20 Quantitative Analyse

20.1 | Was ist quantitative Analyse?

Quantitative Methoden der Datenanalyse werden in den digitalen Geisteswissenschaften in den unterschiedlichsten Bereichen und mit vielfältigen Motivationen eingesetzt. Mit der zunehmenden Verfügbarkeit größerer digitaler Bestände an geisteswissenschaftlich relevanten Daten wird der Einsatz quantitativer Methoden zugleich möglich und erforderlich. Solche Methoden werden unter programmatischen Schlagworten wie »distant reading« (Moretti 2005), »macroanalysis« (Jockers 2011) oder »cultural analytics« (Piper 2016) auch über die digitalen Geisteswissenschaften hinaus diskutiert. Beispiele für quantitative Methoden sind statistische Analysen und das sogenannte maschinelle Lernen sowie die Netzwerkanalyse (s. Kap. 10) und das Topic Modeling (vgl. Blei 2012).

Quantitative Analysemethoden grenzen sich von qualitativen Analysemethoden ab, die Bestandteile und Eigenschaften von Forschungsgegenständen beschreiben und dabei besondere Aufmerksamkeit auf nuancierte Differenzierungen, individualisierende Detailanalysen und herausragende oder beispielhafte Einzelbeispiele legen. Quantitative Analysemethoden hingegen sind in erster Linie darauf ausgerichtet, Merkmale von Forschungsgegenständen zu identifizieren und ihre Häufigkeiten zu erheben, was möglichst klare und teils auch vereinfachende Kategorisierungen erfordert. Auf Grundlage solcher zahlenmäßiger Informationen können dann mit statistischen Verfahren die Forschungsgegenstände beschrieben, Gruppen ähnlicher Gegenstände entdeckt oder automatische Klassifikationen vorgenommen werden. Zwischen qualitativen und quantitativen Methoden gibt es zahlreiche Brücken: So entstehen im Rahmen qualitativer Analysen häufig Hypothesen, die in einer Weise formalisiert und operationalisiert werden können, die sie einer Bearbeitung mit quantitativen Methoden zugänglich macht. Oder die Ergebnisse quantitativer Untersuchungen führen umgekehrt zu neuen Fragestellungen, denen mit qualitativen Methoden vertiefend nachgegangen werden kann (vgl. Jannidis 2010).

Quantitative Analysemethoden grenzen sich zudem vom **Information Retrieval** ab (s. Kap. 19). Letzteres ist damit befasst, wie die in einer Datensammlung explizit vorhandenen Informationen möglichst effizient und präzise abgerufen und die Fundstellen oder relevanten Dokumente in sinnvoller Weise gewichtet und sortiert werden können. Die quantitative Analyse hingegen zielt unter anderem darauf ab, relevante Bestandteile eines Datensatzes oder Strukturen in den Daten überhaupt erst zu identifizieren sowie weiterführende statistische Auswertungen von Datensammlungen durchzuführen.

Ziel dieses Kapitels ist es, einen Überblick über quantitative Methoden der Datenanalyse in den digitalen Geisteswissenschaften zu geben. Hierzu werden zunächst einige Grundlagen der deskriptiven Statistik vermittelt und ihre Nutzung für vergleichende Analysen erläutert. Dann werden die beiden grundlegenden Typen des maschinellen Lernens, das überwachte und unüberwachte Lernen, beschrieben und anhand konkreter Anwendungsbeispiele illustriert.

F. Jannidis et al. (Hrsg.), *Digital Humanities*,
DOI 10.1007/978-3-476-05446-3_20, © Springer-Verlag GmbH Deutschland, 2017

20.2 | Statistische Grundlagen

Merkmalserhebung: Viele quantitative Methoden der Datenanalyse beruhen darauf, dass zunächst für jeden einzelnen Untersuchungsgegenstand verschiedene **Merkmale** erhoben werden, was heute oft automatisch geschieht. Beispielsweise könnte man in einer Sammlung von Gedichten die Häufigkeit jedes Wortes in jedem Gedicht erheben, in einer Briefsammlung die Anzahl verschiedener Orts- und Personennamen pro Brief, oder in einer Gemäldesammlung, welche Objekte in jedem Bild dargestellt sind. Solche Informationen werden üblicherweise in einer Tabelle festgehalten, in der beispielsweise jede Spalte ein Werk (Gedicht, Brief bzw. Gemälde) und jede Zeile ein bestimmtes Merkmal (Wort, Personenname bzw. Objekt) betrifft. Jede Zelle repräsentiert dann den Kreuzungspunkt eines Werks und eines Merkmals und enthält eine zahlenmäßige Charakterisierung des Merkmals in dem Werk. Man nennt eine solche Tabelle **Merkmals-Matrix**, im besonderen Fall der Textanalyse auch Term-Dokument-Matrix. Eine solche Matrix bedeutet einerseits einen Informationsverlust: Beispielsweise werden bei Texten die Reihenfolge der Wörter oder bei Gemälden die Position der Objekte im Bild nicht berücksichtigt. Im Gegenzug erleichtert die Tabellenform aber zahlreiche nützliche Berechnungen: Beispielsweise beantwortet die spaltenweise bzw. zeilenweise Summenbildung die Frage, wie viele Wörter ein bestimmter Text insgesamt enthält bzw. wie oft ein bestimmtes Objekt in der Bildsammlung insgesamt dargestellt wird.

Was ist mit den Merkmalen eines Objekts gemeint? Zur Beantwortung dieser Frage sei hier die Unterscheidung zwischen **Types und Tokens** eingeführt. Im engeren (und auf Texte bezogenen) Sinne sind Tokens die als solche identifizierten Wortformen, wobei es unterschiedliche Auffassungen dessen gibt, wie die Grenzen zwischen Wortformen bestimmt werden und wie Interpunktion und Groß-/Kleinschreibung behandelt werden. Types hingegen sind die unterschiedlichen Tokens. Der Satz »A rose is a rose is a rose.« hat neun Tokens, aber nur fünf Types, wenn man Interpunktion und Groß-/Kleinschreibung berücksichtigt. In einem erweiterten Sinne (der nicht nur für Texte geeignet ist) kann man Tokens als die grundlegende Analyseeinheit verstehen und Types als das jeweils eine solche Einheit beschreibende Kriterium. Noch einmal auf Einzelwörter bezogen, wäre eine einzelne Wortform die Analyseeinheit (Token) und die jeweilige Schreibung das definitorische Kriterium (alle Wortformen gleicher Schreibung gelten als ein Type). Ebenso könnte aber jeder Satz eine Analyseeinheit (Token) sein und das ihn beschreibende Kriterium (Type) beispielsweise, ob es sich um einen einfachen oder komplexen Satz handelt. Oder jedes in einem Bild dargestellte Objekt wäre die Analyseeinheit (Token) und das beschreibende Kriterium (Type) wäre dadurch bestimmt, welches Objekt dargestellt ist. Zu den relevanten Merkmalen eines Untersuchungsgegenstands gehören dann die Häufigkeiten jedes Types, also wie viele der jedem Type entsprechenden Tokens beobachtet werden können. Daraus lassen sich weitere Merkmale ableiten, wie die Gesamtzahl der vorhandenen Types oder Tokens, die schlichte An- oder Abwesenheit eines oder mehrerer Types, oder auch das Verhältnis bestimmter Types zueinander.

Beispiel Romane: Zur Veranschaulichung wird im Folgenden eine Sammlung von zehn Romanen des britischen Autors Arthur Conan Doyle verwendet, von denen vier Kriminalromane sind (mit Sherlock Holmes als Protagonisten), sechs dagegen verschiedenen anderen Romangattungen angehören. Eine Merkmals-Matrix der Doyle-Sammlung erhält man, indem man die Liste aller Types (unterschiedlicher Wortfor-

men) erstellt und dann die Häufigkeit jedes Types in jedem der zehn Romane ermittelt (einen Ausschnitt aus der Matrix zeigt Tabelle 10):

Rang	Type	The Hound of the Baskervilles	The Sign of Four	A Study in Scarlet	The Valley of Fear	Lost World	The Mystery of Cloomber	Poison Belt	Raffles Haw	The Refugees	The White Company	Summe
1	the	3056	2147	2319	1994	4156	2632	1618	1842	7264	9124	37152
2	and	1505	1154	1313	1343	2000	1442	732	1042	3899	5226	19656
3	of	1580	1105	1195	1424	2502	1362	864	1018	3362	4781	19121
4	to	1381	1070	1070	1256	1678	1218	646	936	2606	3131	14992
...	...											...
21	which	417	235	315	307	639	341	203	280	751	844	4332
22	my	420	323	274	221	542	466	175	213	541	918	4093
23	at	335	311	289	341	430	288	172	196	813	821	3996
24	be	328	260	248	302	413	310	189	262	650	763	3725
...	...											...
101	know	115	69	49	104	74	55	32	73	146	142	859
102	eyes	69	36	59	54	77	34	26	49	219	225	848
103	like	51	58	33	59	127	56	47	47	164	204	846
...	...											...
1001	pocket	7	8	14	14	4	6	1	3	7	7	71
1002	extraordinary	8	5	7	5	16	7	3	8	2	8	69
1003	long	9	4	1	10	6	6	0	2	16	9	63
...	...											...
Summe		69802	51583	51218	69676	88221	56534	34544	44910	146303	178233	791024

Tab. 10 Ausschnitt aus der Merkmals-Matrix für die Doyle-Sammlung. Die Matrix enthält die absoluten Häufigkeiten der Types, absteigend sortiert. Es sind die Häufigkeiten für alle zehn Romane, aber nur für Ausschnitte aus verschiedenen Bereichen der Matrix enthalten.

Man kann dieser Matrix bereits einige Informationen entnehmen. Die absolute Häufigkeit eines Types in einem Roman steht jeweils in der entsprechenden Zelle (z. B. kommt das Type ›eyes‹ in *Poison Belt* 26 Mal vor). Die Häufigkeit eines Types in allen Texten ergibt sich aus der Zeilensumme (›eyes‹ kommt insgesamt 848 Mal vor). Unterschiedliche Häufigkeitsbereiche entsprechen unterschiedlichen Wortarten: Die Types der ersten etwa 100 Ränge sind überwiegend Funktionswörter, diejenigen höherer Ränge sind dagegen überwiegend Inhaltswörter. Je höher der Rang eines Types, desto spezifischere Begriffe erscheinen, die dann oft auch gar nicht in allen Texten vorkommen (z. B. kommt ›long‹ in *Poison Belt* nicht vor). Der Merkmals-Matrix ist außerdem die Länge der Texte zu entnehmen; sie entspricht der Spaltensumme (*Lost World* hat beispielsweise eine Länge von 88.221 Tokens). Die Summe

aller Zellen entspricht der Anzahl der Tokens in der Textsammlung (hier: 791.024). Die Länge der Tabelle schließlich zeigt an, wie viele Types vorkommen.

Grundlegende Operationen auf der Merkmals-Matrix

Absolute und relative Häufigkeit: Die absoluten Häufigkeiten haben ihre Berechtigung, allerdings erlauben sie keinen sinnvollen Vergleich der Häufigkeiten zwischen verschiedenen Werken, wenn der Umfang der Werke unterschiedlich ist. **Relative Häufigkeiten** erlauben einen solchen Vergleich und liegen vor, wenn die absoluten Häufigkeiten der Types durch die spaltenweisen Summen (Anzahl der Tokens pro Text, also Textlänge) geteilt werden. Die relative Häufigkeit eines Types in einem Roman beschreibt dann, wie häufig ein bestimmtes Type im Verhältnis zur Länge des Romans anzutreffen ist. Beispielsweise kommt das Type ›know‹ in *Lost World* 74 und in *Raffles Haw* 73 Mal vor, die absolute Häufigkeit ist also fast identisch. Allerdings hat *Lost World* eine Länge von 76.975, *Raffles Haw* dagegen nur von 38.571 Tokens, also etwa die Hälfte. Die relative Häufigkeit und damit der Anteil am Text von ›know‹ ist daher für *Lost World* 74/76975 = 0,0009, für *Raffles Haw* aber 73/38571 = 0,0019 und damit gut doppelt so hoch. Oft multipliziert man alle relativen Häufigkeiten noch um den Faktor 1000, um Häufigkeiten pro 1000 Tokens zu erhalten, was intuitiv besser erfassbar ist. Das Wort ›know‹ kommt in *Lost World* also 0,9 Mal pro 1000 Tokens vor (wir rechnen im Folgenden mit solchen Zahlen).

Die relativen Häufigkeiten erlauben zwar den Vergleich zwischen Texten, häufig möchte man aber genauer wissen, ob ein Type in einem bestimmten Einzeltext einer größeren Sammlung vergleichsweise häufig oder selten vorkommt. Hierfür kann es zweckmäßig sein, die (relative) Häufigkeit des Types in dem Einzeltext mit dem Mittelwert oder Median des Types in der Textsammlung zu vergleichen. Sowohl Mittelwert als auch Median sind sogenannte **Maße der zentralen Tendenz**, die eine Verteilung mehrerer Werte in einem Wert zusammenfassen. Den (arithmetischen) Mittelwert der relativen Häufigkeit eines Types in der Romansammlung berechnet man, indem man die Zeilensumme des Types durch die Anzahl der Romane teilt. Für ›know‹ ergibt sich eine mittlere Häufigkeit von 1,4 (pro 1000 Tokens). Der Mittelwert ist sensibel gegenüber wenigen extrem hohen oder extrem niedrigen Werten und verschiebt sich dann spürbar in deren Richtung. Beim Median ist dies nicht der Fall. Man berechnet den Median, indem man die Häufigkeiten eines Types nach ihrem Wert sortiert und den Wert auswählt, der in der Mitte der geordneten Liste steht. (Hat die Liste eine gerade Anzahl von Einträgen, wird der Mittelwert zwischen den beiden um die Mitte liegenden Werten gebildet.) Für ›know‹ liegt der Median bei 1,2, also etwa 15 % niedriger als der Mittelwert, der von vergleichsweise hohen Werten in *The Hound of the Baskervilles* (1,94) und *Raffles Haw* (1,92) beeinflusst ist. Wenn solche extremen Werte vorliegen, fasst der Median die Verteilung der Type-Häufigkeiten besser zusammen als der Mittelwert. Vergleicht man nun die relative Häufigkeit von ›know‹ in *Lost World* (0,9) mit dem Median des Worts in der Textsammlung (1,2) zeigt sich, dass ›know‹ in *Lost World* bezogen auf die Gesamtsammlung eher selten ist.

Streuungsmaße: Mittelwert und Median beschreiben wesentliche Eigenschaften einer Häufigkeitsverteilung, lassen aber eine weitere wichtige Eigenschaft unberücksichtigt, nämlich wie stark die Häufigkeiten in verschiedenen Texten schwanken. Verschiedene Streuungsmaße beschreiben, wie stark die Häufigkeiten eines Types in

verschiedenen Texten um den Mittelwert (den durchschnittlich zu erwartenden Wert) herum variieren. Zu diesen Maßen gehören insbesondere die Spannweite und die Standardabweichung. Die Spannweite ist schlicht der Abstand zwischen dem höchsten und dem niedrigsten Wert (für ›of‹ im Doyle-Korpus liegt die Spannweite demnach bei 3,3 – 2,4 = 0,9). Die **Standardabweichung** eines Types beschreibt die Streuung genauer und berechnet sich als die Wurzel aus dem Mittelwert der quadrierten Abweichungen jedes Einzelwertes vom Mittelwert aller Werte.

Eine hohe Standardabweichung weist darauf hin, dass die Werte breit um den Mittelwert streuen, während eine niedrige Standardabweichung darauf hinweist, dass die Werte eng um den Mittelwert herum liegen. Die verschiedenen Einzelwerte (hier: die relativen Häufigkeiten pro 1000 Tokens) für ein Type sind auf Grundlage der Matrix zu berechnen (für ›at‹ beträgt der Wert in *Poison Belt* beispielsweise 172/34544 × 1000 = 5,0). Der Mittelwert von ›at‹ in der Doyle-Sammlung beträgt 6,0. Die quadrierte Abweichung eines Einzelwerts vom Mittelwert erhält man, indem man den Mittelwert vom Einzelwert abzieht und diesen Wert quadriert (für ›at‹ in *Poison Belt* also: $(5 - 6)^2 = 1$).

Nun summiert man alle Abweichungen auf, bestimmt ihren Mittelwert und zieht aus dem Ergebnis die Wurzel. Für ›at‹ erhält man so eine Standardabweichung von 0,49. Nimmt man die gleiche Berechnung für ›my‹ vor (das einen ganz ähnlichen Mittelwert wie ›at‹ hat und im Rang direkt davorliegt), ergibt sich eine deutlich höhere Standardabweichung von 1,35. Dies leuchtet auch ein, insofern erwartbar ist, dass das Reflexivpronomen der ersten Person stärker in Abhängigkeit der unterschiedlichen Erzählperspektiven und Romangattungen in der Doyle-Sammlung schwankt, als dies für die Präposition ›at‹ der Fall ist. (Möchte man die Standardabweichungen von Verteilungen vergleichen, die unterschiedliche Mittelwerte haben, empfiehlt es sich, den sog. Variationskoeffizient zu berechnen, indem man die Standardabweichungen durch den Mittelwert der Verteilungen teilt.)

Normalisierung und Standardisierung von Verteilungen

Mittelwert bzw. Median und Standardabweichung charakterisieren zwei wesentliche Aspekte einer Verteilung. Auf ihrer Grundlage kann man zudem eine Normalisierung einer Verteilung vornehmen. Dies dient u. a. dazu, die Verteilungen mehrerer Types besser untereinander vergleichbar zu machen.

Abbildung 58 (oben) zeigt die Verteilungen mehrerer Types aus unterschiedlichen Bereichen der Wortliste als relative Häufigkeiten ohne weitere Normalisierung (die gewählten Types entsprechen denjenigen aus der Merkmals-Matrix in Tabelle 10). Jeder Boxplot in der Abbildung repräsentiert die Verteilung eines Types in den verschiedenen Texten. Der Median der Verteilung ist als dicke horizontale Linie markiert (für ›of‹ liegt er beispielsweise bei 0,028). Die gefüllte Fläche repräsentiert 50 % aller Werte (5 der 10 relativen Häufigkeiten für ›of‹ liegen zwischen 0,026 und 0,030). Diese und die übrigen Werte liegen innerhalb des Bereichs, der von den kurzen horizontalen Linien begrenzt wird (für ›of‹ zwischen 0,024 und 0,033) und zeigen demnach die Spannweite der Verteilung. Gut erkennbar ist der starke Abfall der Häufigkeiten mit zunehmendem Rang, d. h. die Verteilungen haben sehr unterschiedliche Mediane und Mittelwerte, was ihre Vergleichbarkeit erschwert. Die absolute Streuung der Werte nimmt ebenfalls ab, je höher der Rang des Types ist.

Eine einfache **Mittelwert-Normalisierung** erleichtert den Vergleich von Werten

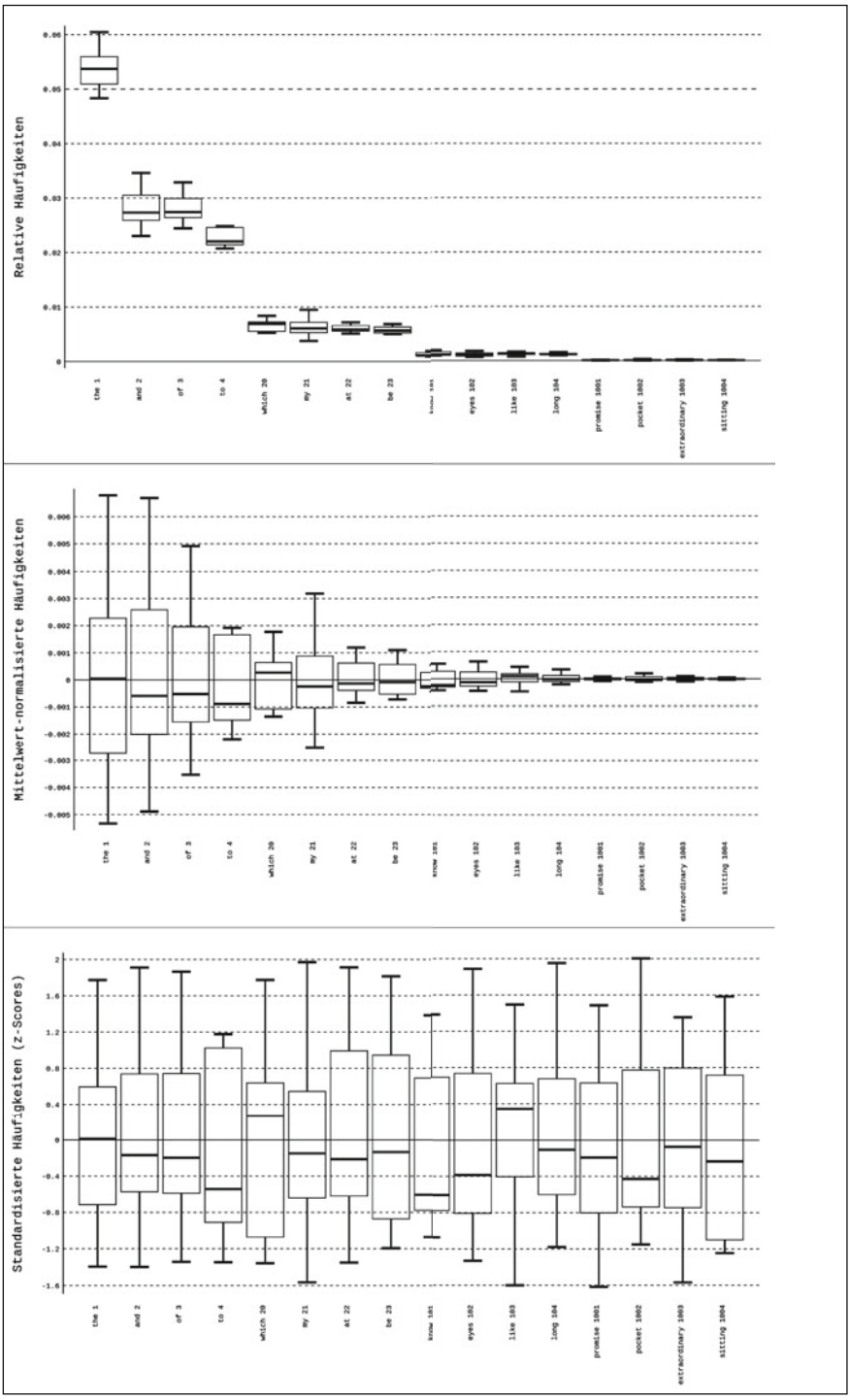

Abb. 58 Die Verteilungen von 16 Types nach relativen Häufigkeiten (oben), mit Mittelwert-Normalisierung (mittig) und als z-scores (unten)

eines Types in verschiedenen Dokumenten (s. Abbildung 58, mittig). Die Formel lautet $x' = x - \mu$, das heißt man zieht von jeder Häufigkeit (x) eines Types in einem Einzeltext den Mittelwert des Types (μ – sprich ›mü‹) in der Textsammlung ab. Die Werte eines Types in einem Einzeltext sind dann positiv oder negativ, je nachdem ob das Type in einem Roman überdurchschnittlich oder unterdurchschnittlich häufig ist. Beispielsweise liegt der Mittelwert für das Type ›of‹ bei 0,028. Zieht man diesen Mittelwert jeweils von den einzelnen Häufigkeiten ab, erhält man die normalisierten Häufigkeiten. Für *The Valley of Fear* ergibt sich dann beispielsweise 0,024 – 0,028 = –0,004, womit ›of‹ in diesem Roman seltener ist als im Mittel der Sammlung. Die so normalisierten Verteilungen haben identische Mittelwerte von Null. Dies erleichtert den Vergleich der Werte, ihre Streuung bleibt aber sehr unterschiedlich.

Um auch die Streuung vergleichbar zu machen, ist es darüber hinaus möglich, eine weitergehende Standardisierung der Verteilungen vorzunehmen und die Werte zu sogenannten **z-scores** zu transformieren (s. Abbildung 58, unten). Die Formel lautet $z = (x - \mu) / \sigma$. Hier wird von jedem Wert (x) der Mittelwert des Types (μ) in der Sammlung abgezogen und das Ergebnis durch die Standardabweichung (σ – sprich ›sigma‹) des Types geteilt. Der z-score beschreibt, wie weit ein Einzelwert vom Mittelwert der Verteilung entfernt ist, ausgedrückt in Standardabweichungen. Jede Verteilung hat nun nicht nur einen identischen Mittelwert von 0, sondern auch eine identische Standardabweichung von 1. Die unterschiedlich starke Richtung der Streuung hängt nun nicht mehr von Rang, Häufigkeit und Standardabweichung des Types ab, sondern bildet einen besser interpretierbaren Hinweis auf das Verhalten des Types in der Textsammlung. Eine solche sogenannte Merkmalsskalierung bietet sich nicht nur bei Worthäufigkeiten an, sondern generell immer dann, wenn Merkmale mit stark unterschiedlichen Häufigkeiten gemeinsam in eine Berechnung eingehen sollen.

Die bis hierher behandelten Kenntnisse der deskriptiven Statistik bilden die Grundlage für die weitere Darstellung quantitativer Methoden in den Geisteswissenschaften.

Kontrastierende Analysen

Für eine große Anzahl an Fragestellungen kann es interessant sein, die gesamte **Datensammlung** in zwei Partitionen genannte Teilmengen aufzuteilen und diese Partitionen kontrastiv zu analysieren, also ihre Unterschiede herauszuarbeiten. Man geht häufig explorativ vor und ermittelt, welche Merkmale für eine Partition (die Zielpartition) gegenüber der anderen Partition (der Vergleichspartition) besonders deutlich überrepräsentiert sind und damit als typisch, charakteristisch oder distinktiv gelten können. Derartige kontrastierende Analysen können nicht nur für Texte, sondern für vielfältige Datentypen durchgeführt werden. Die grundlegende Annahme einer solchen Analyse ist, dass ein Merkmal nicht nur durch seine reine Häufigkeit in der Zielpartition für diese charakteristisch ist, sondern dass dies auch davon abhängt, ob das Merkmal in der Vergleichspartition deutlich weniger häufig ist.

Ein Funktionswort wie ›of‹ mag in Doyles Kriminalromanen sehr häufig vorkommen, da es aber in allen anderen vertretenen Romangattungen ebenfalls sehr häufig ist, kann es nicht als charakteristisch für den Kriminalroman gelten. Ein Wort wie ›murder‹ hingegen, das in Kriminalromanen sehr häufig, in allen anderen Romangattungen aber eher selten ist, kann viel eher als charakteristisch für den Kriminalro-

man gelten. Um diese Intuition algorithmisch abzubilden, werden Testverfahren so konstruiert, dass sie die Häufigkeit eines Types in der Zielpartition in Abhängigkeit seiner Häufigkeit in der Vergleichspartition (oder der gesamten Datensammlung) gewichten. So bekommen diejenigen Types in der untersuchten Teilmenge einen besonders hohen Wert zugewiesen, die in der untersuchten Teilmenge sehr häufig sind und zugleich in den übrigen Teilmengen bzw. der gesamten Sammlung sehr selten sind.

Kontrastmaße: Das denkbar einfachste Vorgehen, das eine solche Gewichtung beinhaltet, betrachtet das **Verhältnis der relativen Häufigkeiten** der Types in zwei Partitionen (vgl. Gries 2010). Für jedes Type ermittelt man die relative Häufigkeit in jeder der beiden Partitionen und dividiert den Wert in der Zielpartition durch den Wert in der Vergleichspartition. Sortiert man anschließend die resultierenden Verhältnisse absteigend, enthält der Anfang der Liste die für die Zielpartition am deutlichsten überrepräsentierten oder präferierten Types, das Ende der Liste dagegen die am deutlichsten unterrepräsentierten oder vermiedenen Types. In der Mitte der Liste finden sich diejenigen Types, die in beiden Partitionen etwa gleich häufig sind, also weder vermieden noch präferiert werden. Prüfen wir mit diesem Instrument einmal die oben formulierte Hypothese zum Verhalten von ›of‹ und ›murder‹. Zunächst erheben wir die mittleren relativen Häufigkeiten pro 1000 Tokens von ›of‹ in den vier Kriminalromanen und in den übrigen sechs Romanen. Sie lauten 26,0 für die Krimis und 29,3 für die übrigen Romane. Das Verhältnis der relativen Häufigkeiten für ›of‹ ist also 26,0/29,3 = 0,88. Die gleiche Berechnung für ›murder‹ ergibt hingegen: 0,34/0,018 = 18,7. Obwohl ›murder‹ eine viel geringere relative Häufigkeit in den Kriminalromanen hat als ›of‹, weist es ein vielfach höheres Verhältnis der relativen Häufigkeiten auf und ist damit deutlich charakteristischer für den Kriminalroman.

Dieses einfache Verfahren kann erste Hinweise geben, doch wird es in den meisten Fällen sinnvoll sein, etwas komplexere Berechnungen vorzunehmen (vgl. den ausgezeichneten Überblick bei Lijffijt et al. 2014). Das bisher dargestellte Verfahren betrachtet die Partitionen als Ganze und vergleicht lediglich zwei Mittelwerte. Dadurch wird die Variabilität der Werte innerhalb der Texte, die eine Partition bilden, nicht berücksichtigt. Je größer die Streuung der Werte aber ist, und je weniger symmetrisch sich die Werte um den Mittelwert verteilen, desto weniger gut repräsentieren die Mittelwerte die Verteilung und desto größer ist die Wahrscheinlichkeit, dass die Verteilungen stark überlappen, selbst wenn sich die Mittelwerte unterscheiden.

Zahlreiche Testverfahren berücksichtigen diese Überlegungen. Beim Welchs t-Test fließt die Standardabweichung und damit ein Maß für die Streuung der Häufigkeiten mit in den Vergleich zweier Partitionen ein (vgl. Oakes 1998, Kapitel 1.3). Das von John Burrows vorgeschlagene **Zeta-Maß** (Burrows 2007) berücksichtigt auf andere Weise die Streuung der Werte bzw. die konsistente Verwendung der Types in zwei Partitionen. Eine Variante des von Hugh Craig weiterentwickelten Zeta (Craig/Kinney 2009) wird hier genauer beschrieben und bringt zwei entscheidende Faktoren ins Spiel:

Erstens werden nicht ganze Partitionen verglichen, sondern einzelne Texte in den beiden Partitionen. Mehr noch, auch die einzelnen Texte werden noch einmal in gleich lange kleinere Segmente (z. B. von 2000 oder 5000 Tokens) aufgeteilt. Dies verfeinert die Analyse insbesondere dann entscheidend, wenn sehr lange Texte wie beispielsweise Romane verglichen werden. Zweitens werden nicht die Häufigkeiten

der Types in den Segmenten erhoben, sondern es wird stattdessen ermittelt, in wie vielen der Segmente ein Type mindestens einmal vorkommt. In den resultierenden Wert fließt damit vor allem ein, wie konsistent das Type in den beiden Partitionen verwendet wird. Für jedes Type wird erhoben, in wie vielen Segmenten der Zielpartition einerseits, der Vergleichspartition andererseits, es vorkommt. Diese Werte werden zu Anteilen umgerechnet, indem sie durch die Anzahl der Segmente der jeweiligen Partition geteilt werden. Und schließlich wird von dem Anteil, zu dem das Type in der Zielpartition vorkommt, der Anteil, zu dem das Type in der Vergleichspartition vorkommt, subtrahiert. Der resultierende Wert liegt zwischen –1 und 1 und drückt aus, wie charakteristisch ein Type für die Zielpartition ist. Der maximale Zeta-Wert von 1 für besonders charakteristische Types entsteht, wenn das Type in allen Segmenten der Zielpartition (Anteil = 1,0) vorkommt, also sehr konsistent verwendet wird, und zugleich in keinem Segment der Vergleichspartition (Anteil 0,0) vorkommt (1 – 0 = 1). Der minimale Zeta-Wert von –1 für extrem uncharakteristische Types entsteht, wenn das Type in keinem Segment der Zielpartition vorkommt, aber in allen Segmenten der Vergleichspartition vorkommt (0 – 1 = –1). Types, die in jedem Segment beider Partitionen mindestens einmal vorkommen, wie äußerst verbreitete Funktionswörter, bekommen einen neutralen Wert von 0 (1 – 1 = 0).

Anwendung: Um Zeta auf die Doyle-Sammlung anzuwenden, können wir die Implementierung des Stilometrie-Pakets ›stylo‹ für die Statistikumgebung R verwenden (vgl. Eder et al. 2016). Wir definieren die vier Kriminalromane als die Zielpartition, die übrigen sechs Romane als die Vergleichspartition. Die Berechnung von Craigs Zeta resultiert in der folgenden Liste der Types mit den höchsten bzw. niedrigsten Zeta-Scores (s. Abbildung 59). Ein Zeta-Wert über Null drückt aus, dass ein Type in den Kriminalromanen überrepräsentiert bzw. für diese besonders charakteristisch ist. Ein Zeta-Wert unter Null drückt aus, dass ein Type in den Kriminalromanen unterrepräsentiert ist bzw. für diese uncharakteristisch ist. Anders ausgedrückt: Es ist charakteristisch für die Kriminalromane, dass diese Wörter selten in ihnen vorkommen.

Es ist unmittelbar erkennbar, dass das Verfahren für diese Textsammlung funktioniert. Die Namen der Protagonisten der Kriminalromane (Sherlock Holmes und Watson) haben die höchsten Scores, weil sie in den Kriminalromanen sehr konsistent, in den übrigen Romanen hingegen überhaupt nicht vorkommen. Und auch viele der weiteren Begriffe sind offensichtlich dem Krimi-Genre zugehörig: ›crime‹, ›police‹, ›detective‹, ›case‹, ›murder‹. Die unterrepräsentierten Begriffe zeigen die Abgrenzung der Kriminalromane von den übrigen Romangattungen auf, wobei man einige der Types sicherlich mit den enthaltenen historischen Romanen assoziieren kann (›lord‹, ›france‹, ›king‹, aber auch ›nay‹, ›ere‹ oder ›hath‹ mit ihren historischen Schreibweisen).

Zum Abschluss dieses Abschnitts noch der Hinweis, dass die hier besprochenen Kennwerte für Verteilungen (wie Mittelwert, Median oder Standardabweichung) und vergleichenden Maße (wie das Verhältnis der relativen Häufigkeiten oder die Zeta-Berechnung) rein deskriptiver Natur sind. Über die Beschreibung von Daten hinaus geht die **Inferenzstatistik**, mit der spezifische Hypothesen über die Daten überprüft werden können. Ein wichtiger Anwendungsbereich der Inferenzstatistik ist die Frage, ob Beobachtungen, die anhand einer spezifischen Datensammlung gemacht wurden, auch auf den Untersuchungsgegenstand insgesamt verallgemeinert werden können. Beispielsweise könnte man festgestellt haben, dass in einer Sammlung von 100 Spielfilmen aus den 1960er Jahren die Western weniger Figuren haben als die

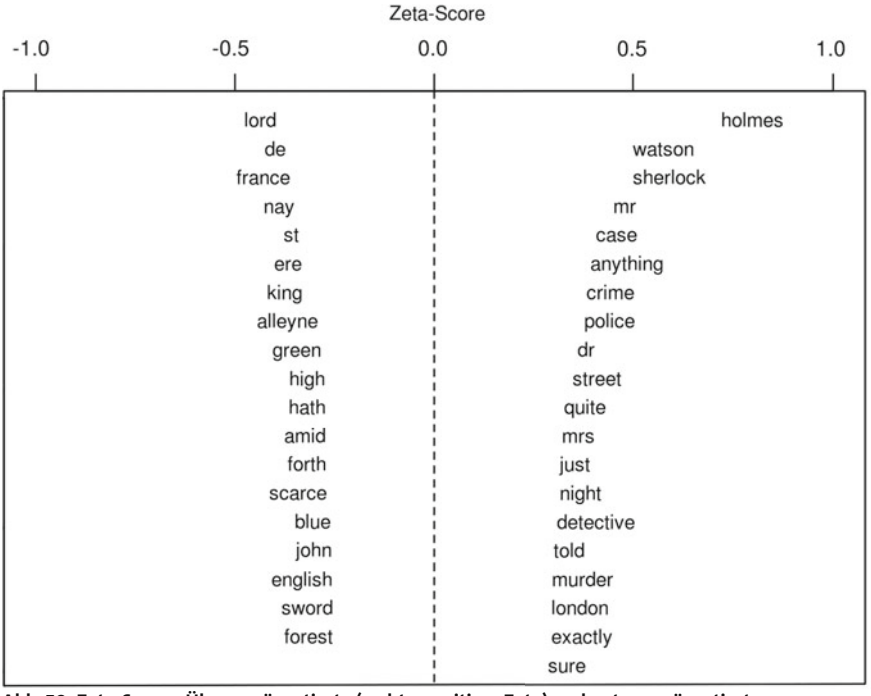

Abb. 59 Zeta-Scores. Überrepräsentierte (rechts, positives Zeta) und unterrepräsentierte (links, negatives Zeta) Types für die Kriminalromane im Vergleich zu den übrigen Romanen in der Doyle-Sammlung

Road Movies. Ist der Unterschied zwischen den Western und den Road Movies nur in der untersuchten Filmsammlung vorhanden, oder gilt dies tatsächlich für die Western und Road Movies der 1960er Jahre insgesamt?

Für die Beantwortung dieser Frage müssen eine Reihe von Faktoren berücksichtigt werden. So spielt es eine Rolle, wie deutlich der Unterschied in der Figurenanzahl ist (eine Frage der sog. Effektgröße), aber auch wie die Figurenanzahl in den Western und Road Movies jeweils verteilt ist (Form und Streuung der Verteilung). Entscheidend ist zudem das Verhältnis zwischen den analysierten 100 Filmen (der sog. Stichprobe) und allen Filmen der 1960er Jahre (der Grundgesamtheit): nur wenn die Stichprobe eine zufällige Auswahl aus allen relevanten Filmen ist – und damit ein sogenanntes ›repräsentatives Sample‹ darstellt (s. Kap. 16.1 zu »Datensammlungen«) –, kann zuverlässig von der Stichprobe auf die Grundgesamtheit geschlossen werden. Eine Reihe von Verfahren für das Testen solcher Hypothesen und zum Schließen von der Stichprobe auf die Grundgesamtheit sind wesentlicher Bestandteil der Inferenzstatistik, können aber im Rahmen dieser Einführung nicht behandelt werden (vgl. hierfür Bortz/Schuster 2010, Kapitel I.6).

20.3 | Maschinelles Lernen

Überwachtes und unüberwachtes Lernen: Die bisher behandelten Verfahren beruhen auf einfachen Algorithmen, d. h. ein Input wird in einer genau festgelegten Folge von Rechenoperationen in ein Output überführt. Allerdings gibt es nicht für alle denkbaren Fragestellungen solche Algorithmen, weil die Fragestellungen oft neu und/oder sehr spezifisch sind und die Zusammenhänge zwischen Input und Output nicht eindeutig formuliert werden können. **Maschinelles Lernen** beruht auf der Grundidee, dass Computer anhand gegebener Beispiele selbständig Zusammenhänge und Muster erkennen und insofern ›lernen‹ können. Dabei kann man mindestens zwei Typen des maschinellen Lernens unterscheiden: überwachtes und unüberwachtes Lernen (für anspruchsvolle Überblicksdarstellungen, vgl. Alpaydin 2010 sowie Han et al. 2012).

Beim **überwachten Lernen** ist es Aufgabe des Computers, aus einer Reihe von Beispielen einen Zusammenhang zwischen den Eigenschaften der Beispiele und den Klassen, denen die Beispiele zugehörig sind, zu erkennen. Es könnte beispielsweise darum gehen, Gemälde aufgrund ihrer Farbprofile einer Epoche zuzuordnen. In dieser Sichtweise zeichnen sich zu untersuchende Gegenstände einerseits durch eine Reihe von Merkmalen aus (hier: die Anteile verschiedener Farben), andererseits durch eine Klassenzugehörigkeit (hier: die Epochen), wobei die möglichen Klassen vorgegeben sind. Der Computer ›lernt‹ dann aus den Beispielen, für die Merkmale und Klassenzugehörigkeit bekannt sind, den Zusammenhang zwischen bestimmten Merkmalsausprägungen und -kombinationen und der Klassenzugehörigkeit.

Die Zusammenhänge zwischen Merkmalen und Klassen können auf vielfältige Weise modelliert sein, unter anderem als Regelsystem (»wenn die Merkmale A und B beide vorhanden sind, dann gehört das Objekt zu Klasse 2«) oder als Set von Wahrscheinlichkeiten (»je stärker Merkmal A ausgeprägt ist, desto höher ist die Wahrscheinlichkeit, dass das Objekt zu Klasse 2 gehört«). Der Computer ermittelt selbständig den **Algorithmus**, mit dem sich aus den Merkmalen als Input die Klassenzugehörigkeit als Output ermitteln lässt. Anschließend kann dieser eigens ermittelte Algorithmus zunächst mit weiteren bereits klassifizierten Daten getestet und dann auf neue Daten angewendet werden, für die die Klassenzugehörigkeit nicht bekannt ist. Diese werden auf dieser Grundlage ebenfalls klassifiziert, so dass bisher unbekannte Informationen vorliegen. Dieses dreischrittige Verfahren beruht auf der Trennung von Lern-, Evaluations- und Anwendungsphase und damit auch auf der Trennung der Trainings- und Evaluationsdaten (die vorab bereits zu den richtigen Klassen zugeordnet wurden) und den Anwendungsdaten (deren Merkmale zwar bekannt sind bzw. leicht erhoben werden können, deren Klassen aber automatisch ermittelt werden sollen).

Das **unüberwachte Lernen** unterscheidet sich in zwei wesentlichen Aspekten vom überwachten Lernen. Erstens gibt es keine Trennung zwischen den verschiedenen Phasen und damit auch keine Trennung zwischen verschiedenen Datensätzen. Vielmehr operiert das Verfahren auf einem einzigen Datensatz. Zweitens werden keine Klassen vorgegeben. Stattdessen ist es das Ziel des unüberwachten Lernens, Regelmäßigkeiten, Korrelationen und andere Zusammenhänge zwischen den Merkmalen für die Bildung von möglichst gut abgegrenzten Gruppen innerhalb der Daten zu nutzen. Statt einen Zusammenhang zwischen Merkmalen und Klassen zu lernen, ermittelt das unüberwachte Lernen also Gruppen von Untersuchungsgegenständen,

die sich untereinander möglichst ähnlich, als Gruppe aber von anderen Gruppen möglichst unterschiedlich sind.

Der Computer ›lernt‹, Gruppen ähnlicher Gegenstände (sogenannte **Cluster**) zu bilden und voneinander zu unterscheiden, ohne diese Cluster aber mit vorgegebenen Klassenlabels zu versehen. Zu entscheiden, worin sich die Kohärenz der Cluster inhaltlich begründet und damit das Ergebnis zu interpretieren, ist dann Aufgabe des Forschers.

Maschinelles Lernen kann dabei mit zwei Motivationen durchgeführt werden: Einerseits können die Datensätze umfangreicher Datensammlungen so **klassifiziert** (überwachtes Paradigma) oder **geclustert** (unüberwachtes Paradigma) werden, wodurch neue Informationen über die Datensätze und die Struktur der Datensammlung entstehen. Andererseits kann eine Analyse der für die Klassifikation oder das Clustering entscheidenden Merkmale oder Regeln auch für ein besseres Verständnis des Untersuchungsgegenstandes genutzt werden.

Maschinelles Lernen wird in vielen Bereichen eingesetzt, die für die digitalen Geisteswissenschaften relevant sind, unter anderem für die *Optical Character Recognition* bei der Volltextdigitalisierung (s. Kap. 12 zu »Digitalisierung«) oder im *Natural Language Processing*. Dieses Forschungsgebiet ist zwischen Informatik, Computerlinguistik und Künstlicher Intelligenz angesiedelt und befasst sich mit der anwendungsbezogenen, computergestützten Verarbeitung natürlicher Sprache. Hier wird maschinelles Lernen eingesetzt, um sprachliche Äußerungen mit linguistischen Kategorien zu annotieren (u. a. Tokenisierung, Lemmatisierung, *Part-of-Speech-Tagging*) oder verschiedenste andere Merkmale zu extrahieren (u. a. Personennamen, Ortsangaben, Zeitausdrücke). Letztlich geschieht dies mit dem Ziel, Anwendungen zu entwickeln, die dazu in der Lage sind, sprachliche Äußerungen inhaltlich zu verstehen oder Sprache zu generieren (vgl. Manning/Schütze 1999).

Auch für die Bearbeitung geisteswissenschaftlicher Fragestellungen wird maschinelles Lernen eingesetzt, was hier exemplarisch anhand zweier konkreter Anwendungsbeispiele vorgestellt wird: Erstens die automatische Klassifikation von Gemälden nach verschiedenen Epochenstilen (ein Beispiel für überwachtes maschinelles Lernen), zweitens die stilometrische Clusteranalyse von Dramentexten (ein Beispiel für unüberwachtes Lernen).

Automatische Klassifikation von Gemälden

Das erste Anwendungsbeispiel für Verfahren des maschinellen Lernens kommt aus der Kunstgeschichte. Die Darstellung folgt einem Beitrag von Babak Saleh und Ahmed Elgammal, zwei US-amerikanischen Informatikern (Saleh/Elgammal 2015). Die Autoren berichten von der Nutzung von *Wikiart* (http://www.wikiart.org), einer digitalen Sammlung von 81.449 Gemälden, um Algorithmen zu entwickeln, die Gemälde auf der Grundlage visueller Merkmale dem passenden Künstler, Epochenstil und Genre zuordnen können. Diese Fragestellung zu automatisieren, ist ein Beispiel für überwachtes maschinelles Lernen. Die denkbaren Klassen liegen bereits fest, fraglich ist die Zuordnung der Gemälde zu diesen Klassen. Anhand der Wikiart-Studie (und mit einem vereinfachenden Fokus auf den Epochenstil der Bilder) wird nun der typische Ablauf des überwachten maschinellen Lernens illustriert.

1. Vorbereitung: Die Datensammlung wird in mehrere Teile aufgeteilt: ein Referenzset für die Trainings- und Evaluationsphase sowie ein Anwendungsset. Das Re-

ferenzset ist in der Regel kleiner als das Anwendungsset, weil die hierfür notwendige Annotation (siehe Folgeschritt) aufwändig ist. In der Wikiart-Studie gab es kein Anwendungsset, da es in der Studie zunächst um die Entwicklung geeigneter Algorithmen ging, noch nicht um die Anwendung auf ganz neue Datensätze. Für die Epochenstil-Klassifikation wurde ein Subset der Gemäldesammlung verwendet, in das nur Gemälde aus 10 verschiedenen Epochenstilen aufgenommen wurden, für die mindestens 1500 Gemälde vorhanden waren (insgesamt 63.691 Gemälde).

2. Annotation: Das Referenzset wird annotiert, das heißt die enthaltenen Datensätze werden manuell den jeweiligen Klassen zugeordnet. Durch diese Annotation wird ein Bestand an Beispielen geschaffen, bei denen die Klassenzugehörigkeit von menschlichen Experten festgestellt wurde und der damit als Referenz dienen kann, sowohl zum Lernen des Zusammenhangs zwischen Merkmalen und Klassen, als auch zur Evaluation des Algorithmus. Je schwieriger die Lernaufgabe, d. h. je komplexer oder indirekter der Zusammenhang zwischen Merkmalen und Klassen, desto mehr Daten müssen in dieser Weise annotiert werden. In der Wikiart-Studie war die Datensammlung bereits annotiert, so dass für jedes Gemälde bereits bekannt war, welchem Epochenstil (u. a. Renaissance, Expressionismus oder Pop Art) es angehört.

3. Merkmalsgenerierung: Nun wird eine geeignete Anzahl von Eigenschaften der Gemälde modelliert, d. h. so beschrieben, dass sie ohne größeren Aufwand automatisch identifiziert oder gemessen werden können, und für jedes Beispiel erhoben. Häufig muss abgewogen werden, wie einfach die Merkmale zu erheben sind, einerseits, und wie vielversprechend sie für die Klassifikation sind, andererseits. Im Fall der hier beschriebenen Studie wurden nicht nur sehr einfache, visuelle, nicht unmittelbar bedeutungstragende Merkmale extrahiert (512 Merkmale, darunter visuelle Kanten mit ihrer Länge und Richtung sowie Farbprofile), sondern auch anspruchsvollere, bedeutungstragende Merkmale (darunter ein Set von 2659 Merkmalen, die im Bild repräsentierten Objekte festhalten, wie Haus, Wolke oder Fisch). Die Erkennung solcher komplexen Merkmale ist selbst eine Aufgabe für Verfahren des maschinellen Lernens, hier konnten die Autoren aber auf bestehende Algorithmen zurückgreifen. Die Daten (Gemälde mit Merkmalen und Klassen) werden nun als eine sehr große Merkmals-Matrix repräsentiert.

4. Trainingsphase: Der Computer ›lernt‹ oder ›entdeckt‹ nun auf der Grundlage der händisch annotierten Beispiele aus dem Trainingsset den Zusammenhang zwischen den Merkmalen und der für das Trainingsset ja bekannten Klassenzugehörigkeit. Hierfür werden *classifier* verwendet, von denen zahlreiche, sehr unterschiedliche Typen zur Verfügung stehen. Im Fall der Wikiart-Studie wurde unter anderem eine Variante der sogenannten »Support-Vector-Machines« eingesetzt (vgl. Han et al. 2011, Kapitel 9.3). Dieser Typ von Klassifikatoren dient dazu, in einem vieldimensionalen Merkmalsraum die Trennung des Raums in Unterräume vorzunehmen, in denen sich jeweils Objekte möglichst nur einer Klasse befinden und dabei möglichst breite Trennbereiche zu identifizieren, in denen keine Objekte sind. Auf diese Weise soll die Klassenzuordnung möglichst eindeutig und zuverlässig erfolgen.

5. Evaluationsphase: Für die Evaluation des gelernten Zusammenhangs wird das händisch annotierte Evaluationsset genutzt. Die Evaluation dient dazu, die Erkennungsqualität einzuschätzen, die der jeweilige Algorithmus auf der Grundlage der Trainingsdaten erreichen konnte, indem die Vorhersagen des Algorithmus mit den tatsächlichen Kategorien verglichen werden. Dabei werden üblicherweise **Precision** (welcher Anteil der Objekte, die einer Klasse zugeordnet wurden, gehören auch tatsächlich dieser Klasse an) und **Recall** (welcher Anteil der Objekte, die einer be-

stimmten Klasse zugehören, wurden auch tatsächlich dieser Klasse zugeordnet) erhoben und zu einer F-Score genannten Einschätzung der Güte des Classifiers kombiniert (s. Kap. 19 zu »Information Retrieval«). Außerdem wird in der Regel eine sogenannte ›Baseline‹ definiert. Damit ist die Klassifikationsgüte gemeint, die man mit einem sehr einfachen oder dem aktuellen Standard-Verfahren erreichen könnte. Ein neuer Algorithmus sollte zumindest besser als eine solche Baseline sein. Die Baseline in der Wikiart-Studie lag bei 0,23. Mit dem SVM-ähnlichen Verfahren und auf Grundlage der 2659 Objekt-Merkmale konnte eine Klassifikationsgüte von 0,27, mit einem anderen Classifier immerhin knapp 0,32 erreicht werden. Das ist zwar eine spürbare Verbesserung gegenüber der Baseline, angesichts der großen Anzahl von Beispielen pro Klasse und der geringen Anzahl von Klassen jedoch trotz der äußerst großen Komplexität der Aufgabe noch weit von einer endgültigen Lösung der Aufgabe entfernt.

6. Anwendungsphase: Schließlich erfolgt die Anwendung des gelernten Zusammenhangs auf die nicht annotierten Beispiele aus dem Anwendungsset. Der Algorithmus annotiert nun automatisch (mit der bei der Evaluation festgestellten Güte) die bisher nicht annotierten Objekte ebenfalls dahingehend, welcher Klasse sie zuzuordnen sind. Dadurch können bisher nicht annotierte Objekte ebenfalls mit einer Klassenzuordnung versehen werden, so dass sie beispielsweise bei einer Schlagwortsuche in einer Datenbank gefunden werden können oder für kontrastive Analysen genutzt werden können. (Dieser Schritt entfiel bei der Wikiart-Studie, bei der es vor allem um die Entwicklung der neuen Methode selbst ging.)

Auch wenn die Herausforderungen bei der Bildklassifikation noch enorm sind: Das überwachte maschinelle Lernen ist ein mächtiges Instrument, um mit verhältnismäßig geringem manuellen Annotationsaufwand auch sehr große Datenbestände nach komplexen Kategorien durchsuchen und analysieren zu können.

Stilometrie mit Burrows' Delta

Die **Stilometrie** ist ein computergestütztes Verfahren der Erhebung stilistischer (oder allgemeiner: sprachlicher) Merkmale und ihrer Häufigkeiten in Texten, sowie der Nutzung dieser Merkmale und Häufigkeiten für die Bestimmung der Ähnlichkeit der Texte zueinander und die Klassifikation oder das Clustering der Texte (für einen Überblick vgl. Juola 2006 und Stamatatos 2009). Ursprünglich wurde die so festgestellte große Ähnlichkeit zweier Texte vor allem als Hinweis dafür interpretiert, dass die beiden Texte vom gleichen Autor geschrieben worden sind, das heißt die Methode wurde für die Autorschaftsattribution eingesetzt. Als Pionierarbeit sind die Studien von Alvar Ellegård (1962) zu einer Reihe unter dem Pseudonym ›Junius‹ erschienener politischer Briefe sowie die Arbeiten von Frederik Mosteller und David Wallace (1963) zu den *Federalist Papers* zu nennen. Ein Meilenstein in der jüngeren Geschichte der Stilometrie ist außerdem ein Artikel von John Burrows, in dem er das stilometrische Distanzmaß ›Delta‹ vorschlägt (Burrows 2002). Seit einigen Jahren wird zunehmend deutlich, dass auf diese Weise erhobene Ähnlichkeiten nicht nur für die Autorschaftsattribution, sondern auch für andere Fragestellungen, insbesondere Fragen der Gattungs- und Epochenzugehörigkeit, von Interesse sind.

Vieldimensionaler Vektorraum: Stilometrische Analysen dieser Art beruhen auf einer Repräsentation der Textsammlung als Merkmals-Matrix, wie sie einleitend bereits beschrieben wurde. Dies erlaubt es, die Sammlung als vieldimensionalen **Vek-**

torraum aufzufassen, was den Vorteil hat, dass dadurch unter anderem die Nutzung verschiedener Ähnlichkeitsmaße für Vektoren ermöglicht wird (s. unten und Kap. 19 zu »Information Retrieval«). Genauer: In diesem Modell wird jeder Text als ein Vektor in einem vieldimensionalen Vektorraum repräsentiert, wobei jedes Type als eine Dimension verstanden wird und die Häufigkeit eines Types in einem Text als der Wert des Vektors dieses Textes in dieser Dimension gilt. Das ermöglicht wiederum, den Grad der Ähnlichkeit bzw. Differenz zwischen einzelnen Datensätzen als räumliche Nähe bzw. Distanz zwischen den Merkmalsvektoren aufzufassen und diese Nähe bzw. Distanz mit verschiedenen Distanzmaßen zu erheben.

Stilometrische Analysen können sowohl in überwachten wie auch in unüberwachten Szenarien eingesetzt werden. In einem überwachten Szenario lernt der Computer den Zusammenhang zwischen einem bestimmten Bereich im Vektorraum und einer Klasse (wie z. B. einem Autor) und kann diese dann auf neue Texte anwenden, d. h. beispielsweise einen Text unbekannter Autorschaft in Abhängigkeit von seiner Position im Vektorraum einem der in der Textsammlung vertretenen Autoren zuordnen. In einem unüberwachten Szenario dagegen steht die Erkenntnis über die mehr oder weniger große Nähe oder Distanz (d. h. Ähnlichkeit oder Unterschiedlichkeit) von mehreren Untergruppen von Texten in der analysierten Textsammlung und die Platzierung einzelner Texte innerhalb solcher Gruppen im Fokus.

Anwendungsbeispiel: Der folgende Anwendungsfall betrifft die Analyse einer Sammlung französischer Komödien aus dem 17. Jahrhundert und ist ein unüberwachtes Szenario. Die hier verfolgte Fragestellung ist die, ob das Stück *Dom Garcie de Navarre* von 1661, das üblicherweise Molière zugeschrieben wird, nicht doch von Pierre Corneille verfasst wurde, wie dies vereinzelt vorgeschlagen wurde (vgl. Schöch 2014). Im Vordergrund steht also die Ähnlichkeit dieses Stücks zu den übrigen, Molière und Pierre Corneille eindeutig zuzuordnenden Stücken. Das konkrete Vorgehen (das sich methodisch an dem inzwischen klassischen Burrows' Delta-Verfahren orientiert), gliedert sich in die folgenden Schritte:

1. Vorbereitung der Textsammlung: Zunächst einmal müssen eine für die verfolgte Fragestellung geeignet zusammengesetzte Textsammlung erstellt und die zu analysierenden Texte so vorbereitet werden, dass sie sinnvoll vergleichbar sind (s. Kap. 16 zu »Aufbau von Datensammlungen«). Insbesondere die Auswahl der zu analysierenden Texte in Abhängigkeit von der Fragestellung gehört zur Vorbereitung der Textsammlung. Im Fall der Komödien-Studie ist die Quelle der digitalen Texte eine dem französischen Theater gewidmete Plattform (http://www.theatre-classique.fr). Die hier verwendete Textsammlung enthält 46 Komödien aus der zweiten Hälfte des 17. Jahrhunderts, die von verschiedenen Autoren (Pierre Corneille, Thomas Corneille, Molière, Regnard und Scarron) verfasst wurden, zudem die strittige Komödie *Dom Garcie de Navarre*.

2. Merkmals-Matrix: Nun werden die Liste aller in der Textsammlung vertretenen Types erstellt, die Häufigkeiten jedes Types in jedem Text ermittelt und die Ergebnisse in einer Merkmals-Matrix festgehalten. Im Fall der Dramenanalyse erfolgen dieser und die weiteren Schritte mit dem stilometrischen Analysetool ›stylo‹, das als ein Paket für die Statistik-Umgebung R zur Verfügung steht (Eder et al. 2016). Für die Komödien-Studie ergibt sich eine Merkmals-Matrix mit 47 Spalten (eine Spalte pro Drama) und 5000 Zeilen (für die häufigsten 5000 Types in der Sammlung; für weitere Types werden die Häufigkeiten nicht erhoben).

3. Merkmals-Behandlung: Erstens wird die Merkmals-Matrix nun nach absteigender Häufigkeit der Types in der Textsammlung insgesamt sortiert und eine Be-

grenzung auf eine bestimmte Anzahl der häufigsten Types (sog. *most frequent words*) festgelegt. Zweitens werden die absoluten Häufigkeiten in ›z-scores‹ überführt (wie im Abschnitt zu »Normalisierung von Verteilungen« beschrieben). Die Begrenzung der Wortliste legt u. a. fest, ob nur Funktionswörter oder auch weitere Wörter in den Vergleich der Texte mit einbezogen werden. Die Transformation der Häufigkeiten in z-scores vermeidet, dass nur die 20 oder 30 häufigsten Types den Merkmalsvektor und damit auch die Ähnlichkeitsbestimmung dominieren bzw. sorgt dafür, dass die Häufigkeitsunterschiede aller berücksichtigter Types in gleicher Weise in die Messung der stilistischen Ähnlichkeit der Texte zueinander eingehen. Im Fall der Komödien-Studie werden die 2000 häufigsten Types berücksichtigt.

4. Distanzmaß: Nun wird ein Distanzmaß angewandt, um die relative stilistische Unterschiedlichkeit bzw. Ähnlichkeit aller Texte zueinander zu ermitteln. Zweck eines Distanzmaßes ist es, die zahlreichen kleinen Unterschiede in der Häufigkeit der Types in zwei Texten zu einem einzigen Wert zusammenzufassen, der dann die zusammengenommene Unterschiedlichkeit bzw. Ähnlichkeit der beiden Texte ausdrückt. Verschiedene Distanzmaße unterscheiden sich darin, wie sie diesen zusammenfassenden Wert ermitteln. Drei wichtige Distanzmaße sind in Abbildung 60 dargestellt.

Bei der **Manhattan-Distanz** werden die absoluten Distanzen zweier Vektoren in jeder einzelnen Dimension addiert. Der Name leitet sich von dem Weg ab, den man in Manhattan wegen der rechtwinklig angelegten Straßenblocks von einem Ort zum anderen nehmen muss. Die Distanz in jeder einzelnen Dimension fließt mit gleichem Gewicht in die Gesamtdistanz ein.

Die **euklidische Distanz** hingegen entspricht der Vogelfluglinie zwischen zwei Orten; hier werden die Distanzen in jeder Dimension erst quadriert, dann summiert, bevor die Wurzel aus der Summe gezogen wird. Große Distanzen in einer einzelnen Dimension werden durch die Quadrierung stärker gewichtet als kleinere Distanzen.

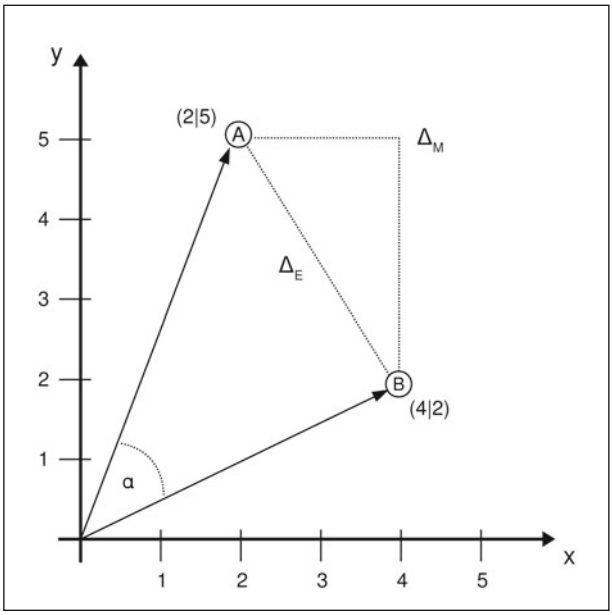

Abb. 60 Drei Distanzmaße. Es liegen zwei Texte A und B vor, sowie zwei Dimensionen x und y. Manhattan-Distanz = $|4–2| + |2–5| = 2 + 3 = 5$. Euklidische Distanz = $(|4–2|^2 + |2–5|^2)^{-2} = (4 + 9)^{-2} = 3{,}6$. Kosinus $\alpha = 41{,}6°$

Beim **Kosinus-Maß** schließlich wird nicht die Distanz zwischen den Vektoren, sondern der Winkel der Vektoren zueinander berechnet. Das wirkt sich dergestalt aus, dass nur die Richtung, nicht aber die Länge eines Vektors für das Ergebnis der Berechnung entscheidend ist. In jedem Fall drückt der resultierende Wert die Nähe der beiden jeweiligen Vektoren im Vektorraum aus und damit, so die Annahme, auch die stilistische bzw. sprachliche Ähnlichkeit der beiden jeweiligen Texte.

Im Fall der Komödien-Studie wurde das klassische Burrows' Delta verwendet, bei dem die standardisierten Häufigkeiten (z-scores) in Kombination mit der Manhattan-Distanz verwendet werden. Es ergibt sich hieraus eine Ähnlichkeits-Matrix von 47 mal 47 Einträgen, in der die Distanzen aller Komödien zueinander festgehalten sind.

5. Clustering und Visualisierung: Um diese immer noch große Menge an Informationen in eine lesbare Form zu bringen, wird die Ähnlichkeitsmatrix nun (mit einem von mehreren denkbaren Verfahren) in eine Cluster-Matrix überführt, auf deren Grundlage ein hierarchisches **Baumdiagramm** (ein sogenanntes **Dendrogramm**) erstellt wird. Im Fall der Komödien-Studie wurde das weit verbreitete Ward-Clustering verwendet, womit sich das in Abbildung 61 gezeigte Dendrogramm ergibt. Es liest

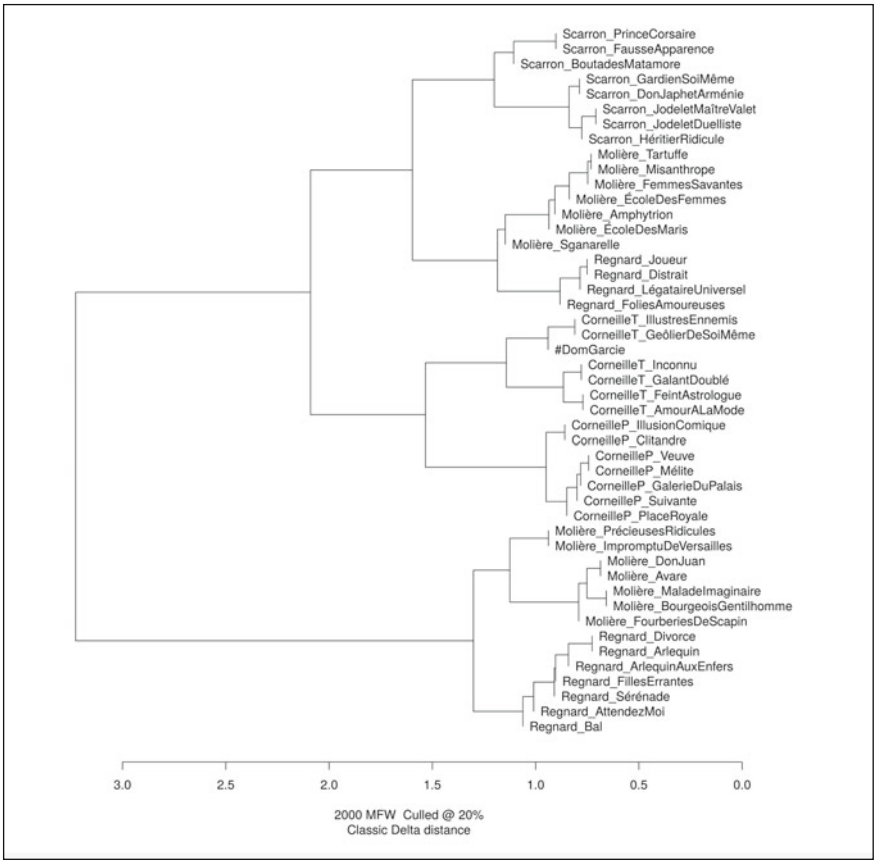

Abb. 61 Stilometrisches Dendrogramm der Ähnlichkeitsbeziehungen zwischen 47 französischen Komödien des 17. Jahrhunderts, darunter ein Stück umstrittener Autorschaft (Label: #DomGarcie)

sich wie folgt: Ganz rechts stehen alle analysierten Komödien, sie sind die Blätter des Baumdiagramms; zwei Dramen, die durch eine Linie direkt verbunden werden, sind sich zueinander ähnlicher, als zu irgendeinem anderen Drama und bilden einen Ast des Diagramms. Zwei Äste, die durch eine Linie direkt verbunden sind, bilden wiederum zwei Gruppen von Dramen, die sich zueinander sehr ähnlich sind. Je weiter links im Diagramm zwei Dramen bzw. Äste aufeinandertreffen, desto weniger ähnlich sind sie sich. Den Wert des Distanzmaßes an einem bestimmten Knotenpunkt kann man bestimmen, indem man vertikal nach unten auf die X-Achse projiziert und dort den entsprechenden Wert abliest. Ein Wert von 0 würde zwei identische Texte kennzeichnen, im Fall der Komödien-Studie liegt die Mehrzahl der Werte zwischen 0,5 und 1,5.

6. Interpretation: Der letzte Schritt ist die Interpretation des Dendrogramms im Licht des über die Textsammlung bekannten Wissens (u. a. Autoren, Gattungen, Entstehungszeit, Inhalt) und dem Wissen um die Parameter des Verfahrens (u. a. Anzahl der Wörter, Distanzmaß). Die im Dendrogramm sichtbaren Ähnlichkeitsbeziehungen und Gruppenbildungen sind nicht als absolute, objektive Ergebnisse zu verstehen, sondern als interpretationsbedürftiges Produkt zahlreicher Faktoren, die in komplexer Weise miteinander interagieren können.

Im Fall der Komödien-Studie wird zunächst einmal deutlich, dass die Textsammlung sich in zwei große Gruppen gliedert, die der Unterscheidung von Komödien in **Prosa** (untere Gruppe mit Komödien von Molière und Regnard) und in **Versform** (obere Gruppe mit Komödien aller Autoren) entsprechen. Innerhalb dieser Gruppen zeigen sich weitere Untergruppen, die jeweils den Komödien eines Autors entsprechen. Mit den gewählten Parametern ist also eine zuverlässige Gruppierung von Stücken nach ihren Autoren möglich. Das umstrittene Stück aber, *Dom Garcie de Navarre*, erscheint an einem überraschenden Ort: nicht bei Molière, aber auch nicht bei Pierre Corneille, sondern in geringem Abstand zu zwei Stücken von Thomas Corneille.

Dieses Ergebnis, das sich auch bei anderen Parametersetzungen in relativ robuster Weise so ergibt, wirft Fragen auf: Hat sich die Forschung zu *Dom Garcie* gleicht doppelt getäuscht, oder spielen andere Faktoren hier eine Rolle? Es stellt sich heraus, dass vermutlich die Gattungszugehörigkeit des Stücks für das Ergebnis entscheidend ist. Denn *Dom Garcie* ist eine sogenannte ›heroische Komödie‹, die Merkmale von Komödie und heroischem Roman vermischt und der Tragikomödie nicht unähnlich ist. Molière hat nur eine einzige Komödie dieses Typs geschrieben, so dass sich in seinem Werk kein unmittelbar verwandtes Stück finden lässt, dem *Dom Garcie* besonders ähnlich wäre. Thomas Corneille dagegen hat mehrere Komödien geschrieben, die als heroische Komödien gelten können oder eine ihnen ähnliche Thematik haben. Dass *Dom Garcie* nun Thomas Corneille zugeordnet wird, ist demnach weniger als Hinweis auf die Autorschaft des Stücks zu werten, denn als Hinweis auf die gattungsbezogene Verwandtschaft des Stücks zu denen Thomas Corneilles. Und es zeigt, dass stilometrisch gewonnene Aussagen zur Autorschaft auch dann von Faktoren wie der literarischen Gattung beeinflusst sein können, wenn es sich um eine eigentlich relativ homogene Textsammlung wie derjenigen der Komödien des 17. Jahrhunderts handelt.

20.4 | Neuere Entwicklungen

Quantitative Verfahren können in den digitalen Geisteswissenschaften für zahlreiche Anwendungsfelder und Fragestellungen eingesetzt werden, unter anderem im Sinne statistischer Beschreibungen von Datensammlungen, der kontrastiven Analyse mehrerer Datensammlungen oder ihrer Teile, der überwachten Klassifikation von Untersuchungsgegenständen oder der explorativen und unüberwachten Gruppenbildung auf der Grundlage der Ähnlichkeit von Untersuchungsgegenständen. Diese und ähnliche Verfahren der statistischen Analyse und des maschinellen Lernens werden für die Bearbeitung vielfältiger geisteswissenschaftlicher Fragestellungen verwendet. Für die digitalen Geisteswissenschaften vielversprechende, neuere Entwicklungen im Bereich quantitativer Verfahren der Datenanalyse sind vektorenbasierte Modelle für die Repräsentation des semantischen Gehalts von Wörtern (*word embeddings* vgl. Mikolov et al. 2013) sowie das ›Deep Learning‹, eine moderne Form neuronaler Netze, die für besonders herausfordernde Aufgaben wie die Objekterkennung in Bildern eingesetzt werden (*Word embeddings* vgl. LeCun et al. 2015). Diese und weitere Verfahren weisen auf das Potential hin, neueste Entwicklungen im Bereich der Künstlichen Intelligenz (vgl. Russel und Norvig 2009) zukünftig auch für die Bearbeitung geisteswissenschaftlicher Fragestellungen einzusetzen.

Literatur

Alpaydin, Ethem: *Introduction to Machine Learning*. Cambridge ²2010.
Baroni, Marco/Evert, Stefan: »Statistical methods for corpus exploitation«. In: Anke Lüdeling/Merja Kytö: *Corpus Linguistics. An International Handbook 2*. Berlin 2009, 777–803.
Blei, David M.: »Probabilistic topic models«. *Communications of the ACM* 55.4, 2012: 77–84.
Burrows, John: »›Delta‹: A measure of stylistic difference and a guide to likely authorship«. In: *Literary and Linguistic Computing* 17/3 (2002), 267–87.
Burrows, John: »All the way through: Testing for authorship in different frequency strata«. In: *Literary and Linguistic Computing* 22/1 (2007), 27–47.
Bortz, Jürgen/Schuster, Christof: *Statistik für Human- und Sozialwissenschaftler*. Berlin 2010.
Craig, Hugh/Kinney, Arthur F.: *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge 2009.
Eder, Maciej/Kestemont, Mike/Rybicki, Jan: »Stylometry with R: A package for computational text analysis«. In: *The R Journal* 16/1 (2016), 1–15.
Ellegård, Alvar: *A Statistical Method for Determining Authorship: The Junius Letters, 1769–1772*. Gothenburg 1962.
Gries, Stefan Th.: »Useful statistics for corpus linguistics«. In: Aquilino Sánchez/Moisés Almela (Hg.): *A Mosaic of Corpus Linguistics: Selected Approaches*. Frankfurt a. M. 2010, 269–291.
Han, Jiawei/Kamber, Micheline/Pei, Jian: *Data Mining: Concepts and Techniques*. Burlington ³2011.
Jannidis, Fotis: »Methoden der computergestützten Textanalyse«. In: Ansgar Nünning/Vera Nünning (Hg.): *Methoden der literatur- und kulturwissenschaftlichen Textanalyse*. Stuttgart/Weimar 2010, 109–132.
Jockers, Matthew L.: *Macroanalysis – Digital Methods and Literary History*. Champaign, Ill. 2013.
Juola, Patrick: »Authorship attribution«. In: *Foundations and Trends in Information Retrieval* 1/3 (2006), 233–334.
LeCun, Yann/Bengio, Yoshua/Hinton, Geoffrey: »Deep learning«. In: *Nature* 521/7553 (2015), 436–444.
Lijffijt, Jefrey/Nevalainen, Terttu/Säily, Tanja/Papapetrou, Panagiotis/Puolamäki, Kai/Mannila, Heikki: »Significance testing of word frequencies in corpora«. In: *Digital Scholarship in the Humanities* 31/2 (2014), 374–97.
Manning, Christopher S./Schütze, Hinrich: *Foundations of Statistical Natural Language Processing*. Cambridge 1999.

Mikolov, Tomas/Chen, Kai/Corrado, Greg/Dean, Jeffrey: »Efficient estimation of word representations in vector space« (2013), http://arxiv.org/abs/1301.3781 (31.10.2016).
Moretti, Franco: *Graphs, Maps, Trees: Abstract Models for a Literary History*. London 2005.
Mosteller, Frederick/Wallace, David L.: »Inference in an authorship problem«. In: *Journal of the American Statistical Association* 58/302 (1963), 275–309.
Oakes, Michael P.: *Statistics for Corpus Linguistics*. Edinburgh 1998.
Piper, Andrew: »There will be numbers«. *Cultural Analytics* 1 (2016). Online: http://culturalanalytics.org/2016/05/there-will-be-numbers/ (31.10.2016).
Russell, Stuart J./Norvig, Peter: *Artificial Intelligence. A Modern Approach*. Boston ^{3}2009.
Saleh, Babak/Elgammal, Ahmed: »Large-scale classification of fine-art paintings: Learning the right metric on the right feature« (2015), http://arxiv.org/abs/1505.00855 (31.10.2016).
Schöch, Christof: »Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik«. In: Christof Schöch/Lars Schneider (Hg.): *Literaturwissenschaft im digitalen Medienwandel*. Beihefte von *Philologie im Netz* 7 (2014), 130–57. Online: http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf (31.10.2016).
Stamatatos, Efstathios: »A survey of modern authorship attribution methods«. In: *Journal of the Association for Information Science and Technology* 60/3 (2009), 538–556.

Christof Schöch