Methodenbeitrag: Textvisualisierung

Jan Horstmann [1]
Jan-Erik Stange [1]
1. Universität Münster

Thema:	Textvisualisierung	DOI:	10.48694/fortext.3772
Jahrgang:	1	Ausgabe:	5
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2018-10-05 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition

Die Textvisualisierung als Teilbereich der Informationsvisualisierung befasst sich mit der visuellen Repräsentation komplexer Textdaten und der Manipulierbarkeit dieser Repräsentation durch interaktive Softwareinterfaces (Card, Mackinlay und Shneiderman 1999). Visuelle Darstellungen können neue Einsichten in Textdaten und deren innere Zusammenhänge liefern. Textvisualisierungen unterstützen sowohl die Kommunikation von Forschungsergebnissen als auch die explorative Analysetätigkeit. Hierbei sind grundlegend drei verschiedene Arten von Daten zu unterscheiden, die für die Visualisierung herangezogen werden können:

1. Der unbearbeitete Text, aus dem mithilfe statistischer Verfahren oder Natural Language Processing Datensätze generiert (vgl. Text Mining) werden. Ein Beispiel hierfür ist die Berechnung von Worthäufigkeiten.
2. Mit Zusatzinformationen (vgl. Metadaten), d. h. mit manuell oder automatisch erzeugten Annotationen (vgl. Annotation) angereicherter Text.
3. Textexterne Metadaten (wie Erscheinungsdatum, Autor*innenangaben, Titel etc.) oder andere Daten, die sich mit Annotationen verknüpfen lassen (etwa Geokoordinaten mit annotierten Ortsnennungen).

In der Praxis kommen zudem häufig auch Kombinationen dieser verschiedenen Datenarten zum Einsatz, die mit unterschiedlichen Visualisierungen jeweils unterschiedliche Perspektiven auf den Text (oder die Textsammlung) ermöglichen.

2. Anwendungsbeispiel

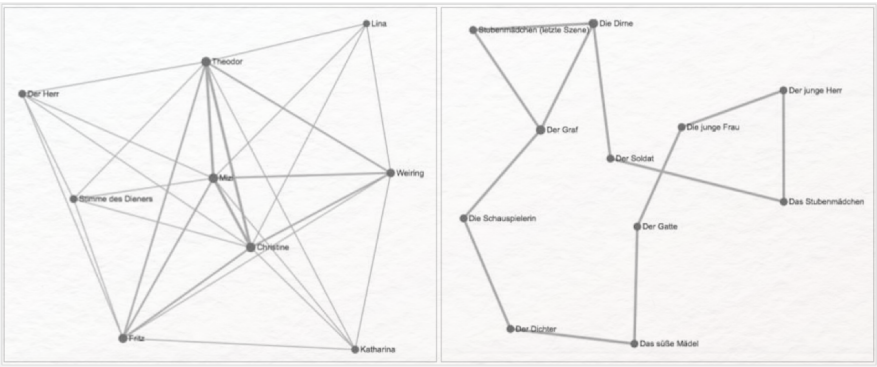

Abb. 1: Figurennetzwerke für Arthur Schnitzlers Dramen Liebelei *(links) und* Reigen *(rechts). Quelle: https://dracor.org/ger*

Sie erforschen Figurenkonstellationen in deutschsprachigen Dramen um 1900. Um deutlich zu machen, dass die Interaktionen der Figuren in den einzelnen Dramen sehr unterschiedlicher Art sind, entscheiden Sie sich, die Figurenkonstellationen als Netzwerke zu visualisieren und einander gegenüber zu stellen. Auch Betrachter*innen, die beispielsweise Arthur Schnitzlers *Reigen* nicht gelesen haben, wird so die besondere Interaktionsstruktur dieses Dramas augenfällig. Weil Sie außerdem deutlich machen wollen, wann im Verlauf der Dramen die jeweili-

gen Figuren auftreten, entscheiden Sie sich für eine Visualisierung der Figurendistribution über den Textverlauf als Balkendiagramm. Für den *Reigen* sieht das beispielsweise so aus:

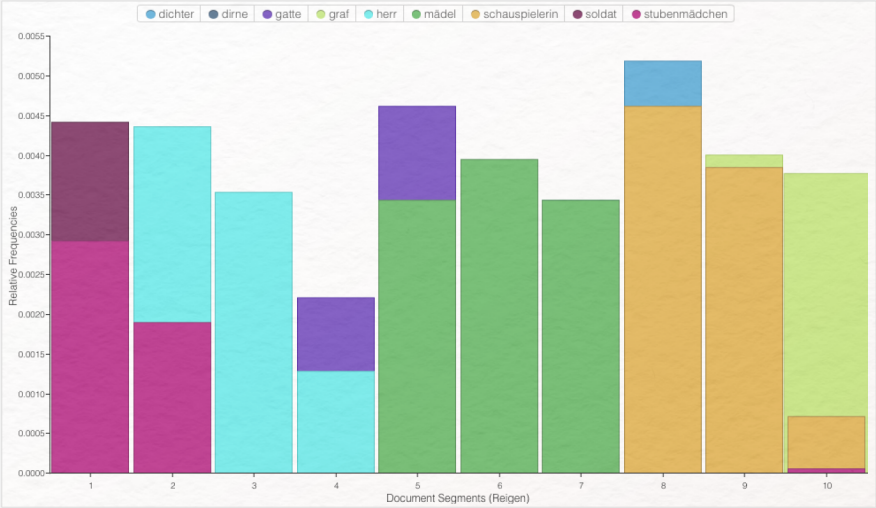

Abb. 2: Stacked Barchart für die Figuren in Arthur Schnitzlers Reigen. *Erzeugt mit https://voyant-tools.org*

Viele weitere Visualisierungsmöglichkeiten werden Ihnen einen jeweils anders gearteten Blick auf Ihre Textsammlung (vgl. Korpus) oder einzelne Texte darin ermöglichen. Je nach Fragestellung sind einzelne Visualisierungsmöglichkeiten mehr oder weniger geeignet. Hermeneutisch anspruchsvollere Fragen lassen sich häufig in der Kombination möglichst reichhaltiger Textannotationen mit den statistischen Textdaten selbst verfolgen und Visualisierungen dieser kombinierten unterschiedlichen Datentypen halten häufig literaturwissenschaftlich relevante Erkenntnisse bereit.

3. Literaturwissenschaftliche Tradition

Die Vorteile, die es mit sich bringt, (textliche) Phänomene zu visualisieren und damit zu veranschaulichen, wurden in der Literatur (und dann auch in der Literaturwissenschaft) schon früh erkannt. Schon Goethe stellte in seinen *Naturwissenschaftlichen Schriften* zur Erkenntnisfunktion von visuellen Eindrücken fest: „Das Ohr ist stumm, der Mund ist taub; aber das Auge vernimmt und spricht. In ihm spiegelt sich von außen die Welt, von innen der Mensch“ (Goethe 1987 WA II, 5 (2), 12).
Solange es Texte gibt, wurden diese mit bildlichen Darstellungen kombiniert und zu diesen in Bezug gesetzt, besonders kunstvoll in der Buchmalerei der Spätantike und des Mittelalters, die mit ihren Bordüren und Drolerien weit über verzierte Initialen hinaus ging und Textinhalt und Verzierung eng miteinander verknüpfte. Wenn eine zeitgenössische Autorin wie Cornelia Funke ihre Bücher mit Illustrationen der darin vorkommenden Figuren und Orte versieht, führt sie diese Tradition fort. Aber auch bereits die Verwendung unterschiedlicher Schriftarten oder -farben (wie beispielsweise in Michael Endes Roman *Die unendliche Geschichte* [1979]) oder die häufige Verwendung von Kursivierungen in John Irvings Romanen gibt visuellen Elementen die Möglichkeit, Textinhalte semantisch anzureichern (Kammer 2014). Künstlerisch in das Zentrum der Aufmerksamkeit gestellt wird die Verbindung von Text und Bild in der ihrerseits auf eine lange Tradition zurückblickenden visuellen Dichtung (Adler und Ernst 1987), bei der die räumliche Verortung von Wörtern oder Wortgruppen bereits zur Aussage des jeweiligen Gedichtes beiträgt.
Diese Formen von Text-Bild-Relationen beziehen sich jedoch auf literarische Primärtexte und nicht selten auf die Präsentation des gesamten Textes. Definieren wir Textvisualisierung jedoch als visuelle *Repräsentation* von Textdaten, wie wir es oben getan haben, tritt die Sekundärliteratur in den Fokus der Aufmerksamkeit, innerhalb derer diese Repräsentation i. d. R. stattfindet.
Die visuelle Repräsentation von Textdaten (oder auch von Metatextdaten) findet in der literaturwissenschaftlichen Forschung mehrere Vorgänger. Einige davon sind so alltäglich, dass sie uns kaum noch als Textvisualisierung bewusst werden, wie z. B. in der Versanalyse, in der es standardisierte Darstellungsweisen bspw. für Hebungen, Senkungen und Zäsuren in der Versfußnotation oder für akzentuierte Silben, Pausen, Taktgrenzen etc. in der Silbenakzentnotation gibt.

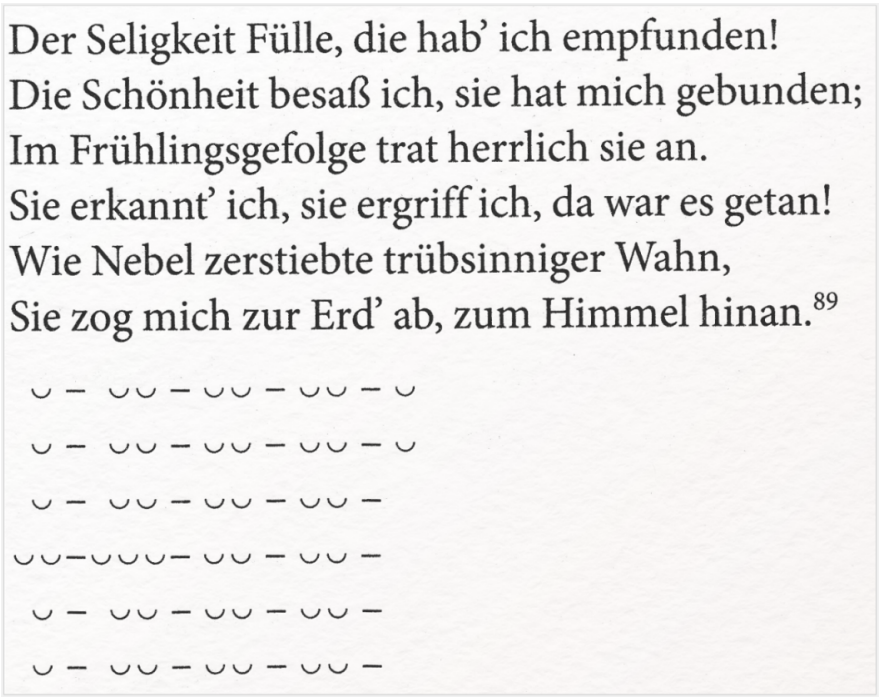

Abb. 3: Visualisierung einer lyrischen Sequenz aus Goethes Drama Pandora*; Mellmann (2007, 88)*

Ebenso häufig werden Figurenkonstellationen als Netzwerke visualisiert. Eder (2007, 239) nennt als mögliche Ebenen der Verbindungen zwischen den einzelnen Figuren 1. handlungsfunktionale Verhältnisse, 2. körperliche, psychische, soziale, symbolische Ähnlichkeiten und Kontraste und 3. die „Positionen innerhalb einer Aufmerksamkeitshierarchie". Damit beschreibt er nichts anderes als die Knoten (= Figuren) und Kanten (= Verbindungen) in einer Netzwerkvisualisierung (vgl. Netzwerkanalyse (Schumacher 2024b)). Die im Anwendungsbeispiel oben dargestellten Netzwerke zeigen jedoch noch eine andere Form von Verbindungen, nämlich die Kopräsenz von Figuren im jeweiligen Drama. In der Dicke der Kanten spiegelt sich dann die von Eder benannte Aufmerksamkeitshierarchie. Ähnlich werden gelegentlich auch Abstammungsverhältnisse literarischer Figuren nicht nur sprachlich erläutert, sondern visuell als Stammbaum dargestellt. Dies geschieht auch mit einzelnen Überlieferungsstufen älterer Texte wie des Nibelungenliedes:

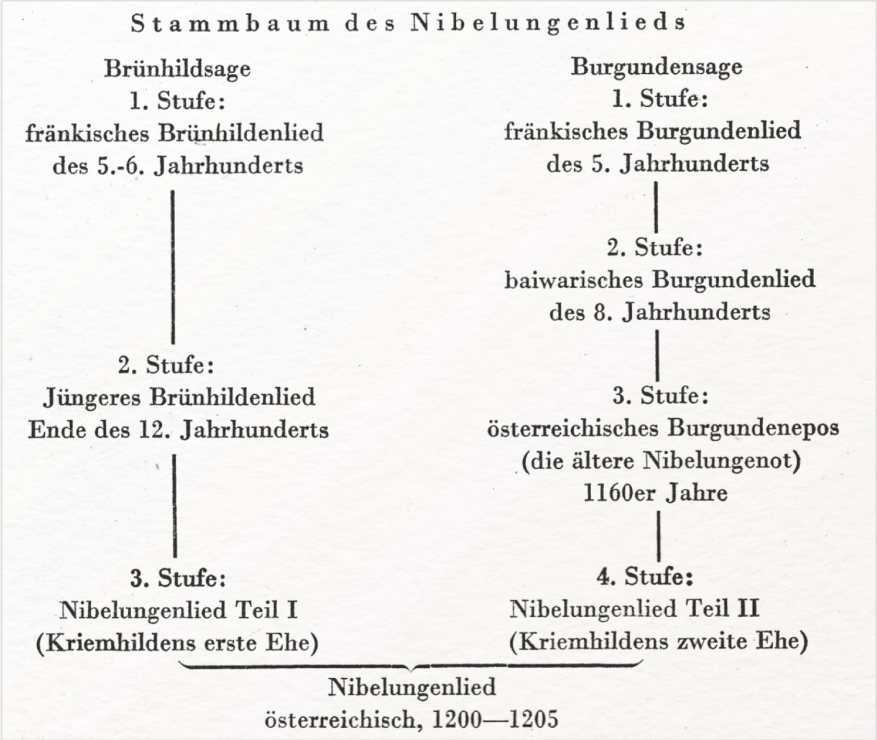

Abb. 4: Stammbaum des Nibelungenlieds bei Heusler (1955, 49)

Häufig nutzt die Forschung auch abstrahierende Darstellungen z. B. von Kommunikationsverhältnissen (wie sie besonders in der Erzähltheorie beliebt sind) oder veranschaulichende Visualisierungen von Zusammenhängen struktureller oder kategorialer Merkmale (wie bspw. von bildhafter vs. bildlicher Versprachlichung von Schmerz in literarischen Texten als Baumdiagramm).

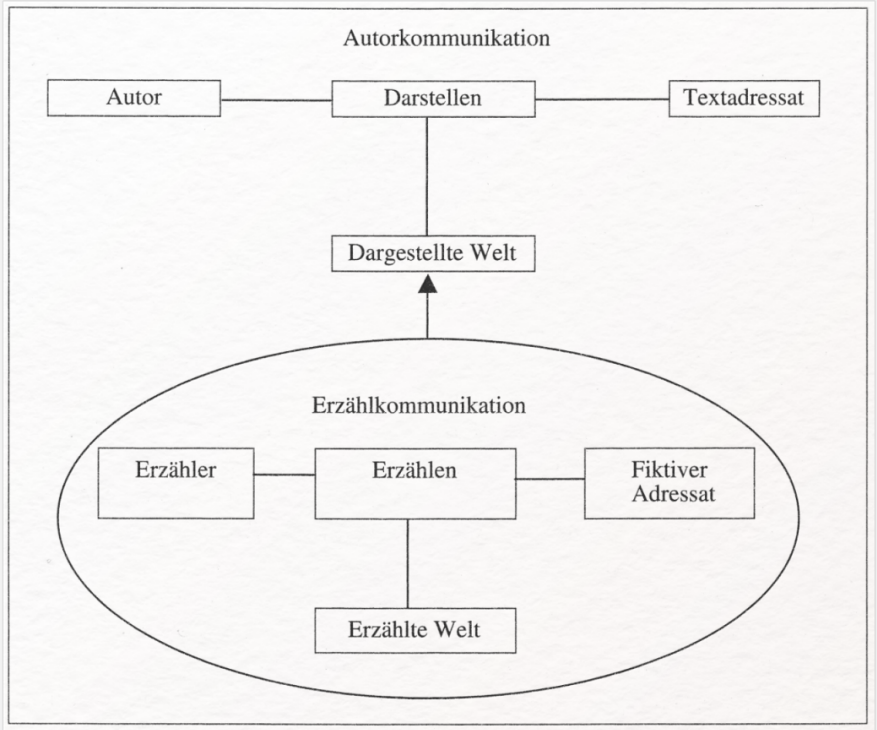

Abb. 5: Visualisierung von Autor- und Erzählkommunikation, Schmid (2007, 176)

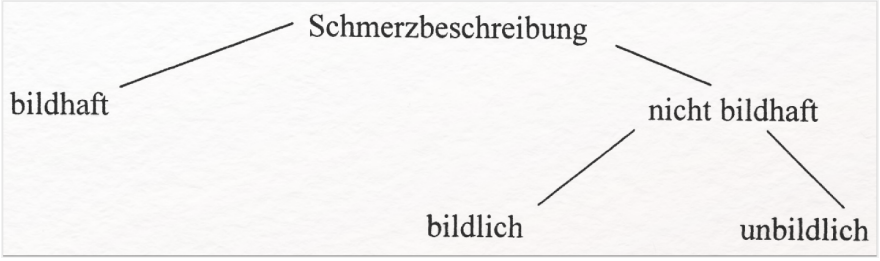

Abb. 6: Visualisierung der Typen von Schmerzbeschreibung, Koller (2000, 129)

Ebenfalls werden in der Forschung Karten oder topografische Repräsentationen von sprachlich erzeugtem Raum bzw. der Bewegung in diesem Raum erstellt und reflektiert – so vergleicht bspw. Ryan (2003) Visualisierungen kognitiver Karten auf Grundlage von García Márquez' Roman *Crónica de una muerte anunciada* (1981), und Piatti (2008) mappt in ihrer Monografie – der 17 Karten beiliegen – fiktive Handlungsräume auf tatsächliche Georäume. Gelegentlich kommt es zudem vor, dass literarische Ekphrasen in der zugehörigen Forschungsliteratur visualisiert werden, d. h. ursprünglich sprachlich beschriebene Bilder erfahren eine direkte visuelle Umsetzung, gelegentlich auch in abstrahierter Form wie bspw. das Schild des Achilles – eine der ältesten und berühmtesten Ekphrasen überhaupt – bei Willcock (1976, 210):

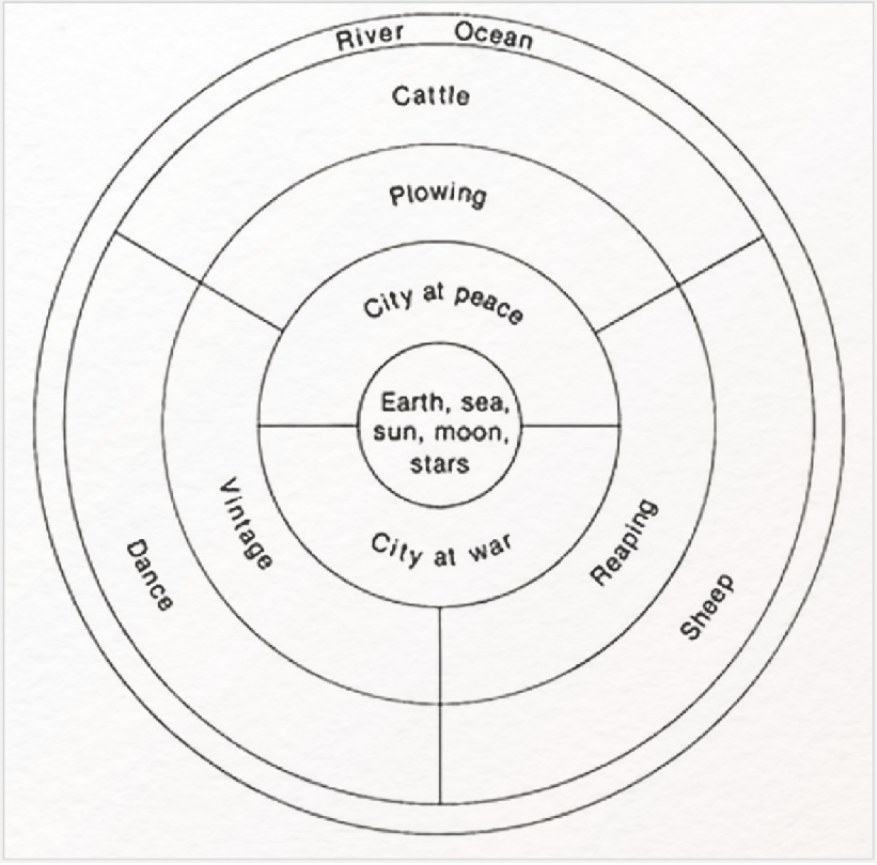

Abb. 7: Abstrahierende Visualisierung von Achilles' Schild bei Willcock (1976, 210)

Es lässt sich in der traditionelleren Literaturwissenschaft eine klare Tendenz hin zu qualitativen Textvisualisierungen ausmachen. In den Digital Humanities sind es vor allem die quantitativen Strukturen von Textdaten, die visualisiert werden. Die digitale Literaturwissenschaft ist momentan daher darum bemüht, Visualisierungsmöglichkeiten zu entwickeln, die qualitative und quantitative Aspekte der Textdaten fruchtbar miteinander verknüpfen (vgl. z. B. das Projekt 3DH), und in diesem Bereich gibt es noch viel Entwicklungspotenzial.

4. Diskussion

Wenngleich analoge Visualisierungen, wie wir gesehen haben, auf eine längere Tradition zurückblicken, so müssen Textvisualisierungen im Sinne der im ersten Abschnitt angeführten Definition als digitale, textdatengetriebene Visualisierungen als ein relativ junges Phänomen in der Literaturwissenschaft verstanden werden. Da die visuelle Repräsentation quantitativer Textdaten im Analogen einen erheblich größeren Aufwand bedeuten würde, ermöglichen die digitalen Methoden neue Perspektiven. Der Hauptvorteil, der durch das digitale Medium geschaffen wird, besteht in der direkten Interaktion und der dadurch geschaffenen Möglichkeit der Exploration. Seifert u. a. (2014) beschreiben Textvisualisierung als „an effective enabler for exploratory analysis, making it a powerful tool for gaining insight into unexplored data sets." Darüberhinaus erlaubt Interaktivität flüssige Bewegungen zwischen Übersichtsdarstellungen, der Darstellung von Teilbereichen und Details (Shneiderman 1996).
Das Setzen von Filtern, das Selektieren von Teilbereichen, das Nebeneinanderstellen dieser verschiedenen Perspektiven auf die Daten ermöglicht es, auf schnelle Art und Weise Vergleiche herzustellen und Schlussfolgerungen zu ziehen (Card, Mackinlay und Shneiderman 1999). Die kognitive Belastung beschränkt sich hierbei auf ein Minimum, da die Textvisualisierung (bzw. allgemeiner Informationsvisualisierung) sich die herausragenden Fähigkeiten des visuellen Wahrnehmungssystems des Menschen zunutze macht (Ware 2012).
Die Vorteile digitaler Textvisualisierung treten dabei nicht nur im *Distant Reading* (vgl. Distant Reading) hervor, auch wenn dieser Bereich in den vergangenen Jahren eine vermehrte Aufmerksamkeit in der Anwendung von Visualisierungstechniken gefunden hat. Auch die im Zusammenhang mit einem *Distant Reading* (vgl. Distant Reading) digital erstellten Annotationen (Manuelle Annotation (Jacke 2024)) sind besonders bei einer hohen Anzahl von Annotationen in ihrem Zusammenhang nicht mehr übersichtlich darzustellen. Textvisualisierung kann hier eingesetzt werden, um eine komprimierte Übersichtsdarstellung zu geben, die nach unterschiedlichen Kriterien strukturiert werden kann.
Typische Anwendungsbereiche von Textvisualisierungen im *Distant Reading* sind beispielsweise stilometrische Analysen (Stilometrie (Horstmann 2024a)), *Häufigkeitsanalysen* (von Wörtern, N-Gramms (vgl. N-Gramm), Annotationen), Darstellungen von Verteilungen von Wörtern oder Annotationen über Textlängen, zeitliches Erscheinen von Werken einer Textsammlung, Netzwerkdiagramme von Figuren, Geovisualisierungen von Ortsnennungen in Texten oder Verteilung von Themen über eine Textsammlung im Topic Modeling. Abbildung 8 zeigt beispielhaft Visualisierungen für diese unterschiedlichen Anwendungsfälle.

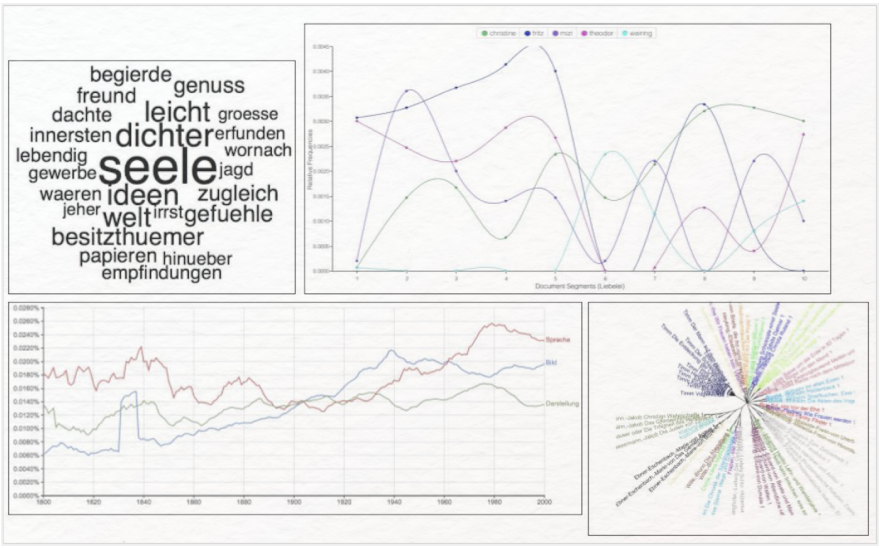

Abb. 8: Visualisierungsbeispiele: Topic Cloud, Distributionsgraphen und stilometrische Analyse

Eine Übersicht über gängige Visualisierungsformen in der digitalen Literaturwissenschaft in den Bereichen *Close Reading* und *Distant Reading* geben Jänicke u. a. (2015). Der Einsatz von Visualisierungen in der digitalen Literaturwissenschaft und den Digital Humanities generell wird jedoch auch kritisch betrachtet. Ein häufig geäußerter Einwand bezieht sich auf die visuelle Darstellung von Ergebnissen literaturwissenschaftlicher Arbeit als vermeintlich objektive Fakten. Drucker (2011) schlägt etwa vor, die geisteswissenschaftliche Konstruiertheit („constructedness") von Daten stärker zu berücksichtigen (in diesem Zusammenhang spricht sie von „capta" statt „data") und geeignete visuelle Darstellungsformen zu verwenden, die in der Lage sind, typische literaturwissenschaftliche Deutungsdimensionen wie Unsicherheit und Mehrdeutigkeit zu kommunizieren. Zunehmend lässt sich ein Zusammenwachsen (vgl. Scalable Reading) des vormaligen Gegensatzpaares quantitativ/qualitativ feststellen. Quantitativ erzeugte Daten werden so etwa durch unterschiedliche Visualisierungen einer kritischen Betrachtungsweise zugänglich, während qualitative Daten in der distanzierten Übersicht dargestellt einen

schnellen Zugriff auf die zugrundeliegenden annotierten Textstellen erlauben.
Auch wenn einfach zu benutzende generische Visualisierungstools grundsätzlich zu begrüßen sind, so stellen sie auch eine gewisse Gefahr da. Gerade bei fehlenden Grundlagenkenntnissen im Bereich der Text- bzw. Informationsvisualisierung können Visualisierungen unwissentlich zu einer Verzerrung der produzierten Ergebnisse führen, z. B. durch die Verwendung eines Kuchendiagramms für Prozentwerte, die nicht Teile eines Ganzen darstellen. Außerdem können Farben unglücklich eingesetzt, Datenattribute mit wenig geeigneten visuellen Variablen repräsentiert oder ganz generell Visualisierungsformen verwendet werden, die eine bestimmte Datenstruktur voraussetzen und bei abweichenden Strukturen eine verzerrende Aussage zur Folge haben können.
Insgesamt lässt sich konstatieren, dass das Potenzial interaktiver Textvisualisierungstools noch nicht ausgeschöpft ist. So wird Textvisualisierung in der Praxis vornehmlich als Repräsentation eines vermeintlichen Endergebnisses eines Forschungsprozesses eingesetzt, während der Prozess selbst, den häufig die Produktion einer Vielzahl von Visualisierungen und deren Interpretation ausmacht, ignoriert wird. Die Stärke von Visualisierungen nicht nur zu kommunizieren, sondern auch Exploration zu ermöglichen und das Entstehen von Schlussfolgerungen zu befördern, wird hier nur im Prozess für die Forschenden sichtbar, nicht jedoch für das spätere Publikum der Forschungsarbeit. Ein vermehrter Einsatz von Visualisierungen auch als wesentlicher Teil der Argumentation steht derzeit noch aus.

5. Technische Grundlagen

Die Bandbreite an Software, mit der sich Textvisualisierungen erzeugen lassen, ist groß. Grundlegend lassen sich generische Tools und Programmiersprachen unterscheiden. Generische Tools bieten unterschiedliche Standardvisualisierungsformen an, für ihre Nutzung sind in der Regel keine tiefergehenden Programmierkenntnisse notwendig. Programmiersprachen hingegen erlauben statistische Auswertungen oder eine sprachliche Analyse von Texten und/oder eine Ausgabe berechneter Daten in unterschiedlichen Visualisierungen.
In ihrer einfachsten Variante erlauben generische Tools eine unmittelbare automatische Analyse und Visualisierung von Texten. Ein solches Beispiel ist die niederschwellige Webapplikation Voyant (Flüh 2024), die beispielsweise mithilfe von *Distant Reading* (vgl. Distant Reading) Worthäufigkeiten visualisiert oder Verteilungen von Wörtern über den Gesamttext als Liniendiagramme darstellt. Manuell erstellte Annotationen in digitalen Texten lassen sich mit Tools wie CATMA (Schumacher 2024a) analysieren und auf verschiedene Weisen visualisieren.
Einige Programmiersprachen haben sich für die Textanalyse und -visualisierung in der Praxis besonders bewährt. Mit R lassen sich beispielsweise *Topic Models* generieren (mehr zu Topic Modeling siehe Horstmann (2024b)) und in Form von *Word Clouds* visualisieren. Python ist besonders für die Textanalyse im Bereich *Natural Language Processing* geeignet und erlaubt mit entsprechenden Bibliotheken auch eine Visualisierung der erzeugten Daten.
Mit der Javascript-Bibliothek D3 lassen sich unterschiedlichste Datensätze (nicht nur Textdaten) in vorgefertigten Visualisierungen darstellen, aber auch vollkommen neue Visualisierungen erstellen. Die Anwendung aller dieser Sprachen erfordert Programmierkenntnisse im Allgemeinen und grundlegende bis fortgeschrittene Kenntnisse der jeweiligen Sprache im Speziellen. Für eine Vielzahl dieser Tools existiert im Web eine umfangreiche Sammlung an Tutorials und Dokumentationen, die als Lernmaterial genutzt werden können. Weiterführende Literatur und Links finden Sie am Ende des Artikels. Für einen adäquaten Einsatz von Visualisierungstools ist die Kenntnis einiger Grundprinzipien der Textvisualisierung entscheidend. Zunächst sollten Sie sich fragen, was für Daten Sie repräsentieren möchten. Abhängig davon, was Sie an einem Text oder einer Textsammlung analysieren möchten, arbeiten Sie mit unterschiedlichen Datensätzen (siehe Abschnitt 1). Normalerweise besteht ein Datensatz aus einzelnen Entitäten (in der Regel Zeilen in der Datentabelle), die sich wiederum aus einzelnen Datenattributen zusammensetzen. Menge und Art der Attribute haben Einfluss auf die Visualisierungstypen, die für Ihre Visualisierung in Betracht kommen.
Man unterscheidet drei grundlegende Datenattribute: quantitativ, kategorisch (auch: nominal) und ordinal. Unter quantitativen Daten versteht man numerische Daten, mit denen Berechnungen angestellt werden können, wie beispielsweise die Häufigkeit von Wörtern in einem Text. Geografische oder zeitliche Daten können als Sonderform quantitativer Daten betrachtet werden, die gewissen formalen Beschränkungen unterliegt. Im Gegensatz dazu spricht man von kategorischen Daten, wenn diese nicht-numerisch sind und voneinander unabhängige Werte darstellen. Ein Beispiel hierfür sind etwa die unterschiedlichen Namen der Figuren in einem Text. Ordinale Daten ähneln kategorischen Daten, unterscheiden sich aber durch die in Ihnen zum Ausdruck kommende Reihenfolge. Eine Bewertung wie „klein, mittel, groß“ fällt beispielsweise unter diese Bezeichnung. Die bei den jeweiligen Datenattributen auftretenden Werte können zudem Relationen untereinander haben, die ebenfalls im Datensatz kodiert sein können. Um diese unterschiedlichen Datenattribute zu repräsentieren, stehen verschiedene sogenannte „Visuelle Variablen“ (Bertin 1974) zur Verfügung, von denen die wichtigsten in Abbildung 9 aufgeführt sind.

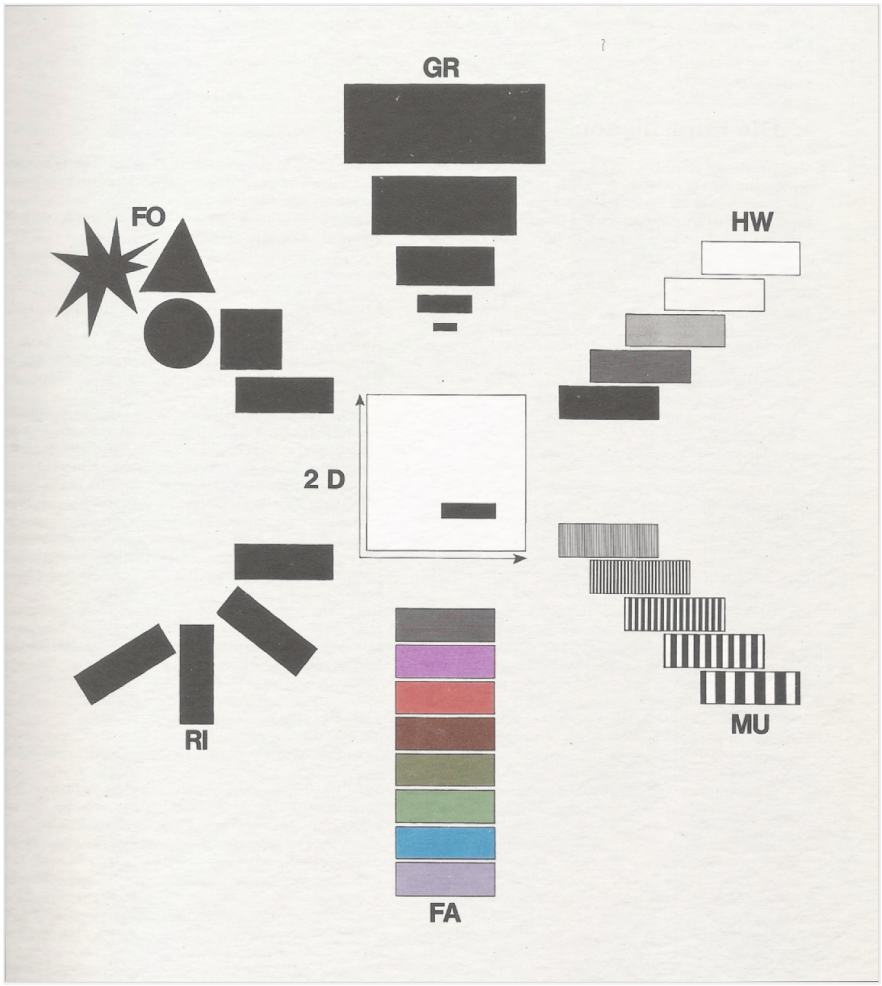

Abb. 9: Visuelle Variablen nach Bertin (1974, 51)

Im einzelnen sind dies: Form, Größe, Helligkeitswert, Muster, Farbe, Richtung und Position. Um visuelle Variablen und Gestaltgesetze sinnvoll anwenden zu können, sind einige Regeln zu beachten. Je nach Datenattribut sind visuelle Variablen unterschiedlich gut geeignet, um Datenwerte zu repräsentieren: Während sich z. B. die visuelle Variable Farbe besonders eignet, um kategorische Werte darzustellen, so ist sie nicht geeignet, um quantitative Werte zu kodieren. In der untenstehenden Tabelle ist eine Übersicht über sinnvolle Kodierungen als visuelle Variablen der einzelnen Datenattribute in Spalten abgebildet (mit von oben nach unten abnehmender Eignung). Die genannten visuellen Variablen wurden durch weitere ergänzt, die in Spezialfällen zur Anwendung kommen können. Im Normalfall ist die oben dargestellte Auswahl von sieben Variablen ausreichend.

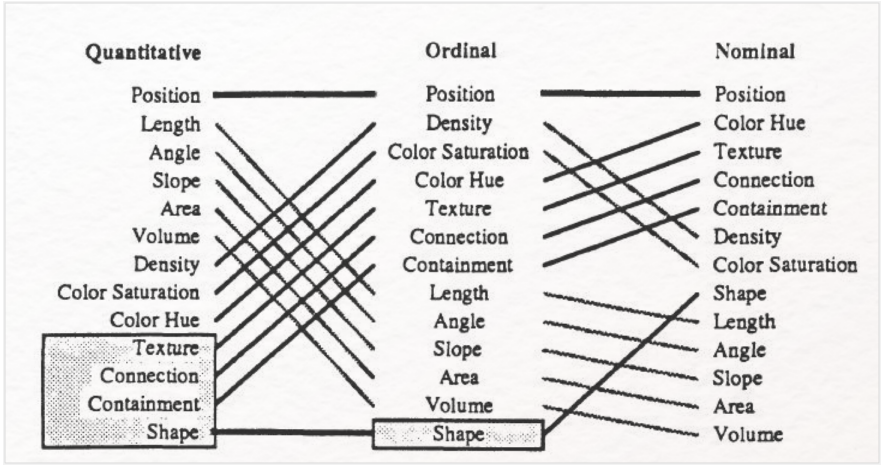

Abb. 10: Eignung ausgewählter visueller Variablen für verschiedene Datenattribute, von oben nach unten abnehmend (Mackinlay 1986)

Visualisierungen, die Relationen darstellen, wie beispielsweise Netzwerke oder Hierarchien, orientieren sich an den sog. Gestaltgesetzen. Für die Textvisualisierung sind hier besonders die Gesetze *Nähe, Ähnlichkeit, Verbundenheit* und *Gemeinsame Regionen* wichtig:

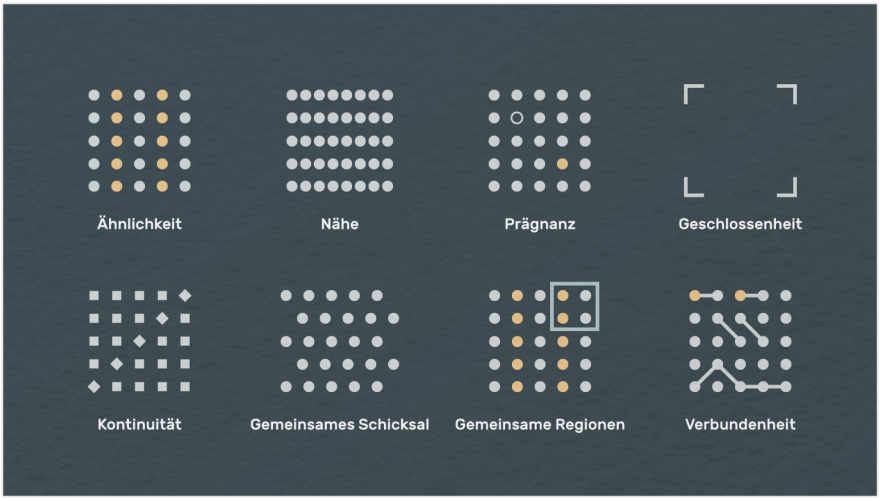

Abb. 11: Gestaltgesetze nach Wertheimer 1923. Quelle: https://userinterfacedesign.ch/gestaltgesetze/

Auch im Bereich der *Interaktion* haben sich im Laufe der Zeit einige Techniken etabliert, deren Eignung sich in der Praxis besonders bewährt hat Normalerweise gestatten generische Tools nur sehr eingeschränkt eine Beeinflussung der Interaktionen, die mit einer Textvisualisierung möglich sind. Auch in *R* und *Python* findet die Interaktion eher auf Code-Ebene statt. Die Gestaltung und Programmierung eines vollständigen User-Interfaces, bei dem die Interaktionen frei zu bestimmen sind, kann in der Regel nur unter Beteiligung von Programmierer*innen und Designer*innen mit Webtechnologien wie HTML/CSS und Javascript umgesetzt werden. Dennoch sollen die wichtigsten wissenschaftlichen Erkenntnisse kurz vorgestellt werden.
Ein Prinzip, das sich in der Praxis immer wieder bewährt hat, ist das sogenannte *Shneiderman's Visual Information Seeking Mantra*: „Overview first, zoom and filter, then details-on-demand" (Shneiderman 1996). Visualisierungen sollten zunächst einen Überblick über die Daten vermitteln, durch Interaktionen wie Zoomen und Filtern die Auswahl eines Teilbereiches erlauben und schließlich Detailinformationen für einzelne Datenpunkte bei Bedarf anzeigen (z. B. durch einen Mausklick).
Ein weiteres Prinzip ist das *Semantic Zooming*. Hierbei handelt es sich um eine hilfreiche Strukturierung zoombarer Interfaces, wie sie z. B. von Google Maps eingesetzt werden: Mit jeder Zoomstufe, die man weiter hineinzoomt, werden mehr Details auf der Karte angezeigt. Auf diese Weise lässt sich vermeiden, dass sich eine Vielzahl von visuellen Elementen oder Beschriftungen überlagert, die auf einer bestimmten Zoomstufe noch gar nicht relevant ist.
Bei mehreren Visualisierungen, die durch die Daten miteinander in Verbindung stehen, bietet sich schließlich *Brushing+Linking* als Prinzip an, um die Verbindung auch visuell deutlich zu machen. Mit *Brushing* wird die Auswahl eines Teilbereichs der Daten einer Visualisierung bezeichnet, mit *Linking* die Hervorhebung dieser Daten in anderen verknüpften Visualisierungen, die den gleichen Datensatz mit einer anderen Perspektive repräsentieren (z. B. andere Datendimensionen oder ein gänzlich anderer Visualisierungstyp). Eine Vielzahl von Websites und Infografiken geben grundlegende Hilfestellung zur Auswahl geeigneter Visualisierungen. Durch Eingabe von Informationen zu Ihrem Datensatz (oder das Verfolgen von Pfaden in einer Baumstruktur) schränken Sie die Zahl in Frage kommender Visualisierungen schrittweise ein.

Externe und weiterführende Links

- 3DH-Projekt: https://web.archive.org/save/https://web.archive.org/save/http://threedh.net/ (Letzter Zugriff: 26.06.2024)

Generische Visualisierungstools

- Voyant: https://web.archive.org/save/https://voyant-tools.org/ (vgl. Voyant (Flüh 2024)) (Letzter Zugriff: 26.06.2024)
- CATMA: https://web.archive.org/http://catma.de/ (vgl. CATMA (Schumacher 2024a)) (Letzter Zugriff: 26.06.2024)
- Prism: https://web.archive.org/save/https://www.graphpad.com/features (Letzter Zugriff: 26.06.2024)

- Data Wrapper: https://web.archive.org/save/https://www.datawrapper.de/ (Letzter Zugriff: 26.06.2024)
- Plot.ly: https://web.archive.org/https://plotly.com (Letzter Zugriff: 26.06.2024)
- RAW: https://web.archive.org/save/http://rawgraphs.io/ (Letzter Zugriff: 26.06.2024)

R: The R project for Statistical Computing

- R-Webseite: https://web.archive.org/save/https://www.r-project.org (Letzter Zugriff: 26.06.2024)
- Tutorials und Einführungen:
  - A gentle Introduction to Text Mining Using R: https://web.archive.org/save/https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/ (Letzter Zugriff: 26.06.2024)
  - Introduction to Text Analysis and Topic Modeling with R: http://www.matthewjockers.net/materials/dh-2014-introduction-to-text-analysis-and-topic-modeling-with-r/ (Letzter Zugriff: 26.06.2024)

Python

- Python-Webseite: https://web.archive.org/save/https://www.python.org/ (Letzter Zugriff: 26.06.2024)
- Natural Language Toolkit: https://web.archive.org/save/http://www.nltk.org/ (Letzter Zugriff: 26.06.2024)
- Tutorials und Einführungen:
  - Python Programming for the Humanities: https://web.archive.org/save/http://www.karsdorp.io/python*course/ (Letzter Zugriff: 26.06.2024)
  - Natural Language Processing with Python (Buch): https://web.archive.org/save/http://www.nltk.org/book/ (Letzter Zugriff: 26.06.2024)
  - Text Analysis with Topic Models: https://web.archive.org/save/https://de.dariah.eu/text-analysis-with-topic-models (Letzter Zugriff: 26.06.2024)

D3

- D3-Webseite: https://web.archive.org/save/https://d3js.org (Letzter Zugriff: 26.06.2024)
- Tutorials und Einführungen:
  - Überblick und kurze Einführung D3: https://web.archive.org/save/https://d3js.org/#introduction (Letzter Zugriff: 26.06.2024)
  - Sammlung von Tutorials: https://web.archive.org/save/https://github.com/d3/d3/wiki/Tutorials (Letzter Zugriff: 26.06.2024)

Textvisualisierung und Visualisierung im Allgemeinen

Die im Folgenden aufgeführten Webseiten und Grafiken leisten Hilfestellung bei der Wahl geeigneter Visualisierungen für unterschiedliche Datensätze

- Principles of Information Visualization: https://web.archive.org/save/http://www.themacroscope.org/?page_id=469 (Letzter Zugriff: 26.06.2024)
- Data and Design Handbook: https://web.archive.org/save/https://trinachi.github.io/data-design-builds/copyright-page01.html (Letzter Zugriff: 26.06.2024)
- Data Viz Project: https://web.archive.org/save/https://datavizproject.com/ (Letzter Zugriff: 26.06.2024)

Bibliographie

Adler, Jeremy und Ulrich Ernst. 1987. *Text als Figur. Visuelle Poesie von der Antike bis zur Moderne*. Weinheim: Acta humaniora, VCH.

Bertin, Jacques. 1974. *Graphische Semiologie: Diagramme, Netze, Karten*. Berlin: de Gruyter.

Card, Stuart K., Jock Mackinlay und Ben Shneiderman. 1999. *Readings in information visualization: Using vision to think*. San Francisco: Kaufmann.

Drucker, Johanna. 2011. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 005, Nr. 1 (10. März).

Eder, Jens. 2007. Figurenkonstellation. In: *Metzler Lexikon Literatur. Begriffe und Definitionen*, hg. von Dieter Burdorf, Christoph Fasbender, und Burkhard Moenninghoff, 239. Stuttgart, Weimar: Metzler.

Flüh, Marie. 2024. Toolbeitrag: Voyant. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3775, https://fortext.net/tools/tools/voyant.

Goethe, Johann W. 1987. *Goethes Werke (1887-1919). Hrsg im Auftrage der Großherzogin Sophie von Sachsen. Abteilung II. Weimarer Ausgabe*. 2. Aufl. Bd. 5. Weimar: DTV.

Heusler, Andreas. 1955. *Nibelungensage und Nibelungenlied. Die Stoffgeschichte des deutschen Heldenepos*. Dortmund: Ruhfus.

Horstmann, Jan. 2024a. Methodenbeitrag: Stilometrie. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 1. Stilometrie (26. Februar). doi: 10.48694/fortext.3769, https://fortext.net/routinen/methoden/stilometrie.

———. 2024b. Methodenbeitrag: Topic Modeling. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3717, https://fortext.net/routinen/methoden/topic-modeling.

Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema und Gerik Scheuermann. 2015. On close and distant reading in digital humanities: A survey and future challenges. In: *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association.

Kammer, Stephan. 2014. Visualität und Materialität der Literatur. In: *Handbuch Literatur & Visuelle Kultur*, hg. von Claudia Benthien und Brigitte Weingart, 31–47. Berlin, Boston: de Gruyter.

Koller, Erwin. 2000. Unbildliche, bildliche und bildhafte Versprachlichung von Schmerz (bei A. Döblin, R. Musil, Th. Mann und M. Walser). In: *Bild im Text - Text und Bild*, hg. von Ulla Fix und Hans Wellmann, 129–153. Heidelberg: Winter.

Mackinlay, Jock. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5, Nr. 2: 110–141.

Mellmann, Katja. 2007. Versanalyse. In: *Handbuch Literaturwissenschaft*, hg. von Thomas Anz, 2: Methoden und Theorien:81–97. Stuttgart, Weimar: Metzler.

Piatti, Barbara. 2008. *Die Geografie der Literatur. Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.

Ryan, Marie-Laure. 2003. Cognitive Maps and the Construction of Narrative Space. In: *Narrative Theory and the Cognitive Sciences*, hg. von David Herman, 214–242. Stanford: CSLI Publications.

Schmid, Wolf. 2007. Textadressat. In: *Handbuch Literaturwissenschaft*, hg. von Thomas Anz, 1: Gegenstände und Grundbegriffe:171–181. Weimar: Metzler.

Schumacher, Mareike. 2024a. Toolbeitrag: CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3761, https://fortext.net/tools/tools/catma.

———. 2024b. Methodenbeitrag: Netzwerkanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3759, https://fortext.net/routinen/methoden/netzwerkanalyse.

Seifert, Christin, Vedran Sabol, Wolfgang Kienreich, Elisabeth Lex und Michael Granitzer. 2014. Visual analysis and knowledge discovery for text. In: *Large-Scale Data Analystics*, 189–218. New York: Springer.

Shneiderman, Ben. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In: *In IEEE Symposium on Visual Languages*, 336–343.

Ware, Colin. 2012. *Information Visualization: Perception for Design*. 3. Aufl. Waltham: Elsevier.

Wertheimer, Max. 1923. Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung* 4, Nr. 1: 301–350.

Willcock, Malcolm M. 1976. *A Companion to the Iliad: Based on the Translation by Richmond Lattimore*. Chicago, London: The University of Chicago Press.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch

auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

N-Gramm Unter N-Gramm versteht man in der Linguistik eine Sequenz von *N* aufeinanderfolgenden Fragmenten/Einheiten in einem Text. So gibt es beispielsweise Bigramme, Trigramme etc. Diese Fragmente können Buchstaben oder Phoneme sein. Der Satz „Marie erforscht Literatur digital“ kann zum Beispiel folgendermaßen in Bigramme, drei wortbasierte N-gramme mit je zwei Wörtern, aufgeteilt werden: „Marie erforscht“, „erforscht Literatur“ und „Literatur digital“.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Scalable Reading Die Kombination aus Distant Reading- und Close Reading-Methoden, angewandt auf einen Untersuchungsgegenstand, wird als Scalable Reading bezeichnet.

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf

Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.