# Eröffnung ZfdW

23. November 2018

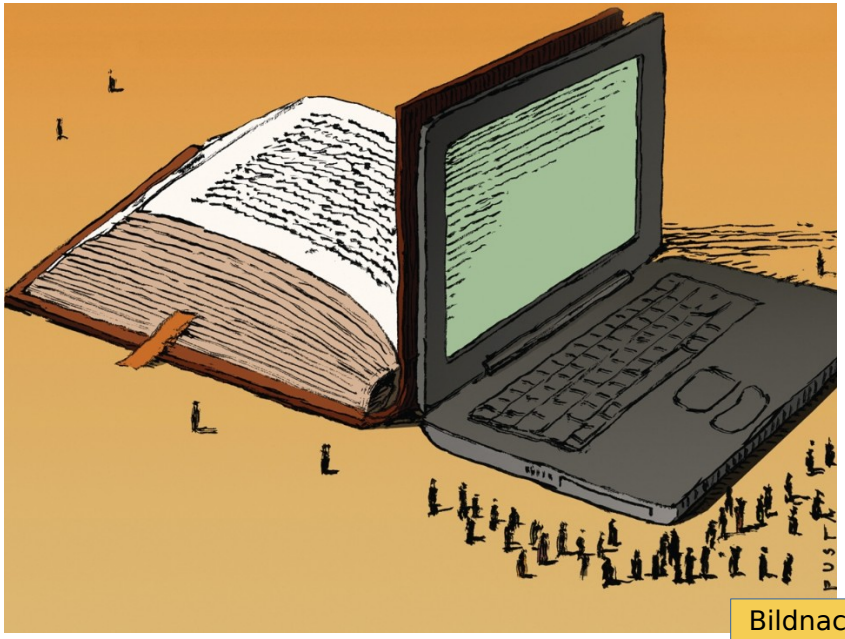
Fritz Kliche

# Übersicht

- Seminar „Wahlkampf in (a)sozialen Medien“
- Methoden der quantitativen Textanalyse
  - Termextraktion
  - Burrows‘ Zeta

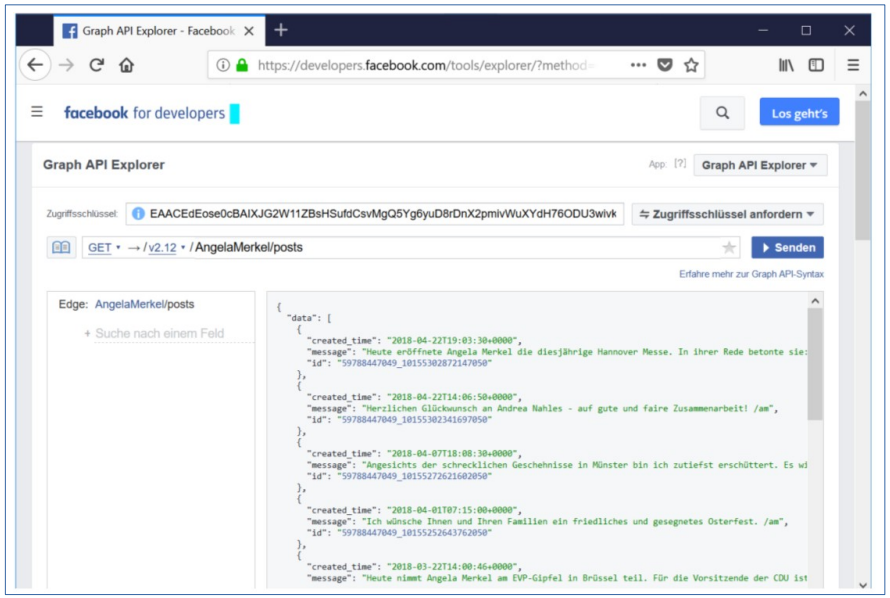
# Seminar „Wahlkampf in (a)sozialen Medien“

- Seminar in Anlehnung an laufendes Forschungsprojekt
- Kooperation Politikwissenschaft und Computerlinguistik
- Digitaler Wandel:
  - Computerlinguistische Methoden für politikwissenschaftliche Fragen
  - Konzept der Digital Humanities



Bildnachweis:
https://mitpress.mit.edu/books/war-learning

# Seminar Wa(s)M: Das Facebook-Korpus




# Seminar Wa(s)M: Das Facebook-Korpus

```
<doc id=294 subcorpus="AngelaMerkel" db_id="10152198616677050_1394383917241010" LOC="O"
PER="PER" ORG="O" MISC="O" date=201612" likecount=2 messagetype="comment" language="de"
length=143 cross="-" frequser=0 ptv=1 neg=1 sentdiff=0 sharecount=2>
<s>     O
Bitte   NN      Bitte   O
Frau    NN      Frau    O
Kanzlerin       NN      Kanzlerin       O
.       $.      .       O
Hilft   VVFIN   helfen  O
evtl.   NE      evtl.   O
Herrn   NN      Herr    O
Sch<C3><A4>uble NE      Sch<C3><A4>uble I-PER
aber    ADV     aber    O
hat     VAFIN   haben   O
leider  ADV     leider  O
die     ART     die     O
Linke   NN      Linke   O
auch    ADV     auch    O
schon   ADV     schon   O
festgestellt    VVPP    feststellen     O
```

**Begriffe**

- Korpus
- Sub-/Teilkorpus
- „Messages“ (Begriff von FB)
- Metadaten

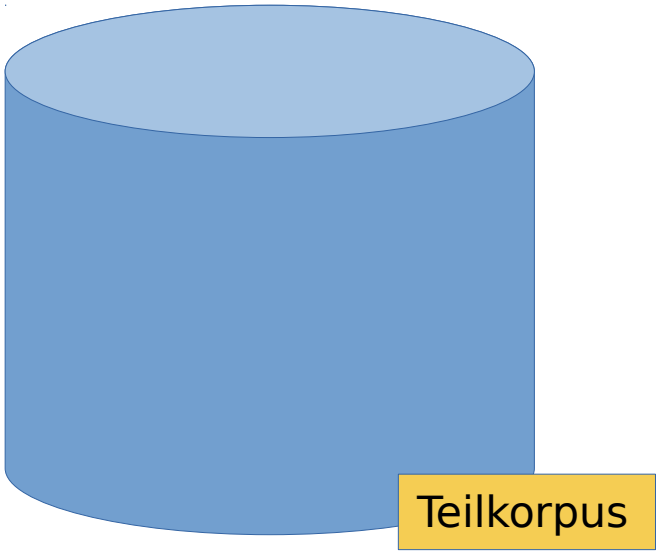
- Annotationen
  Im Beispiel verwendete Annotationen:
  - Wortart
  - Lemma
  - Eigenname

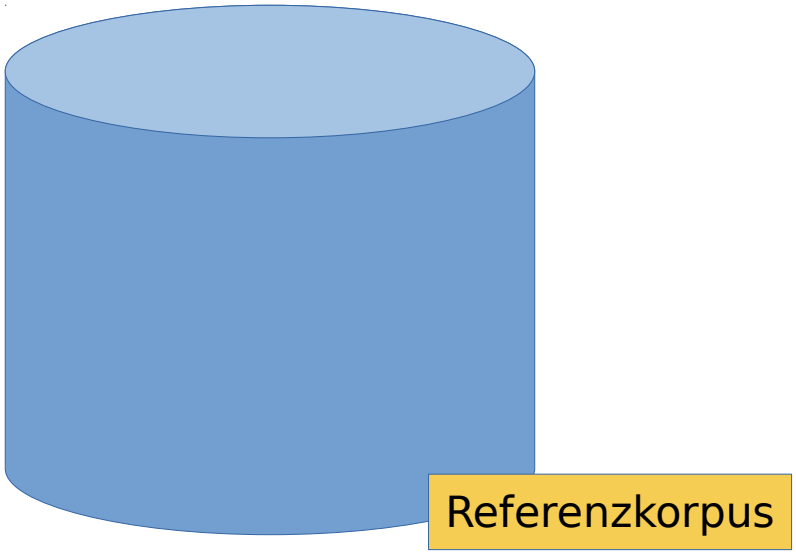
# Seminar Wa(s)M: Das Facebook-Korpus

- AngelaMerkel 530.520
- CDU 204.026
- CSU 246.238
- joachimherrmanncsu 3.053
- B90DieGruenen 127.847
- GoeringEckardt 31.016
- Cem 106.837
- linkspartei 120.629
- sahrawagenknecht 333.454
- DietmarBartschMdB 19.537
- FDP 100.273
- lindnerchristian 211.783
- SPD 156.911
- martinschulz 336.652
- alternativefuerde 761.036
- aliceweidel 121.634

- Messages insgesamt: 3.411.448
- Posts: 9.081
- Kommentare: 1.851.945
- Subkommentare: 1.550.420
- Tokens (d. h.: Wörter und Satzzeichen): 113.639.602

# Erste Idee: Termextraktion




$$\textit{Weirdness Ratio}(x) = \frac{\dfrac{\text{Anzahl aller Vorkommen von "Einkommen" im Teilkorpus}}{\text{Anzahl aller Vorkommen von [N.*] im Teilkorpus}}}{\dfrac{\text{Anzahl aller Vorkommen von "Einkommen" im Vergleichskorpus}}{\text{Anzahl aller Vorkommen von [N.*] im Vergleichskorpus}}}$$

Schäfer et al, 2015

# Erste Idee: Termextraktion

| ##### Partei und Spitzenkandidaten, mit cross="-", Pattern: [ADJ.*] [N.*] | | | | | |
|---|---|---|---|---|---|
| ##__linke | | ##__spd | | ##__gruene | |
| 1665 | sozial Gerechtigkeit [#2 | 1956 | sozial Gerechtigkeit [#1 | 433 | deutsch Politiker [#2086 |
| 1038 | falsch Partei [#75882-#7 | 1450 | letzt Jahr [#143913-#14 | 365 | türkisch Politik [#10637 |
| 698 | deutsch Volk [#44491-#4 | 1205 | lieb Herr [#148238-#149 | 352 | deutsch Politik [#20504- |
| 648 | letzt Jahr [#159153-#159 | 983 | geehrter Herr [#76467-#7 | 306 | deutsch Volk [#22539-#2 |
| 598 | einzig Partei [#64531-#6 | 883 | groß Koalition [#94347-# | 302 | letzt Jahr [#72953-#732 |
| 512 | bedingungslos Grundein | 835 | herzlich Glückwunsch [# | 255 | froh Weihnachten [#393 |
| 481 | herzlich Glückwunsch [# | 721 | leer Versprechung [#141 | 236 | lieb Grüne [#74528-#747 |
| 476 | geehrt Frau [#90929-#91 | 611 | klein Mann [#131940-#1 | 228 | eigen Land [#27787-#280 |
| 461 | groß Koalition [#111038- | 568 | lieb Martin [#149678-#15 | 210 | cem özdemir [#15807-#1 |
| 460 | frau Wagenknecht [#812 | 552 | gut Gruß [#103743-#104 | 203 | ganz Welt [#41513-#417 |
| 459 | groß Teil [#114035-#114 | 527 | deutsch Volk [#39442-#3 | 201 | herzlich Glückwunsch [# |
| 430 | ganz Welt [#89593-#900 | 484 | gut Mann [#106276-#106 | 172 | lieb Herr [#74787-#7495 |
| 410 | nah Wahl [#179289-#179 | 437 | jung Mensch [#125852-# | 171 | grün Partei [#52911-#530 |
| 409 | nah Jahr [#177971-#178 | 409 | heiß Luft [#113673-#114 | 170 | grün Politik [#53168-#53 |
| 401 | lieb Frau [#161729-#162 | 404 | bedingungslos Grundein | 170 | türkisch Volk [#107504-# |
| 393 | eigen Land [#56132-#569 | 391 | a | | |
| 391 | eigen Volk [#58997-#593 | 362 | v | | |
| 391 | etabliert Partei [#71968- | 333 | k | | |
| 386 | klein Mann [#147141-#14 | 331 | l | | |
| 376 | eigen Partei [#57273-#57 | 320 | e | | |
| 301 | gut Idee [#120246-#1205 | 318 | k | | |
| 300 | links Politik [#167558-#1 | 315 | c | | |

| ##__fdp | | ##__cdu | | ##__csu | | ##__afd | |
|---|---|---|---|---|---|---|---|
| 1446 | herzlich Glückwunsch [ | 3743 | herzlich Glückwunsch [ | 440 | herzlich Glückwunsch [# | 2386 | deutsch Volk [#77816-#8 |
| 822 | lieb Herr [#106501-#1073 | 2567 | deutsch Volk [#64946-#( | 339 | heiß Luft [#32936-#3327 | 1606 | herzlich Glückwunsch [# |
| 681 | frei Demokrat [#48743-#1 | 1043 | frau merkel [#106590-#1 | 292 | letzt Jahr [#41976-#4226 | 879 | eigen Volk [#101466-#102 |
| 491 | letzt Jahr [#102332-#102 | 1042 | letzt Jahr [#201187-#201 | 252 | deutsch Volk [#12878-#1 | 834 | letzt Jahr [#231798-#2326 |
| 472 | geehrter Herr [#54460-#5 | 1012 | lieb Frau [#205231-#206 | 245 | nah Jahr [#46318-#4656 | 824 | eigen Land [#97439-#982 |
| 400 | gut Mann [#75151-#7555 | 975 | geehrt Frau [#120534-#1 | 224 | nah Wahl [#46796-#4701 | 784 | etabliert Partei [#118547- |
| 396 | gut Idee [#73825-#74220 | 917 | ganz Welt [#118491-#11 | 192 | deutsch Autofahrer [#10 | 739 | nah Jahr [#255737-#2564 |
| 353 | wichtig Thema [#169202 | 890 | eigen Volk [#83599-#844 | 165 | inner Sicherheit [#35839 | 666 | arm Deutschland [#3554 |
| 324 | groß Koalition [#66079-# | 803 | gut Frau [#150057-#1508 | 158 | eigen Land [#15911-#160 | 664 | deutsch Sprache [#7493 |
| 291 | jung Mensch [#90519-#9 | 733 | froh Weihnachten [#1118 | 146 | doppelt Staatsbürgersch | 654 | AfD Wähler [#7409-#8062 |
| 272 | inner Sicherheit [#87229 | 691 | neu Jahr [#226067-#226 | 143 | lieb Herr [#43156-#4329 | 576 | ganz Welt [#141080-#141 |
| 224 | nah Jahr [#114206-#114 | 656 | eigen Land [#80430-#810 | 134 | arm Deutschland [#5060 | 563 | deutsch Bürger [#65401-# |
| 222 | gut Abend [#69779-#700 | 596 | besorgt Bürger [#40965-# | 131 | eigen Volk [#16666-#167 | 461 | einzig Partei [#109059-#1 |
| 211 | liberal Partei [#104975-# | 595 | nah Jahr [#219154-#219 | 118 | deutsch Bürger [#11269 | 458 | groß Teil [#169839-#1702 |
| 201 | erst Mal [#41833-#42033 | 576 | inner Sicherheit [#17548 | 110 | ganz Welt [#23629-#237 | 453 | deutsch Pass [#71958-#7 |
| 199 | groß Problem [#67054-# | 469 | groß Teil [#144509-#1449 | 104 | deutsch Pass [#12055-# | 453 | eigen Meinung [#98755-# |
| 192 | richtig Weg [#134846-#1 | 433 | WARNING Frau [#16986 | 103 | lieb CSU [#42794-#4289 | 434 | gut Nacht [#180318-#180 |
| 187 | kalt Progression [#91398 | 433 | deutsch Bürger [#56728- | 100 | klar Wort [#38061-#3816 | 433 | nah Wahl [#257807-#258 |
| 186 | sozial Marktwirtschaft [ | 394 | jung Mann [#182215-#18 | 98 | geehrter Herr [#24056-#2 | 408 | jung Mann [#208270-#208 |
| 179 | gut Bildung [#71467-#71 | 382 | froh Ostern [#111352-#1 | 98 | leer Versprechung [#414 | 387 | deutsch Kultur [#69626-# |
| 177 | erst Linie [#41647-#4182 | 361 | gut Kanzlerin [#152542-# | 83 | wichtig Thema [#66944-# | 387 | dumm Mensch [#87893-# |
| 177 | sozial Gerechtigkeit [#1 | 347 | offen Grenze [#233116-# | 81 | etabliert Partei [#19679-# | 379 | gut Beispiel [#174606-#1 |

# Zweite Idee: Burrows‘ Zeta

- Ausgangspunkt: Stilometrie
  - Die Stilometrie bezeichnet „die Anwendung quantitativer Methoden zur Erfassung und Klassifizierung stilistischer Merkmale von Texten“.

Viehhauser, 2015

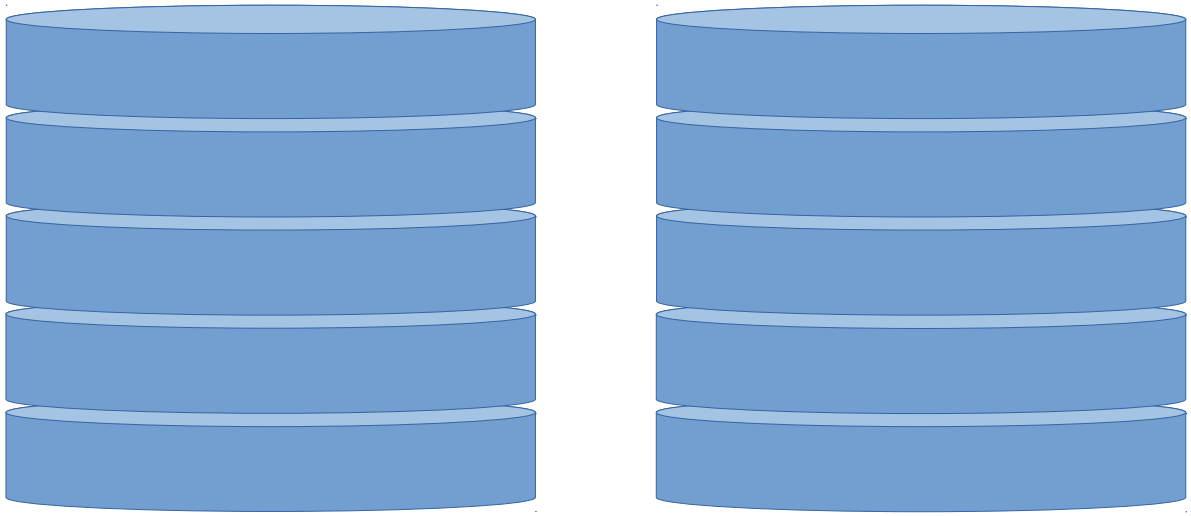
- Eine Anwendung ist *authorship attribution*,
  d. h. die Zuordnung eines Textes, dessen Autor unbekannt oder unsicher ist, zu einem Autor mit ähnlichen stilistischen Merkmalen.

- Ein Maß für den kontrastiven Vergleich zweier Korpora/Texte mit dem Ziel einer stilometrischen Analyse ist Burrows‘ Zeta.

Schöch, 2018

# Zweite Idee: Burrows‘ Zeta

- Die zu vergleichenden Korpora/Texte werden – im Gegensatz zum vorgestellten Verfahren zur Termextraktion – zunächst in Einheiten segmentiert, beispielsweise in Abschnitte einheitlicher Länge (in Tokens).



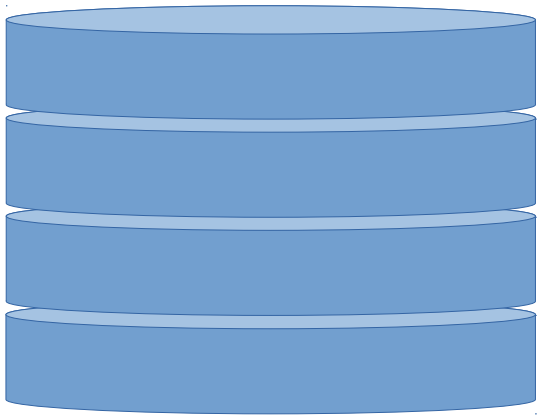
# Zweite Idee: Burrows‘ Zeta

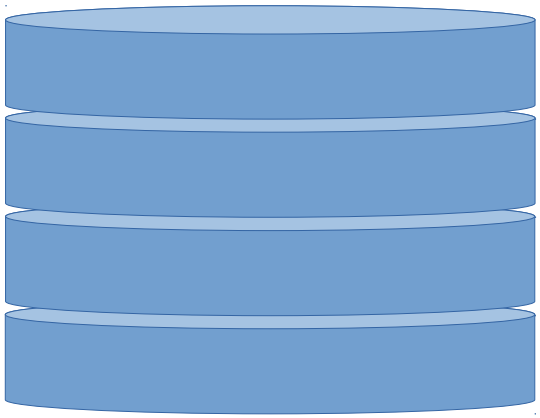
- Die häufigsten Tokens der Textdaten werden ermittelt.
- Für jedes dieser Tokens wird für beide zu vergleichenden Korpora/Texte bestimmt, in wie vielen Segmenten sie mindestens einmal vorkommen.

$$dp_i(Z) = \frac{df_i(Z)}{n(Z)} \quad \text{bzw.} \; dp_i(V) = \frac{df_i(V)}{n(V)}$$

wobei

| | |
|---|---|
| $dp_i(Z)$: | die *document proportions* des Tokens *i* im Korpus/Text *Z*. |
| $df_i(Z)$: | die *document frequency* des Tokens *i* im Korpus/Text *Z.* |
| n(Z): | die Anzahl der Segmente in Z |
| V: | das Vergleichskorpus/der Vergleichstext |





# Zweite Idee: Burrows‘ Zeta

- Damit ergibt sich Zeta:

$$Zeta_i = dp_i(Z) - dp_i(V)$$

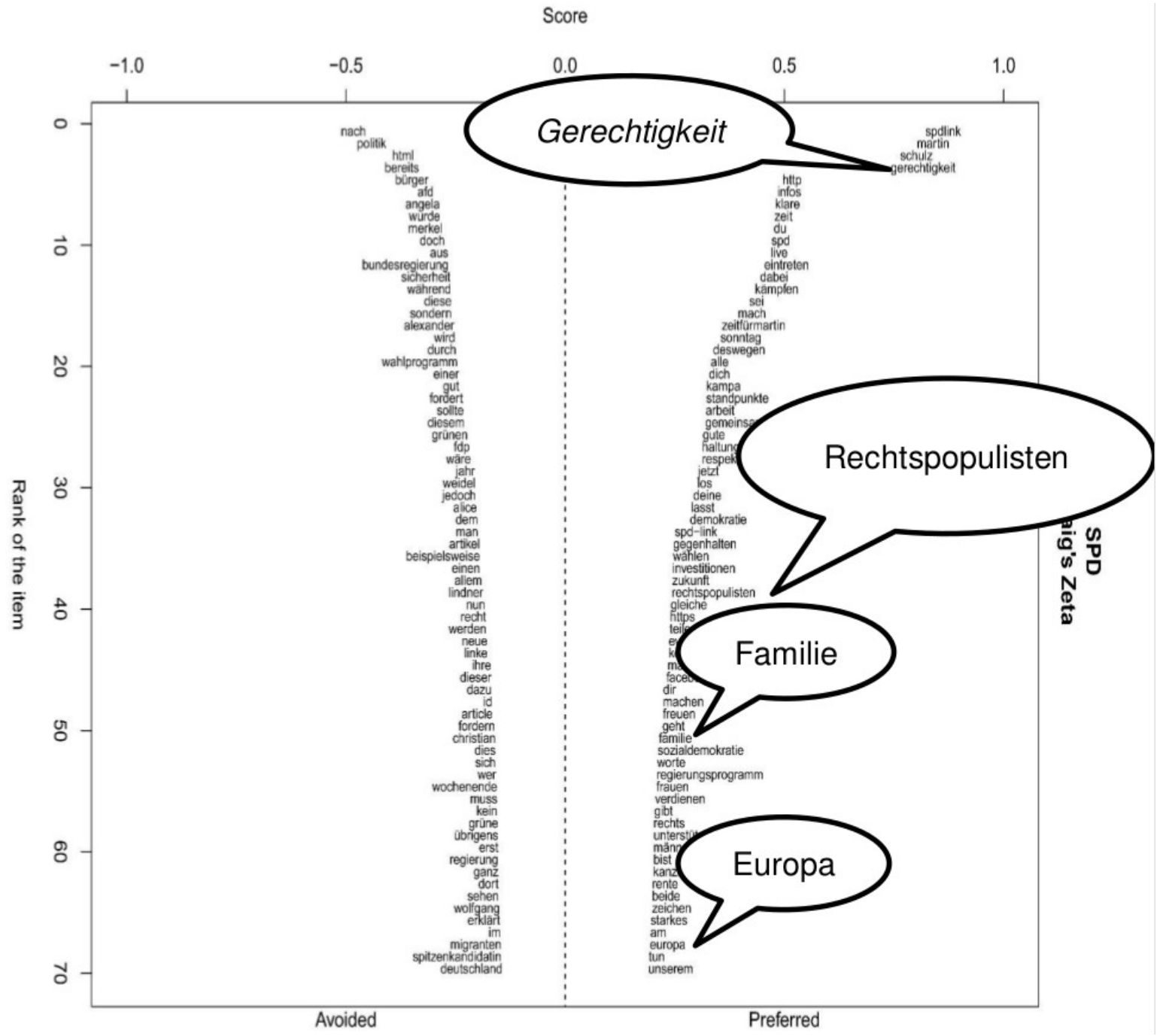
- Zeta liefert Ergebnisse von -1 bis 1.
- In den resultierenden Wert fließt vor allem ein, wie konsistent ein Token verwendet wird, und nicht allein, wie häufig es vorkommt.

# Zweite Idee: Burrows‘ Zeta

- Umsetzung mit STYLO
  - R-Paket zur stilometrischen Analyse. Eder et al., 2016
  - Bedingung:
    - Die Texte liegen in *plain text* vor.
    - Die Texte sind lemmatisiert.

  - Hinweis: *stopword removal* ist nicht erforderlich.

# Einige Ergebnisse



# Einige Ergebnisse

Vergleich der Ergebnisse
Termextraktion – Burrows' Zeta

| ##__afd | |
|---|---|
| 2386 | deutsch Volk [#77816-#8 |
| 1606 | herzlich Glückwunsch [# |
| 879 | eigen Volk [#101466-#10 |
| 834 | letzt Jahr [#231798-#232 |
| 824 | eigen Land [#97439-#982 |
| 784 | etabliert Partei [#118547- |
| 739 | nah Jahr [#255737-#2564 |
| 666 | arm Deutschland [#3554 |
| 664 | deutsch Sprache [#7493 |
| 654 | AfD Wähler [#7409-#8062 |
| 576 | ganz Welt [#141080-#141 |
| 563 | deutsch Bürger [#65401- |
| 461 | einzig Partei [#109059-#1 |
| 458 | groß Teil [#169839-#1702 |
| 453 | deutsch Pass [#71958-#7 |
| 453 | eigen Meinung [#98755-# |
| 434 | gut Nacht [#180318-#180 |
| 433 | nah Wahl [#257807-#258 |
| 408 | jung Mann [#208270-#208 |
| 387 | deutsch Kultur [#69626-# |
| 387 | dumm Mensch [#87893-# |
| 379 | gut Beispiel [#174606-#1 |

# Fragen

- Erweiterung der Merkmale:
  - Statt Tokens:
  - N-grams?
  - POS-Patterns?
  - Weitere Merkmale?

- Anwendung computerlinguistischer Methoden in Seminaren
  - Umgang mit R
  - Umgang mit der Kommandozeile
  - Verwendung bestehender Tools

# Übersicht


- Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. R Journal, 8(1): 107-121.
- Schäfer, J., Rösiger, I., Heid, U. und Dorna, M. (2015). Evaluating noise reduction strategies for terminology extraction. In den *Proceedings der TIA 2015*. Granada.
- Schöch, C. (2018). Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie. In T. Bernhart u. a. (Hrsg.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven* (S. 77–94). Berlin: De Gruyter.
- Viehhauser, G. (2015). Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte. In C. Baum und T. Stäcker (Hrsg.): *Grenzen und Möglichkeiten der Digital Humanities*. *Sonderband der Zeitschrift für digitale Geisteswissenschaften.*