Stilometrie

Eine in den Digital Humanities häufig genutzte Methode zur Analyse von Textkorpora, insbesondere zum Vergleich von Texten zur Überprüfung auf Ähnlichkeit oder zur Zuordnung zu Epochen und auch zur Klassifizierung nach Autorschaft, ist die Stilometrie. „In der digitalen Stilometrie werden Texte oder Textpassagen auf der Grundlage statistischer Verteilungen“ [1] (i. d. R. der häufigsten Wörter, MFW = most frequent words) stilistisch miteinander verglichen, wobei damit die stilometrische Analyse univariat ist. Bei den MFW können alle Wörter, auch die Funktionswörter wie ‚und‘, ‚oder‘, ‚in‘, ‚der‘, ‚die‘, ‚das‘ usw. oder nur die Inhaltswörter genutzt werden. Auf diese Weise „lässt sich die stilistische Entwicklung oder Differenzierung eines literarischen Textes, eines Œuvres, oder gar einer ganzen Epoche quantitativ nachvollziehen. Insbesondere werden stilometrische Methoden neben Autorschaftsattributionen und Epochendifferenzierungen zur Genreklassifikationen oder auch in der forensischen Linguistik eingesetzt.“ (vgl. Horstmann, J. 2018)

Basis des stilistischen Vergleichs ist die Ähnlichkeit zweier Texte, die mit Hilfe eines Distanzmaßes gemessen wird. Hier hat sich das sogenannte Burrows‘ Delta-mean Distanzmaß (Burrows, J. 2002, S. 267) bewährt, es können aber auch das Euklidische oder das Cosinus Maß genutzt werden. Je kleiner die Distanz zwischen zwei Texten ist, desto ähnlicher sind sie sich. Zusätzlich wird auch die Distanz eines Textes zum Mittelwert des Korpus gemessen. Wenn das Korpus z.B. aus Texten einer bestimmten Epoche besteht, kann der analysierte Text dieser Epoche zugeordnet werden, falls die Distanz des Textes zum Distanz-Mittelwert des Korpus gering ist. Die Distanzmessung kann noch andere Erkenntnisse zutage fördern; u.a. kann man erkennen, ob der Text von einer Frau oder einem Mann geschrieben wurde.

In dem System R steht das package stylo zur digitalen Anwendung der Stilometrie zur Verfügung, s.u. In den Studien von Jannidis und Weitin, die in dem vorliegenden Artikel beschrieben sind, werden Anwendungsbeispiele der Stilometrie vorgestellt. Das Ergebnis des stilistischen Vergleichs kann mit einer speziellen Netzwerkgrafik, dem Simmelian Backbone Network, mithilfe des Tools visone visualisiert werden. Für die Visualisierung kann auch die Clusteranalyse innerhalb von stylo oder ein Microsoft Excel©-Diagramm genutzt werden; hierzu sind in dem vorliegenden Buch ebenfalls Beispiele angegeben.

Mit dem Delta-mean Distanzmaß (vgl. Büttner 2017) werden die Distanzen zweier Texte wie folgt berechnet:

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right| \qquad \text{vereinfacht:} \qquad \Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - B_i}{\sigma_i}\right|$$

mit:
$\Delta_{(AB)}$ = Distanz der beiden zu vergleichenden Texte
n = Anzahl der MFWs (most frequent words)
A, B = die beiden zu vergleichenden Texte
Ai = relative Häufigkeit des Wortes i im Text A (= Anzahl Wort / Anzahl aller Wörter im Text)
Bi = relative Häufigkeit des Wortes i im Text B,
µi = relative (mittlere) Häufigkeit des Wortes i im Korpus
σi = Standardabweichung (der Häufigkeiten) des Wortes i im Korpus

In der Formel wird für jedes Wort i der Wortliste (MFWs) die relative Häufigkeit Ai (bzw. Bi für Text B) berechnet und dann der Mittelwert μi, also die relative mittlere Häufigkeit des Wortes i im Korpus, abgezogen. Diese Differenz wird normiert, indem sie durch die Standardabweichung (Streuung) des Wortes dividiert wird, um die Werte vergleichbar zu machen. Der Ausdruck (Ai-μi)/σi wird auch Z-Score (des Wortes) genannt und gibt an, um wie viele Standardabweichungen der betreffende Häufigkeitswert (des Wortes) vom Durchschnitt im Korpus abweicht.

Für die Distanz zwischen den beiden Texten (bezüglich aller Wörter der Wortliste) werden nun die Differenzen der beiden Z-Scores der beiden Texte gebildet, aufsummiert und durch die Anzahl der MFWs geteilt. Dies entspricht wieder einer Mittelwertbildung. Da die Differenzen auch negativ sein können, werden sie mit der Betragsbildung | | ins Positive gebracht.

Weitere Distanzmaße

Neben dem Delta-mean Distanzmaß (das auch Manhattan-Maß genannt wird, weil die Abstände rechtwinklig wie die Straße von Manhattan angeordnet sind) gibt es noch weitere, die für die Ähnlichkeit von Texten herangezogen werden können. Dazu gehören u.a. das Euklidische und das Cosinus Abstandsmaß. Folgend Grafik zeigt die drei Abstandsmaße im Vergleich als Vektoren im n-dimensionalen Raum (hier 2-dim.); n ist dabei die Anzahl der MFW.

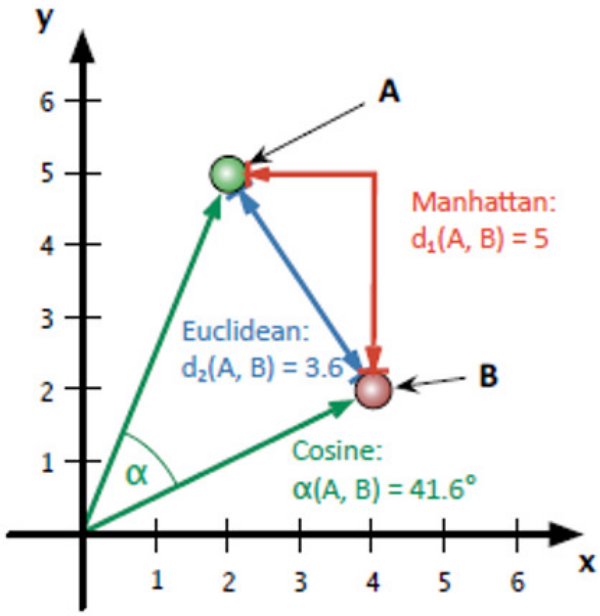

Abbildung: Vergleich der Abstandsmaße (Manhattan = Burrows Delta-mean, vgl. Weitin 2021, S. 62)

Anwendung der Stilometrie und Auswertung der Distanzen

Ein wesentlicher Unterschied zwischen der im Beitrag Femmer/Lucke genutzten NLP-Methode und der Stilometrie ist, dass bei der Stilometrie nur **ein** Kriterium für die Analyse genutzt wird (univariat) und zwar die Distanz auf Basis von Worthäufigkeiten, während bei der NLP-Methode mehrere Kriterien herangezogen werden (multivariat).

Bei der Analyse mit Stilometrie ist folgendes Vorgehen angebracht.

1. Als erstes werden die zu untersuchenden Texte heruntergeladen. Dazu bietet sich die Seite von Gutenberg an[2], unter der man aus einer Vielzahl von Literaturtexten den gewünschten Text kapitelweise per copy/paste runterladen und ihn in einer Datei im txt-Format speichern kann. Sämtliche Texte werden in einem Ordner corpus abgelegt.
2. Für die Nutzung des Statistikprogramms R werden das ‚System R' selbst und das GUI ‚RStudio' sowie das package ‚stylo' runtergeladen und installiert[3]. Eine gute Dokumentation findet man in forText[4] und auch in dem Buch von Prof. Weitin[5]. In diesem Buch ist ein Link enthalten, unter dem man die kompletten Projektdaten (Korpora, R-Skripte, Ergebnisdateien, Visualisierungen, u.s.w.) runterladen kann.
3. Zur Analyse wird ein R-Skript ausgeführt, das alle Texte des Korpus importiert, die Distanzen berechnet und die generierten Ergebnisdateien in einen result-Ordner schreibt. Dabei werden Konfigurationsparameter eingestellt, u.a. die MFW-Anzahl (z.B. 500), der Culling-Wert (20), die Auswahl des Distanzmaßes (Delta-mean), die Sprache (German) uva. Mit dem Cullingwert 20 wird festgelegt, dass ein MFW nur dann verwendet wird, wenn es in mindestens 80% der Texte vorkommt.

   Zu den von R erstellten Ergebnisdateien gehören

   a. eine txt-Datei mit der automatisch erzeugten wordlist der MFW; alternativ kann die wortlist auch mit eigenen Wörtern vorgegeben werden
   b. eine txt-Datei im csv-Format (character separated values) mit den relativen Häufigkeiten der MFW-Wörter pro Text
   c. die stylo-Konfigurationsdatei mit den eingestellten Parametern
   d. die Distanztabelle als Adjazenzmatrix im txt-Format
   e. optional eine csv-Datei mit den Z-Scores
4. Zum Schluss werden die Ergebnisse auf Basis der Distanztabelle visualisiert, wobei es dazu wieder viele Alternativen gibt.

Eine Variante der Visualisierung ist die Darstellung der Distanzen mit einer Microsoft Excel©-Tabelle, wobei die Distanzen zur besseren Ansicht mit 100 multipliziert werden. Die Zellen mit den Distanzwerten werden mit Graustufen oder farbig markiert (mittels bedingter Formatierung), wobei gilt, je dunkler ein Grau- bzw. Farbwert ist, desto größer ist die Distanz. Falls Farben genutzt werden, bedeuten grüne und helle Farben geringe, rote und dunkle Farben große Distanzen.

---

[2] URL https://www.projekt-gutenberg.org/

[3] Die Projektseite von R findet sich unter https://www.r-project.org/. Der Download der Installationsdatei ist unter https://cran.r-project.org/mirrors.html zu finden.
RStudio kann von https://posit.co/download/rstudio-desktop/ runtergeladen werden.

[4] https://fortext.net/

[5] Weitin, T. 2021: Digitale Literaturgeschichte, Springer Berlin Heidelberg

Texte	HauptmannE	HolzPapaHar	KafkaEinLan	KafkaEinTrau	KafkaErstesL	KafkaVorDer	MusilMannO	StifterGranit
HauptmannBahnv		96	95	110	112	119	78	94
HolzPapaHamlet	96		99	102	105	102	100	112
KafkaEinLandarzt	95	99		99	102	97	99	103
KafkaEinTraum	110	102	99		88	89	119	125
KafkaErstesLeid	112	105	102	88		88	111	118
KafkaVorDemGes	119	102	97	89	88		119	127
MusilMannOhnel	78	100	99	119	111	119		93
StifterGranit	94	112	103	125	118	127	93	

Abb. Distanzen mit Grautönen zwischen je zwei Texten mit Hilfe des Delta-mean Distanzmaßes (mit Excel selbst erstellte Abbildung auf Basis der Distanzdatei)

Texte	HauptmannE	HolzPapaHar	KafkaEinLan	KafkaEinTrau	KafkaErstesL	KafkaVorDer	MusilMannO	StifterGranit
HauptmannBahnv		96	95	110	112	119	78	94
HolzPapaHamlet	96		99	102	105	102	100	112
KafkaEinLandarzt	95	99		99	102	97	99	103
KafkaEinTraum	110	102	99		88	89	119	125
KafkaErstesLeid	112	105	102	88		88	111	118
KafkaVorDemGes	119	102	97	89	88		119	127
MusilMannOhnel	78	100	99	119	111	119		93
StifterGranit	94	112	103	125	118	127	93	

Abb. Distanzen mit Farben zwischen je zwei Texten mit Hilfe des Delta-mean Distanzmaßes (mit Excel selbst erstellte Abbildung auf Basis der Distanzdatei)

Kafkas Texte unterscheiden sich, wie im NLP-Programm, nicht nennenswert von Texten des Naturalismus (z.B. hat *Ein Landarzt* den Distanzwert 95 zu Hauptmanns Erzählung *Bahnwärter Thiel)*, aber schon stärker vom Realismus (Stifters *Granit*, hier findet sich der Wert 103). Unter den Texten Kafkas in Bezug auf die Ähnlichkeit zu anderen Texten zeigt sich eine hohe Varianz. So unterscheidet sich die Erzählung *Vor dem Gesetz* (Türhüter-Legende) nach der Messung mit der Stilometrie-Methode am meisten von dem Realismustext *Granit* (Distanz 127) und auch stark von den naturalistischen Erzählungen (*Bahnwärter Thiel*: 119 und *Papa Hamlet*: 102) und sogar mit einem Wert von 119 von dem Musil-Text *Mann ohne Eigenschaften*, der derselben Epoche (Moderne) zugeordnet wird. Das mag ein Hinweis auf die vielerorts wahrgenommene Besonderheit dieses Textes von Kafka sein, der im literaturwissenschaftlichen Kontext viel diskutiert wurde und methodologische Debatten im Bereich der Literaturwissenschaft (in der Folge seiner paradigmatischen Inanspruchnahme durch Derridas exemplifizierenden Textanalyse) ausgelöst hat. Auch im NLP zeigt sich bei diesem Text eine auffällige sprachliche Besonderheit in Bezug auf die Kategorie „long sentences“. Die Distanzen zwischen Musils Text zu den Texten des Naturalismus (*Bahnwärter Thiel* mit Distanz 78 und *Papa Hamlet* mit 100) sind prototypisch für Texte der Moderne.

Die sogen. ‚Clusteranalyse‘ des R-stylo-Programms zeigt das Ergebnis als Dendogramm-Plot. Dabei gilt: „Texte am gleichen Ast sind sich (im Sinne der Stilometrie) stilistisch ähnlich. Je mehr Gabelungen zwischen zwei Texten liegen, desto unähnlicher sind sie sich“, vgl. forText[6] oder Rißler2010[7]

[6] https://fortext.net/routinen/lerneinheiten/stilometrie-mit-stylo

[7] Rißler-Pipka, Nanette: Quantitative Textanalyse/Stilometrie. © 2010 Universität Tübingen, URL: https://d-nb.info/1140764195/34

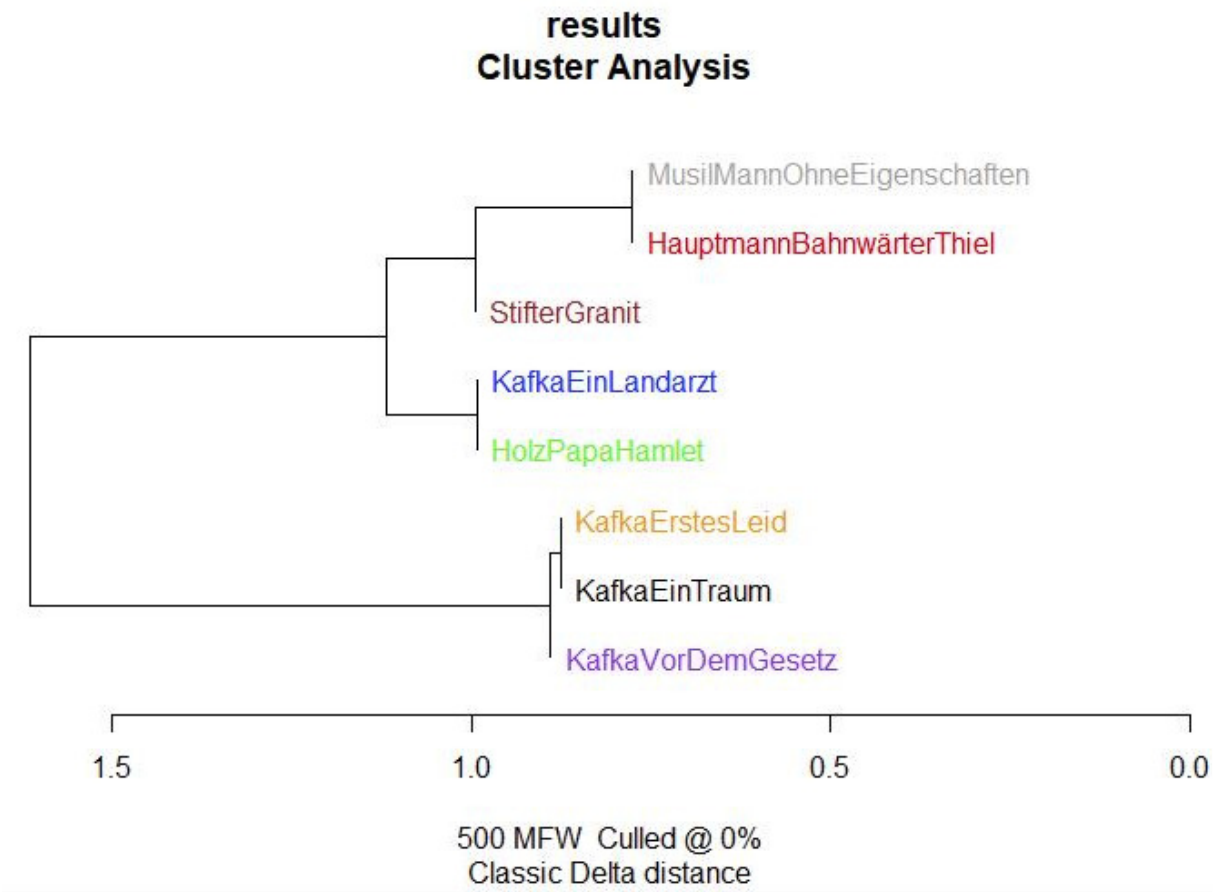

Abb. Clusteranalyse für Korpus mit 8 Literaturtexten (eigene Abbildung erstellt mit R, RStudio und der Visualisierung mit stylo-plot)

Für einen **erweiterten** Korpus mit zusätzlichen Literaturtexten zeigt die Clusteranalyse folgendes Dendogramm:

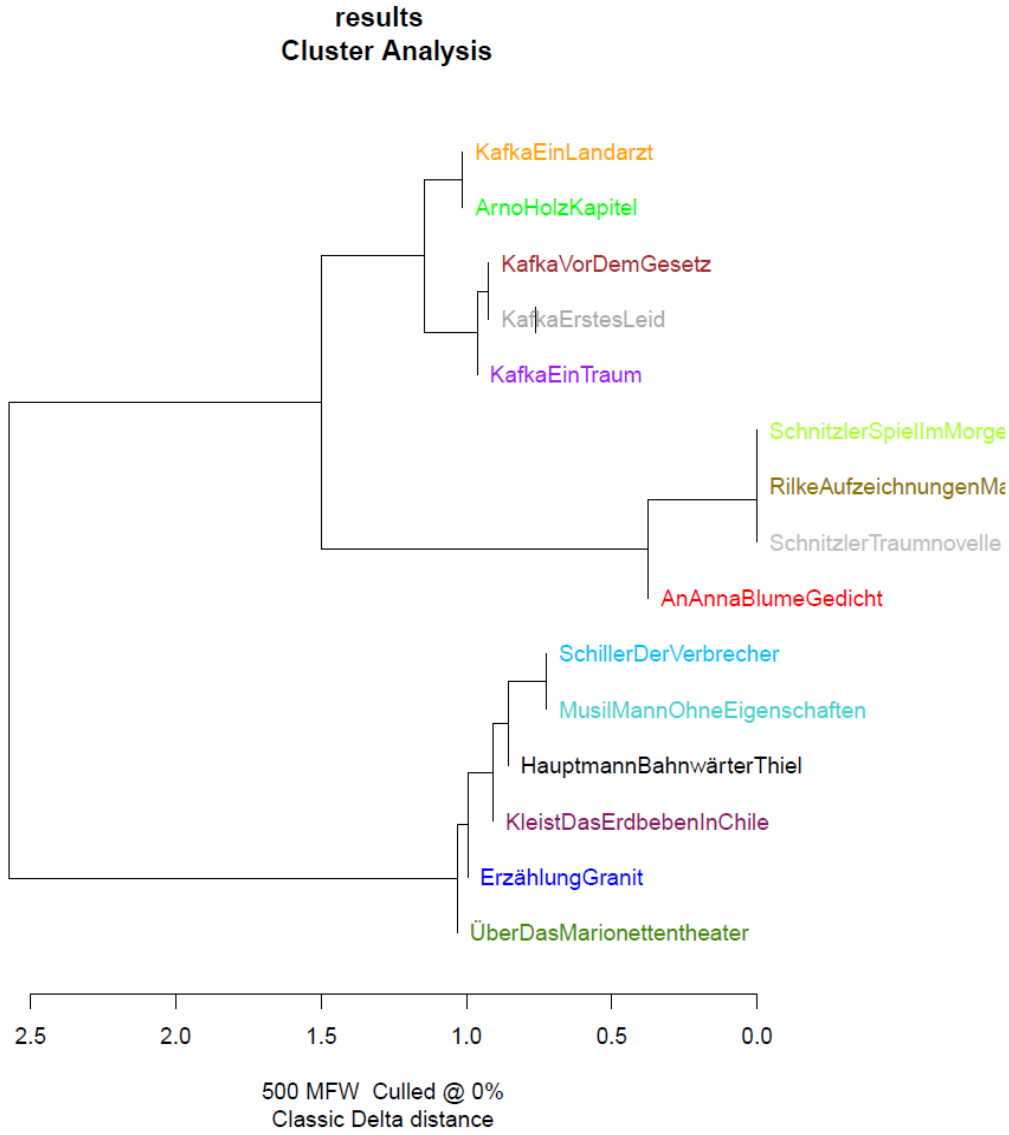

Abb. Clusteranalyse für erweiterten Korpus (eigene Abbildung erstellt mit R, RStudio und der Visualisierung mit stylo-plot).

Eine weitere Alternative zur Visualisierung der Distanzen ist die Nutzung der Netzwerkanalyse. Dazu bieten sich einige Visualisierungstools an, z.B. Visone[8] oder Gephi[9], die aus dem Bereich der Soziologie kommen. Als unterlagertes Netzwerk hat sich das Simmelian Backbone Network[10] bewährt, das Ähnlichkeiten von Novellen, Romanen usw. auf Basis von Distanzen besonders anschaulich anzeigt.

In dem folgenden Beispiel werden die Farbstufen der Knoten und Kanten nach den Delta-Werten skaliert. An den Kanten liegt der Wert für die paarweise Ähnlichkeit zweier Texte an; eine dunkle, dicke Verbindung zeigt eine große, eine helle wenig Ähnlichkeit. An den Knoten wird die durchschnittliche Ähnlichkeit jedes einzelnen Textes zum Gesamtkorpus (Delta-mean) markiert, wobei ein dunkler Knoten hohe Korpusähnlichkeit abbildet, ein heller Knoten

[8] https://visone.ethz.ch/html/about.html
[9] https://gephi.org/users/
[10] Nick, Bobo, Conrad Lee, Pádraig Cunningham, und Ulrik Brandes (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, 525–532. ASONAM ’13. New York, NY: ACM.

dagegen anzeigt, dass diese Novelle sich vom Mainstream der Textsammlung abhebt (vgl. Weitin 2021, S. 25)

Für die Visualisierung der Distanzen mit visone sollte die txt-Distanztabelle (500 MFW und 20% Culling) noch in eine csv-Datei geändert und eine Attributdatei erstellt werden; für beides steht bei Weitin je ein R-script zur Verfügung.

Visone liegt als Java-jar-Datei vor und wird mit

java -jar <visone.jar>, z.B. java -jar visone-2.25.jar

gestartet. Dabei ist <visone.jar> der Name der aktuellen jar-Datei von Visone, z.B. visone-2.25.jar. Die jar-Datei muss sich in dem aktuellen Ordner befinden, ansonsten muss der Pfad angegeben werden. Als erstes werden die Distanzdatei (distance_table_500mfw_20c.csv) und die Attributdatei (delta-mean.csv) importiert, wobei folgende Einstellungen vorgenommen werden:

data format = adjacency matrix
delimiter= space
encoding = UTF-8 statt Windows Default.

Beim Import wird eine Grafik erzeugt, die durch Modifikation der Parameter angepasst werden kann, so dass sie am Ende folgendes Simmelian Backbone Network darstellt:

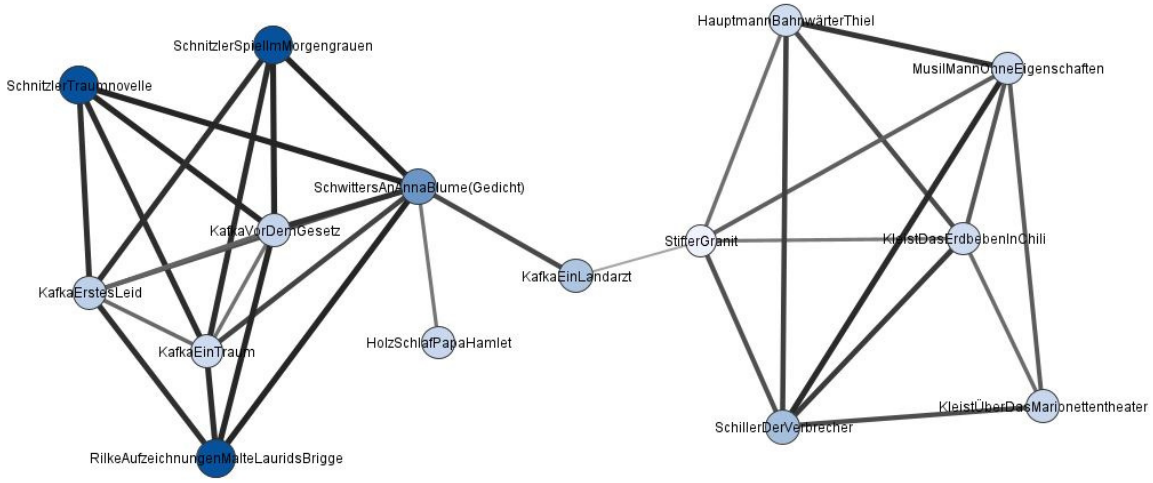

Abb. Visualisierung der Distanzen zwischen je zwei Texten mit dem Simmelian Backbone Network von 15 ausgewählten deutschsprachigen Texten (s.o.) mit den 500 häufigsten Wörtern; (Delta-mean Distanz, 500 MFW, Culled 0%, eigene Abbildung mit Visone erstellt)

In einem Download-Bereich sind weitere Korpora enthalten, ein Korpus mit 8 Texten (Korpus3), (sowie ein weiterer Korpus mit 29 Literaturtexte. Für sämtliche Korpora stehen unter diesem Link alle Analyseergebnisse, Visualisierungen und R-Skripte als Dateien zum Download bereit.

Vergleich zur NLP-Analyse.

Beim Vergleich der beiden Methoden NLP und Stilometrie finden wir teilweise nur eine geringe Übereinstimmung der Ergebnisse.

In der Cluster Analysis sind KafkaEinLandarzt und HolzPapaHamlet direkt nebeneinander und damit sehr ähnlich. In der NLP-Methode bei Imprecise Phrases dagegen haben diese beiden Texte die größte Differenz: 3,4 % und 6,4 %. Bei Smell Density finden wir bei NLP 80,1 % und 55,4 %.