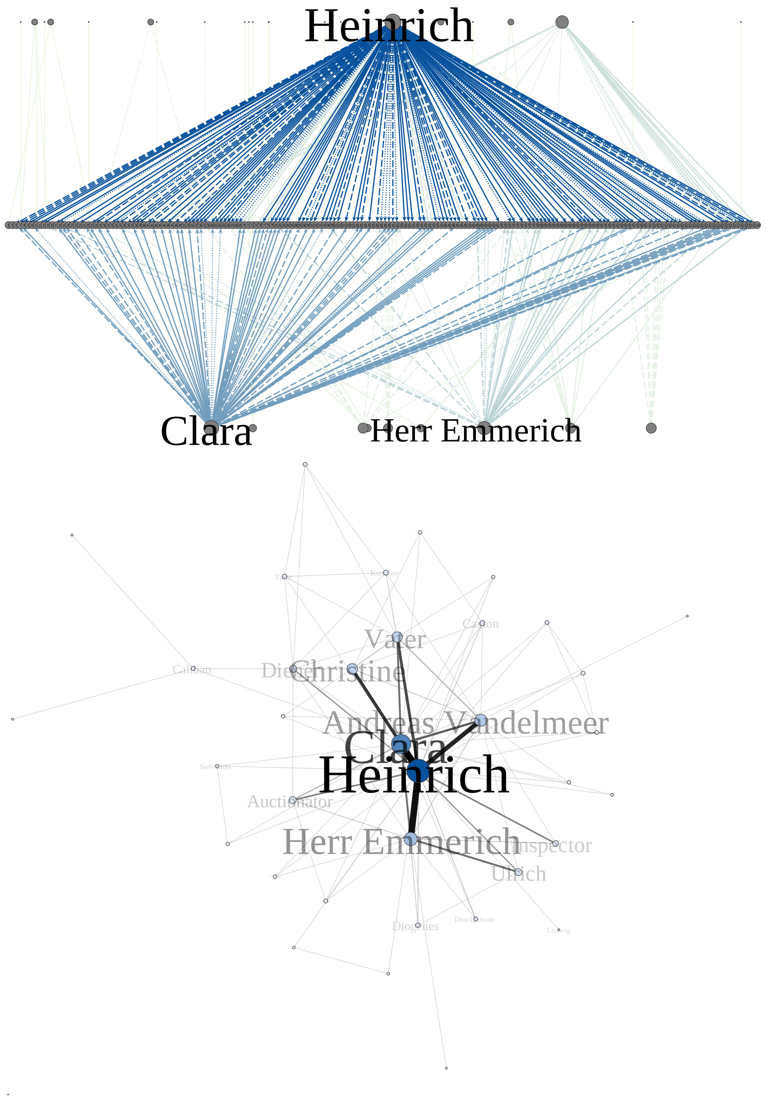


Storm
Eine Malerarbeit

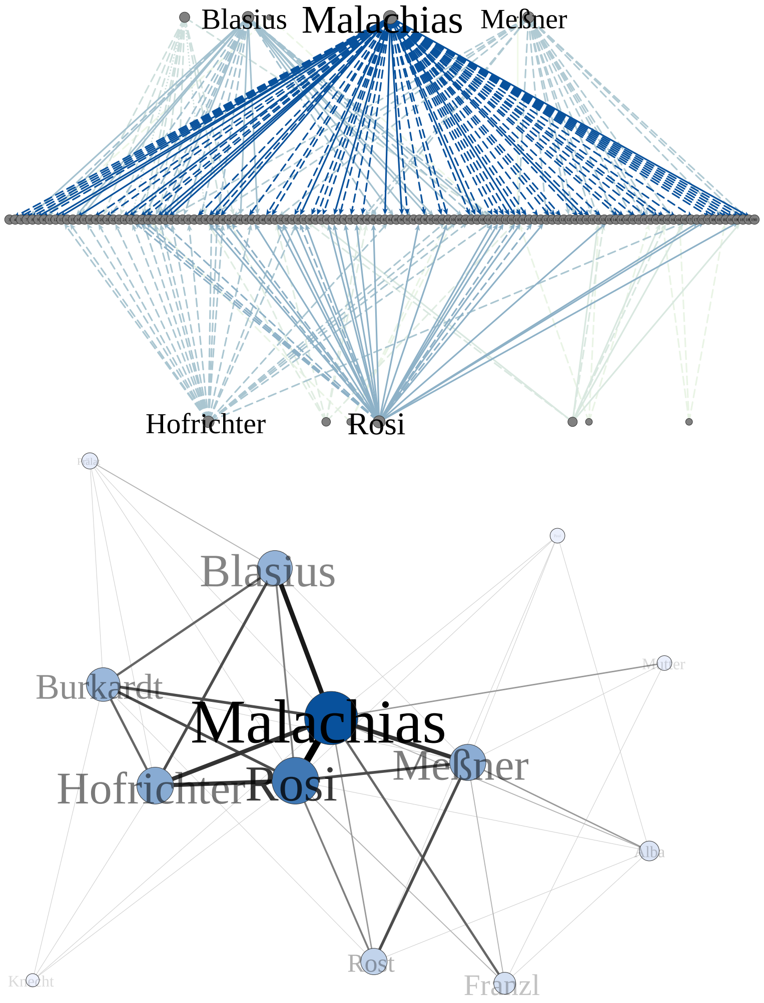



Tesche
Der Enten-Piet
Bertold Zorg
Mynheer Verkolyn
Peter Waterhout
Galinda
Jan von Amsterdam
Tieck
Die Gemälde
alte Walther
Eduard
Sophie
Prinzen
Eulenböck
Jan von Amsterdam
Baldus Sachtervanst
Mynheer van Ternswaard
Hendrik Zorg
Galinda
Mynheer Verkolyn
Bertold Zorg
Peter Waterhout
Lora
Sara Zorg
Drudje
Matrose
Gesinde
Diener
Steuermann
Scharfrichtermeister Balzer von Gouda
Meisterfrau Balzer von Gouda
Domenichino
Guido
Johannes
Quintin Messys
Buchhalter
Tizian
alten deutschen Meister
Herr von Essen
Dietrich
Musikdirektor Henne
Sophie
Händler Erich
Herr von Eisenschlicht
Eulenböck
alte Walther
Eduard
Prinzen
Diener
alten Herrn von Eisenschlicht
Barbier
Raffael
Salvator Rosa
Höllenbreughel
Rubens
Gimpel
Dante
Julio Romano
Herr der Gemälde
Kammerdiener des Grafen Alten

Tieck
Des Lebens Überfluss

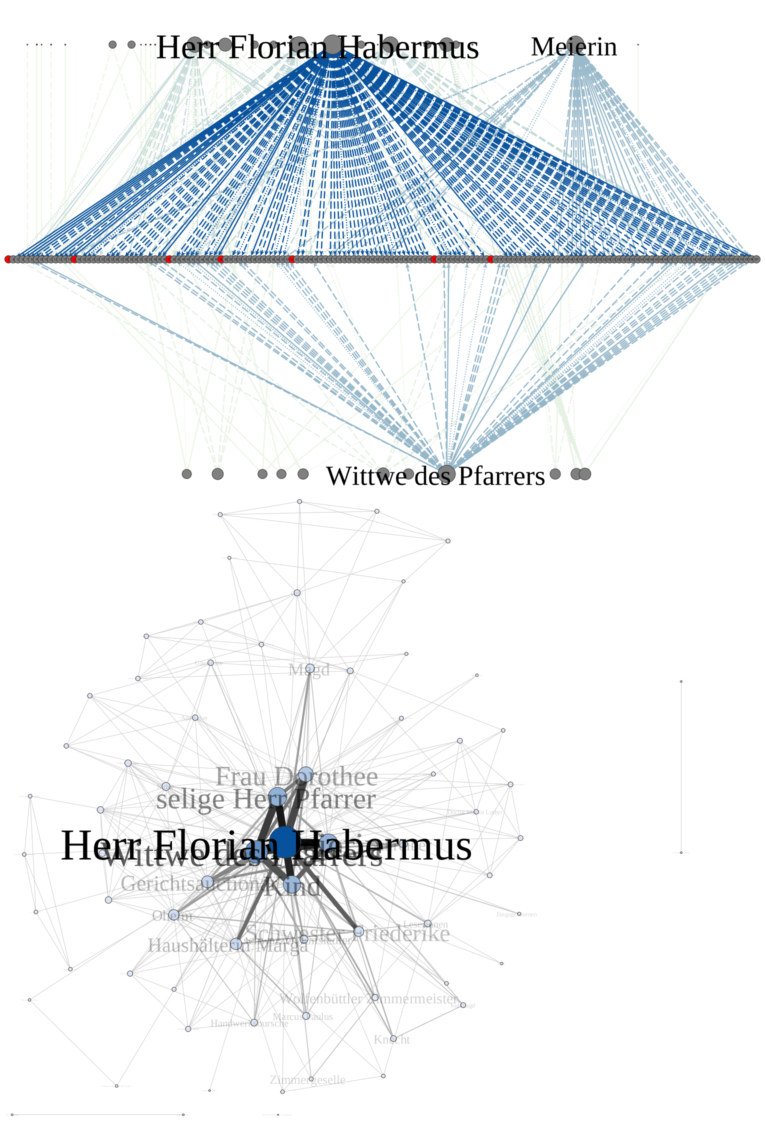


Traun
Der Gebirgspfarrer



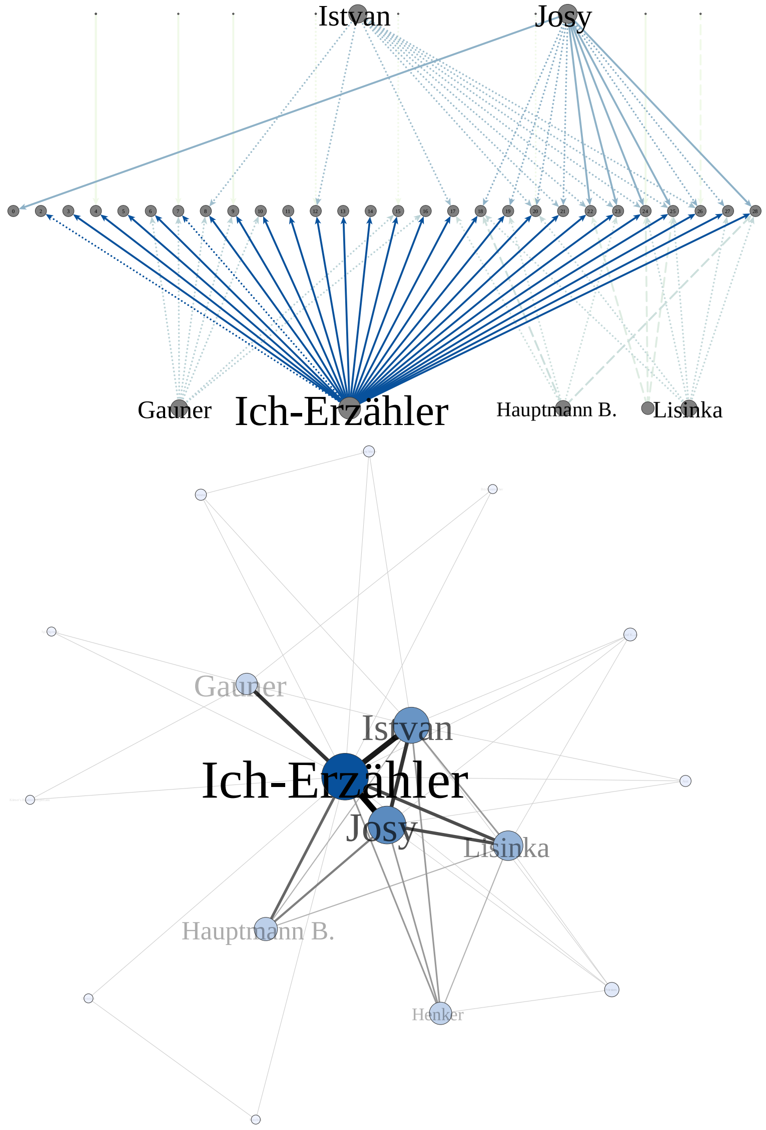
Waldmüller

Es ist nicht gut, daß der Mensch allein sei

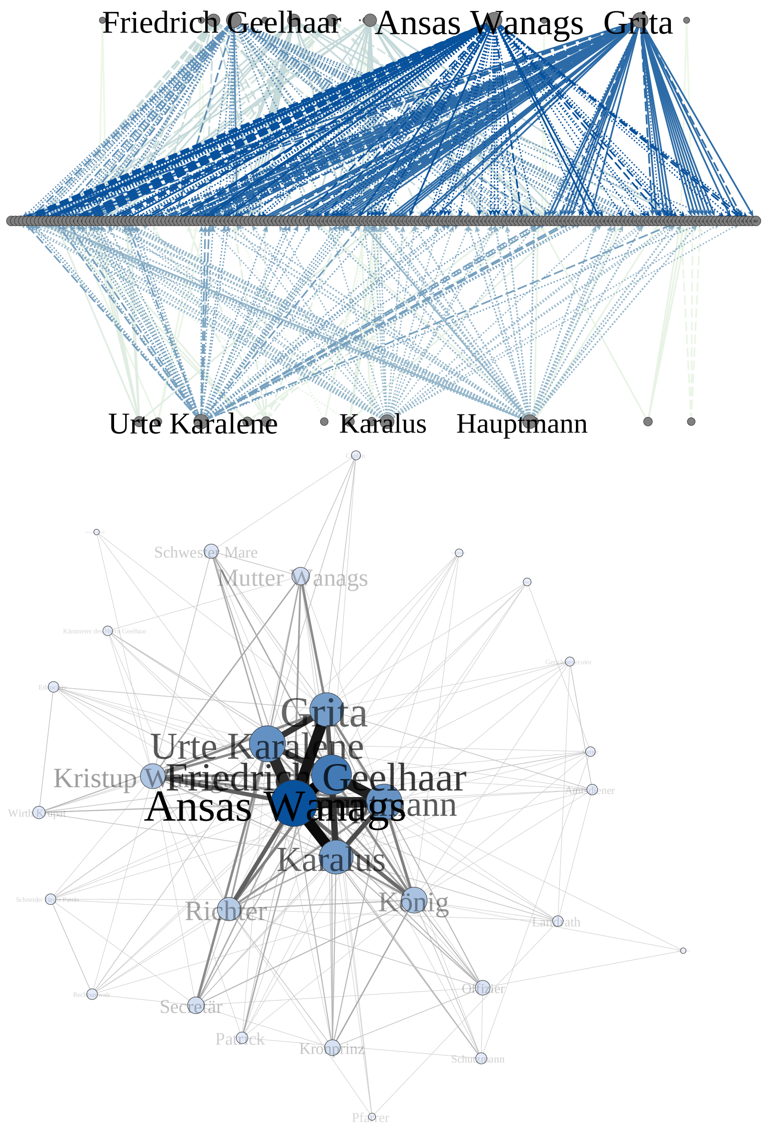


Wallner

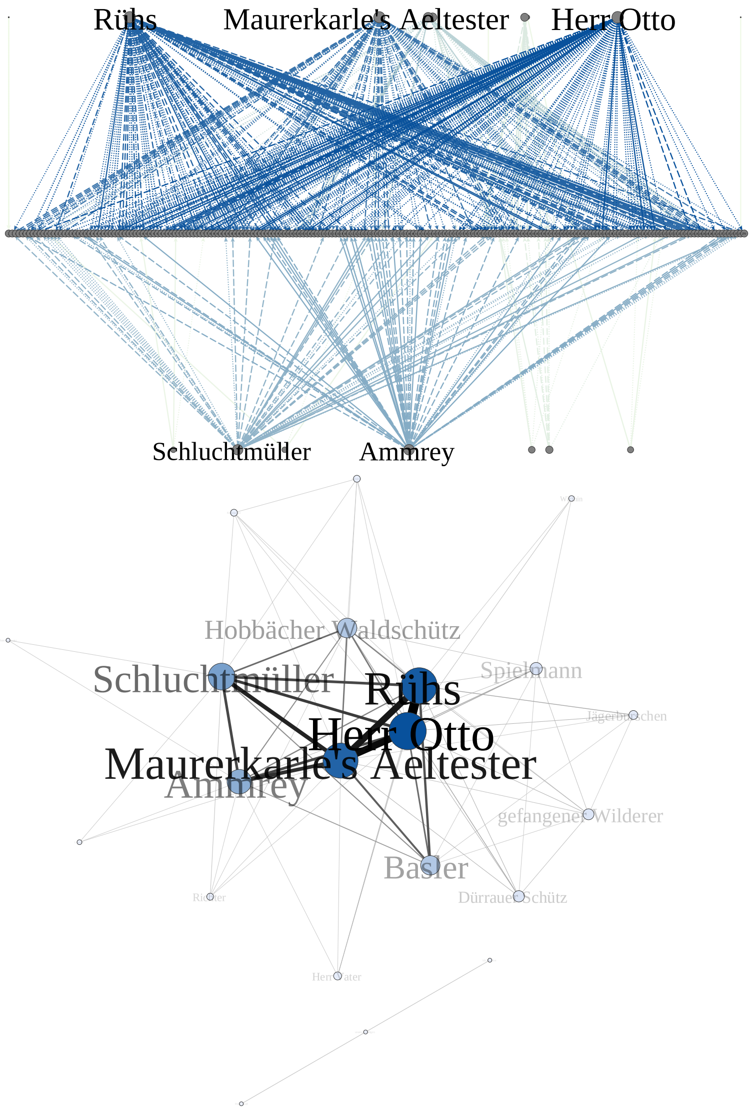
Der arme Josy



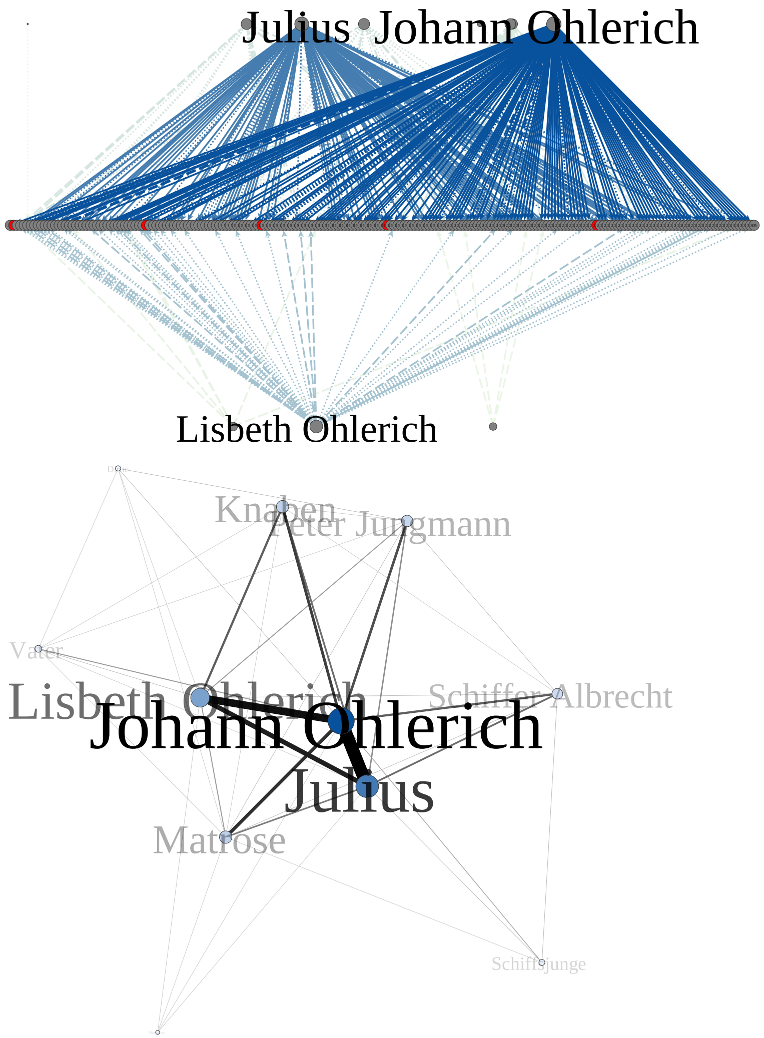
Wichert
Ansas und Grita



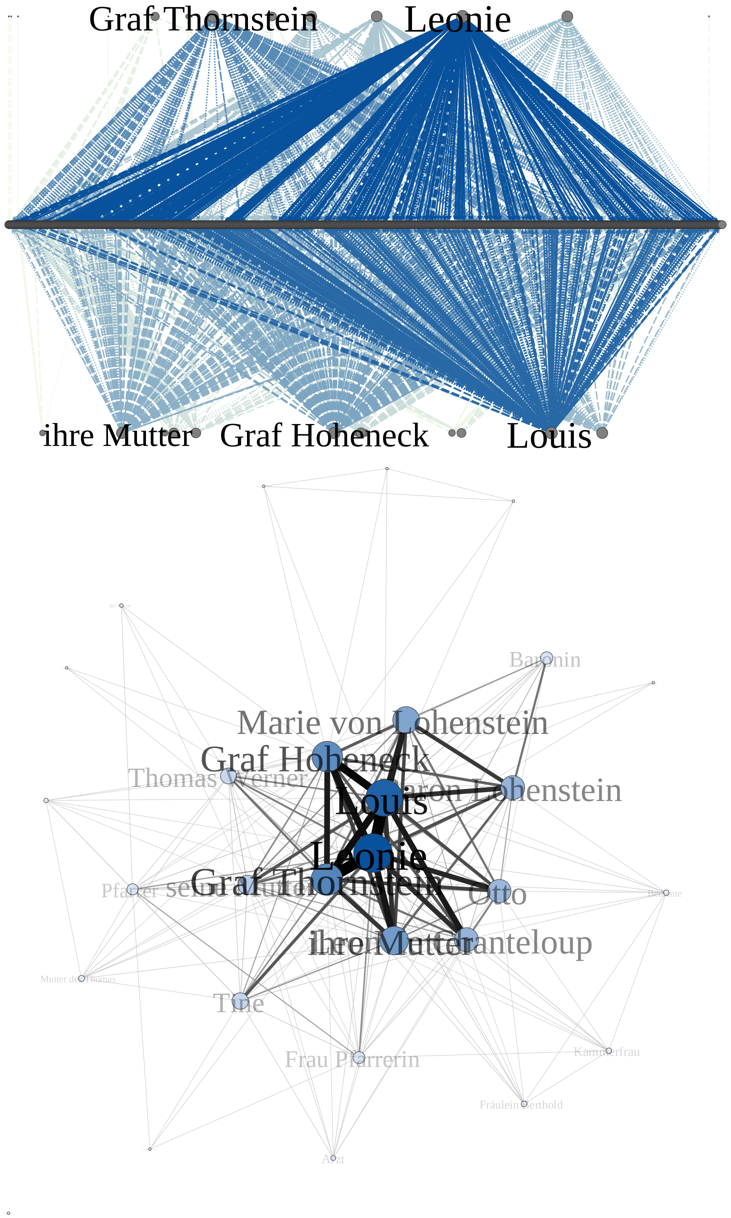
Widmann
Die katholische Mühle



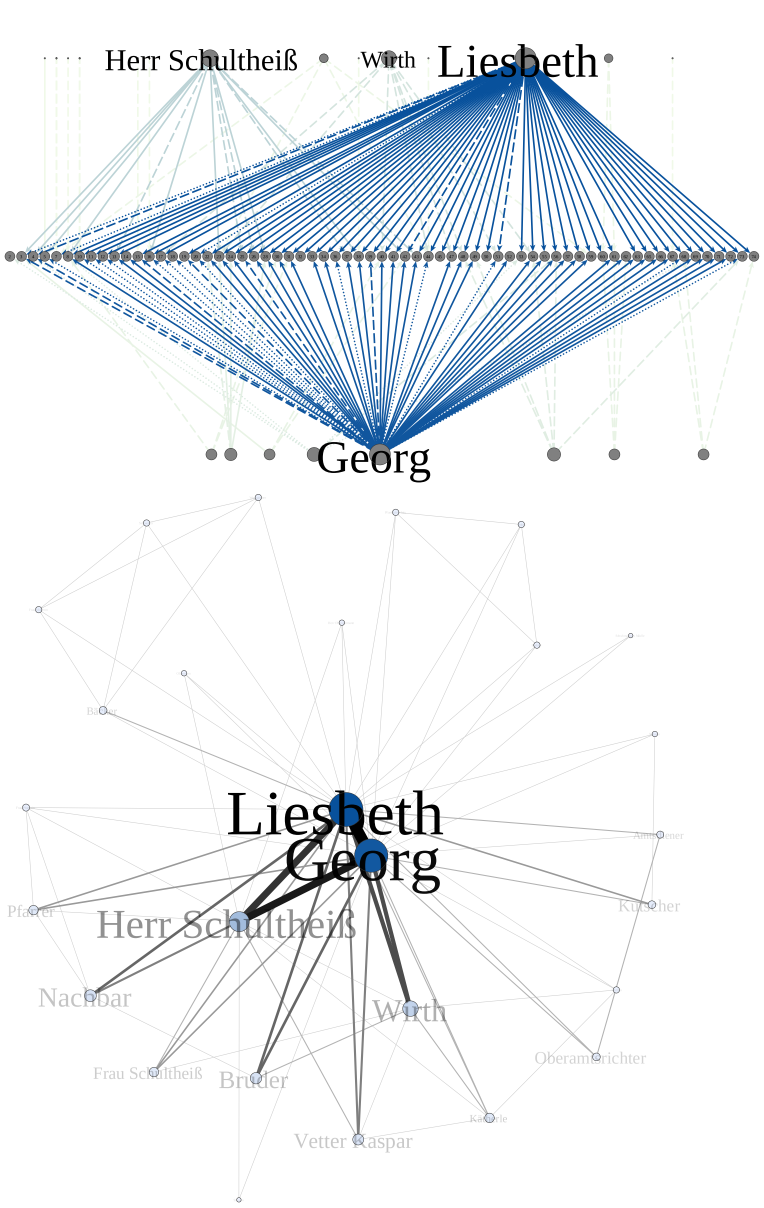
Wilbrandt
Johann Ohlerich



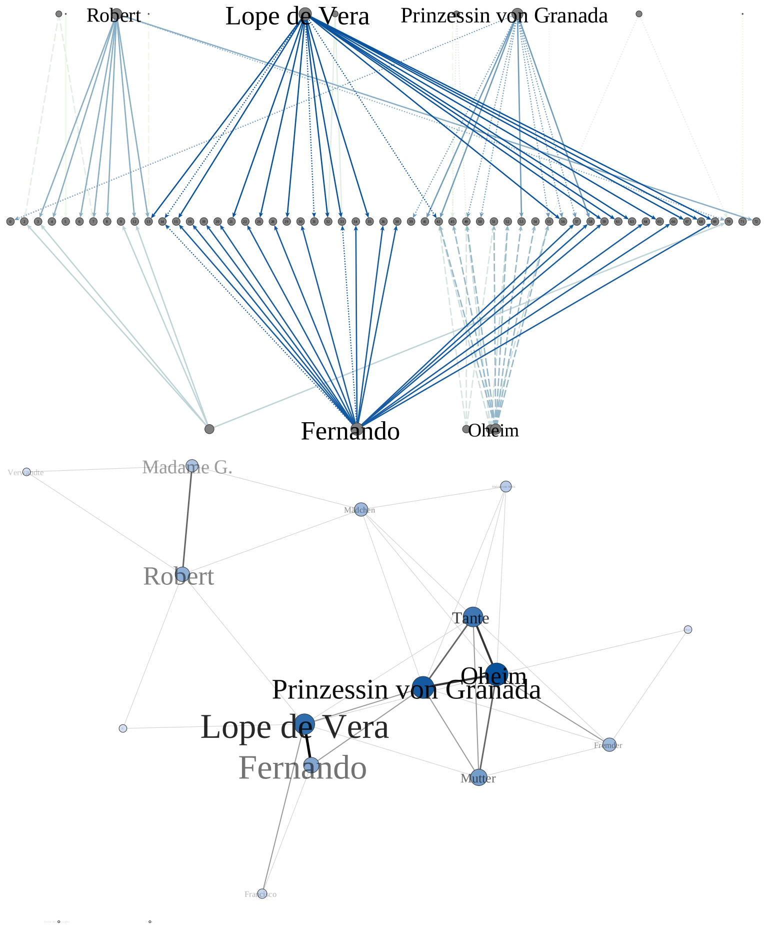
Wild
Eure Wege sind nicht meine Wege



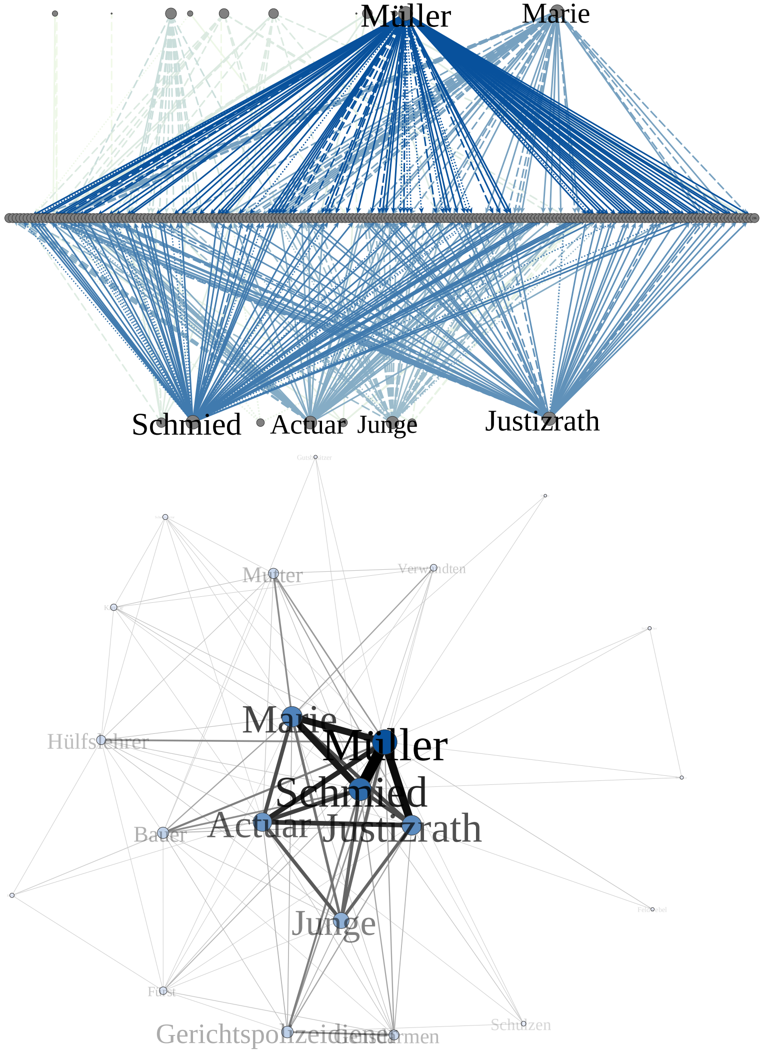
Wildermuth
Streit in der Liebe und Liebe im Streit



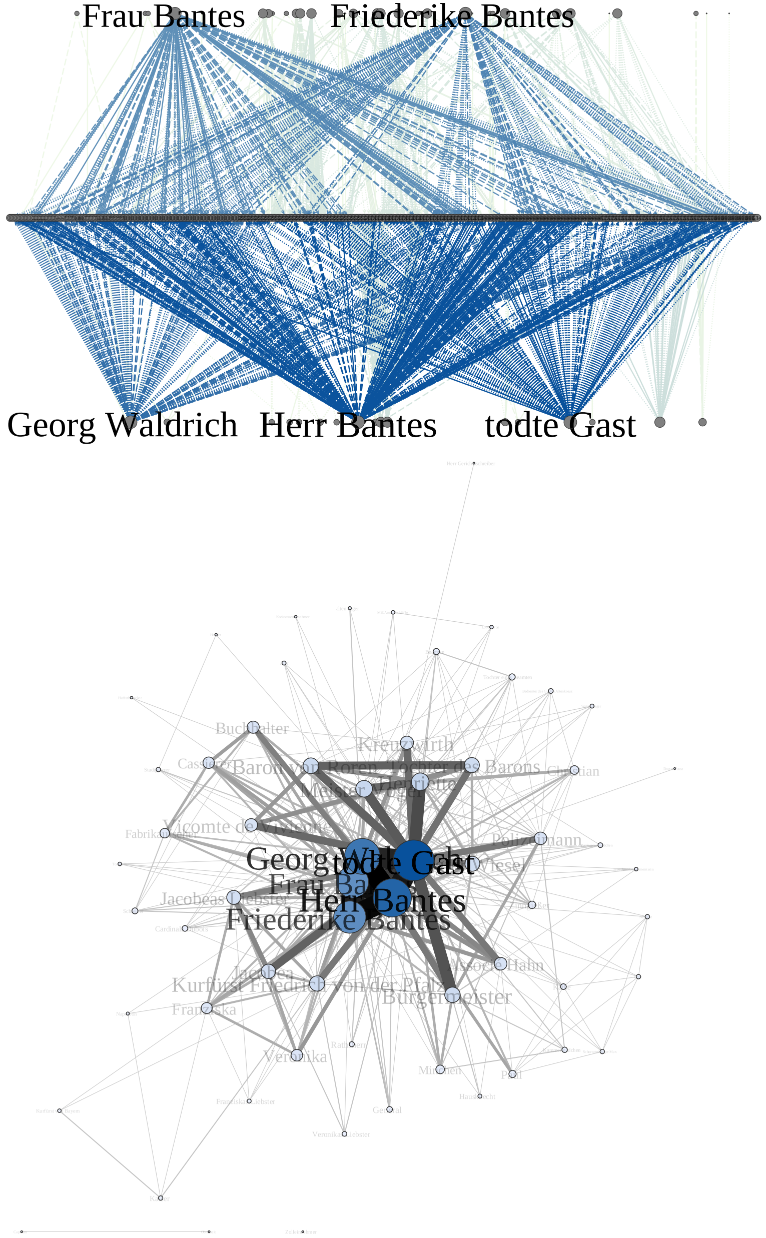
Wolf
Der Stern der Schönheit



Ziegler
Saat und Ernte



Zschokke
Der tote Gast



author	**text length**	**# paragraphs**	**# characters**	**average weighted degree**	**group degree centralization**	**entropy**
Alexis	22920	292	36	37.39	0.703	8.34
Andolt	20129	420	56	17.57	0.544	7.16
Arnim	7907	20	22	22.64	0.287	4.37
Auerbach	65328	973	96	56.67	0.796	12.55
Berthold	25235	290	44	56.45	0.563	11.62
Brentano	12367	146	49	34.57	0.553	7.95
Chamisso	19958	195	15	42.53	0.604	8.99
Dincklage	7647	110	17	27.29	0.476	7.63
Droste-Hüllshof	16428	209	29	36.34	0.499	8.93
Eichendorff	14536	269	30	22.4	0.632	5.1
Ense	16314	96	13	34.77	0.5	7.29
Fräulein v. Wolf	18022	104	12	59.33	0.236	8.35
Frey	22198	318	26	54.85	0.63	9.54
Gall	13764	380	40	36.0	0.585	9.43
Gerstäcker	9679	289	21	20.76	0.478	6.23
Glümer	14441	301	29	65.59	0.403	10.07
Goethe	7967	68	16	9.0	0.833	6.38
Goldammer Hochzeitsnacht	3526	39	6	16.0	0.2	4.42
Goldammer Wiedersehen	5464	151	13	15.38	0.509	4.48
Gotthelf Koppigen	44191	71	72	41.28	0.655	12.79
Gotthelf Notar	12666	26	16	32.62	0.379	7.68
Grillparzer	15290	80	14	33.57	0.455	8.4
Grimm	17892	202	15	51.87	0.593	8.65
Grosse	27960	592	47	55.96	0.597	11.29
Hackländer	13191	217	22	21.64	0.6	7.09
Halm	14301	53	33	28.55	0.663	10.43

author	text length	# paragraphs	# characters	average weighted degree	group degree centralization	entropy
Alexis	8256	82	15	29.33	0.604	8.11
Hauff	16788	285	57	28.98	0.834	9.82
Heyden	10555	94	21	19.33	0.702	7.1
Heyse	32365	304	34	61.88	0.611	10.65
Höfer	14072	86	27	29.19	0.659	9.63
Hoffmann	23796	104	55	24.91	0.54	9.53
Holtei	6887	96	24	19.5	0.648	7.09
Horner	18178	264	13	79.85	0.348	9.38
Immermann	29127	218	67	24.63	0.869	10.48
Kähler	11051	302	38	33.21	0.776	9.51
Keller	26026	97	36	30.78	0.652	8.26
Kinkel G.	13384	106	44	20.14	0.709	9.42
Kinkel J.	14502	198	40	18.6	0.718	8.44
Kleist	13204	23	22	36.64	0.275	6.63
Kompert	44239	1034	29	93.72	0.567	9.97
Kopisch Karnevalsfest	12625	164	44	15.45	0.622	5.8
Kopisch Träumer	13881	186	30	35.07	0.564	7.54
Kruse	21328	340	25	22.8	0.6	6.31
Kürnberger	12434	106	23	19.04	0.503	7.01
Kugler	12886	73	22	33.82	0.382	9.33
Kurz	25164	350	66	13.97	0.416	8.17
Lewald	27296	257	42	37.71	0.577	10.98
Lohmann	15962	138	25	64.64	0.448	11.84
Lorm	9567	123	11	24.0	0.6	5.78
Ludwig	11343	140	11	7.27	0.578	3.46
Meißner	19101	453	43	22.19	0.769	7.6
Meyr	48433	330	48	59.04	0.711	13.0
Mörike	19946	297	70	18.57	0.719	8.17
Mügge	36901	699	25	114.24	0.63	11.74
Müller	30147	283	16	46.75	0.495	8.06

author	**text length**	**# paragraphs**	**# characters**	**average weighted degree**	**group degree centralization**	**entropy**
Pichler	17596	277	61	14.62	0.555	6.93
Raabe	15112	298	39	26.1	0.524	7.69
Reich	9904	41	11	34.0	0.311	7.33
Riehl	5007	65	17	10.35	0.714	6.29
Roquette	24547	258	29	92.21	0.548	13.5
Rumohr	17521	117	35	24.06	0.612	9.46
Sacher-Masoch	17064	583	56	14.86	0.617	7.17
Schefer	25996	345	50	41.96	0.616	10.7
Scheffel	2934	58	13	4.46	0.318	3.98
Schmid	19909	269	53	39.47	0.616	9.87
Schreyvogel	20801	166	33	40.06	0.753	11.54
Schücking	24892	529	52	41.62	0.583	9.82
Spindler	14430	49	36	39.17	0.636	11.36
Sternberg	21841	175	36	40.5	0.531	10.01
Stifter	19866	208	16	44.12	0.637	8.65
Storm	9442	195	11	56.0	0.333	8.37
Tesche	23093	473	23	107.22	0.434	11.75
Tieck Gemälde	25691	262	61	33.67	0.632	11.31
Tieck Überfluß	17919	240	37	12.81	0.64	6.86
Traun	6548	191	13	22.0	0.53	5.14
Waldmüller	18740	199	75	16.93	0.697	8.87
Wallner	3509	29	17	11.88	0.783	6.17
Wichert	22871	264	34	37.88	0.53	7.93
Widmann	15140	289	22	39.36	0.439	8.02
Wilbrandt	18371	306	12	47.0	0.333	6.47
Wild	61403	1208	29	149.45	0.425	10.49
Wildermuth	7071	75	27	15.78	0.795	7.9
Wolf	3200	73	17	6.71	0.275	4.54
Ziegler	14082	240	23	37.3	0.614	8.28
Zschokke	32384	635	67	49.19	0.668	11.43

Table A.3: Text and network properties for each novella character network.