Stilometrie

Stilometrie ist eine Disziplin der quantitativen Literaturwissenschaft. Es handelt sich dabei um die operationalisiert-quantitative Untersuchung (*-metrie*) des textuellen Phänomens des Stils (*Stilo-*). Heutzutage werden stilometrische Untersuchungen so gut wie immer digital ausgeführt, sodass man die Stilometrie auch der digitalen Literaturwissenschaft zurechnen kann. Um Missverständnisse zu vermeiden, kann man auch explizit von *digitaler Stilometrie* sprechen.

Vorgehen der Stilometrie

Eine Grundvoraussetzung der Stilometrie ist die operationalisierend-quantitative Betrachtung von *Stil*.[1] Hierzu ist es notwendig, Stil zu abstrahieren und auf eindeutig identifizierbare textuelle Merkmale zu reduzieren. Eine häufig angewandte Variante ist die Betrachtung der häufigsten Wörter eines Textes (der *most frequent words*) – dieses Verfahren wird u.a. in der *Autorenattribution* angewandt, wenn also ein anonymer Text bzw. ein Text, dessen Autor unbekannt oder zweifelhaft ist, zu bestimmen ist.[2]

Auch wenn sich nicht zweifelsfrei auf theoretischer Ebene begründen lässt, warum die Auffassung von Stil in Form der häufigsten Wörter empirisch korrekte Ergebnisse liefern kann, wurde die Betrachtung der *most frequent words* bereits mehrfach erfolgreich eingesetzt.[3]

Andere Möglichkeiten, das Phänomen des Stils zu operationalisieren sind z.B. *n-grams* bzw. *n-Gramme*, d.h. mehrere aufeinanderfolgende Elemente werden zusammen betrachtet. So kann man den Satz *Io vado a casa.* als drei *Bigramme* auf Wortebene auffassen: {*Io vado*; *vado a*; *a casa*}. Selbstverständlich sind n-Gramme auch auf Zeichenebene, auf Satzebene oder auf anderen Ebenen möglich. Weitere Formen der Operationalisierung von Stil können die Betrachtung der Satzlänge oder der Interpunktion sein. In komplexen stilometrischen Studien werden häufig mehrere Stil-Operationalisierungen miteinander kombiniert.

Distanzmaße in stilometrischen Untersuchungen

Grundsätzlich kann man Stilometrie dahingehend auffassen, dass statistische Berechnungen auf den extrahierten Stildaten ausgeführt werden. Hierbei kommen häufig sogenannten Distanzmaße zum Einsatz. Mit Hilfe dieser Distanzmaße lässt sich mathematisch ausdrücken, wie ähnlich oder unähnlich sich die Stildaten verschiedener Entitäten (z.B. Texten) sind. Wenn man beispielsweise die *most frequent words* aus mehreren Texten ermittelt hat und in eine mathematische *Matrix* (oder in einen mathematischen *Vektor* – in der Mathematik werden Matrizen mit einer Spalte bzw. einer Zeile als *Vektor* bezeichnet) überführt hat, kann man mithilfe von Distanzmaßen die Unterschiedlichkeit (also die Distanz) der verschiedenen Matrizen berechnen.

Eines der am häufigsten verwendeten Distanzmaße in der Literaturwissenschaft ist das *Delta*-Maß, das von John Burrows explizit zur Distanzberechnung im Bereich der Textanalyse entwickelt wurde.[4] Selbstredend existieren aber auch andere Distanzmaße, so das *Euklidsche* Distanzmaß oder das *Manhattan* Distanzmaß.[5] Mithilfe der Distanzwerte lassen sich nun Aussagen zu den analysierten Textdaten treffen. Häufig lagert man noch eine weitere Verarbeitung der Distanzwerte nach, z.B. eine *Clusterung* mithilfe eines unüberwachten Lernalgorithmus oder die Anwendung einer *PCA* (*Principal Component Analysis* bzw. *Hauptkomponentenanalyse*) – ein unüberwachtes Verfahren zur Dimensionsreduktion.

(Überwachtes) Maschinelles Lernen in der Stilometrie

Während die erwähnten Verfahren der *Clusterung* und der *PCA* bereits zum (unüberwachten) maschinellen Lernen zählen, sind auch überwachte Lernverfahren anwendbar. Dies kann u.a. nützlich sein, wenn man den zu analysierenden Texten ein eindeutiges Label zuweisen möchte, d.h. wenn man die Texte eindeutig klassifizieren möchte. Hierzu können auf die aus den Texten extrahierten Textdaten

(z.B. eine Matrix der *most frequent words*) entsprechend überwachte Lernalgorithmen angewandt werden. Beispiele sind *k-nearest neighbor* oder *Support Vector Machines* (*SVM*). Diese Algorithmen benötigen, um überwacht klassifizieren zu können, sogenannte Trainingsdaten. D.h. man stellt ihnen Referenzdaten zur Verfügung, für die das Label jeweils bekannt ist. So kann man den Algorithmen Texte vorgeben, deren Autor bereits bekannt ist. Ausgehend von den Trainingsdaten kann der Algorithmus schließlich Schätzungen abliefern, welches Label die zu analysierenden Texte haben dürften.

Digitale Tools zur Durchführung stilometrischer Untersuchung

Eines der beliebtesten Tools zur Durchführung stilometrischer Untersuchungen ist *stylo*[6] – ein Paket in der Programmiersprache *R*, das verschiedene Routinen aus dem Bereich der Stilometrie zur Verfügung stellt und auch über eine graphische Benutzeroberfläche aufrufbar macht.

Neben R ist auch *Python* eine gängige und häufig verwendete Programmiersprache für stilometrische Untersuchungen, da Pythons Stärken u.a. in der Verarbeitung (großer) Datenmengen, aber auch im Bereich des maschinellen Lernens liegt. Bekannte Python-Anwendungen sind die Implementierung des Kontrastiv-Maßes *Zeta*[7] oder die Umgebung *Pystyl* (die jedoch offensichtlich nicht intensiv weiterentwickelt wird, der letzte Commit fand vor sechs Jahren statt). Grundsätzlich ist es von Vorteil, wenn man bei der Anwendung stilometrischer Verfahren über Grundkenntnisse in der Programmierung verfügt, um die Vorgehensweise der Tools nachvollziehen zu können und um ggf. eigene Routinen programmieren zu können.

Literaturnachweise

1. Vgl. SAVOY, Jacques 2020: *Machine learning methods for stylometry. Authorship attribution and author profiling*, Cham, S. 3–5.[↩]

2. Vgl. u.a. BURROWS, John 2002: „‚Delta‘: a Measure of Stylistic Difference and a Guide to Likely Authorship“, in: *Literary and Linguistic Computing*, Bd. 17, 3, S. 267–287.[↩]
3. Vgl. u.a. JUOLA, Patrick 2013: „How a Computer Program Helped Show J.K. Rowling write A Cuckoo’s Calling“, in: *Scientific American*, Bd. 20, August 2013, https://www.scientificamerican.com/article/howa-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/#.[↩]
4. Vgl. u.a. BURROWS, John 2002: „‚Delta‘: a Measure of Stylistic Difference and a Guide to Likely Authorship“, in: *Literary and Linguistic Computing*, Bd. 17, 3, S. 267–287.[↩]
5. Vgl. hierzu u.a. JANNIDIS, Fotis/PIELSTRÖM, Steffen/SCHÖCH, Christof 2015: „Improving Burrows’ Delta – An empirical evaluation of text distance measures“, in: *Digital Humanities Conference 2015*, Sydney, o.S.[↩]
6. Vgl. EDER, Maciej/RYBICKI, Jan/KESTEMONT, Mike 2016: „Stylometry with R: a package for computational text analysis“, in: *R Journal*, Bd. 8, 1, S. 107–121. Die Software bzw. ihr Quellcode ist online unter: https://github.com/computationalstylistics/stylo einsehbar.[↩]
7. Vgl. u.a. SCHÖCH, Christof 2018: „Zeta für die kontrastive Analyse literarischer Texte: Theorie, Implementierung, Fallstudie“, in: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*, hg. von edited Toni Bernhart, Marcus Willand, Sandra Richter and Andrea Albrecht, Berlin/Boston, S. 77–94.[↩]