Literaturgeschichte als Wissensgraph

Automatische Erhebung literaturhistorisch relevanter Informationen aus Volltexten am Beispiel von französischen Romanen des XVIII. Jahrhunderts

Julia Röttgermann

2025-04-08

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Stilometrische Analyse von Französischen Romanen 1751-1800

> The close reader sees things in a text — single moments and large amorphous movements —to which computer programs give no easy access. The computer, on the other hand, reveals hidden patterns and enables us to marshal hosts of instances too numerous for our unassisted powers.
>
> — Burrows (2002); pp. 696

1.1 Einleitung: Was ist Stilometrie?

Was ist Stilometrie und wozu können wir die Methode hinsichtlich einer datenbasierten Literaturgeschichte einsetzen? Die Methode, die Anwendung sowohl in den Digitalen Geisteswissenschaften als auch in der forensischen Linguistik findet, nutzt linguistische Merkmale, statistische Modelle, und computergestützte Verfahren, um stilistische Muster in schriftlichen Texten zu identifizieren und zu vergleichen. Eines der meist verbreiteten Ziele der Stilometrie ist es dabei, die Autorschaft eines Textes mit computationellen Mitteln nachzuweisen (Burrows 2002; Hoover 2010; Rotari / Jander / Rybicki 2021). Daneben gibt es jedoch auch Studien, die die statistischen sprachlichen Eigenschaften hinsichtlich von Gender-Unterschieden, Phänomene wie die individuellen Veränderungen eines Schreibstils im Laufe der Zeit oder in unterschiedlichen literarischen Gattungen analysieren (Holmes 1998; Jannidis / Lauer 2014; Schöch 2014; Weidman / Pastor 2021).
Der Begriff Stilometrie setzt sich aus den beiden Wortteilen "Stil" und "-metrie" zusammen, also eine literarische oder linguistische Kategorie (Stil) und das Konfix "-metrie", das suggeriert, dass etwas gezählt wird.[1] Zunächst stellt sich daher

[1] Morozov (1854-1946) verwendet erstmals den Begriff ‚stilometrija /*Stilometrie*' in Russland [Köhler / Altmann / Piotrowski (2005); pp. 37 ff.]. Für eine allgemeine Einführung in die

die Frage: Was ist Stil? Dazu haben (Herrmann / Van Dalen-Oskam / Schöch 2015) eine Definition formuliert, die man sowohl für computationelle Methoden als auch im klassischen literaturwissenschaftlichen Sinne anwenden kann: "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (Herrmann / Van Dalen-Oskam / Schöch 2015). Stilometrie nutzt das Zählen stilistischer Merkmale in Form der Verwendungshäufigkeit bestimmter Wörter ('most frequent words'), um Aussagen zur Textähnlichkeit zu generieren, die unter bestimmten Bedingungen Rückschlüsse auf die Autorschaft zulassen.[2]

Das vorliegende Kapitel beginnt damit, die Anfänge bzw. wegweisende Studien der Stilometrie zu beschreiben. Als Pionierstudien der Stilometrie werden die Studie zu den *Federalist Papers* (Mosteller / Wallace 1963) und Burrow's Delta (Burrows 2002) zusammengefasst. Als weitere theoretische Grundlage werden verschiedene Distanzmaße und zudem die Relevanz des R-Pakets 'stylo' innerhalb der Community der Digital Humanities beschrieben.
Im Anschluss beschreibe ich einen konkreten Anwendungsfall. Dazu untersuche ich eine Fallstudie zu einem Roman mit ungeklärter Autorschaft aus der Textsammlung roman18 in mehreren Schritten: Verwendung eines Subkorpus von roman18, Analyse mit um Texte eines potentiellen Autors erweitertem Datensatz, Hierarchical Wards Clustering, Bootstrap Consensus Tree. Eine überraschende Erkenntnis liefert der Vergleich aus Clustering und numerischen Distanzwerten der stilometrischen Analyse. Abschließend beschreibe und diskutiere ich die Modellierung in Wikibase unter Verwendung von 'ranks' und SPARQL-Abfragen zu Gemeinsamkeiten stilometrisch naher Werke.

Keywords: Stilometrie, Federalist Papers, Delta, stylo, Hierarchical Wards Clustering, Bootstrap Consensus Tree, Principal Component Analysis, Wikibase, preferred rank, SPARQL.

Methode der Stilometrie im Kontext der Digital Humanities cf. (Laramée 2018; Horstmann 2024; Eder / Rybicki / Kestemont 2016).

[2] Hinzufügen könnte man außerdem, dass der Computer hier völlig neue Möglichkeiten eröffnete. Gingen Mosteller und Wallace in Teilen ihrer Berechnung noch händisch vor, ist es heute eine Selbstverständlichkeit, eine stilometrische Studie von einem Computer berechnen zu lassen. "The most important technological advances in authorship studies have arisen from the computer." stellen sie auch selbst 20 Jahre nach ihrer wegweisenden Studie in einem Zusatzkapitel fest [Mosteller / Wallace (1963); pp. 268].

1.2 Pionierstudien und Toolentwicklung

Eine Studie von Mosteller und Wallace 1964 zu den *Federalist Papers* verschaffte der Stilometrie den wissenschaftlichen Durchbruch: Die *Federalist Papers* sind eine 1787-1788 veröffentlichte Sammlung von 85 Essays zur politischen Theorie. In diesen 900 bis 3500 Wörter langen Schriften, die während der Debatte über die Ratifizierung der Verfassung der Vereinigten Staaten verfasst wurden, werden die Argumente für das Regierungssystem dargelegt, dass die USA schließlich angenommen haben und unter dem sie bis zum heutigen Tag leben. Die Autorschaft dieser Artikel, unterzeichnet mit dem Pseudonym "Publius", war lange Zeit unklar (Laramée 2018). Mithilfe der Analyse der Wortfrequenzen gelang es Mosteller und Wallace, die ungeklärte Zuordnung aus mehreren möglichen Autoren aufzulösen [Kenny (1982); pp.8–9]. Die Bedingungen in der *Federalist Paper*-Studie können sicherlich auch als ideal gelten, werden doch mehrere Parameter - wie Kontext, Gender und Gattung - konstant gehalten.

Ein weiterer bedeutender Beitrag zur Stilometrie stammt von John Burrows, der das Distanzmaß ›Delta‹ entwickelte, mithilfe dessen sich für einen anonymen Text eine Zuordnung zu einem Korpus bekannter Texte und Autor:innen treffen lässt. Seine Studie, die multivariate statistische Verfahren wie Cluster- oder Principal Component Analyse (PCA) einsetzt, und die daraus resultierenden Erkenntnisse haben bedeutende Impulse für die Stilometrie und die Autorschaftsattribution in den Computational Literary Studies geliefert. Burrows Delta verwendet den Manhattan-Abstand und hat sich als besonders effektiv erwiesen, um Autorschaftsfragen oder die Beteiligung einzelner Autoren an umfangreichen Textsammlungen zu klären (Büttner et al. 2017). Basierend auf Burrows Delta wurden weitere Varianten zur Verbesserung vorgeschlagen, beispielsweise der Vorschlag, statt der Manhattan-Distanz den Kosinus-Winkel zwischen den Vektoren zu verwenden (Smith / Aldridge 2011).[3]

Bezüglich der Toolentwicklung kommt Maciej Eder, Jan Rybicki und Mike Kestemont der Verdienst zu, ein eingängiges, leicht bedienbares Tool für die Statistikumgebung R entwickelt zu haben, dass auch zur weiten Verbreitung der Methode in den Digital Humanities und darüber hinaus geführt hat (Horstmann 2024): "Stylo" (Eder / Rybicki / Kestemont 2016).

[3]Das Projekt "***Zeta und Konsorten.*** Distinktivitätsmaße für die Computational Literary Studies." am Trier Center for Digital Humanities hat es sich zum Ziel gesetzt, ein tieferes Verständnis verschiedener Distinktivitätsmaße zu erreichen und Verbesserungen in deren Implementierung und Anwendung vorzuschlagen (cf. Du / Dudar / Schöch 2022).

1.3 Theoretische Grundlagen: Distanzmaße

Im Kontext der Textanalyse werden verschiedene Distanzmaße verwendet, um die Unterschiede zwischen Texten zu quantifizieren [Schöch (2017); pp. 294–295]. Die **Manhattan-Distanz** misst die absolute Distanz zweier Vektoren in jeder einzelnen Dimension. Die Summe dieser absoluten Distanzen ergibt die Gesamtdistanz, wobei jede Dimension gleich gewichtet wird (Figure **??**).

Die **Euklidische Distanz** ähnelt der Vogelfluglinie zwischen zwei Punkten. Hierbei werden die Distanzen in jeder Dimension quadriert, summiert und schließlich die Quadratwurzel aus der Gesamtsumme gezogen. Dies führt dazu, dass größere Distanzen in einer einzelnen Dimension durch die Quadrierung stärker gewichtet werden, was in der stilometrischen Autorschaftsattribution besonders relevant ist, da häufig vorkommende Wörter einen überproportionalen Einfluss haben können. Bei der euklidischen Distanz haben die höchstfrequenten Wörter besonders viel Einfluss.

Der **Kosinus-Abstand** wird als eine Form der Vektor-Normalisierung betrachtet, da bei der Berechnung des Winkels zwischen Vektoren die Länge der Vektoren keine Rolle spielt. Im Gegensatz zu Manhattan- und Euklidischem Abstand ermöglicht der Kosinus-Abstand eine Bewertung, die unabhängig von der Länge der Vektoren ist.

In Bezug auf die stilometrische Autorschaftsattribution betonen (Büttner et al. 2017), dass das charakteristische stilistische Profil eines Autors eher in der qualitativen Kombination bestimmter Wortpräferenzen zu finden ist. Dies bezieht sich auf das grundlegende Muster von über- bzw. unterdurchschnittlich häufigem Gebrauch von Wörtern, anstatt auf die Amplitude dieser Abweichungen. "Delta" in der stilometrischen Autorschaftsattribution nutzt die Manhattan-Distanz und erweist sich als erfolgreich, da es strukturelle Unterschiede in den sprachlichen Vorlieben eines Autors erfasst, ohne stark von der Intensität des Autorenprofils in einem bestimmten Text beeinflusst zu werden (Evert et al. 2017).

Nachdem nun wegweisende Studien der Stilometrie vorgestellt und theoretische Grundlagen der Methode erläutert wurden, stellt sich die Frage, wozu stilometrische Analysen im Kontext einer datenbasierten Literaturgeschichte beitragen können. Eine der häufigsten Anwendungsfälle stellt sicherlich die Autorschaftsattribution dar. Ob J. K. Rowling (Juola 2015) oder Shakespeare (Craig / Kinney 2009) lassen sich hier viele Anwendungsstudien zitieren.
Stilometrie lässt sich jedoch neben der Analyse von Autorschaftssignalen auch zur Bestimmung von Gattungszugehörigkeit einsetzen (Schöch 2014) oder in der diachronen Betrachtung einer stilistischen Handschrift eines einzelnen Autors

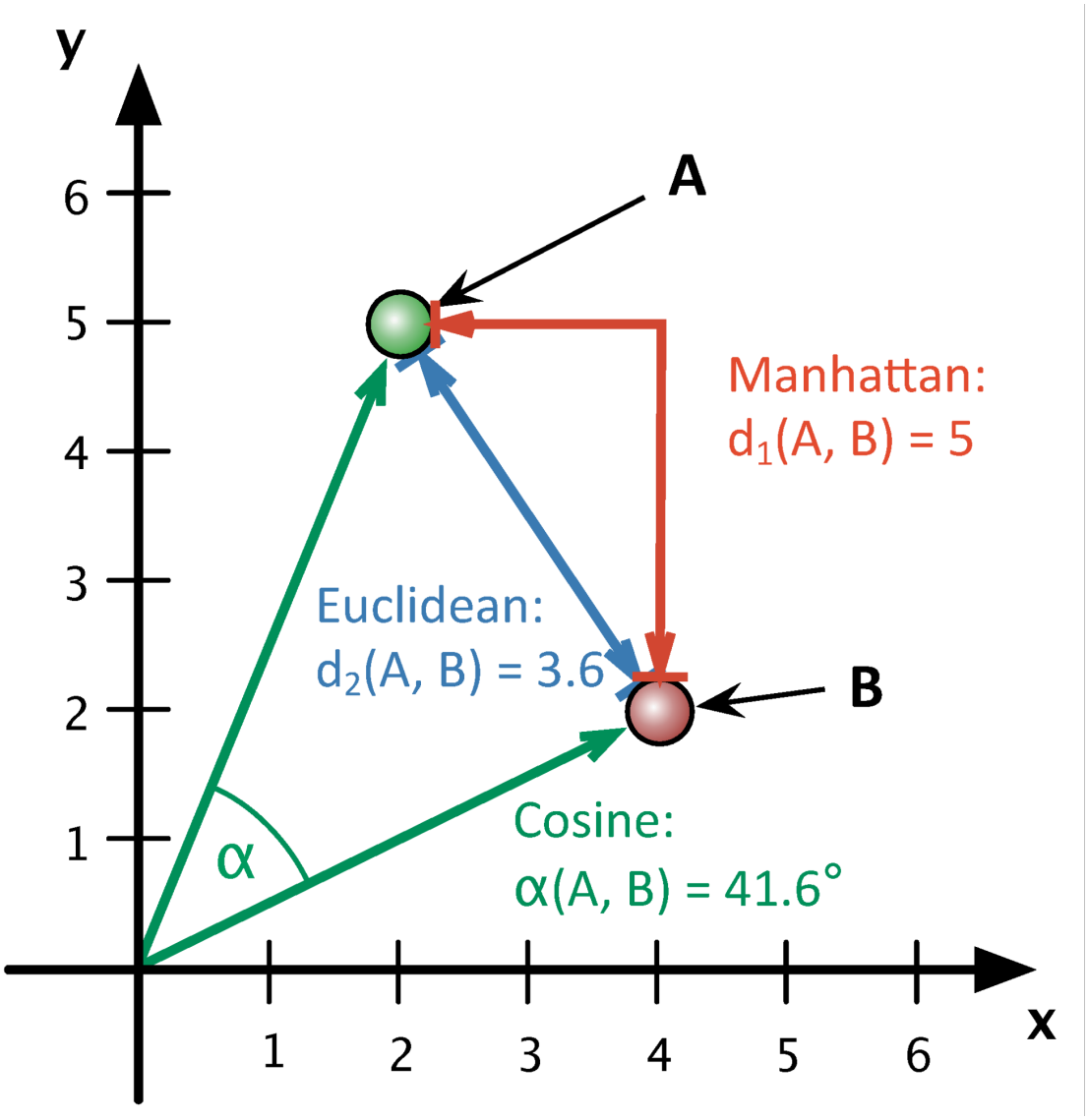

Figure 1.1: Distanzmaße: Manhattan-Distanz, Euklidische Distanz und Kosinus-Distanz [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

oder Autorin, beispielsweise in der Frage: Gibt es so etwas wie “late style” (Reeve; Rebora / Salgaro 2018)?

Wieso funktioniert die Methode Stilometrie? Dies lässt sich durch einen interessanten historischen Vergleich in der kunsthistorischen Forschung veranschaulichen, den Mike Kestemont beschreibt. Der Wechsel von Inhaltswörtern zu Funktionswörtern in Studien zur Autorschaftsattribution findet hier eine bemerkenswerte Parallele:

“Giovanni Morelli (1816-1891) was among the first to suggest that the attribution of, for instance, a
Quattrocento painting to some Italian master, could not happen based on ‘content’ […] Morelli thought it better to restrict an authorship analysis to discrete details such as ears, hands and feet” (Kestemont 2014: pp.61)

Auch in der Malerei lassen sich über die Analyse von unwillkürlich getroffenen Entscheidungen, wie der Art, die Hände, Ohren oder Füße zu malen, Rückschlüsse auf die Autorschaft ziehen. Das Zitat unterstreicht die Faszination darüber, dass Menschen, sei es in Texten oder Gemälden, unbewusste Entscheidungen treffen, die in der Summe Aufschlüsse über ihre künstlerische Handschrift geben, so auch bei der Verwendung bestimmter Funktionswörter.

Im nächsten Abschnitt sollen mithilfe von Stilometrie innerhalb des Korpus französischer Romane 1751-1800 (roman18) Aussagen zur möglichen Autorschaft untersucht werden; Erzählform oder thematische Ausrichtung sind ebenfalls im Graphen als Statements hinterlegt und könnten theoretisch in einem weiteren Schritt auch mit in Betracht gezogen werden, werden jedoch in der hier vorliegenden Anwendung ausgeklammert. Stilometrie wird in der hier vorliegenden Untersuchung mit dem Tool Stylo in R durchgeführt (Eder / Rybicki / Kestemont 2016).

1.4 Fallstudie: Wer ist der Autor von *L’Étourdi* (1784)?

Ein Roman mit abweichenden Angaben zur Autorschaft aus der Romansammlung roman18 (Röttgermann 2023) soll nun als Anschauungsbeispiel der stilometrischen Analyse dienen. *L’Étourdi* ist ein Roman aus dem Jahr 1782. Laut bibliographischen Metadaten (Martin / Mylne / Frautschi 1977, pp. 272) ist der Autor André-Robert Andréa de Nerciat (Figure **??**).[4] Die für die Generation des TEI-Volltexts

[4] Außerdem werden zwei weitere mögliche Autorschaften genannt: Neufville-Montador oder Marquis de Sade (Martin / Mylne / Frautschi 1977, pp. 272).