Delta und Merkmalsselektion

Welche Wörter unterscheiden arabisch-lateinische Übersetzer?

Andreas Büttner[1] Thomas Proisl[2]

[1]Institut für Philosophie
Universität Würzburg

[2]Professur für Korpuslinguistik
Universität Erlangen-Nürnberg

<philtag n="13"/>, 26.02.2016

Gliederung

Gliederung

Einleitung

Arabisch-lateinische Übersetzungen im 12. Jahrhundert

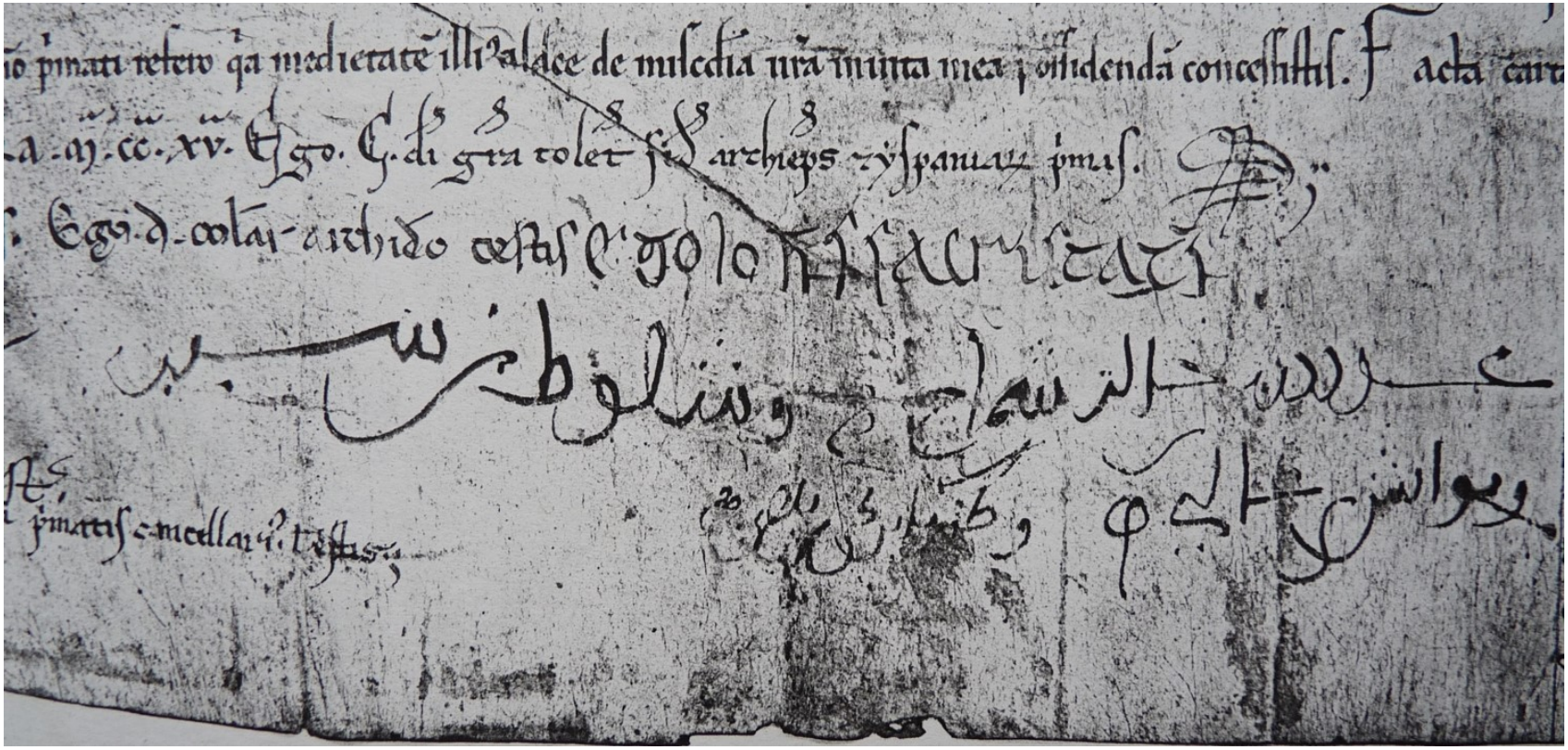


Abbildung: Urkunde Toledo, 20. Dezember 1177, ACT V.12.P.1.1
(F. Hernández, Los cartularios de Toledo: catálogo documental, 1985).

Einleitung

Zum Textcorpus

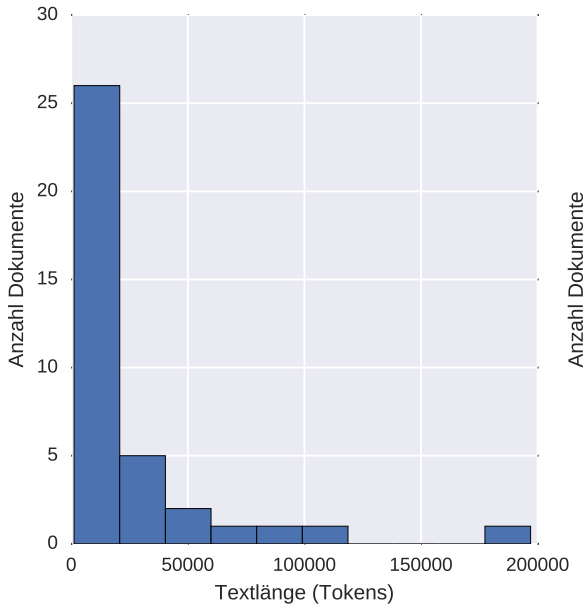
- inzwischen 98 Texte digitalisiert
- Texte von 11 verschiedenen Übersetzern, 41 anonym
- zum Teil Zusammenarbeit von Übersetzern
- Philosophie, Mathematik, Astronomie, Astrologie, Medizin, Geologie und Metereologie, aber auch religiöse, magische und alchemistische Traktate
- Ziel: Identifizierung der anonymen Übersetzer

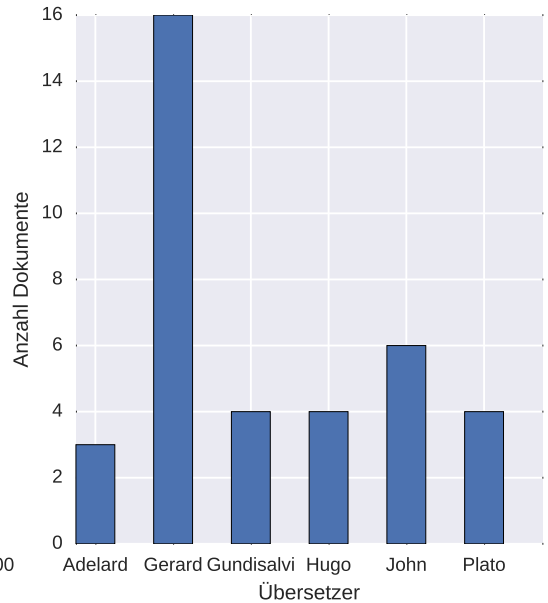
Gliederung

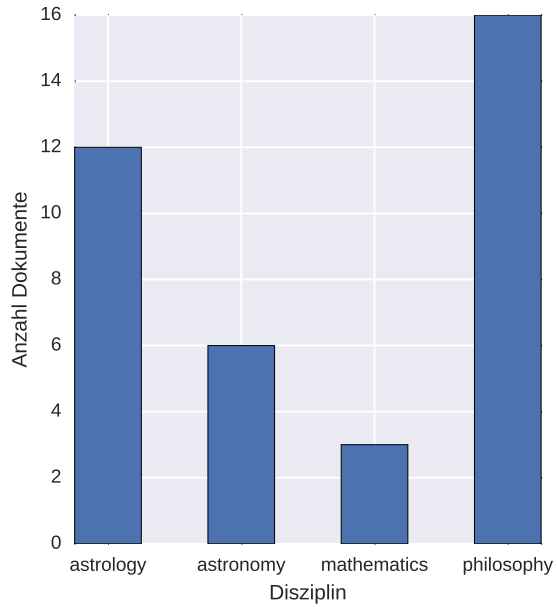
Stilometrie (MFW)

Testkorpus

- Auswahl von 37 Texten
- keine Überarbeitungen (soweit bekannt)
- Textlänge mindestens 750 Wörter
- mindestens drei Texte pro Übersetzer/Disziplin



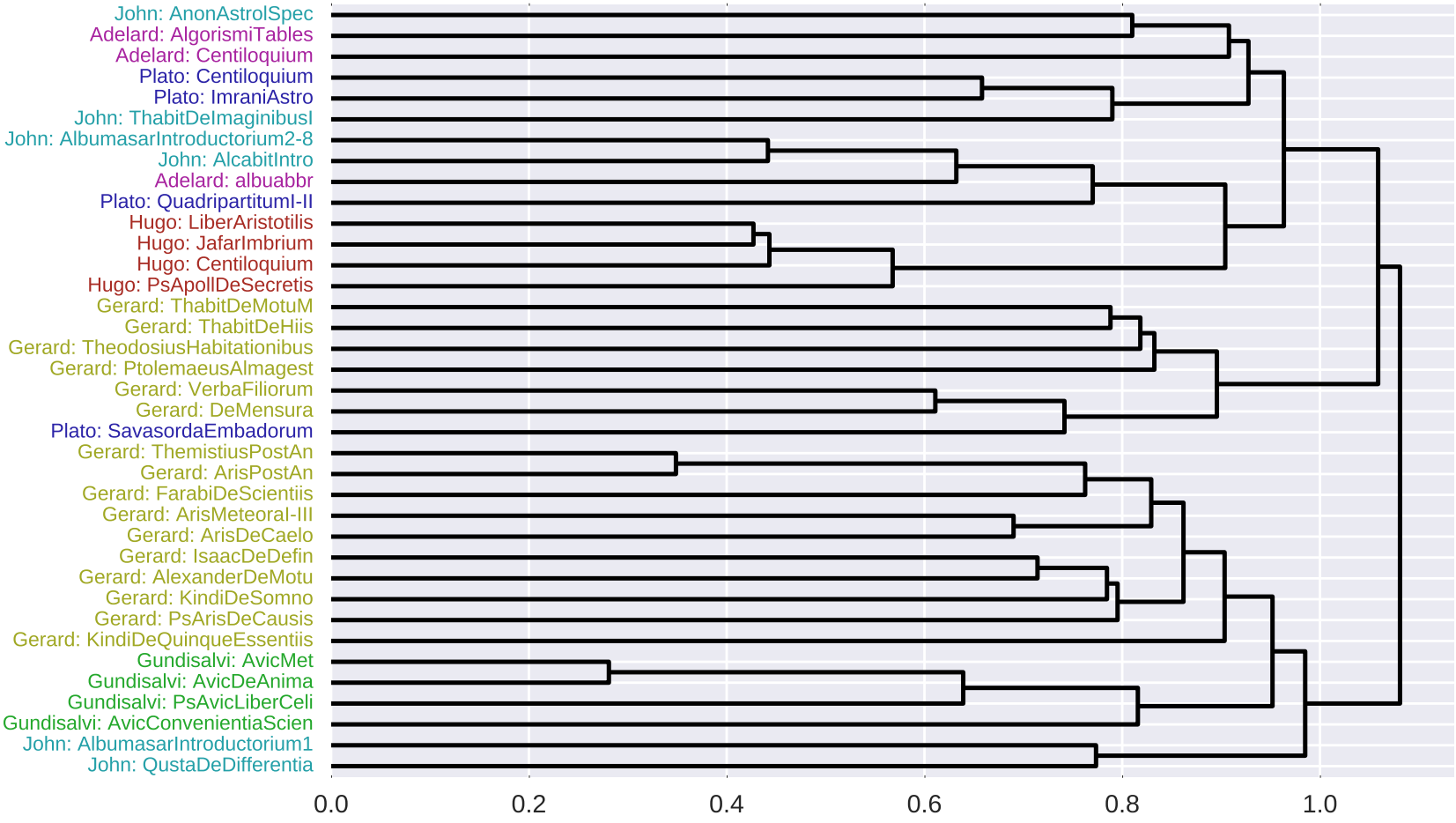




Stilometrie (MFW)

das Problem: Übersetzer

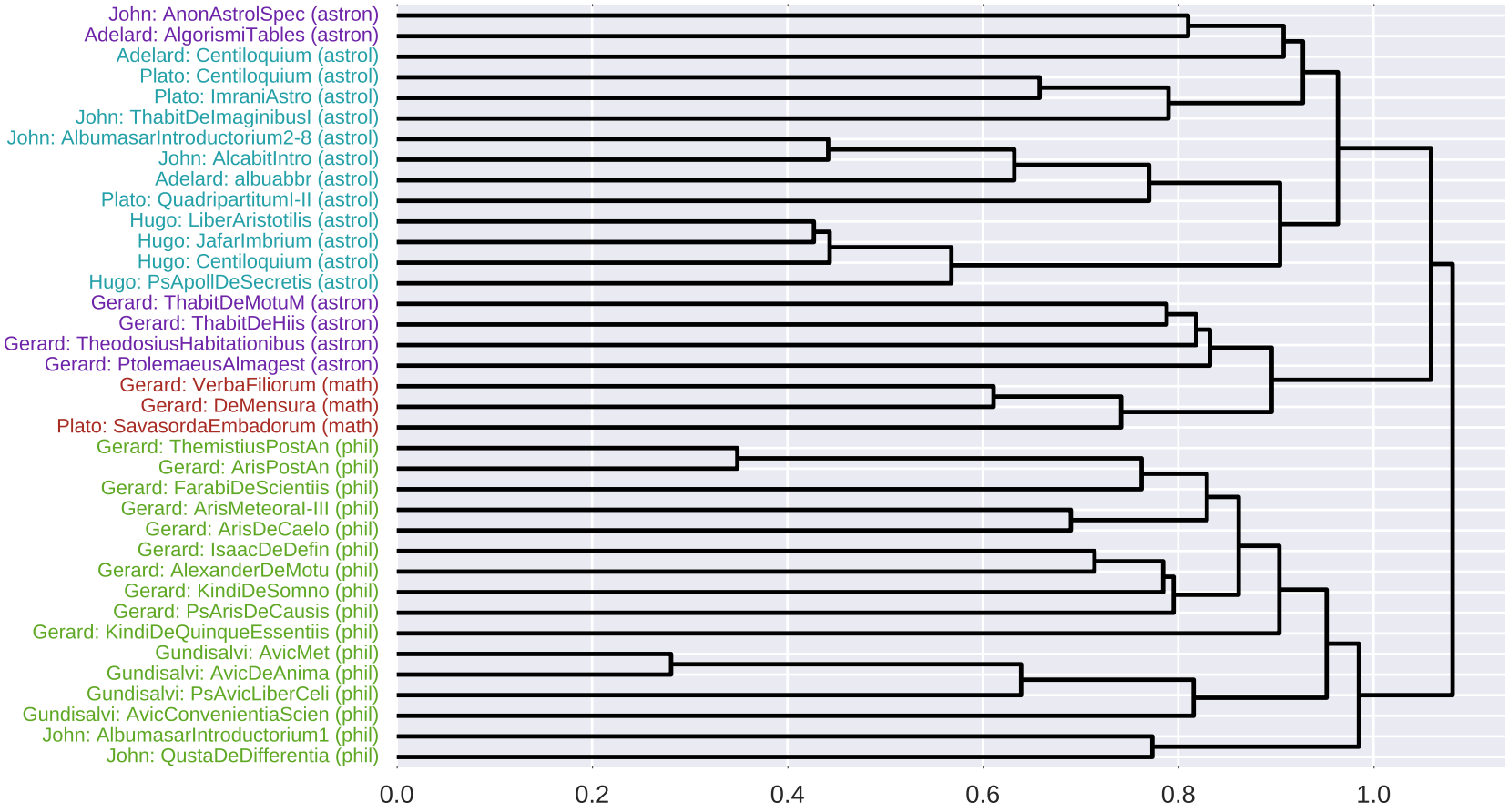
Cosinus/Weighted-Clustering auf Z-scores der 500 häufigsten Wörter, Farben nach Übersetzer.



Stilometrie (MFW)

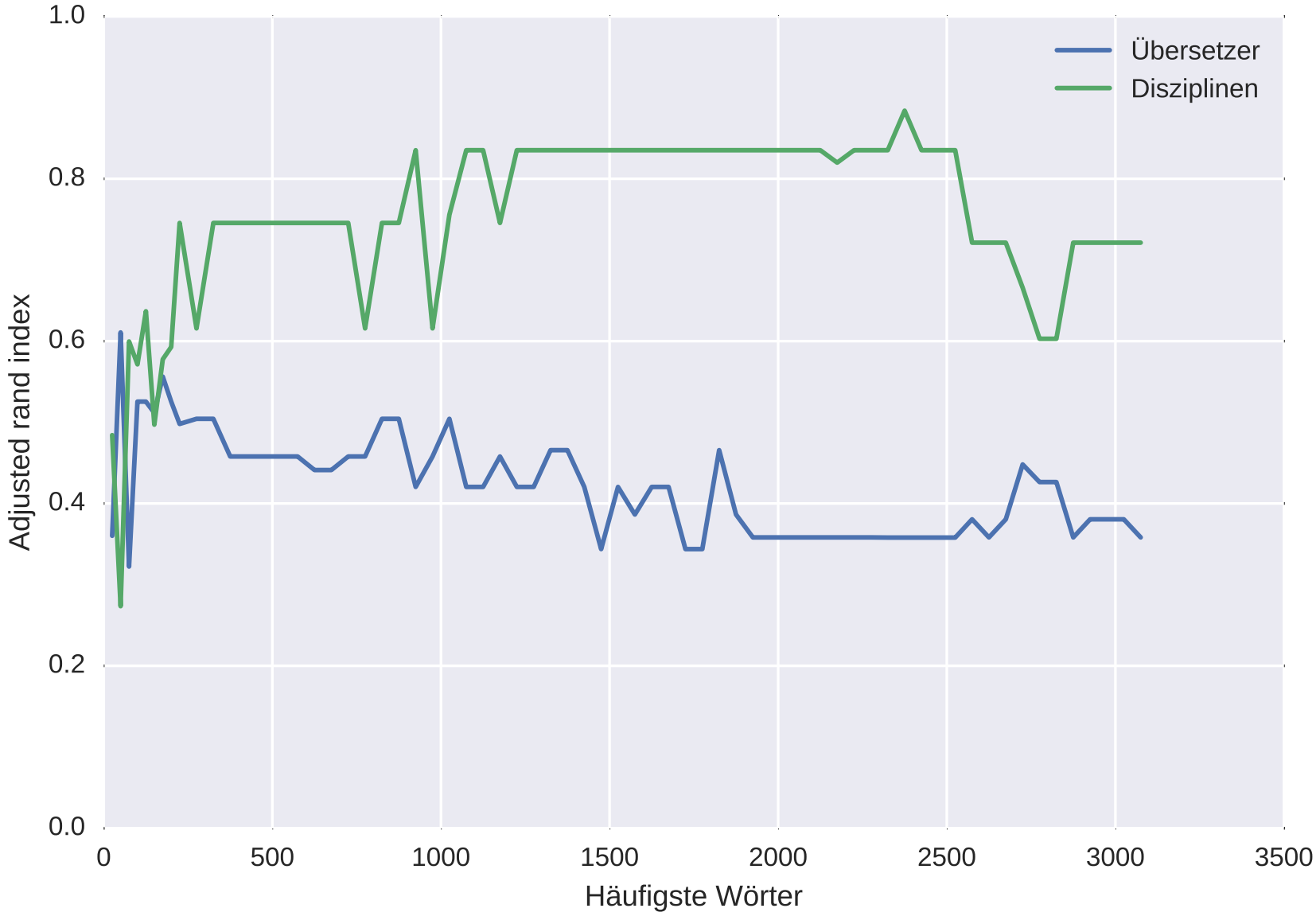
das Problem: Disziplinen

Cosinus/Weighted-Clustering auf Z-scores der 500 häufigsten Wörter, Farben nach Disziplin.



Stilometrie (MFW)

das Problem: Übersetzer vs. Disziplinen



Stilometrie (MFW)

das Problem: Zusammenfassung

- Clustering nach Disziplinen funktioniert relativ gut
- inhaltliche Information überlagert Übersetzer
- Wie lassen sich Übersetzer- von Disziplineninformationen trennen?

Gliederung

Rekursive Merkmalseliminierung

Methode

- von Guyon et al. (2002) vorgeschlagene Methode
  - möglichst kleine Teilmenge von Merkmalen
  - trotzdem möglichst gute Ergebnisse
- basiert auf überwachtem Lernverfahren (üblicherweise *Support Vector Classifier*)
  - d.h. wahre Autoren bzw. Übersetzer müssen bekannt sein (zumindest für Teilkorpus)
- mögliche Alternative zu *most frequent words* zur Autorschaftszuschreibung (Evert et al. 2015)

Rekursive Merkmalseliminierung

Details

Algorithmus

1. trainiere Klassifikator auf Merkmalsmenge (= Zuweisung von Gewichten)
2. entferne k Merkmale mit niedrigsten absoluten Gewichten (*pruning*)
3. gewünschte Anzahl von Merkmalen erreicht?

   ja: fertig

   nein: gehe zu 1.

- alternativ: Kreuzvalidierung nach jedem *pruning* Schritt zur Bestimmung der optimalen Merkmalsmenge

Rekursive Merkmalseliminierung

Anwendung

Unser Vorgehen

- schrittweise Verkleinerung der Merkmalsmenge auf 500 Merkmale
  - auf 10.000 Merkmale in 1.000er Schritten
  - auf 1.000 Merkmale in 200er Schritten
  - auf 500 Merkmale in 25er Schritten
- anschließend Bestimmung der optimalen Merkmalsmenge
  - reduzieren der Merkmalsmenge in 1er Schritten
  - 3-fache Kreuzvalidierung

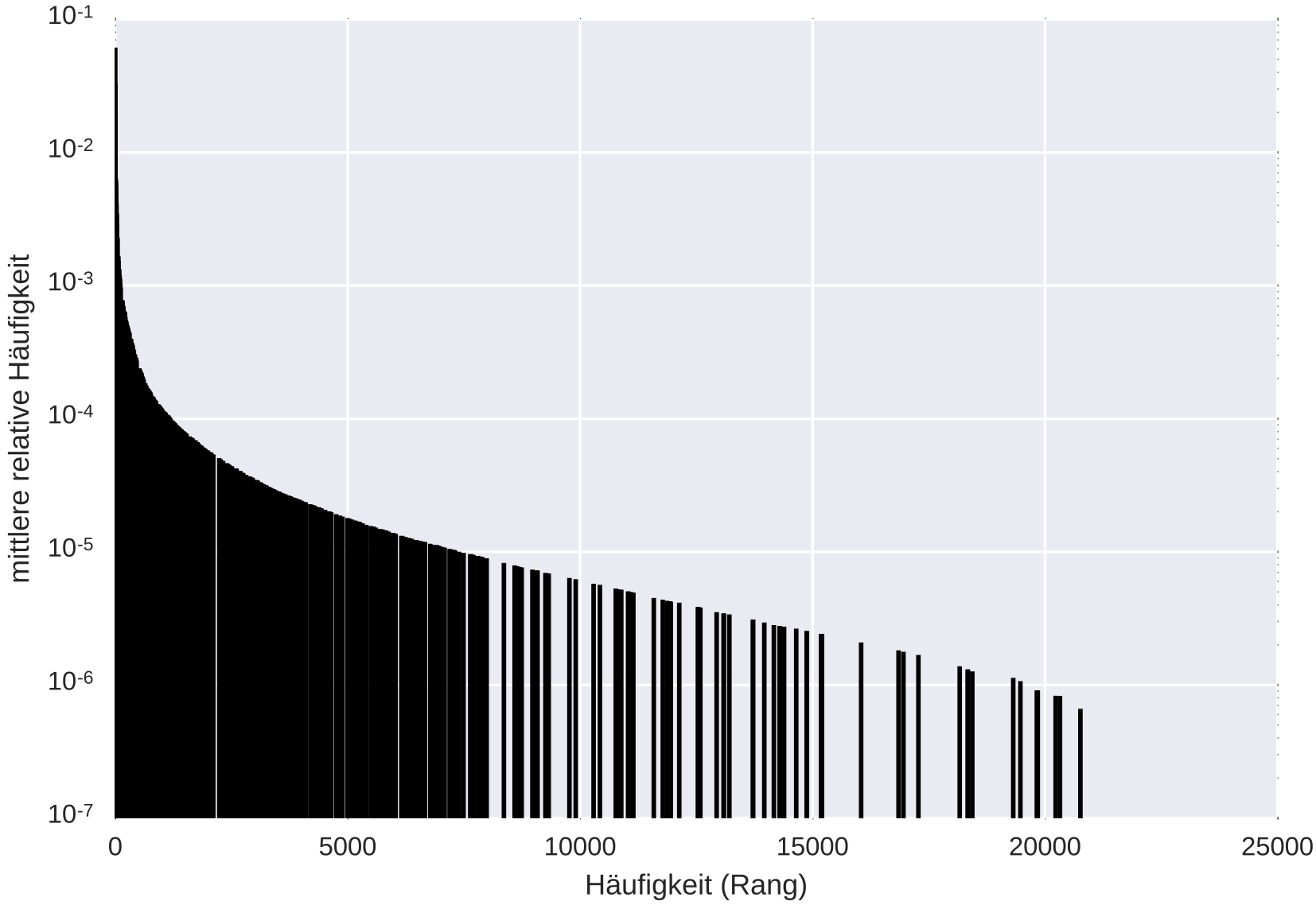
Rekursive Merkmalseliminierung

Merkmalsmengen

- Trainingssets für Übersetzer/Disziplinen
- Optimale Merkmalsmenge Übersetzer: 495 Wörter
- Optimale Merkmalsmenge Disziplinen: 485 Wörter
- auf dem Testset korrekte Aufteilung nach Übersetzern/Disziplinen ($ARI = 1$)

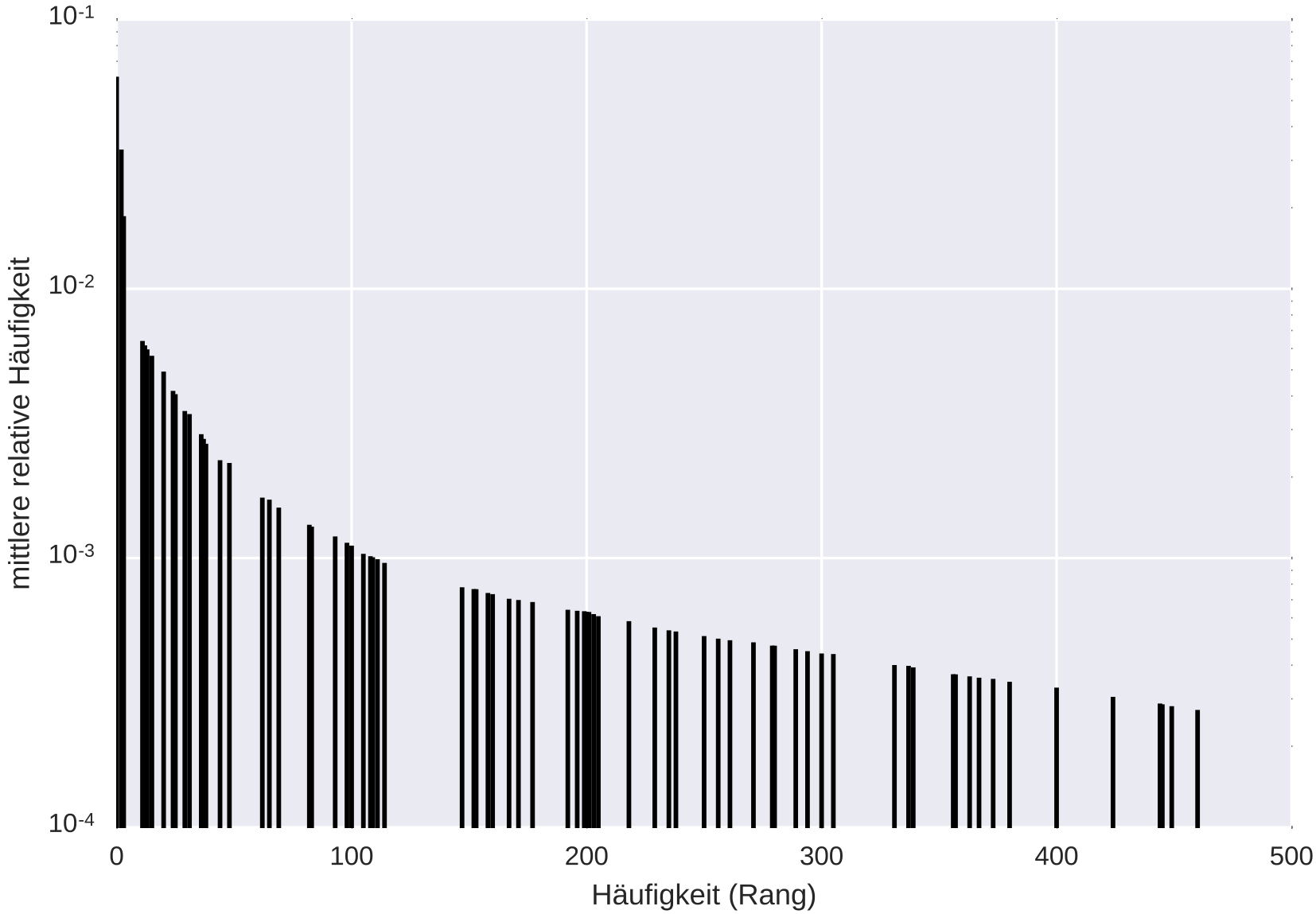
Rekursive Merkmalseliminierung

Merkmalsmenge – Übersetzer



Rekursive Merkmalseliminierung

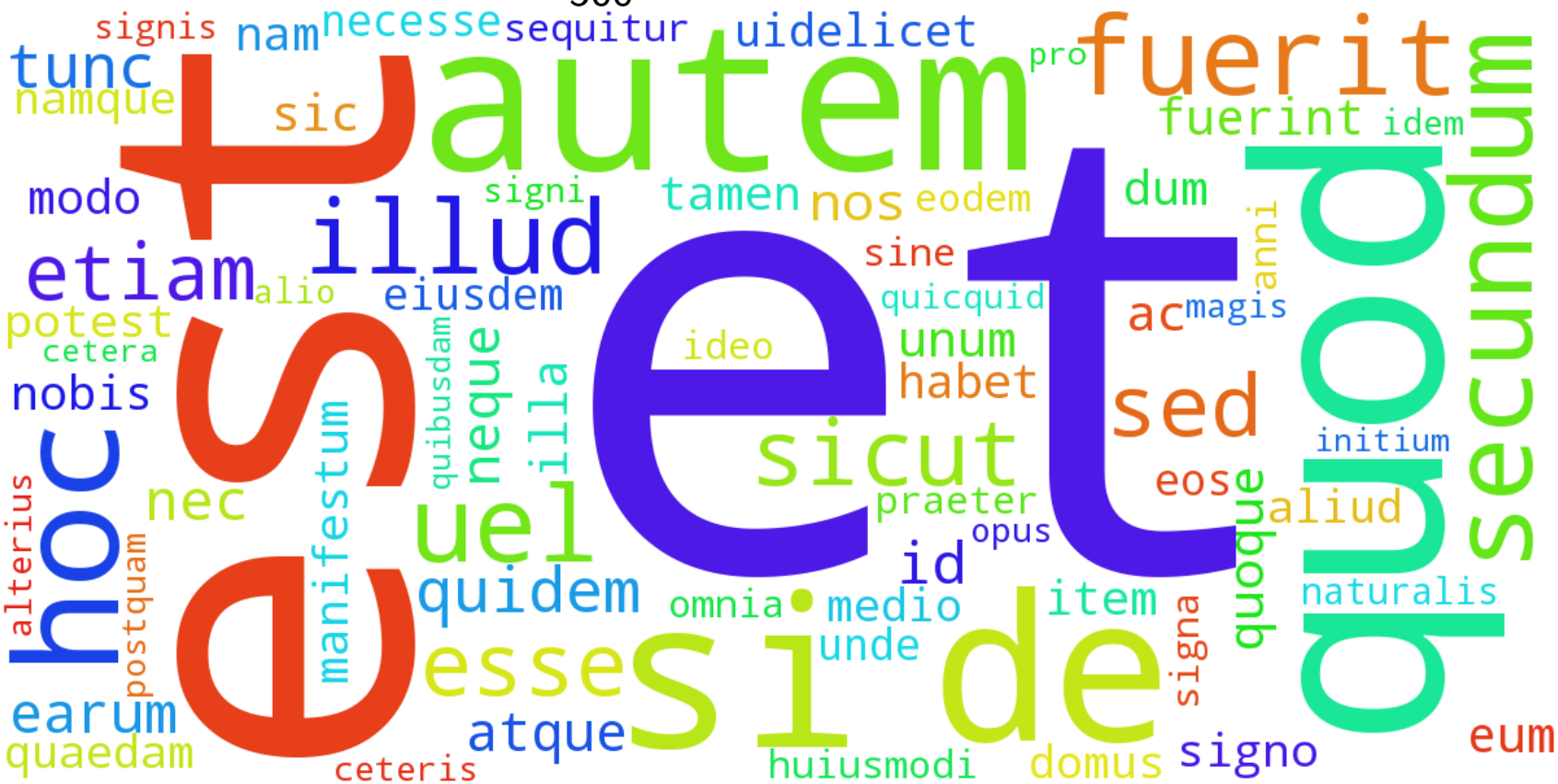
Merkmalsmenge – Übersetzer (Detail)



Rekursive Merkmalseliminierung

Übersetzerwörter

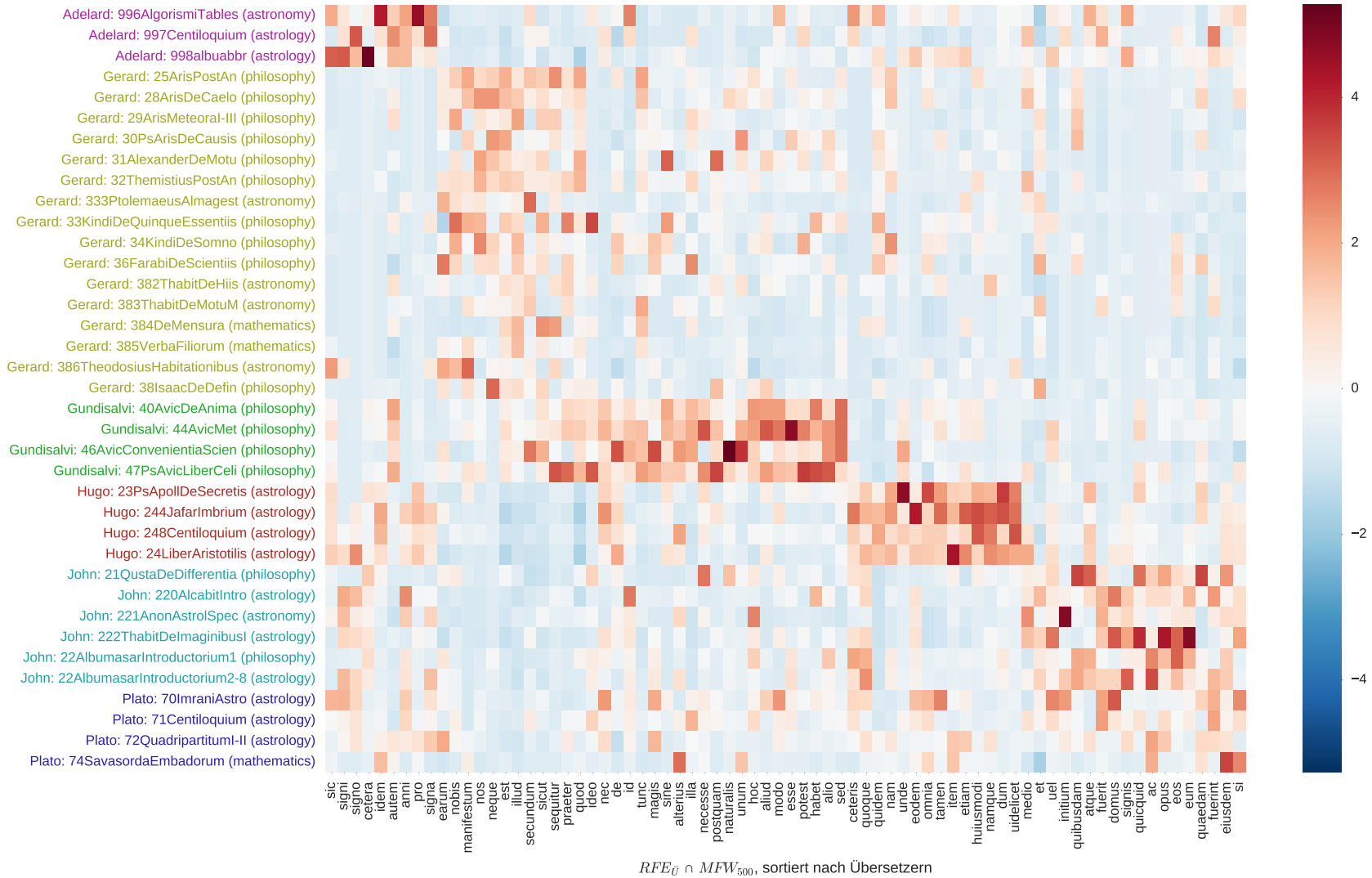
Schnittmenge (74 Wörter) zwischen Übersetzerwörtern ($RFE_{Ü}$) und Menge der 500 häufigsten Wörter (MFW_{500}): $ARI = 0,824$ – deutlich besser als MFW_{500}.



Rekursive Merkmalseliminierung

Übersetzerwörter

Terme lassen sich dem Übersetzer zuordnen, in dessen Texten der Mittelwert ihrer z-Werte am höchsten ist.



$RFE_{\ddot{U}} \cap MFW_{500}$, sortiert nach Übersetzern

Rekursive Merkmalseliminierung

Differenzmengen

- Wenn die 74 Wörter in $\text{MFW}_{500} \cap \text{RFE}_{\text{Ü}}$ so gut für Übersetzer funktionieren (ARI=0,824), was ist mit den 426 restlichen Wörtern in MFW_{500}?

Differenzmenge mit MFW_{500} ($\text{MFW}_{500} \setminus \text{RFE}_{\text{Ü}}$)

- 426 Wörter
- (ARI für Übersetzer: 0,284)
- ARI für Disziplinen: 0,795
  - besser als mit allen MFW_{500} (0,746)!

Rekursive Merkmalseliminierung

Differenzmengen

- Wenn die 123 Wörter in $MFW_{500} \cap RFE_D$ so gut für Disziplinen funktionieren (ARI=0,836), was ist mit den 377 restlichen Wörtern in MFW_{500}?

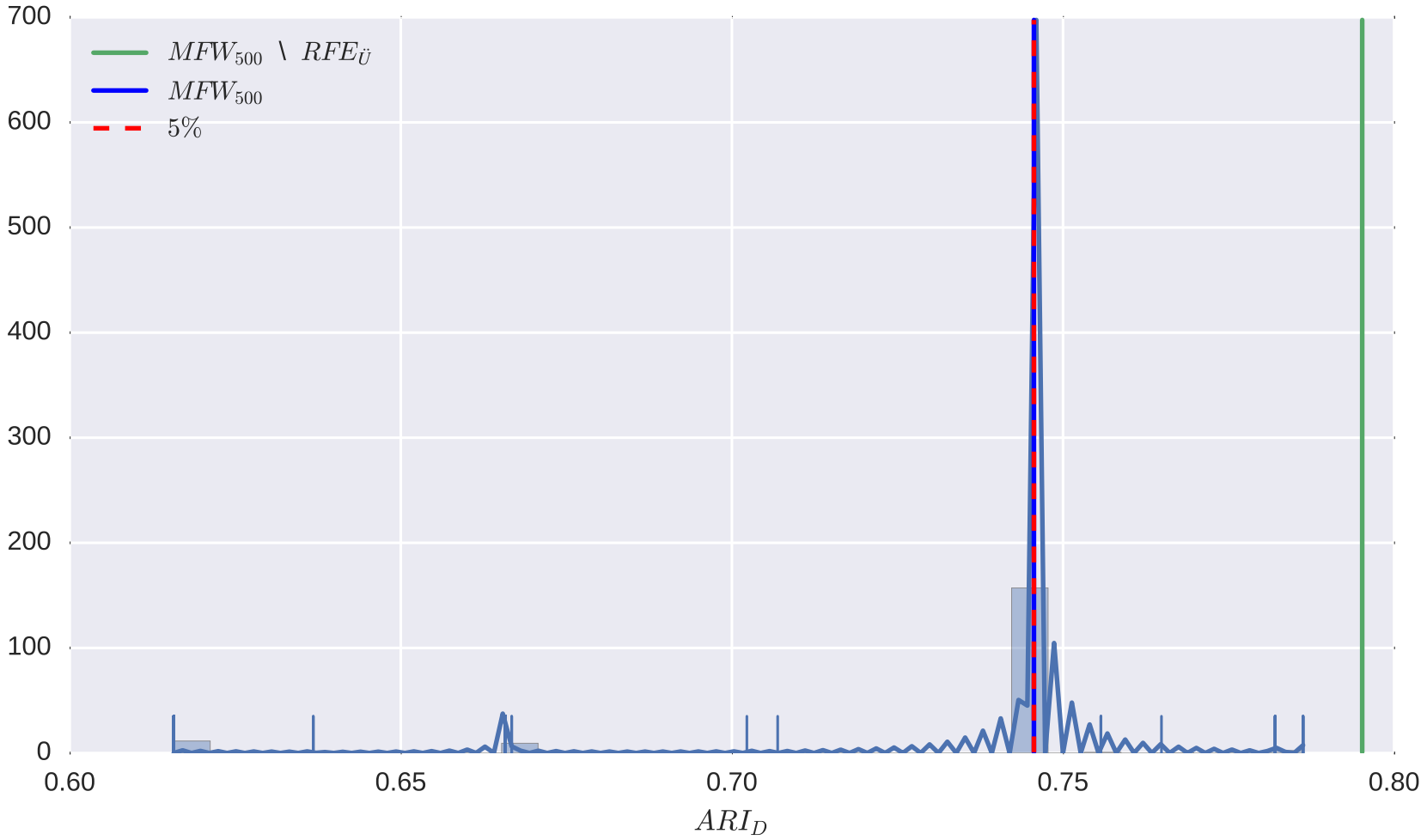
Differenzmenge mit MFW_{500} ($MFW_{500} \setminus RFE_D$)

- 377 Wörter
- (ARI für Disziplinen: 0,593)
- ARI für Übersetzer: 0,526
  - besser als mit allen MFW_{500} (0,458)!

Rekursive Merkmalseliminierung

Differenzmengen – besser als Zufall?

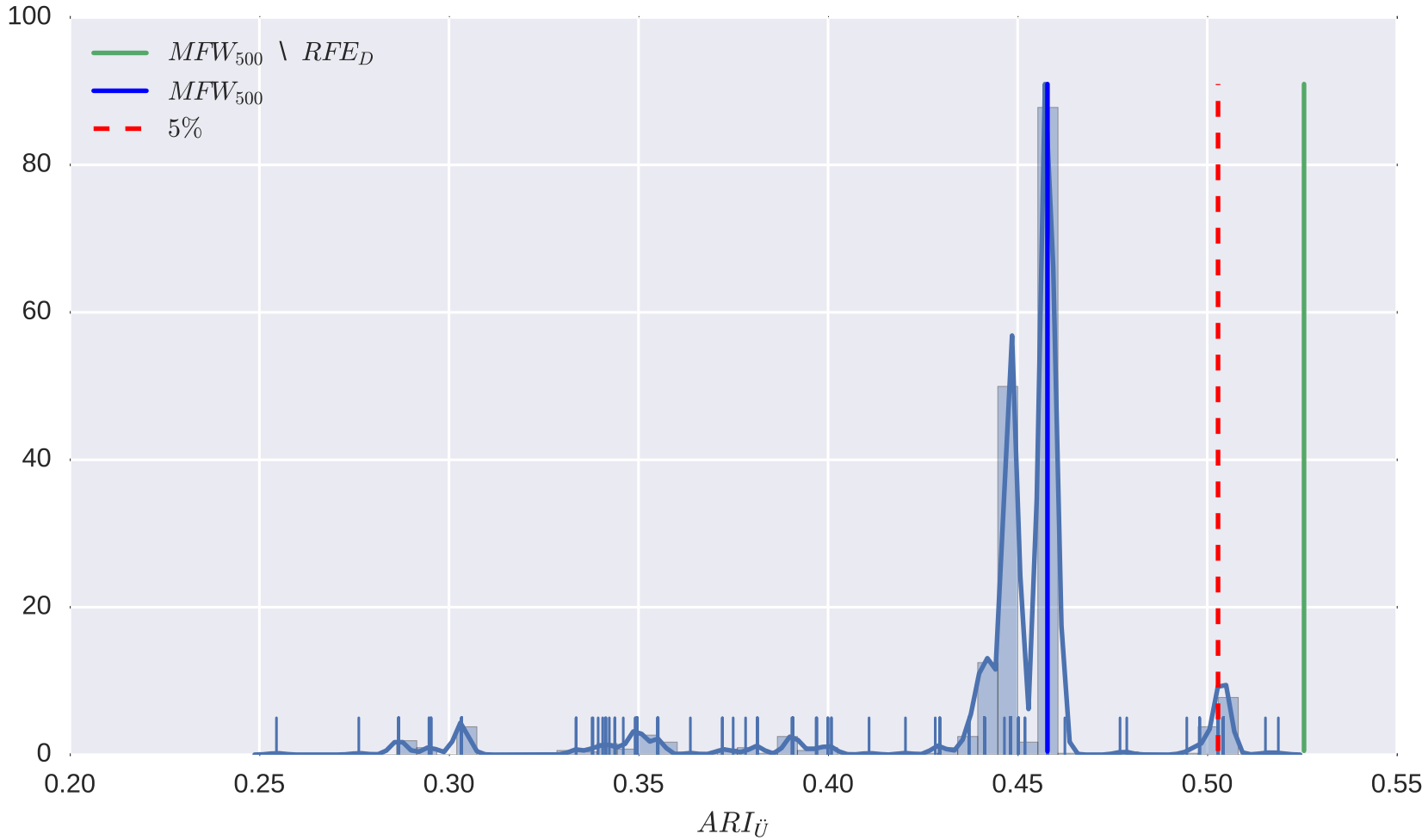
Ohne Übersetzerwörter: Vergleich von ($MFW_{500} \setminus RFE_{Ü}$) mit 1000 zufällig ausgewählten Teilmengen der gleichen Größe.



Rekursive Merkmalseliminierung

Differenzmengen – besser als Zufall?

Ohne Disziplinenwörter: Vergleich von ($\text{MFW}_{500} \setminus \text{RFE}_{\text{D}}$) mit 1000 zufällig ausgewählten Teilmengen der gleichen Größe.



Gliederung

Ergebnisse

Überblick

Merkmale	$ARI_{Ü}$	ARI_{D}
MFW_{500}	0,458	0,746
$RFE_{Ü}$	1,000	0,310
RFE_{D}	0,223	1,000
$MFW_{500} \cap RFE_{Ü}$	0,824	0,228
$MFW_{500} \cap RFE_{D}$	0,189	0,836
$MFW_{500} \setminus RFE_{Ü}$	0,284	0,795
$MFW_{500} \setminus RFE_{D}$	0,526	0,593

Bei anderen Qualitätsmaßen als dem ARI, nämlich dem V-measure-score (Homogenität und Vollständigkeit) sowie dem mittleren Silhouettenkoeffizienten, ist der gleiche Trend zu bemerken.

Ergebnisse

Fazit

Hauptergebnis

- Partitionierung der MFW in zwei Teilmengen möglich:
  - Teilmenge 1: bessere Identifikation der Übersetzer als Gesamtmenge
  - Teilmenge 2: bessere Identifikation der Disziplinen als Gesamtmenge
- MFW tragen teils Übersetzer-, teils Disziplinsignal!

Weitere Ergebnisse

- rekursive Merkmalseliminierung wirksame Methode zur Bestimmung von übersetzer- und disziplintypischen Merkmalen
- Kondensierung der Wortliste als philologische Chance

Zukünftige Forschung

- Anwendung auf Mehrwortgruppen
- Anwendung auf andere Textkorpora
- unüberwachte Partitionierung der Merkmale

Gliederung

Literatur (1)

- Argamon, Shlomo (2008): „Interpreting Burrows's Delta: Geometric and Probabilistic Foundations.“ Literary and Linguistic Computing 23/2: 131–147. doi:10.1093/llc/fqn003
- Burrows, John (2002): „Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship.“ Literary and Linguistic Computing 17/3: 267–287. doi:10.1093/llc/17.3.267
- Eder, Maciej (2015): „Does size matter? Authorship attribution, small samples, big problem.“ Digital Scholarship Humanities 30/2: 167–182. doi:10.1093/llc/fqt066
- Eder, Maciej, Mike Kestemont, Jan Rybicki (2013): „Stylometry with R: a suite of tools.“ In: Digital Humanities 2013: Conference Abstracts. Lincoln: University of Nebraska, 487–489. http://dh2013.unl.edu/abstracts/ab-136.html

Literatur (2)

- Evert, Stefan, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Christof Schöch, Thorsten Vitt (2015): „Towards a better understanding of Burrows's Delta in literary authorship attribution.“ In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature. Association for Computational Linguistics, 79–88. `http://www.aclweb.org/anthology/W/W15/W15-0709.pdf`
- Guyon, Isabelle, Jason Weston, Stephen Barnhill, Vladimir Vapnik (2002): „Gene Selection for Cancer Classification using Support Vector Machines.“ Machine Learning 46/1: 389–422. `doi:10.1023/A:1012487302797`
- Hasse, Dag Nikolaus, Andreas Büttner (in Vorbereitung): „Notes on the Identity of the Latin Translator of Avicenna's Physics and on Further Anonymous Translations in Twelfth-Century Spain.“ Vorabversion: `https://go.uniwue.de/hassevigoni`

Literatur (3)

- Hoover, David L. (2004a): „Testing Burrows's Delta.“ Literary and Linguistic Computing 19/4: 453–475. doi:10.1093/llc/19.4.453
- Hoover, David L. (2004b): „Delta Prime?“ Literary and Linguistic Computing 19/4: 477–495. doi:10.1093/llc/19.4.477
- Jannidis, Fotis, Steffen Pielström, Christof Schöch, Thorsten Vitt (2015): „Improving Burrows' Delta - An empirical evaluation of text distance measures.“ In: Digital Humanities Conference 2015, Sydney. http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html
- Kestemont, Mike, Kim Luyckx, Walter Daelemans, Thomas Crombez (2012): „Cross-Genre Authorship Verification Using Unmasking.“ English Studies 93/3: 340–356. doi:10.1080/0013838X.2012.668793

Literatur (4)

- Rybicki, Jan (2012): „The great mystery of the (almost) invisible translator: stylometry in translation.“ In: M. Oakley and M. Ji (Hrsg.): Quantitative Methods in Corpus-Based Translation Studies. Amsterdam: John Benjamins, 231–248. https://sites.google.com/site/computationalstylistics/preprints/Rybicki%20Great%20Mystery.pdf
- Rybicki, Jan, Maciej Eder (2011): „Deeper Delta across genres and languages: do we really need the most frequent words?“ Literary and Linguistic Computing 26/3: 315–321. doi:10.1093/llc/fqr031
- Schöch, Christof (2013): „Fine-Tuning our Stylometric Tools: Investigating Authorship and Genre in French Classical Drama.“ In: Digital Humanities 2013: Conference Abstracts. Lincoln: University of Nebraska, 383–386. http://dh2013.unl.edu/abstracts/ab-270.html

Literatur (5)

- Smith, Peter W. H., Aldridge, W. (2011): „Improving Authorship Attribution: Optimizing Burrows' Delta Method.“ Journal of Quantitative Linguistics 18/1: 63–88. doi:10.1080/09296174.2011.533591
- Stamatatos, Efstathios, Nikos Fakotakis, George Kokkinakis (2000): „Automatic Text Categorization in Terms of Genre and Author.“ Computational Linguistics 26/4: 471–497. doi:10.1162/089120100750105920