Christof Schöch (Würzburg)

Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik

The digital age, by making large amounts of text available to us, prompts us to develop new and additional reading strategies supported by the use of computers and enabling us to deal with such amounts of text. One such "distant reading" strategy is stylometry, a method of quantitative text analysis which relies on the frequencies of certain linguistic features such as words, letters or grammatical units to statistically assess the relative similarity of texts to each other and to classify texts on this basis. This method is applied here to French drama of the seventeenth century, more precisely to the now famous "Corneille / Molière-controversy". In this controversy, some researchers claim that Pierre Corneille wrote several of the plays traditionally attributed to Molière. The methodological challenge, it is shown here, lies in the fact that categories such as authorship, genre (comedy vs. tragedy) and literary form (prose vs. verse) all have an influence on stylometric distance measures and classification. Cross-genre and cross-form authorship attribution needs to distinguish such competing signals if it is to produce reliable attribution results. This contribution describes two attempts to accomplish this, parameter optimization and feature-range selection. The contribution concludes with some more general remarks about the use of quantitative methods in a hermeneutic discipline such as literary studies.

Einleitung: Die digitale Wende und neue Verfahren der Textanalyse

In der Frühen Neuzeit ereignete sich mit der Erfindung des Buchdrucks mit beweglichen Lettern eine Revolution in den technischen Möglichkeiten der Reproduktion von Texten, die mittelfristig auch signifikante Veränderungen in den Modalitäten der Rezeption von Texten mit sich brachte. Diese durch sozioökonomische Entwicklungen geförderten Veränderungen sind dadurch gekennzeichnet, dass die intensive Lektüre, bei der sehr wenige oder gar ein einziger Text wiederholt gelesen wurde, nach und nach ersetzt wurde durch eine extensivere Lektürepraxis, bei der eine zunehmenden Zahl unterschiedlicher Texte in diversen Kontexten rezipiert werden (Wittmann 1999).

Mit der digitalen Wende, die wir derzeit erleben, werden die Mittel der Repräsentation, Reproduktion, des Austauschs und der Manipulation von Texten erneut radikal verändert.[1] Insbesondere durch zahlreiche groß angelegte Digitalisierungs-

[1] Selbstverständlich ist die digitale Wende nicht auf Texte beschränkt, sondern bezeichnet eine technologische und gesellschaftliche Entwicklung, im Zuge derer eine Vielzahl alltäglicher Tätigkeiten, Erfahrungen und Wissenspraktiken von digitalen Medien geprägt sind. Der Endpunkt die-

Projekte weltweit erweitert sich das digital verfügbare kulturelle Erbe in Textform ständig, die Texte sind untereinander vernetzt und liegen in fluiden Formen vor. Immer größere Anteile des kulturellen Erbes liegen in elektronischer Form buchstäblich unter unseren Fingerspitzen und fordern uns dazu auf, sie zu nutzen. Letztlich hat hier eine Revolution im wörtlichen Sinne einer Umkehrung von Hierarchien stattgefunden, weil die entscheidende, knappe Ressource nicht mehr der verfügbare Text und seine Inhalte sind, sondern die verfügbare Zeit und Aufmerksamkeit für Texte und andere Träger von Informationen. Wie wir alle selbst erleben, wird das Verhältnis zwischen verfügbarer Lebens- und Lesezeit und vorhandenem Lesestoff immer ungünstiger.

Im Bereich der Literaturwissenschaft ist das Spannungsfeld zwischen extensiver Lektüre, also der flüchtigen Lektüre sehr vieler, unterschiedlicher Texte einerseits, und dem "close reading", also der detaillierten Lektüre und Interpretation weniger oder wenig umfangreicher Texte andererseits, erhalten geblieben. Beide Modalitäten der Lektüre sind für die literaturwissenschaftliche Analyse von Sammlungen tausender oder gar Millionen von Texten nicht geeignet (siehe Crane 2006). Der zunehmend verbreitete Wunsch, die Grenzen von einschränkenden und historisch determinierten Kanons zu durchbrechen, wird von unseren begrenzten kognitiven Kapazitäten unterminiert. Digitale Methoden der computergestützten, quantitativen Textanalyse, wie beispielsweise die Stilometrie oder das Topic Modeling, die man mit Franco Moretti dem Paradigma des "distant reading" zuordnen oder mit Matthew Jockers als zwei Modalitäten der "macroanalysis" bezeichnen könnte,[2] sind Strategien, dieses Dilemma zu lösen oder es zumindest möglichst geschickt zu umgehen. Beide Methoden sind im größeren Kontext der "digitalen Geisteswissenschaften" zu sehen.[3]

---

ser Entwicklung ist von dem britischen Informationstheoretiker Luciano Floridi als "life in the infosphere" bezeichnet worden (Floridi 2010: 14). Die Digitalisierung der Wissenschaft ist, wie Christine Borgman in ihrem Buch über *Scholarship in the Digital Age* herausgearbeitet hat, zugleich eine technologische und soziale Herausforderung (Borgman 2010). Auch die deutsche Romanistik entdeckt zunehmend die Möglichkeiten digitaler Ressourcen, Methoden und Tools (Stierle 2013).

[2] Der Begriff "distant reading" geht auf Franco Moretti (2000) zurück und wurde durch Moretti 2005 weithin bekannt; den Begriff “macroanalysis” hat jüngst Matthew Jockers (2013) geprägt.

[3] Die digitalen Geisteswissenschaften (oder "digital humanities") sind im Schnittfeld von Geisteswissenschaften, Medienwissenschaften und Informatik angesiedelt. Zum Konzept und zur Geschichte dieser "Interdisziplin" oder dieses Methodenfeldes siehe McCarty 1999 und Unsworth 2002. Einführende Texte zu verschiedenen Methoden bietet der *Companion to Digital Humanities* (Siemens et al. 2004), einen Überblick zur "Computerphilologie" findet man bei Jannidis 2007.

Die Stilometrie meint computergestützte Verfahren der quantitativen Erhebung stilistischer Merkmale für die Klassifikation von Texten. Traditionell für die Attribution von Autorschaft eingesetzt, wird aktuell deutlich, dass die Stilometrie auch für die Literaturgeschichtsschreibung und Gattungstheorie interessante Perspektiven eröffnet.[4] Topic Modeling bezeichnet ein probabilistisches Verfahren der automatischen Extraktion von Gruppen thematisch verwandter Begriffe aus großen Textsammlungen und der Analyse der zeitlichen oder gattungsabhängigen Distribution solcher Gruppen innerhalb der untersuchten Textsammlung.[5]

Ziel des vorliegenden Beitrags ist es, das Verfahren der Stilometrie vorzustellen und von der Erprobung stilometrischer Klassifikationsverfahren im Bereich des französischen Theaters des siebzehnten Jahrhunderts zu berichten. Diese Erprobung ist im Kontext eines der methodischen Schwerpunkte des europäischen Infrastruktur-Projekts DARIAH (siehe DARIAH-DE 2012 und http://de.dariah.eu) zu sehen, der sich mit der computergestützten Analyse großer Textsammlungen befasst, wobei das Ziel dieser Aktivitäten letztlich ist, auf eine größere Verbreitung quantitativer Methoden der Textanalyse in den Geisteswissenschaften und eine bessere Vernetzung der GeisteswissenschaftlerInnen verschiedener Disziplinen hinzuwirken, die in Europa an der computergestützten Analyse großer Textsammlungen arbeiten.[6] Für die Erprobung stilometrischer Klassifikationsverfahren im Bereich des französischen Theaters des siebzehnten Jahrhunderts wurde das stylo-Paket eingesetzt, das von Maciej Eder und Jan Rybicki für die Statistikumgebung R entwickelt wurde (Eder & Rybicki 2011).

Zunächst soll das Verfahren der Stilometrie knapp vorgestellt werden. Ausgangspunkt für alles Weitere wird dann die mittlerweile berühmte Corneille / Molière-Kontroverse sein. Im Zentrum der Kontroverse steht die Frage, ob Corneille möglicherweise einige oder gar viele der traditionell Molière zugesprochenen Werke

[4] Für einen historischen und systematischen Überblick zur Stilometrie siehe Holmes 1994 und Juola 2006. Für die Beziehung zwischen Stilometrie und Literaturgeschichte siehe Jannidis & Lauer 2013.

[5] Für eine knappe Einführung siehe Templeton 2011. Für einen erhellenden Überblick siehe Blei 2011.

[6] Zu diesem Schwerpunkt gehört unter anderem die Entwicklung einer Beispielanwendung, die Textgrids *Digitale Bibliothek* [http://www.textgridrep.de] mit dem Analyse- und Explorationstool *Voyant Tools* ([http://www.voyant-tools.org], Sinclair & Rockwell 2013) verbindet, so dass literatur- und kulturwissenschaftliche Fragestellungen damit bearbeitet werden können; vgl. [https://de.dariah.eu/digivoy]; außerdem die Erarbeitung eines Überblicks über zentrale methodische Ansätze wie Stilometrie und Topic Modeling und deren aktuell zentrale methodische Fragen; schließlich die Durchführung von Experten-Meetings und Workshops zum Thema Textanalyse.

verfasst hat. Eine solche Frage erfordert die Lösung sehr konkreter Probleme, wie das des relativen Einflusses von Autorschaft und Gattungszugehörigkeit bei stilometrischen Klassifikationssaufgaben. Von zwei konkreten Versuchen, solche Fragen methodisch in den Griff zu bekommen, wird hauptsächlich die Rede sein. Das erste stilometrische Experiment betrifft die Parameter-Optimierung auf der Grundlage des Corneille / Molière-Korpusses. Das zweite stilometrische Experiment betrifft etwas grundsätzlichere Versuche der "Signal-Trennung", ebenfalls bezogen auf das französische Theater der Klassik. Beide Experimente werfen allgemeinere Fragen nach der Verlässlichkeit und Nachvollziehbarkeit statistischer Verfahren in den Philologien auf, und damit auch nach deren Relevanz für philologische Fragestellungen.

1 Stilometrische Verfahren der Textanalyse

Die Stilometrie ist eines von mehreren Verfahren, die dem Bereich der quantitativen Textanalyse zugerechnet werden können. Der Begriff Stilometrie bezeichnet dabei computergestützte Verfahren der Erhebung stilistischer Merkmale und ihrer Häufigkeiten in Texten, sowie der Nutzung dieser Merkmale und Häufigkeiten für die Klassifikation von Texten. Die Methode selbst geht auf Überlegungen zurück, die noch vor dem Zeitalter des Computers liegen. Schon im Jahr 1851 hat der britische Mathematiker Augustus de Morgan stilometrische Prinzipien konzipiert, indem er den Vergleich der durchschnittlichen Wortlängen in verschiedenen Texten zur Feststellung der Autorschaft anonymer Texte vorschlug (vgl. Juola 2006: 240). Die von de Morgan vorgeschlagene Methode wurde von Mendenhall (1887) erstmals erprobt. Und der polnische Philosoph Wincenty Lutosławski definierte in seiner Schrift *Principes de stylométrie appliqués à la chronologie des œuvres de Platon* von 1898 die Methode der "Stilometrie" im modernen Sinne, nämlich als die "recherche d'affinités stylistiques" (vgl. hierzu Pawłowski & Pacewicz 2004). Mit der Verbreitung des Computers seit den 1960er Jahren war die technische Grundlage dafür geschaffen, dass stilometrische Verfahren systematisch eingesetzt wurden.

Als Pionierarbeiten sind die Studien von Frederik Mosteller und David Wallace (1963) zu den *Federalist Papers* und von John Burrows (1987) zur Figurenrede im Werk von Jane Austen zu nennen. Einen deutlichen Aufschwung und größere

Verbreitung haben quantitative Verfahren der Analyse von Sammlungen literarischer Texte jedoch erst in den letzten zehn Jahren erlebt.[7] Dies liegt sicherlich daran, dass vermehrt geeignete digitale Texte vorliegen und daran, dass es zunehmend nutzerfreundlichere und leistungsfähigere Werkzeuge gibt. Immer mehr LiteraturwissenschaftlerInnen experimentieren mit solchen Verfahren und die Menge an Erfahrungswerten und Einsatzmöglichkeiten steigt. Dabei gibt es heute mehrere etablierte Anwendungsfelder der Stilometrie: Am bekanntesten ist sicherlich die Autor-Attribution; hier werden für Texte, deren Autor unbekannt oder umstritten ist, mögliche Autoren festgestellt. Auch für die chronologische Einordnung von Einzeltexten bekannter Autoren in deren Gesamtwerk wird die Stilometrie eingesetzt. Ein Anwendungsfeld außerhalb der Philologien ist die digitale Forensik, bei der es um die Authentifizierung von Texten oder die Plagiats-Detektion geht. Neuere Entwicklungen in der Literaturwissenschaft setzen auf die Stilometrie und andere quantitative Verfahren, um die Gattungstheorie und die Literaturgeschichte neu zu denken.

In der Tat ist die Stilometrie trotz ihres Namens im Kern ein Verfahren, das der Literaturgeschichte nicht weniger nahesteht als der Stilistik: Es geht der Stilometrie nicht in erster Linie um die Beschreibung von Texteigenschaften oder um die Definition oder Charakterisierung eines Autoren- oder Epochenstils. Vielmehr nutzt die Stilometrie bestimmte Textmerkmale, um die relative Ähnlichkeit oder Differenz verschiedener Texte zu bestimmen und um auf dieser Grundlage Texte zu klassifizieren oder zu gruppieren. Die der Klassifikation zugrundeliegenden Kategorien sind dabei nicht auf die Autoren der Texte beschränkt, sondern können sich auch auf die Gattung oder Untergattung, auf die Epochenzugehörigkeit oder auf das Geschlecht der Autoren beziehen. Da es sich um ein sog. unüberwachtes Klassifizierungsverfahren handelt, sind die Zielkategorien oder Gruppen nicht vorgegeben, sondern ergeben sich aus der jeweils festgestellten Ähnlichkeit der Texte. Dass sich die Autorkategorie nicht immer isolieren lässt, ist gerade eines der Probleme – oder auch eines der Ergebnisse – der aktuellen methodischen Debatten in diesem Feld. In der Tat sind die methodischen Schwierigkeiten und Unwägbarkeiten immer noch enorm, gerade für Literaturen in anderen Sprachen als

[7] Nicht zuletzt hat die Methode kürzlich auch eine literarische Behandlung erfahren, und zwar in Mitzi Morris' Roman *Poetic Justice* aus dem Jahr 2012, den man wohl als den ersten stilometrischen Kriminalroman in der Geschichte der Gattung bezeichnen kann (vgl. Schöch 2013).

dem Englischen, und dies trotz einiger wichtiger Arbeiten zu verschiedenen europäischen Sprachen (u.a. Van Dalen-Oskam & Van Zundert 2007, Rybicki & Eder 2011, Kestemont et al. 2012).

In den nun folgenden Ausführungen soll es daher um zwei Dinge gehen: Einerseits um methodische Fragen, die sich vor allem um das Problem der Klassifikation von Texten drehen; andererseits um einen literaturwissenschaftlichen Anwendungsfall. Bevor ich zu einem solchen Anwendungsfall komme, sind in Bezug auf die stilometrische Verfahrensweise noch einige grundlegende technische Hinweise notwendig. Um ein bestimmtes stilometrisches Verfahren zu beschreiben, kann man drei zentrale Aspekte herausgreifen: Erstens, welche sprachlichen Merkmale die Grundlage bilden; zweitens, welche Methode zur Berechnung der statistischen Ähnlichkeit der Texte angewandt wird; und drittens, welche Visualisierungstechnik gewählt wird. In der Tat ist ein genaues Verständnis der Funktionsweise dieser Verfahrens entscheidend, damit wir uns nicht dem Algorithmus als undurchschaubarer, magischer "black box" ausliefern. Die Ergebnisse des stilometrischen Verfahrens und ihre Interpretation hängen zu sehr von den Details seiner Durchführung ab.

Die sprachlichen Merkmale, die für stilometrische Verfahren verwendet werden, sind meistens Oberflächenphänomene, die sich leicht feststellen und quantifizieren lassen, also Wörter (*types* oder *tokens*) oder Buchstaben und ihre Frequenzen. Eine geeignete Vorbereitung der Texte durch linguistische Annotation vorausgesetzt, können aber auch die Frequenzen grammatikalischer Kategorien Grundlage stilometrischer Verfahren sein. Zudem können Maße wie die durchschnittliche Satzlänge, die durchschnittliche Wortlänge, der Anteil bestimmter Wortklassen am Gesamttext oder die N-Gramme von Buchstaben, Wörtern oder Wortklassen berücksichtigt werden. Wie direkt diese Merkmale mit Phänomenen verknüpft sind, die bei der literaturwissenschaftlichen Lektüre von Einzeltexten beobachtet werden können, kann stark variieren. In der Regel geht es zunächst darum, für jeden der untersuchten Texte die Frequenz der gewählten Merkmale festzustellen. Auf dieser Grundlage kann man eine vergleichende Matrix der normalisierten Frequenzen jedes Merkmals in allen Texten erstellen (Abb. 1: Frequenz-Matrix). Man kann dieser Matrix gewissermaßen das stilometrische Profil jedes Textes entnehmen; dieses Profil kann man sich als Vektor in einem multi-dimensionalen Raum vorstellen, in dem jedes Merkmal eine Dimension und jede Frequenz ein

Wert für diese Dimension ist. Jeder Text zeichnet einen anderen Vektor durch diesen multi-dimensionalen Raum. Es ist aus der Tabelle so beispielweise erkennbar, dass die relativen Frequenzen von "de" sich bei Malet und Simenon unterscheiden (rund 4.1 in beiden Malet-Texten, und rund 3.5 in beiden Simenon-Texten), während sie sich für "la" kaum systematisch unterscheiden.

	A	B	C	D	E
1		Malet_JonnyMetal	Malet_120RueDeLaGare	Simenon_GrandePerche	Simenon_MaigretLognon
2	de	4.11874621680659	4.07162171531581	3.45314106703016	3.56895279411436
3	il	1.48169592336237	2.23947861055259	2.63768323887414	2.83324761287149
4	la	2.27123193936363	2.17187825001733	2.40811153358682	2.13555759296943
5	le	2.23438692528357	2.04014421410248	1.88200970897004	1.97007983183882
6	et	1.90804537200305	1.70734243915968	1.49939020015783	1.58098347458575
7	à	1.65539384688265	1.74547597587187	1.73374464930531	1.93877322838167
8	je	2.36334447456378	2.16494487970603	1.49699882822775	1.22542990675105
9	un	1.84751427744296	1.79574291062886	1.59743644929096	1.68161184284085
10	l	1.6238124062426	1.61027525480136	1.29612358610135	1.52060645363268
11	pas	1.05534647472169	1.21160646190113	1.62135016859172	1.47364654844697
12	que	1.21062189120194	1.21853983221244	1.50417294401798	1.34842013461839
13	en	1.19219938416191	1.29654024821466	1.21242556854868	1.20754041906125
14	vous	1.38958338816222	1.43867433959648	2.03984025635507	1.56533017285717
15	d	1.16061794352186	1.13880607363239	1.19329459310807	0.930253359869407
16	les	1.05534647472169	0.793870900644803	1.0187244422125	1.17847000156533
17	est	0.763218148801221	0.951605075227068	1.4515627615563	1.18741474541023
18	une	1.10798220912177	1.10760590723151	1.17416361766746	1.00851986851227
19	ne	0.794799589441272	0.93253830687097	1.06176913695387	1.01075605447349
20	qu	0.692159907361107	0.650003466685156	0.985245235191429	1.14269102618574
21	n	0.715845987841145	0.760937391666089	1.04263816151326	0.91683624410206
22	elle	0.521093770560834	0.379602024544131	1.22438242819906	0.400277287059192
23	dans	0.86848961760139	0.684670318241697	0.954157400100438	0.809499317963282
24	était	0.713214201121141	0.800804270956112	0.839371547456776	0.93696191775308
25	a	0.528989130720846	0.650003466685156	1.18612047731784	1.12256535253472
26	des	0.589520225280943	0.506136032725508	0.726977066743191	0.925780987946958
27	ce	0.821117456641314	0.757470706510435	0.700671975512351	0.684272904134708
28	qui	0.684264547201095	0.61186992997296	0.789152736925174	0.914600058140835
29	avait	0.455299102560728	0.506136032725508	0.817849200086089	0.800554574118384
30	s	0.552675211200884	0.539069541704222	0.676758256211589	0.722288065475525
31	c	0.484248756480775	0.551202939749012	0.698280603582275	0.612714953375523
32	du	0.810590309761297	0.617069957706441	0.561972403567927	0.570227420112257

Abb. 1: Frequenz-Matrix am Beispiel von Romanen Léo Malets und Georges Simenons

Der zweite Aspekt der stilometrischen Methode ist es, die stilometrischen Profile oder Vektoren aller Texte so miteinander zu vergleichen, dass man ein Maß der relativen Nähe oder Distanz der einzelnen Texte zueinander bekommt. Ein wichtiger Parameter, fast der wichtigste von allen, ist die Länge der Wortliste bzw. des Vektors, die berücksichtigt wird. In der Pionierzeit der Stilometrie wurden in der Regel nur die 30–50 häufigsten Wörter berücksichtigt, also reine Funktionswörter ohne semantischen Inhalt. Mittlerweile gibt es hierzu verschiedenste Vorschläge, die bis zur Berücksichtigung der gesamten Wortliste reichen. Die Unterstützung durch den Computer bedeutet jedenfalls, dass die Entscheidung über die Länge der Wortliste nicht mehr von arbeitsökonomischen Faktoren abhängt. Außerdem gibt es systematische Untersuchungen zur Qualität der resultierenden Klassifikationen in Abhängigkeit von der Länge der Wortlisten, und das für unterschiedliche Sprachen und Textsorten (Rybicki & Eder 2011). Diese zeigen vor allem, dass es

keine universelle Lösung für diesen Parameter gibt, sondern dass die angemessene Länge der Wortliste von Sprachen und Gattungen abhängt und stark variieren kann.

	A	B	C	D	E
1		Malet_JohnnyMetal	Malet_120RueDeLaGare	Simenon_GrandePerche	Simenon_MaigretLognon
2	Malet_JohnnyMetal	0	456.281391084859	715.371430641871	692.032721739592
3	Malet_120RueDeLaGare	456.281391084859	0	664.664891571222	666.833849835716
4	Simenon_GrandePerche	715.371430641871	664.664891571222	0	505.818228388631
5	Simenon_MaigretLognon	692.032721739592	666.833849835716	505.818228388631	0
6					

Abb. 2: Distanz-Matrix am Beispiel von vier Romanen Léo Malets und George Simenons

Entscheidend ist dann, welches mathematische Modell für die Berechnung der Ähnlichkeiten zur Anwendung kommt, d.h. welches Distanz-Maß verwendet wird, um die Vektoren der verschiedenen Texte zu vergleichen. Ergebnis dieser Berechnungen ist die Distanz-Matrix, in der für jeden Text ein Distanz-Wert zu allen anderen Texten festgehalten ist (Abb. 2: Distanz-Matrix). Hier ist deutlich erkennbar, dass die Distanzwerte für zwei Texte desselben Autors wesentlich niedriger liegen als für zwei Texte unterschiedlicher Autoren.

Unterschiedliche Distanz-Maße geben bestimmten Teilen der Wortliste ein je unterschiedliches Gewicht in der Ermittlung der Distanz-Werte: Bei der euklidischen Distanz beispielsweise haben die höchstfrequenten Wörter besonders viel Einfluss, während der Einfluss dann stark abfällt; speziell für die Stilometrie entwickelte Distanz-Maße wie 'Burrows' Delta' (Burrows 2002) oder 'Eder's Delta' (cf. Eder & Rybicki 2011) definieren dagegen keinen, respektive einen sanfteren Abfall der Gewichtung (siehe hierzu u.a. Schöch 2012). Für französische Texte gilt, dass 'Eder's Delta' recht eindeutig die besten Resultate ergibt; andererseits ist die Frage der besten Länge der Wortliste für verschiedene Gattungen hier recht offen.

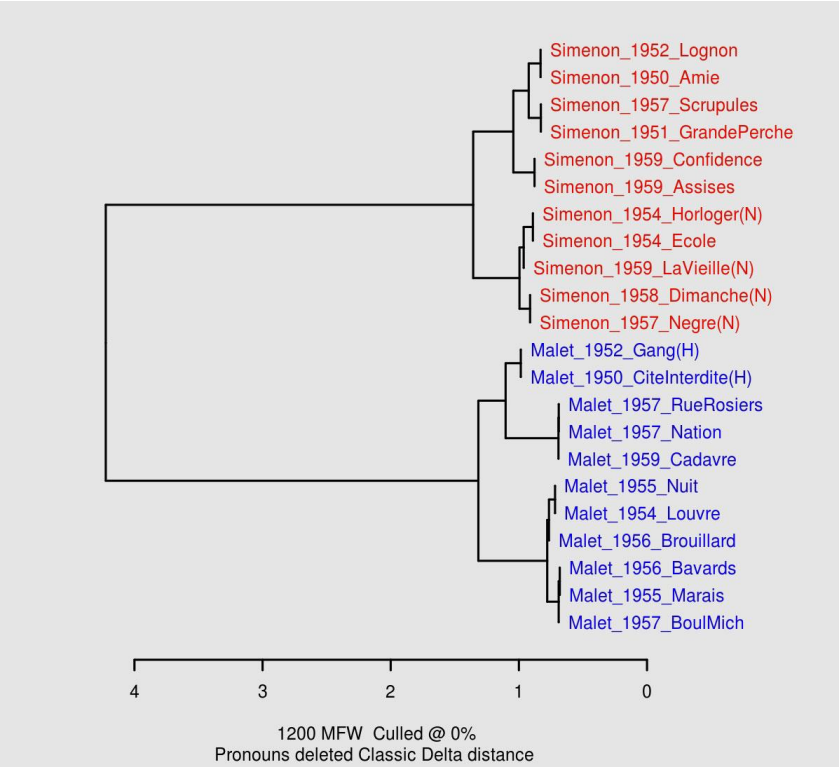

Abb. 3: Cluster Analysis Graph;
Romane von Léo Malet und Georges Simenon aus den 1950er Jahren

Auf der Grundlage der Distanzmatrix können dann die Ähnlichkeitsbeziehungen der Texte zueinander ermittelt und visualisiert werden. Auch hierfür gibt es verschiedene Methoden. Eine grundlegende und relativ transparente Visualisierung beruht auf der "Cluster Analysis" und der Visualisierung als Dendrogramm (Abb. 3: Cluster Analysis Graph). In dem am weitesten verbreiteten, hierarchisch vorgehenden Ward-Verfahren werden die Texte in "nearest neighbor"-Paare gruppiert, also jeder Text zunächst mit dem einen ihm ähnlichsten anderen Text; diese Paare werden wiederum zu größeren Gruppen zusammengefasst. In dem resultierenden Dendrogramm drückt sich die Distanz zwischen zwei Texten in ihrer Entfernung auf der horizontalen Achse aus. Eine stärker synthetisierende Visualisierung beruht auf der Technik des "bootstrapping" (Abb. 4: Bootstrap Consensus Tree). Hier werden mehrere Dendrogramme gewissermaßen miteinander verglichen und nur die "nächsten Nachbarn", die ein bestimmtes Maß der Übereinstimmung in allen Dendrogrammen zeigen, werden berücksichtigt. Dadurch ist diese Visualisierung verlässlicher, allerdings verliert man auch Informationen, insofern die jeweilige Distanz zwischen Texten, die im Dendrogram abgebildet ist, hier nicht mehr ersichtlich ist. Eine weitere Technik der Distanz-Berechnung und Visualisie-

rung ist die "Principal Component Analysis", die auf dem Prinzip der Reduktion des vieldimensionalen Raumes auf mehrere unabhängige Dimensionen beruht und auf die weiter unten noch eingegangen wird. Gemein ist allen diesen Methoden, dass sie statistische Verfahren der Komplexitätsreduktion anwenden, die Gruppierungen und Muster ermitteln oder erkennbar machen.

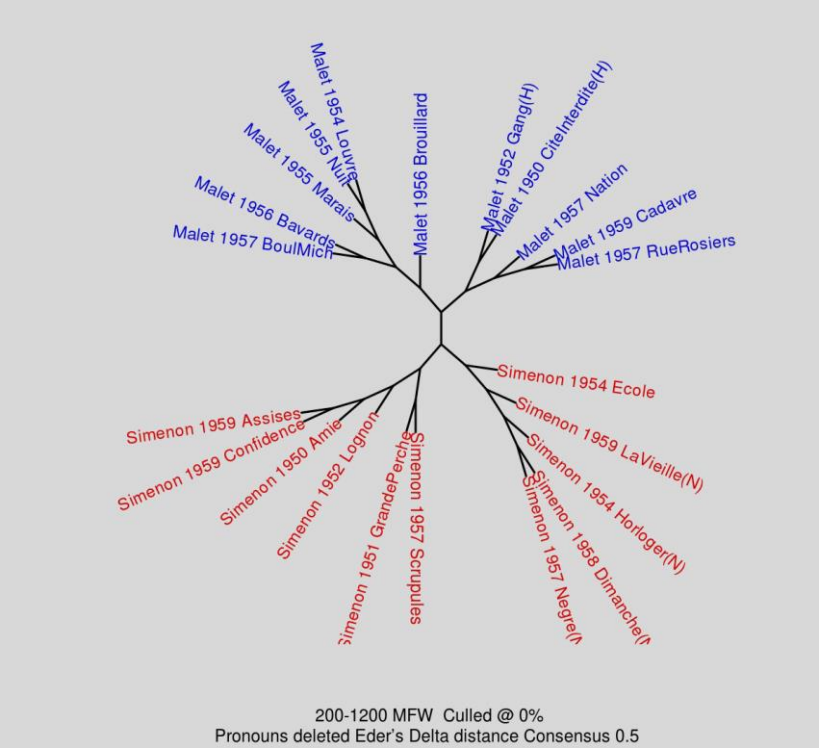

Abb. 4: Bootstrap Consensus Tree;
Romane von Léo Malet und Georges Simenon aus den 1950er Jahren

So streng mathematisch-empirisch das Verfahren selbst ist, so sehr sind die resultierenden Graphen interpretationsbedürftige Gebilde. In manchen Fällen sind sie sehr abhängig von den jeweils gewählten Parametern, in anderen Fällen sind sie erstaunlich stabil. Manchmal leuchten die Graphen sofort ein, weil sie schon bekannte Klassifizierungen bestätigen. In anderen Fällen weichen sie von etablierten Klassifizierungen ab oder werfen ganz neue Fragen auf. Das kann sich auf die Texte selbst beziehen, wie im Falle der bereits genannten Kriminalromane: Warum wird dieser eine Maigret-Roman von Georges Simenon (*Maigret à l'École* in Abb. 3 und 4) mit einer Gruppe von ansonsten eindeutig unterschiedenen *romans durs* desselben Autors klassifiziert? Oder eben auch auf die Methode: Welche Merkmale, welche Parameter, sind für diese Klassifikation verantwortlich? Oder überwiegt einfach der Einfluss chronologischer Nähe im Werk Simenons über die

verschiedenen Roman-typen? Ohne Zweifel erfordern solche Fragen weiterhin unsere ganze literaturwissenschaftliche Kompetenz.

2 Autorschaft und Gattungszugehörigkeit im Theater der französischen Klassik

Ausgangspunkt für die nun folgenden Anwendungsbeispiele ist die mittlerweile berühmte Corneille / Molière-Kontroverse, die einen faszinierenden Testfall für stilometrische Verfahren darstellt. Im Zentrum der Kontroverse steht die Frage, ob Corneille einige oder sehr viele der traditionell Molière zugesprochenen Werke verfasst hat.

2.1 Die Corneille / Molière-Kontroverse

Der französische Dichter Pierre Louÿs behauptete 1919 in einem kurzen Artikel, dass Corneille eindeutig der Autor des *Amphytrion* sei, ein Stück, das üblicherweise Molière zugeschrieben wird (Louÿs o.J.). In einem weiteren Artikel behauptete der Dichter ferner, dass Molière überhaupt keines der Stücke, die ihm normalerweise zugerechnet werden, geschrieben habe, sondern dass Corneille während Molières gesamter Karriere dessen *ghost writer* gewesen sei. Pierre Louÿs stützt seine These einerseits auf die Chronologie – die Karriere des Autors Molière beginnt beispielsweise ausgerechnet im Jahr 1658, als er sich längere Zeit bei Corneille in Rouen aufhält – und auf die Lektüre der Texte, wo der Dichter meint, in den Versen Molières das Vokabular und den Stil Corneilles identifizieren zu können. Die Debatte erhielt neuen Schwung und bald auch erstaunliche Polemik, als Dominique und Cyril Labbé, zwei Spezialisten für Lexikometrie und Statistik, im Jahr 2001 ihre stilometrischen Analysen der Werke Corneilles und Molières veröffentlichten, unterstützt von detailreichen biographischen Nachforschungen durch Denis Boissier (2004).[8]

Labbé und Labbé (2001) kommen zu einigen überraschenden Ergebnissen. Erstens behaupten sie, dass das Stück *Dom Garcie*, das normalerweise Molière zugeschrieben wird, den Stücken Corneilles viel ähnlicher sei als anderen Stücken von Molière und also Corneille zugeschrieben werden müsse. Zweitens behaupten sie,

[8] Einen guten Einstiegspunkt in die Debatte, die u.a. in Form von gegensätzlich positionierten Webseiten ausgetragen wurde, bietet *Fabula.org* (Atelier de théorie littéraire 2012). Die Debatte erreichte die Aufmerksamkeit selbst der *New York Times* (Zanganeh 2003).

dass die beiden Stücke *Le Menteur* und *La suite du Menteur*, die Corneille zugeschrieben werden, einigen Stücken von Molière sehr ähnlich seien; da es aber keine Zweifel an Corneilles Autorschaft der beiden Stücke gibt, müssten auch die anderen Stücke, die normalerweise Molière zugeschrieben werden, von Corneille sein. Diese Ergebnisse haben deutliche Kritik hervorgerufen, einerseits eher lautstark und polemisch vom akademischen Establishment, insbesondere von etablierten Molière-Spezialisten, andererseits aber durchaus auch nachdrücklich von erfahrenen Stilometrie-Experten, allen voran Étienne Brunet, der zahlreiche Details der spezifischen Anwendung der Methode kritisiert, um die Methode selbst zu verteidigen.[9]

In der Tat weist die Methode von Labbé und Labbé einige Besonderheiten auf. Die Autoren verwenden zur Berechnung der Distanzen das euklidische Distanz-Maß, bei dem nur wenige hochfrequente Merkmale ein sehr großes Gewicht haben und die Gewichtung dann sehr stark abfällt. Dieses Distanz-Maß ist in der Stilometrie mittlerweile eher unüblich, weil es sich nicht bewährt hat. Außerdem verwenden die Autoren ein lemmatisiertes Korpus, in dem nicht die *token*-Frequenzen verwendet werden, sondern die *type*-Frequenzen. Dadurch sinkt im Grunde die verfügbare Informationsmenge, denn zahlreiche stilistische Informationen werden nivelliert. Schließlich verwenden die Autoren die gesamte Wortliste, und nicht nur einen mehr oder weniger umfangreichen Teil der Wortliste. Dies kann unter Umständen, wie Eder und Rybicki (2011) gezeigt haben, insbesondere bei französischen Texten, zu einer Verschlechterung der Ergebnisse führen. Allerdings wird dieser Effekt durch die Verwendung der euklidischen Distanz relativiert. Die Kombination einer vollständigen Wortliste mit der euklidischen Distanz leuchtet allerdings nicht ein, da sich beide Entscheidungen gegenseitig relativieren. Zudem unterschätzen Labbé und Labbé einerseits die besondere und sehr konventionalisierte Textform des klassischen französischen Theaters, die im Kontext von Sprachnormierung und "doctrine classique" allen Autoren einen gemeinsamen und relativ engen stilistischen und thematischen Rahmen vorgibt (Génétiot 2005) und stilometrische Klassifikationen somit erschwert. Sie blenden anderer-

[9] Siehe Brunet 2004. Ein neuerer, linguistisch geprägter Ansatz zur Corneille / Molière-Kontroverse ist Marusenko & Rodionova 2010. Ein unmittelbar von den Labbé'schen Arbeiten angeregter Beitrag zu Shakespeare und Middleton ist Merriam 2003.

seits Faktoren wie die Unterscheidung von Vers- und Prosaform, die für systematische stilistische Variation verantwortlich zu sein scheinen, weitgehend aus.

Wie nun unter anderem zu zeigen sein wird, liegt eine aktuelle Herausforderung der Stilometrie darin, das Verhältnis der Signale von Autorschaft, Gattung und Form zueinander zu verstehen und Methoden zu erproben oder zu entwickeln, um diese Signale zu unterscheiden (siehe auch Jockers 2013). Genau dieses Problem scheint entscheidend für die stilometrische Entscheidung über die Autorschaft von Corneille und Molière. Denn während Corneille Tragödien und Komödien geschrieben hat, ist Molière ein reiner Komödienautor. Und während Corneille nur Stücke in Versform geschrieben hat, gibt es von Molière Komödien in Vers- und in Prosaform. Vergleicht man nur die Verskomödien, kann man sich zwar auf ein relativ homogenes Korpus stützen, erkauft diesen Vorteil aber damit, dass ein gewisser Teil der relevanten Texte unberücksichtigt bleibt. Dadurch verringert sich die Verlässlichkeit der Methode, was insbesondere bei den im Vergleich zu Romanen kurzen Theaterstücken problematisch sein kann. Berücksichtigt man alle vorhandenen Texte der beiden Autoren, erhöht sich zwar die Textmenge und der Abdeckungsgrad, zugleich wird aber die Textsammlung heterogener, was schwer abzuschätzende Auswirkungen auf die Textklassifikation haben kann.

Die Schwierigkeit liegt hier darin, dass noch nicht ausreichende Erfahrungswerte für diverse stilometrische Verfahren im Bereich des Französischen im Allgemeinen, und des klassischen französischen Dramas in Vers und Prosa im Besonderen, vorliegen. Das heißt, man kann nicht auf etablierte Verfahren und Parameter zurückgreifen, die verlässliche Ergebnisse garantieren würden. Davon sind alle Parameter der stilometrischen Analyse betroffen, sowohl die Vorbereitung der Texte, die Auswahl der relevanten linguistischen Merkmale, die Anwendung bestimmter Distanzmaße und die Klassifikationsmethoden.

2.2 Korpusbasierte Parameter-Optimierung

Ein erster Ansatz zur Lösung dieses Problems ist eine Vorgehensweise, die man korpusbasierte Parameter-Optimierung nennen könnte. Dabei wird die zur Debatte stehende Textsammlung in zwei Teile aufgeteilt: Einerseits ein Hauptkorpus mit unstrittigen Texten mindestens zweier Autoren, andererseits eine kleine Anzahl strittiger Texte, bei denen unklar ist, welcher der Autoren des Hauptkorpus sie

geschrieben hat. Dann wird durch systematische Variation relevanter Parameter diejenige Kombination von Parametern gesucht, mit denen die unstrittigen Texte am besten unterschieden werden. Anschließend werden die strittigen Texte jeweils einzeln wieder in die Untersuchung eingebracht, was eine relativ zuverlässige Klassifikation erlauben sollte. Allerdings können die Distanzmaße und die Klassifikationsmethoden unter Umständen enorm sensibel sein und es ist nicht auszuschließen, dass das Hinzufügen einzelner Texte zur ursprünglichen Textsammlung dann nicht-optimale Ergebnisse produziert. Generell ist die Verlässlichkeit stilometrischer Verfahren schwer einzuschätzen.

Ein Beispiel mit Texten von Molière, Pierre Corneille und weiteren Autoren der französischen Klassik erlaubt es, das Verfahren zu illustrieren. Alle 54 verwendeten Komödientexte stammen aus dem ausgezeichneten Textarchiv *Théâtre classique* von Paul Fièvre (2007–2013), das die Texte unter anderem in einem XML-Format anbietet, das den Empfehlungen der *Text Encoding Initiative* (Burnard & Bauman 2007) folgt. Die in diesem Format vorliegenden strukturellen Auszeichnungen erlauben es, die Texte automatisch so vorzubereiten, dass nur die Figurenrede berücksichtigt wird. Dem Dramentext im engeren Sinne nicht zugehörige Passagen wie Vorworte und Anmerkungen sind automatisch entfernt worden, aber auch Szenenanweisungen und insbesondere Sprechernamen, die durch ihre hohe Frequenz und Spezifizität möglicherweise die Klassifikation ungebührlich vereinfacht hätten.

Die verwendete stilometrische Anwendung ist ein Paket für die Statistikumgebung R, das Maciej Eder und Jan Rybicki entwickelt haben (Rybicki & Eder 2011). Die hier verwendete Methode beruht auf Wortfrequenzen, das verwendete Distanzmaß ist 'Eder's Delta' und das Ergebnis wird mit einem Cluster Analysis Graph visualisiert. Die Parameter-Optimierung bezieht sich vor allem auf die Länge der berücksichtigten Wortliste. Für die hier beschriebene Untersuchung wurden die drei Texte, für die Labbé und Labbé neue Attributionen vorgeschlagen haben, als strittig eingestuft und zunächst vom Hauptkorpus entfernt; das sind *Dom Garcie* von Molière und *Le Menteur* und *La Suite du Menteur* von Corneille. Alle anderen Komödien von Corneille und Molière sowie die Komödien einiger weiterer Autoren der gleichen Zeit bilden das Hauptkorpus.

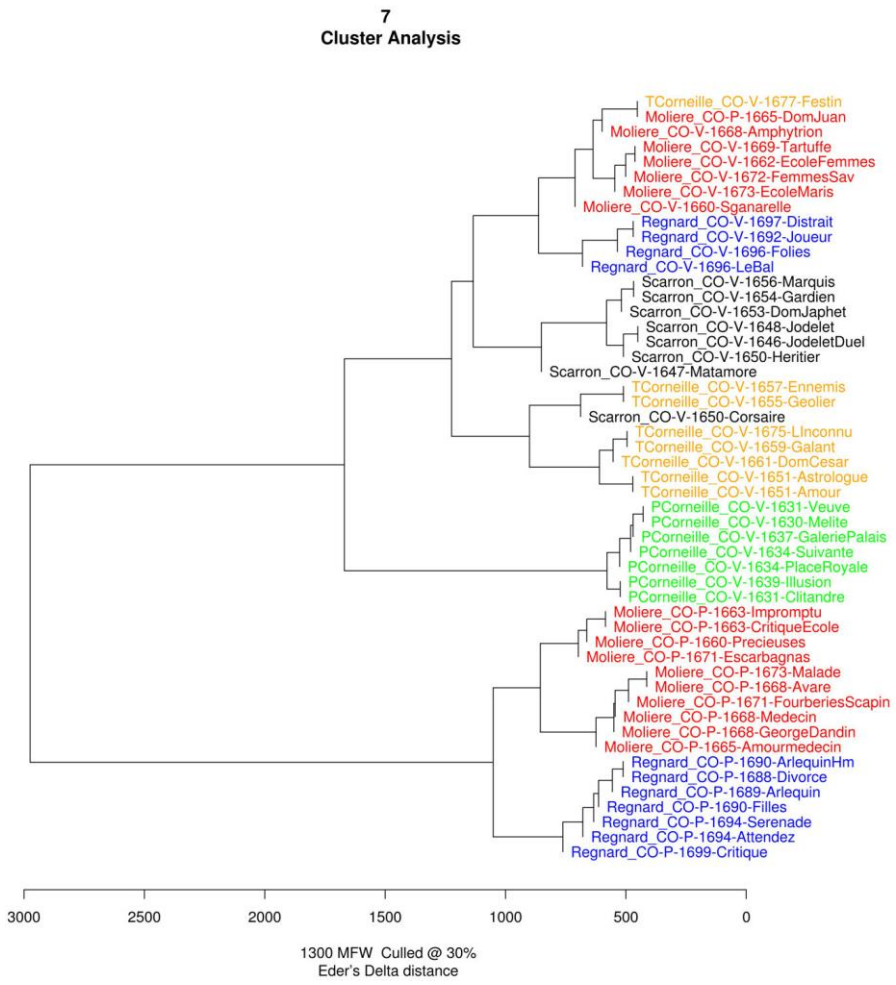

Abb. 5: Cluster Analysis Graph für 51 Dramentexte (1300 Wörter, 30% Culling)

Für das Hauptkorpus werden nun mehrere Durchläufe mit verschiedenen Parametern gemacht und die resultierenden Cluster Analysis Graphen daraufhin untersucht, wie gut sie die unstrittigen Texte nach Autor klassifizieren. Wenn man besondere Parameter zunächst unberücksichtigt lässt, ergibt sich folgende Konstellation: Bis zu einer Länge der Wortliste von 1100 Wörtern werden vier Stücke (zwei von Pierre Corneille und zwei von Scarron) nicht korrekt klassifiziert, alle anderen Stücke werden aber korrekt klassifiziert. Bei einer Länge der Wortliste von 1200 oder 1300 Wörtern werden auch die beiden letzten Corneille-Stücke korrekt klassifiziert, die beiden Scarron-Stücke werden aber nach wie vor als den Stücken anderer Autoren besonders ähnlich klassifiziert. Außerdem ist ein neuer "Fehler" aufgetreten, denn eine Verskomödie von Thomas Corneille erscheint mitten in den Prosakomödien von Molière. Verwendet man bei einer Länge der Wortliste von 1300 außerdem nur die Wörter, die in mindestens 30% der Texte einmal vorkommen, verbessert sich die Klassifikation noch minimal: ein

weiteres Scarron-Stück (*Les Boutades du Capitaine Matamore*) wird an der richtigen Stelle eingeordnet, so dass nur ein Scarron-Stück und ein Stück von Thomas Corneille mit den Stücken anderer Autoren gruppiert werden (Thomas Corneilles *Le Festin de Pierre* mit Molière, Scarrons *Le Prince corsaire* mit Thomas Corneille; vgl. Abb. 5: Cluster Analysis Graph für 51 Dramentexte). Die Manipulation weiterer Parameter, wie beispielsweise die Löschung der Pronomina aus der Wortliste, bringt dagegen keine Verbesserung mehr.

Man kann bis hierhin mehrere Ergebnisse festhalten. Erstens kann das Verfahren für bekannte Texte durch die Wahl der Parameter erstaunlich weit optimiert werden, denn von 51 Texten wurden 49 korrekt klassifiziert, was einer Erfolgsrate von 96 Prozent entspricht. Zweitens zeigt sich in den Graphen, dass die Unterscheidung von Prosa und Vers offensichtlich eine massiv differenzierende Kategorie ist, die die Autorenkategorie klar überlagert. Untersuchungen jenseits dieser Grenze scheinen problematisch. Drittens gibt es zwei Texte, nämlich die beiden Verskomödien von Scarron, die sich nicht korrekt klassifizieren lassen, egal mit welchen Parametern. Das kann an den Texten selbst liegen (eventuell zeichnen sich diese Stücke durch einen für Scarron ungewöhnlichen Stil oder ungewöhnliche Themen aus, und / oder durch für einen anderen Autor typischen Stil oder Themen), oder sie markieren aufgrund anderer Eigenschaften einen Punkt, an dem die stilometrische Methode keine korrekte Klassifikation ermöglicht.

Mit diesen Parametern, die insbesondere die beiden strittigen Autoren Pierre Corneille und Molière zuverlässig voneinander trennen, kann man nun die strittigen Texte wieder in das untersuchte Korpus hineinnehmen, und zwar jeweils einzeln, um das Gleichgewicht des Korpus möglichst wenig zu stören. Das Ergebnis ist durchaus überraschend, denn alle drei Stücke zeigen eine starke Affinität mit den Verskomödien von Thomas Corneille, nicht aber mit denen von Pierre Corneille. Entweder sind diese drei strittigen Texte tatsächlich weder von Pierre Corneille noch von Molière, sondern von Thomas Corneille geschrieben worden, oder das Genre- und Formsignal sind hier so dominant, dass das Autorsignal sich nicht manifestiert. Schließlich ist denkbar, dass die drei strittigen Stücke den Stücken von Thomas Corneille thematisch so stark verwandt sind, dass dieses eigentlich unerwünschte Signal das Autoren-Signal überlagert.

Um einmal bei *Dom Garcie* als Beispiel zu bleiben: das Stück ist im Grunde eine *comédie héroïque*, wie Thomas Corneille einige geschrieben hat, weist also eine

gattungs- und themenbezogene Nähe zu diesem Autor auf. Diesem Erklärungsansatz kann man mit den *Voyant-Tools* (Sinclair & Rockwell 2013) nachgehen (Abb. 6: Vergleich dreier Gesamtwerke mit *Voyant Tools*). Hierbei wird das Gesamtwerk der drei Autoren jeweils als Teilkorpus aufgefasst, woraufhin für jedes Teilkorpus die sogenannten 'distinktiven' Wörter extrahiert werden, also diejenigen Wörter, die im Vergleich zum Gesamtkorpus besonders häufig sind. Dieses Verfahren zeigt, dass die zehn Wörter mit der höchsten absoluten Frequenz in *Dom Garcie* (*cœur*, *point*, *trop*, *âme*, *madame*, *prince*, *voir*, *yeux*, *peut*, *doit*; Funktionswörter ausgeschlossen) recht stark mit den distinktiven Wörtern in Thomas Corneilles Werk übereinstimmen (*madame*, *monsieur*, *voir*, *inconnu*, *prince*: drei Übereinstimmungen) aber nur schwach mit denen in Molières Werk (*oui*, *femme*, *point*, *faire*, *monsieur*: eine Übereinstimmung).

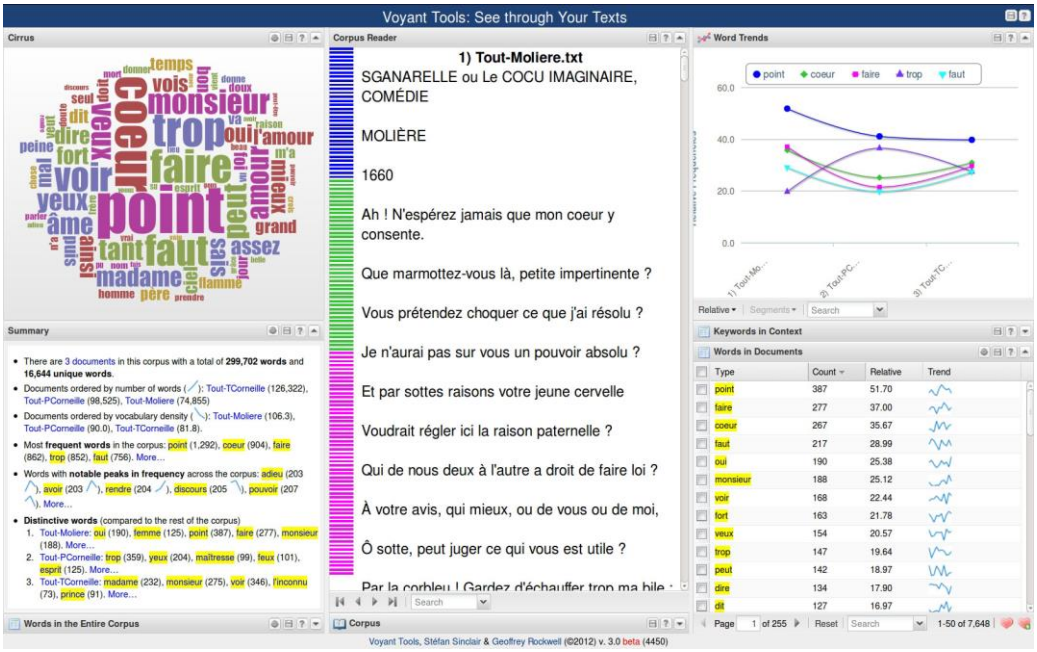

Abb. 6: Vergleich dreier Gesamtwerke mit *Voyant Tools*

Durch diesen sehr oberflächlichen Vergleich der Texte auf thematische Ähnlichkeit hinaus, der nur ein Ausgangspunkt für eingehendere thematische Vergleiche sein kann, sieht man aber auch: Die bisher verfolgte stilometrische Clustering-Methode etabliert immer nur den am wenigsten unterschiedlichen Kandidaten als 'nearest neighbor'. Damit ist aber nicht gesagt, dass die strittigen Texte nicht einem weiteren Autor, der im untersuchten Korpus nicht vertreten ist, eigentlich noch ähnlicher sind (für eine Lösung dieses Problems siehe Eder 2013). Die Frage

der konkurrierenden Signale von Autor, Gattung und Form ist damit aber noch nicht gelöst.

2.3 Trennung der Signale von Autor, Genre und Form

Der zweite Ansatz zur Lösung der Corneille / Molière-Kontroverse liegt daher in dem Versuch, die Signale von Autor, Genre und Form voneinander zu trennen. Würde dies gelingen, könnte man kontrolliert nur eines dieser Signale, in diesem Fall das des Autors, für eine erneute Klassifikation nutzen, in der sich dann Genre und Form nicht "störend" auswirken würden. Es gibt für dieses Problem noch keine etablierte Lösung; einen eher einfachen Lösungsansatz möchte ich hier schildern und an einem Beispiel illustrieren (siehe hierzu auch Schöch 2013b).

Die Ausgangshypothese für diese Experimente ist die, dass bestimmte Teile der Wortfrequenzliste tendenziell für bestimmte Signale verantwortlich sind. Man kann beispielsweise annehmen, dass die Kategorie der Gattung stärker als die Kategorie der Autorschaft mit semantisch-inhaltlichen Wörtern assoziiert ist und dass umgekehrt Autorschaft vor allem mit rein funktionalen Wörtern assoziiert ist, die aufgrund ihrer großen Häufigkeit im allerersten Teil der Wortfrequenzliste dominieren. Es handelt sich hierbei allerdings um vereinfachende Annahmen, die beispielsweise auch unterschlagen, dass die Frequenzen einzelner Wörter in einer Textsammlung nicht unabhängig voneinander sind; zudem ist die Stärke der stilometrischen Verfahren ja gerade, die Werte sehr vieler Einzelmerkmale kumuliert zu betrachten.

Der hier verfolgte Ansatz verfährt daher in mehreren Schritten. Zunächst einmal wird eine Textsammlung definiert, die relativ homogen ist, und deren Texte sich nur in zwei entscheidenden Faktoren unterscheiden: in diesem Fall handelt es sich um französische Theaterstücke, die alle aus dem siebzehnten Jahrhundert stammen, aus der Zeit zwischen 1631 und 1677, und die alle in Versform geschrieben sind. Die Stücke unterscheiden sich in ihren Autoren – Pierre Corneille und Thomas Corneille – und nach ihrer Untergattung, denn es handelt sich um Komödien und Tragödien. Die Sammlung enthält eine ausgeglichene Anzahl von Komödien und Tragödien der beiden Autoren: insgesamt 32 Stücke, je 8 Komödien und 8 Tragödien von Pierre Corneille und von Thomas Corneille.

Für diese Textsammlung wird eine Wortfrequenzliste erstellt und die Texte werden mehrfach, jeweils auf der Grundlage eines anderen, kleinen Teils der Wortfrequenzliste klassifiziert. Die Klassifikation zielt darauf ab, jeweils 'nächste Nachbarn' zu identifizieren, d.h. jeweils Paare von Stücken, die den niedrigsten Wert stilometrischer Differenz aufweisen. Diese Paare werden dann daraufhin untersucht, ob sie sich nach den Kategorien Autor und Gattung gleichen oder unterscheiden. Für jedes Paar sind vier Konstellationen möglich: (1) Autor und Gattung sind gleich, ein Paar besteht also beispielsweise aus zwei Komödien von Pierre Corneille; (2) Nur der Autor ist gleich, d.h. zwei Stücke von Pierre Corneille oder Thomas Corneille, aber eine Komödie und eine Tragödie; (3) Nur die Gattung ist gleich, also zwei Komödien, aber eine von Pierre Corneille, eine von Thomas Corneille; (4) Weder Autor noch Gattung stimmen überein, das Paar besteht also beispielsweise aus einer Komödie von Pierre Corneille und einer Tragödie von Thomas Corneille.

Wenn man nun die jeweilige Anzahl dieser vier Arten von 'nächsten Nachbarn' für jeden Durchgang mit einem anderen Teil der Wortfrequenzliste ermittelt und visualisiert, erhält man ein Diagramm, das die relativen Einflüsse der Signale von Autor und Gattung aufzeigen könnte. Es sind zwei Ergebnisse denkbar: entweder gibt es einen klaren Trend über das Wortfrequenzspektrum hinweg sowie einen deutlichen Grenzbereich oder es gibt eine sehr viel komplexere Gemengelage ohne klare Trends.

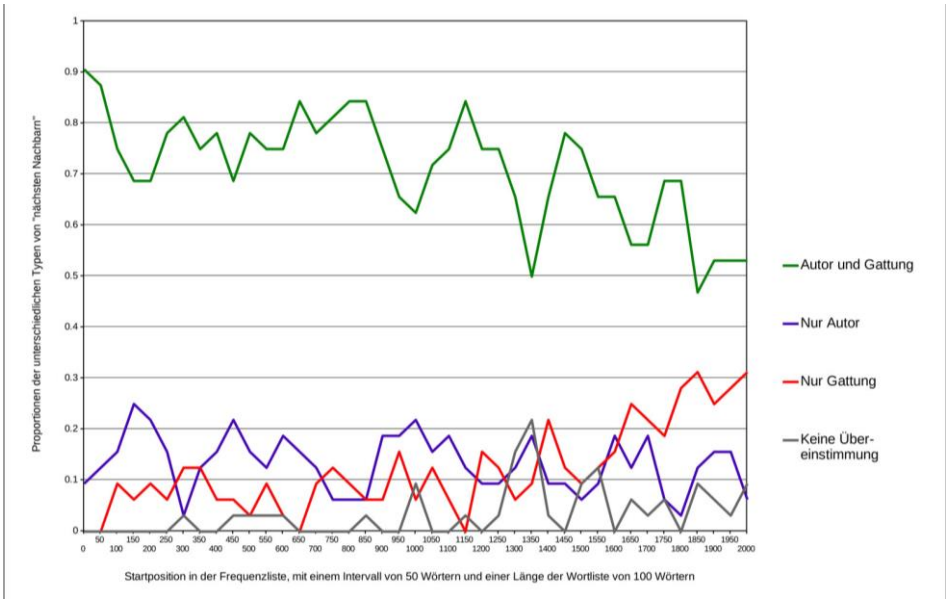

Abb. 7: Autor- und Gattungssignal für 32 Theaterstücke

Für das gerade beschriebene Beispiel ergibt sich das oben stehende Diagramm, wenn man jeweils einen Abschnitt der Frequenzliste berücksichtigt, der 100 Wörter lang ist, und in 50-er Schritten durch die Frequenzliste geht, und das vom Beginn der Wortliste bis zum zweitausendsten Wort (Abb. 7: Autor- und Gattungssignal für 32 Theaterstücke). Wie man sieht, gibt es nur eine relativ klare Tendenz: Der Anteil der 'nächsten Nachbarn', die sowohl nach Autor als auch nach Gattung korrekt klassifiziert wurden (grüne Linie), nimmt nach und nach ab. Eine möglicherweise sichtbare Tendenz ist, dass die nächsten Nachbarn, die nur dem Autor nach korrekt klassifiziert wurden (blaue Linie), bis zur etwa 1150. Position leicht gegenüber den Gattungs-Nachbarn (rote Linie) überwiegen und sich danach eine unklare Gemenge-Lage zeigt. Addiert man jeweils die Werte der "Autor+Gattung"-Paare mit denen der "nur Autor"-Paare einerseits, und denen der "nur-Gattung"-Paar andererseits, zeigt sich allerdings, dass die beiden resultierenden Linien so nah beieinander sind, dass letztlich kein signifikanter Trend erkennbar scheint.

Man könnte aus diesen Ergebnissen schließen, dass wir es bei den Kategorien von Autor und Gattung eben tatsächlich nicht mit voneinander unabhängigen Faktoren zu tun haben, bzw. dass zumindest die sprachlichen Signale und Ähnlichkeitsrelationen, die sich mit dieser konkreten stilometrischen Methode ermitteln lassen, keine Differenzierung nach Autor und Gattung zulassen. Die entscheidende Schwäche des eben vorgestellten Ansatzes ist, dass er die Annahme trifft, dass sich größere, zusammenhängende Teile der Wortliste mit bestimmten Kategorien wie Autorschaft, Gattungszugehörigkeit oder Form in Korrelation setzen lassen. Obwohl dies nicht der Fall zu sein scheint, bedeutet dies nicht, dass sich die Signale von Autoren, Gattungen und Formen nicht doch unterscheiden lassen.

In der Tat weist ein statistisches Analyseverfahren darauf hin, dass die relativen Frequenzen bestimmter Wörter tatsächlich mit einzelnen Kategorien wie Autorschaft oder Gattungszugehörigkeit verbunden sind und sich voneinander unabhängige Signale aus den Frequenzdaten extrahieren lassen. Dieses statistische Verfahren ist die sogenannte "Principal Component Analysis" (PCA), ein weit verbreitetes komplexitätsreduzierendes Verfahren zur Entdeckung von Trends und Mustern in umfangreichen, unübersichtlichen Daten. Das Grundprinzip von PCA ist es, aus einem hochdimensionalen Datensatz mehrere voneinander unabhängige

Dimensionen zu extrahieren, die verborgene, approximative Korrelationen in den Daten sichtbar machen (vgl. Binongo & Smith 1999). Wendet man eine solche PCA auf die Wortfrequenzvektoren der bereits verwendeten Sammlung von Komödien und Tragödien von Pierre und Thomas Corneille an, ergibt sich ein aufschlussreiches Bild (Abb. 8: Principal Component Analysis).

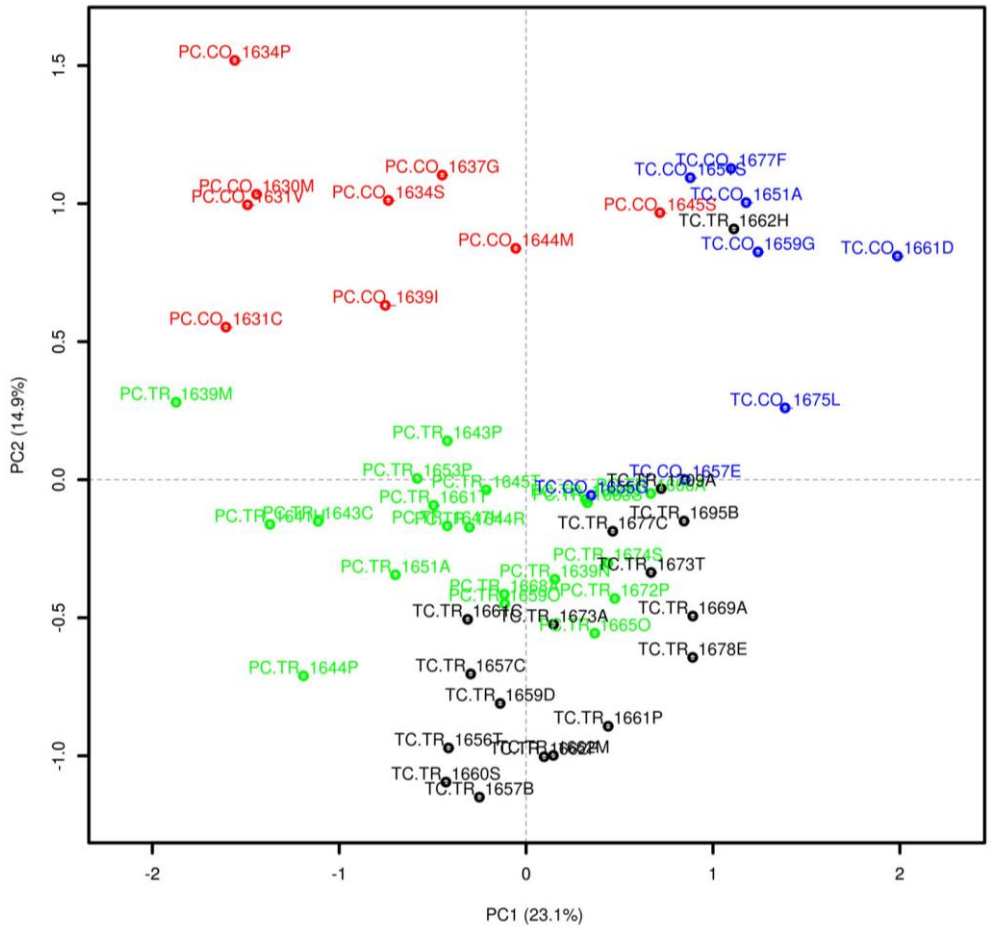

Abb. 8: Principal Component Analysis
(Kürzel: PC = Pierre Corneille, TC = Thomas Corneille, CO = comédies, TR = tragédies)

Das Interessante an der obigen Abbildung ist nun, dass die beiden wesentlichen Komponenten, die auf der horizontalen und vertikalen Achse abgebildet werden, sehr stark mit den Kategorien Autorschaft und Gattungszugehörigkeit korrelieren: Die erste Komponente auf der horizontalen Achse trennt die Stücke tendenziell nach Autoren (negative Werte in der linken Hälfte entsprechen Pierre Corneille, positive Werte in der rechten Hälfte entsprechen Thomas Corneille); die zweite Komponente auf der vertikalen Achse trennt die Stücke recht deutlich nach Gat-

tungen (positive Werte in der oberen Hälfte entsprechen Komödien, negative Werte in der unteren Hälfte entsprechen Tragödien).

Zunächst einmal spricht dies dafür, dass die Signale von Autorschaft und Gattung eben doch zumindest in Teilen unabhängig voneinander sind. Allerdings ist auch festzustellen, dass die Trennung von Autoren im Bereich der Komödien wesentlich deutlicher ist als im Bereich der Tragödien. Dies scheint Ausdruck davon zu sein, dass die Tragödie zumindest in der untersuchten Zeit noch stärker als die Komödie thematischen und stilistischen Konventionen unterworfen ist, die eine Differenzierung von Autorenstilen relativ schwierig macht. Anstatt einen Zusammenhang zwischen größeren zusammenhängenden Abschnitten der Wortliste mit den Signalen von Autoren und Gattung zu postulieren, muss von einem komplexeren Zusammenhang zwischen unterschiedlichen Mengen von breit über die Wortliste verteilten Einzelwörtern ausgegangen werden, die jeweils mit den Zielkategorien korrelieren.

Hier lässt sich vorübergehend nur bilanzieren, dass bisher keine wirkliche Lösung gefunden werden konnte und die hier vorgestellten, bisherigen Versuche noch keine Auflösung der Molière / Corneille-Kontroverse erlauben. Zugleich haben die bisherigen Versuche zu methodischen Fragen geführt, die zur Zeit von mehreren Forschern im Bereich der digitalen Textanalyse untersucht werden, so dass man auf baldige Fortschritte der Diskussion hoffen darf. Man könnte hier Matthew Jockers (2013) sowie Fotis Jannidis und Gerhard Lauer (2013) nennen. Sie interessieren sich aktuell für die Frage, wie und ob man die stilistischen Signale oder Signaturen von Autorschaft und Gattungszugehörigkeit entwirren kann. Unter den relevanten Verfahren für dieses Problem ist das Prinzip der Merkmalsauswahl mit "Zeta", das es erlaubt, gezielt die distinktiven Wörter für Texte aus zwei vorgegebenen, bekannten Kategorien zu ermitteln.[10] Es beruht auf dem Prinzip, distinktive Wörter für zwei Teilkorpora nicht nach relativer Gesamtfrequenz der Wörter, sondern nach ihrer Präsenz in einer mehr oder weniger großen Anzahl der Textabschnitte zu bemessen, in die sich die Teilkorpora gliedern lassen. Clustering-Verfahren auf der Grundlage so strategisch erstellter Wortlisten sind vielversprechend, eventuell allerdings schwer verallgemeinerbar. Weitere vielver-

[10] Das Verfahren wurde von John Burrows (2007) entwickelt, von Hugh Craig (Craig & Kinney 2009) weiterentwickelt und ist in dem "stylo"-Paket für R von Maciej Eder und Jan Rybicki verfügbar.

sprechende Verfahren sind die "unmasking"-Technik, die Mike Kestemont in einigen allerdings nicht ganz identisch gelagerten Fällen erfolgreich eingesetzt hat (Kestemont et al. 2012), sowie im Kontext von überwachten Methoden des maschinellen Lernens das Verfahren des "information gain" (siehe hierzu Witten et al. 2011).

3 Fazit: einige übergeordnete Gesichtspunkte

Neben dieser noch relativ konkreten, auf eine spezifische methodische Fragestellung bezogene Problematik werfen stilometrische Verfahren der Textklassifikation einige grundsätzliche Fragen auf, die im weiteren Kontext der digitalen Erweiterung unseres Methodenrepertoires in den Philologien zu sehen sind. Drei dieser Fragen seien abschließend kurz angerissen.

Erstens die Frage nach der Relevanz stilometrischer Methoden für die Literaturwissenschaft. Für die Gattungstheorie könnte es meiner Meinung nach interessant sein, herauszufinden, welche sprachlichen Merkmale auf der Mikroebene mit der Gattungszugehörigkeit von Texten korrelieren. Während stilistische und thematische Merkmale auf diese Weise erfasst werden können, bleiben andere Kriterien wie dominante Handlungsmuster unberücksichtigt. Ebenso interessant scheint mir, die Ergebnisse computergestützter Klassifikationsverfahren (insbesondere, wenn der Faktor des individuellen Autors ausgeklammert werden könnte), mit etablierten Gattungs- und insbesondere Untergattungs-Klassifikationen zu vergleichen. Hier wird es dann auch für die Literaturgeschichtsschreibung interessant: Gerhard Lauer und Fotis Jannidis haben kürzlich bereits festgestellt, dass es beispielsweise um 1800 eine Gruppe weiblicher Autoren gibt, deren Texte von stilometrischen Verfahren nicht mit einer der etablierten Autorengruppen der Zeit assoziiert werden, sondern eine eigene Gruppe bilden, die bisher von der Literaturgeschichte gar nicht als solche wahrgenommen worden ist.

Zweitens die Frage der Transparenz der Verfahren für uns selbst. Solange wir *close reading* betreiben, scheinen wir die Kontrolle darüber zu behalten, wie wir von einer oder mehreren Beobachtungen am Text zu einer Interpretation des Textes kommen. In Wirklichkeit ist der hermeneutische Akt, der präzise Detailbeobachtungen, komplexe Syntheseleistungen, lange Argumentationsketten und kreative Sinnstiftung zusammenbringt, uns selbst und anderen gegenüber weniger

transparent und nachvollziehbar, als uns vielleicht lieb wäre. Doch während in der digitalen Philologie vieles formalisiert, expliziert und offengelegt wird, weil es dadurch mit Hilfe des Computers manipulierbar wird, entstehen doch zugleich ganz neue blinde Flecken: beispielsweise gilt es, die statistischen Verfahren zu durchschauen, die die nicht mehr übersehbaren Datenmengen zusammenfassen und transformieren; oder es gilt, die Mechanismen der Visualisierung zu verstehen, die ebenfalls Trends und Muster in unseren "Textdaten" aufzeigen. Hier sind allerdings, so meine ich, geisteswissenschaftliche Kernkompetenzen wieder enorm relevant, beispielsweise der Umgang mit Hyperkomplexität, die Fähigkeit zu Kontextualisierung, Syntheseleistung und Sinnstiftung oder die Fähigkeit zu selbstkritischer Reflexion.

Drittens möchte ich die Frage der im Kontext der digitalen Philologien notwendigen Kompetenzen aufwerfen, die hier direkt anschließt. Wir als "digitale Philologen" müssen nicht zugleich auch vollwertige Programmierer und / oder Statistiker werden. Im Gegenteil scheint mir wesentlich, dass wir unsere disziplinären Stärken und Spezialisierungen in einen Forschungsprozess einbringen, der von kleinen und äußerst heterogenen, inhärent transdisziplinären Forschergruppen oder von informell oder auch nur indirekt kooperierenden Forschern vorangetrieben wird. Entscheidend wird allerdings sein, das Kommunizieren über die Fächergrenzen hinweg verstärkt zu lernen und zu praktizieren und den jeweils anderen immer ein Stück weit entgegen zu kommen. Wenn wir als Geisteswissenschaftler lernen, präzisen Pseudo-Code zu schreiben und ein Grundverständnis von Datenformaten und Programmiersprachen besitzen, ist den Informatikern sehr geholfen; wenn Statistiker Visualisierungen entwickeln, die die entscheidenden Eigenschaften bestimmter Algorithmen veranschaulichen oder Ergebnisse auf innovative Weise aufzeigen, dann ist uns Geisteswissenschaftlern sehr geholfen.[11] Und wenn im Dialog von Geisteswissenschaftlern und Informatikern Anwendungen entstehen, deren Funktionsweise wir verstehen und die wir verwenden können, auch ohne diese Anwendungen selbst programmieren zu können, kann Fortschritt stattfinden. In jedem Fall gilt hier wiederum, dass wir eine geisteswissenschaftliche Kernkompetenz einbringen können und sollten; nämlich die Fähigkeit, mit Alteritätserfahrungen umzugehen und kommunikative Brücken zu bauen. Wenn dies

[11] Ein wunderbares Beispiel hierfür ist das "visual analytics tool", das von Jeong et al. 2009 beschrieben wird und die Principal Component Analysis veranschaulicht.

gelingt, könnte die Zusammenarbeit von Geisteswissenschaftlern und Informatikern durchaus neue und wichtige Einblicke in die Beziehung zwischen den Komödien Pierre Corneilles, Molières und anderer Autoren erlauben.

Bibliographie

Atelier de théorie littéraire (Hg., 2012): "Dossier Corneille-Molière", in: *Fabula.org*. [http://www.fabula.org/atelier.php?Corneille-Moliere]

Binongo, José Nilo / M.W.A. Smith (1999): "The Application of Principal Component Analysis to Stylometry", *Literary and Linguistic Computing* 14. 4, 445–466.

Blei, David (2011): "Introduction to Probabilistic Topic Models", in: *Communication of the ACM*.

Boissier, Denis (2004): *L'affaire Molière. La grande supercherie littéraire*. Paris: Jean-Cyrille Godefroy.

Borgman, Christine (2010): *Scholarship in the Digital Age : Information, Infrastructure, and the Internet*. Cambridge, MA / London: MIT Press.

Brunet, Étienne (2004): "Où l'on mesure la distance entre les distances." *Texto!*, 04 / 2004. [http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html]

Burnard, Lou / Bauman, Syd (Hg., 2007): *Guidelines for Electronic Text Encoding and Interchange*. Charlotteville: TEI Consortium. [http://www.tei-c.org/Guidelines/P5/]

Burrows, John (1987): *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Oxford & New York: Clarendon Press & Oxford University Press.

Burrows, John (2002): "Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing* 17.3, 267–287.

Burrows, John (2007): "All the Way Through: Testing for Authorship in Different Frequency Strata", in: *Literary and Linguistic Computing* 22.1, 27–47. [http://llc.oxfordjournals.org/content/22/1/27]

Crane, Gregory (2006): "What Do You Do with a Million Books?", in: *D-Lib Magazine*. [http://www.dlib.org/dlib/march06/crane/03crane.html]

Craig, Hugh / Kinney, Arthur (2009): *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

DARIAH-DE (2012): *DARIAH-DE. Aufbau von Forschungsinfrastrukturen für die eHumanities.* Göttingen: Niedersächsische Staats- und Universitätsbibliothek. [http://dx.doi.org/10.3249/webdoc-3589]

Eder, Maciej (2013): "Bootstrapping Delta: a Safety-net in Open-Set Authorship Attribution", *Digital Humanities 2013: Conference Abstracts*, Lincoln: University of Lincoln. [http://dh2013.unl.edu/abstracts/ab-135.html]

Eder, Maciej / Rybicki, Jan (2011): "Stylometry with R", in: *Digital Humanities 2011: Conference Abstracts*. Stanford: Stanford University, 308–11. [http://sites.google.com/site/computationalstylistics/]

Fièvre, Paul (Hg., 2007–2013): *Théâtre classique.* [http://www.theatre-classique.fr/]

Floridi, Luciano (2010): *Information: a Very Short Introduction*. Oxford / New York: Oxford University Press.

Génétiot, Alain (2005): *Le Classicisme*. Paris: PUF.

Holmes, David (1994): "Authorship attribution", in: *Computers and the Humanities* 28.2, 87–106.

Jannidis, Fotis (2007): "Computerphilologie", in: Thomas Anz (Hg.): *Handbuch Literaturwissenschaft*. Bd. 2. Stuttgart: Metzler, 27–40.

Jannidis, Fotis / Lauer, Gerhard (2014): "Burrows Delta and its Use in German Literary History", in: Erlin, Matt / Tatlock, Lynne (Hg.): *Distant Readings – Descriptive Turns. Topologies of German Culture in the Long Nineteenth Century*. Rochester: Camden House (im Druck).

Jeong, Dong Hyun / Ziemkiewicz, Caroline / Ribarsky, William / Chang, Remco (2009): "Understanding Principal Component Analysis Using a Visual Analytics Tool", *US Korea Conference 2009: Mathematics: Fundamentals and Applications*. [http://www.vrissue.com/portfolio/pdf/UKC2009.pdf]

Jockers, Matthew (2013): *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Juola, Patrick (2006): "Authorship Attribution", in: *Foundations and Trends in Information Retrieval* 1.3, 233–334.

Kestemont, Mike / Luyckx, Kim / Daelemans, Walter / Crombez, Thomas (2012): "Cross-Genre Authorship Verification Using Unmasking", in: *English Studies* 93.3, 340–356.

Labbé, Cyril / Labbé, Dominique (2001): "Inter-textual Distance and Authorship Attribution Corneille and Molière", in: *Journal of Quantitative Linguistics* 8.3, 213–231.

Louÿs, Pierre (o.J.): "'L'auteur d'*Amphitryon*' (*Le Temps*, 16.10.1919)", in: *L'Affaire Corneille-Molière*. [http://corneille-moliere.org/pageshtml/plamphitryon.html]

Merriam, Thomas (2003): "An Application of Authorship Attribution by Intertextual Distance in English", in: *Corpus* 2. [http://corpus.revues.org/35?&id=35]

Marusenko, Mikhail / Rodionova, Elena (2010): "Mathematical Methods for Attributing Literary Works When Solving the 'Corneille-Molière'-Problem", in: *Journal of Quantitative Linguistics* 17.1, 30–54.

McCarty, Willard (1999): "Humanities Computing as Interdiscipline", in: *Is Humanities Computing an Academic Discipline?* Charlottesville: IATH, University of Virginia. [http://www.iath.virginia.edu/hcs/mccarty.html]

Mendenhall, T.C. (1887): "The Characteristic Curves of Composition", in: *Science Supplement* IX.214, 237–246. [http://www.sciencemag.org/content/ns-9/214S/237.full.pdf]

Moretti, Franco (2000): "Conjectures on World Literature", in: *New Left Review* 1, 54–68. [http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature]

Moretti, Franco (2005): *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Mosteller, Frederick / Wallace, David L. (1963): "Inference in an Authorship Problem", in: *Journal of the American Statistical Association* 58.302, 275–309. [https://www.jstor.org/stable/2283270]

Pawłowski, Adam / Pacewicz, Artur (2004): "Wincenty Lutosławski (1863–1954): Philosophe, helléniste ou fondateur sous-estimé de la stylométrie?", in: *Historiographia Linguistic*a XXXI.2/3, 423–447.

Rybicki, Jan / Eder, Maciej (2011): "Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?", in: *Literary and Linguistic Computing* 26.3, 315–321. [http://llc.oxfordjournals.org/content/26/3/315]

Schöch, Christof (2012): "Beyond the black box, or: understanding the difference between various statistical distance measures", in: *The Dragonfly's Gaze*. [http://dragonfly.hypotheses.org/101]

Schöch, Christof / Jannidis, Fotis (2013a): "Quantitative Text Analysis for Literary History – Report on a DARIAH-DE Expert Workshop", in: *DARIAH-DE Working Papers* 2. Göttingen: DARIAH-DE. [https://de.dariah.eu/working-papers-publications]

Schöch, Christof (2013b): "Fine-tuning our Stylometric Tools: Investigating Authorship, Genre, and Form in French Classical Theater", in: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Lincoln. [http://dh2013.unl.edu/abstracts/ab-270.html]

Schöch, Christof (2013c): "A Stylometric Murder Mystery, or: *Poetic Justice* by Mitzi Morris", in: *The Dragonfly's Gaze*, 27.01.2013. [http://dragonfly.hypotheses.org/225]

Siemens, Ray / Unsworth, John / Schreibman, Susan (Hg., 2004). *A Companion to Digital Humanities*. Oxford: Blackwell, 2004. [http://www.digitalhumanities.org/companion/]

Sinclair, Stéfan / Rockwell, Geoffrey (2013): *Voyant Tools*, version 3.0. [http://www.voyant-tools.org]

Stierle, Karlheinz (2013): "Romanistik als Passion", *Frankfurter Allgemeine Zeitung*, 27.09.2013. [http://www.faz.net/aktuell/feuilleton/bilder-und-zeiten/philologie-romanistik-als-passion-12594341.html] (Kurzfassung des Eröffnungsvortrags zum Romanistentag 2013 in Würzburg)

Templeton, Clay (2011): "Topic Modeling in the Humanities: An Overview." *MITH Blog*. [http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/]

Unsworth, John (2002): "What Is Humanities Computing and What Is Not?", in: *Jahrbuch für Computerphilologie 4*, 71–83. [http://computerphilologie.digital-humanities.de/jg02/unsworth.html]

Van Dalen-Oskam, Karina / van Zundert, Joris (2007): "Delta for Middle Dutch. Author and Copyist Distinction in Walewein", in: *Literary and Linguistic Computing* 22.3, 345–362.

Witten, Ian / Frank, Eibe / Hall, Mark (2011): *Data Mining: Practical Machine Learning Tools and Techniques*. 3. Auflage. San Francisco: Morgan Kaufmann.

Wittmann, Reinhard (1999): "Gibt es eine Leserevolution am Ende des 18. Jahrhunderts?", in: Chartier, Roger / Cavallo, Guglielmo (Hg.): *Die Welt des Lesens. Von der Schriftrolle zum Bildschirm.* Frankfurt am Main / New York / Paris: 419–454.

Zanganeh, Lila Azam (2003): "Not Molière! Ah, Nothing Is Sacred", in: *The New York Times*, 06.09.2003. [http://www.nytimes.com/2003/09/06/books/not-moliere-ah-nothing-is-sacred.html]