Toolbeitrag: Stylo

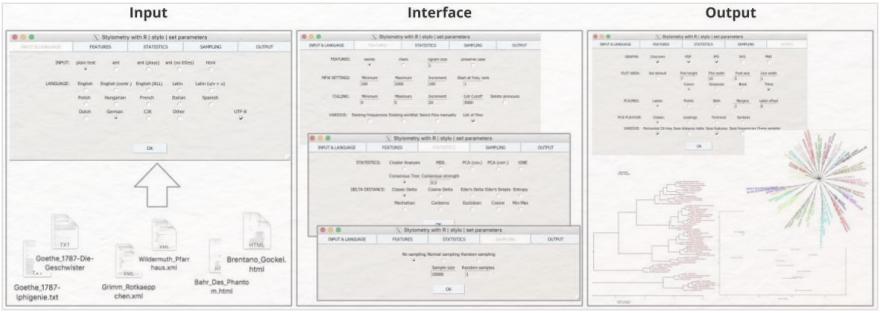
Jan Horstmann [1]

1. Universität Münster

forTEXT

Thema:		DOI:	10.48694/fortext.3770
Jahrgang:	1	Ausgabe:	
Erscheinungsdatum:	2024-02-26	Erstveröffentlichung:	2019-01-07 auf fortext.net
Lizenz:	CC BY SA		open access

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



Der Stylo-Workflow: Nach dem Öffnen der grafischen Benutzeroberfläche lassen sich Textsammlungen in den Formaten TXT (empfohlen), XML oder HTML hochladen. Die einzelnen Tabs führen Sie durch die möglichen Einstellungen der stilometrischen Analyse. Heruntergeladen und spezifiziert werden die ausgewählten Visualisierungen in verschiedenen Dateiformaten wie PDF und/oder JPG.

- **Systemanforderungen:** Vorherige Installation von R bzw. RStudio (und für Mac-User*innen XQuartz); von dort kann das „stylo"-Package entweder aus CRAN oder GitHub installiert werden, oder auch aus einer lokal gespeicherten Datei (hierfür müssen Sie vorab zusätzliche R-Packages installieren: tcltk2, ape, class, e1071, pamr, tsne); Stylo ist mit allen Betriebssystemen nutzbar; zur Installation der Packages benötigen Sie eine Internetverbindung, ab dann kann Stylo offline verwendet werden (zur Verwendung von Stylo in der Programmiersprache Python siehe Calvo Tello 2017)
- **Stand der Entwicklung:** Version 0.6.9 (Oktober 2018)
- **Herausgeber:** Maciej Eder, Mike Kestemont, Jan Rybicki
- **Lizenz:** Kostenfrei, Open Source
- **Weblink:** https://github.com/computationalstylistics/stylo
- **Im- und Export:** Import: Alle Dateien einer Textsammlung (vgl. Korpus) sollten das gleiche Format haben; die Entwickler empfehlen das TXT (vgl. Reintext-Version)-Format (UTF8-codiert (vgl. Unicode/UTF-8)); auch möglich, aber weniger erprobt, sind HTML und TEI-XML; die einzelnen Dateien müssen nach dem Muster Kategorie_Titel.txt gespeichert werden, z. B. Bachmann_Malina.txt; Export: Ausgabe jedes Durchlaufs als txt-Datei; Visualisierungen: PDF, JPG, PNG, SVG (nützlich zur Einbindung in HTML-Codes oder weitere Verarbeitung)
- **Sprachen:** Optimiert für Englisch, Latein, Polnisch, Ungarisch, Französisch, Italienisch, Spanisch, Holländisch, Deutsch, CJK (Chinesisch/Japanisch/Koreanisch auf Basis gleicher Zeichen), außerdem gibt es die Option „other" (inwiefern die Methode auch mit anderen Sprachen stabil läuft, sollte selbst getestet werden)

1. Für welche Fragestellungen kann Stylo eingesetzt werden?

Mit Stylo lassen sich alle Fragen der Stilometrie (Horstmann 2024) bearbeiten. Dazu gehören vor allem Fragen der Autorschaftsattribution, Genre- oder Epochenklassifikationen, stilistische Entwicklungen eines Autorinnenoeuvres usw. Das Tool ermöglicht dabei die Anwendung unterschiedlicher in der Stilometrie diskutierter Algorithmen.

2. Welche Funktionalitäten bietet Stylo und wie zuverlässig ist das Tool?

Funktionen:

- Stilistische Vergleichsanalyse von Texten oder Textsegmenten anhand der häufigsten Wörter (MFW)
- Verschiedene gängige Statistiken sind implementiert
- Verschiedene Preprocessing-Maßnahmen sind im Programm inbegriffen, unter anderem ein Tokenizer (vgl. Type/Token) und eine Pronomen-/Stoppwortliste für mehrere Sprachen (mit der Möglichkeit, die jeweiligen Wörter aus den Texten zu löschen)
- Variable Visualisierungsformen der Ergebnisse

Zuverlässigkeit: Stylo funktioniert je nach Größe Ihrer Textsammlung und vorgenommener Voreinstellungen zügig und zuverlässig. Die errechneten Ergebnisse sind je nach Datengrundlage und manuell vorgenommenen Voreinstellungen so gut wie die zugrunde gelegten stilometrischen Algorithmen (z. B. „Burrows' Delta"). Diese Algorithmen sind nicht von oder für Stylo selbst entwickelt worden, sondern finden lediglich im Tool Verwendung. Als mathematische Formeln ergeben sie immer nur Annäherungswerte an tatsächliche Phänomene.

3. Ist Stylo für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / -
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	teilweise
Leichter Einstieg	-
Handbuch vorhanden	✓
Handbuch aktuell	✓
Tutorials vorhanden	✓
Erklärung von Fachbegriffen	teilweise
Gibt es eine gute Nutzerbetreuung?	✓

Die Stylo zugrunde liegende Idee ist es, uns ohne Coding-Kenntnisse (vgl. Commandline) zu ermöglichen, hochfunktionale und komplexe Algorithmen der Stilometrie zu verwenden. Grundlegende Kenntnisse in der Programmiersprache R sind jedoch zum Installieren und Starten des Tools vonnöten. Das Preprocessing findet ebenfalls innerhalb des Programms und unabhängig von der jeweiligen Sprache statt. Eine Sammlung von Folien bildet ein Tutorial zum Einstieg. Das Handbuch ist darum bemüht, Fachbegriffe zu erklären, für Anfänger*innen mögen einige Erklärungen aber zu technisch sein. Bei Fragen oder Problemen gibt es zwar keine Helpdeskfunktion, Sie können die Entwickler von Stylo aber über GitHub und Twitter kontaktieren. Verhältnismäßig schnelle Hilfe erhält man auch in einem Google-Forum der Computational-Stylistics-Gruppe.

4. Wie etabliert ist Stylo in den (Literatur-)Wissenschaften?

Stylo ist für die digitale Stilometrie eines der etabliertesten Tools. Es findet in zahlreichen Projekten zur Autorschaftsattribution oder zum geschlechtsspezifischen Schreiben Anwendung. Methodisch reflektiert nutzen auch kombinierte Ansätze von Distant- (vgl. Distant Reading) und Close Reading-Verfahren (sog. „mixed-methods (vgl. Scalable Reading)") das Tool (Herrmann 2017). Wie die meisten digitalen Textanalysetools findet jedoch auch Stylo keine Erwähnung in Publikationen von Zeitschriften der traditionelleren Literaturwissenschaft.

5. Unterstützt Stylo kollaboratives Arbeiten?

Nein. Stylo wird als R-Package auf dem eigenen Computer ausgeführt und ermittelte Ergebnisse müssen individuell verteilt und diskutiert werden.

6. Sind meine Daten bei Stylo sicher?

Ja. Es werden keine personenbezogenen Daten erhoben. Da Stylo auf dem eigenen Rechner genutzt wird, müssen Sie Ihre Texte zudem nirgendwo hochladen, um sie stilometrisch zu erforschen.

Externe und weiterführende Links

- Google-Forum: https://web.archive.org/save/https://groups.google.com/forum/#!forum/computationalstylistics (Letzter Zugriff: 20.02.2024)
- R Project: https://web.archive.org/save/https://www.r-project.org/ (Letzter Zugriff: 20.02.2024)

- R Studio: https://web.archive.org/save/https://www.rstudio.com/ (Letzter Zugriff: 20.02.2024)
- Stylo auf Github: https://web.archive.org/save/https://github.com/computationalstylistics/stylo (Letzter Zugriff: 20.02.2024)
- Stylo Tutorial: https://web.archive.org/save/https://computationalstylistics.github.io/stylo_nutshell/ (Letzter Zugriff: 20.02.2024)
- Twitter Maciej Eder: https://web.archive.org/save/https://twitter.com/MaciejEder (Letzter Zugriff: 20.02.2024)
- XQuartz: https://web.archive.org/save/https://www.xquartz.org/ (Letzter Zugriff: 20.02.2024)

Bibliographie

Calvo Tello, José. 2017. Using Stylo in Python. Juni. https://cligs.hypotheses.org/577 (zugegriffen: 12. November 2018).

Eder, Maciej. 2017. Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities* 32, Nr. 1: 50–64. doi: 10.1093/llc/fqv061, (zugegriffen: 13. November 2018).

Eder, Maciej, Jan Rybicki und Mike Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis. *The R Journal* 8, Nr. 1: 107–121. https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf (zugegriffen: 12. November 2018).

Herrmann, Berenike J. 2017. In a text bed with Kafka. Introducing a mixed-method approach to digital stylistics. *Digital Humanities Quarterly* 11, Nr. 4. http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html (zugegriffen: 12. November 2018).

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

Glossar

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

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

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

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

CSV CSV ist die englische Abkürzung für *Comma Separated Values* . Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv` , sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben

die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

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

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

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

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

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

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

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

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

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

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

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

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

Stoppwortliste Stoppwörter sind hochfrequente Wörter, meist Funktionswörter, die, aufgrund ihrer grammatisch bedingten Häufigkeit, beispielsweise die Ergebnisse von inhaltlichen oder thematischen Analysen verzerren können. Deshalb werden diese Wörter, gesammelt in einer Stoppwortliste, bei digitalen Textanalysen meist nicht berücksichtigt.

SVG SVG steht für *Scalable Vector Graphics* und ist ein freies, standardisiertes Dateiformat, das Bilddateien bezeichnet, die als 2D-Vektorgrafiken größenunabhängig reproduziert werden können. Bei SVG-Dateien wird im Gegensatz zu anderen Bildgrafiken somit die Auflösung der Abbildung beim Vergrößern nicht schlechter. Es basiert auf den Strukturen von XML und wird dazu verwendet, Bilddaten zu repräsentieren.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

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

Unicode/UTF-8 Unicode ist ein internationaler Standard, der für jedes Schriftzeichen oder Textelement einen digitalen Code festlegt. Dabei ist UTF-8 die am weitesten verbreitete Kodierung für Unicode-Zeichen. UTF-8 ist die international standardisierte Kodierungsform elektronischer Zeichen und kann von den meisten Digital-Humanities-Tools verarbeitet werden.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.