Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Burrows' „Delta“ verstehen

Stefan Evert, Thomas Proisl
Friedrich-Alexander-Universität Erlangen-Nürnberg

Fotis Jannidis, Steffen Pielström, Isabella Reger, Christof Schöch, Thorsten Vitt
Julius-Maximilians-Universität Würzburg

<philtag n="13"/>, Würzburg, 26. Februar 2016

Authorship attribution

(Juola 2006; Koppel *et al.* 2008; Stamatatos 2009)

- Vordergründiges Ziel: Identifikation unbekannter oder strittiger Autoren anhand stilistischer Merkmale
  - Federalist Papers: Hamilton vs. Madison (Mosteller and Wallace 1963)
  - Gab es Shakespeare wirklich?
  - Robert Galbraith (*The Cuckoo's Calling*) = J. K. Rowling
    http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/

- Interessanter: Was sind die charakteristischen stilometrischen Merkmale eines bestimmten Schriftstellers?
  - ☞ Merkmale, die besonders zur erfolgreichen Attribution beitragen

- Können auf diese Weise auch gattungsspezifische Merkmale gefunden und von autorenspezifischen unterschieden werden?

Authorship attribution

(Juola 2006; Koppel *et al.* 2008; Stamatatos 2009)

- Autorschaftsattribution als Klassifikation
  - vorgegebene Liste von in Frage kommenden Autoren
  - Trainingsdaten: bekannte Texte dieser Autoren
  - überwachte Lernverfahren ➜ relevante Merkmale + Gewichtung
  - Evaluation: Genauigkeit (accuracy)

- Autorschaftsattribution als Clustering
  - unüberwachtes Lernverfahren auf Basis einer Textsammlung
  - Texte des gleichen Autors sollen in ein Cluster gruppiert werden
  - Basis: stilometrisches Abstandsmaß für Texte
  - Evaluation: adjusted Rand index (ARI, Hubert and Arabie 1985)

  ☞ kann auch zur Klassifikation genutzt werden (*nearest neighbour*)

- Einfaches und erfolgreiches Abstandsmaß: Burrows Delta

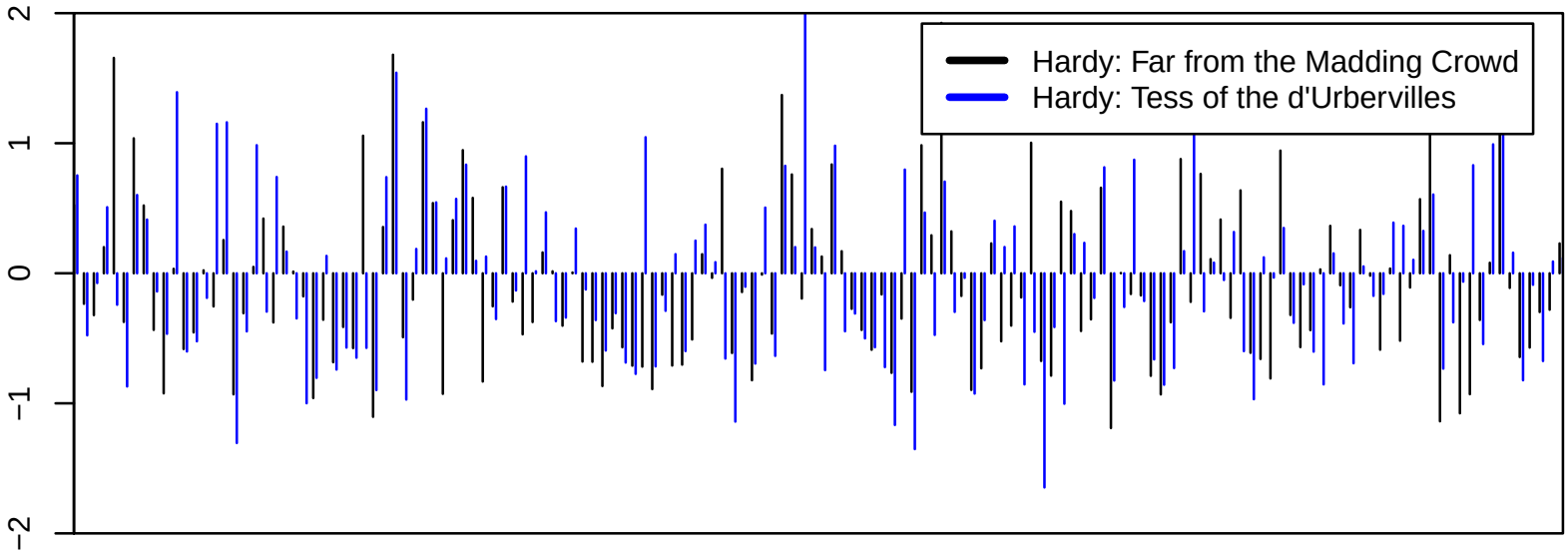
Burrows Delta (Δ_B)

(Burrows 2002)

- Relative Häufigkeiten der 100–5000 häufigsten Wörter (MFW) als stilistischer „Fingerabdruck“ eines Autors

standardisierte z-Werte: $$z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i}$$

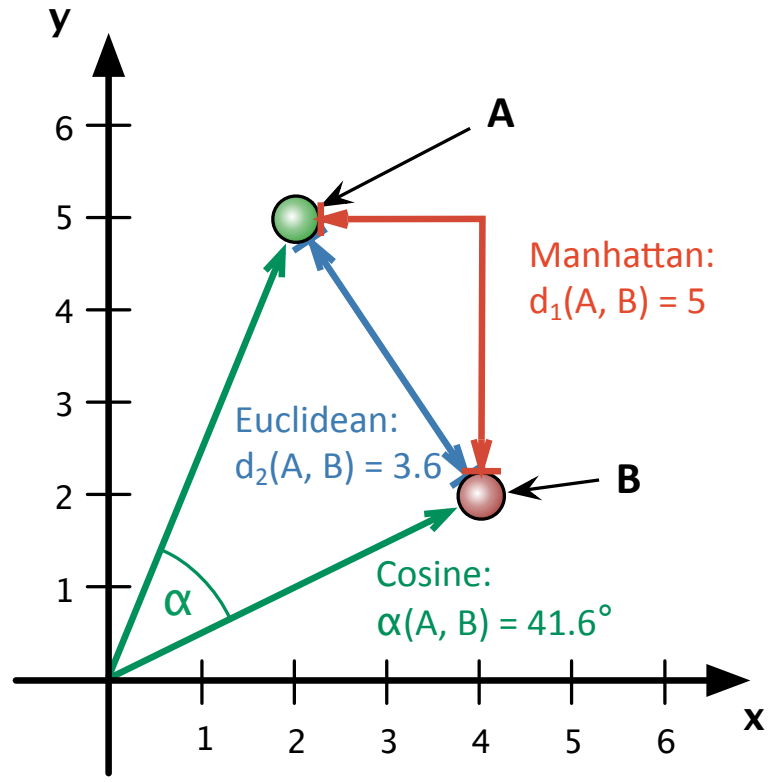
	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*	
z(Madding Crowd) = (	.53,	-.23,	-.32,	.20,	1.66,	-.37,	1.04,	.52,	-.44,	-.92,	.03,	…)
z(Tess of the d'U.) = (	.75,	-.48,	-.08,	.51,	-.24,	-.87,	.60,	.41,	-.14,	-.47,	1.39,	…)
z(Oliver Twist) = (	1.05,	.15,	-.71,	-.56,	.37,	-1.01,	-.06,	-.74,	-.28,	.48,	-.94,	…)



Abstandsmaße für Texte

- Ähnlichkeit zwischen zwei „Fingerabdrücken“, d.h. Vektoren $\mathbf{z}(D_1)$ und $\mathbf{z}(D_2)$, muss quantifiziert werden
- Vektor = Punkt in einem n_w-dimensionalen Raum
  - n_w = Anzahl MFW = Dimensionalität des Vektors

Metrik = geometrisches Abstandsmaß



Die Delta-Familie

(Burrows 2002; Hoover 2004; Argamon 2008; Smith and Aldridge 2011)

- Burrows Delta = Manhattan-Abstand (Burrows 2002)

$$\Delta_B(D_1, D_2) = \|\mathbf{z}(D_1) - \mathbf{z}(D_2)\|_1 = \sum_{i=1}^{n_w} |z_i(D_1) - z_i(D_2)|$$

- Quadratic Delta = Euklidischer Abstand (Argamon 2008)

$$\Delta_Q(D_1, D_2) = \|\mathbf{z}(D_1) - \mathbf{z}(D_2)\|_2^2 = \sum_{i=1}^{n_w} (z_i(D_1) - z_i(D_2))^2$$

- Cosine Delta = Winkelabstand (Smith and Aldridge 2011)

$$\cos \Delta_\angle(D_1, D_2) = \frac{\sum_{i=1}^{n_w} z_i(D_1) \cdot z_i(D_2)}{\|\mathbf{z}(D_1)\|_2 \cdot \|\mathbf{z}(D_2)\|_2}$$

Delta verstehen

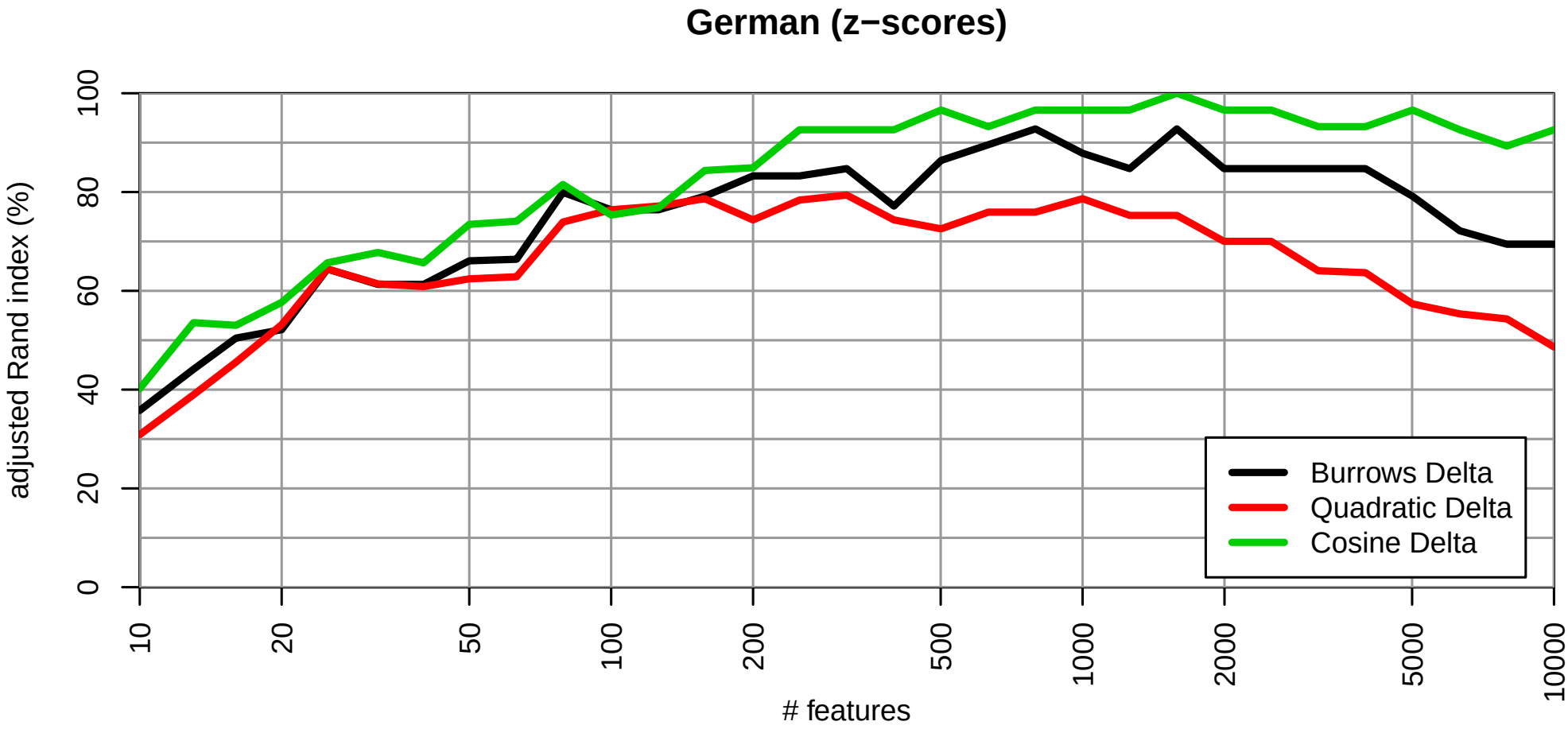
„In theory, theory and practice are the same. In practice, they are not.“

- Empirische Studie zu Delta-Maßen (Jannidis *et al.* 2015)
  - drei Romankorpora (Deutsch, Englisch, Französisch)
  - jeweils 75 Romane von 25 verschiedenen Autoren
  - Zeitraum: frühes 19. Jhd. bis Mitte des 20. Jhd.
  - unüberwachtes Clustering in 25 Gruppen mit allen bekannten Delta-Varianten und $n_w = 100, 1000, 5000$ MFW
  - ☞ Beste Maße: Burrows Delta Δ_B und Cosine Delta $\Delta_\angle$

- Gemeinsame Folgestudie (Evert *et al.* 2015)
  - systematische Untersuchung von n_w und anderen Parametern
  - besseres Clusteringverfahren, Evaluation mit ARI
  - ☞ Längennormalisierung spielt entscheidende Rolle

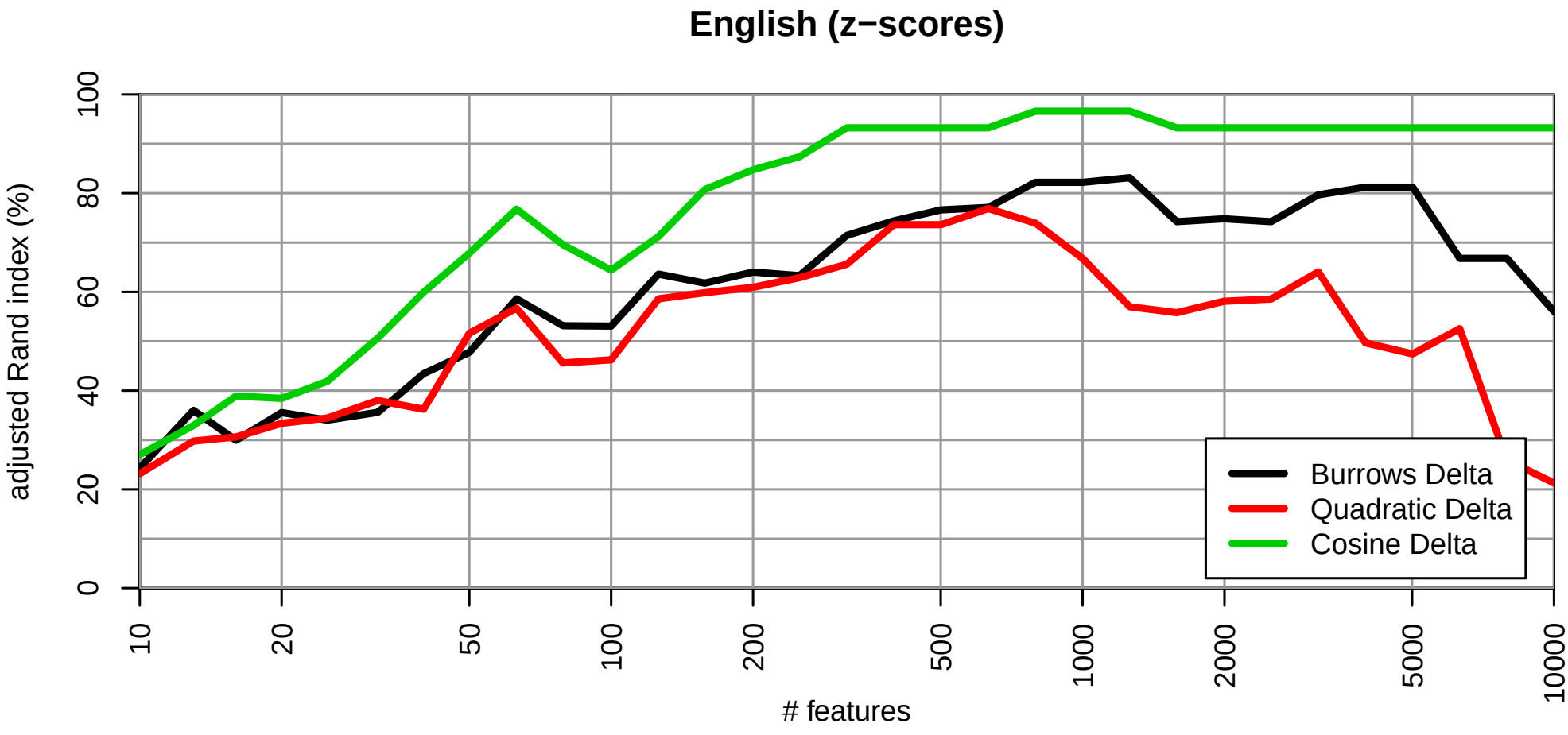
- Neue Ergebnisse (philtag / DHd 2016)
  - Gründe für positive Auswirkung der Längennormalisierung
  - Hypothesen: (i) Outlier vs. (ii) „Schlüsselprofil“

Parameter: Anzahl n_w von MFW



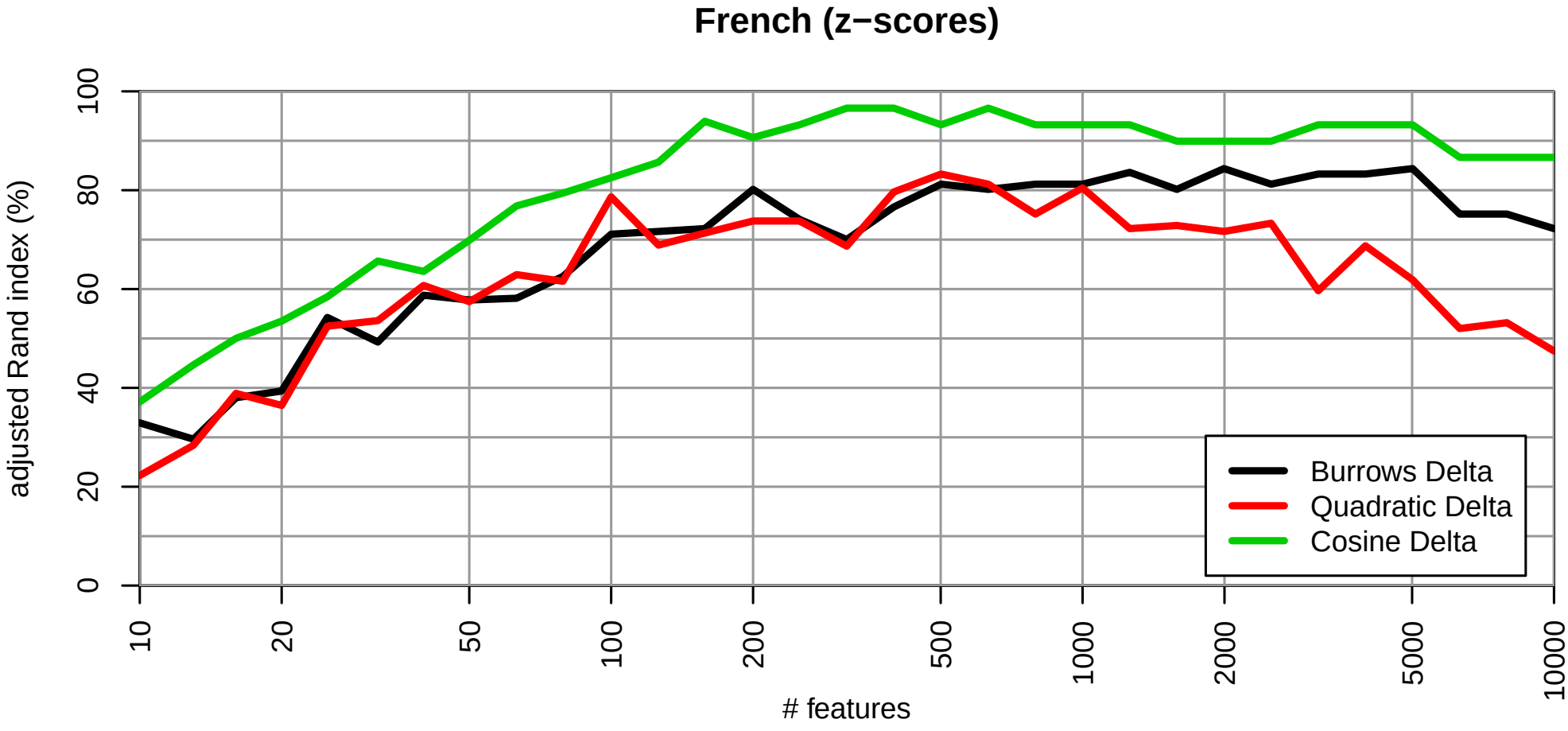
(im folgenden nur noch Ergebnisse für deutsches Romankorpus)

Parameter: Anzahl n_w von MFW



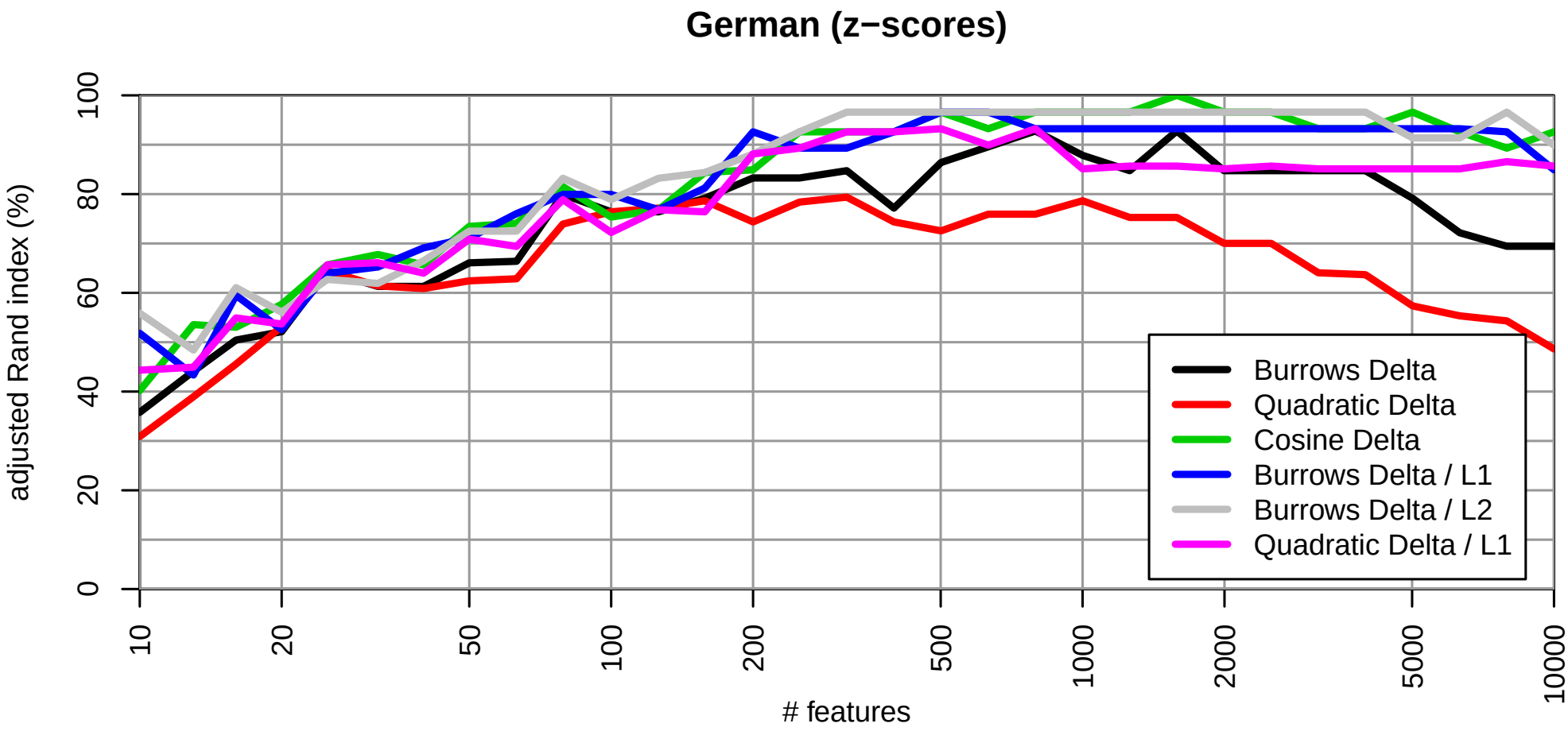
(im folgenden nur noch Ergebnisse für deutsches Romankorpus)

Parameter: Anzahl n_w von MFW



(im folgenden nur noch Ergebnisse für deutsches Romankorpus)

Parameter: Längennormalisierung der Vektoren



Längennormalisierung ist entscheidend für gutes Clustering!

Warum ist Längennormalisierung so wichtig?

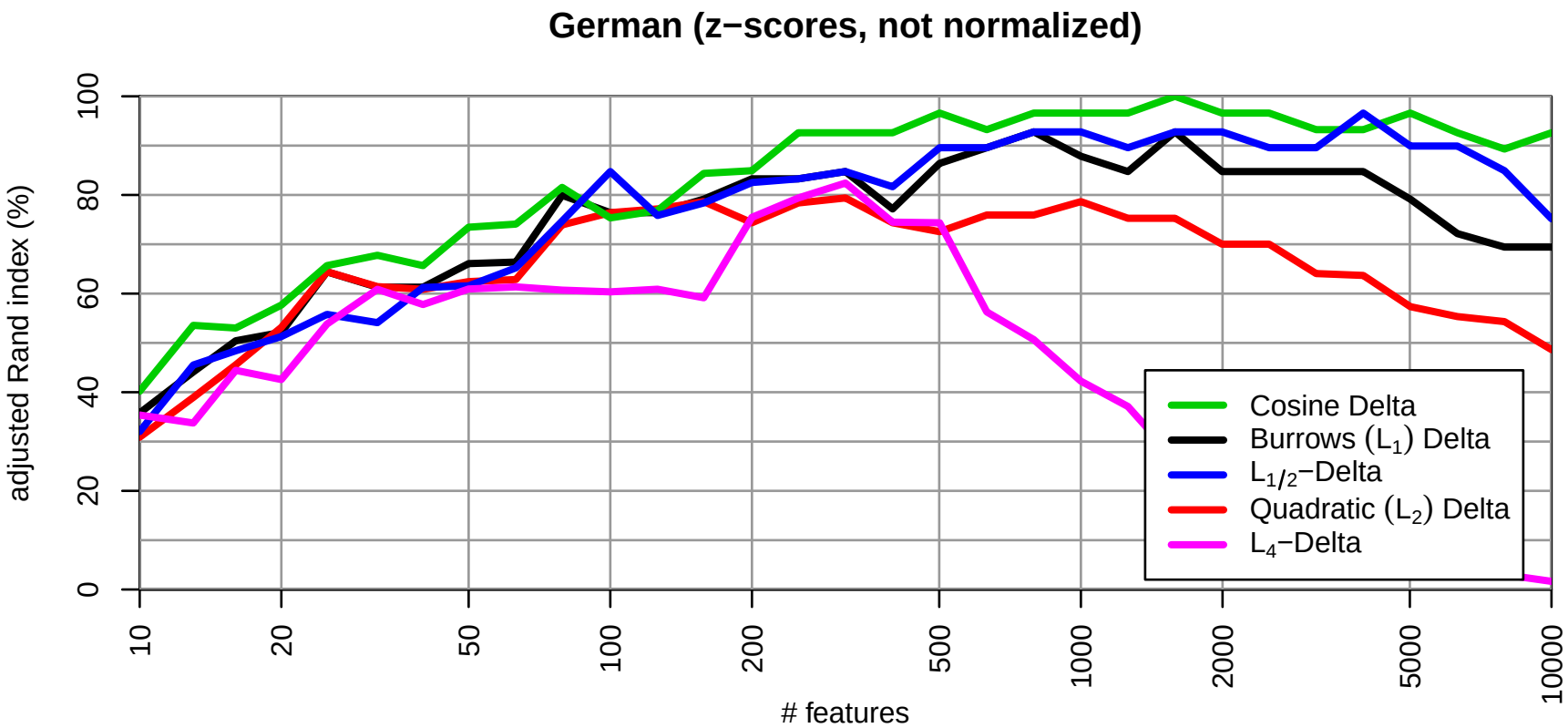
- Hypothese 1: wenige „extreme“ z-Werte (outlier) üben großen Einfluss auf Abstandsmaße aus und verzerren die Ergebnisse
  - derartige Ausreißer sind oft charakteristisch für einzelnen Text

- Hypothese 2: charakteristisches Autorenprofil findet sich im grundsätzlichen Muster positiver und negativer Abweichungen, unabhängig vom Grad ihrer Ausprägung („Schlüsselprofil“)

Evidenz für Hypothese 1: Minkowski-Abstand

- Minkowski-Delta als Verallgemeinerung: $\Delta_B = \Delta_1$, $\Delta_Q = \Delta_2$
- Δ_p ist umso anfälliger für Ausreißer, je größer p wird

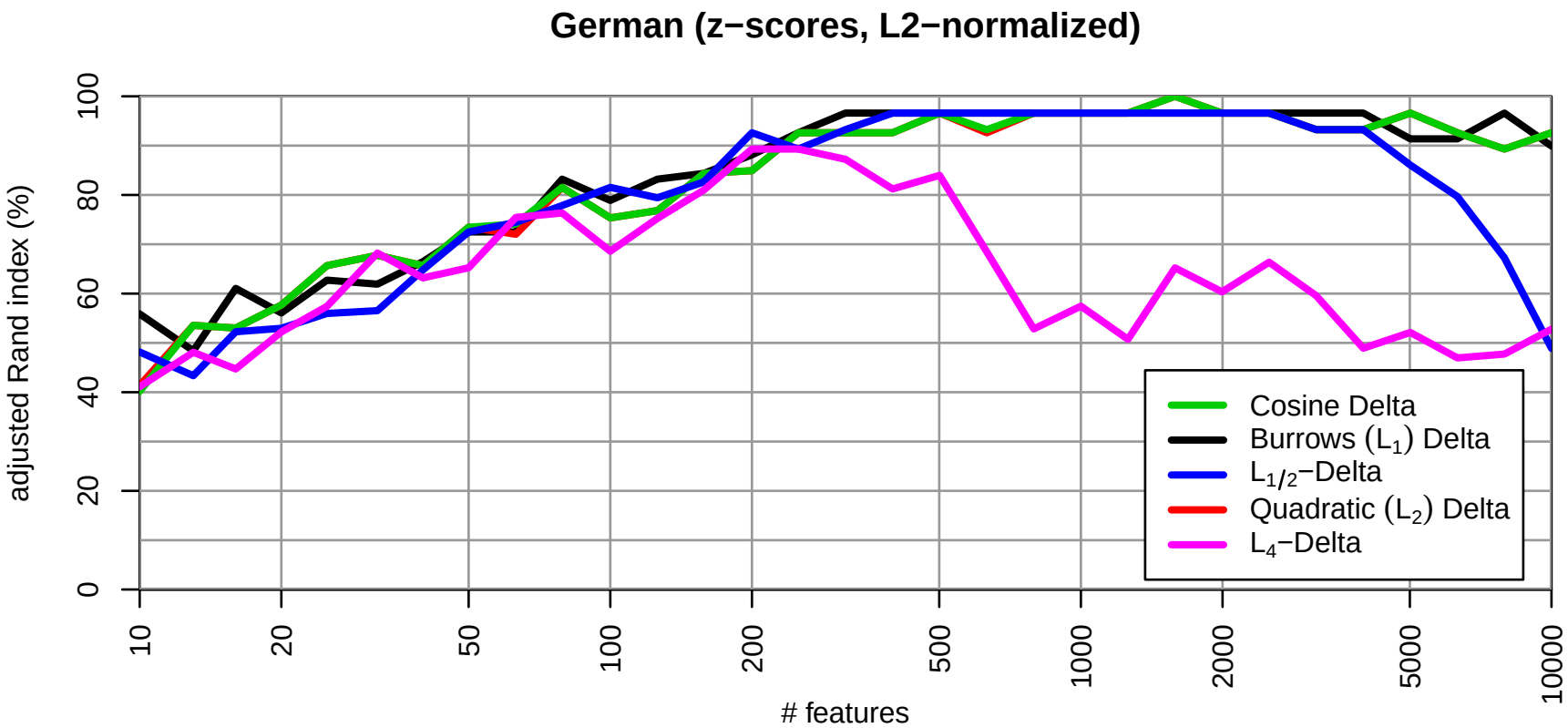
$$\Delta_p(D_1, D_2) = \|\mathbf{z}(D_1) - \mathbf{z}(D_2)\|_p = \sqrt[p]{\sum_{i=1}^{n_w} |z_i(D_1) - z_i(D_2)|^p}$$



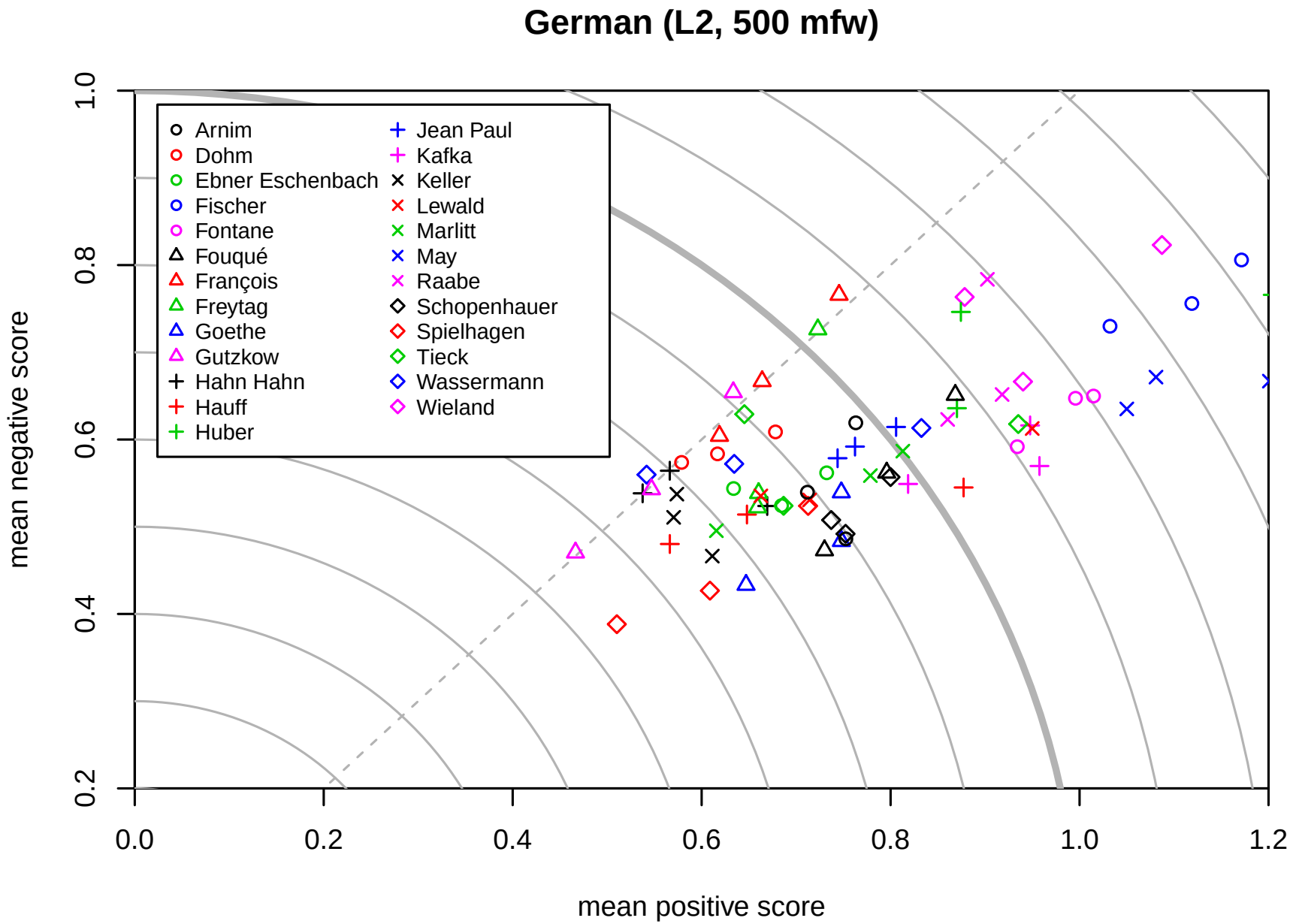
Evidenz für Hypothese 1: Minkowski-Abstand

- Minkowski-Delta als Verallgemeinerung: $\Delta_B = \Delta_1$, $\Delta_Q = \Delta_2$
- Δ_p ist umso anfälliger für Ausreißer, je größer p wird

$$\Delta_p(D_1, D_2) = \|\mathbf{z}(D_1) - \mathbf{z}(D_2)\|_p = \sqrt[p]{\sum_{i=1}^{n_w} |z_i(D_1) - z_i(D_2)|^p}$$

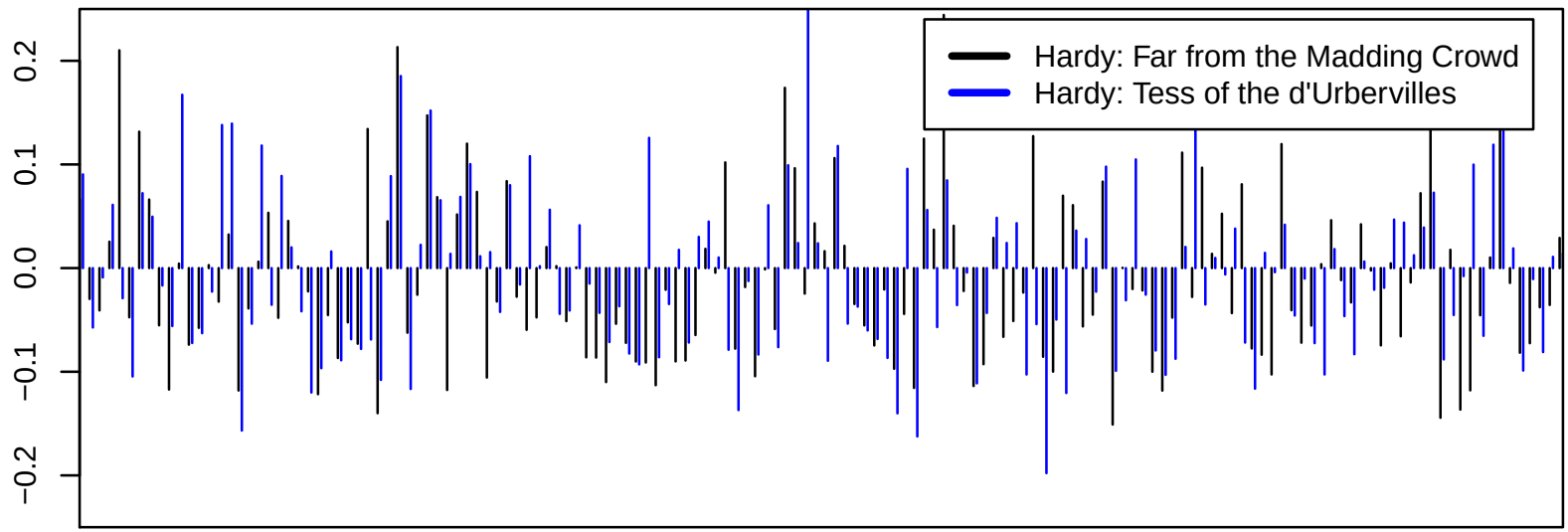


Evidenz für Hypothese 2: Vektorlänge (Norm)



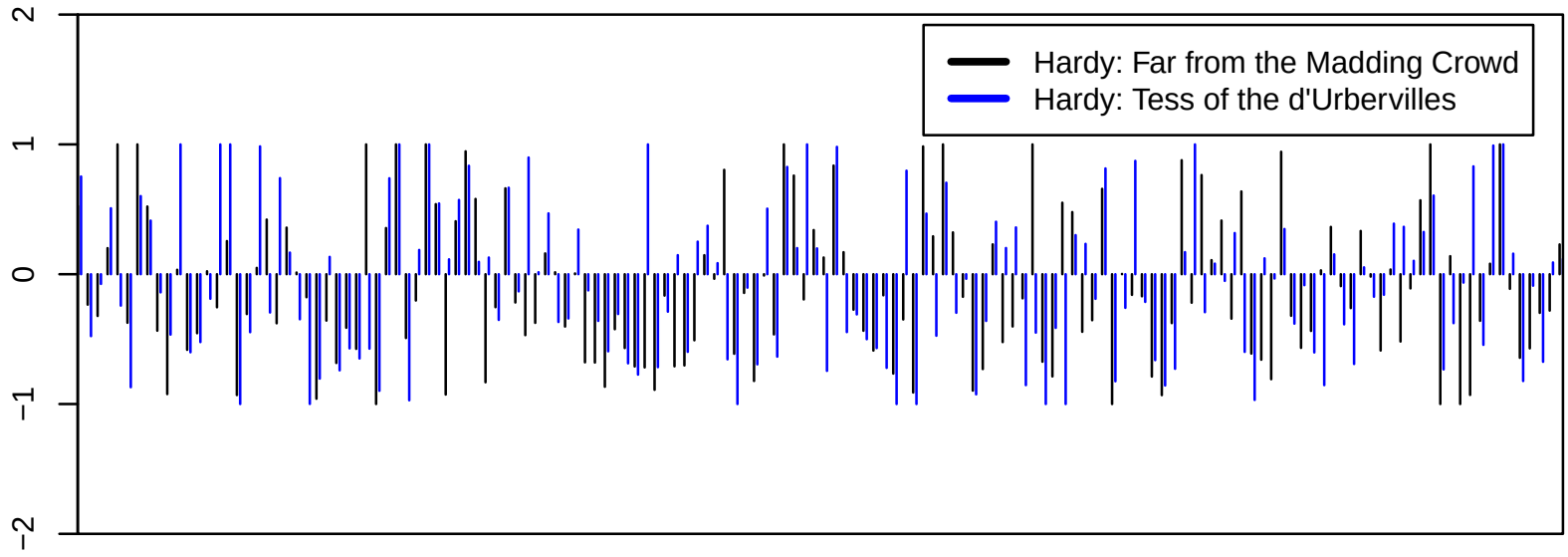
Neuer Ansatz: Transformation der z-Werte

- Längennormalisierung der Vektoren



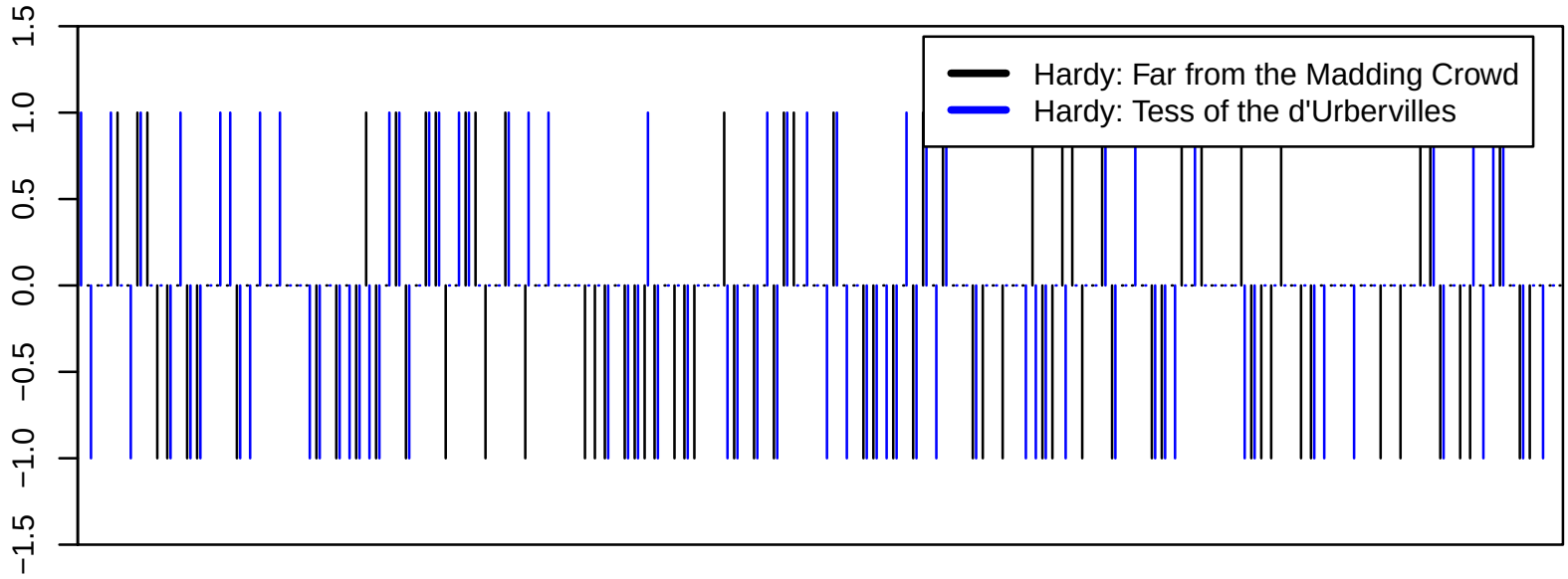
Neuer Ansatz: Transformation der z-Werte

- Abschneiden von Ausreißern (z.B. $|z| > 1$) H1



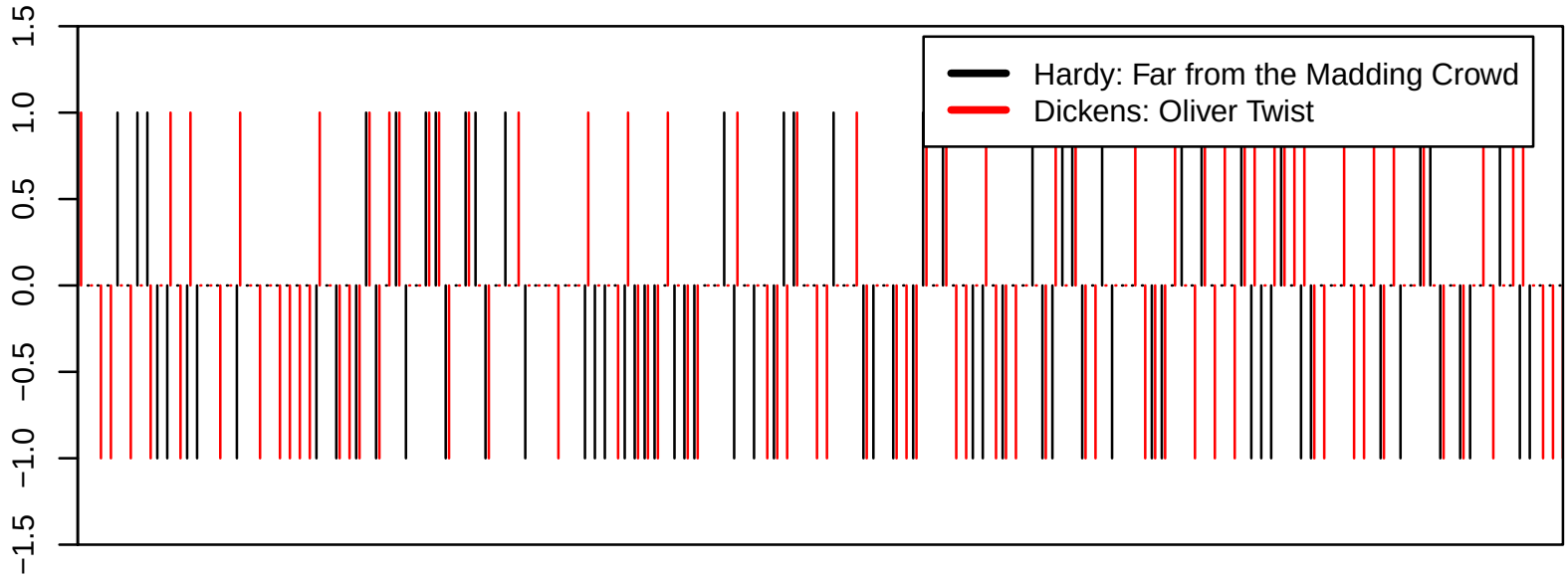
Neuer Ansatz: Transformation der z-Werte

- Ternarisierung: -1 (selten) / 0 (normal) / $+1$ (häufig) H2

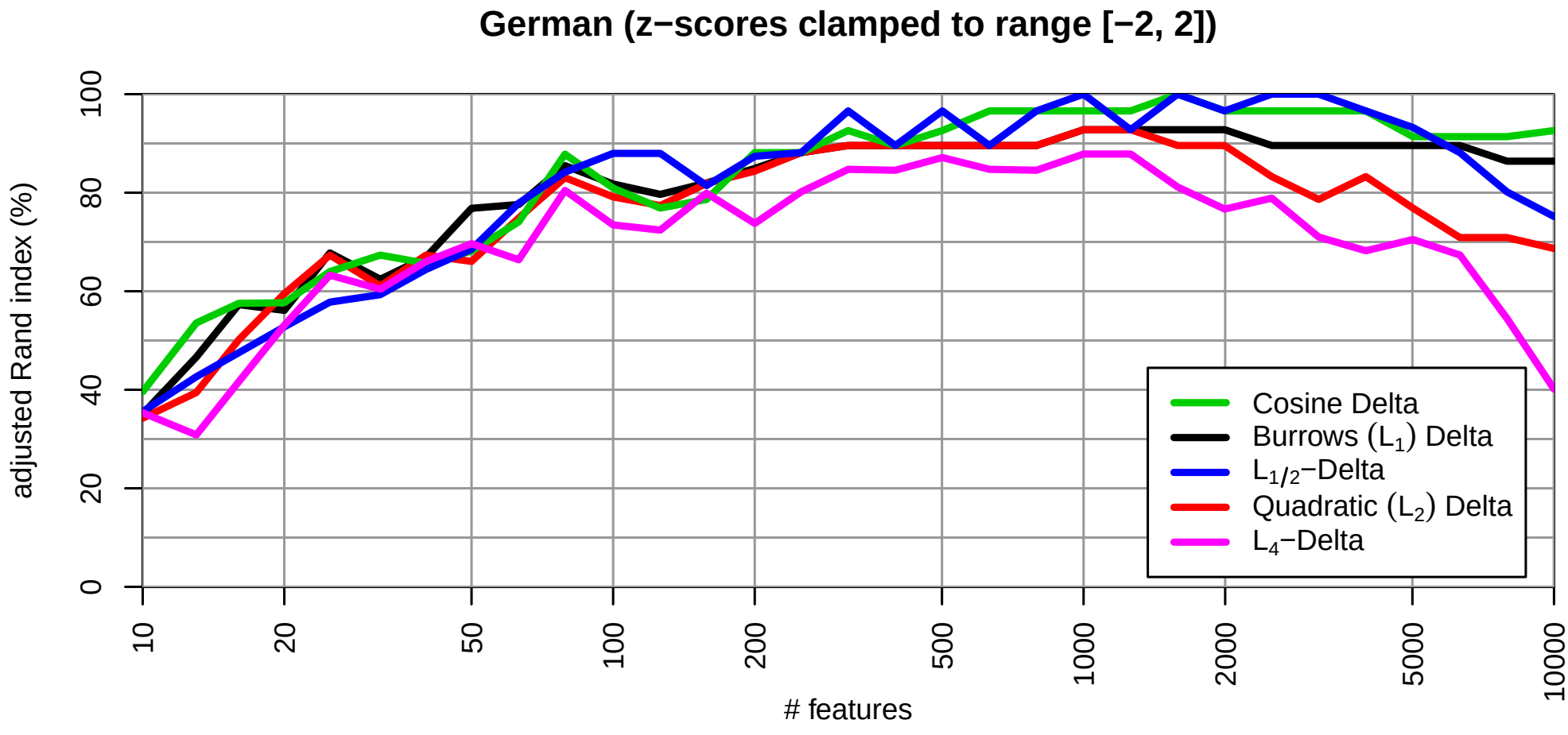


Neuer Ansatz: Transformation der z-Werte

- Ternarisierung: -1 (selten) / 0 (normal) / $+1$ (häufig) H2

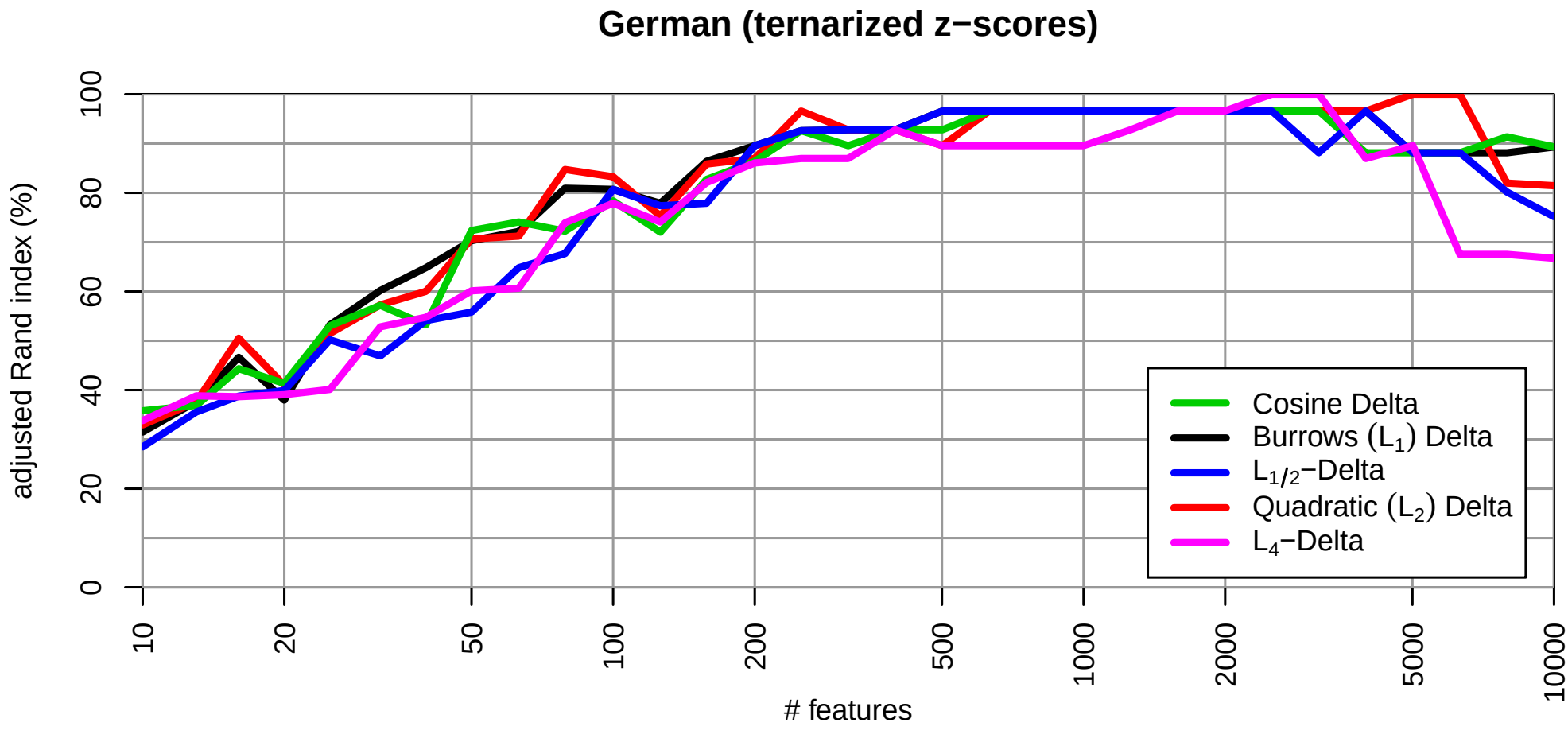


Parameter: Transformation der z-Werte



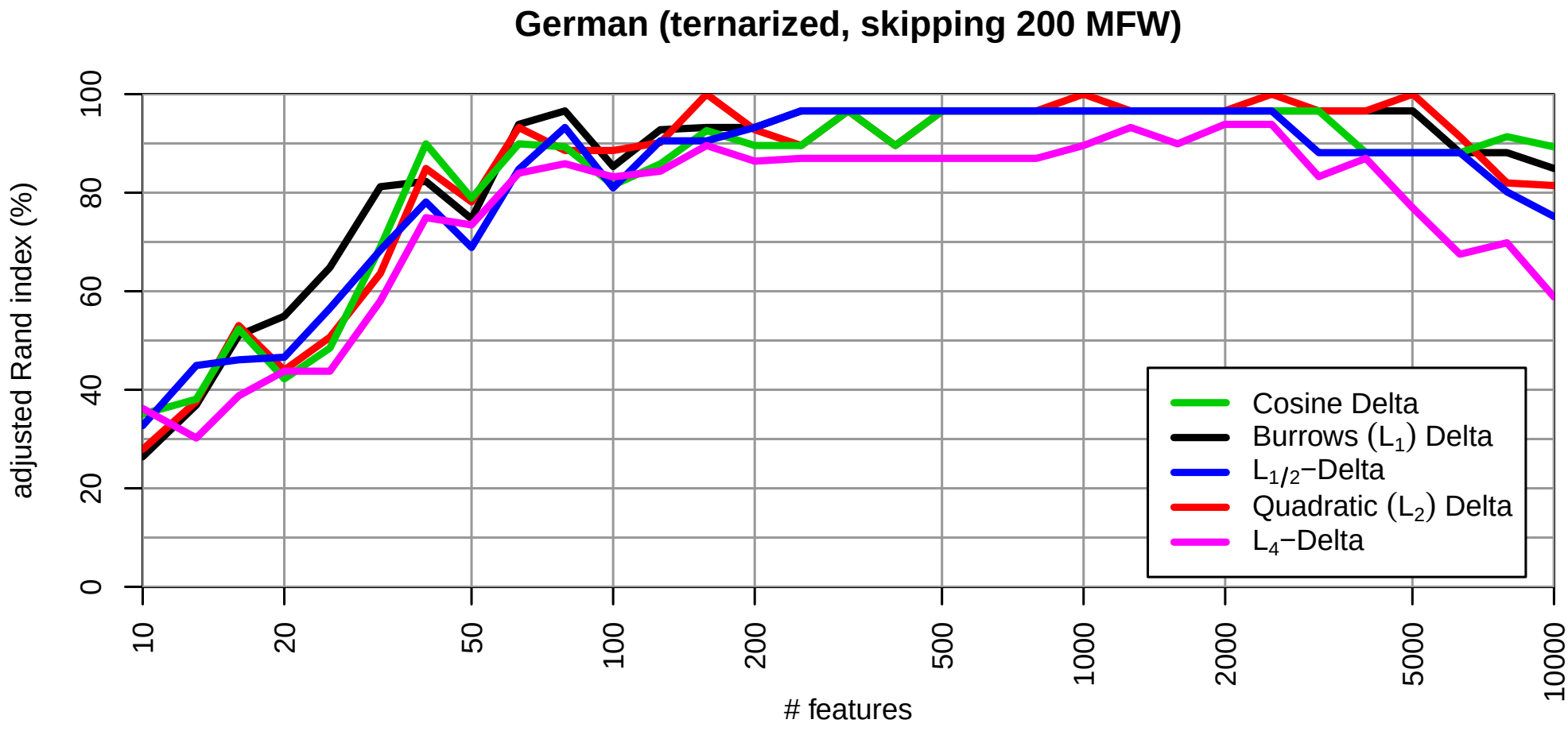
Abschneiden von Ausreißern ($|z| > 2$) H1

Parameter: Transformation der z-Werte



Ternarisierung (je 33%) H2

Parameter: Transformation der z-Werte



Ternarisierung (je 33%, *ohne* 200 MFW)

Fazit

- Empirische Befunde stützen „Schlüsselprofil“-Hypothese 2
- Ternarisierte Vektoren sehr robust und $\Delta_{\angle}$ ebenbürtig
- Abschneiden von Ausreißern ➜ nur leichte Verbesserung
- Simulationsexperiment: Erzeugen von zufälligen Ausreißern als „*noise*“ verschlechtert die Ergebnisse kaum

Nächste Schritte

- Einfluss unterschiedlicher Clusteringverfahren
- Experimente mit Textfragmenten (Textlänge, Konsistenz)
- Wortartenfilter (z.B. Funktionswörter), Lemmatisierung, . . .
- Relevanz einzelner Wörter (Beitrag zu Abstandsmaßen)
- Anwendung auf interpretierbare stilometrische Merkmale?

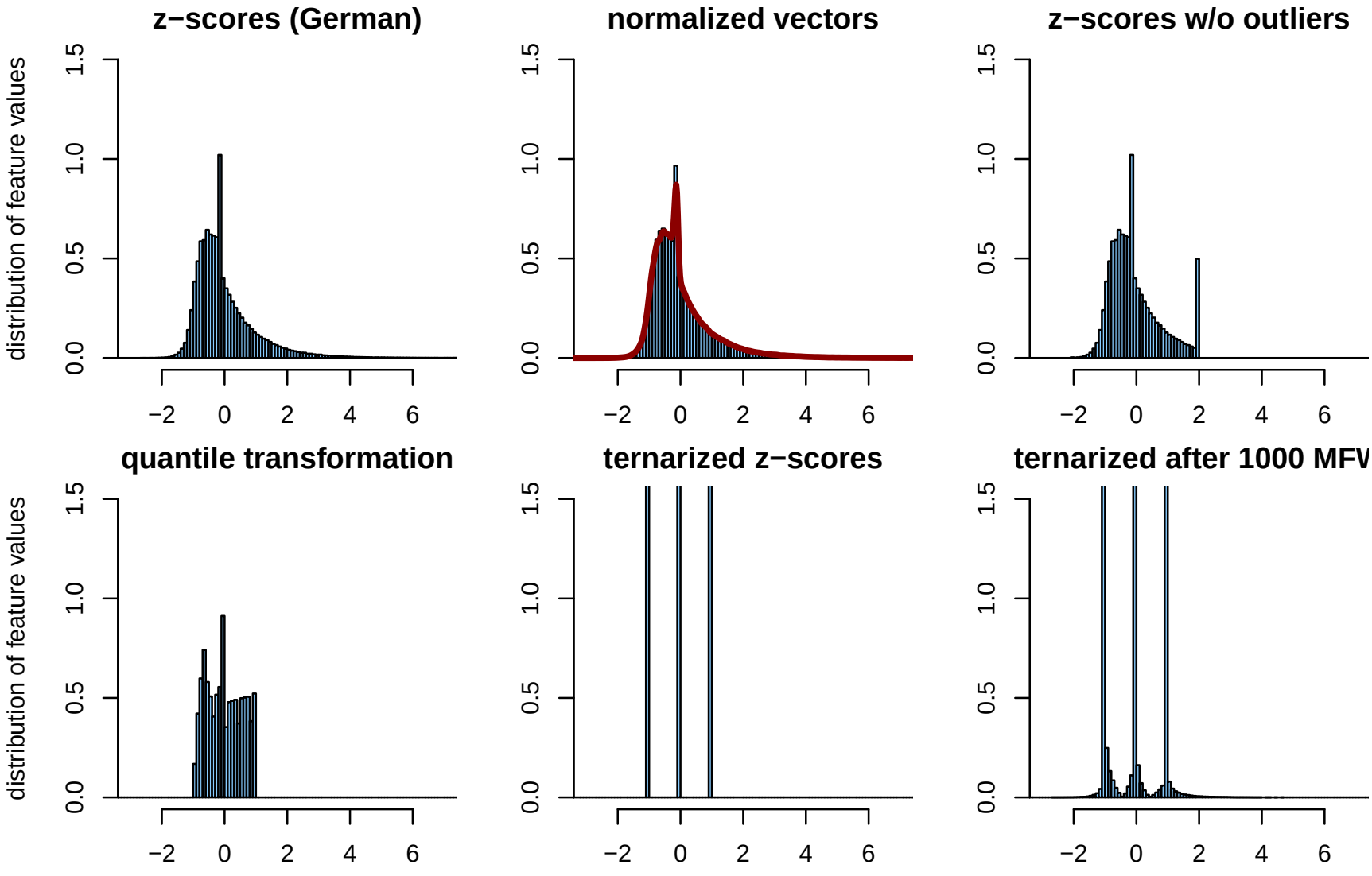
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

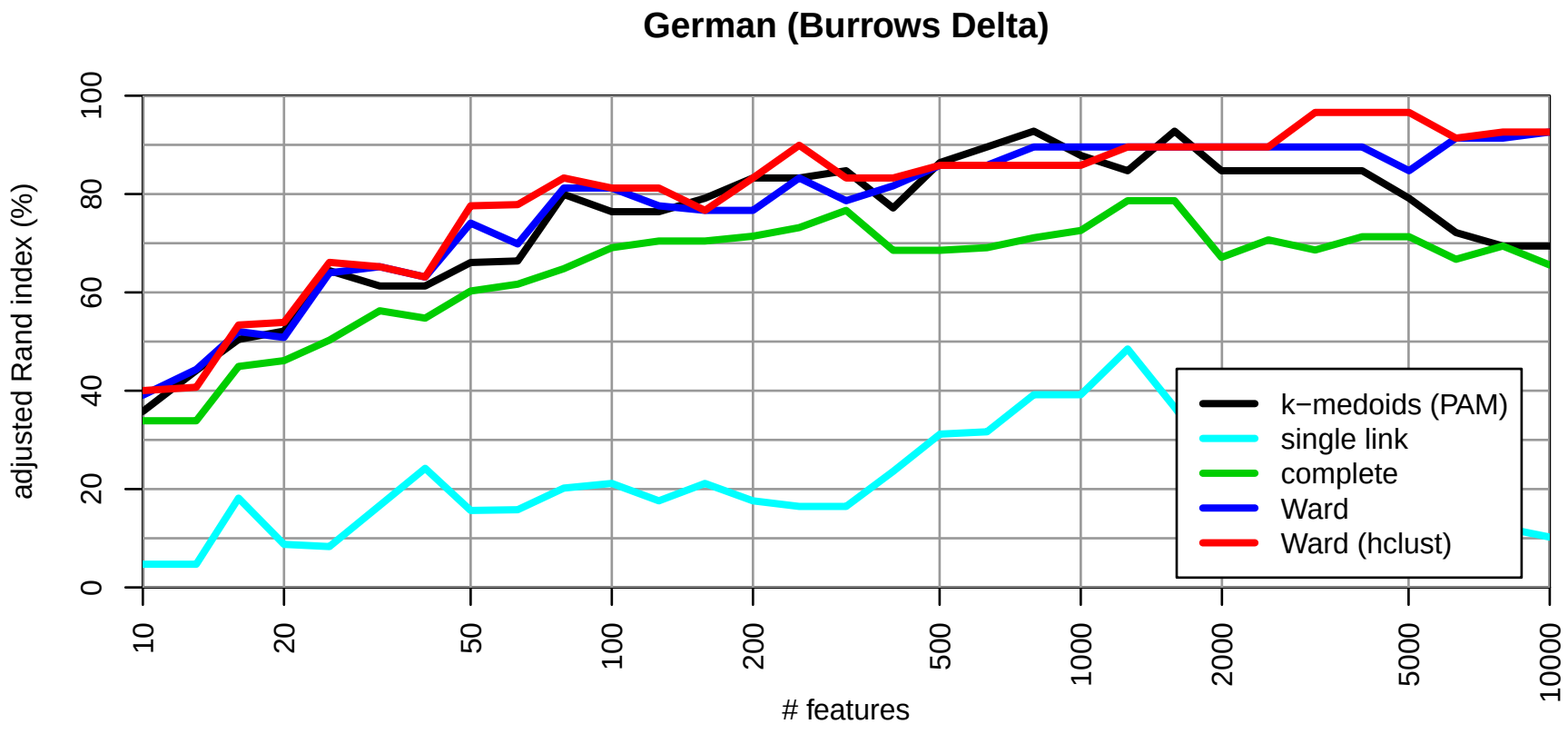
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Dank & Diskussion

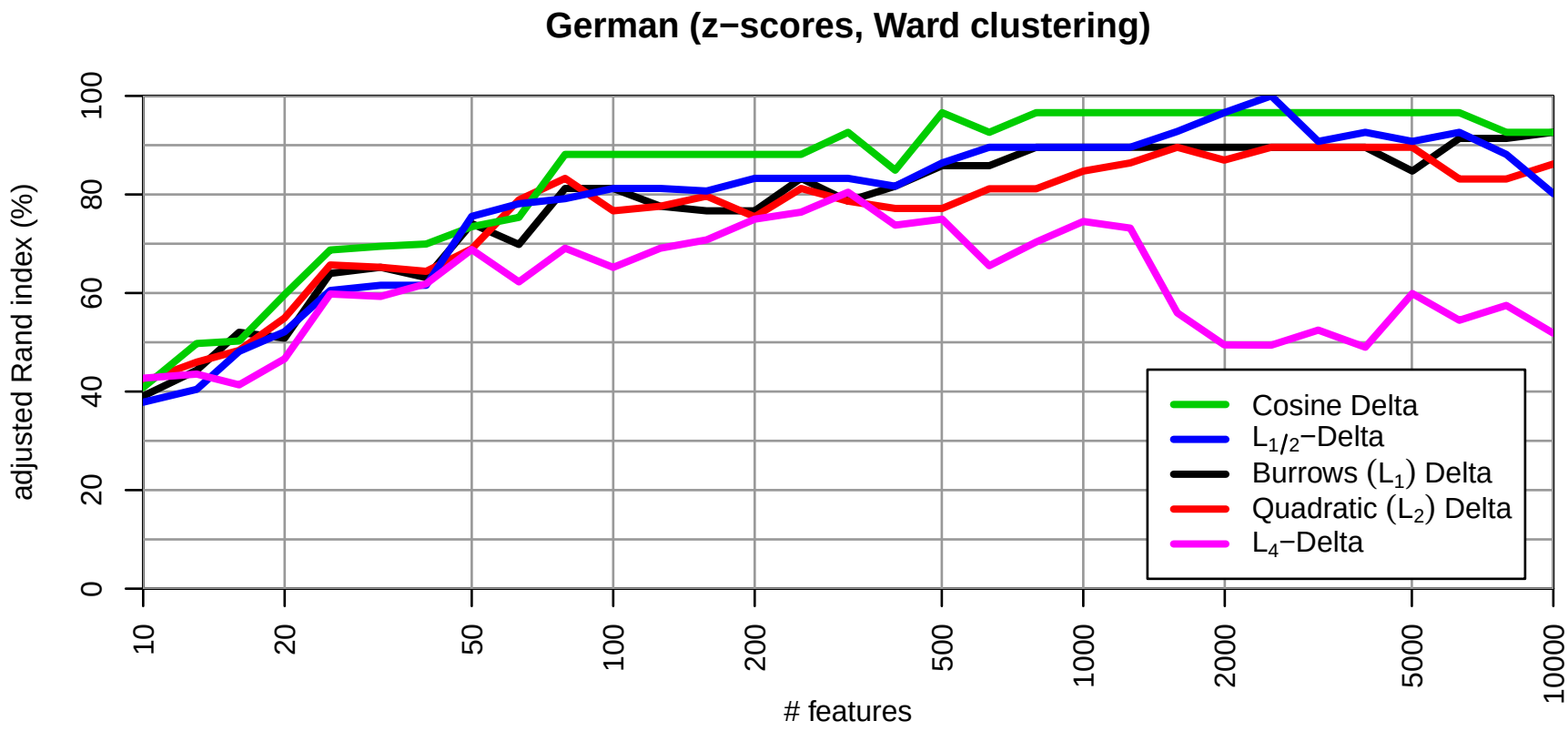
Auswirkung der Transformationen auf Verteilung



Vergleich von Clustering-Verfahren

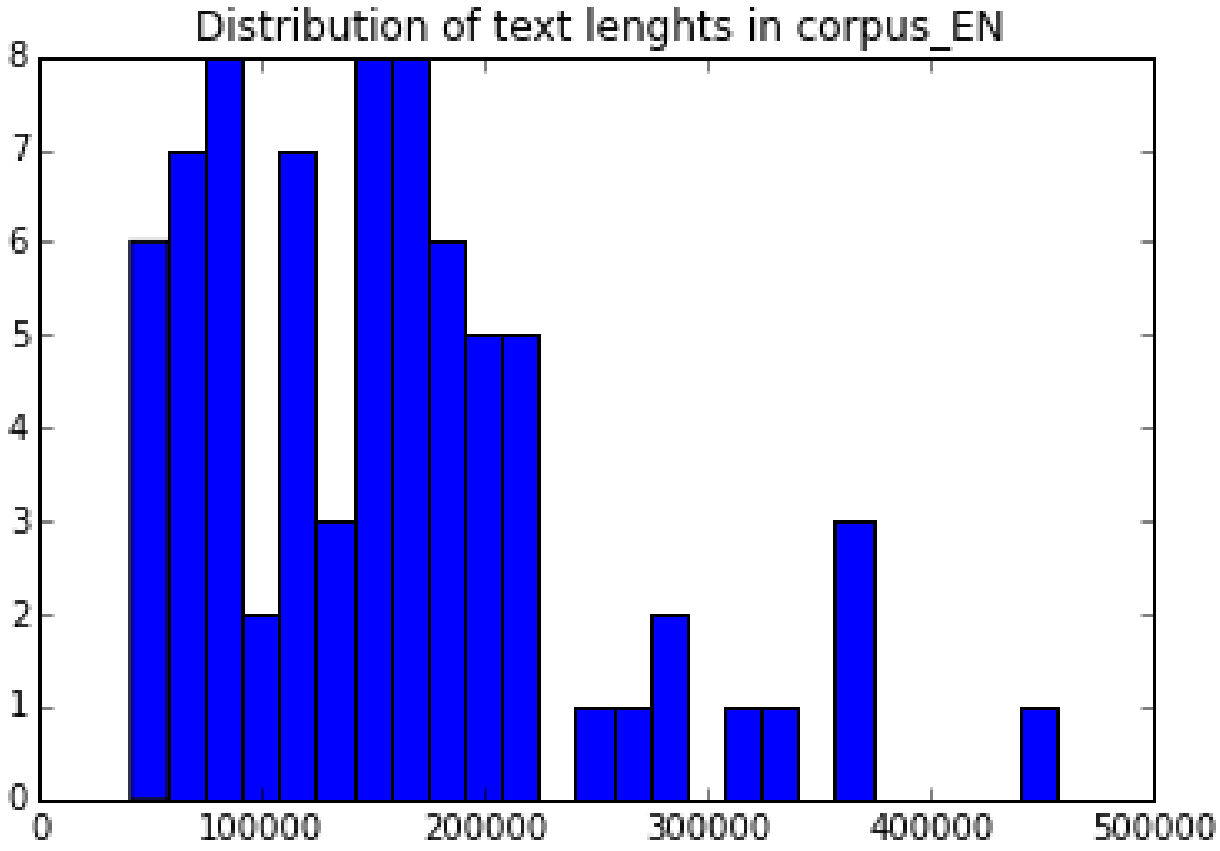


Clustering-Verfahren: PAM vs. Ward

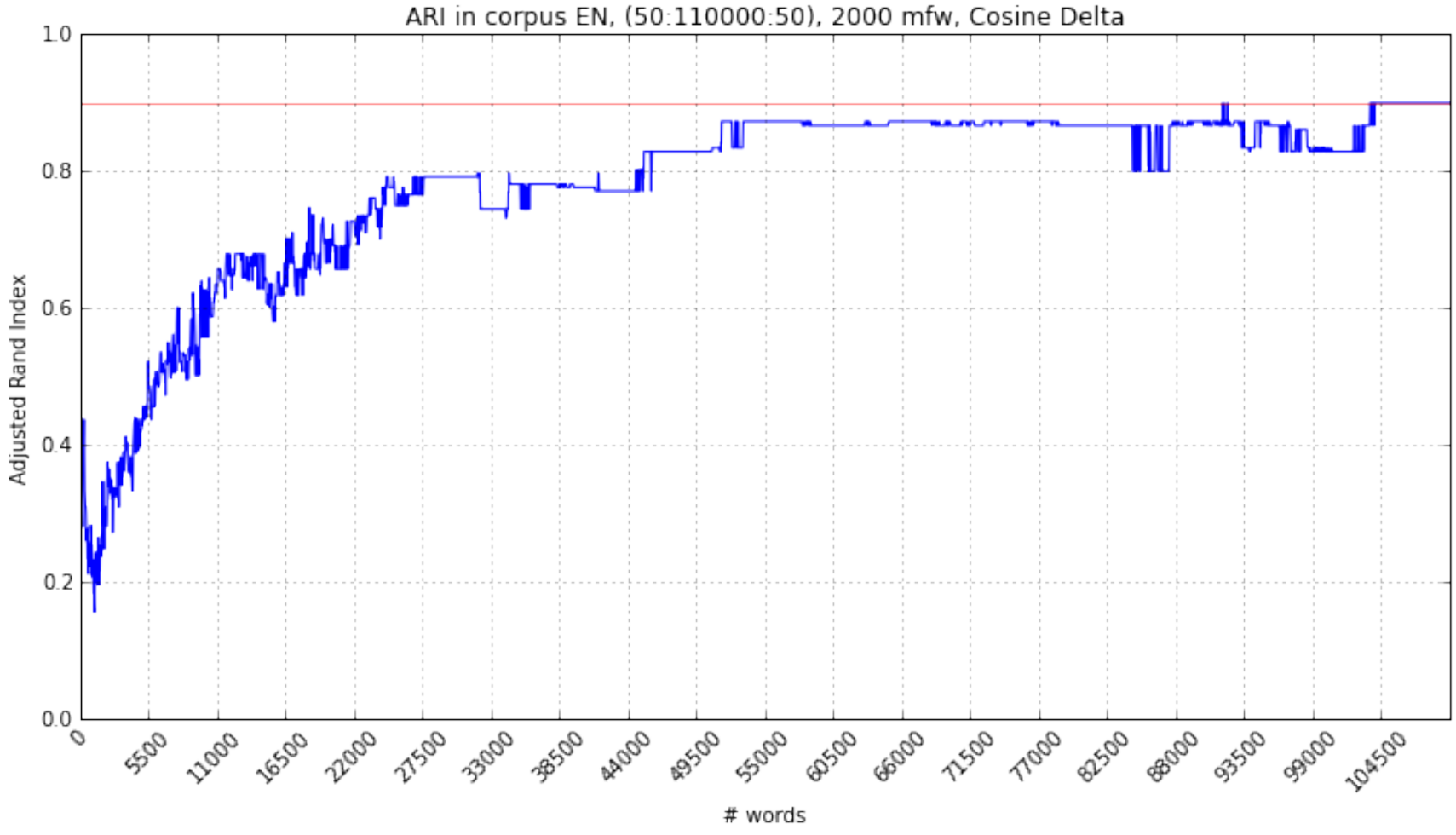


Learning curves

- Experimente wurden mit den ganzen Romanen durchgeführt
- Funktioniert Delta auch bei kürzeren Texten?
  - ☞ zwei Fallstudien mit Cosine Delta $\Delta_{\angle}$ und $n_w = 2000$ MFW

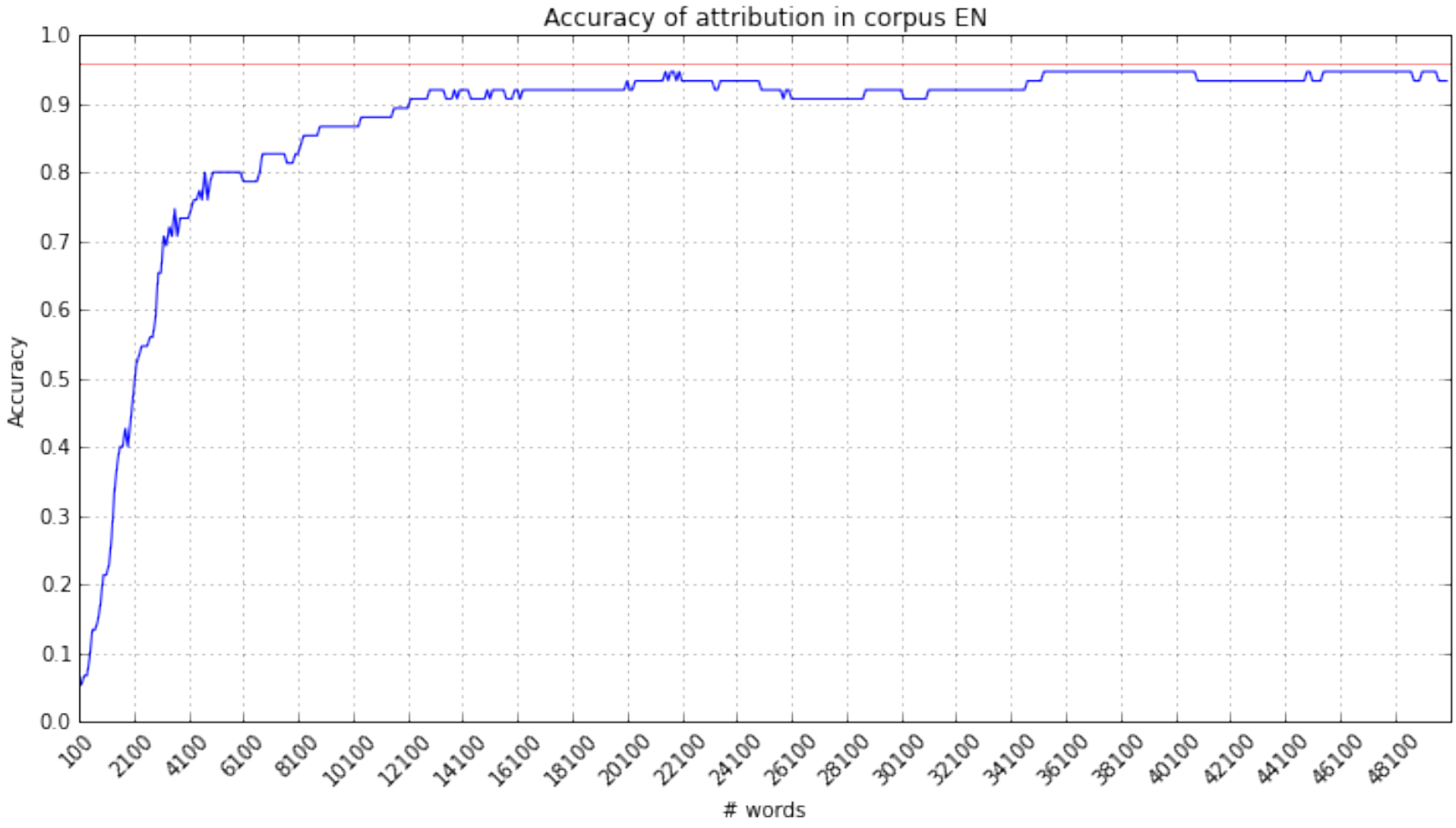


Learning curves: Clustering



alle Texte auf die angegebene Zahl von Token gekürzt

Learning curves: Klassifikation



ein Text gekürzt, andere Romane ungekürzt als Trainingsdaten

References I

Argamon, Shlomo (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, **23**(2), 131 –147.

Burrows, John (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, **17**(3), 267 –287.

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.

Hoover, David L. (2004). Delta Prime? *Literary and Linguistic Computing*, **19**(4), 477 –495.

Hubert, Lawrence and Arabie, Phipps (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

Jannidis, Fotis; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2015). Improving Burrows' Delta - An empirical evaluation of text distance measures. In *Digital Humanities Conference 2015*, Sydney.

Juola, Patrick (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3), 233–334.

Koppel, M.; Schler, J.; Argamon, S. (2008). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, **60**(1), 9–26.

References II

Mosteller, Frederick and Wallace, David L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, **58**(302), 275–309.

Smith, Peter W. H. and Aldridge, W. (2011). Improving authorship attribution: Optimizing Burrows' Delta method. *Journal of Quantitative Linguistics*, **18**(1), 63–88.

Stamatatos, Efstathios (2009). A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, **60**(3), 538–556.