Romeo, Freund des Mercutio: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren

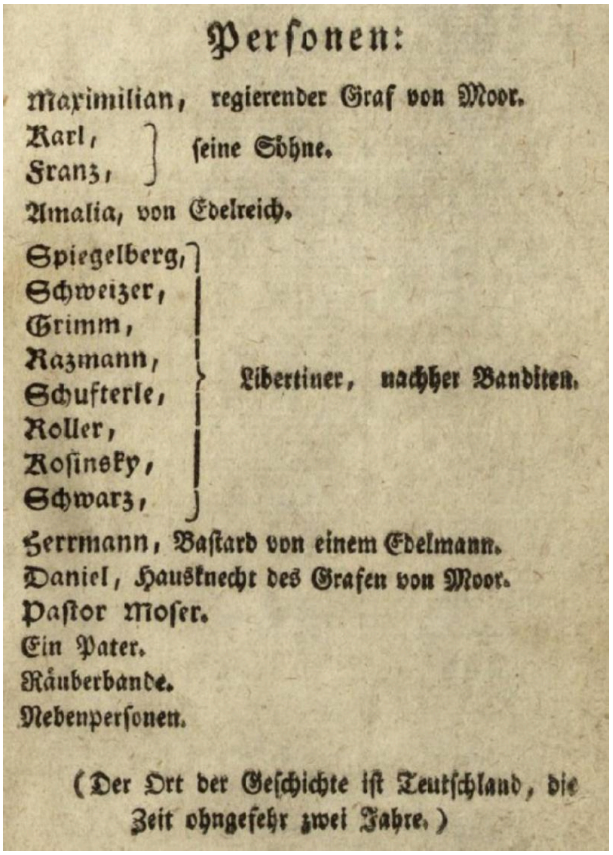
Wiedmer, Nathalie
nathalie.wiedmer@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pagel, Janis
janis.pagel@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Reiter, Nils
nils.reiter@uni-koeln.de
Universität Stuttgart, Deutschland; Universität Köln, Deutschland

Einleitung

In diesem Beitrag stellen wir eine Methode vor, um Informationen über Figurenrelationen in dramatischen Texten, die innerhalb der *dramatis personae* (Figurenverzeichnis) sprachlich kodiert sind, zu extrahieren und maschinenlesbar im TEI/XML vorzuhalten. Das Figurenverzeichnis kann als Paratext (Genette 1993) dem Nebentext zugerechnet werden, ist jedoch literaturwissenschaftlich, von Einführungswerken abgesehen, noch so gut wie nicht erschlossen.[1] Das Figurenverzeichnis steht zwar unabhängig vom eigentlichen Text am Anfang, kann jedoch bereits Figuren- bzw. Textwissen vermitteln, indem die Figuren nach sozial-politischem Stand, Familienzugehörigkeit oder nach anderen Gruppierungen geordnet sind (vgl. Abbildung 1). Häufig lässt sich an der Positionierung eines Names im Figurenverzeichnis auch die Wichtigkeit der betreffenden Figur im Drama ablesen (Pangallo 2015, 91). Durch diese Strukturierung ist es teilweise möglich, schon vorab auf zentrale Konfliktpotentiale des Textes zu schließen (Jeßing 2015, 79–80). Darüberhinaus kann das Figurenverzeichnis laut Pfister und Asmuth auch der Ort erster auktorialer Bewertungen oder Hinweise sein und dient somit nicht nur der reinen Vorstellung der Figuren und ihrer Strukturen untereinander (Pfister 2001, 95; Asmuth 2016, 85).


Personen:

Maximilian, regierender Graf von Moor.
Karl, Franz, } seine Söhne.
Amalia, von Edelreich.
Spiegelberg, Schweizer, Grimm, Razmann, Schufterle, Roller, Rosinsky, Schwarz, } Libertiner, nachher Banditen.
Herrmann, Bastard von einem Edelmann.
Daniel, Hausknecht des Grafen von Moor.
Pastor Moser.
Ein Pater.
Räuberbande.
Nebenpersonen.

(Der Ort der Geschichte ist Teutschland, die Zeit ohngefehr zwei Jahre.)

Abbildung 1: Figurenverzeichnis in Die Räuber (Friedrich Schiller, 1781)

Das Verfahren – und dessen Implementierung in einem Python-Skript – ist auch für in Zukunft digitalisierte Dramen anwendbar, und wird von uns als quelloffene Software zur Verfügung gestellt. Es ist vergleichsweise einfach auf neue Sprachstufen oder Genres anpassbar und liefert – auch bei nicht-perfekten Ergebnissen – eine gute Vorlage. Eine Evaluation des Verfahrens erfolgt auf ungesehenen Testdaten. Außerdem veröffentlichen wir einen Datensatz mit extrahierten Figurenrelationen aus deutschsprachigen Dramen, die manuell validiert und korrigiert wurden. Diese Daten werden zur einfachen und breiten Nutzung im TEI-Format in das GerDraCor[2] eingespeist. Schlussendlich beschreiben wir beispielhaft zwei Analyseszenarien in denen die Daten neue Einblicke bieten (können).

Automatische Extraktion von Figurenrelationen

Unsere Methode unterscheidet zwischen sieben Kategorien von Figurenrelationen (Tabelle 1). Ausschlaggebend für die Zuordnung zu einer der Kategorien sind Signalwörter wie “Vater”, “Kammerdiener”, “Geschwister” etc. Diese Signalwörter

werden in einer kontextfreien Grammatik der entsprechenden Kategorie zugeordnet.

Tabelle 1: Figurenrelationen

Relationen Label	gerichtet/ungerichtet	Beschreibung
parent_of	directed	Eine Figur ist Elternteil einer anderen
lover_of	directed	Liebesbeziehungen (unverheiratet)
related_with	directed	Familienbeziehungen (außer Eheleute)
associated_with	directed	Figuren, die miteinander anderweitig verbunden sind (z.B. Diener, Kindermädchen etc.)
siblings	undirected	Figuren, die mindestens ein gemeinsames Elternteil haben
spouses	undirected	verheiratete oder verlobte Figuren
friends	undirected	Freundschaftsbeziehungen

Kontextfreie Grammatiken bezeichnen in der Informatik eine Sammlung aller syntaktisch korrekten Programme einer Programmiersprache (Böckenhauer und Hromkovi# 2013, 177). Die formalisierte Art, in der die Grammatik alle Regeln einer Programmiersprache enthält, erlaubt es, automatisierte Syntaxanalysen von Programmen durchzuführen (Böckenhauer und Hromkovi# 2013, 177). Die Regeln werden mit Hilfe zweier Alphabete beschrieben: Das Terminalalphabet enthält alle Wörter einer Sprache, wohingegen das Nichtterminalalphabet Variablen enthält, die vorgeben, auf welche Art und Weise die Wörter kombiniert werden können (Böckenhauer und Hromkovi# 2013, 178).

Wir nutzen eine solche Grammatik, um drei verschiedene Zeilenarten im Figurenverzeichnis zu unterscheiden, bei denen es sich um Nichtterminale handelt. Alle in den Sätzen vorkommenden Tokens sind Terminale, deren Kombination und Anzahl Aufschluss darüber gibt, um was für eine Art von Zeile es sich jeweils handelt. Auf diese Weise können auch zeilenübergreifende Relationen erkannt werden.

Zu Beginn des Programmablaufs werden die in GerDraCor vorhandenen Figuren-IDs zusammen mit dem Figurenverzeichnis ausgelesen und gespeichert. Da wir die Beziehungen zwischen den Figuren ausschließlich anhand der Angaben im Figurenverzeichnis konstruieren, muss der Dramentext nicht extra eingelesen werden. Daraus ergibt sich die Beschränkung, dass jegliche Beziehungen, die nicht im Figureverzeichnis explizit gemacht werden, vom Programm auch nicht erkannt werden können. Es geht demnach ausschließlich darum, das Personenverzeichnis maschinenlesbar und -interpretierbar zu machen. So ignoriert das Programm beispielsweise auch alle Zeilen, die eine Gruppe von Figuren als Kollektiv einführt, da diese als “Nummern oder als anonyme Angehörige von Untergruppen” (Schlaffer 1972, 11) meistens keine eigenen Namen haben und auch keine explizit gemachten Beziehungen.[3]

Anschließend werden alle Tokens jeder Zeile des Figurenverzeichnisses daraufhin untersucht, ob es sich dabei um Figurennennungen oder Signalwörter handelt und die Grammatik einem Parser übergeben, der die Zeilen des Figurenverzeichnisses in Baumstrukturen überführt (Abbildung 2).

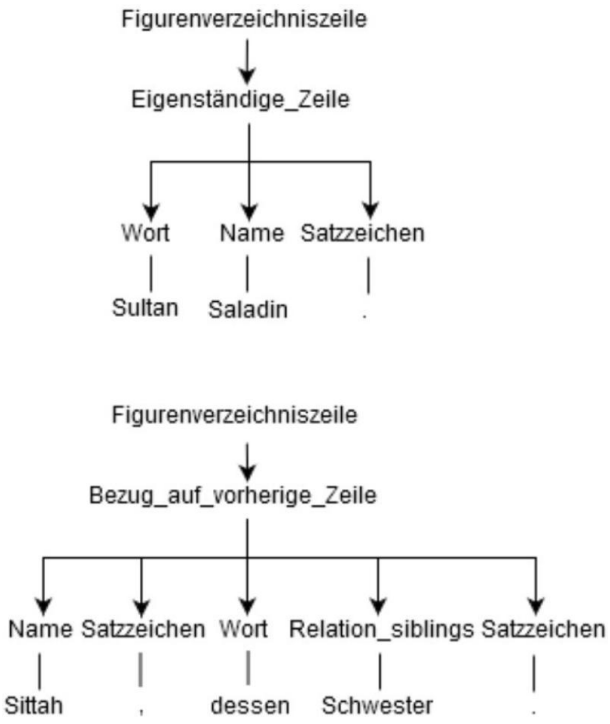


Abbildung 2: Zwei reduzierte Baumstrukturen für Figuren aus Nathan der Weise.

Aus den erstellten Baumstrukturen werden einzelne Informationen ausgelesen, die grundlegend für die Erkennung der Figurenrelationen sind. Zuerst wird überprüft, wie viele IDs sich in einer Zeile befinden. Die erste oder einzige wird zur Erstellung späterer Relationen abgespeichert. Befindet sich in einer Zeile zusätzlich zu einer ID noch ein Signalwort für eine Figurenrelation, bezieht sich die Zeile in der Regel auf die vorangegangene, wie beispielsweise in *Nathan der Weise*:

<castItem corresp="#saladin">Sultan Saladin.</castItem>

<castItem corresp="#sittah">Sittah, seine Schwester.</castItem>

Die zweite Zeile enthält neben dem Namen noch das Signalwort “Schwester”, das auf die Beziehungsart siblings hinweist, eine ungerichtete Relation. Da keine zweite Figurenbezeichnung in der Zeile vorkommt, entnimmt das Programm als zweiten Part für die Geschwisterbeziehung den Namen bzw. die daraus abgeleitete ID saladin aus der vorherigen Zeile:

<relation name="siblings" mutual="#sittah #saladin" />

Wenn die beiden benötigten IDs für das Erstellen der Figurenrelation feststehen, wird die Art der Relation durch das Auslesen des Signalworts aus der Baumstruktur festgestellt. Danach werden daraus die Zeilen mit den Figurenrelationen erstellt und diese anschließend in die jeweilige TEI-Version des Textes geschrieben.

Befindet sich in einer Zeile eine zweite Figuren-ID, bezieht sich die Zeile nicht auf eine vorangegangene, sondern stellt selbst den zweite Bezugspunkt der Relation. Das ist beispielsweise bei der Figur “Camillo Rota” in *Emilia Galotti* der Fall:

<castItem corresp="#camillo_rota">Camillo Rota, einer von des Prinzen Räten.</castItem>

Die erste erkannte ID ist camillo_rota, die zweite der_prinz, abgeleitet aus “des Prinzen”. Die IDs werden in gerichtete Relationen mit aktivem und passivem Part überführt:

<relation name="associated_with" active="#camillo_rota" passive="#der_prinz" />

Das Programm arbeitet dabei ausschließlich mit den IDs. Dafür ist es nicht nötig, dass Figurennamen explizit als Namen oder Adelstitel als Titel erkannt werden. Es geht ausschließlich darum aus den einzelnen Wörtern einer Zeile im Figurenverzeichnis Namen bzw. Namensteile und Titelangaben herauszufiltern, die den IDs entsprechen, um die Zeilen einer oder mehreren Figuren zuordnen zu können.

Um auch IDs zu erkennen, die sich geringfügig von den Namensnennungen im Figurenverzeichnis unterscheiden, überprüft das Programm pro Wort eine Reihe an Varianten. So trennt es beispielsweise vom oben gennannten Wort “Prinzen” das Suffix ab und überprüft, ob ein Artikel Teil der ID ist. So kann “des Prinzen” der ID “der_prinz” zugeordnet werden. In manchen Fällen funktioniert diese Abwandlung aber nicht so reibungslos. In *Der Eheteufel auf Reisen* wird eine Figur im Figurenverzeichnis mit dem Namen “Gustel” eingeführt, wohingegen die ID „gustchen“ lautet. Die ID orientiert sich hier an der Namensform, die im Stück tatsächlich verwendet wird und nicht an der Bezeichnung im Figurenverzeichnis. Das führt dazu, dass das Programm die ID “gustchen” nicht dem Wort “Gustel” zuordnen kann, da sie sich zu stark unterscheiden.

Evaluation

Um die Methode zu evaluieren, wurden die automatisch erzeugten Relationen manuell nachkorrigiert und so ein Goldstandart erzeugt. Im Schnitt bearbeiteten die Korrektoren 12 Texte pro Stunde. Beim Abgleich der automatisch erzeugten Ergebnisse mit dem Goldstandart lag der Macro-Average-Recall Wert bei 0,3 (Standardabweichung: 0,3) und der Wert von Macro-Average-Precision bei 0,55 (Standardabweichung: 0,4), was einen Macro-Average-F-Score von 0,49 (Standardabweichung: 0,25) ergibt.

Korpus

GerDraCor ist ein deutsches Dramenkorpus, das nach TEI-P5 Standarts kodiert ist und im Dezember 2019 474 Dramen enthält, die im Zeitraum von 1730 bis 1940 veröffentlicht wurden (Fischer u. a. 2019). Es ist Teil des größeren DraCor (Fischer u. a. 2019), das als *Programmable Corpus* darauf ausgelegt ist, durch Community-Anstrengungen korrigiert und verbessert werden zu können (Fischer u. a. 2019, 195). Da auf einem Fork von GerDraCor gearbeitet wurde, können die automatisch erzeugten Figurenrelationen dem Korpus unproblematisch hinzugefügt werden. Zusätzlich wurden die Relationen, wie bereits beschrieben, manuell nachkorrigiert, um eine erhöhte Qualität für die Nachnutzung zu gewährleisten.

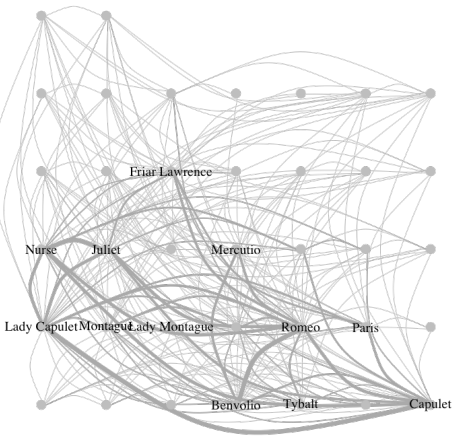
Im Rahmen der manuellen Nachkorrektur wurden außerdem interessante Fälle identifiziert. So wird etwa eine Gruppe von Figuren in dem oben abgebildeten Figurenverzeichnis von Schillers *Die Räuber* als “Libertiner, nachher Banditen” bezeichnet, wodurch Informationen aus der späteren Handlung des Stückes vorweggenommen werden. Diese Art der Vorwegnahme findet sich außerdem in Stücken von Grabbe (*Herzog Theodor von Gothland*, Panizza (*Das Liebeskonzil*) und Uhland (*Ludwig der Bayer*). In Kaisers *Stadt und Land* hingegen wird mit der Zeile “Erster Bergmann, später Michael” keine Entwicklung in der Handlung, sondern eine Veränderung der Sprecherbezeichnung markiert. Vorwegnahmen mit Bezug auf veränderliche Beziehungen zwischen Figuren konnten nicht festgestellt werden.

Analyseszenarien

Wir stellen im folgenden zwei Analysen vor, in denen von den automatisch extrahierten Relationen Gebrauch gemacht wird, sowohl eine Einzeltext- als auch eine Korpusanalyse. Diese illustrieren Möglichkeiten, die Relationen in der Textanalyse zu berücksichtigen.

Im ersten Beispiel betrachten wir Shakespeares *Romeo and Juliet* in der derzeit auf dracor.org verfügbaren Fassung.[4] Zunächst können die Relationen visualisiert werden. Abbildung 3 zeigt das Figurennetzwerk nach Kopräsenz auf der linken und das Netzwerk, das sich aus den sozialen Beziehungen ergibt auf der rechten Seite. Zur besseren Lesbarkeit wurde ein geeigneter Layout-Algorithmus angewendet. Dabei ist zunächst interessant, dass die beiden Familien keineswegs unverbunden sind:

Über Mercutio (Freund von Romeo) und Paris (Verlobter von Julia) sind beide mit dem Prinzen verbunden.



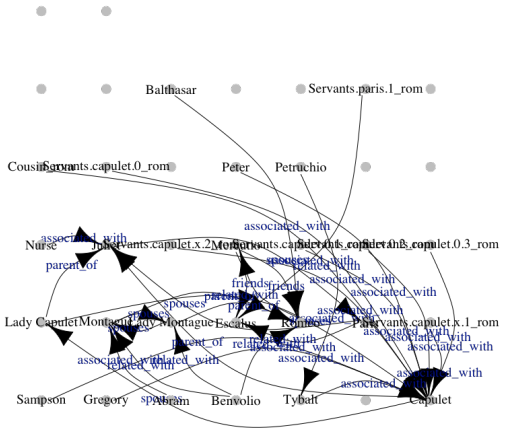


Abbildung 3: Figurennetzwerke nach Kopräsenz (oben) und Relationen (unten). Zur besseren Übersichtlichkeit wurden die Figuren auf feste Positionen gesetzt, die oben und unten gleich sind. Bezeichnungen werden nur gezeigt wenn der Grad groß genug ist (oben) oder sie an einer Beziehung beteiligt sind (unten).

Auch wenn Abbildung 3 eine gewisse Symmetrie suggeriert, ist diese keineswegs gegeben wenn wir die Redeanteile nach Familien aufschlüsseln, wie es aus den Annotationen ebenfalls direkt möglich ist. Abbildung 4 zeigt die aggregierten Redeanteile der Figuren, wobei Figuren, die durch Verwandtschaft oder Arbeitsverhältnis zu einer der Familien gehören, zusammengefasst wurden (mit Ausnahme von Mercutio und Paris, die beide mit dem Prinzen verwandt sind). Es zeigt sich, dass Angehörige der Familie Capulet etwas weniger als doppelt so viele Wörter äußern als Angehörige der Familie Montague.

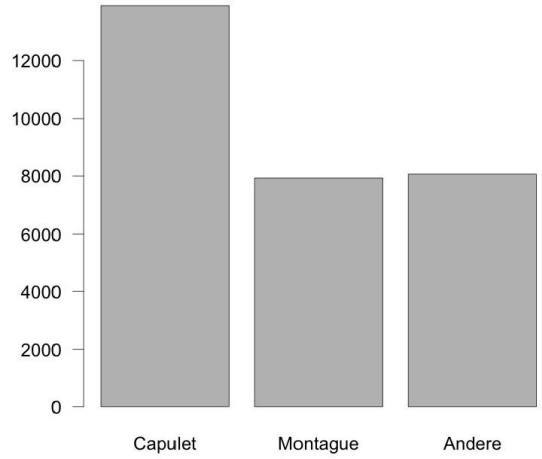


Abbildung 4: Redeanteile nach Familie

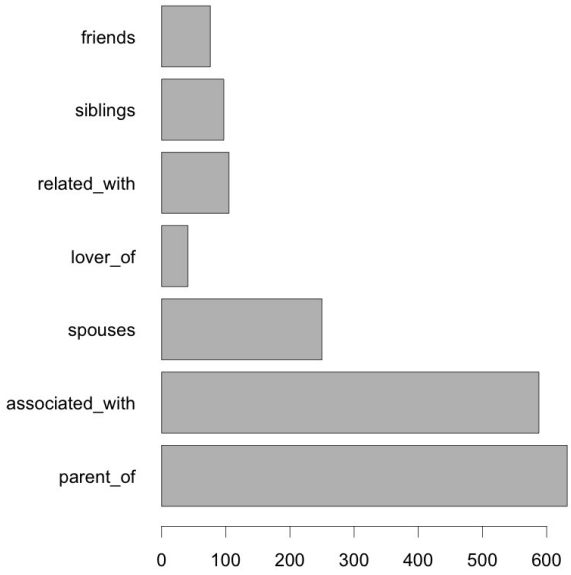


Abbildung 5: Verteilung der Relationen im Gesamtkorpus

Betrachtet man das annotierte Gesamtkorpus stellt man fest, dass die Relationen ungleich verteilt sind. Während Ehen/Verlobungen, Elternschaft und sonstige Assoziationen relativ häufig vorkommen, spielen Geliebte, sonstige Verwandtschaften, Freundschaften und Geschwister eine vergleichsweise kleine Rolle.[5]

In Abbildung 6 sehen wir die Anzahl der Relationen bestimmter Typen ins Verhältnis gesetzt zur Großgattung (Komödie/Tragödie). Dabei wurden die Angaben auf den Titeln der Dramen übernommen und leicht vereinheitlicht (z.B. Bürgerliches Trauerspiel → Tragödie oder

Zauberlustspiel → Komödie). Dabei ist zu konstatieren, dass Median und erstes Quartil bei 0 für alle Dramen bei 0 liegen: Viele Dramen weisen keine Beziehungsdefinition auf (oder sie konnten nicht automatisch identifiziert werden, siehe Fußnote). Größere oder signifikante Abweichungen zwischen den Gattungen gibt es nicht, egal welche Relation betrachtet wird. Lediglich die Relation spouses scheint im Figurenverzeichnis von Komödien häufiger genannt zu werden.

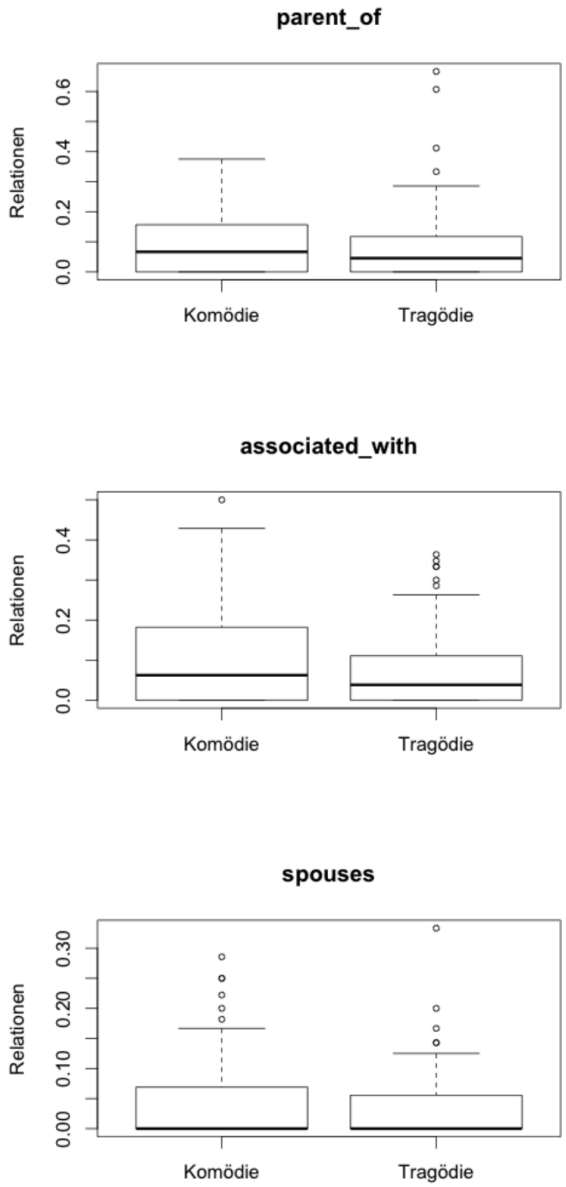


Abbildung 6: Anzahl typisierter Relationen nach Gattung

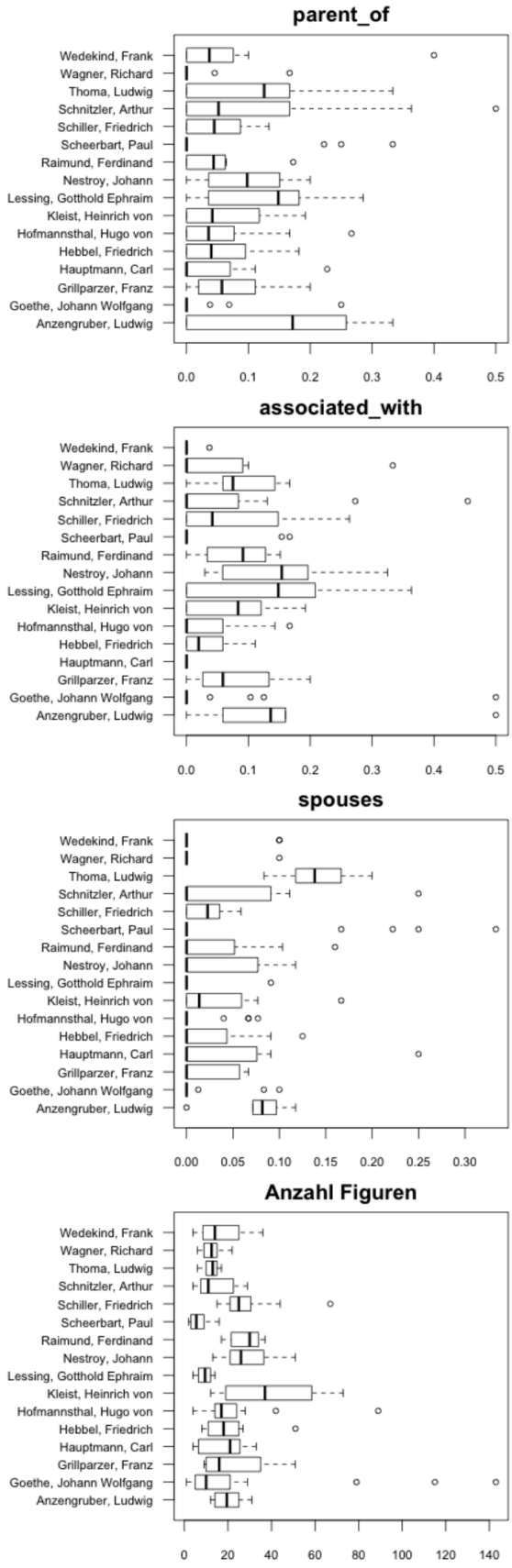


Abbildung 7: Anzahl typisierter Relationen nach Autor. Zur besseren Übersicht wurden nur Autoren

berücksichtigt, die mindestens durch fünf Dramen vertreten sind

Eine Verteilung der genannten Relationen nach Autor zeichnet jedoch ein anderes Bild (Abbildung 7). Bestimmte Autoren, vor allem Ludwig Anzengruber (1839-1889) und Johann Nestroy (1801-1862), haben klare Tendenzen dazu, mehr Relationen im Figurenverzeichnis zu nennen. Beide verfassen tendenziell Possen und Komödien.

Fazit

Mit den von uns bereitgestellten maschinenlesbaren Informationen ermöglichen wir Analysen dramatischer Figuren, die die als bekannt vorausgesetzten Informationen im Figurenverzeichnis mit berücksichtigen können. Neben den oben skizzierten Analysen können die Informationen auch in inhaltliche Analysen einfließen und etwa die soziale Nähe mit der Bühnennähe korrelieren o.ä.

Kontextfreie Grammatiken haben sich hier – trotz der bekannten Schwächen im Bezug auf natürliche Sprache – als effizienter Formalismus herausgestellt, um die Figurenverzeichnisse maschinenlesbar zu machen. Wir halten dieses Verfahren für geeignet, um auch in anderen Kontexten mit semi-strukturierten Textdaten zu arbeiten, wo aufgrund der begrenzten Menge ein maschinelles Lernverfahren nur bedingt zum Einsatz kommen kann.

Fußnoten

1. Beispielsweise spielt das Figurenverzeichnis im kürzlich erschienenen (Tonger-Erk, Werber, und Baum 2018), aber auch in (Genette 1993) quasi keine Rolle.
2. https://dracor.org
3. Für mehr Informationen vergleiche (Schlaffer 1972, 11)
4. https://github.com/dracor-org/shakedracor/blob/d569dc9886b3d1951f23b0454a3d7103e4cdf1bb/tei/romeo-and-juliet.xml
5. Die konkreten Ergebnisse wurden auf den *vollautomatisch erzeugten Relationen* erzielt.

Bibliographie

Asmuth, Bernhard. (2016). *Einführung in die Dramenanalyse*. Stuttgart: J.B. Metzler Verlag.

Böckenhauer, Hans-Joachim / Juraj Hromkovi# (2013): *Formale Sprachen: Endliche Automaten, Grammatiken, lexikalische und syntaktische Analyse*. Zürich: Springer.

Fischer, Frank / Ingo Börner / Mathias Göbel / Angelika Hechtl / Christopher Kittel / Carsten Milling / Peer Trilcke (2019): „Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor“. In *Proceedings of DHd*. https://doi.org/10.5281/zenodo.2596095 .

Genette, Gérard (1993): *Palimpseste. Die Literatur auf zweiter Stufe*. Frankfurt am Main: Suhrkamp.

Jeßing, Benedikt (2015): *Dramenanalyse. Eine Einführung*. Berlin: Erich Schmidt Verlag.

Pangallo, Matteo (2015): „‚I will keep and character that name‘: Dramatis Personae Lists in Early Modern Manuscript Plays“. *Early Theatre* 18 (2): 87–118. https://doi.org/10.12745/et.18.2.1166 .

Pfister, Manfred (2001): *Das Drama*. München: Wilhelm Fink.

Schlaffer, Hannelore (1972): *Dramenform und Klassenstruktur. Eine Analyse der dramatis persona "Volk"*. Stuttgart: J.B. Metzler Verlag.

Tonger-Erk, Lily / Nils Werber / Constanze Baum (Hrsg.) (2018): „Hauptsache Nebentext. Regiebemerkungen im Drama“. *Zeitschrift für Literaturwissenschaft und Linguistik* 48 (3).