EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT

Lehrstuhl für Romanische Philologie

Quantitative Textanalyse: Stilometrie

Autoren und Gattungen

07.06.2017, XVII.Tagung der DGAVL in Bochum,
Workshop / Digital Humanities in der Literaturwissenschaft
PD Dr. Nanette Rißler-Pipka

Quantitative Textanalyse

Funktioniert auch ohne Comupter:
Fokus liegt nicht auf dem Besonderen (qualitativen), sondern das Besondere ergibt sich aus der Häufigkeit (Quantität)

Daten + Schnelligkeit

Moretti: „today, we can replicate in a few minutes investigations that took a giant like Leo Spitzer months and years of work" TRUE?

Von Spitzer zu Moretti…

Stilistik

- Untersuchung von sprachlichen Besonderheiten, Unterschieden, Merkmalen
- Kann linguistisch oder literaturwissenschaftlich sein
- In der Romanistik weit über das Fach hinausgehende Tradition durch die Forschungen Leo Spitzers, aber auch E. R. Curtius‘ und E. Auerbachs
- CLOSE READING

Stilometrie

- Quantitative Textanalyse seit 19.Jh.
- Autorschaftstattribution in forensischer Linguistik: digitaler Fingerabdruck – und Literaturwissenschaft
- Statistische Verfahren: Verteilung der „Most Frequent Words“, um daraus Distanzwerte zwischen Texten zu erkennen
- DISTANT READING

Stil + Metrik

Stilelemente lassen sich zählen: Vokabular, grammatikalische Einheiten, Satzzeichen, Buchstaben

Dennoch schwierig: Was macht eigentlich den Stil eines Autors aus?

Was macht eigentlich den Stil eines Autors aus?

- Vokabular?
  - Vorliebe für bestimmte Begriffe: z.B. Angebetete, Geliebte, Freundin (auch Kontext und Zeit-abhängig)
- Phrasen?
  - Vorliebe für ganze Phrase: z.B. Picasso „es bleibt nichts übrig als“ („no hay más hacer que“)
- Satzeichen?
  - Verzicht auf Interpunktion
- Worthäufigkeiten?
  - Wie oft werden diejenigen Worte im Vergleich zu anderen Texten verwendet, die in jedem Text vorhanden sind – unabhängig vom Thema (Most frequent words, Funktionswörter)

Entscheidung: Was wird gezählt, gemessen, verglichen?

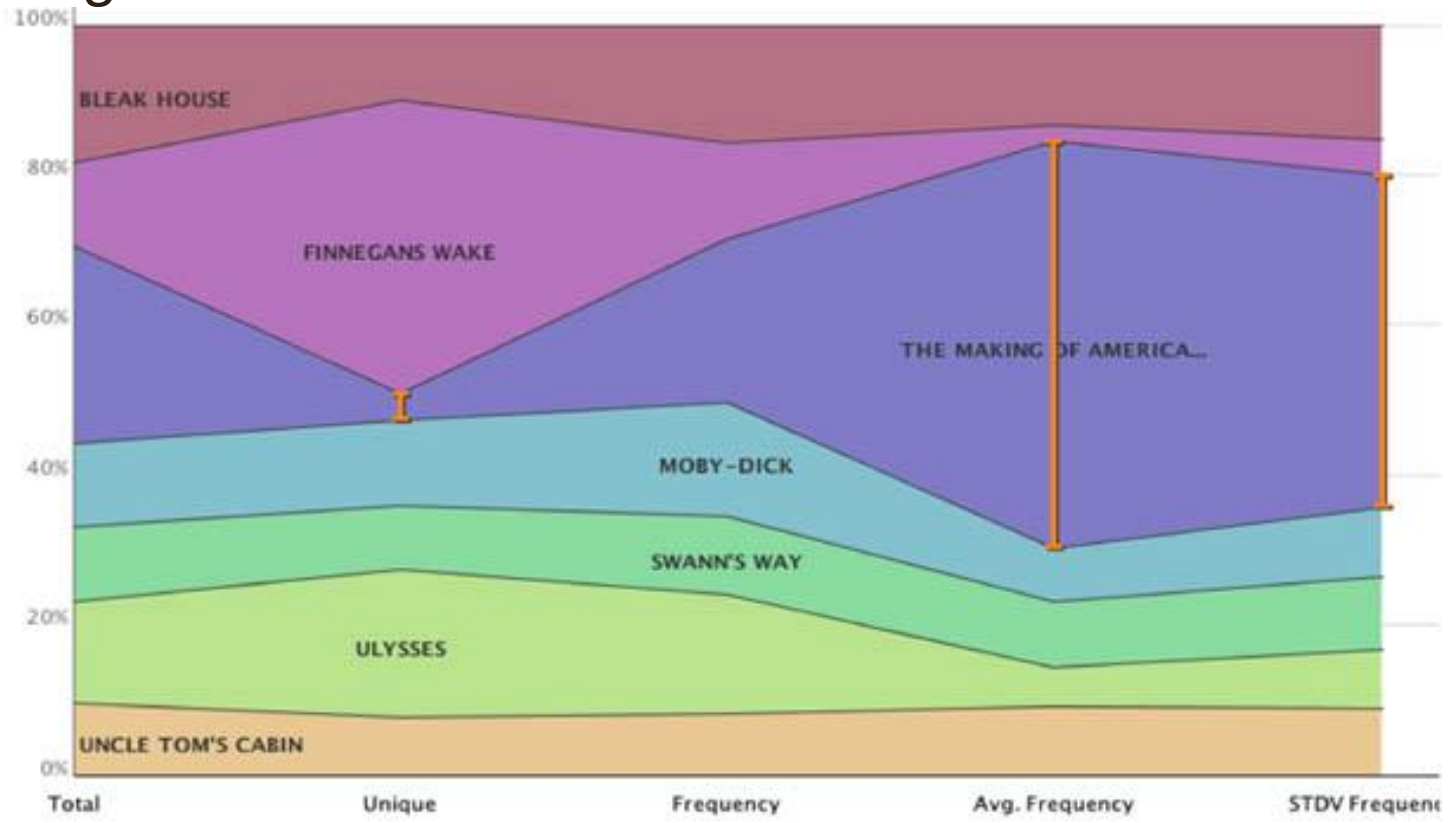
- Fragestellung: Autorschaftsattribution
  1. Anzahl von Kandidaten für einen unbekannten Text (classification): anhand von einem Set bekannter Texte der Kandidaten können deren Stilmerkmale per Machine Learning ermittelt werden, d.h. features werden anhand des Training-Sets festgelegt und dann am unbekannten Text getestet (welchem dieser Text am nächsten kommt, gilt als wahrscheinlicher Autor)
  2. Anzahl von unbekannten Texten (clustering): kein Training-Set möglich, da der Autor außerhalb des Korpus liegen kann Daher Clustering durch Distanzmaß – aufgrund der Ähnlichkeiten zwischen den Texten: Frage welches Distanzmaß?

Vgl. Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015).

Entscheidung: Was wird gezählt, gemessen, verglichen?

- Fragestellung: Stilbesonderheit (Outlier)
  1. Warum lassen sich manche Autoren in keine Epoche und zu keiner künstlerischen Bewegung gruppieren: Sie gelten als „Outlier“, bilden ihren eigenen Stil und grenzen sich bewusst ab
  2. Gerade Texte der Avantgarde-Literatur zeichnen sich durch Besonderheiten aus: Neologismen (die nur in diesem Text vorkommen), Wörter, die nur einmal im Werk verwendet werden, Wiederholung (als Stilmittel), Satzzeichen-Verzicht, Zahlengebrauch, etc.

Tanya Clement: Table of word frequencies from texts comparable in size or composition date to Stein's *The Making of Americans*

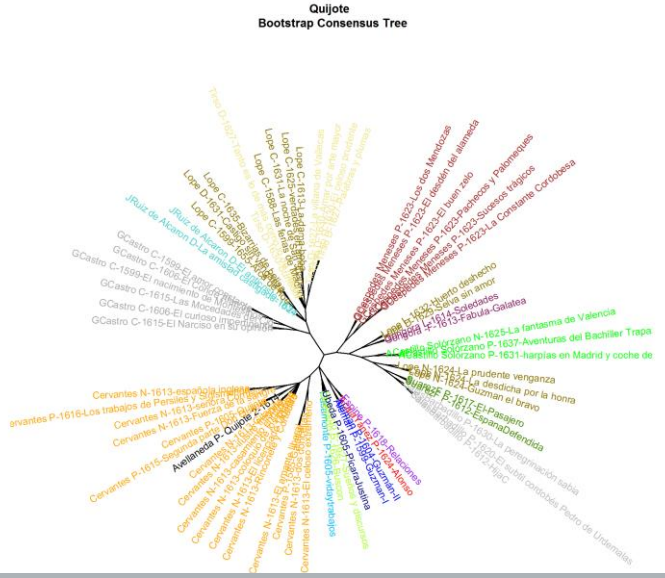


Tanya Clement: https://dlsanthology.commons.mla.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/

Tanya Clement: Table of word frequencies from texts comparable in size or composition date to Stein's *The Making of Americans*

- Problem:
- Textauswahl: Dickens: *The Bleak House* (1852); Joyce: *Ullysses* (1922); Joyce: *Finnegans Wake* (1939); Melville: *Moby Dick* (1851); Proust: *Swann's Way* (1913); Stowe: *Uncle Tom's Cabin* (1852); Stein: *Making of Americans* (1925)
- Wenig Texte, zeitliche Differenz, stilistische Differenz und Übersetzung – Vergleichbarkeit?

Tanya Clement: https://dlsanthology.commons.mla.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/



Stilometrie

Funktionsweise

Anwendungsbeispiele

Autorschaftsattribution und Stilfragen

Stilometrie: Funktionsweise

- Fragestellung: Autorschaftsattribution
  1. 1963/64 finden die Statistiker Frederic Mosteller und David Wallace anhand der *Federalist Papers* heraus, dass die Messung der „Most frequent Words“ das entscheidende Kriterium für die Autorschaftsattribution ist.
  2. Von allen vorgestellten *features* (=computermessbaren Stilelementen) ist das Einfachste (Berechnung der Verteilung der MFW pro Dokument) zugleich das Zuverlässigste.
  3. Für ein Textkorpus werden zunächst die gesamten MFW ausgezählt (wordlist), dann die Häufung pro Dokument (frequencies) und daraus lässt sich dann mithilfe eines Distanzmaßes der räumliche Abstand zwischen den Texten berechnen (distance table).

Stilometrie: Beispiel *Federalist Papers*

- *Federalist Papers:* Zeitungsartikel („Föderalistenartikel“, 85 Stück in versch. Blättern NY 1787/88 ersch.); 3 Autoren: Hamilton, Jay, Madison, unterzeichnen alle mit Pseudonym „Publius“: Frage wer hat welchen Text verfasst?
- 12 umstrittene Texte (nur von Hamilton und Madison) wurden von Mosteller/Wallace untersucht
- Mithilfe der 3 Funktionswörer „by, from, to“ wird versucht die autorentyp. Verwendung heraus zu finden – im Gegensatz zum Inhaltswort „war“, das sich als unbrauchbar erweist
- Wichtige Erkenntnis für Stilometrie:

As we implied in discussing the word *war*, the words we want to use are non-contextual ones, words whose rate of use is nearly invariant under change of topic. For this reason, the little filler words, called function words (see Table

Frederick Mosteller, David L. Wallace: *Inference and Disputed Authorship. The Federalist*. Addison-Wesley, Reading MA 1964, p.280

Stilometrie: Beispiel *Federalist Papers*

- Häufigste Wörter in Texten sind Funktionswörter
- Beispiel *Federalist Papers*: 1-22: the,of, to, and, in, a, be, that, it, is, which, by, as, this, would, for, have, will, or, not, from, their
- ABER: in sehr kurzen Texten wie den *Federalist Papers*: kommt schon an Platz 29-30: states, government (Inhaltswörter)
- Auch aus diesem Grund haben Mosteller/Wallace nur bestimmte Wörter zum Vergleich heran gezogen

A major weakness of this study is that it does not have the serious protection against contextual words provided for the main study discussed later. Partly this is a lapse in our work, but partly it comes from the limited amount of material available when we use a calibrating set as well as a selection and weighting set of papers.

Frederick Mosteller, David L. Wallace: *Inference and Disputed Authorship. The Federalist*. Addison-Wesley, Reading MA 1964, p.282

Stilometrie: Korpus – Auswahl - Vorbereitung

- Daher: Vorüberlegung zum Korpus – Vergleichbarkeit

1. Texte sollten einheitlich sein: Sprache, Epoche, Länge
2. Texte sollten in mind. Länge (Wortanzahl von ca. 5000) und in ausreichender Anzahl (erhöht die Aussagekraft)
3. Datenformat: txt (Datenqualität?)
4. Von einem Autor sollten mind. 2 Texte im Korpus vertreten sein
5. Zunächst sollte in einem Test, die Funktionstüchtigkeit der Einstellungen überprüft werden: Wie sicher funktioniert das Clustering bei bekannten Autoren?

Frederick Mosteller, David L. Wallace: *Inference and Disputed Authorship. The Federalist*. Addison-Wesley, Reading MA 1964, p.282

Stilometrie: mit stylo-Paket für R

- R ist ein Statistik-Programm, das in vielen Disziplinen zum Einsatz kommt
- In der quantitativen Textanalyse sucht man ein Tool zur statistischen Auswertung der Worthäufigkeiten in Texten

- Folgende Fragen soll das Tool beantworten:

1. Welche sind die meistgenutzen Wörter im Korpus (Textsammlung)? Listen von 10-5000
2. Wie ist deren relative Häufigkeit?
   Ergebnis statistischer Wert, der die Häufigkeit von Wort x im Verhältnis zur Gesamtmenge der Wörter im Korpus angibt (stylo erstellt dazu eine Frequency-Tabelle)

Stilometrie: mit stylo-Paket für R

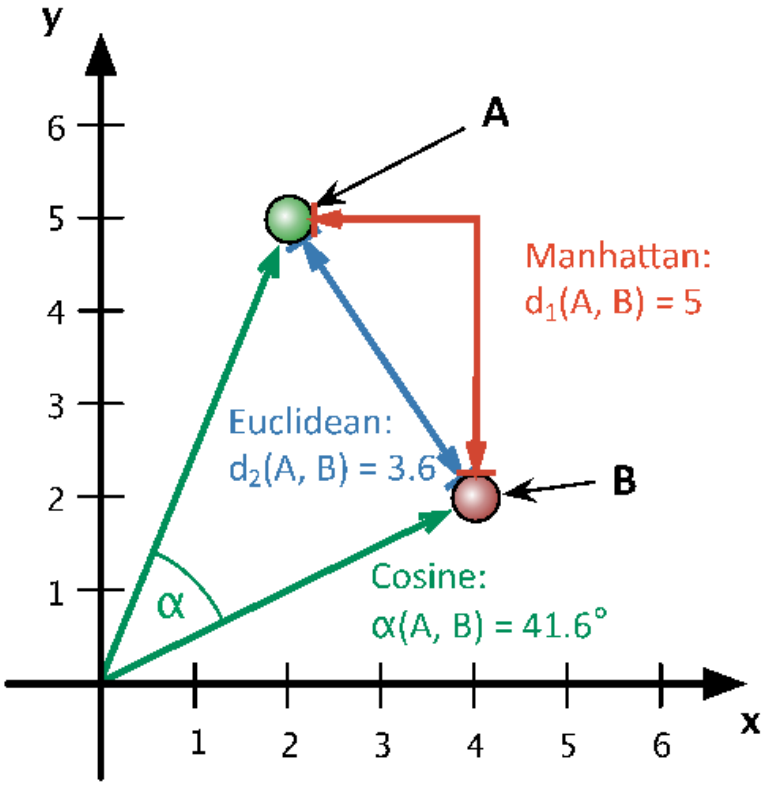
- Folgende Fragen soll das Tool beantworten:

3. Die Worthäufigkeiten sollen für jedes Wort und jeden Text im Vergleich analysiert werden, d.h. welche rel. Häufigkeit hat Wort x im Text y und welche im Text z? Dazu wird eine „frequencies-analyzed“-Tabelle erstellt
4. Wie ist auf Grundlage der rel. Häufigkeit die Distanz der Texte im Korpus untereinander? Zu dieser Berechnung wird ein statistisches Distanzmaß benötigt. Das überträgt die nummerischen Abstände zwischen den Häufigkeitswerten in räumliche Abstände: hier z.B. das euklidische Delta

$$\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2}$$

Stilometrie: Frage des Deltas

- Euclidian: einfache, ähnliche Distanzen (nicht geeignet für Texte mit so unterschiedlichen Distanzwerten wie „der“ und „Diskurs“
- Manhattan: ähnl. Nachteile wie eukl. Distanz, weil Größenunterschiede nicht abgefangen werden
- Cosine: Größenunterschiede werden durch Effekt der Vektor-Normalisierung abgefangen

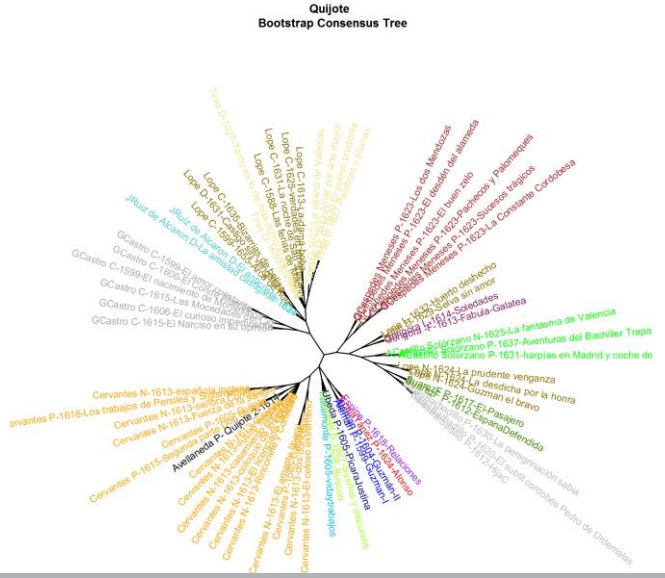


Vgl. Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015). Eder/ Rybicki/ Kestemont: `Stylo': a package for stylometric analyses (2104)

Stilometrie: stylo-Paket für R

https://sites.google.com/site/computationalstylistics/stylo

- Maciej Eder, Jan Rybicki, Mike Kestemont: (2013): “Stylometry with R: a suite of tools”, in: Digital Humanities 2013: Conference Abstracts, University of Nebraska, Lincoln, 487-89.
- Vorteile des stylo packages: GUI-Oberfläche ermöglicht es auch Laien, mit R zu arbeiten
- Eder entwickelte auch ein eigenes Delta, das sich besonders zur Autorschaftsattribution eignet



Stilometrie

Anwendungsbeispiele

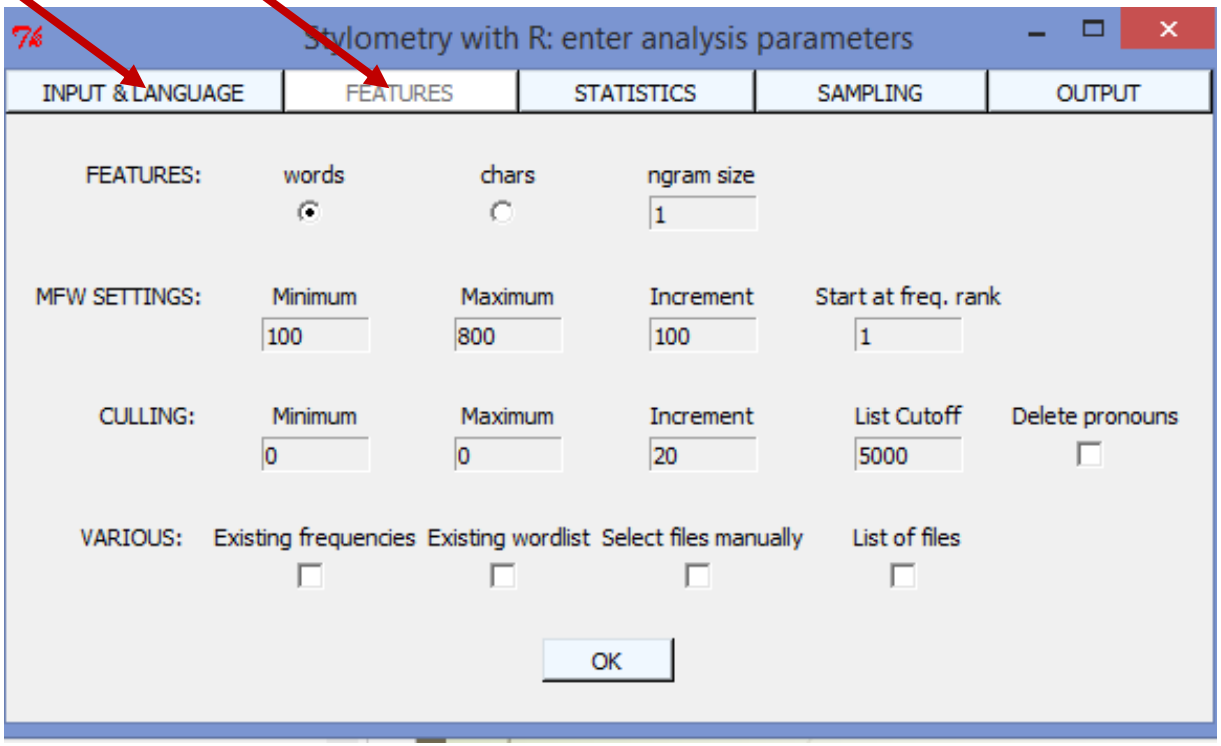
1. „Federalist Papers“
2. Der Spanische pikareske Roman: Autorschaft des *Lazarillo de Tormes*

Stilometrie: Federalist Papers

1. 85 Texte, Wöteranzahl zwischen 1000 und 5000
2. Alle Texte als Ordner „corpus“ abspeichern (txt-Format) und Titel der Dateien mit Name_Jahr oder Nummer oder Titel
3. Stylo library in R aufrufen: Verzeichnis auswählen (setwd) und stylo starten stylo ()
4. Nun erhält man folgendes GUI

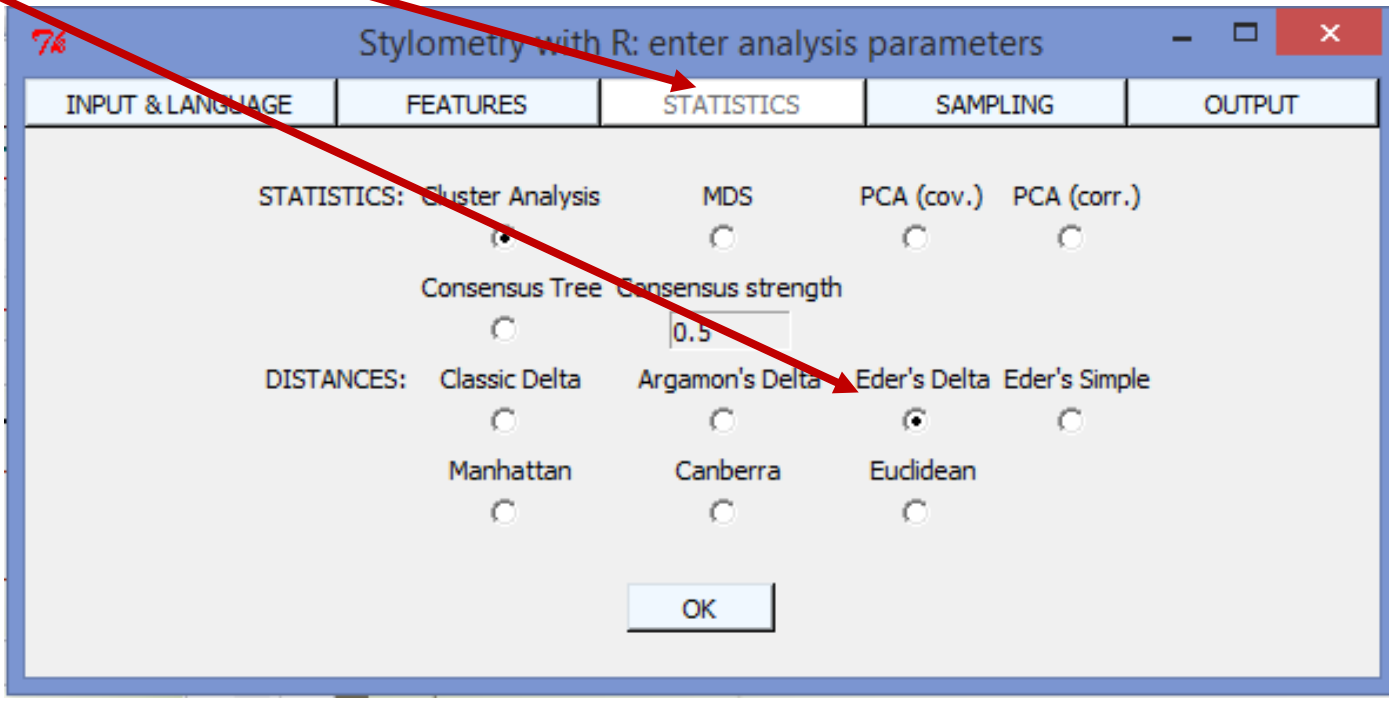
Stilometrie: Federalist Papers

- Sprachauswahl, Features: words, MFW-Settings klein auswählen,
- Culling unnötig
- Existing frequ. etc. irrelevant



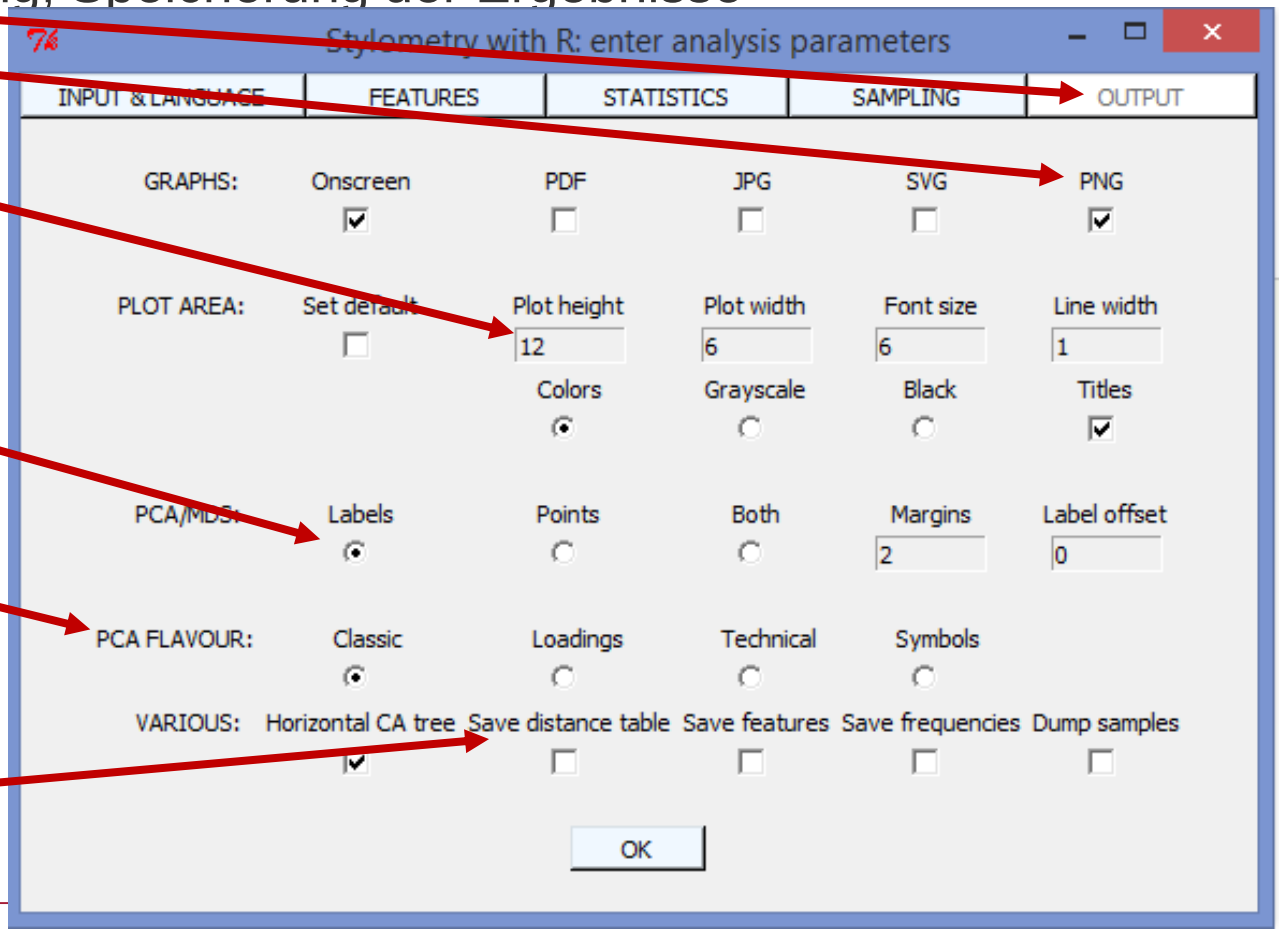
Stilometrie: Federalist Papers

- Statistics: Berechnung: Cluster Analysis
- Delta: untersch. auspro-bieren, hier: Eder‘s

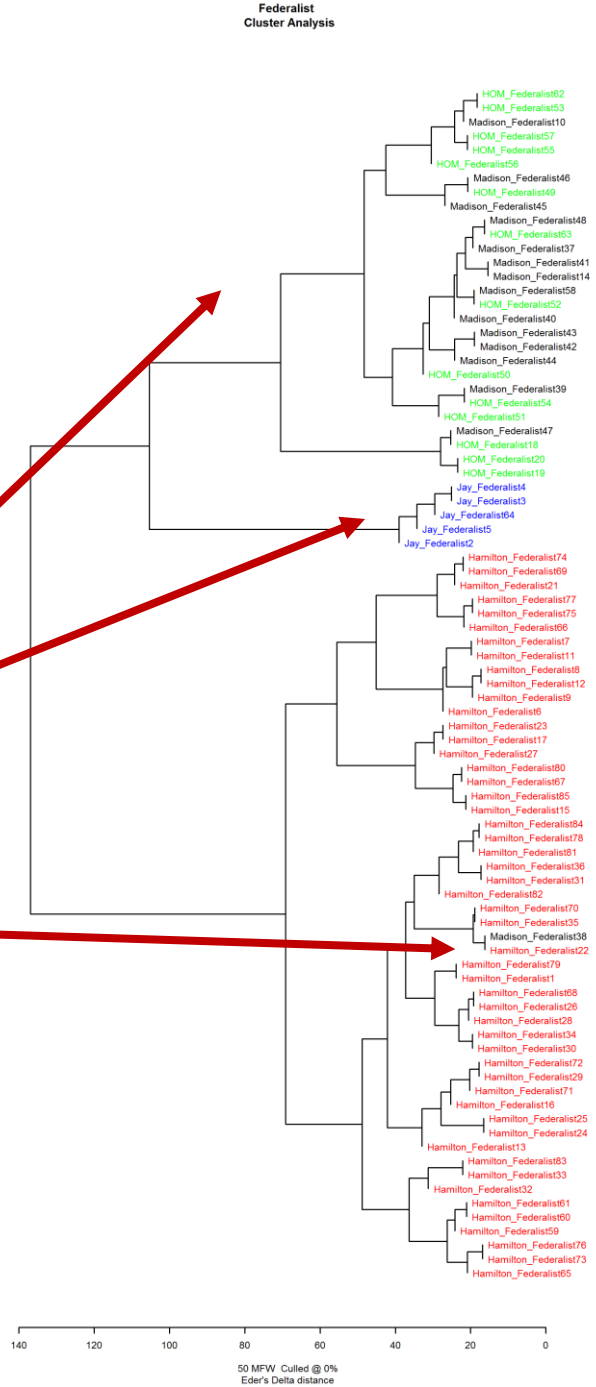


Stilometrie: Federalist Papers

- Output: Darstellung, Speicherung der Ergebnisse
- PNG: Grafik wird gespeichert Größenangaben wichtig für Lesbarkeit
- Labels: Beschriftung mit Titel
- Nur bei PCA wichtig
- Speicherung der Tabellen wichtig



Stilometrie: *Federalist Papers*: Ergebnisse



Alle drei Autoren wurden erkannt und geclustert:
Grün: Hamilton/Madison (gemeinsam?)
Blau: Jay
Rot: Hamilton
Schwarz: Madison

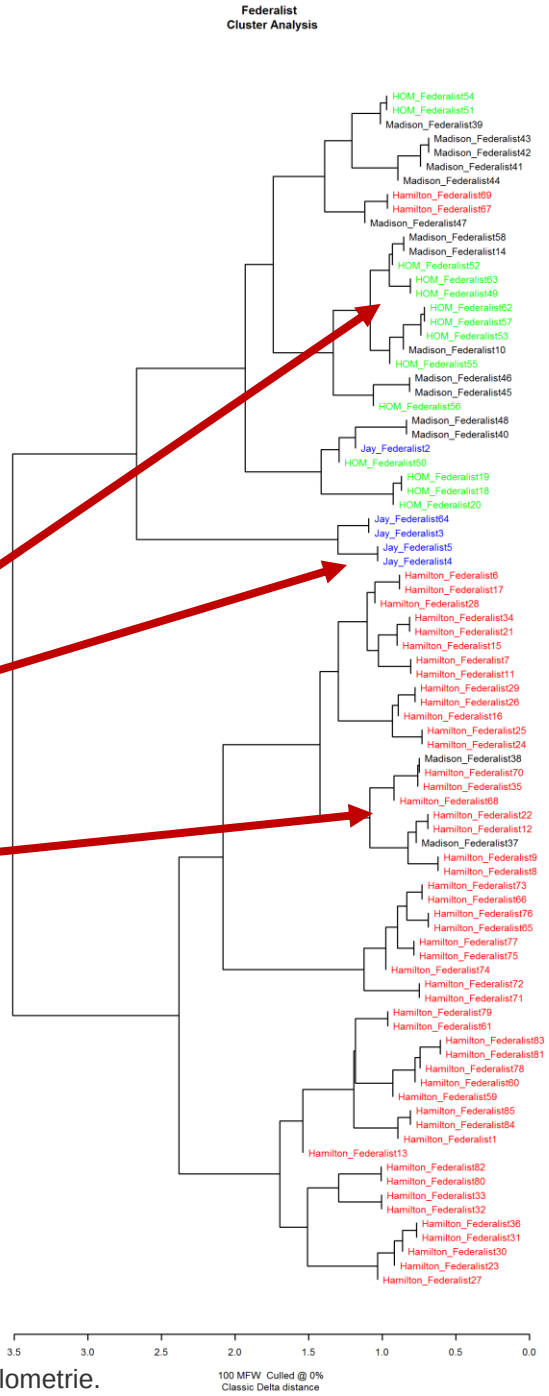
- Interpretation: Gemeinschaftswerke eher Madison zuzuordnen – ein Madison-Text eher Hamilton?

Parameter: nur 50 MFW und Eder‘s Delta; Cluster-Analysis: Dendogramm

Stilometrie: *Federalist Papers*: Ergebnisse

Alle drei Autoren wurden erkannt und geclustert:
Grün: Hamilton/Madison (gemeinsam?)
Blau: Jay
Rot: Hamilton
Schwarz: Madison

- Interpretation:
  Mehr Unsicherheiten?
  Jay plötzlich als Autor unter grün/sch?
  Konsistent: grün eher Madison



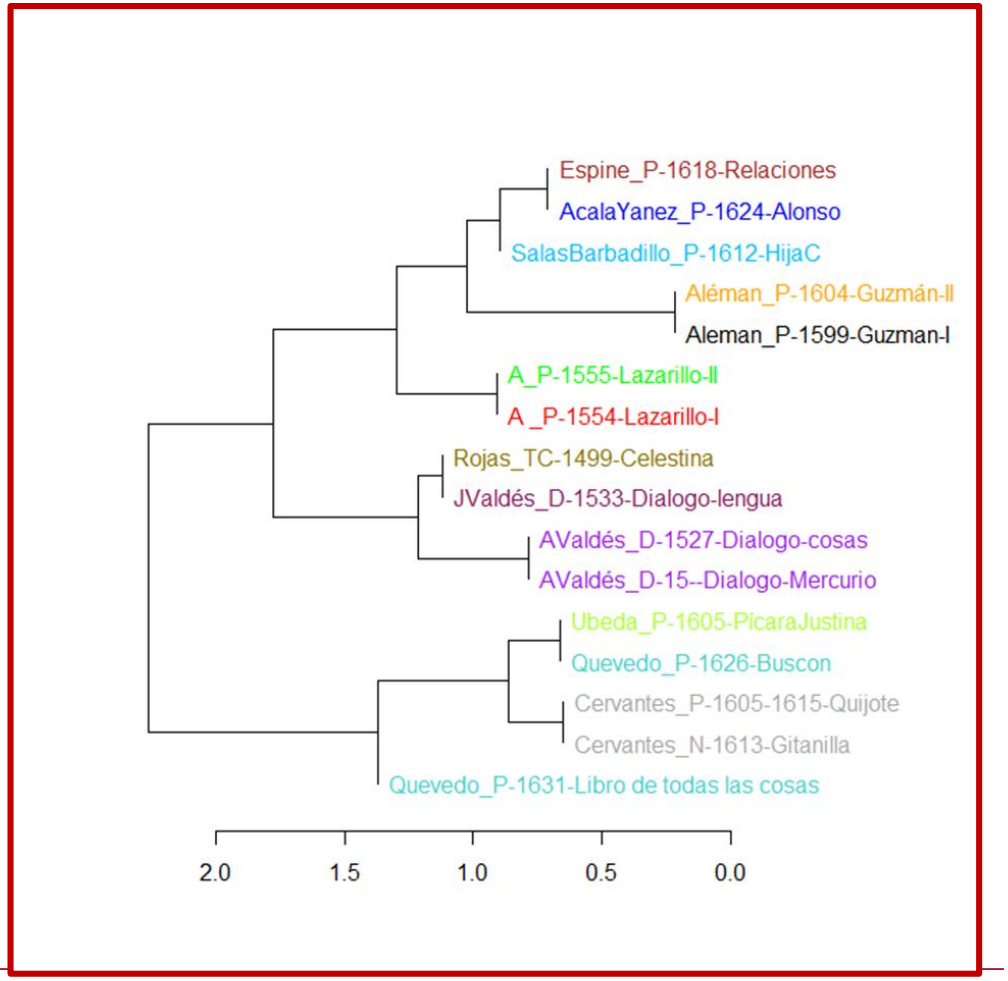
Parameter: nur 100 MFW und Classic Delta; Cluster-Analysis: Dendogramm

Stilometrie: Der spanische pikareske Roman: Autorschaft *Lazarillo de Tormes*

1. Ansteigende Textanazahl, verschiedene Kombinationen, Wöteranzahl zwischen 5000 und 140.000
2. Pícaro-Romane und andere spanische Romane des 16./17. Jahrhunderts
3. Sprache beachten kein Alt-Spanisch

Stilometrie: Der spanische pikareske Roman: Autorschaft *Lazarillo de Tormes*

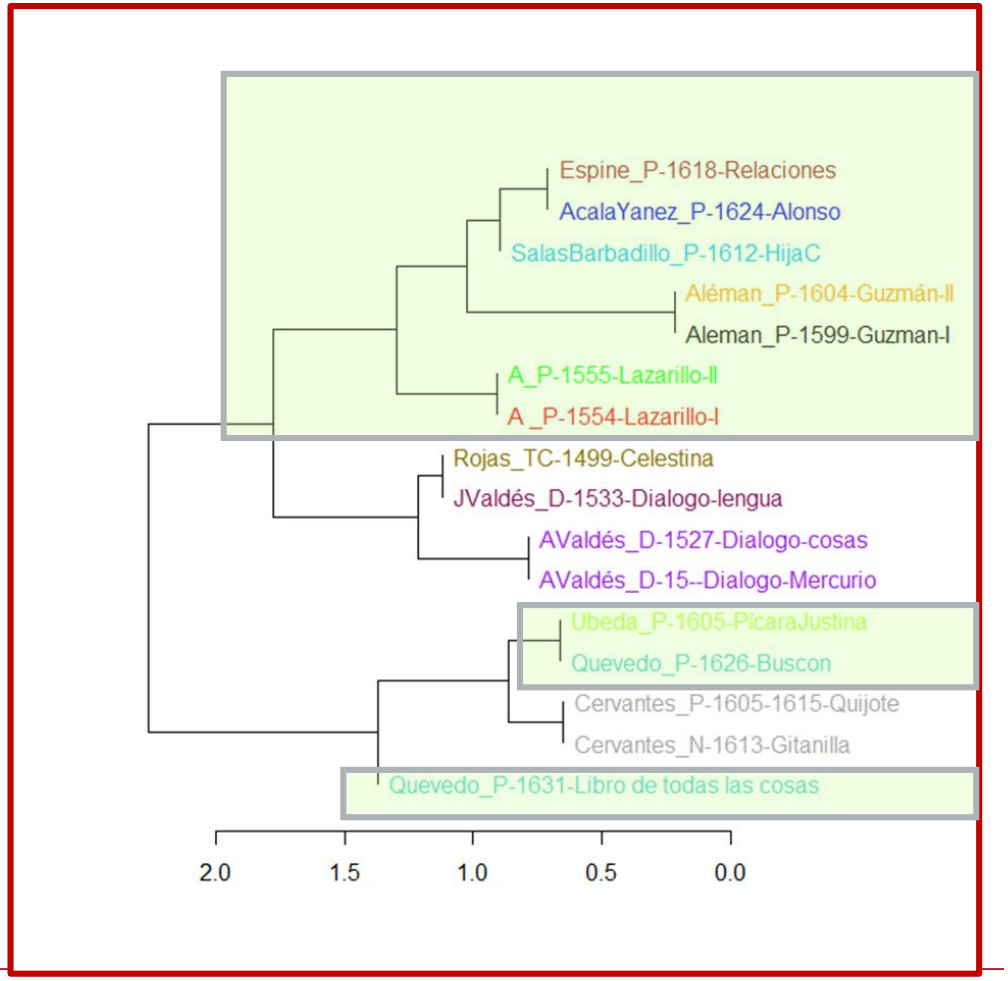
Auswertung von 16 Texten mit stilistischer Ähnlichkeit zum *Lazarillo:* (Parameter: 100-800 MFW, Classic Delta)



Stilometrie: Der spanische pikareske Roman: Autorschaft *Lazarillo de Tormes*

Ergebnisse / Interpretation:

1. Trotz Einzelgänger bildet die Gruppe des pikaresken Romans ein Cluster
2. *Lazarillo* wird keinem Autorschaftskandidaten zugeordnet (auch nicht Alfonso de Valdés)
3. Quevedos *Buscón* und Ubedas Pícara Justina bilden ein Cluster
4. Falschzuordnung von Quevedos „Libro de…“



Conclusion

1. Trotz Einfachheit der Bedienung, und der features (MFW), komplexes Verfahren (Übung, ausprobieren)
2. Bei nicht-lit. Texten, Autorschaft leichter zu klären: denn in Literatur erfordern Textsorten, Genres, Zensur, etc., dass der Stil wandelbar ist
3. Daher das Wichtigste: Interpretieren, Interpretieren, Interpretieren
4. Keineswegs „objektiv“ oder „intersubjektiv“ – aber Bedingungen des Experiments können eindeutig offen gelegt und reproduziert werden
5. Kein Massenphänomen, sondern für bestimmte Erkenntnisinteressen wertvoll

References

Burrows, J. „‚Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17, Nr. 3 (1. September 2002): 267–87. doi:10.1093/llc/17.3.267.

Calvo Taller, José. „ENTENDIENDO DELTA DESDE LAS HUMANIDADES". *Caracteres*, o. J., 140.

Clement, Tanya. "Text Analysis, Data Mining, and Visualizations in Literary Scholarship". In *Literary Studies in the Digital Age. An Evolving Anthology*, (2013). doi: 10.1632/lsda.2013.8

Eder, Maciej, M Kestemont, und J Rybicki. „Stylometry with R: a suite of tools". In *Digital Humanities 2013. Conference Abstracts*, herausgegeben von Lincoln: University of Nebraska-Lincoln, 487–89.

Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015). "Towards a better understanding of Burrows's Delta in literary authorship attribution". In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, CO.

Herrmann, Berenike, J., Karina van Dalen-Oskam, und Christof Schöch. „Revisiting Style, a Key Concept in Literary Studies". *Journal of Literary Theory* 9, Nr. 1 (2015): 25–52.

Mosteller, Frederick, David L. Wallace. *Inference and Disputed Authorship. The Federalist*. Addison-Wesley, Reading MA 1964.

Schöch, Christof. „Quantitative Analyse". *Einführung in die Digital Humanities*, 2017, 279-298

Danke.

Kontakt:

PD Dr. Nanette Rißler-Pipka
Wilhelmstraße 50, 72074 Tübingen
Telefon: +49 7071 29-72394
nanette.rissler-pipka@uni-tuebingen.de