Lerneinheit: Stilometrie mit Stylo

Jan Horstmann
1. Universität Münster

Thema:	Stilometrie	DOI:	10.48694/fortext.3771
Jahrgang:	1	Ausgabe:	1
Erscheinungsdatum:	2024-02-26	Erstveröffentlichung:	2019-05-20 auf fortext.net
Lizenz:			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.

Eckdaten der Lerneinheit

- Anwendungsbezug: 67 deutschsprachige Texte
- Methode: Stilometrische Analyse
- Angewendetes Tool: Stylo
- Lernziele: Installation von R, RStudio und des Stylo-Packages, Anwendung unterschiedlicher stilometrischer Analysemethoden, Interpretation der Visualisierungen
- Dauer der Lerneinheit: ca. 90 Minuten
- Schwierigkeitsgrad des Tools: mittel

Bausteine

- Anwendungsbeispiel
  Welche Texte werden analysiert? Untersuchen Sie ein Korpus von 67 deutschsprachigen Texten aus dem 19. und beginnenden 20. Jahrhundert auf stilistische Verwandtschaft.
- Vorarbeiten
  Was muss vor der stilometrischen Analyse getan werden? Lernen Sie, wie Sie sich das Korpus herunterladen, die Software R und RStudio installieren, eine Session einrichten und das Stylo-Package hinzufügen.
- Funktionen
  Welche Funktionen bietet Stylo Ihnen für die stilometrische Analyse des Korpus? Lernen Sie ausgewählte Analysefunktionen von Stylo kennen und lösen Sie Beispielaufgaben.
- Lösungen zu den Beispielaufgaben
  Haben Sie die Beispielaufgaben richtig gelöst? Hier finden Sie Antworten.

1. Anwendungsbeispiel

In dieser Lerneinheit werden wir anhand einer Textsammlung von 67 deutschsprachigen Texten (Romane, Romanauszüge, Dramen etc.) aus dem 19. und beginnenden 20. Jahrhundert die grundsätzlichen Funktionen der stilometrischen Textanalyse kennenlernen. Die Texte stammen zum Großteil aus dem literarischen Kanon; einer ist jedoch der anonym erschienene Text *Schwester Monika* und wir werden versuchen, herauszufinden, welchem Autor/welcher Autorin dieser Text evtl. zugesprochen werden könnte. In der Stilometrie (Horstmann 2024a) werden hierzu die statistischen Verteilungen der häufigsten Wörter (*most frequent words*) in den Texten miteinander verglichen (vgl. Text Mining); diese Verteilungsmuster sind oft autor*innenspezifisch. Einer der zentralen Anwendungsfälle der Stilometrie ist daher die Autorschaftsattribution, ebenso kommt sie aber bei stilistischen Gattungs-, Genre- oder Epochenklassifizierungen zum Einsatz. Das Tool Stylo (Horstmann 2024b) wird am häufigsten für diese Methode verwendet. Es integriert die gängigen Algorithmen stilometrischer Verfahren in einer um Nutzerfreundlichkeit bemühten grafischen Benutzeroberfläche (vgl. GUI).

2. Vorarbeiten

Zunächst laden Sie sich die Sammlung der **67 Texte** auf Ihren Rechner. Gehen Sie dazu über diesen Link auf die Github-Seite der Computational Stylistics Group, die das Stylo-Package für die Statistiksoftware R entwickelt hat. Dort klicken Sie auf „Clone or download“ und dann auf „Download ZIP“ (vgl. Abb. 1).

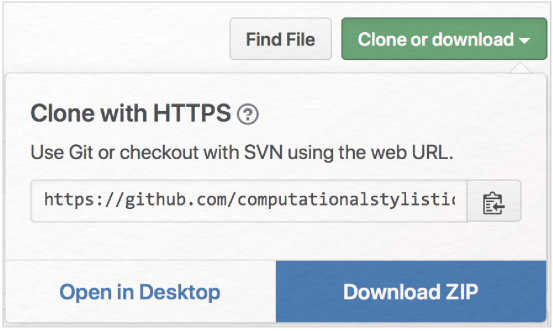

Abb. 1: Herunterladen der Textsammlung von Github als ZIP-Datei

In Ihrem Downloadordner finden Sie nun einen Ordner namens „68_german_novels-master“. Speichern Sie diesen Ordner an einem Ort auf Ihrem PC, an dem Sie ihn leicht wiederfinden können. Sie sehen in diesem Ordner einen weiteren Ordner, der die 67 Texte im TXT-Format enthält, das für Stylo ideal ist. Die Dateinamen folgen dem Muster „Kategorie_Titel“. Die Kategorie sind hier die Autor*innennamen, genauso gut könnte man hier Genres, Genderbezeichnungen, Übersetzer*innennamen etc. angeben. Wichtig ist, dass alle Dateien in einer Textsammlung diesem Muster entsprechend benannt werden, damit sie von Stylo im Analyseergebnis korrekt visualisiert werden (jede Kategorie bekommt dort eine andere Farbe) und dass die Texte einer Sammlung alle das gleiche Dateiformat haben. Da Stylo nur die Worte in ihrer Häufigkeit und Verteilung analysiert, d. h. lediglich die Textoberfläche betrachtet, werden bei dieser Methode keine Metadaten benötigt. Eine Textsammlung im XML- oder auch HTML-Format (vgl. HTML) kann im Tool zwar verarbeitet werden, die besten Ergebnisse werden jedoch mit Dateien im Plaintext-Format (vgl. Reintext-Version) (TXT) erzielt. Eine weitere günstige Voraussetzung für die stilometrische Analyse von Textsammlungen mit Stylo ist die UTF-8-Codierung (vgl. Unicode/UTF-8) der Texte, die beim vorliegenden Korpus ebenfalls gegeben ist.

Nun müssen Sie sich noch die **Software R installieren**, in der Stylo ausgeführt werden soll. Dazu folgen Sie diesem Link, klicken unter der Überschrift „Getting Started“ auf „Download R“ und wählen einen sog. „CRAN Mirror“, der in der Nähe Ihres Standortes liegt (wie z. B. unter Germany die GWDG Göttingen). CRAN (d. h. Comprehensive R Archive Network) ist ein internationales Server-Netzwerk (vgl. Server), aus dem Sie die Packages, die Sie in R gebrauchen werden, downloaden. Einen Mirror zu wählen, der topografisch in der Nähe liegt, minimiert daher den Rechenaufwand für das Netzwerk. Auf der sich öffnenden Seite klicken Sie auf denjenigen Downloadlink, der Ihrem Betriebssystem entspricht (vgl. Abb. 2).

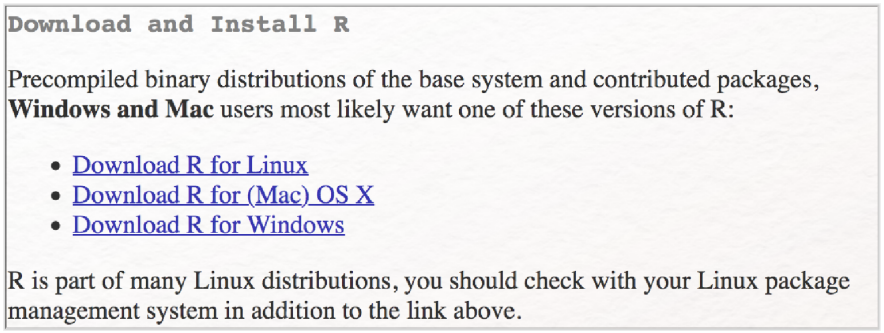

Abb. 2: Download von R für die verschiedenen Betriebssysteme

Auf der sich öffnenden Seite klicken Sie (als Mac-User*in) auf die PKG-Datei, die Ihrer Mac-OS-Version entspricht (z. B. R-3-6-0.pkg) bzw. (als Windows-User) auf „install R for the first time“ und dann direkt oben auf „Download R 3.6.0 for Windows“ (bzw. die jeweils aktuelle Version). Als Mac-User*in müssen Sie, damit Sie Stylo benutzen können, zusätzlich noch XQuartz installieren, klicken Sie dafür hier und laden sich die angezeigte DMG-Datei herunter. Sie müssen im Anschluss an die Installation von XQuartz Ihren Computer einmal neu starten, damit sich die Benutzeroberfläche von Stylo öffnen kann.

Öffnen Sie die heruntergeladenen Dateien (PKG für Mac, EXE für Windows) und folgen den Anweisungen des Installationsassistenten. Mac-User*innen machen dies ebenfalls mit der heruntergeladenen XQuartz-Installationsdatei (und starten anschließend ihren Mac neu). Sollten Sie beim Öffnen dieser Dateien eine Warnmeldung bekommen, kann es sein, dass Sie für das Programm eine Ausnahme in Ihren Sicherheitseinstellungen hinzufügen müssen. Wie das geht, können Sie in unseren Videos für Mac (forTEXT 2019a) und Windows (forTEXT 2019b) nachschauen. Da es insbesondere für Einsteiger*innen sehr viel leichter ist, R innerhalb einer **grafischen**

Benutzeroberfläche zu bedienen, die Ihnen neben der Console noch zahlreiche weitere Bedienoptionen bietet, installieren Sie sich nun noch **RStudio**, indem Sie diesem Link folgen, dem Downloadlink unterhalb der Open-Source-Desktopversion folgen und anschließend unter der Überschrift „Installers for Supported Platforms“ den Ihrem Betriebssystem entsprechenden Link klicken (z. B. „RStudio 1.2.1335 - Windows 7+“, vgl. Abb. 3). Die heruntergeladene Installationsdatei führen Sie wie schon zuvor bei R per Klick aus.

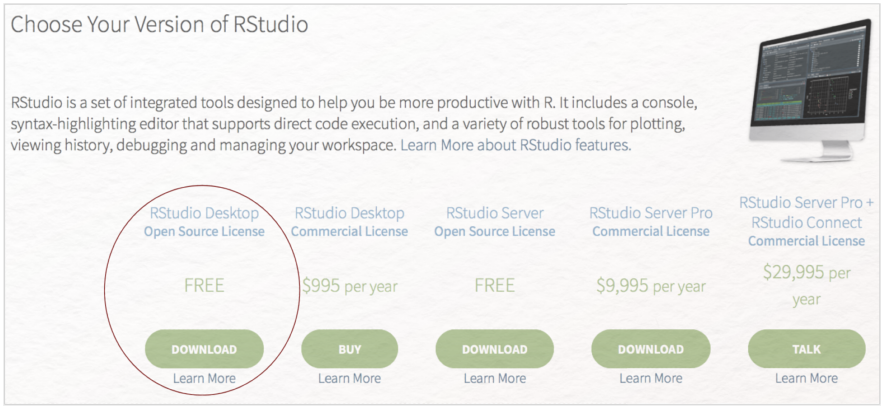

Abb. 3: Installation der Open-Source-Desktopversion von RStudio

In Ihrem Programmeordner sollten sich nun die Programme R und RStudio befinden. Öffnen Sie nun RStudio, indem sie auf das Programmicon klicken (vgl. Abb. 4). Die Benutzeroberfläche von RStudio öffnet sich im Anschluss (vgl. Abb. 5).

Abb. 4: Programmicon von RStudio

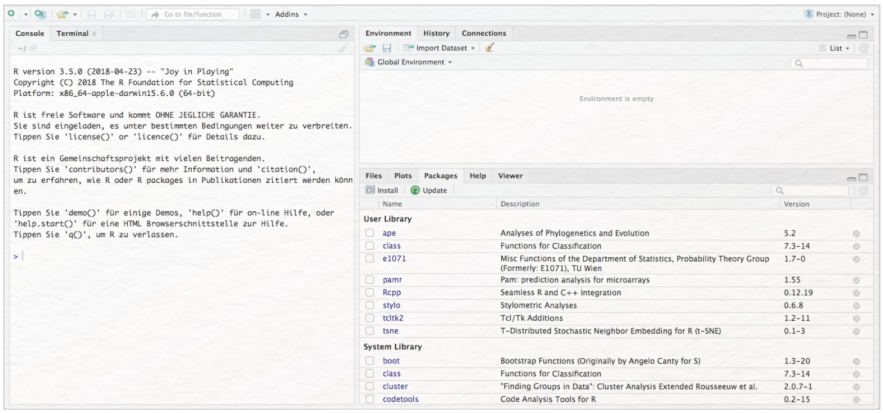

Abb. 5: Benutzeroberfläche von RStudio mit geöffnetem „Packages“-Reiter

Als letzten Schritt zur Vorbereitung der stilometrischen Analyse müssen Sie nun noch das **Stylo-Package in RStudio installieren**. Klicken Sie dazu im rechten unteren Panel auf den Reiter „Packages“ (vgl. Abb. 5) und dann auf den Button „Install“. Es öffnet sich ein Fenster wie in Abbildung 6.

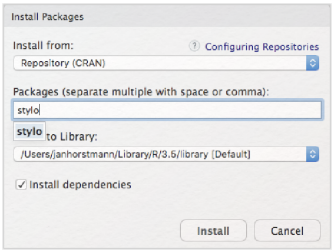

Abb. 6: Installation des Stylo-Packages in RStudio

Sie begegnen hier erneut dem CRAN-Repositorium, das Ihnen aus der Installation von R bereits vertraut ist. Aus diesem Repositorium werden Packages jeweils aktuell installiert. Tippen Sie in die „Packages"-Zeile „stylo" ein, wird Ihnen auch schon das entsprechende Package vorgeschlagen. Anschließend bestätigen Sie die Installation mit einem Klick auf „Install".
Nun erscheint im rechten unteren Panel von RStudio unter „User Library" das Package „stylo", das Sie mit einem Häkchen versehen und dadurch aktivieren (vgl. Abb. 7). Durch diese Aktion öffnet sich (sind Sie Mac-User*in) automatisch ebenfalls XQuartz.

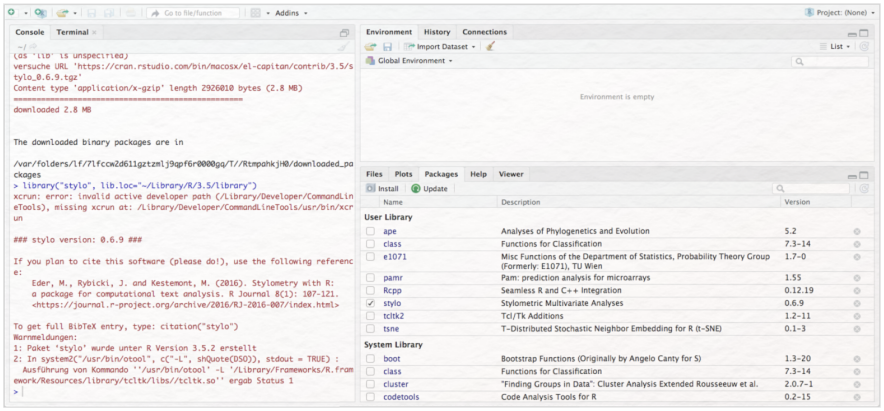

Abb. 7: Aktivierung des Stylo-Packages in RStudio

Sie haben nun alle Komponenten installiert, die Sie für eine erfolgreiche stilometrische Untersuchung benötigen: das Korpus, R, RStudio und das Stylo-Package. Sollte es im Folgenden zu Fehlern kommen, könnte es sein, dass Sie noch zusätzlich die Packages „tcltk2" und „ape" installieren müssen. Einige Systeme benötigen diese, um die Ergebnisvisualisierungen korrekt darstellen zu können. Diese Packages installieren Sie genau so wie das Stylo-Package zuvor.Um loslegen zu können, müssen Sie RStudio lediglich noch mitteilen, auf welchen Datensatz (d. h. unsere zuvor heruntergeladene Textsammlung) es zugreifen soll. Klicken Sie dazu in der oberen Menüleiste auf „Session", hovern Sie über „**Set Working Directory**" und klicken dann auf „Choose Directory…" (vgl. Abb. 8).

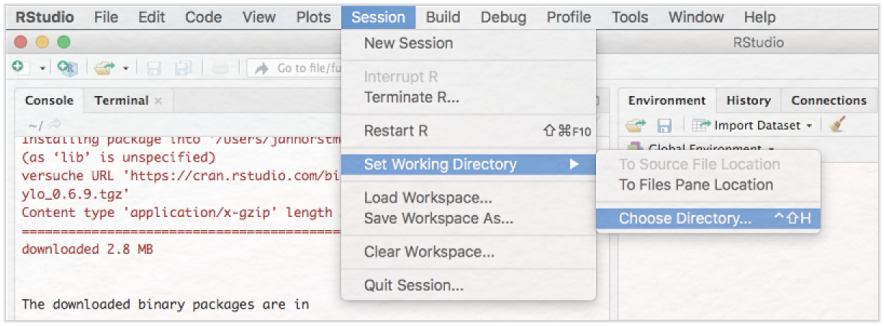

Abb. 8: Mit „Set Working Directory" legen Sie fest, auf welchen Datensatz RStudio in der jeweiligen Sitzung zugreifen soll

Navigieren Sie im sich öffnenden Fenster zum Ordner „68_german_novels-master" und bestätigen Ihre Auswahl mit einem Klick auf den Button „Open". Achtung: Wählen Sie hier *nicht* den Unterordner „corpus" aus! In der Console (linken Panel von RStudio) erscheint anschließend eine Meldung, die den Ort angeben wird, an dem Sie

das Korpus gespeichert haben: „>setwd("~/Documents/Stylo/68_german_novels-master")“. Nun haben Sie alle nötigen Vorarbeiten erfolgreich hinter sich gebracht und sind bereit für die digitale stilometrische Analyse der Texte.

3. Funktionen

Um die grafische Benutzeroberfläche von Stylo zu öffnen, klicken Sie in der Console hinter das blaue „>“-Zeichen, tippen „stylo()“ (ohne die Anführungszeichen) und drücken anschließend die Enter-Taste. Das „>“-Zeichen in der Console zeigt immer an, dass R bereit für eine Eingabe Ihrerseits ist. Solange das „>“-Zeichen nicht angezeigt wird, hat R die aktuell durchgeführte Aufgabe noch nicht abgeschlossen. Es öffnet sich ein Fenster (vgl. Abb. 9).

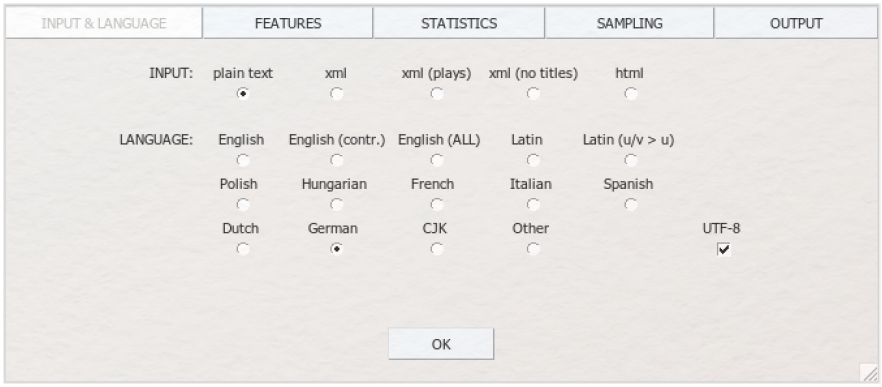

Abb. 9: Das grafische User-Interface von Stylo mit dem Modul INPUT & LANGUAGE

Die Benutzeroberfläche von Stylo gliedert sich in fünf Module: „INPUT & LANGUAGE“, „FEATURES (vgl. Feature)“, „STATISTICS“, „SAMPLING“ und „OUTPUT“. Wir werden im Folgenden die meisten Einstellungsmöglichkeiten der einzelnen Module kennenlernen. Tipp: Wenn Sie über die einzelnen Optionen hovern, die unter den jeweiligen Modulen aufgelistet werden, erscheinen kurze Erläuterungen der jeweiligen Funktion, sog. Tooltips. (Achtung: Ein Klick auf den „OK“-Button unten startet die stilometrische Analyse Ihrer Textsammlung. Dieser Button sollte erst betätigt werden, wenn die Einstellungen in sämtlichen fünf Modulen angepasst worden sind. Sämtliche Einstellungen eines Durchgangs werden in einer Datei namens „stylo_config.txt“ im ausgewählten Ordner (der Working Directory) gespeichert, sodass eine Analyse bei Bedarf exakt wiederholt werden kann. Diese Datei sorgt außerdem dafür, dass die Einstellungen im Stylo-GUI (vgl. GUI) bei jedem neuen Durchgang zu Beginn noch so sind wie beim jeweils vorherigen.)

Im **Modul INPUT & LANGUAGE** legen Sie fest, mit welcher Datengrundlage Stylo operieren soll. Die Dokumente in unserem Korpus sind im TXT-Format, deutschsprachig und UTF-8 codiert, Sie stellen folglich „plain text“ und „German“ ein (und aktivieren als Windows-Nutzer*in außerdem das Kästchen bei UTF-8).

Aufgabe 1: Was ist in den Spracheinstellungen der Unterschied zwischen „English“ und „English (ALL)“ und was bedeutet „Latin (u/v > u)“?

Im **Modul FEATURES** (vgl. Abb. 10) legen Sie die grundsätzlichen Einstellungen Ihres jeweiligen Analysedurchgangs fest.

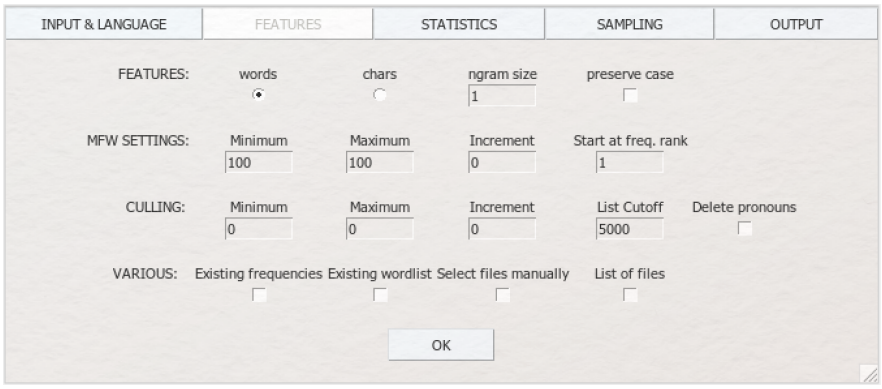

Abb. 10: Das grafische User-Interface von Stylo mit dem Modul FEATURES

Unter dem Unterpunkt „**FEATURES**“ können Sie entscheiden, ob Sie die stilometrische Untersuchung auf Wort- oder Buchstabenebene („words“ oder „chars“) ausführen wollen. Entscheidet man sich für die Buchstabenebene, sollte man auf jeden Fall Buchstaben_folgen_ untersuchen lassen; dies erreicht man mit der Festlegung einer N-Gramm-Größe. N-Gramms (vgl. N-gramm) funktionieren sowohl auf Buchstaben- wie auch auf Wortebene. Bei der Analyse auf Wortebene bietet sich jedoch eine sehr kleine N-Gramm-Größe an, da nur die Verteilung gleicher Wortkombinationen (und nicht gleicher Wörter) untersucht werden und man somit nur eine sehr limitierte Datengrundlage hätte. Die Untersuchung auf Buchstabenebene bietet sich vor allem an, wenn man ein „schmutziges“ Korpus hat, d. h. Texte mit orthografischen Fehlern, die z. B. im Digitalisierungsprozess durch die OCR entstehen können. Ist im Stylo-GUI die Option „preserve case“ aktiviert, werden Groß- und Kleinbuchstaben weiterhin voneinander getrennt; ist sie nicht aktiviert, werden für die Analyse alle Buchstaben in Kleinbuchstaben umgewandelt (vgl. Preprocessing). Es ist nicht nur sprachen- sondern auch korpusabhängig, welche Einstellung hier die besten Ergebnisse bringt.

Der Unterpunkt „**MFW SETTINGS**“ regelt die Einstellungen für die *most frequent words* (MFW). Auf Grundlage der Analyseergebnisse werden Sie vor allem in dieser Kategorier nach jedem Analysedurchgang neue Einstellungen vornehmen. Für den Anfang bietet es sich an, sowohl bei „Minimum“ als auch bei „Maximum“ die Voreinstellung beizubehalten und die Zahl 100 nach jedem Durchgang jeweils zu erhöhen. 100 bedeutet in diesem Fall, dass im Analysedurchgang die 100 häufigsten Wörter aus allen Texten miteinander verglichen werden. Die „Increment“-Einstellung (deutsch: Zuwachs) ist solange irrelevant, wie bei „Minimum“ und „Maximum“ die gleiche Zahl eingetragen wird. Wir kommen im Modul „STATISTICS“ im Zusammenhang mit dem sog. Bootstrap-Verfahren noch einmal auf diesen Parameter zurück. Im Feld „Start at freq. rank“ können Sie festlegen, ob und wenn ja wie viele der häufigsten der MFW aus der Analyse ausgeschlossen werden sollen. Lassen Sie die voreingestellte 1 in diesem Fall einfach stehen; sie bedeutet, dass alle 100 häufigsten Wörter berücksichtigt werden.

Im Unterpunkt „**CULLING**“ (deutsch: auslesen, aussondern) können Sie festlegen, ob textspezifisch besondere Wörter aus der Analyse ausgeschlossen werden sollen. Stylo berechnet Wortfrequenzlisten und auf Basis dieser Listen die für die jeweiligen Texte typischsten Wörter, d. h. alle Wörter, die nur in dem jeweiligen Text vorkommen. Von diesen Wörtern können Sie mittels Culling einen bestimmbaren Prozentsatz aus der Analyse ausschließen. Ein Culling-Wert von bspw. 20 bedeutet somit, dass nur Wörter, die in wenigstens 20 % der Texte vorkommen, analysiert werden, ein Wert von 100, dass nur Wörter berücksichtigt werden, die in sämtlichen Texten der Sammlung vorkommen. Auch hier ist die „Increment“-Einstellung insbesondere im Zusammenhang mit dem Bootstrap-Verfahren, auf das wir zurückkommen, relevant. Für den Anfang müssen Sie in dieser Lerneinheit beim Culling keine Einstellungen vornehmen.

Die Einstellungen bei „List Cutoff“ und „Delete pronouns“ stehen in keiner Verbindung zum Culling, auch wenn das Interface diesen Eindruck erweckt. Der Wert bei „List Cutoff“ bestimmt, ab welchem Punkt der Worthäufigkeitstabelle die jeweils seltenen Wörter der Texte nicht weiter berücksichtigt werden. Die voreingestellte 5000 können Sie in dieser Lerneinheit einfach stehen lassen. Die Option „Delete pronouns“ ist eher für erfahrene Nutzer*innen gedacht. Mit ihr werden Pronomen ausgeschlossen (wichtig ist dabei, dass die Sprache im „INPUT & LANGUAGE“-Modul korrekt eingestellt wurde).

Die Optionen unter „**VARIOUS**“ bleiben für diese Lerneinheit inaktiv und werden aus diesem Grund hier nicht en detail besprochen. Die Einstellungen erlauben Ihnen, Ergebnisse aus einer Analyse mit der Folgeanalyse zu verknüpfen, oder nur ausgewählte Texte aus Ihrem Korpus in der Analyse zu berücksichtigen. Ein umfangreiches, englischsprachiges Manual zu allen Funktionen in Stylo findet sich bereitgestellt durch die Computational Stylistics Group hier.

Aufgabe 2: Welche Vor- und Nachteile hat es, die sog. *case sensitivity* zu berücksichtigen, d. h. „preserve case“ bei einer Analyse zu aktivieren? Welche Wörter sind vermutlich in den meisten Texten die MFW?

Das Modul „**STATISTICS**“ (vgl. Abb. 11) bietet Ihnen die Möglichkeit, aus einer Vielzahl von Algorithmen (statistischen Verfahren und Distanzmaßen) auszuwählen.

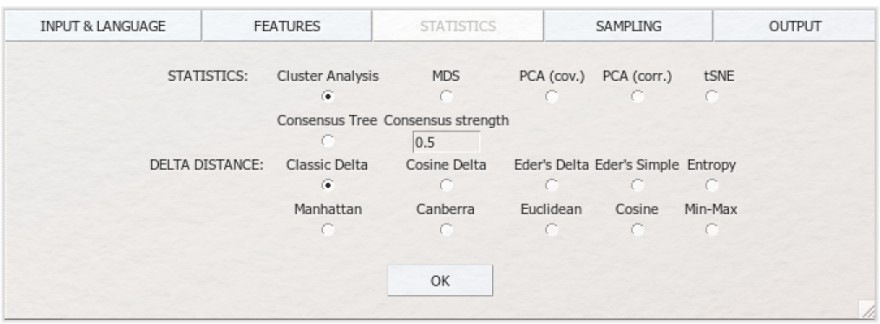

Abb. 11: Das grafische User-Interface von Stylo mit dem Modul STATISTICS

Das grundlegendste statistische Verfahren ist die *Cluster Analysis*, die das Korpus nach Ähnlichkeiten durchsucht und ähnliche Texte zusammen „clustert", d. h. eine Gruppe bilden lässt. Die Visualisierungsform dieser Analyse ist ein Dendrogramm. Die multidimensionale Skalierung (*MDS*), die Principal Component Analysis (*PCA*) (vgl. PCA) und die t-distributive stochastische Nachbareinbettung (t-distributive stochastic neighbour embedding; *tSNE*) sind Verfahren der Dimensionsreduktion und lassen ähnliche Texte innerhalb eines zweidimensionalen Koordinatensystems beieinander clustern. Der *Consensus Tree* (oder auch „Bootstrap Consensus Tree") ist eine runde Visualisierungsform, die mehrere Clusteranalysedurchgänge mit unterschiedlich vielen MFW bzw. Culling-Parametern in einer Ergebnisvisualisierung vereinigt. Dieses Verfahren wird daher auch „Bootstrap" genannt (was wörtlich mit „Stiefelriemen" übersetzt werden kann - wie die Schnürsenkel eines Schuhs werden die einzelnen Analysedurchläufe ineinander gewebt und am Ende festgezogen). Ähnliche Texte werden hier an einem Ast clusternd dargestellt. Wenn Sie „Consensus Tree" auswählen, müssen Sie im Modul FEATURES beim MFW-Setting zusätzlich angeben, mit wie vielen MFW die Analyse beginnen („Minimum"), mit wie vielen sie aufhören soll („Maximum") und wie viele Wörter bei jeder einzelnen Analyse hinzukommen sollen („Increment"). Ebenso erfordert dieses Verfahren unterschiedliche Minimum- und Maximumangaben beim Culling, insofern Sie die Cullingfunktion nutzen wollen (ansonsten lassen Sie dort einfach überall die „0" stehen). Für den Consensus Tree (und nur für ihn) muss zusätzliche eine „Consensus strength" angegeben werden. Hier können Sie eine Zahl zwischen 0,4 und 1 eingeben. 0,4 bedeutet, dass Verbindungen in mindestens 40 % der Analysedurchgänge gegeben sein müssen, um auch in der Consensus-Tree-Visualisierung als Verbindung angezeigt zu werden. 1 bedeutet entsprechend, dass dies in 100 % der Durchläufe gegeben sein muss. Die voreingestellte 0,5 können Sie hier zunächst ebenfalls stehen lassen; diese Zahl spielt solange keine Rolle, wie eine der anderen Statistiken ausgewählt ist. Wir konzentrieren uns hier auf die Clusteranalyse.

Hinter den unterschiedlichen Distanzmaßen, die neben dem Menüpunkt „DELTA DISTANCE" aufgelistet werden, verstecken sich die Algorithmen, mit denen die Distanzen, d. h. im Grunde die Ähnlichkeiten zwischen den Texten, berechnet werden. Diese Sektion bildet das mathematische Herzstück von Stylo, soll in dieser einführenden Lerneinheit jedoch nicht en detail besprochen werden. Je nach Sprache und Einstellung (Wörter oder N-Gramms) bieten sich unterschiedliche Distanzmaße an. Für unsere deutschsprachige Textsammlung sind insbesondere die Maße „Classic Delta" (auf John Burrows zurückgehend) und „Cosine Delta" (an der Universität Würzburg entwickelt) geeignet. Euclidean und Manhattan Distance sollten im Zusammenhang mit der Analyse von Worthäufigkeiten eher vermieden werden; Canberra Distance bietet sich insbesondere für lateinische Texte an, etc. Eine genauere Beschreibung der einzelnen Distanzmaße finden Sie im Stylo-Handbuch (S. 15-17).

Das Modul „**SAMPLING**" (vgl. Abb. 12) bietet Ihnen die Möglichkeit, die Texte in Ihrer Sammlung in gleich große Segmente zu zerschneiden.

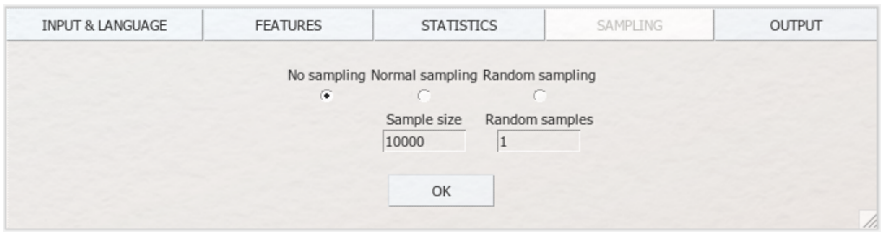

Abb. 12: Das grafische User-Interface von Stylo mit dem Modul SAMPLING

Diese Segmentierung, die aus literaturwissenschaftlicher Sicht erst einmal blasphemisch anmutet, kann die Ergebnisse der stilometrischen Analyse durchaus verbessern, insbesondere wenn man eine Sammlung untersucht, die Texte mit unterschiedlicher Länge enthält. Um statistische Verteilungen von häufigen Wörtern in Texten zu analysieren, ist es immer sinnvoll, ähnlich lange Texte (bzw. Textsegmente) miteinander zu vergleichen. Im Feld „Sample size" wird bei aktivierter Funktion „Normal sampling" angegeben, wie groß die Segmente sein sollen - angegeben in Wortmenge. Noch experimenteller wird es mit der Option „Random sampling": ein Verfahren, bei dem die Texte nicht nur in gleich große Segmente zerschnitten, sondern die Wörter aus den Texten auch noch zufällig ausgewählt und zu einem Segment zusammengesetzt werden (dieses Verfahren ist unter dem Namen „bag of words" bekannt). Auf diese Weise können gelegentlich noch bessere Zuschreibungen erreicht werden. Bei dieser Option geben Sie im Feld „Random samples" die Menge der Segmente an, in die ein jeder Text in Ihrer Sammlung zerschnitten werden soll. Keine Sorge: Ihre Texte werden nicht wirklich zerstört und die Dateien in Ihrem Ordner verbleiben in ihrem ursprünglichen Zustand, Sampling ist eine rein rechnerische Aktion. Für unsere Beispielanalyse lassen Sie die voreingestellte Option „No sampling" aktiviert.
Nun haben Sie die wichtigsten Einstellungen von Stylo (INPUT & LANGUAGE, FEATURES, STATISTICS und SAMPLING) kennengelernt, die Sie für Ihre erste stilometrische Analyse brauchen. Sie müssen nur noch bestimmen, wie Ihnen die Ergebnisse angezeigt und wie sie gespeichert werden sollen. Das geschieht im fünften Modul „**OUTPUT**" (vgl. Abb. 13.).

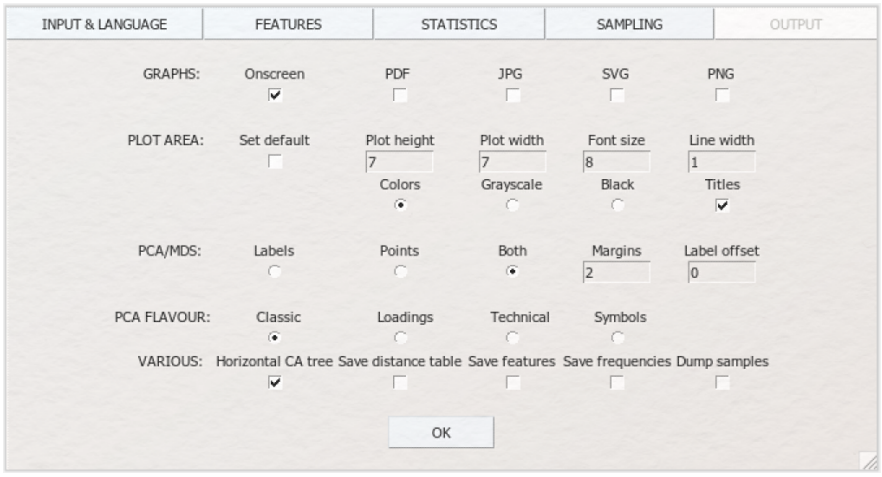

Abb. 13: Das grafische User-Interface von Stylo mit dem Modul OUTPUT

Unter „GRAPHS“ können Sie auswählen, in welchen Formaten die Ergebnisvisualisierung gespeichert werden soll. „Onscreen“ zeigt Ihnen den Graphen lediglich programmintern im RStudio-Modul rechts unten unter „Plots“ an. Sämtliche anderen Formate werden als Dateien in Ihrem „Working Directory“-Ordner auf Ihrem Computer gespeichert, in unserem Fall also unter „68_german_novels-master“. Insbesondere bei iterativen Analysen mit unterschiedlichen MFW- oder Culling-Einstellungen kann es dabei schnell zu einer großen Menge an gespeicherten Dateien kommen; wählen Sie die Einstellungen daher mit Bedacht. Tipp: Wählen Sie zunächst „Onscreen“, inspizieren Sie die Visualisierung nach der Analyse und wiederholen Sie den Durchgang mit den gleichen Einstellungen dann noch einmal, um die gleiche Visualisierung als Datei im gewünschten Format zu speichern.
Sie finden im Modul OUTPUT noch etliche weitere Einstellungsmöglichkeiten, um die zu erstellenden Visualisierungen nach Ihren Vorstellungen zu modifizieren, die wir hier nicht detailliert vorstellen werden. Die Tooltips und das Stylo-Handbuch geben Ihnen hier bei Bedarf weitere Hinweise. Viele der OUTPUT-Einstellungen beziehen sich auf spezifische statistische Verfahren. Die Einstellungen in den entsprechenden Zeilen bei OUTPUT sind daher so lange irrelevant, wie nicht das entsprechende statistische Verfahren unter STATISTICS ausgewählt wurde.

Aufgabe 3: Es ist soweit. Klicken Sie nun auf den „OK“-Button und starten damit Ihre erste stilometrische Analyse. Wenn Sie anschließend die Konsole in RStudio beobachten, können Sie verfolgen, wie der Algorithmus arbeitet. Wenn ein „>“-Symbol erscheint, ist die Analyse abgeschlossen. Dies kann abhängig von der Rechenleistung Ihres PCs durchaus eine Weile dauern. Unter „Plots“ rechts von der Konsole wird Ihnen anschließend ein Dendrogramm angezeigt, das Sie mit einem Klick auf „Zoom“ vergrößern und inspizieren können (*Hinweis*: Sollte es passieren, dass Ihnen ein leeres Dendrogramm angezeigt wird (d. h. nur Äste und keine Schrift), liegt das höchstwahrscheinlich daran, dass Sie die Schriftart „Arial“ nicht installiert und aktiviert haben. Diese Schriftart gehört i. d. R. zum Standardset der aktivierten Schriftarten, im neuen Betriebssystem von MAC OS (Mojave) ist dies aber bspw. nicht der Fall. Schriftarten können Sie leicht im Internet finden und im Format TTF downloaden). Wie interpretieren Sie das Ergebnis?

Aufgabe 4: Experimentieren Sie mit den besprochenen Einstellungsmöglichkeiten. Wie können Sie beim Testen unterschiedlicher MFW-Zahlen Zeit sparen? Welche Einstellungen scheinen Ihnen für die gegebene Textsammlung gut geeignet?

Aufgabe 5: Ihnen wird bei der Bearbeitung der vorigen Aufgabe immer wieder der anonyme Text *Schwester Monika* ins Auge gefallen sein. Können Sie diesen Text einem der anderen Autoren zuordnen? Wenn nicht, wo könnten die Schwierigkeiten liegen?

4. Lösungen zu den Beispielaufgaben

Aufgabe 1: Was ist in den Spracheinstellungen der Unterschied zwischen „English“ und „English (ALL)“ und was bedeutet „Latin (u/v > u)“?

Die Tooltips (die Sie durch Hovern erhalten) geben Ihnen die Antworten: Bei der Einstellung „English (ALL)“ werden im Gegensatz zur Einstellung „English“ Verkürzungen wie „don’t“ und zusammengesetzte Wörter wie „light-headed“ nicht aufgesplittet. Im Modus „Latin (u/v > u)“ werden alle Us und alle Vs als U gewertet.

Aufgabe 2: Welche Vor- und Nachteile hat es, die sog. *case sensitivity* zu berücksichtigen, d. h. „preserve case“ bei einer Analyse zu aktivieren? Welche Wörter sind vermutlich in den meisten Texten die MFW?

Eine aktivierte „preserve case“-Option ist besonders für deutschsprachige Texte relevant. Sie verhindert, dass alle Buchstaben in Kleinbuchstaben umgewandelt werden. Der Vorteil davon ist beispielsweise, dass Wörter wie „spinne“ bzw. „Spinne“ nicht als gleiches Wort gewertet werden, was je nach Textsammlung einen wichtigen Vorteil darstellen kann. Ein entscheidender Nachteil ist jedoch, dass Worte am Satzanfang mit einem großen Buchstaben beginnen und daher in diesen Fällen als anderes Wort klassifiziert werden, als wenn sie innerhalb des Satzes auftauchen. Welche Einstellung für Ihre jeweilige Textsammlung sinnvoller ist, können Sie durch Ausprobieren herausfinden. Die MFW in Texten sind eigentlich immer Funktionswörter wie „und“, „aber“, „denn“, „ich“, „er“, „sie“, „es“ etc. In unserem Beispielkorpus sind das in absteigender Häufigkeit „und“, „die“, „der“, „zu“, „in“, „er“, etc.

Aufgabe 3: Es ist soweit. Klicken Sie nun auf den „OK“-Button und starten damit Ihre erste stilometrische Analyse. Wenn Sie anschließend die Konsole in RStudio beobachten, können Sie verfolgen, wie der Algorithmus arbeitet. Wenn ein „>“-Symbol erscheint, ist die Analyse abgeschlossen. Dies kann abhängig von der Rechenleistung Ihres PCs durchaus eine Weile dauern. Unter „Plots“ rechts von der Konsole wird Ihnen anschließend ein Dendrogramm angezeigt, das Sie mit einem Klick auf „Zoom“ vergrößern und inspizieren können *Hinweis*: Sollte es passieren, dass Ihnen ein leeres Dendrogramm angezeigt wird (d. h. nur Äste und keine Schrift), liegt das höchstwahrscheinlich daran, dass Sie die Schriftart „Arial“ nicht installiert und aktiviert haben. Diese Schriftart gehört i. d. R. zum Standardset der aktivierten Schriftarten, im neuen Betriebssystem von MAC OS (Mojave) ist dies aber bspw. nicht der Fall. Schriftarten können Sie leicht im Internet finden und downloaden (im Format TTF). Wie interpretieren Sie das Ergebnis? Die erste Analyse erstellt Ihnen das Dendrogramm in Abbildung 14.

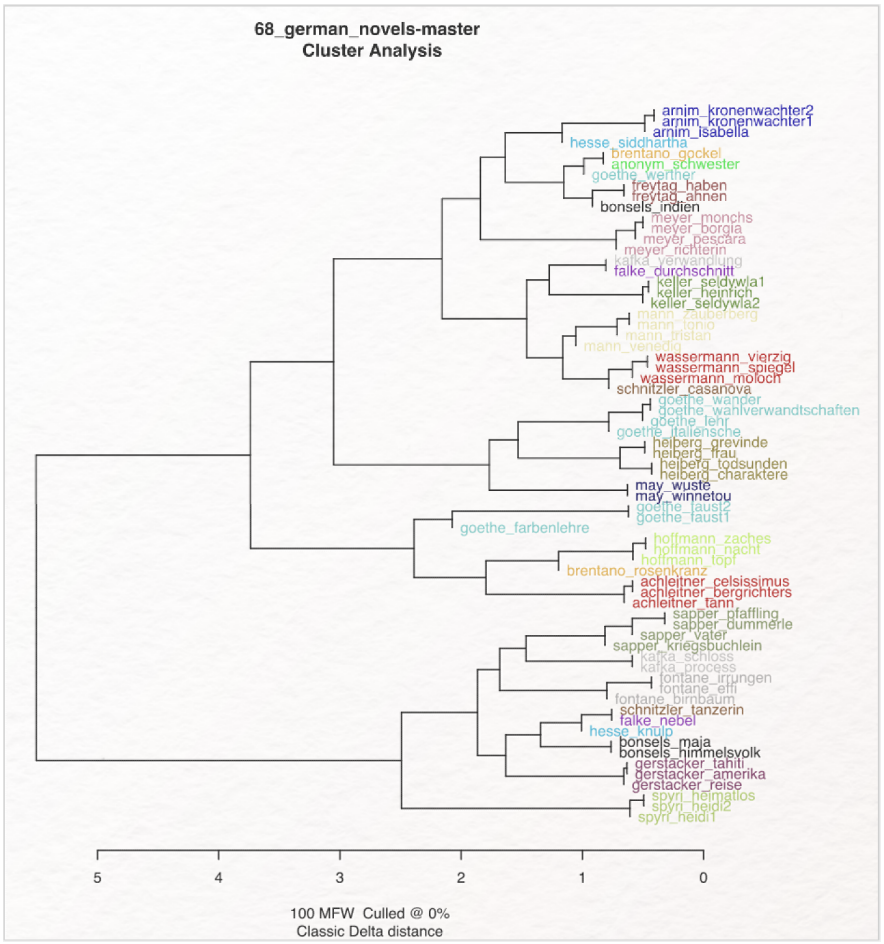

Abb. 14: Clusteranalyse von 67 deutschsprachigen Texten mit den 100 häufigsten Wörtern (Classic Delta Distance)

Texte am gleichen Ast sind sich (im Sinne der Stilometrie) stilistisch ähnlich, je mehr Gabelungen zwischen zwei Texten liegen, desto unähnlicher sind sie sich. Die vertikale Anordnung spielt dabei keine Rolle (es wäre eine Fehlinterpretation, die Texte von Spyri und Arnim für maximal unterschiedlich zu halten, weil sie am weitesten auseinander liegen - wenngleich sie sich auch alles andere als ähnlich sind). Die Zahlen in der horizontalen Leiste unten geben die Werte des jeweils ausgewählten Distanzmaßes an (hier Classic Delta Distance). Im Zuge einer Autorschaftsattribution ist es sinnvoll, wenn Texte des gleichen Autors am gleichen Ast clustern. Wir sehen in diesem ersten Beispiel, dass das bei vielen Autor*innen schon gut funktioniert. Das Vorkommen von Goethe,

Kafka, Brentano, Hesse, Bonsels und Schnitzler an unterschiedlichen nicht zusammengehörenden Ästen zeigt, dass sich die Texte dieser Autoren noch nicht einheitlich klassifizieren lassen. Es kann einem zudem aufstoßen, dass Goethes *Farbenlehre* mit den *Faust*-Dramen und nicht mit den Prosatexten Goethes zusammen clustert, ist doch eine stilometrische Analyse auch zur Gattungsidentifikation geeignet. Daraus kann geschlossen werden, dass die Einstellungen für das Beispielkorpus noch nicht ideal sind.

Aufgabe 4: Experimentieren Sie mit den besprochenen Einstellungsmöglichkeiten. Wie können Sie beim Testen unterschiedlicher MFW-Zahlen Zeit sparen? Welche Einstellungen scheinen Ihnen für die gegebene Textsammlung gut geeignet?

Mit den „Minimum"-, „Maximum"- und „Increment"-Einstellungen lassen sich mehrere Analysen in einem Durchgang durchführen. Sie können somit z. B. bei „Minimum" 100, bei „Maximum" 3000 und bei „Increment" 100 einstellen und die erstellten Visualisierungen anschließend vergleichend analysieren. Im „Plots"-Panel von RStudio machen Sie das über die Pfeiltasten. Wenn Sie unterschiedliche MFW-Einstellungen in einer einzigen Visualisierung zusammenfassen möchten, wählen Sie im STATISTICS-Modul „Consensus Tree" aus und stellen ebenfalls in den FEATURES unterschiedliche Zahlen bei „Maximum" und „Minimum" und ein bestimmtes „Increment" ein.

Was „gute" Einstellungen sind, ist sicherlich subjektiv. Wir interpretieren Einstellungen als „gut", wenn möglichst alle Texte einer Autorin/eines Autors am gleichen Ast clustern. Für die Clusteranalyse erreicht man mit 1500 Wörtern ein recht überzeugendes Ergebnis (vgl. Abb. 15). Ab etwa 2500 MFW lässt sich beobachten, dass Thomas Manns *Zauberberg* nicht mehr mit den anderen Mann-Texten clustert, die Einstellung eignet sich für die vorliegende Textsammlung daher nicht mehr so gut.

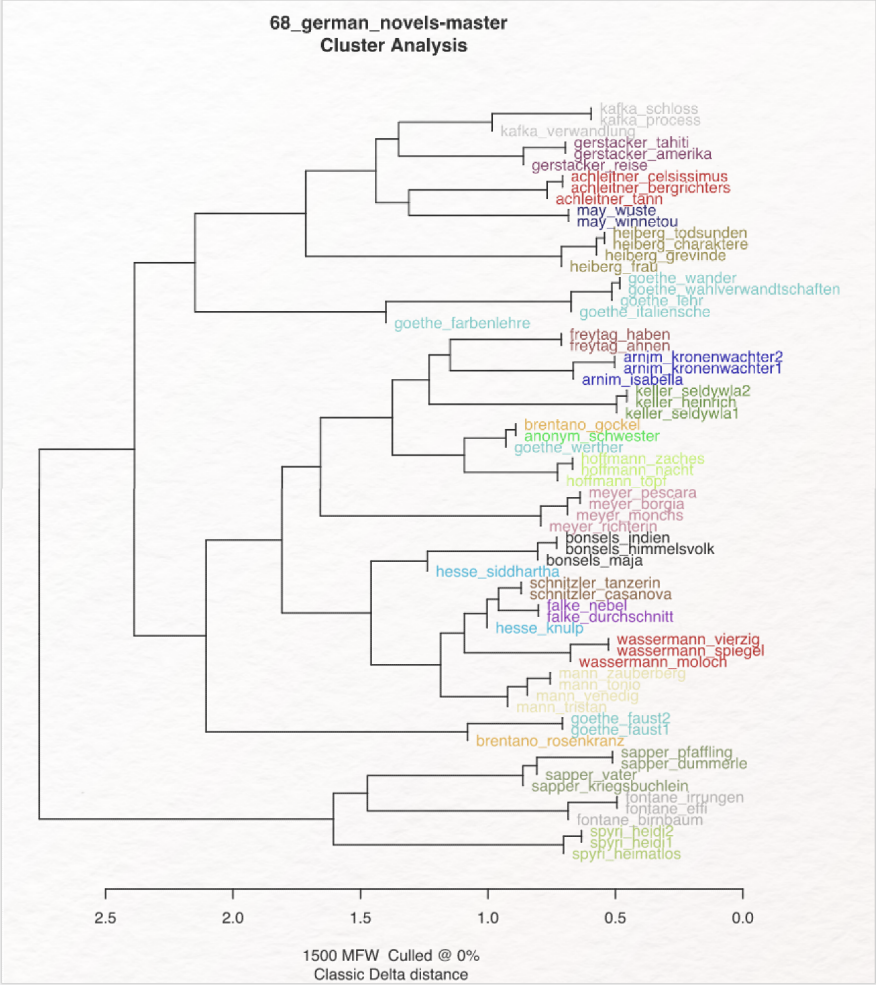

Abb. 15: Clusteranalyse von 67 deutschsprachigen Texten mit den 1500 häufigsten Wörtern (Classic Delta Distance)

Der Bootstrap Consensus Tree sieht bei 1000-2500 MFW ebenfalls überzeugend aus (vgl. Abb. 16). Bei dieser Visualisierung gilt ebenfalls, dass die Lokalisierung auf der Fläche (oben/unten bzw. rechts/links) keine semantische Aussagekraft hat; es kommt lediglich darauf an, ob Texte eines Autors bzw. einer Gattung gemeinsam

an einem Ast clustern. Wir sehen, dass mit einer Einstellung von 1000-2500 MFW beinahe sämtliche Texte Autor*innencluster bilden. Goethes *Faust*-Dramen clustern zudem nicht mehr zusammen mit der in Prosa geschriebenen *Farbenlehre*. Die zwei Texte Brentanos clustern bei fast sämtlichen Einstellungen nicht zusammen. Der Grund dafür ist die Gattungsdifferenz, die hier stärker zu Buche schlägt als etwaige Autorenstilmerkmale: Die Gedichtsammlung *Romanzen vom Rosenkranz* clustert nachvollziehbar zusammen mit den einzigen (ebenfalls in Versen geschriebenen) dramatischen Texten des Korpus (Goethes *Faust I* & *II*) und nicht zusammen mit anderen Prosatexten.

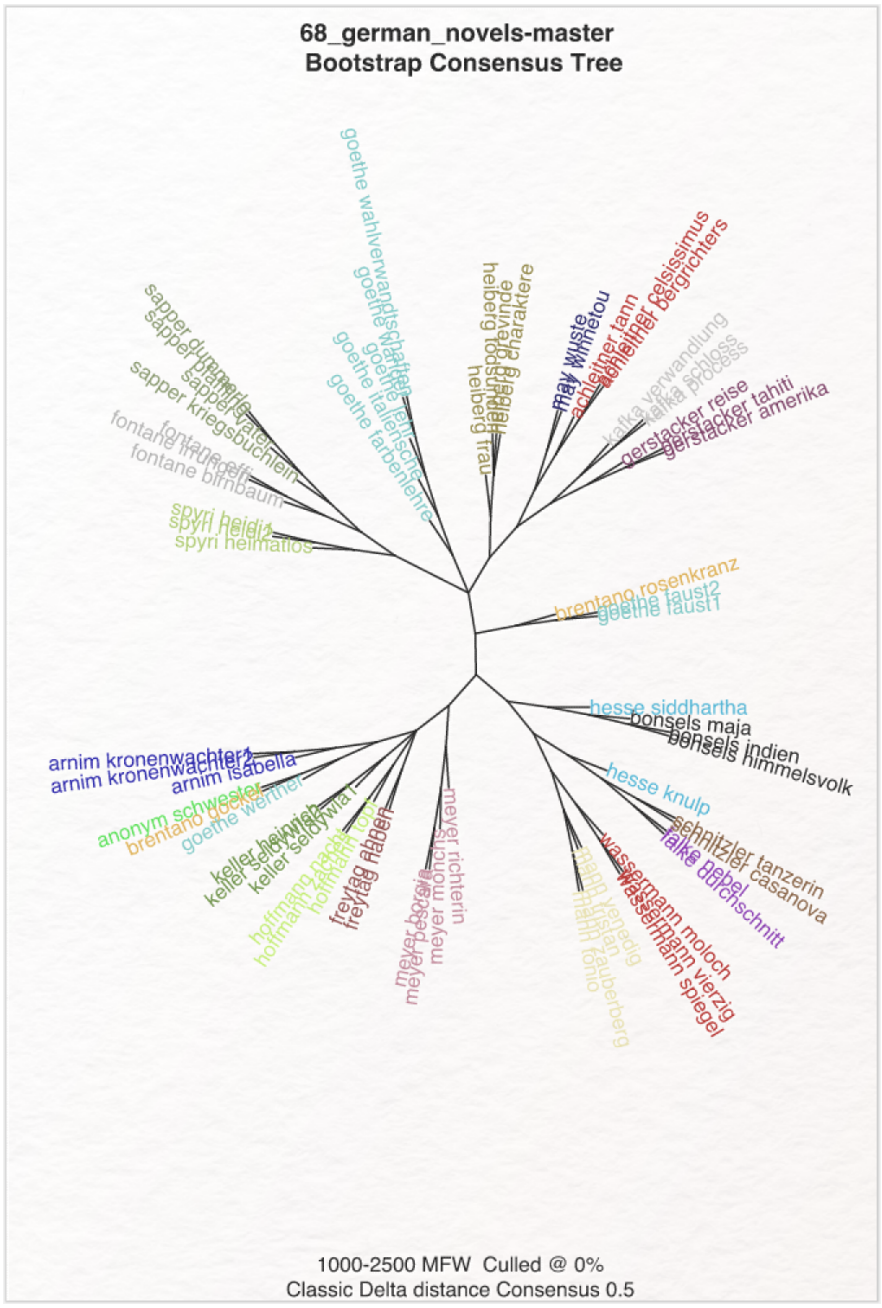

Abb. 16: Bootstrap Consensus Tree von 67 deutschsprachigen Texten mit 1000-2500 der häufigsten Wörter (Classic Delta Distance)

Aufgabe 5: Ihnen wird bei der Bearbeitung der vorigen Aufgabe immer wieder der anonyme Text *Schwester Monika* ins Auge gefallen sein. Können Sie diesen Text einem der anderen Autoren zuordnen? Wenn nicht, wo könnten die Schwierigkeiten liegen?

Der Text *Schwester Monika* (1815) clustert in den meisten Einstellungen sowohl zusammen mit Brentanos Märchen *Gockel, Hinkel und Gackeleia* (1838) als auch mit Goethes *Die Leiden des jungen Werthers* (1774), nicht jedoch zusammen mit den Texten E. T. A. Hoffmanns, dem der Text gelegentlich zugeschrieben wird. Vermutlich wurde *Schwester Monika* aber weder von Goethe noch von Brentano geschrieben, sondern von einer Autorin/einem Autor,

der nicht in der Sammlung enthalten ist. Die häufige Zuordnung zu Brentano und Goethe in der vorliegenden Textsammlung liegt daran, dass in *Schwester Monika* etliche literarische Zitate enthalten sind, die der Idee eines „Autorenstils“ selbstverständlich entgegenlaufen. Auch mit automatisierter Unterstützung konnte daher das Rätsel um diesen anonym erschienenen Text bislang nicht gelüftet werden. Sie sind nach dieser Lerneinheit fähig, sich mit unterschiedlichen Textsammlungen und verschiedenen Einstellungen in Stylo an dieser Debatte zu beteiligen.

Externe und weiterführende Links

- Github Computationsal Sytlistics Group, 67 Texte: https://web.archive.org/save/https://github.com/computationalstylistics/68_german_novels (Letzter Zugriff: 20.02.2024)
- Stylo-Handbuch: https://web.archive.org/save/https://github.com/computationalstylistics/stylo_howto/blob/master/stylo_howto.pdf (Letzter Zugriff: 20.02.2024)
- R Project: https://web.archive.org/save/https://www.r-project.org (Letzter Zugriff: 20.02.2024)
- Tutorial: Sicherheitsausnahme für Internetprogramme Hinzufügen (Windows): https://doi.org/10.5281/zenodo.11074222 (Letzter Zugriff: 20.02.2024)
- Tutorial: Sicherheitsausnahme für Internetprogramme Hinzufügen (Mac): https://doi.org/10.5281/zenodo.11074232 (Letzter Zugriff: 20.02.2024)

Bibliographie

forTEXT. 2019a. Tutorial: Sicherheitsaufnahme für Internetprogramme Hinzufügen (Mac). 19. Januar. https://doi.org/10.5281/zenodo.11074232.

———. 2019b. Tutorial: Sicherheitsausnahme für Internetprogramme Hinzufügen (Windows). 25. Januar. https://doi.org/10.5281/zenodo.11074222.

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

———. 2024b. Toolbeitrag: Stylo. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 1. Stilometrie (26. Februar). doi: 10.48694/fortext.3770, https://fortext.net/tools/tools/stylo.

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.

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.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktions-

einheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

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.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

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.

N-gramm Unter N-Gramm versteht man in der Linguistik eine Sequenz von N aufeinanderfolgenden Fragmenten/Einheiten in einem Text. So gibt es beispielsweise Bigramme, Trigramme etc. Diese Fragmente können Buchstaben oder Phoneme sein. Der Satz „Marie erforscht Literatur digital“ kann zum Beispiel folgendermaßen in Bigramme, drei wortbasierte N-gramme mit je zwei Wörtern, aufgeteilt werden: „Marie erforscht“, „erforscht Literatur“ und „Literatur digital“.

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.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

PCA PCA steht für *Principal Component Analysis* . Die Hauptkomponentenanalyse ist ein komplexes, statistisches Verfahren zur Reduktion und Veranschaulichung umfangreicher Datensätze.

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.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

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.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

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).

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

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.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

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.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.