Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Statistical Significance in Literary Authorship Attribution

based on work in the e-humanities research centre Kallimachos

Prof. Dr. Stefan Evert

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg

with

Frimi Dimpel[1], Fotis Jannidis[2], Steffen Pielström[2], Thomas Proisl[1], Isabella Reger[2], Christof Schöch[2], Thorsten Vitt[2]

[2]Julius-Maximilians-Universität Würzburg

Manchester, 15 May 2019

Outline

Literary authorship attribution

Burrows's Delta

Empirical evaluation

Understanding Delta

Statistical significance

Stylometry & authorship attribution

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

- Stylometry aims to characterise the style of a literary author in terms of quantitiative features, e.g.
  - sentence length, word length, frequency classes, ...
  - vocabulary richness (type-token distribution)
  - syntactic complexity
  - frequency of function words, syntactic structures
  - spelling preferences, synonym choice

Stylometry & authorship attribution

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

- Stylometry aims to characterise the style of a literary author in terms of quantitiative features, e.g.
  - sentence length, word length, frequency classes, ...
  - vocabulary richness (type-token distribution)
  - syntactic complexity
  - frequency of function words, syntactic structures
  - spelling preferences, synonym choice

- Important application: literary authorship attribution
  - The Federalist Papers (Mosteller and Wallace 1963)
  - Who wrote the Bixby latter? (Grieve *et al.* 2018)
  - Did Shakespeare really exist? (Thisted and Efron 1987)
  - MHG poetry: *Die Halbe Birne* (Dimpel 2018)
  - 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/

Stylometry & authorship attribution

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

- Stylometry aims to characterise the style of a literary author in terms of quantitiative features, e.g.
  - sentence length, word length, frequency classes, ...
  - vocabulary richness (type-token distribution)
  - syntactic complexity
  - frequency of function words, syntactic structures
  - spelling preferences, synonym choice

- Important application: literary authorship attribution
  - The Federalist Papers (Mosteller and Wallace 1963)
  - Who wrote the Bixby latter? (Grieve *et al.* 2018)
  - Did Shakespeare really exist? (Thisted and Efron 1987)
  - MHG poetry: *Die Halbe Birne* (Dimpel 2018)
  - 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/

☞ Often proxy task for identification of stylometric fingerprint

Stylometry & authorship attribution

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

- Authorship attribution as classification task
  - closed set of candidate authors for unkown text
  - training set of texts with known authorship
  - supervised machine learning algorithm (deep, if you must)
  - evaluation: classification accuracy

Stylometry & authorship attribution

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

- Authorship attribution as classification task
  - closed set of candidate authors for unkown text
  - training set of texts with known authorship
  - supervised machine learning algorithm (deep, if you must)
  - evaluation: classification accuracy

- Authorship attribution as clustering task
  - given set of unknown texts
  - group texts written by same author into cluster
  - measure of text (dis)similarity + clustering algorithm
  - evaluation: adjusted Rand index (ARI)

Stylometry & authorship attribution

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

- Authorship attribution as classification task
  - closed set of candidate authors for unkown text
  - training set of texts with known authorship
  - supervised machine learning algorithm (deep, if you must)
  - evaluation: classification accuracy

- Authorship attribution as clustering task
  - given set of unknown texts
  - group texts written by same author into cluster
  - measure of text (dis)similarity + clustering algorithm
  - evaluation: adjusted Rand index (ARI)

- ☞ Clustering is the more general approach
  - successful clustering ➜ classification by majority vote
  - k-NN classifier based on text similarity measure

A (very) brief history of authorshop attribution

A (very) brief history of authorshop attribution

- Comparison of word frequencies (Mosteller and Wallace 1963)
  - based on sophisticated Bayesian statistical model

A (very) brief history of authorshop attribution

- Comparison of word frequencies (Mosteller and Wallace 1963)
  - based on sophisticated Bayesian statistical model

- Machine learning from wide range of stylometric features (Juola 2006; Stamatatos 2009)
  - large number of lexical, syntactic, semantic, lexico-statistical and character-level features
  - ML algorithm selects the most informative features

A (very) brief history of authorshop attribution

- Comparison of word frequencies (Mosteller and Wallace 1963)
  - based on sophisticated Bayesian statistical model

- Machine learning from wide range of stylometric features (Juola 2006; Stamatatos 2009)
  - large number of lexical, syntactic, semantic, lexico-statistical and character-level features
  - ML algorithm selects the most informative features

- The Delta method (Burrows 2002)
  - simple frequency comparison of a few hundred common words
  - essentially Mosteller and Wallace (1963) without the math
  - but also without need for closed training corpus

Outline

Literary authorship attribution

Burrows's Delta

Empirical evaluation

Understanding Delta

Statistical significance

Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW) form a "fingerprint" of an author's style

relative frequency: $f_i(D)$

	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*
f(Madding Crowd) =	(.051,	.029,	.026,	.027,	.027,	.016,	.016,	.014,	.011,	.008,	.010, ...)
f(Tess of the d'U.) =	(.053,	.027,	.027,	.028,	.020,	.013,	.015,	.014,	.012,	.009,	.018, ...)
f(Oliver Twist) =	(.055,	.032,	.024,	.023,	.022,	.012,	.014,	.011,	.011,	.013,	.005, ...)

Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW) form a "fingerprint" of an author's style

relative frequency: $f_i(D)$

	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*
f(Madding Crowd) =	.051,	.029,	.026,	.027,	.027,	.016,	.016,	.014,	.011,	.008,	.010, ...)
f(Tess of the d'U.) =	.053,	.027,	.027,	.028,	.020,	.013,	.015,	.014,	.012,	.009,	.018, ...)
f(Oliver Twist) =	.055,	.032,	.024,	.023,	.022,	.012,	.014,	.011,	.011,	.013,	.005, ...)

‘‘It shows you didn’t care a bit about me, and were ready to desert
can not recollect - I only saw her for a day or two.’’ ‘‘It was lig
e the girl had been in her grave for a few days, when the interposi
d to him, spoke to him for a few minutes, and then all three parted
with stakes of all sizes. For a few seconds the wayfarer stood with
eral good character. On Sundays he was a man of misty views, rather
way!’’ ‘‘I wish you’d show yourself a man of spirit, and not sit wh
ves which is so conspicuous in a woman’s gardening, and which flowe

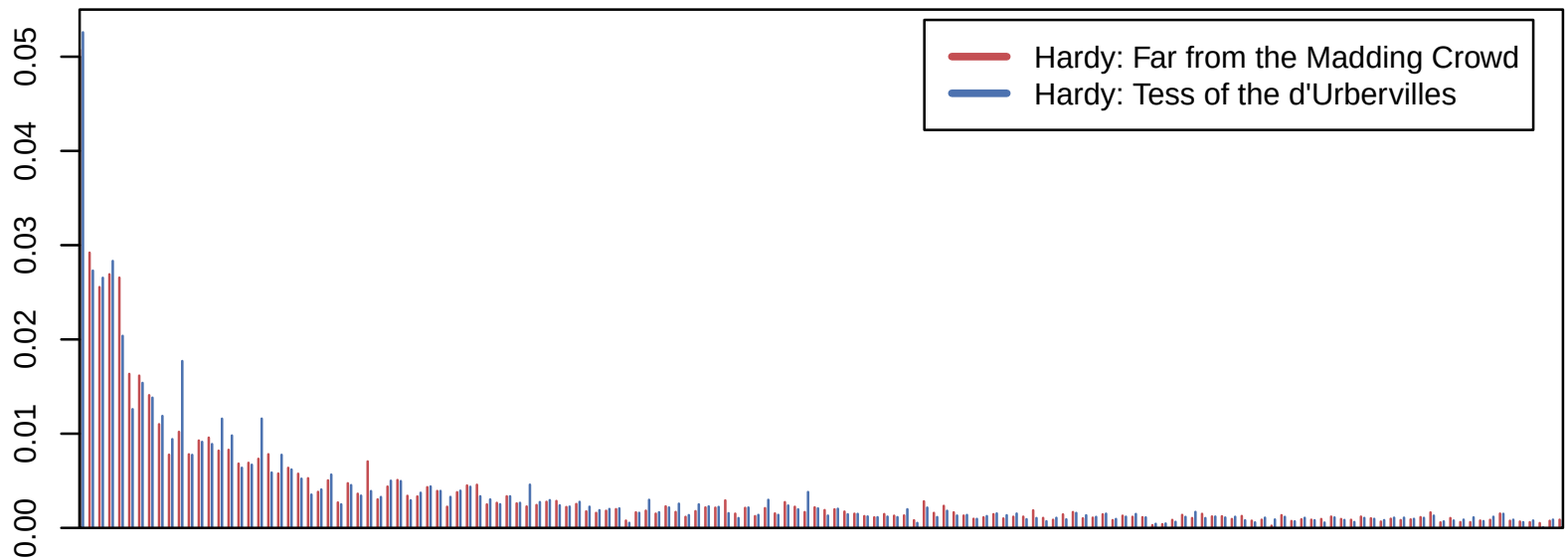
Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW) form a "fingerprint" of an author's style

relative frequency: $f_i(D)$

	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*
f(Madding Crowd) =	(.051,	.029,	.026,	.027,	.027,	.016,	.016,	.014,	.011,	.008,	.010, ...)
f(Tess of the d'U.) =	(.053,	.027,	.027,	.028,	.020,	.013,	.015,	.014,	.012,	.009,	.018, ...)
f(Oliver Twist) =	(.055,	.032,	.024,	.023,	.022,	.012,	.014,	.011,	.011,	.013,	.005, ...)



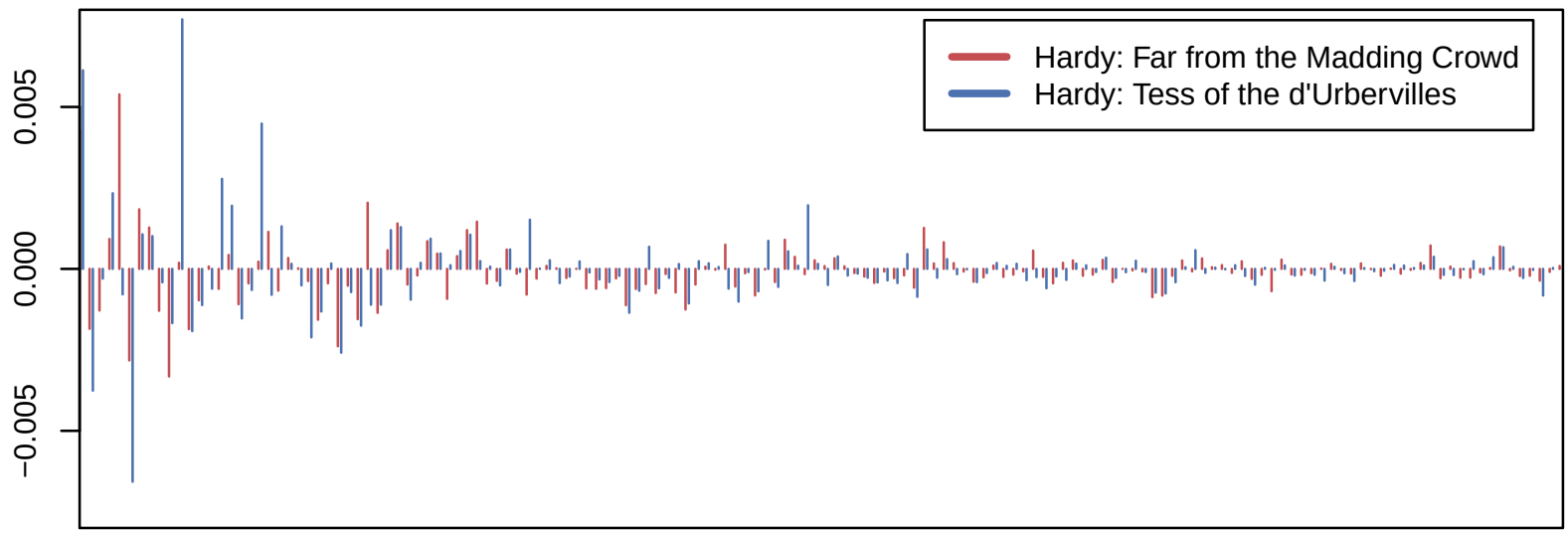
Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW) form a "fingerprint" of an author's style

deviation from mean: $\bar{f}_i(D) = f_i(D) - \mu_i$

	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*
$\bar{\mathbf{f}}$(Madding Crowd) =	.004,	-.002,	-.001,	.001,	.005,	-.003,	.002,	.001,	-.001,	-.003,	.000,
$\bar{\mathbf{f}}$(Tess of the d'U.) =	.006,	-.004,	.000,	.002,	-.001,	-.007,	.001,	.001,	.000,	-.002,	.008,
$\bar{\mathbf{f}}$(Oliver Twist) =	.009,	.001,	-.003,	-.003,	.001,	-.008,	.000,	-.002,	-.001,	.002,	-.005,



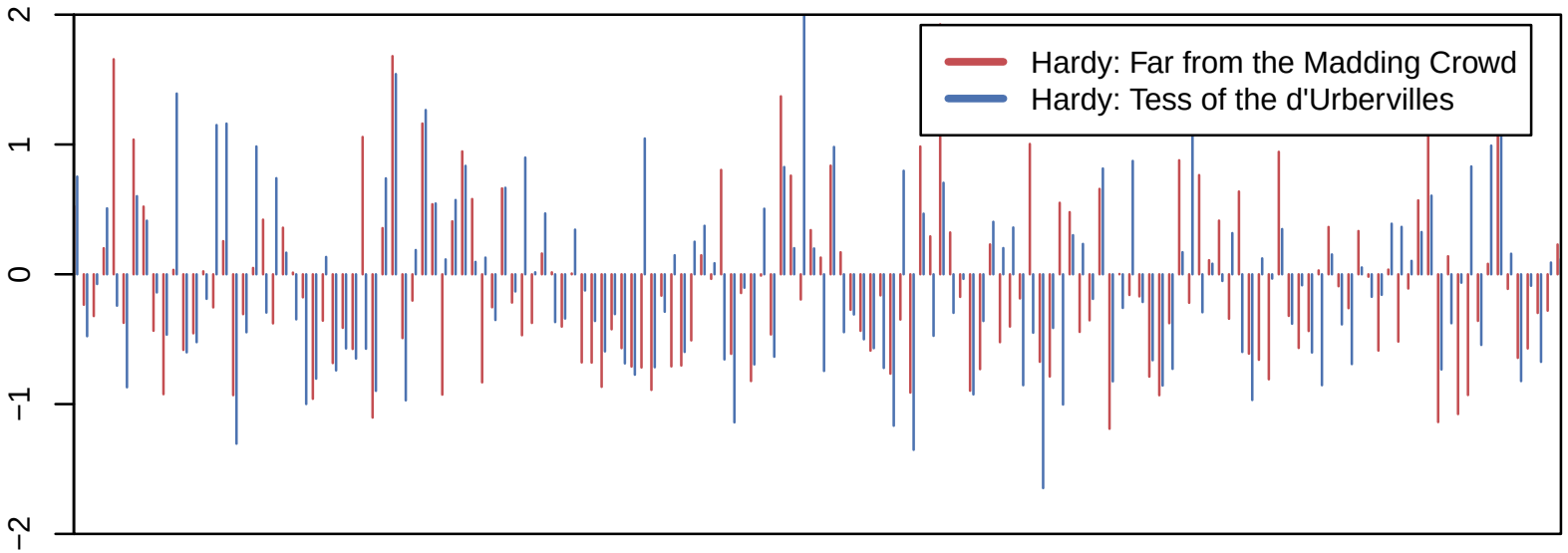
Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW) form a "fingerprint" of an author's style

$$\text{standardised z-scores:} \quad z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i}$$

	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,	...)



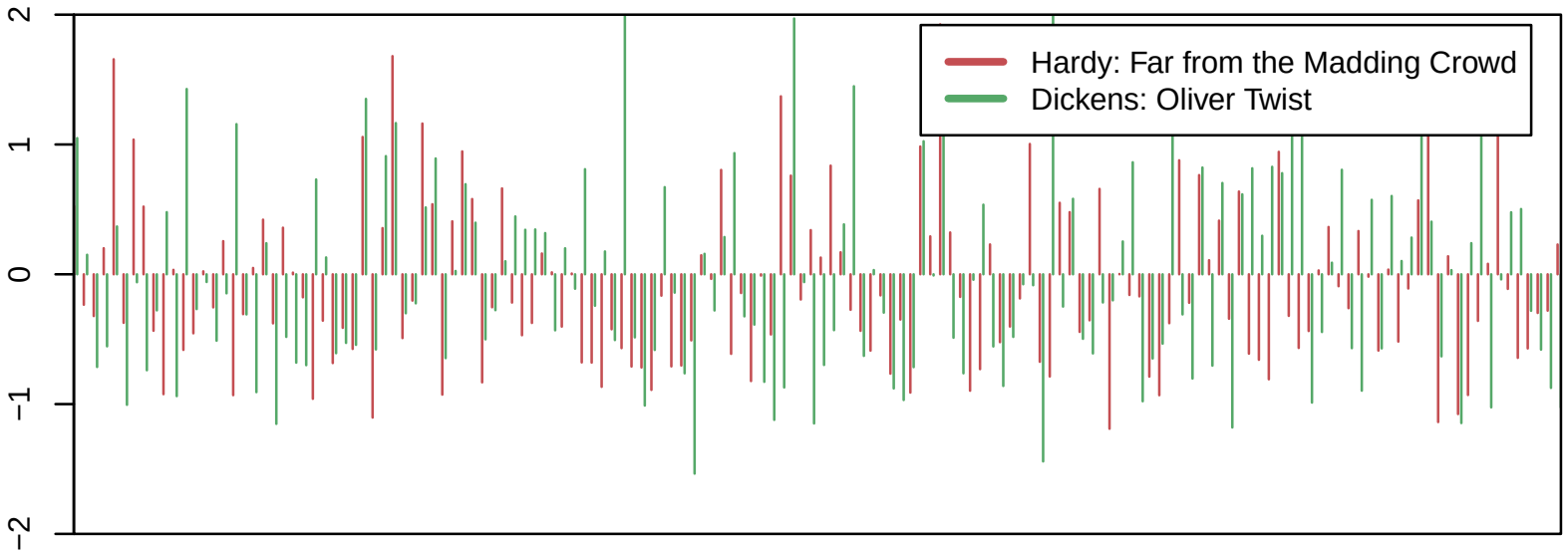
Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW) form a "fingerprint" of an author's style

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

	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,	...)



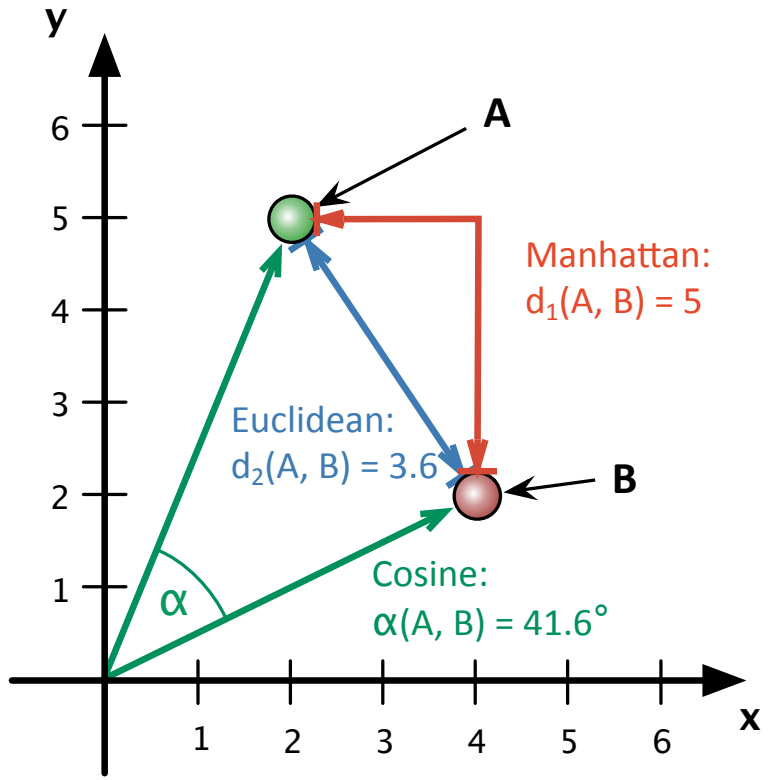
Distance metrics for texts

- Need to quantify similarity of "fingerprints" of two texts, i.e.
- Vectors of z-scores $\mathbf{z}(D_1), \mathbf{z}(D_2) \in \mathbb{R}^{n_w}$
  - n_w = number of MFW = dimensionality of vector

Distance metrics for texts

- Need to quantify similarity of "fingerprints" of two texts, i.e.
- Vectors of z-scores $\mathbf{z}(D_1), \mathbf{z}(D_2) \in \mathbb{R}^{n_w}$
  - n_w = number of MFW = dimensionality of vector

metric = measure of geometric distance



The family of Delta measures

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

- Burrows's Delta = Manhattan distance (Burrows 2002)

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

The family of Delta measures

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

- Burrows's Delta = Manhattan distance (Burrows 2002)

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

- Quadratic Delta = Euclidean distance (Argamon 2008)

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

The family of Delta measures

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

- Burrows's Delta = Manhattan distance (Burrows 2002)

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

- Quadratic Delta = Euclidean distance (Argamon 2008)

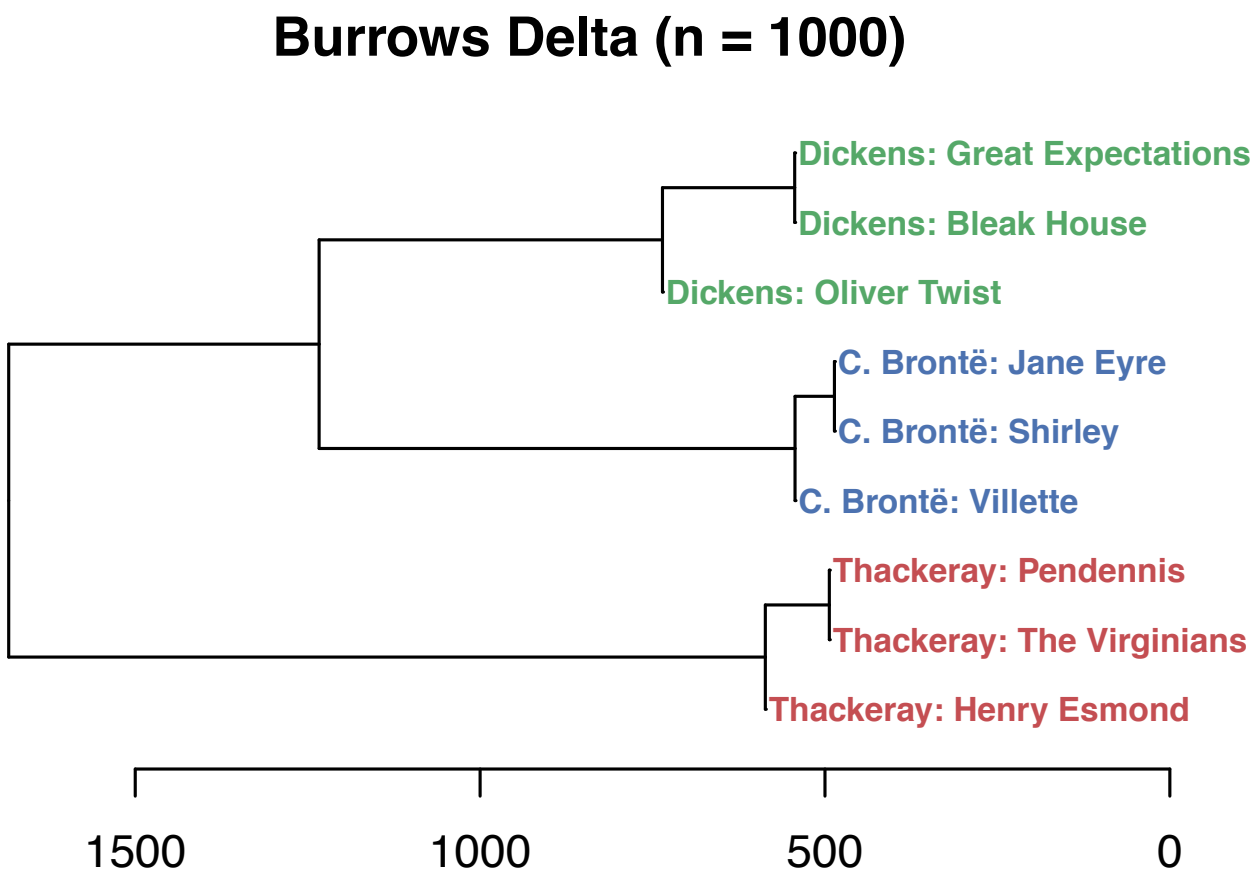
$$\Delta_Q(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_2^2 = \sum_{i=1}^{n_w} (z_i(D) - z_i(D'))^2$$

- Cosine Delta = angular distance (Smith and Aldridge 2011)

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

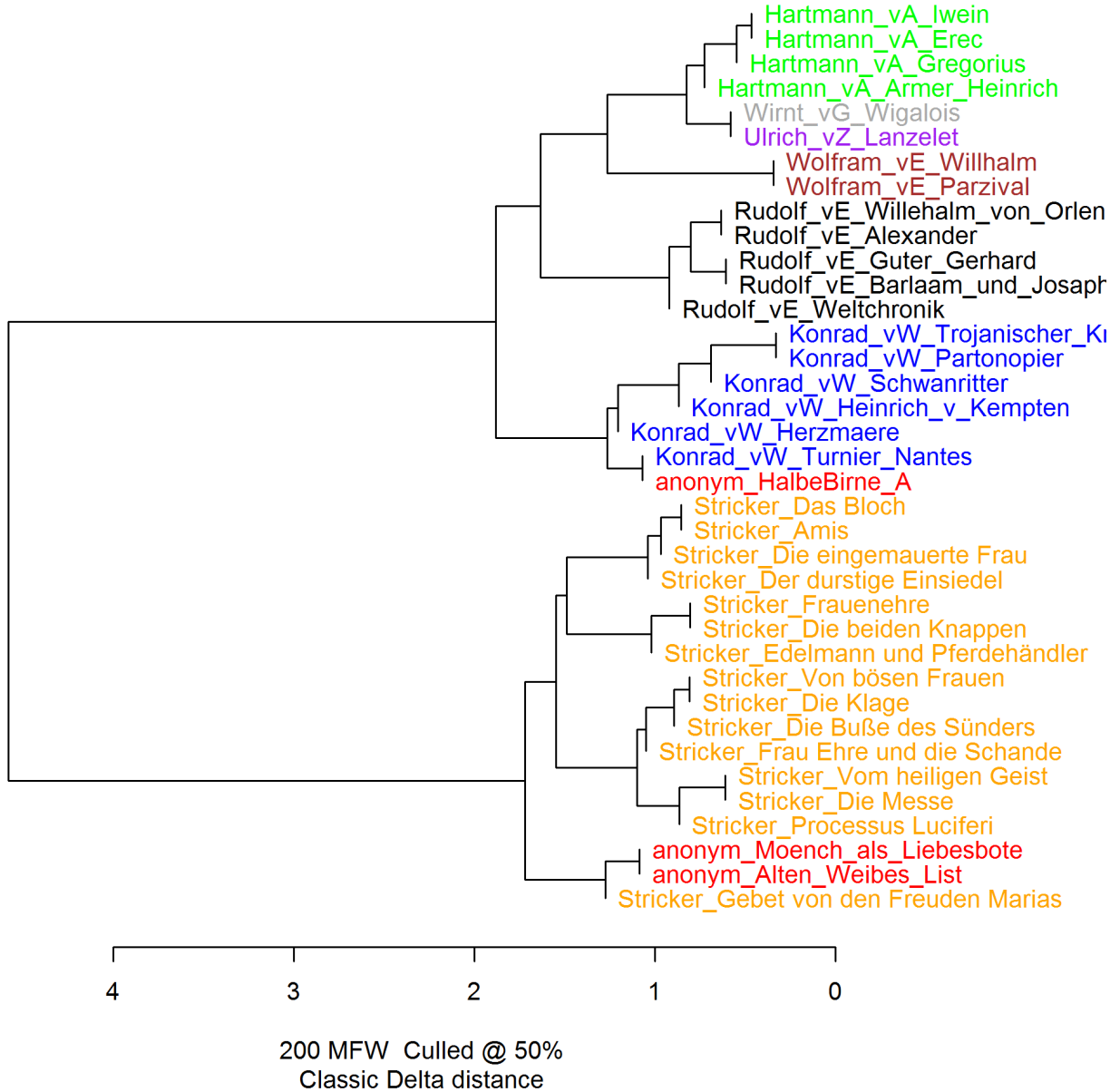
Hierarchical clustering based on Δ

(Evert *et al.* 2017, Fig. 3)



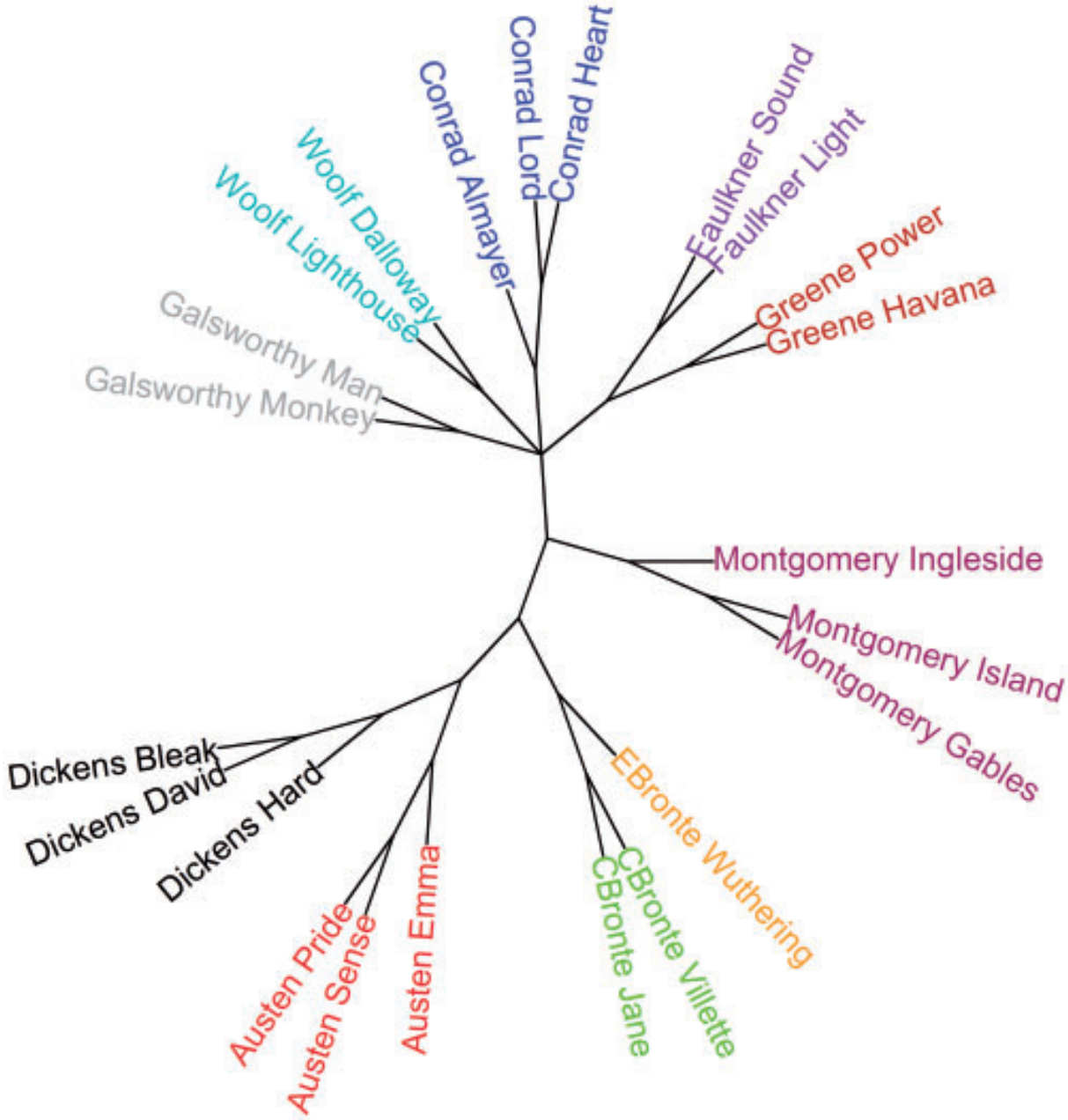
Hierarchical clustering based on Δ

(Büttner *et al.* 2017, Fig. 10)



Hierarchical clustering based on Δ

(Rybicki and Heydel 2013, Fig. 1)



Research questions

1. Can Δ accurately identify the author of a text?
2. How many MFW should we use?
3. What is the effect of other parameters (e.g. Δ_B vs. Δ_Q)?
4. Which words form the characteristic fingerprint of an author?
5. Is this fingerprint mixed with other "signals" (e.g. genre)?
6. Can we apply Δ to short texts?
7. How robust is Δ against "noise" in the data (e.g. OCR errors)?
8. Why does Δ work so well?

Research questions

1. Can Δ accurately identify the author of a text?
2. How many MFW should we use?
3. What is the effect of other parameters (e.g. Δ_B vs. Δ_Q)?
4. Which words form the characteristic fingerprint of an author?
5. Is this fingerprint mixed with other "signals" (e.g. genre)?
6. Can we apply Δ to short texts?
7. How robust is Δ against "noise" in the data (e.g. OCR errors)?
8. Why does Δ work so well?

9. Far too rarely: Are the results statistically significant?

Outline

Literary authorship attribution

Burrows's Delta

Empirical evaluation

Understanding Delta

Statistical significance

Empirical evaluation

"In theory, theory and practice are the same. In practice, they are not."

Data set: 19th-century novels (Jannidis *et al.* 2015)

- three literary corpora (German, English, French)
- each corpus: 25 authors $\times$ 3 novels $=$ 75 novels
- from early 19th century to mid 20th century
- ca. 10M tokens per language
- individual texts: 25k – 900k tokens

Empirical evaluation

"In theory, theory and practice are the same. In practice, they are not."

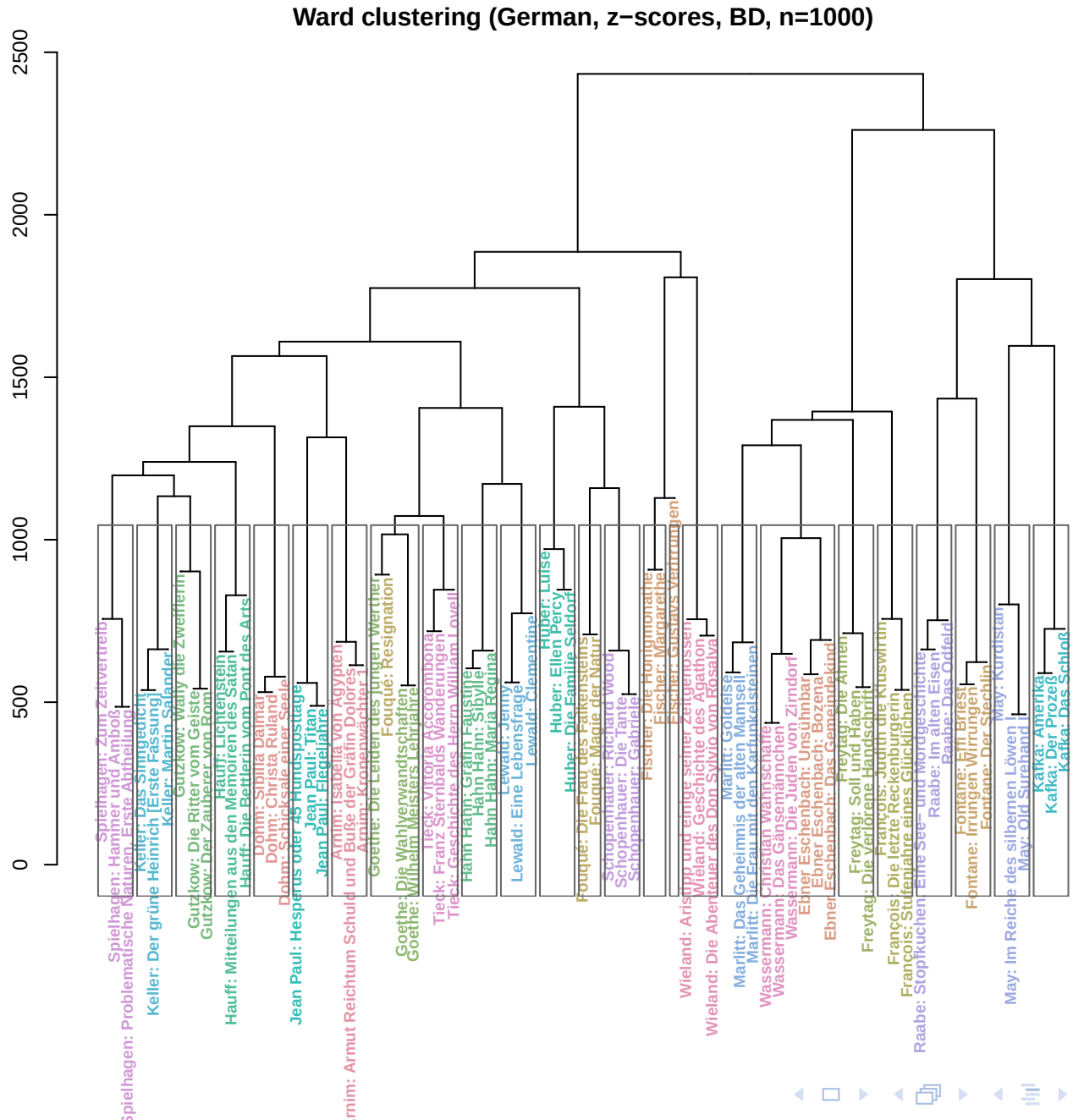
Data set: 19th-century novels (Jannidis *et al.* 2015)

- three literary corpora (German, English, French)
- each corpus: 25 authors $\times$ 3 novels $=$ 75 novels
- from early 19th century to mid 20th century
- ca. 10M tokens per language
- individual texts: 25k – 900k tokens

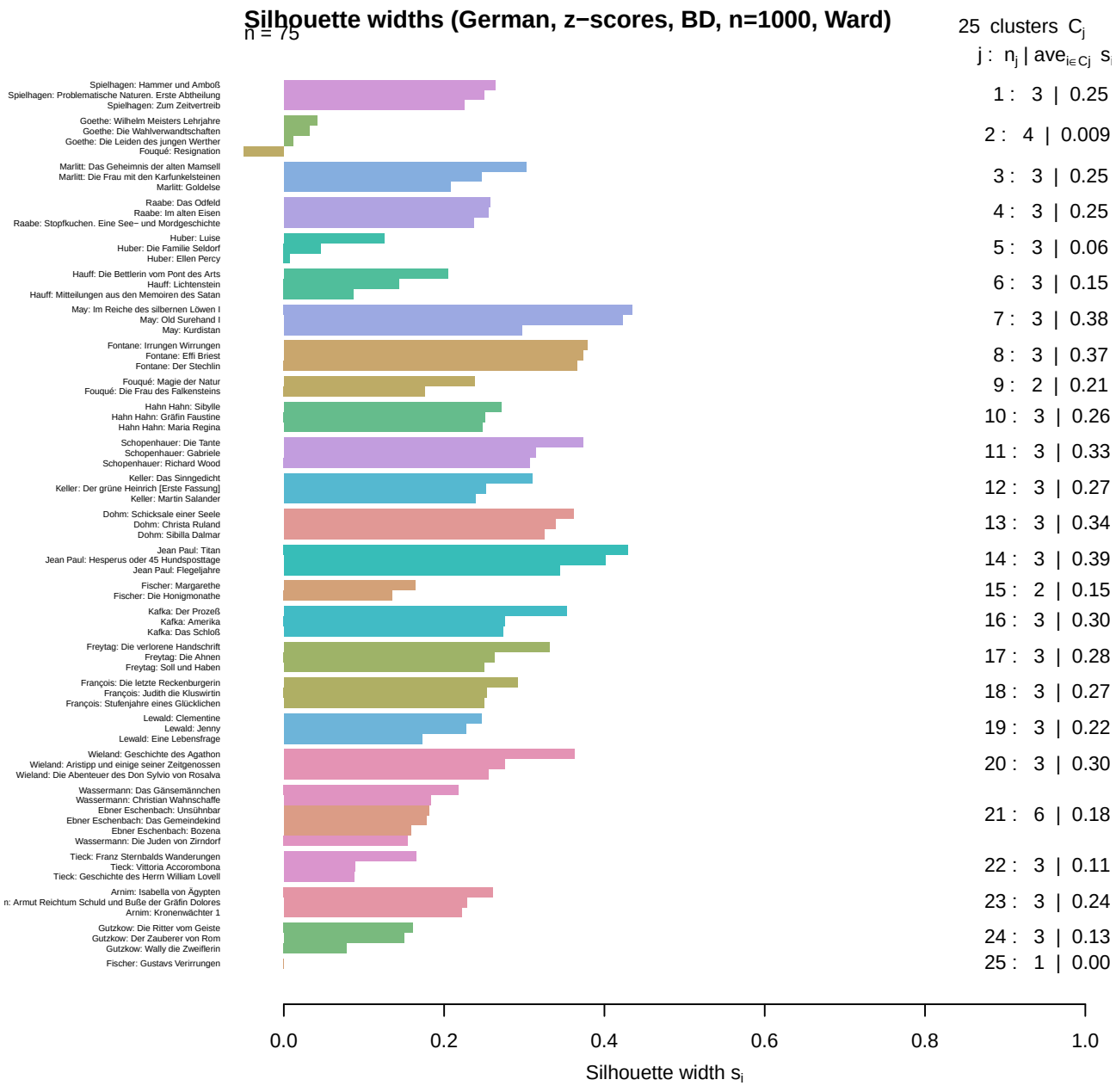
Evaluation studies

- Jannidis *et al.* (2015) – all known variants of Δ, $\approx$ purity
- Evert *et al.* (2017) – focus on n_w and feature scaling, ARI
- Büttner *et al.* (2017) – combination with further data sets

Clustering example (ARI = 90%)

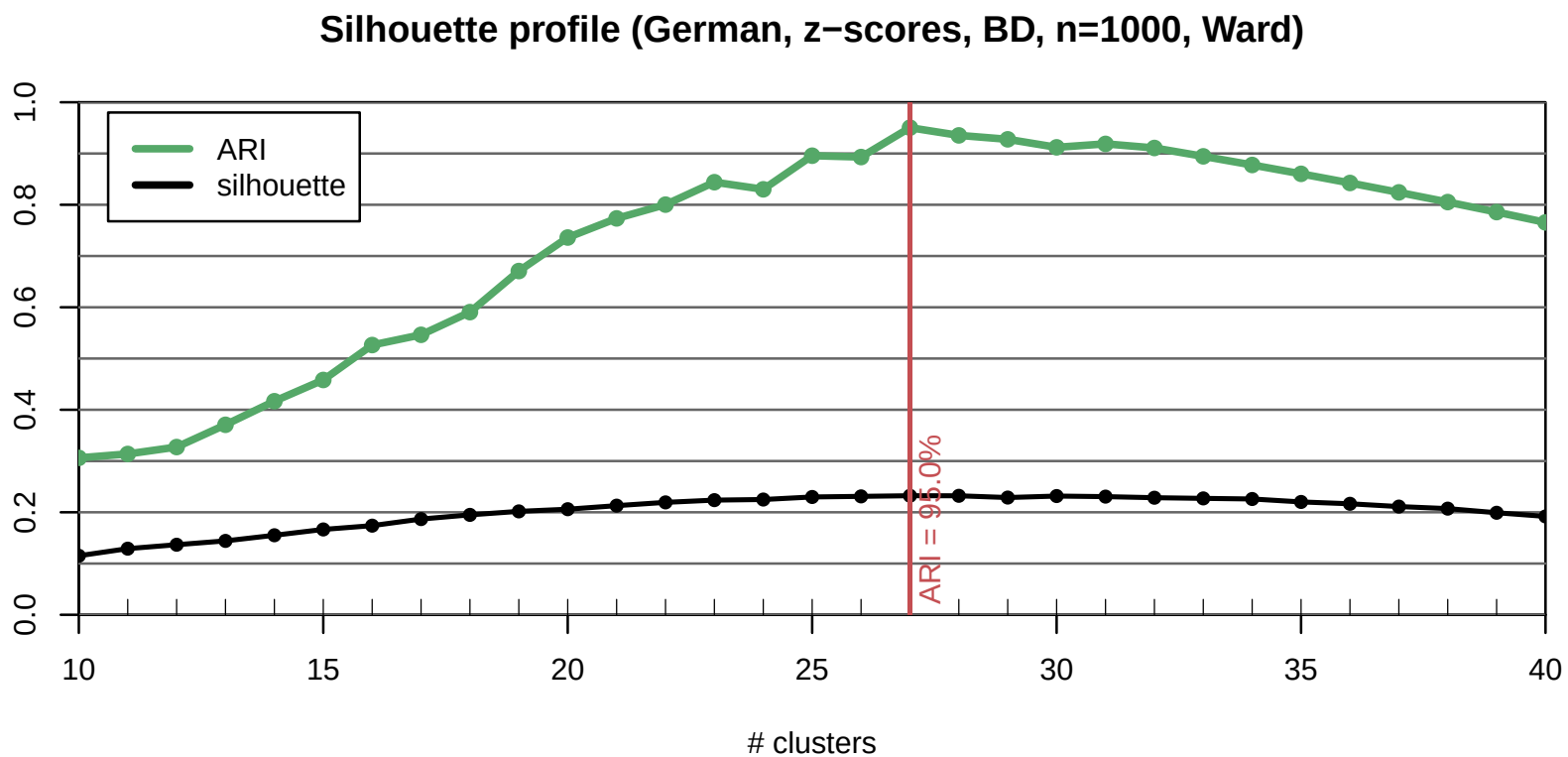


Can we determine the number of clusters automatically?

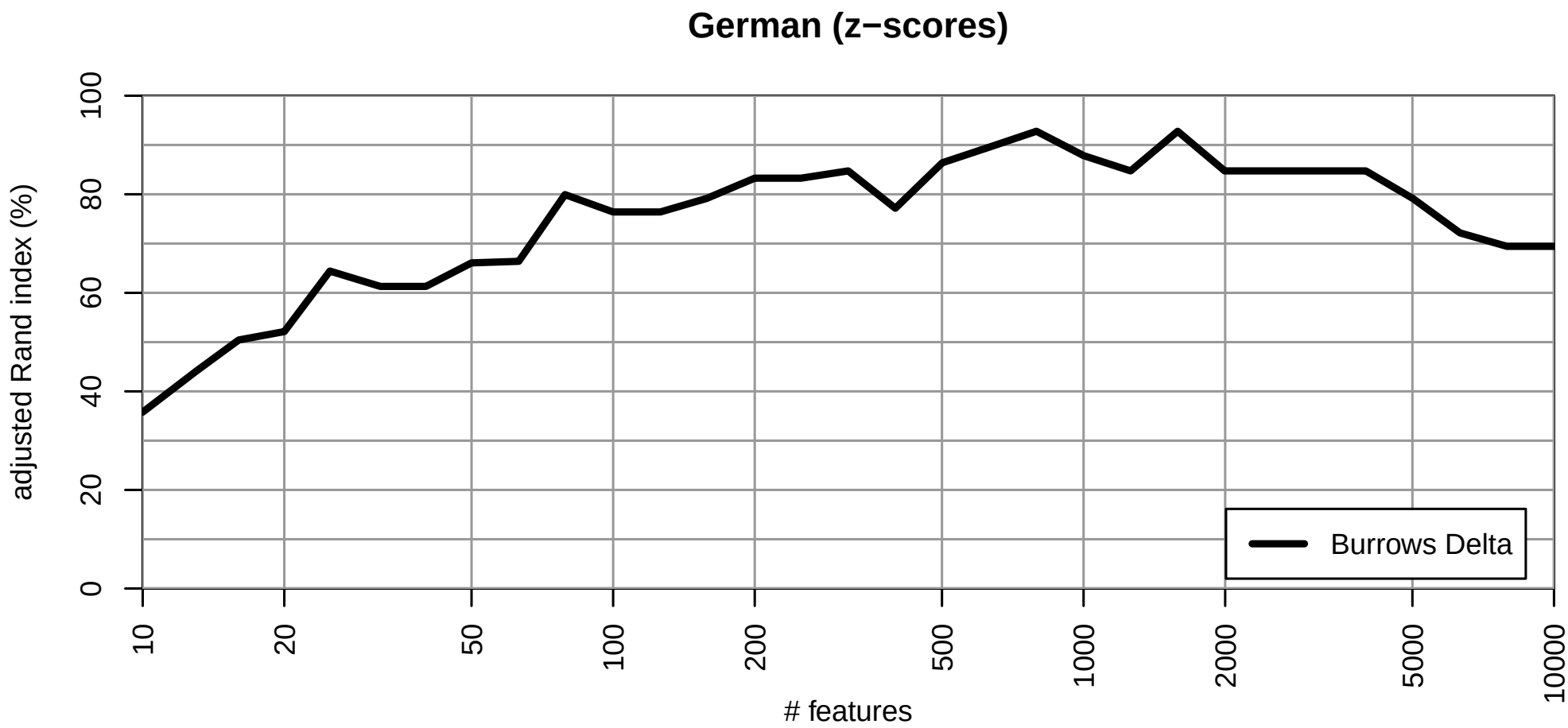


Can we determine the number of clusters automatically?

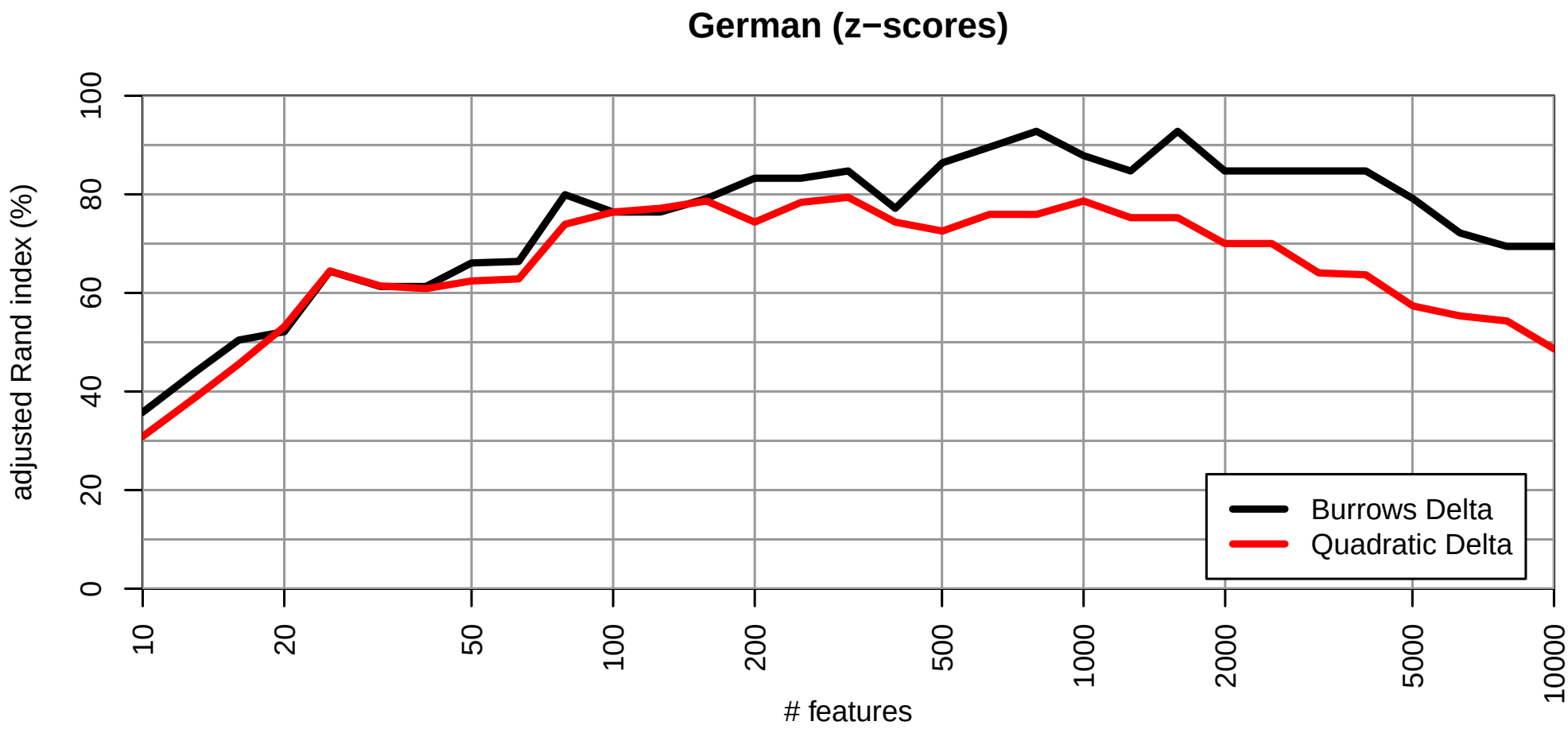
- Average silhouette width is standard criterion for unsupervised model selection in clustering
- Works well on our data:
  usually $k > 25$ clusters, but better ARI than for $k = 25$
- Justifies further experiments with pre-determined $k = 25$



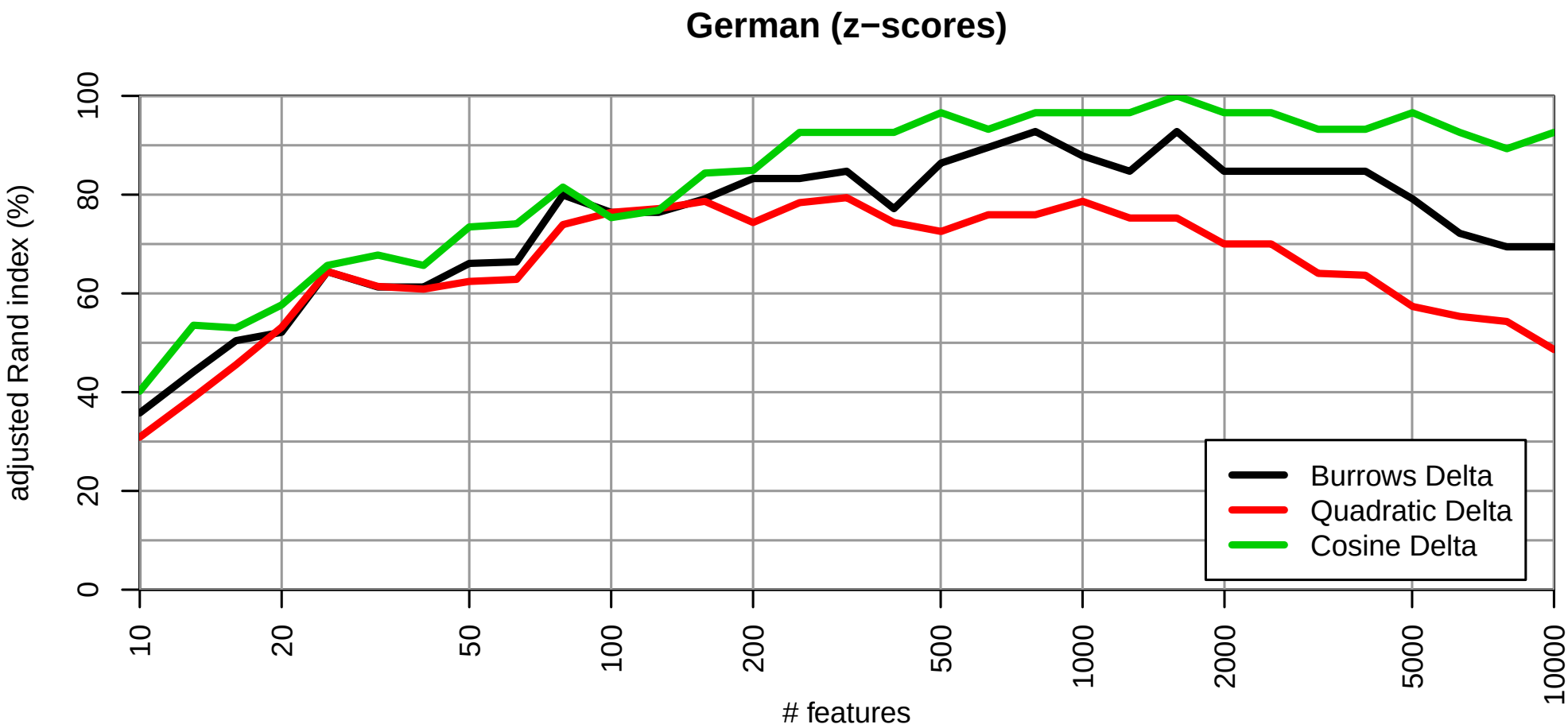
Parameter: Number n_w of MFW



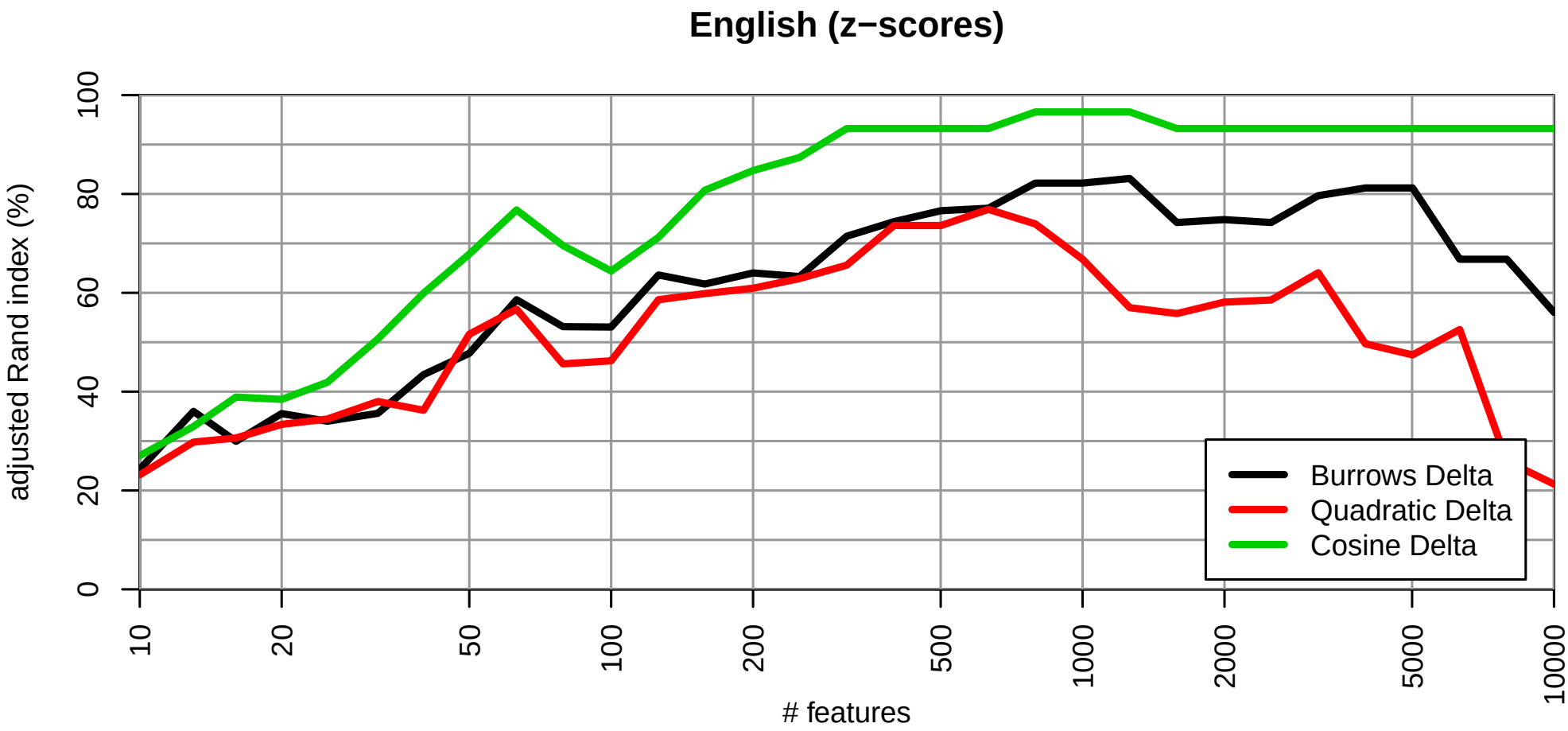
Parameter: Number n_w of MFW



Parameter: Number n_w of MFW

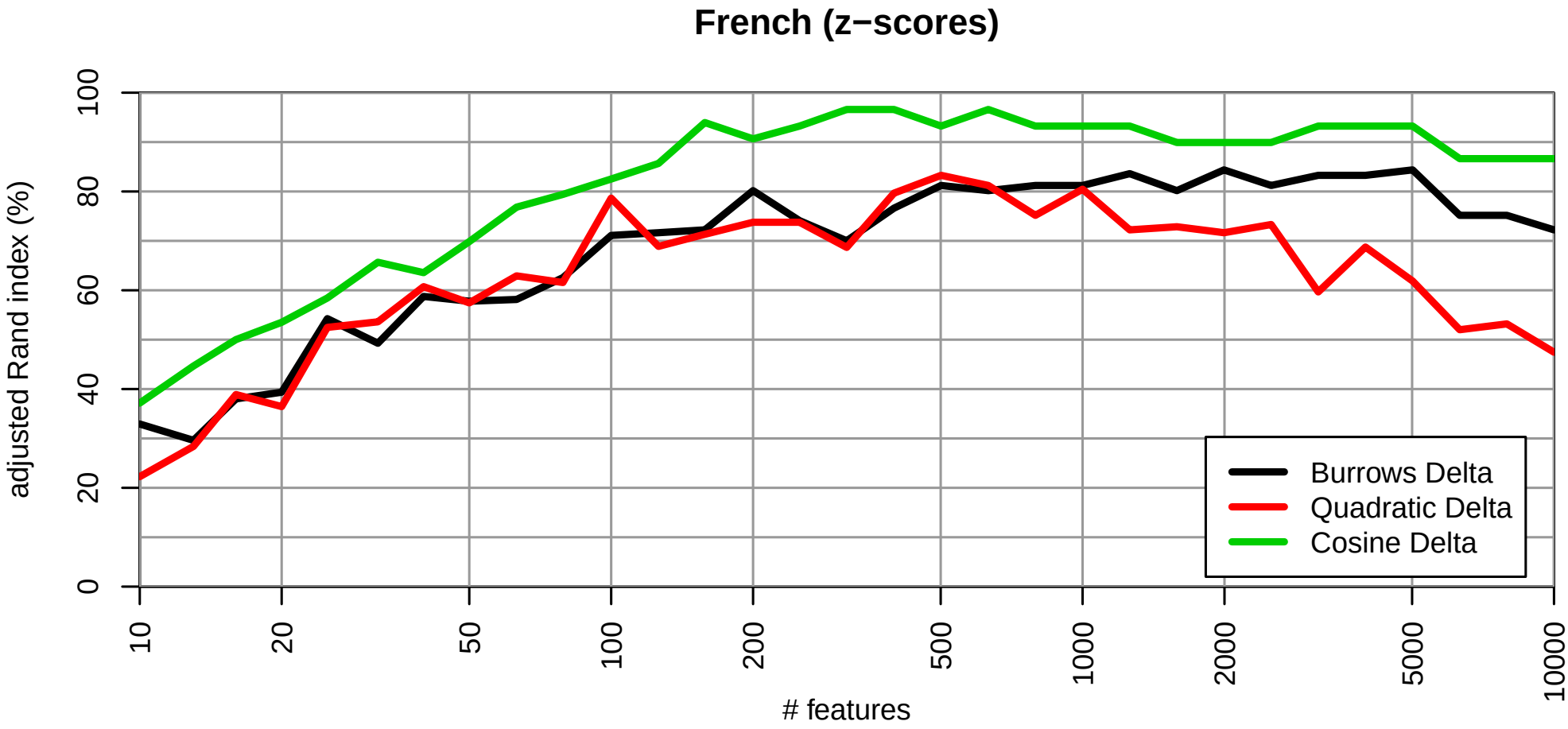


Parameter: Number n_w of MFW



(following slides will show results for German corpus only)

Parameter: Number n_w of MFW

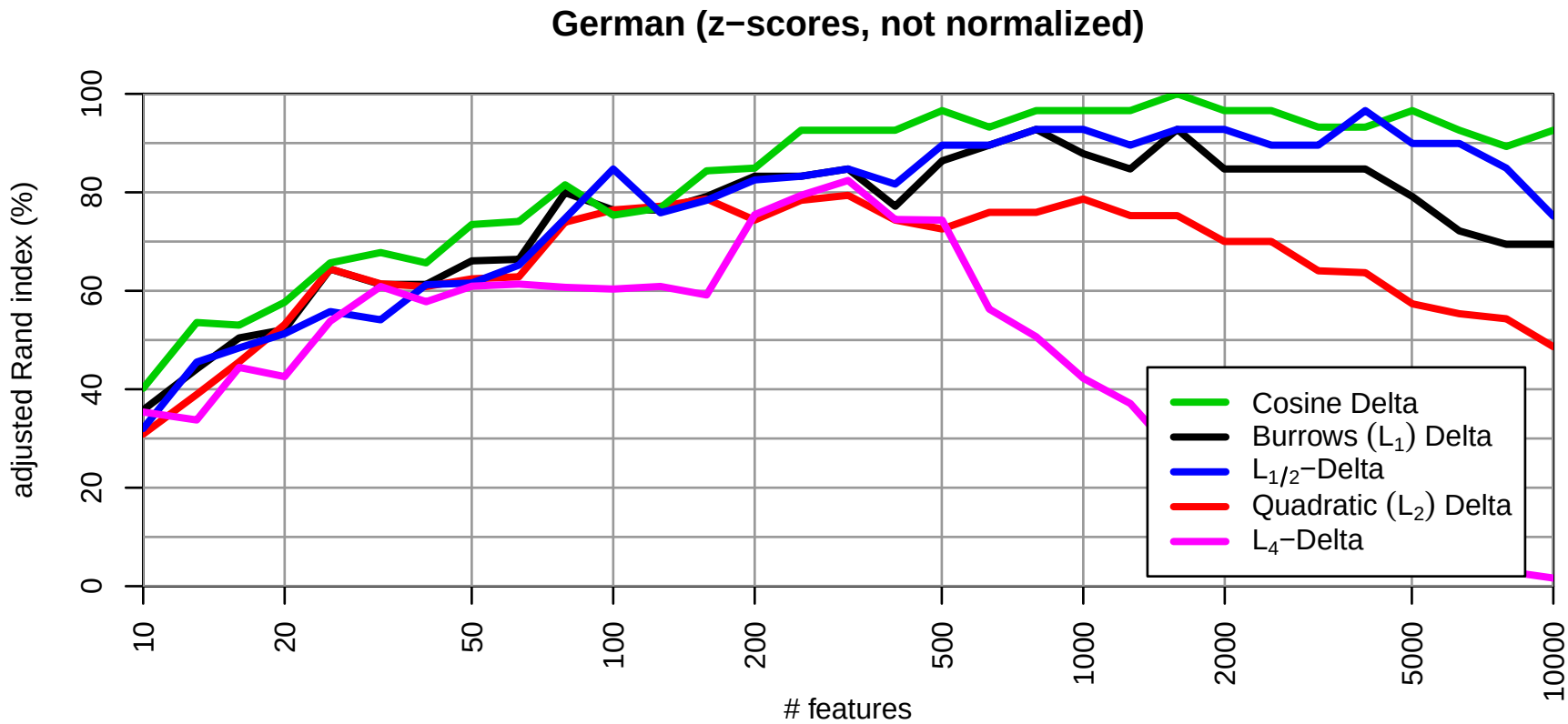


(following slides will show results for German corpus only)

Parameter: Distance metric

- Generalised Minkowski Delta: $\Delta_B = \Delta_1$, $\Delta_Q = \Delta_2$
- Δ_p becomes less "robust" with increasing $p \in [0, \infty]$

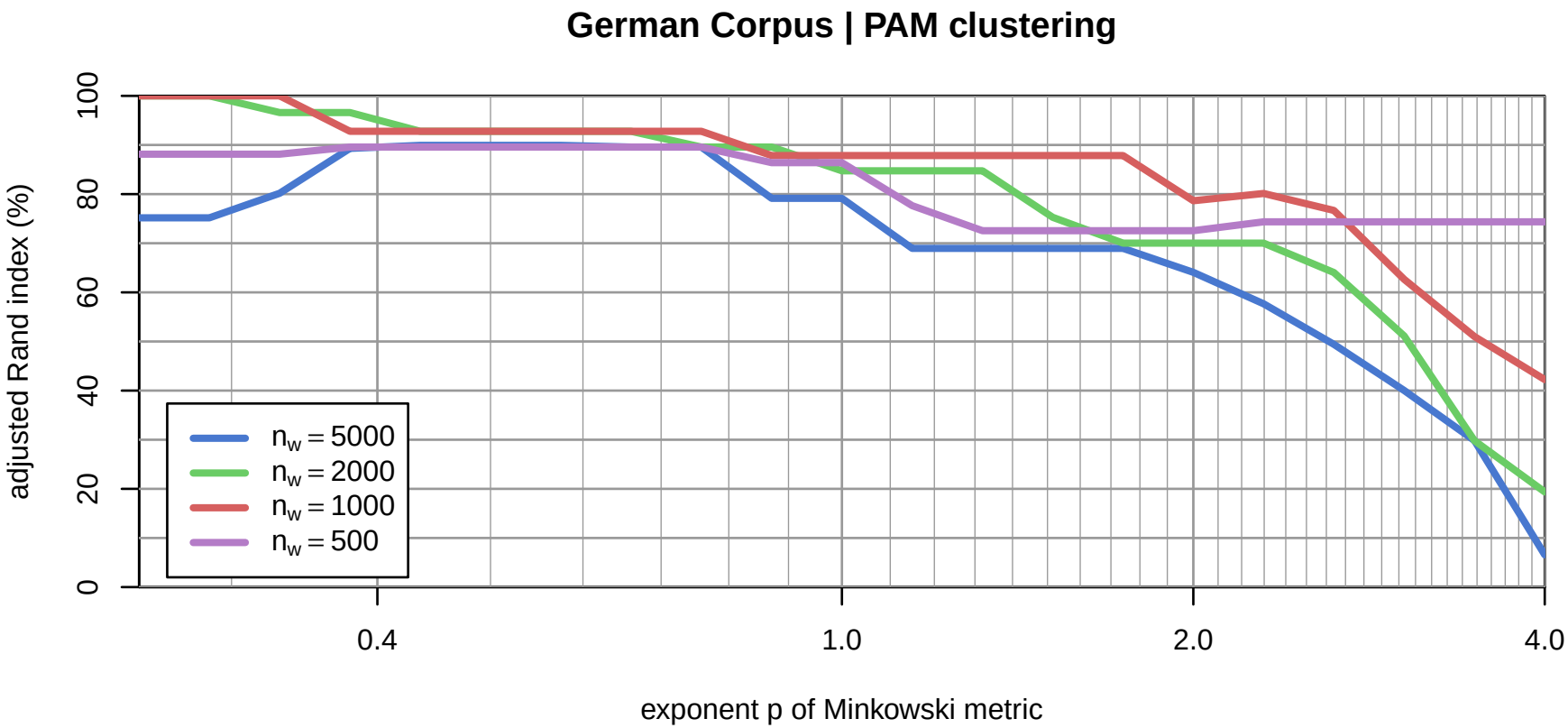
$$\Delta_p(D_1, D_2) = \|\mathbf{z}(D_1) - \mathbf{z}(D_2)\|_p = \sqrt[p]{\sum_{i=1}^{n_w} |z_i(D_1) - z_i(D_2)|^p}$$



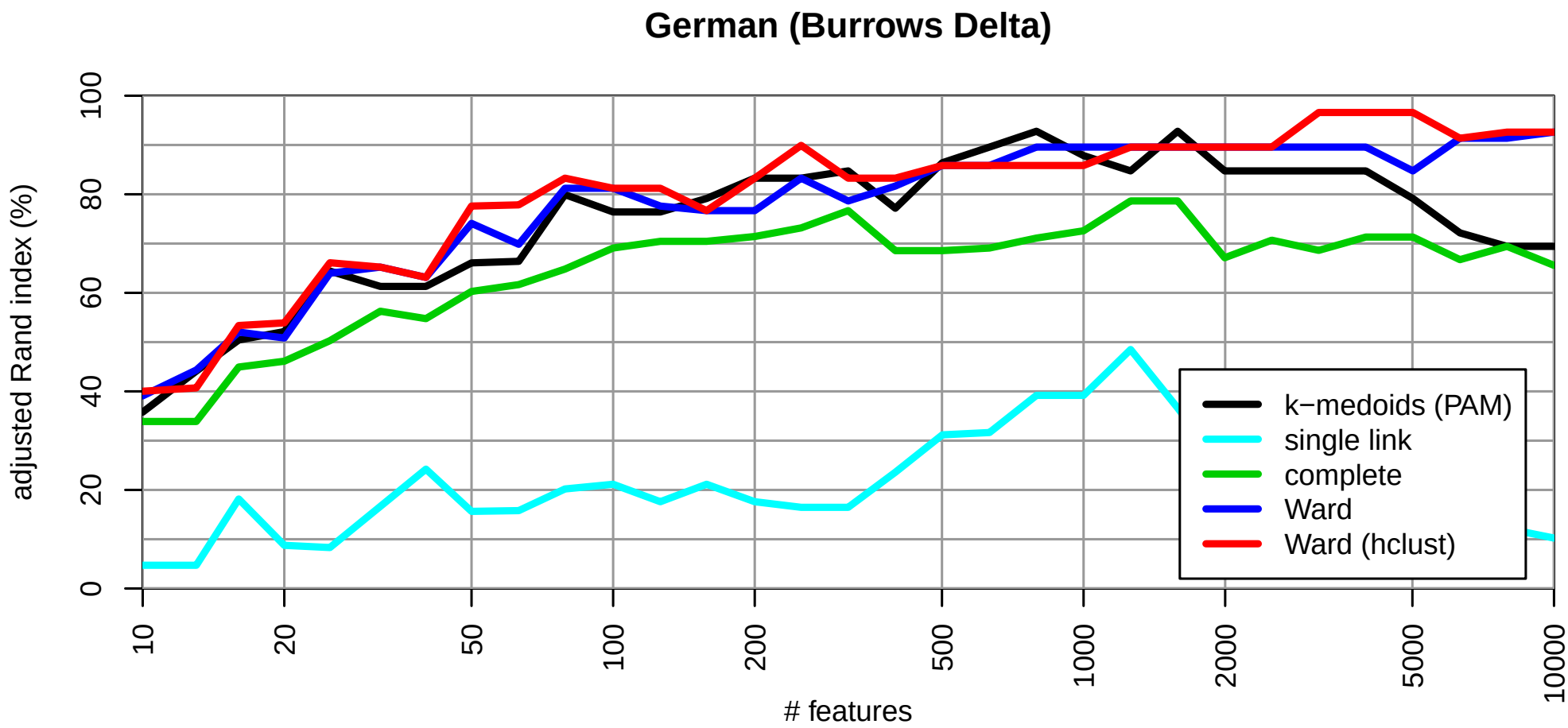
Parameter: Distance metric

- Generalised Minkowski Delta: $\Delta_B = \Delta_1$, $\Delta_Q = \Delta_2$
- Δ_p becomes less "robust" with increasing $p \in [0, \infty]$

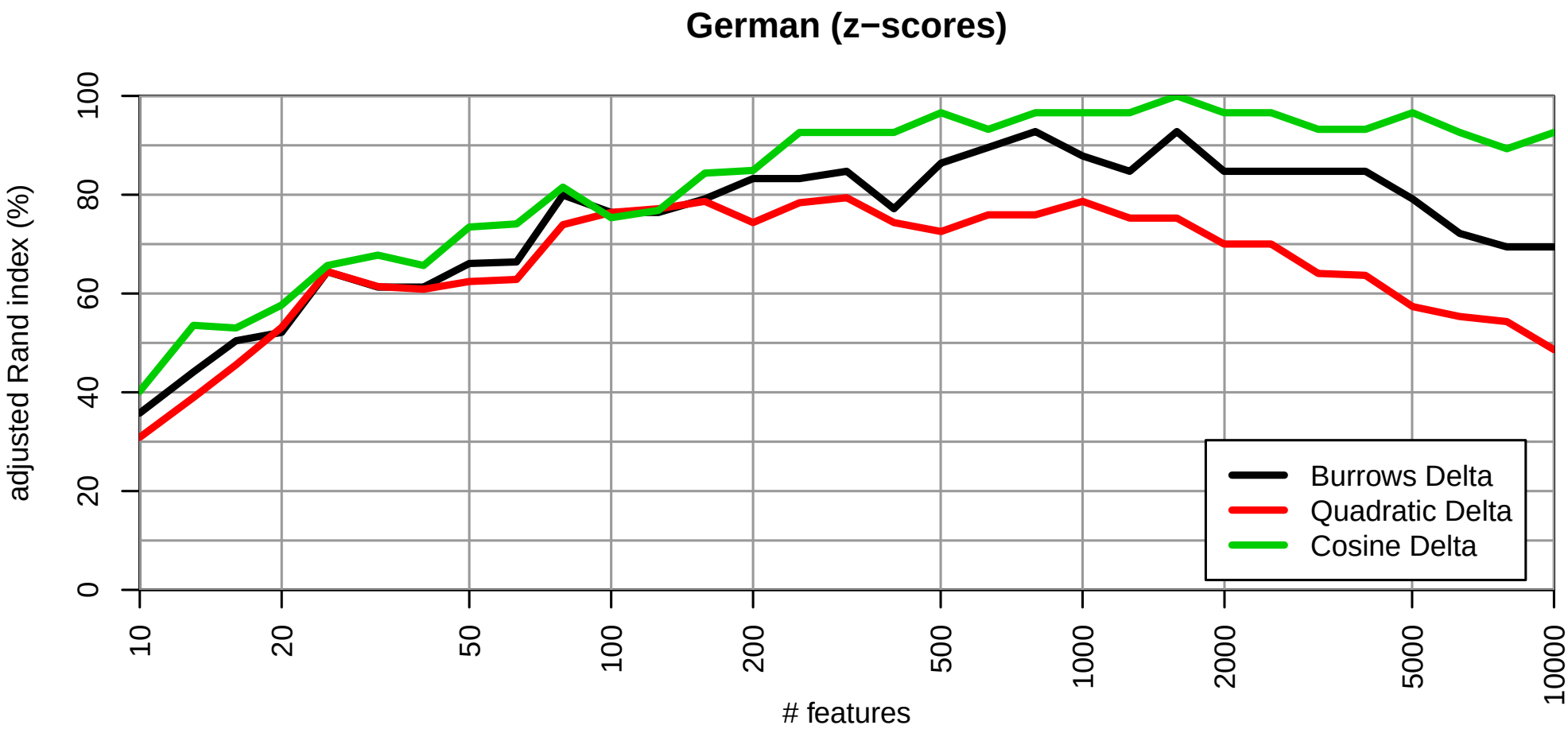
$$\Delta_p(D_1, D_2) = \|\mathbf{z}(D_1) - \mathbf{z}(D_2)\|_p = \sqrt[p]{\sum_{i=1}^{n_w} |z_i(D_1) - z_i(D_2)|^p}$$



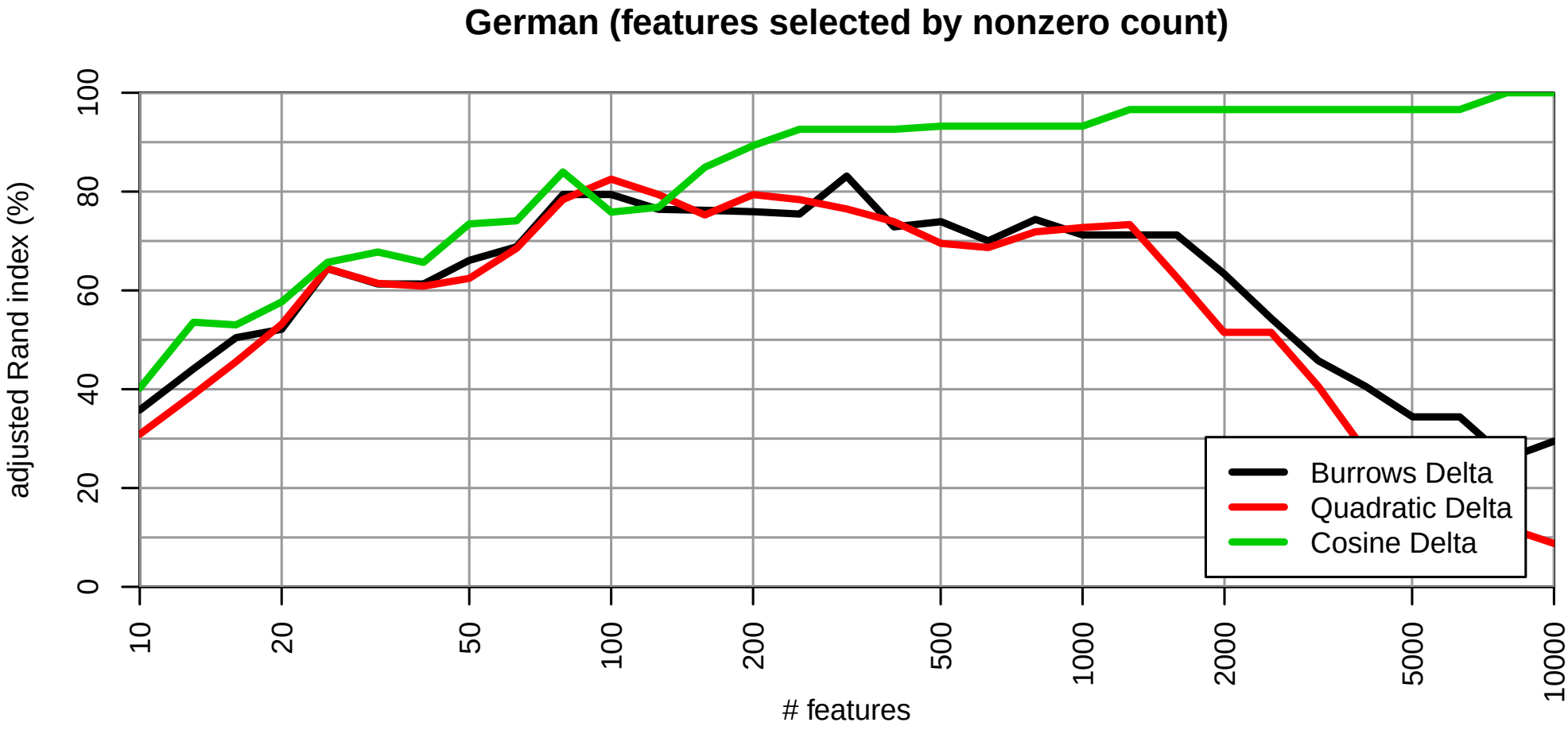
Parameter: Clustering method



Parameter: Feature selection

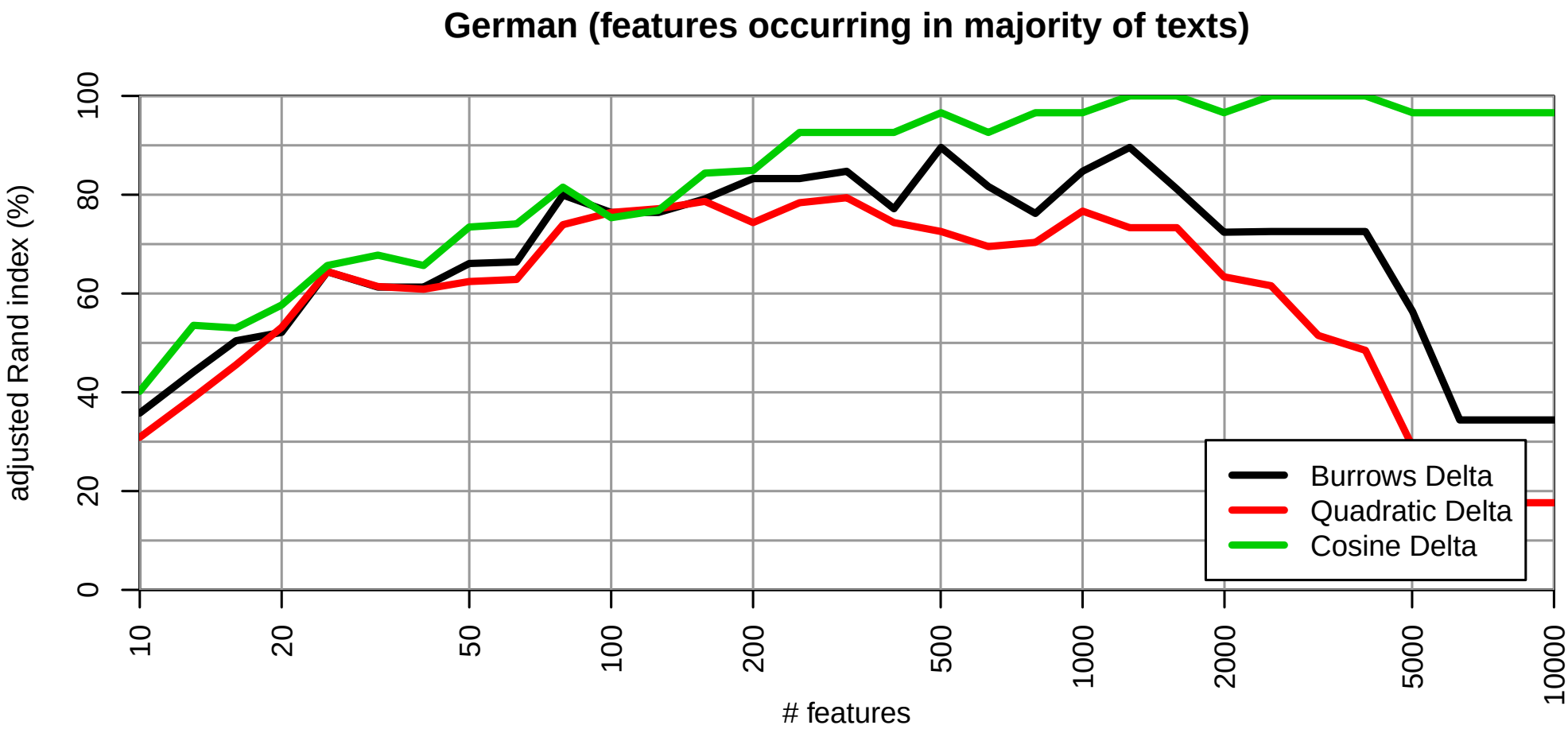


Parameter: Feature selection



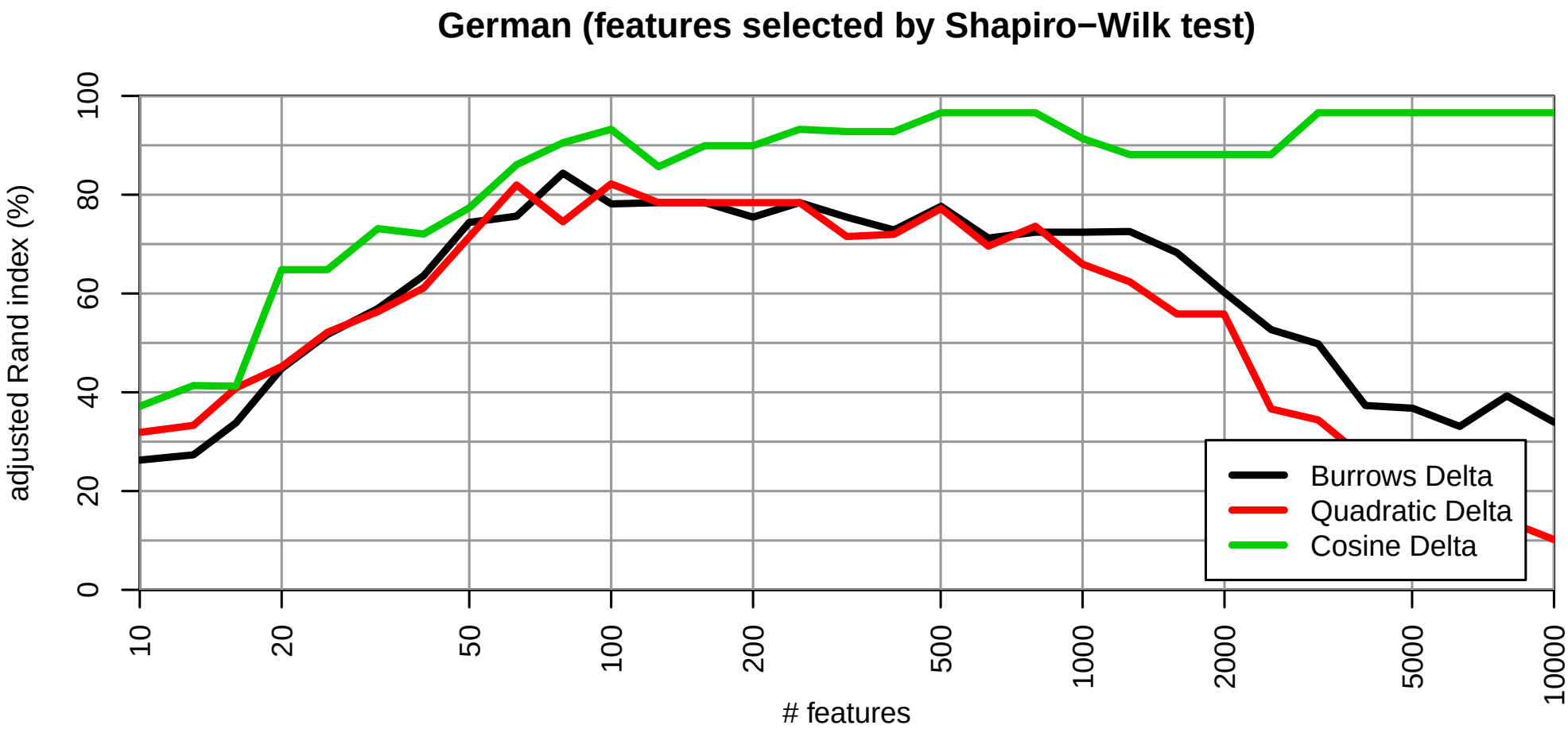
☞ prefer words that occur in most texts

Parameter: Feature selection



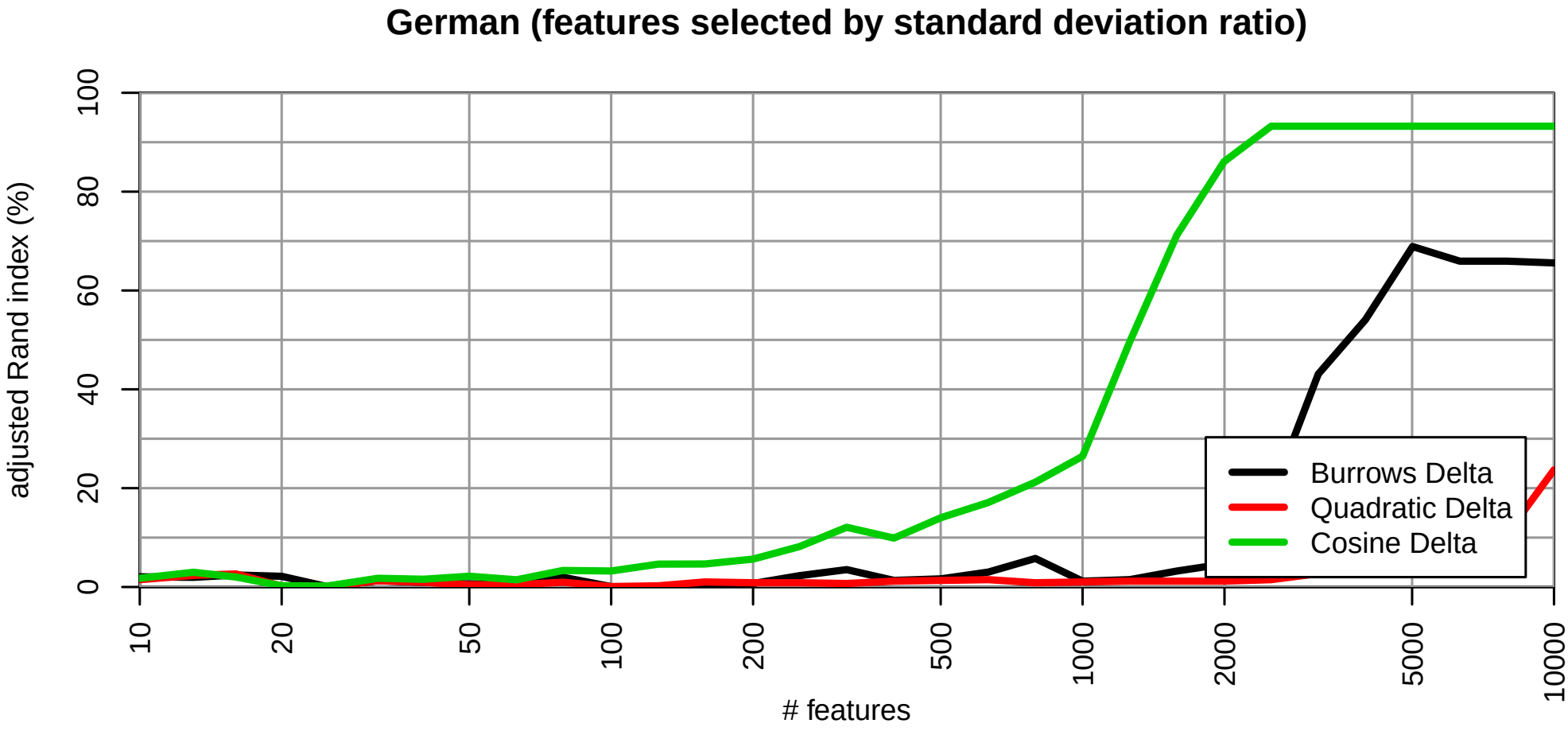
☞ MFW, but must occur in more than half of the texts

Parameter: Feature selection



☞ prefer words with approximately normal distribution

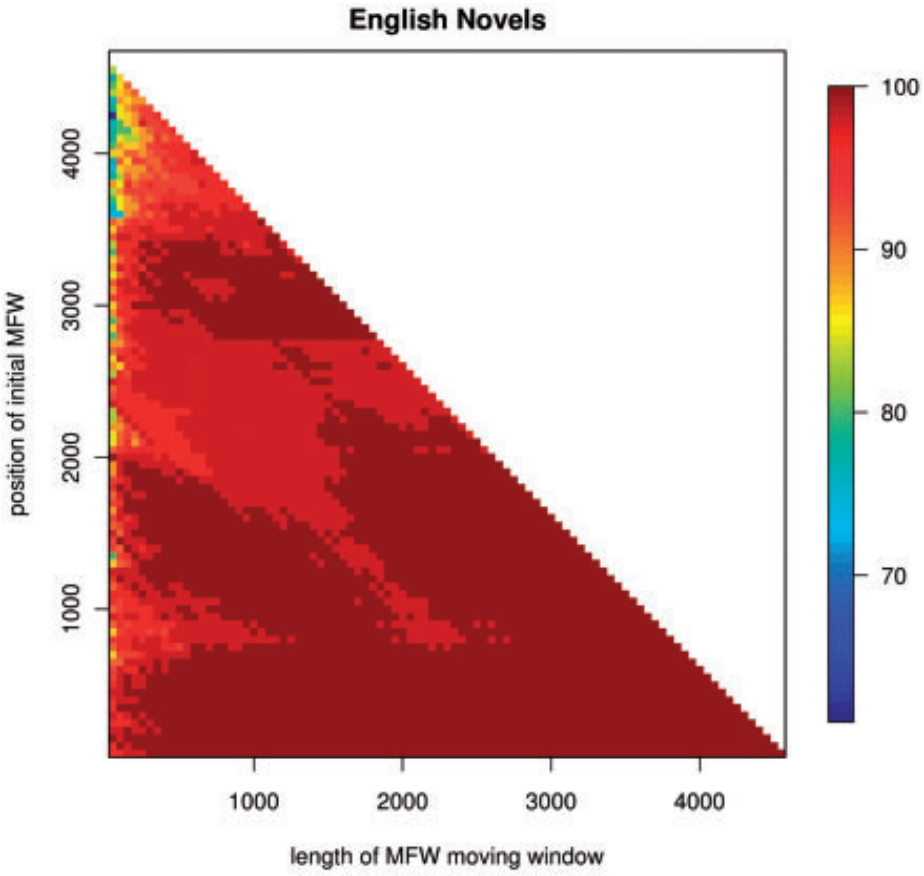
Parameter: Feature selection

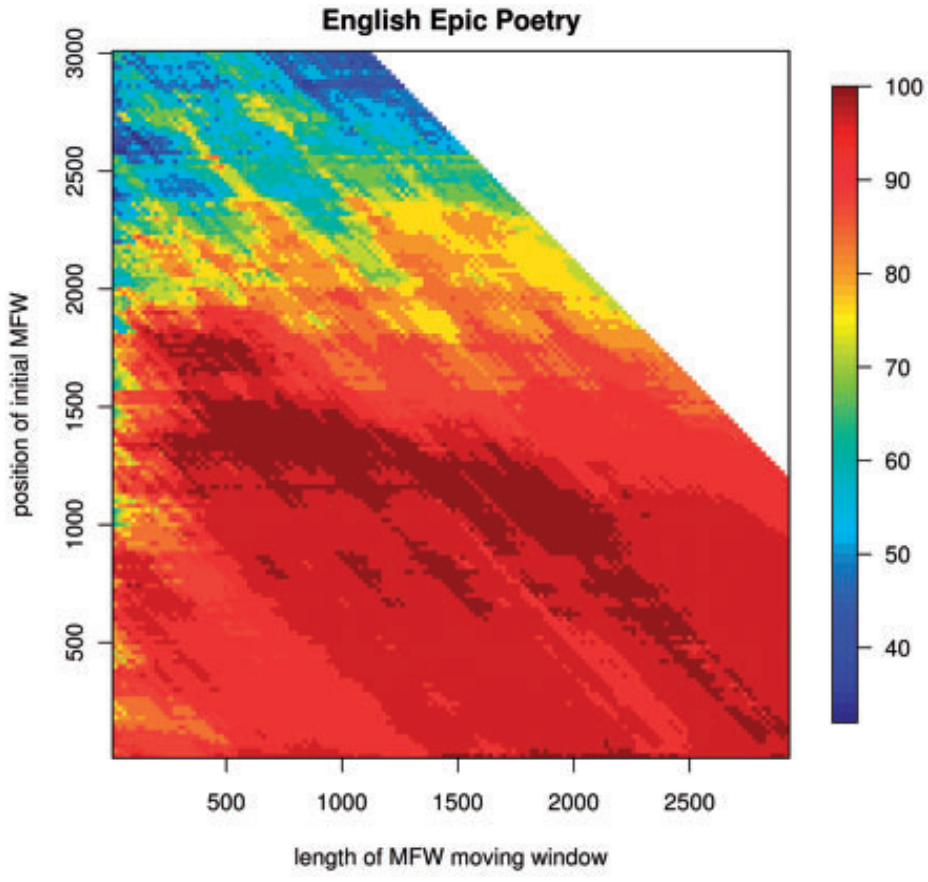


☞ prefer words with unexpectedly large standard deviation

Parameter: Skipping the MFW

(Rybicki and Eder 2011, Fig. 1+2)





65 English novels

32 epic poems in English

attribution accuracy of 1-NN classifier

Parameter: Character n-grams instead of MFW

n	all characters	only alpha-numeric	within words
2	0.9226	0.8510	0.8515
3	***0.9534***	0.9276	0.9242
4	0.9473	*0.9299*	*0.9329*
5	0.9407	0.9250	0.9320
6	0.9349	0.9078	0.9296
7	0.9432	0.9043	0.9220

ARI averaged over different n_w

(German novels)

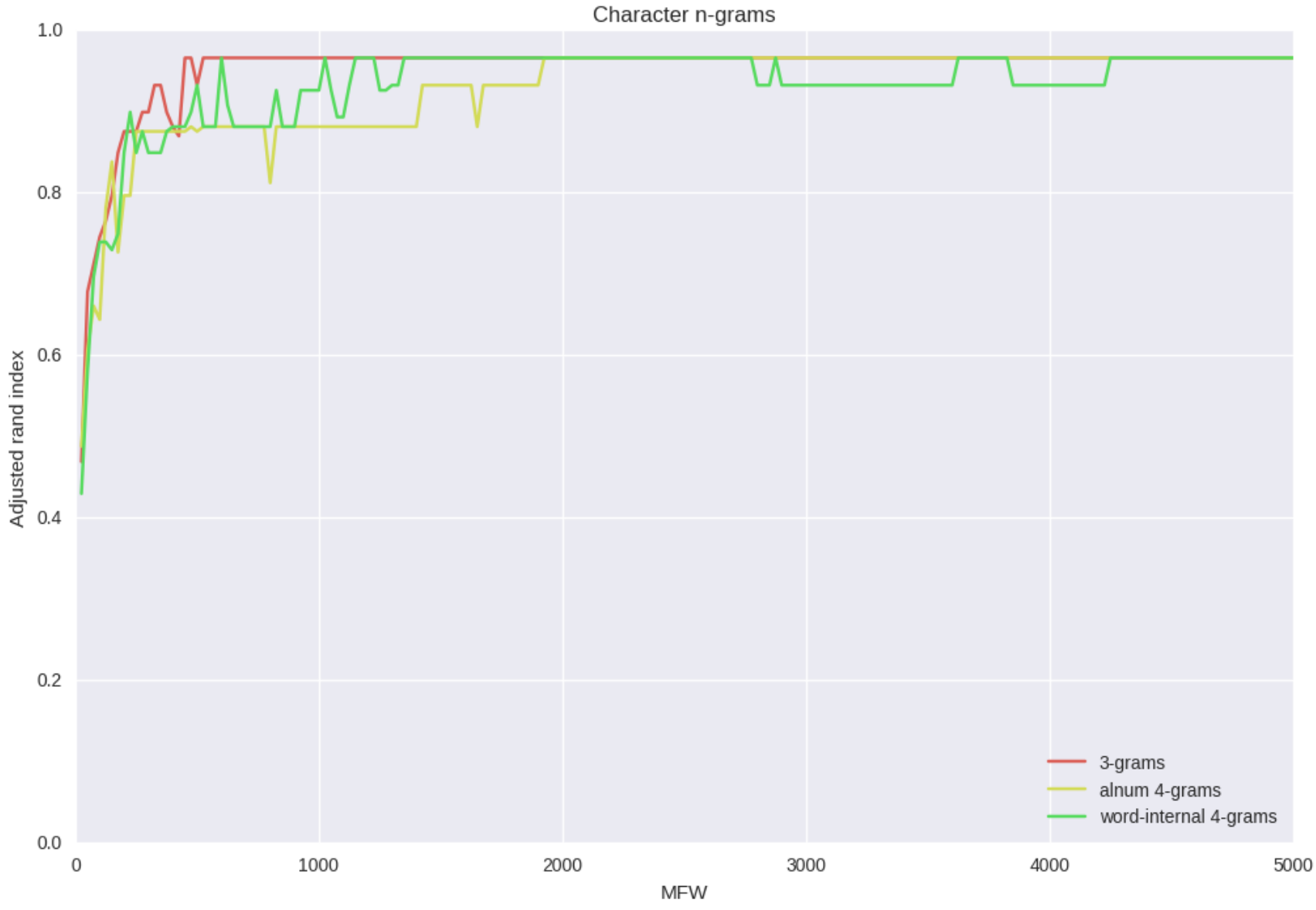
Parameter: Character n-grams instead of MFW

n	all characters	only alpha-numeric	within words
2	0.9226	0.8510	0.8515
3	***0.9534***	0.9276	0.9242
4	0.9473	*0.9299*	*0.9329*
5	0.9407	0.9250	0.9320
6	0.9349	0.9078	0.9296
7	0.9432	0.9043	0.9220

ARI averaged over different n_w

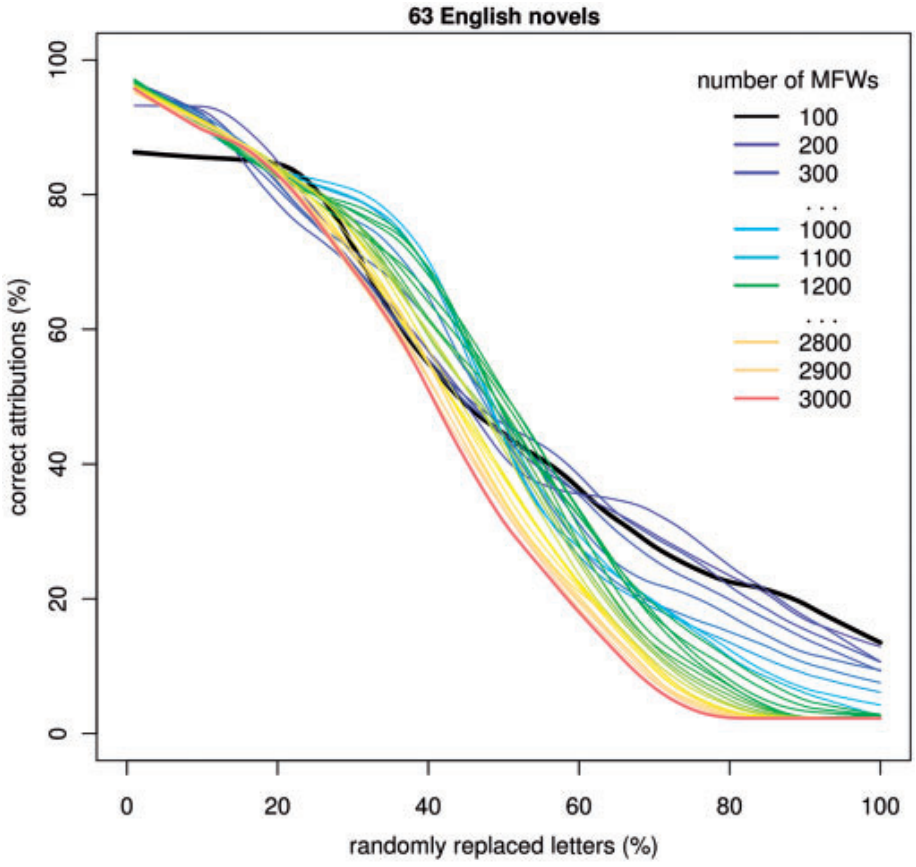
(German novels)

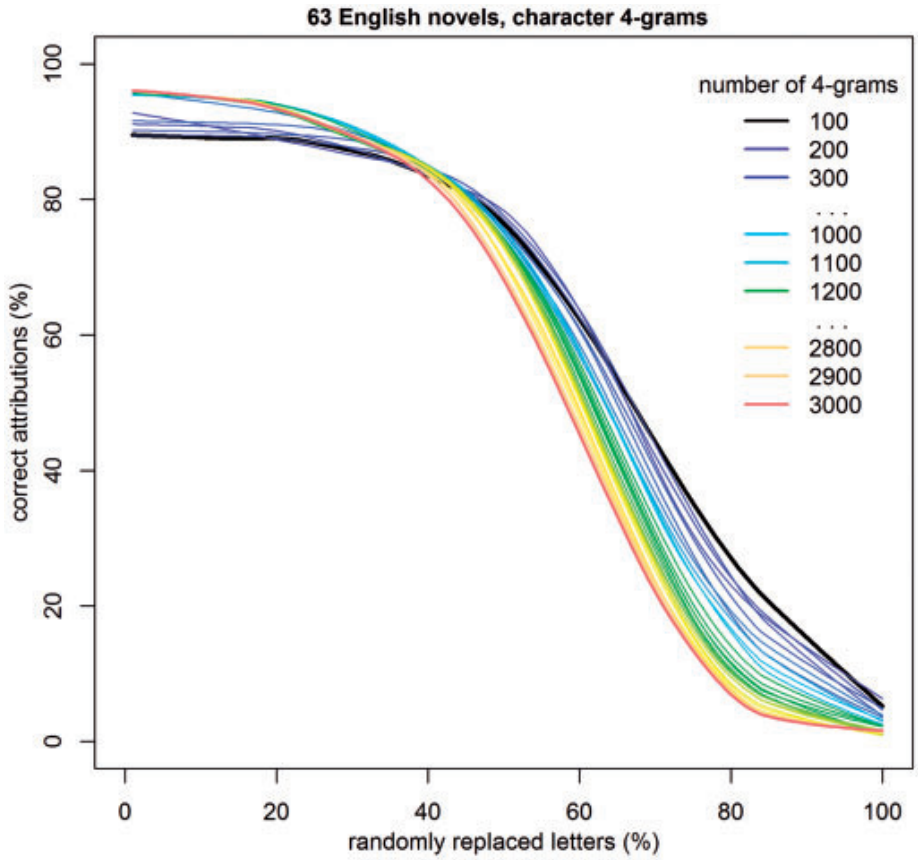
Parameter: Character n-grams instead of MFW



Robustness against noise

(Eder 2013, Fig. 1+3)



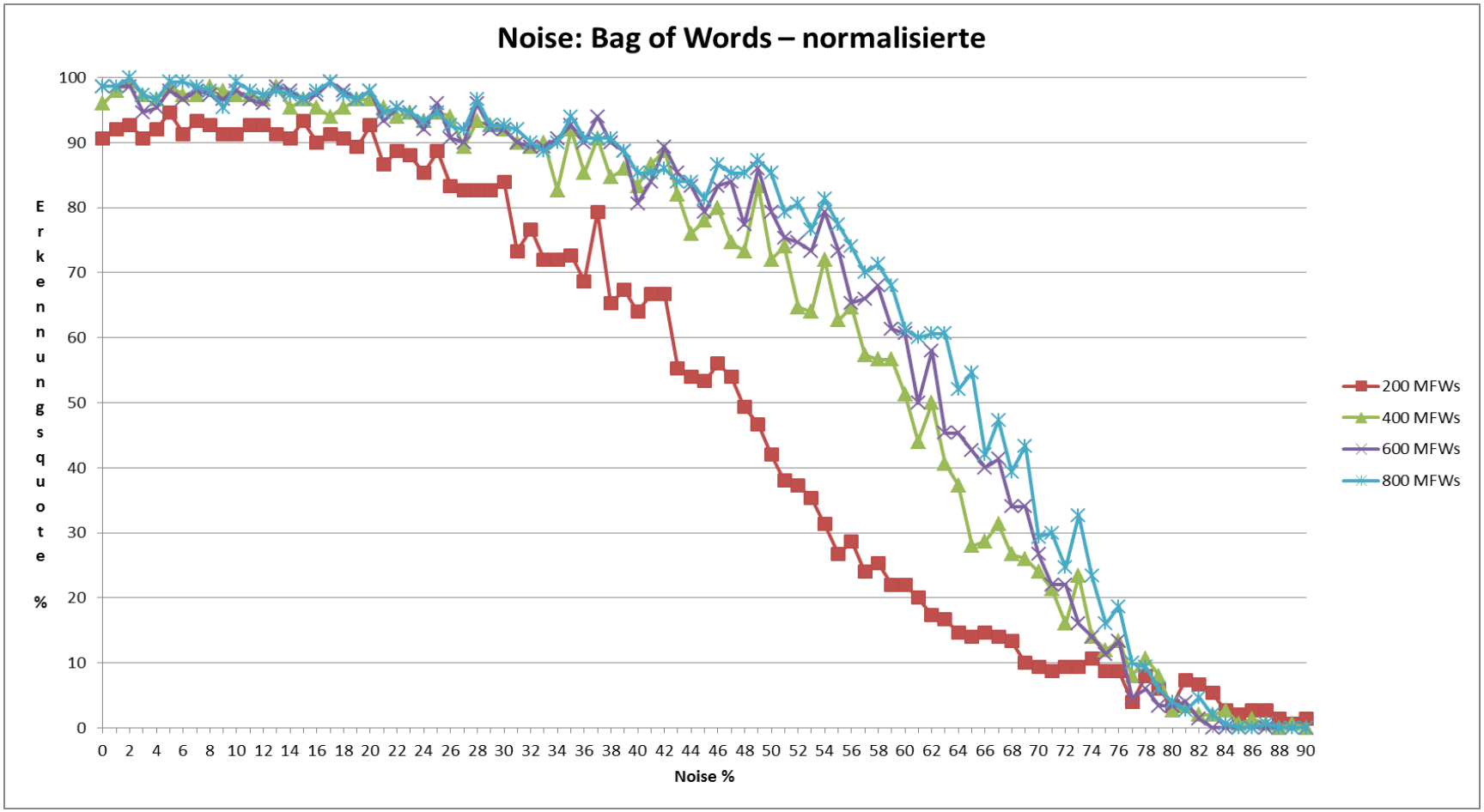


MFW character 4-grams

effect of simulated OCR noise

Robustness against noise

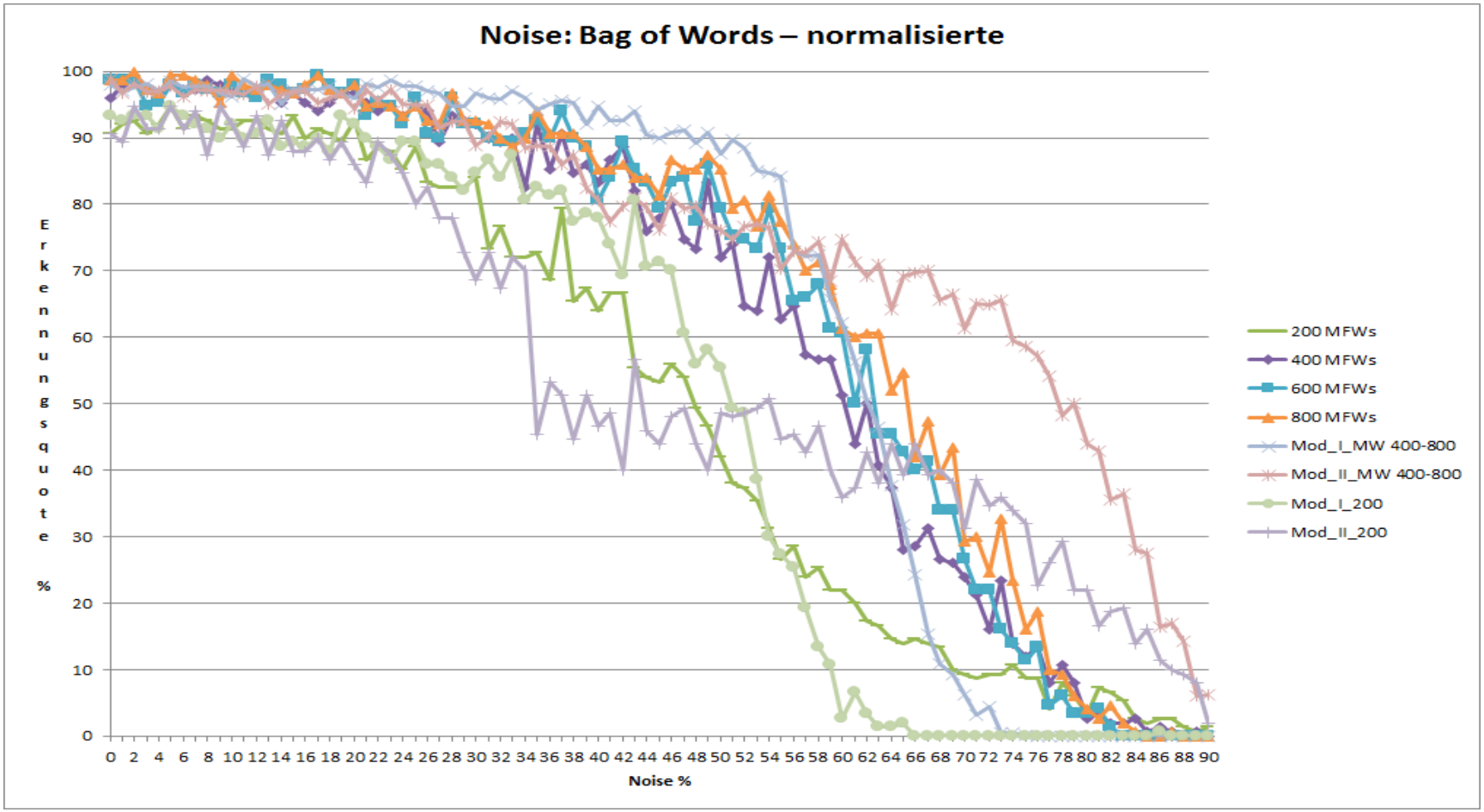
(Büttner *et al.* 2017, Fig. 18+21)



simulated noise in MHG poetry, Δ on 200–800 MFW

Robustness against noise

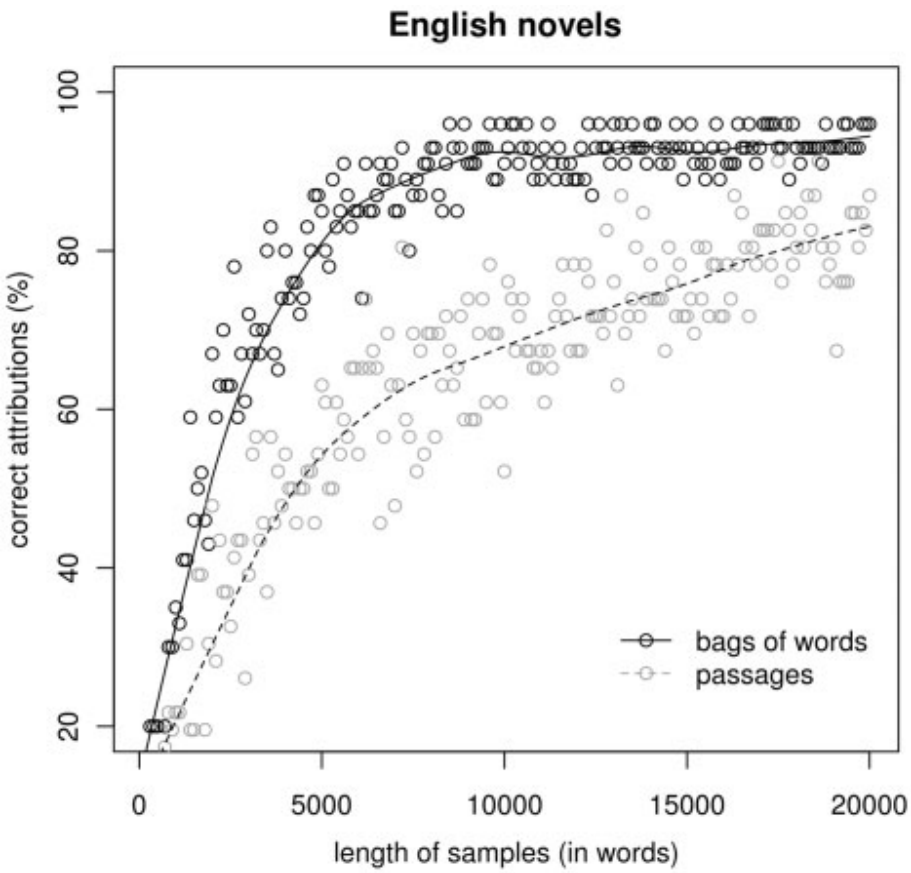
(Büttner *et al.* 2017, Fig. 18+21)

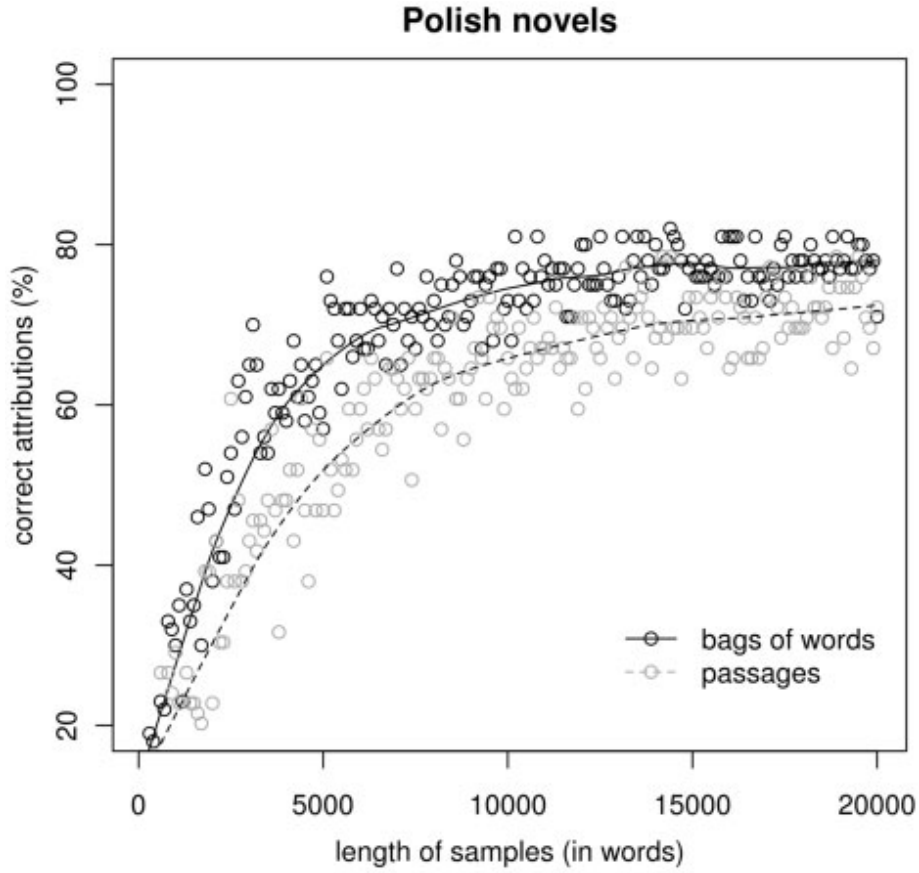


simulated noise in MHG poetry, Δ on 200–800 MFW

Delta and text length

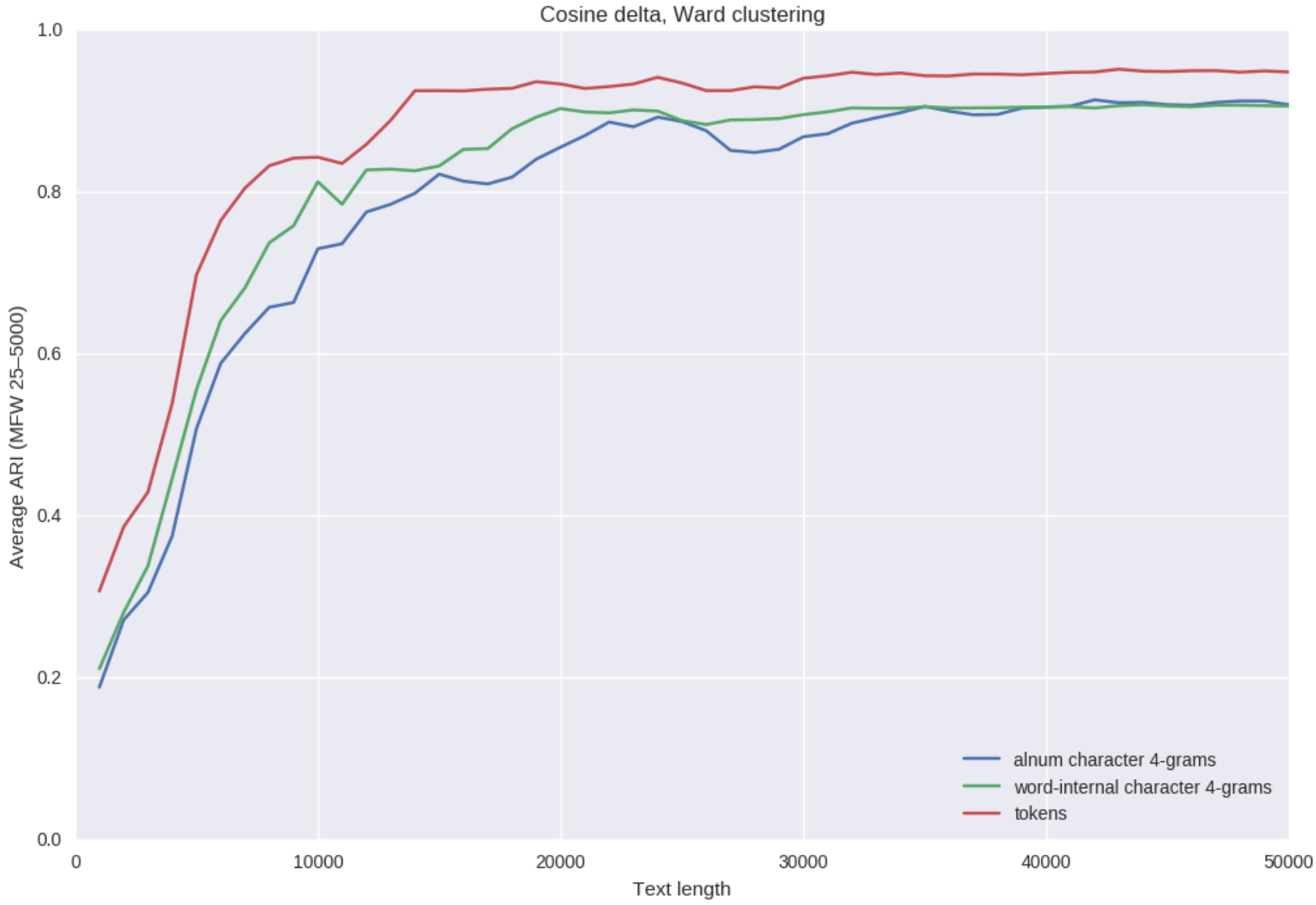
(Eder 2015, Fig. 1+4)





texts shortened by random sampling / excerpt, Δ on 200 MFW

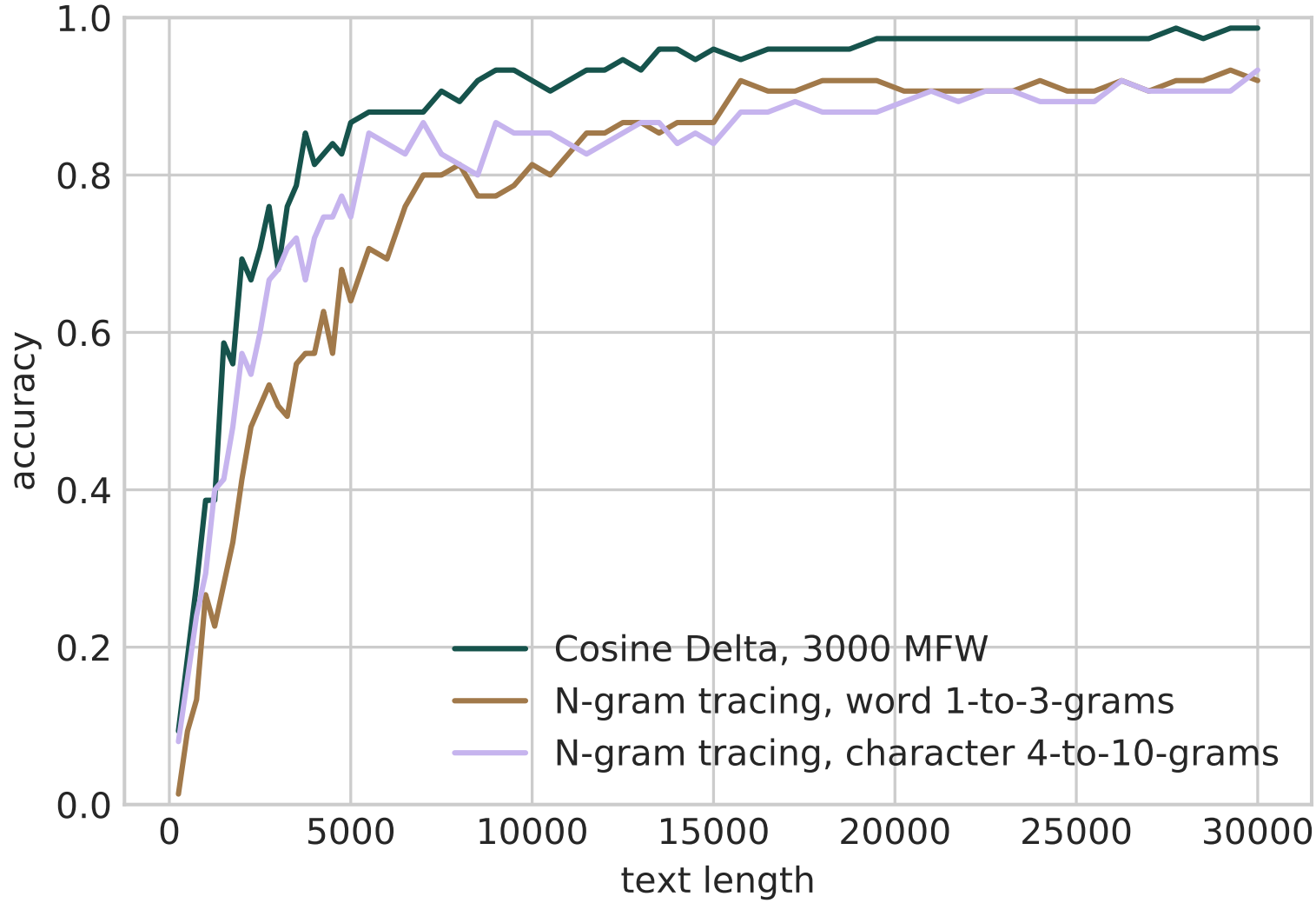
Delta and text length



(German novels)

Delta vs. n-gram tracing on shortened texts

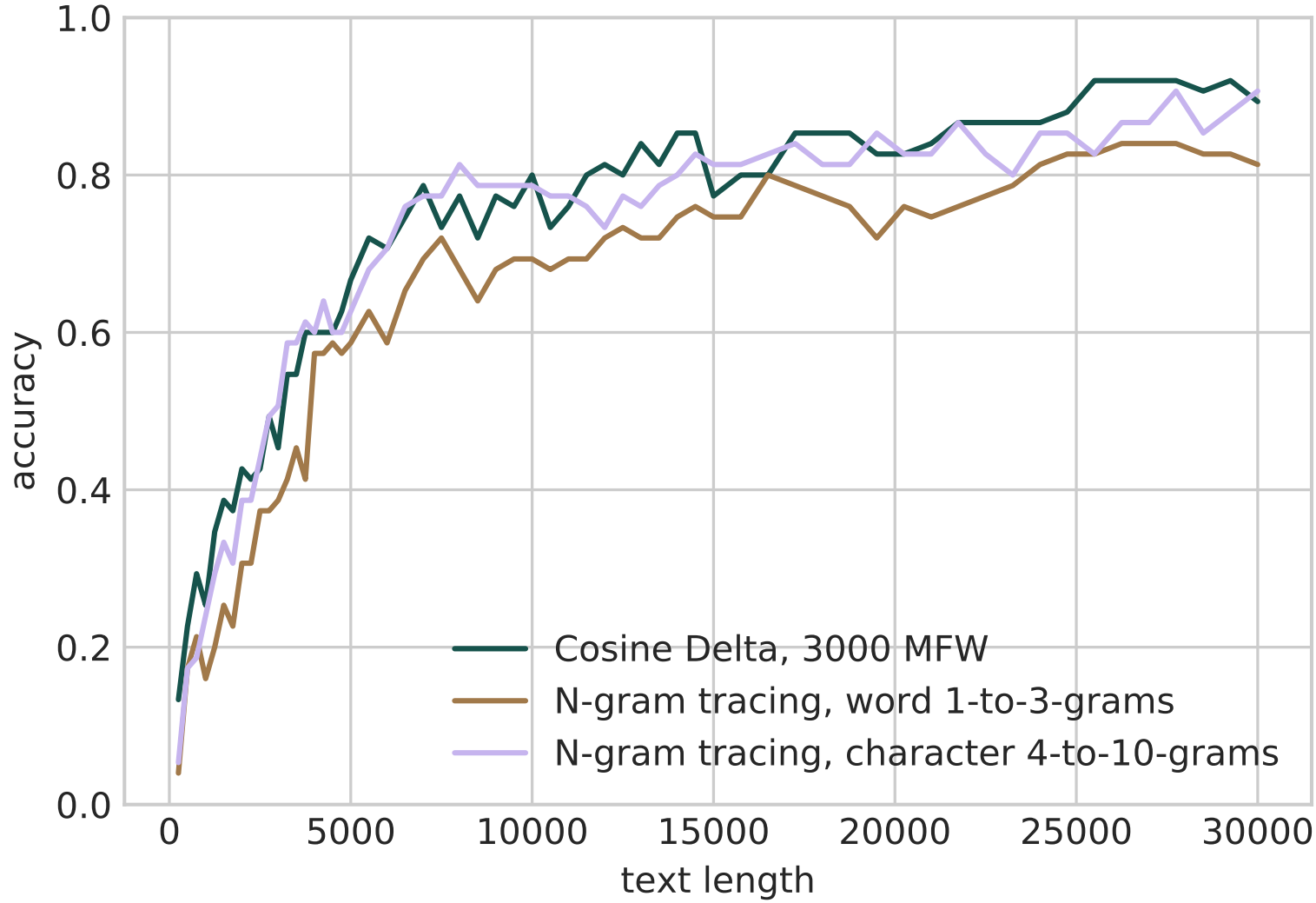
(Grieve *et al.* 2018; Proisl *et al.* 2018)



German novels, all texts shortened

Delta vs. n-gram tracing on shortened texts

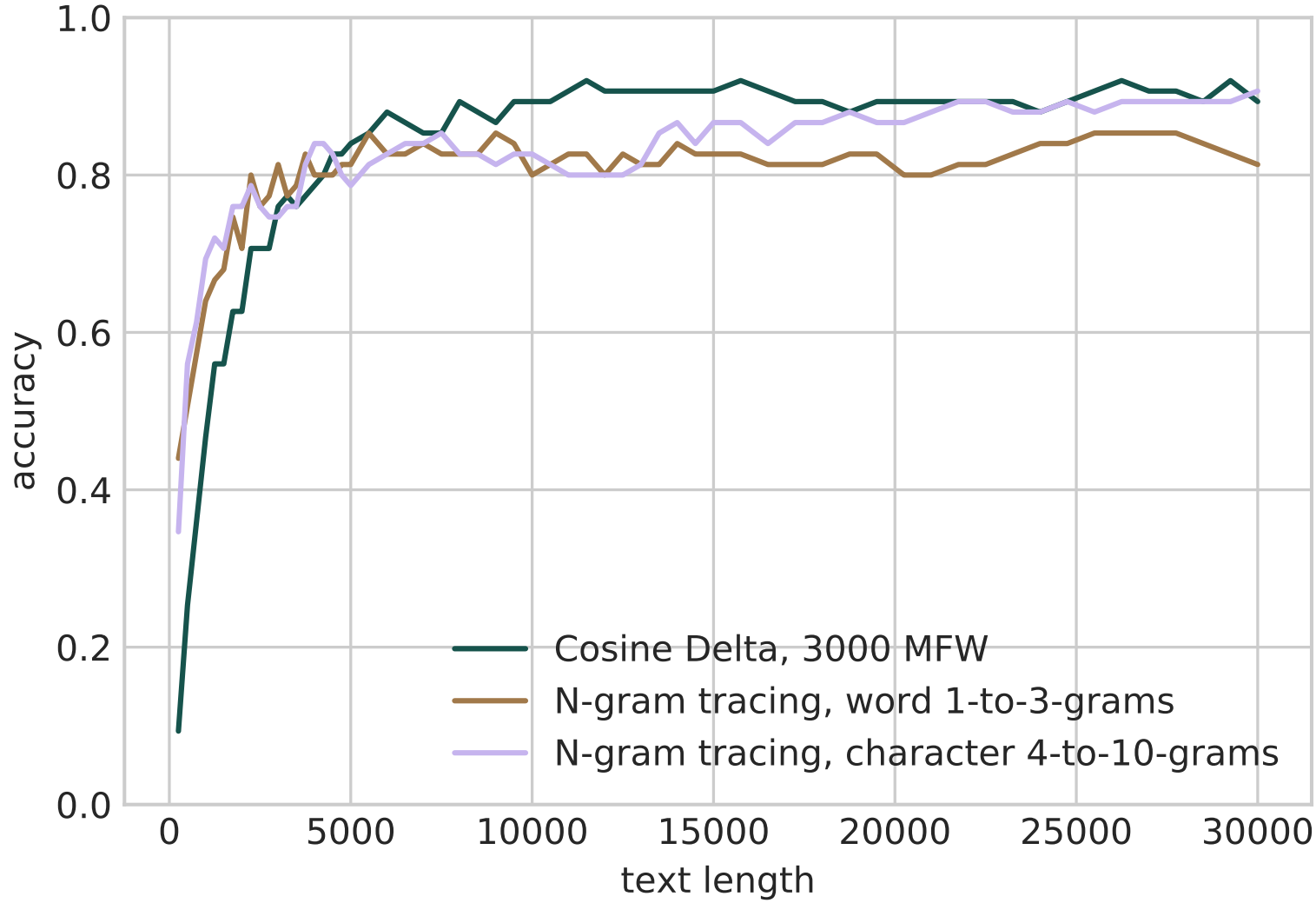
(Grieve *et al.* 2018; Proisl *et al.* 2018)



English novels, all texts shortened

Delta vs. n-gram tracing on shortened texts

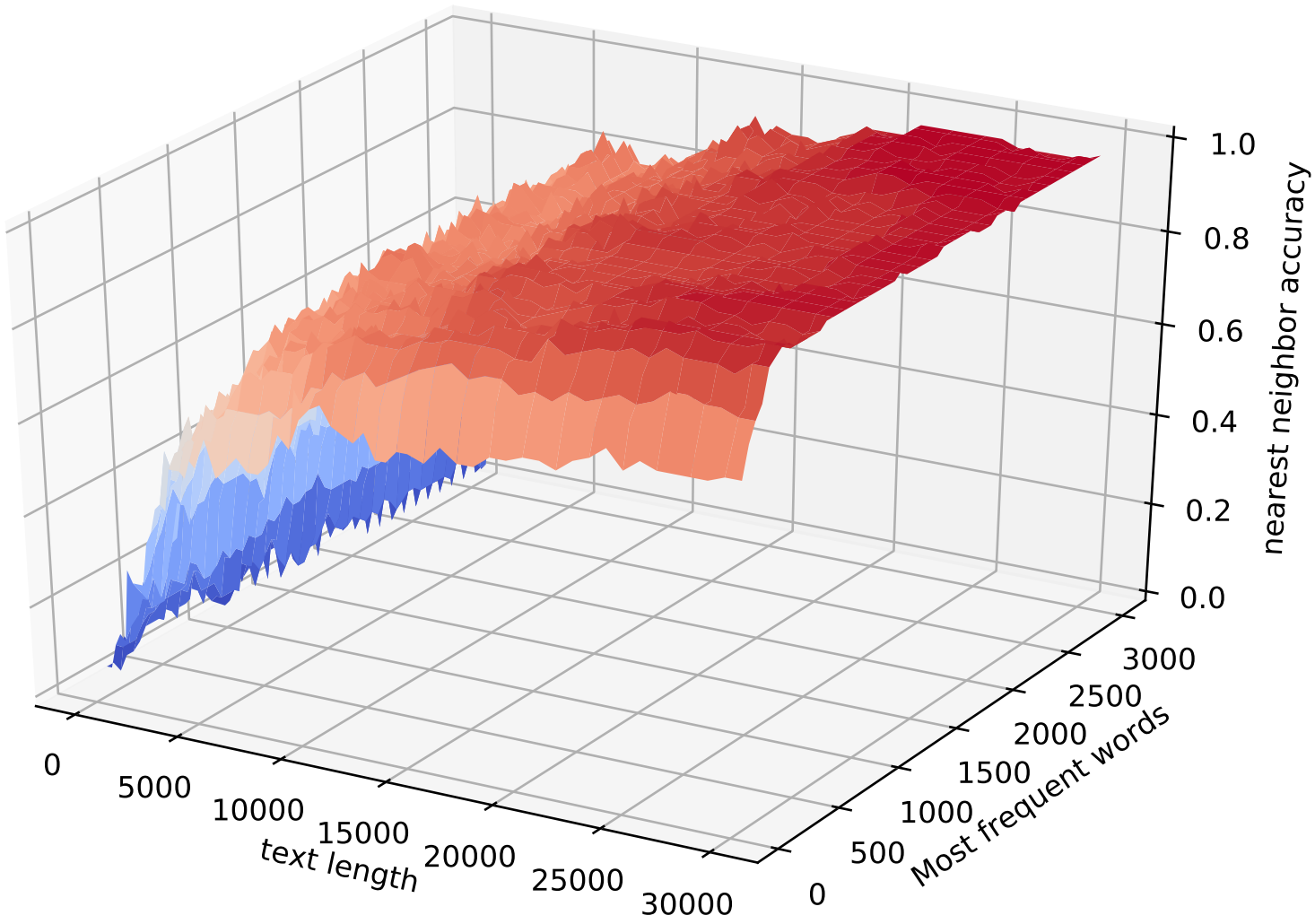
(Grieve *et al.* 2018; Proisl *et al.* 2018)



English novels, only disputed text shortened

Delta vs. n-gram tracing on shortened texts

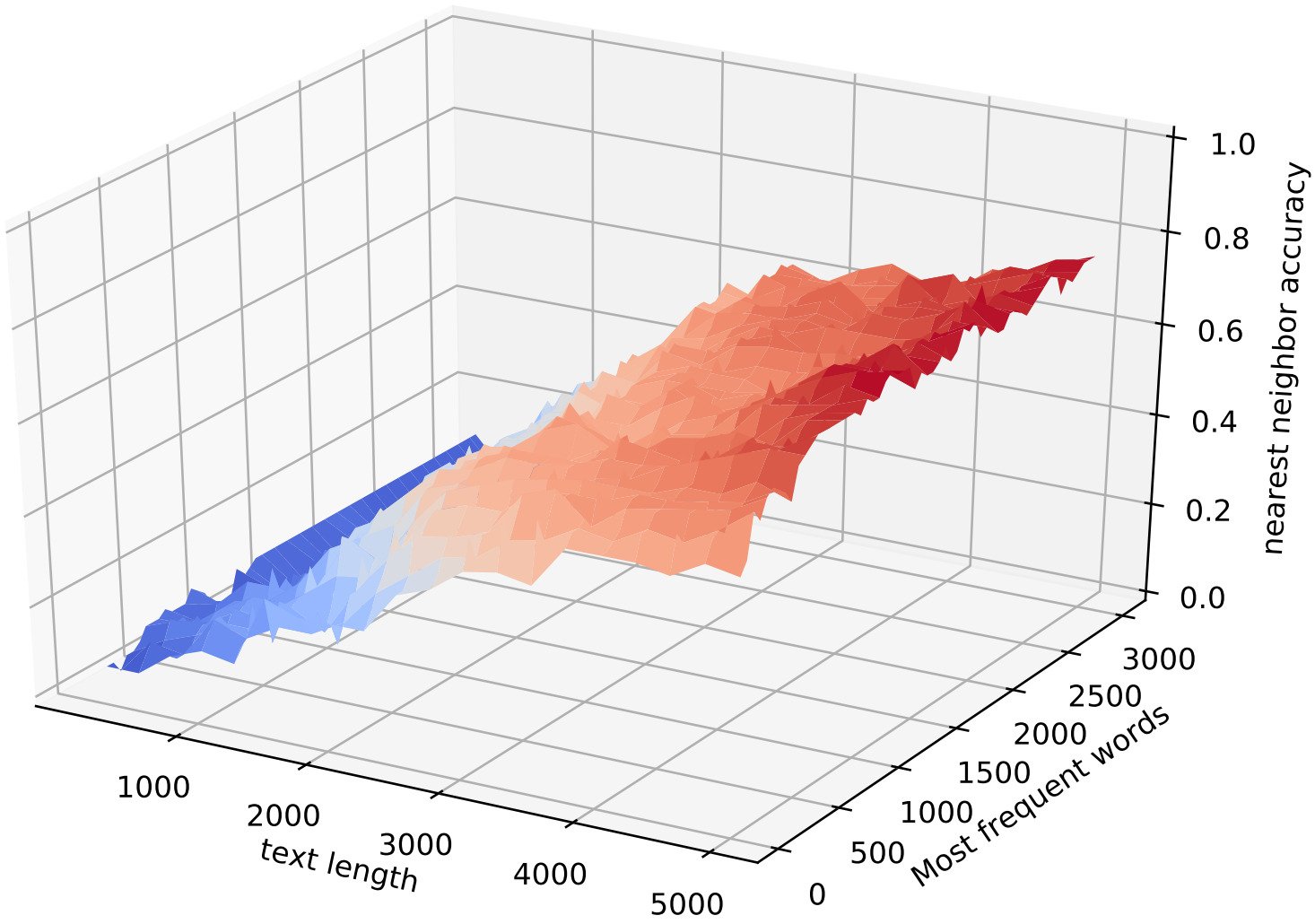
(Grieve *et al.* 2018; Proisl *et al.* 2018)



German novels: are fewer MFW better for short texts?

Delta vs. n-gram tracing on shortened texts

(Grieve *et al.* 2018; Proisl *et al.* 2018)



German novels: are fewer MFW better for short texts?

Outline

Literary authorship attribution

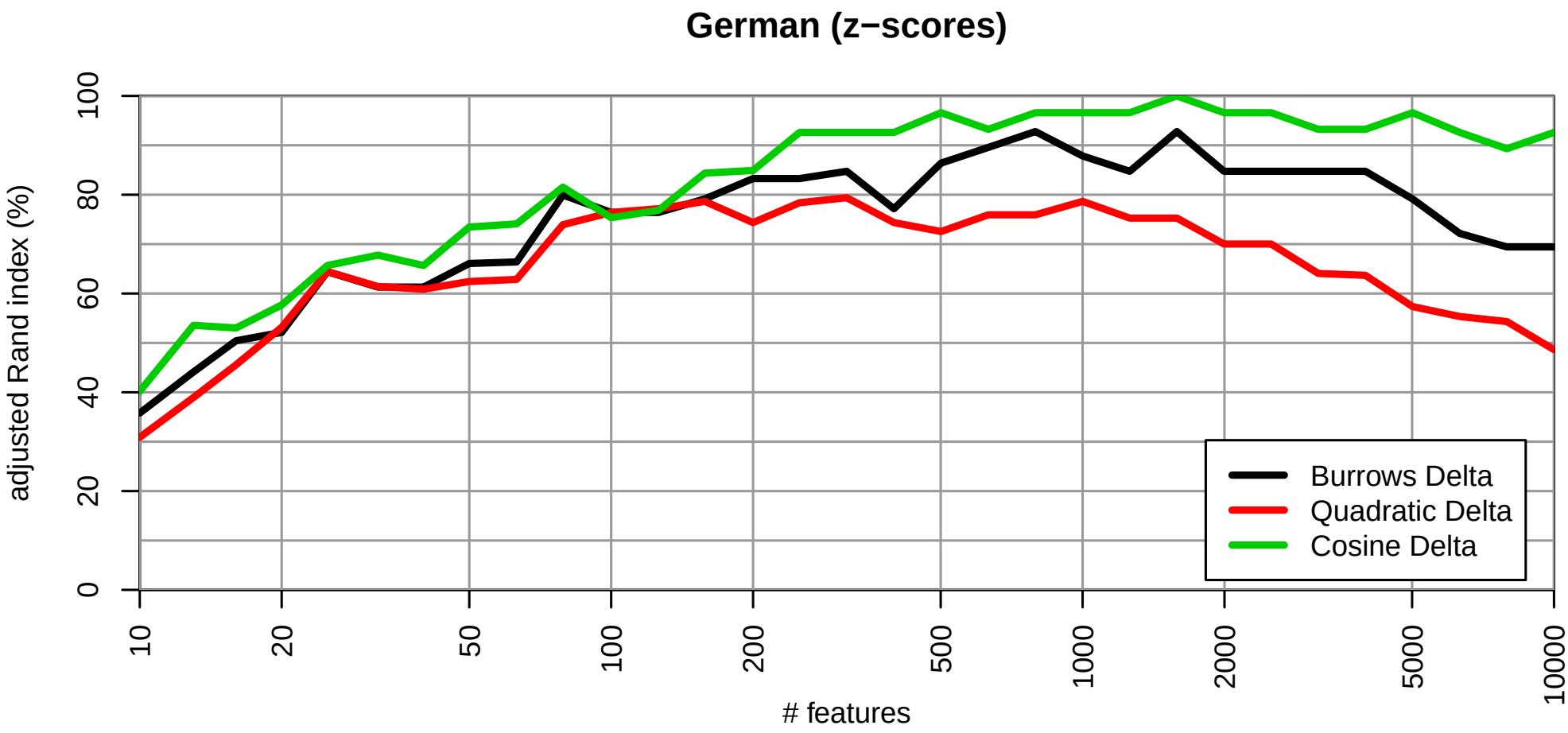
Burrows's Delta

Empirical evaluation

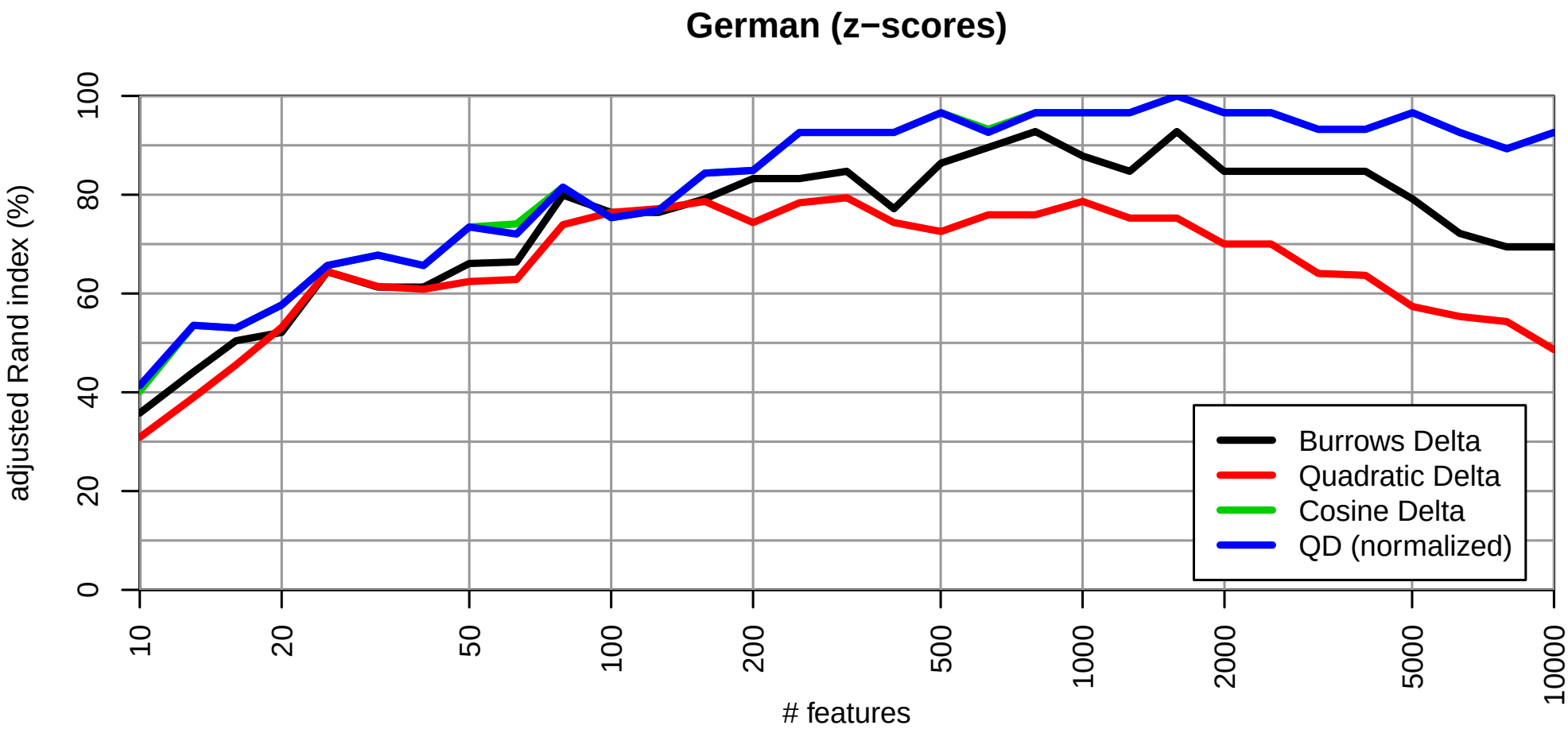
Understanding Delta

Statistical significance

A crucial riddle . . .



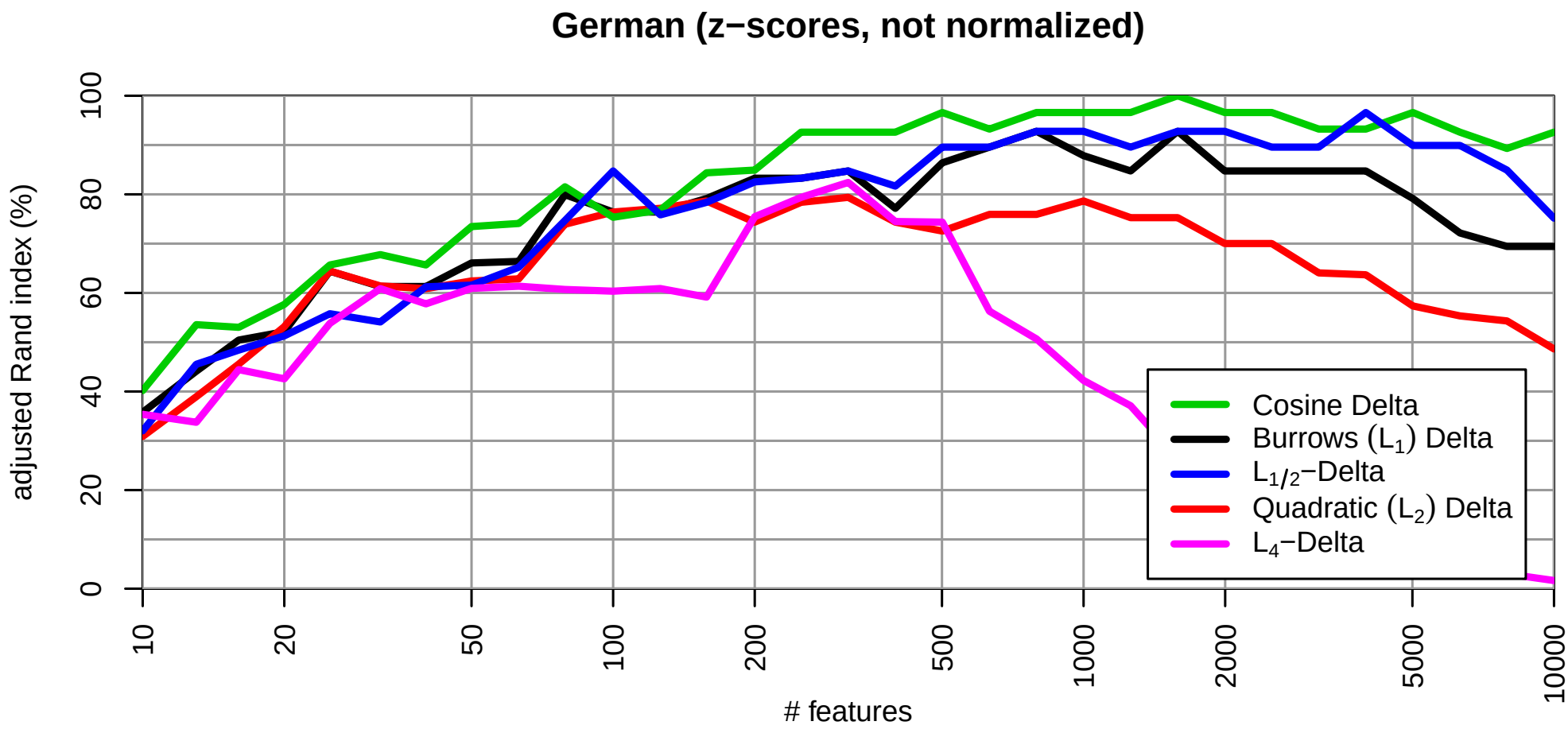
A crucial riddle . . .



Euclidean metric (Δ_Q) + vector normalization = Cosine Delta ($\Delta_\angle$)

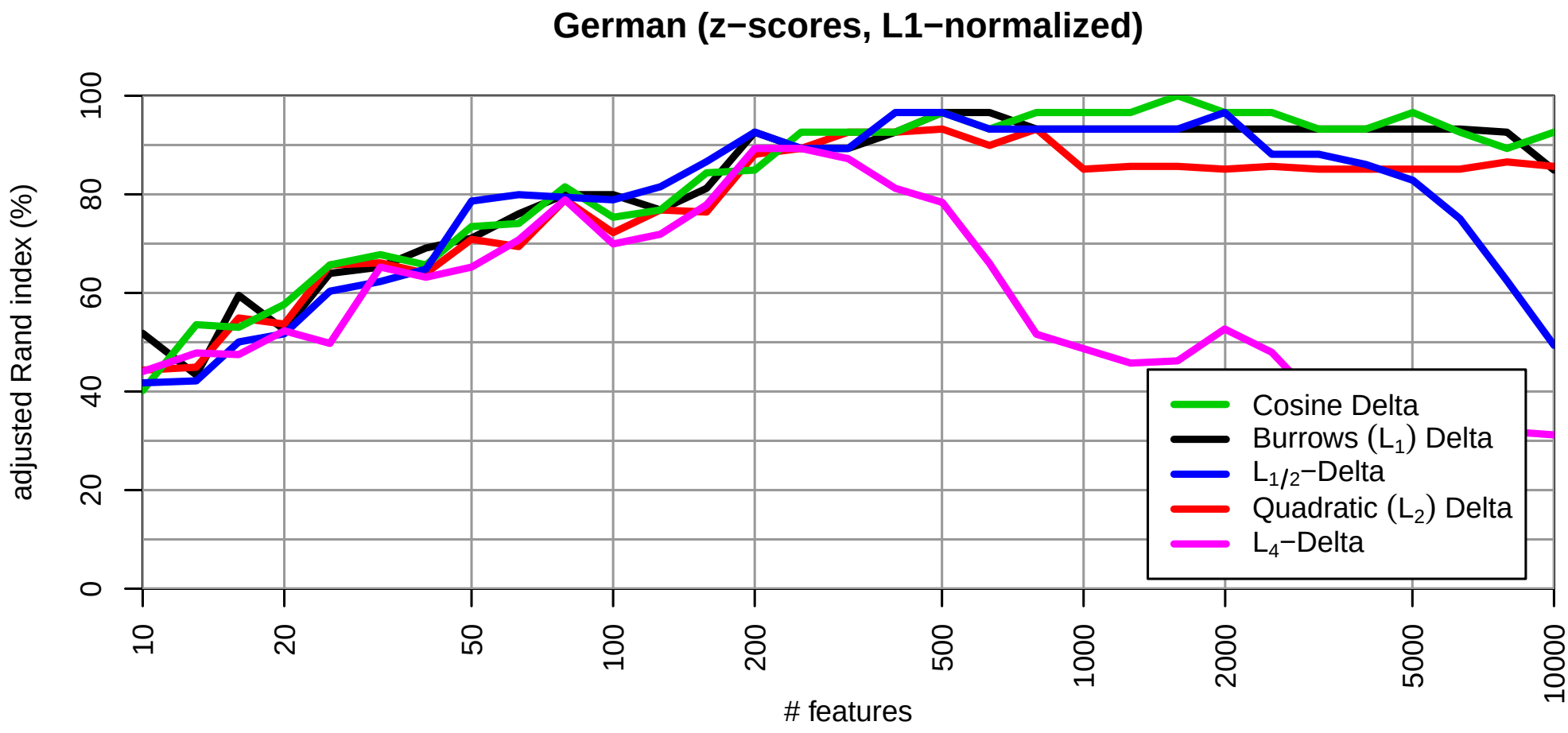
i.e. normalization turns worst Delta measure (Δ_Q) into the best one ($\Delta_\angle$)

The effect of vector length normalization



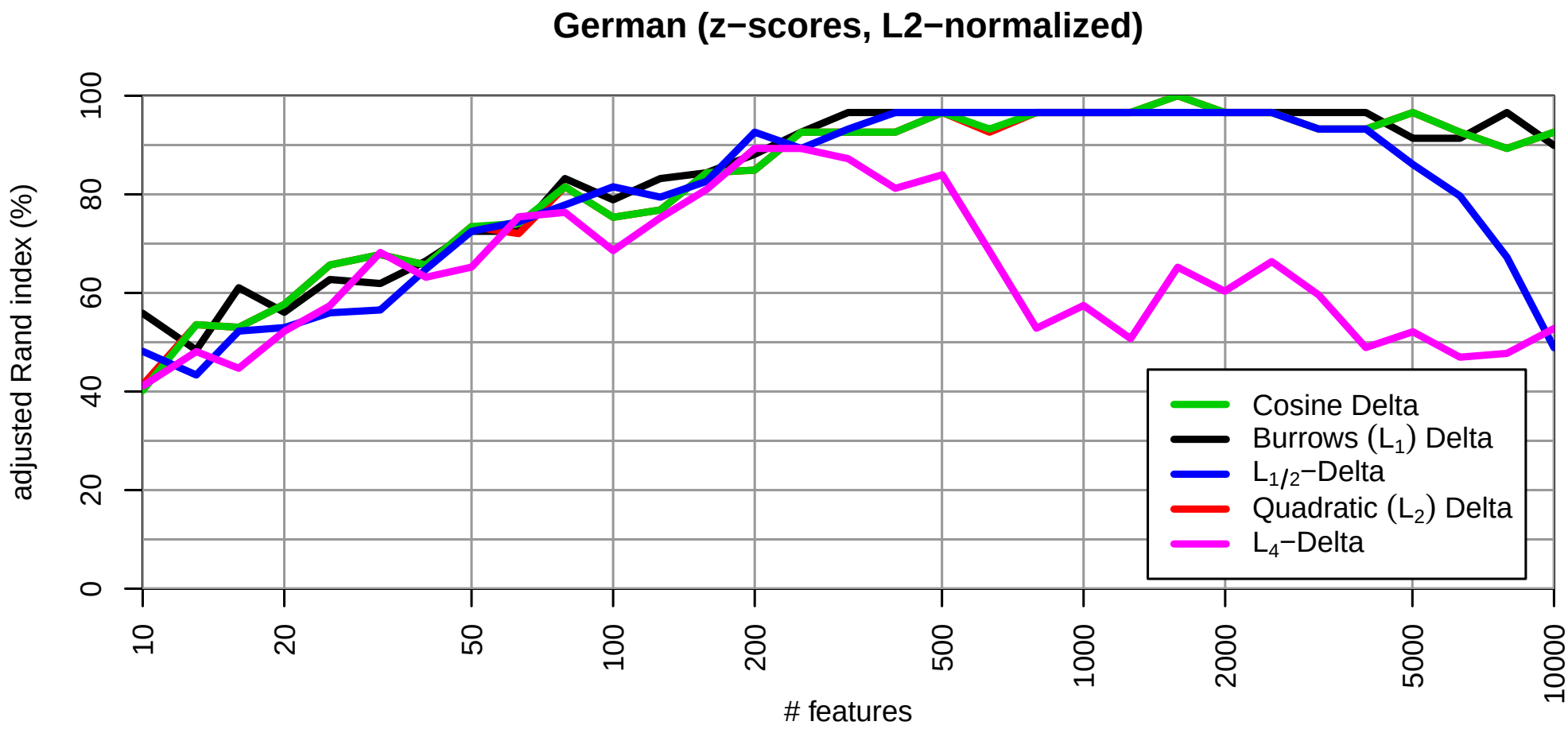
- Minkowsky Δ_p becomes less "robust" for increasing $p \in [0, \infty]$

The effect of vector length normalization



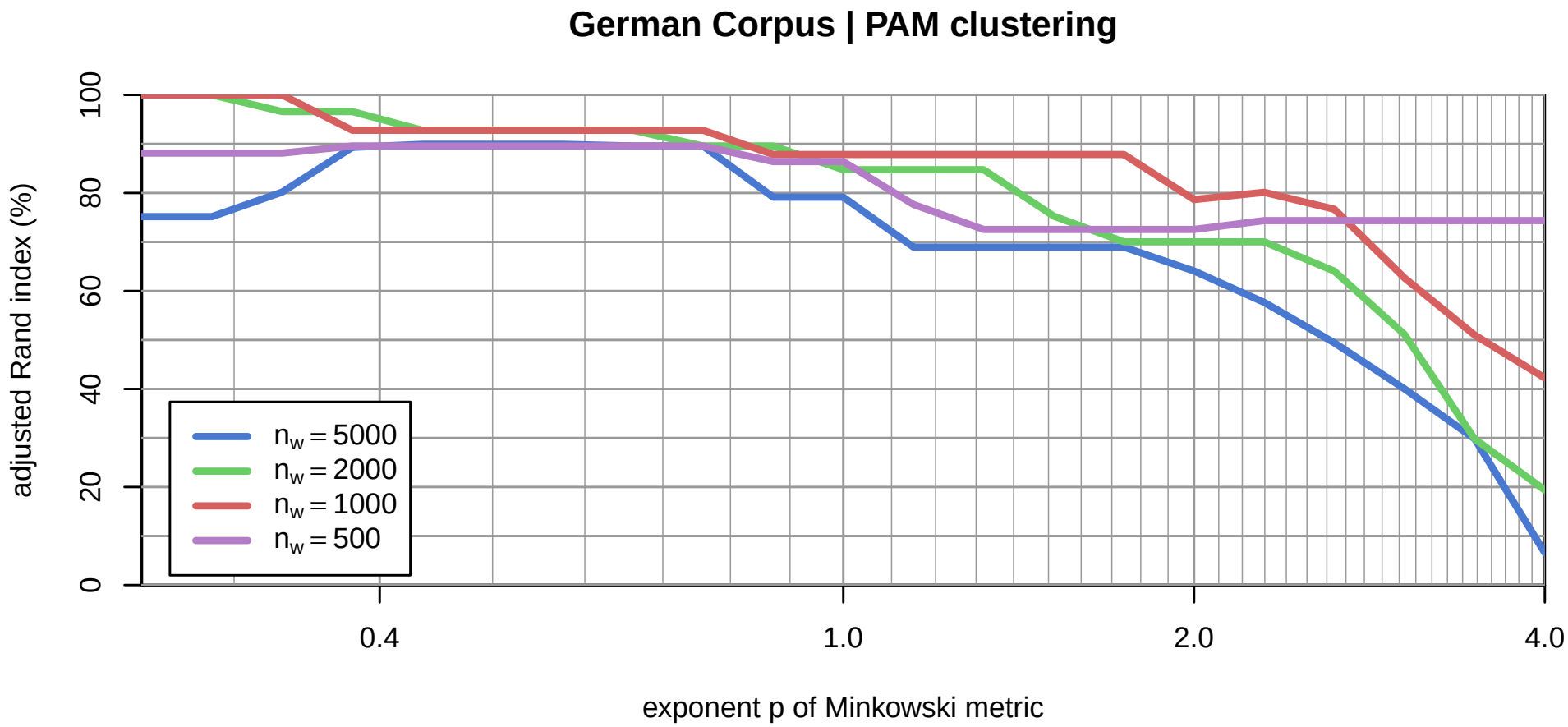
- Minkowsky Δ_p becomes less "robust" for increasing $p \in [0, \infty]$
- Vector length normalization $\|\mathbf{z}\|_p = 1$ (usually $p = 1, 2$)

The effect of vector length normalization



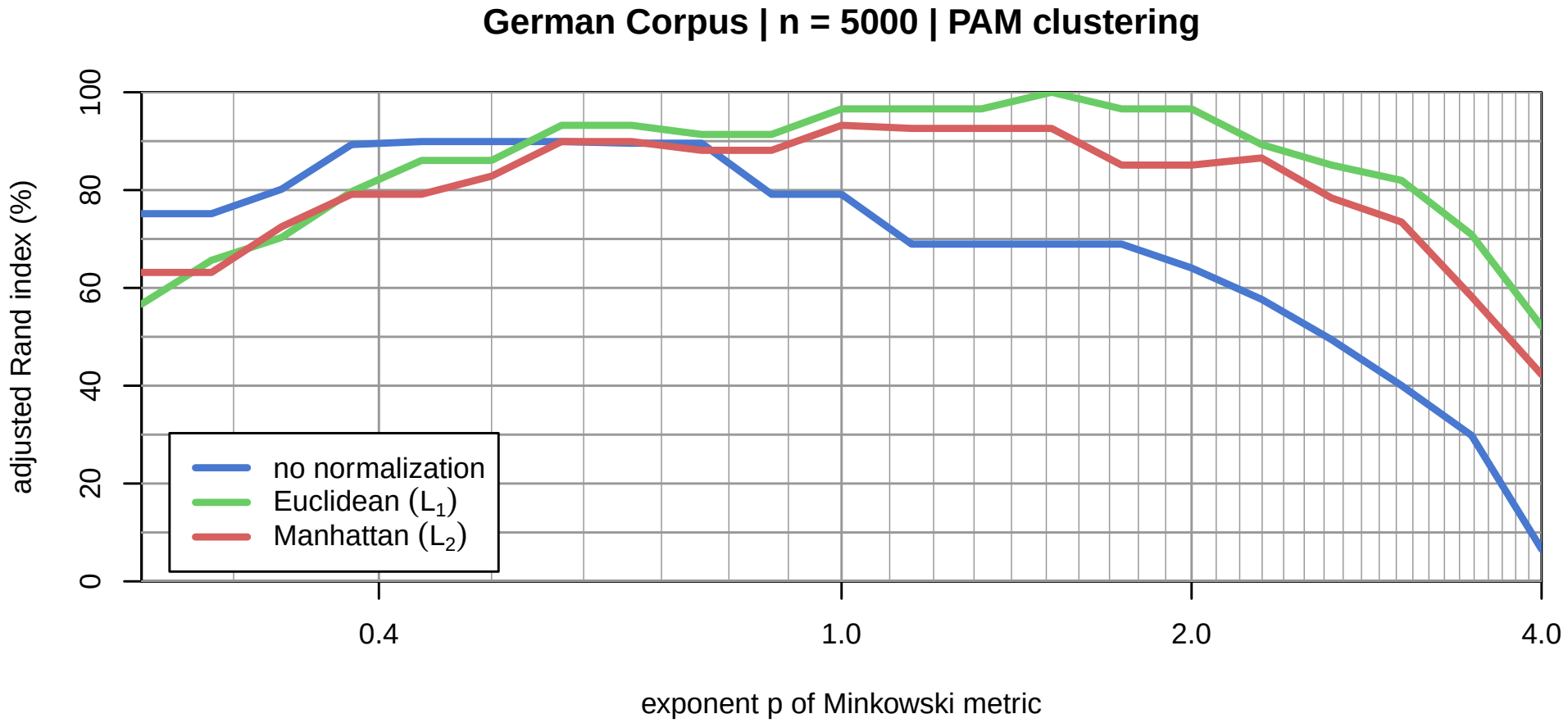
- Minkowsky Δ_p becomes less "robust" for increasing $p \in [0, \infty]$
- Vector length normalization $\|\mathbf{z}\|_p = 1$ (usually $p = 1, 2$)

The effect of vector length normalization



- Minkowsky Δ_p becomes less "robust" for increasing $p \in [0, \infty]$
- Vector length normalization $\|\mathbf{z}\|_p = 1$ (usually $p = 1, 2$)

The effect of vector length normalization

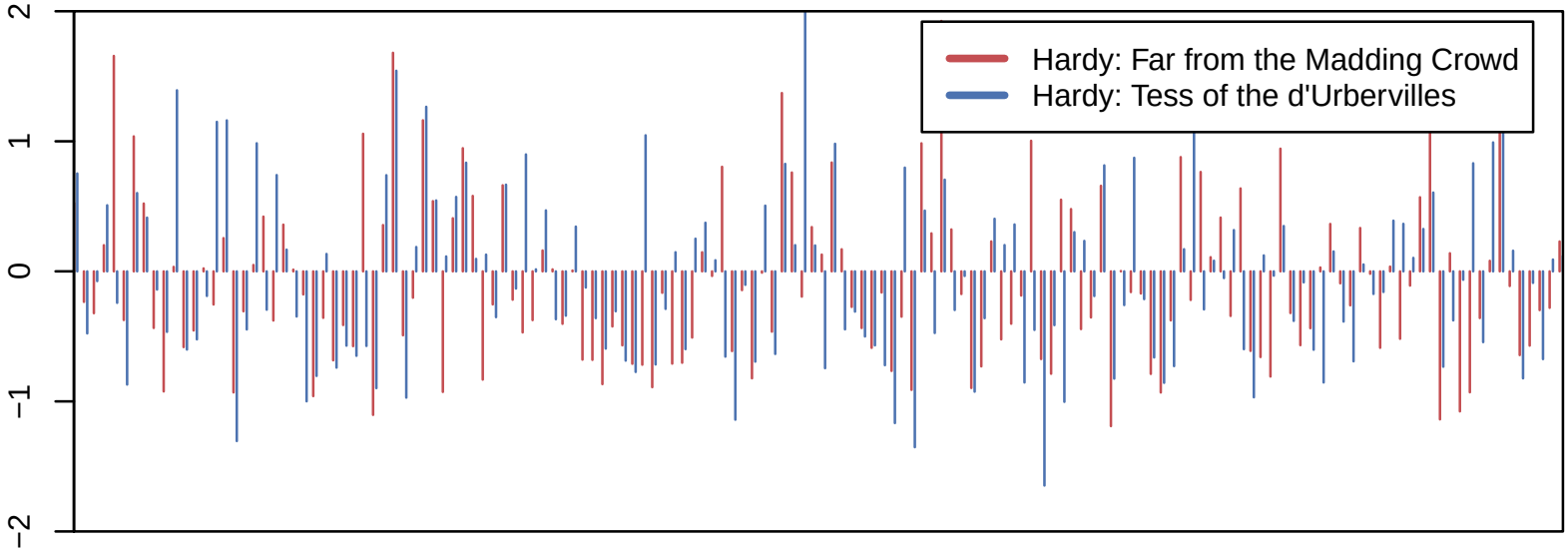


- Minkowsky Δ_p becomes less "robust" for increasing $p \in [0, \infty]$
- Vector length normalization $\|\mathbf{z}\|_p = 1$ (usually $p = 1, 2$)

Why is vector normalization so important?

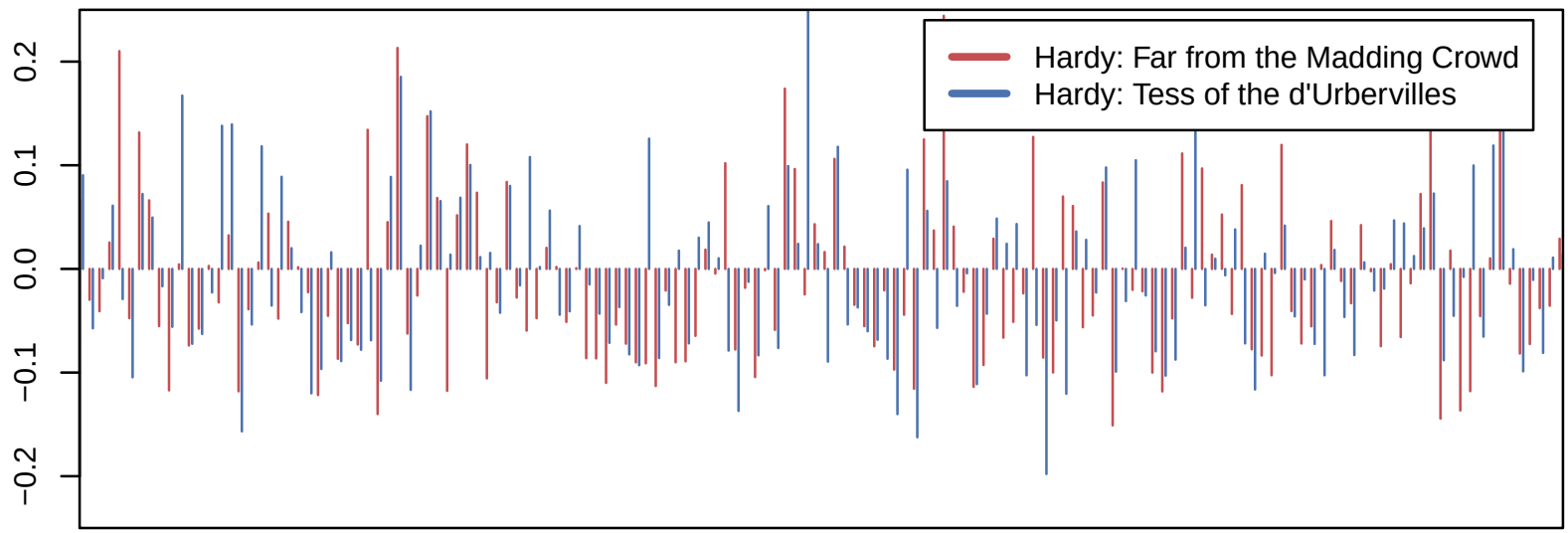
- Hypothesis 1: few "extreme" z-scores (outliers) distort metric
  - such outliers will often be characteristic for a particular text rather than an author
- Hypothesis 2: fingerprint of author = pattern of over- and underuse, regardless of magnitude of deviations ("key profile")

Approach: compare different transformations of z-scores



Approach: compare different transformations of z-scores

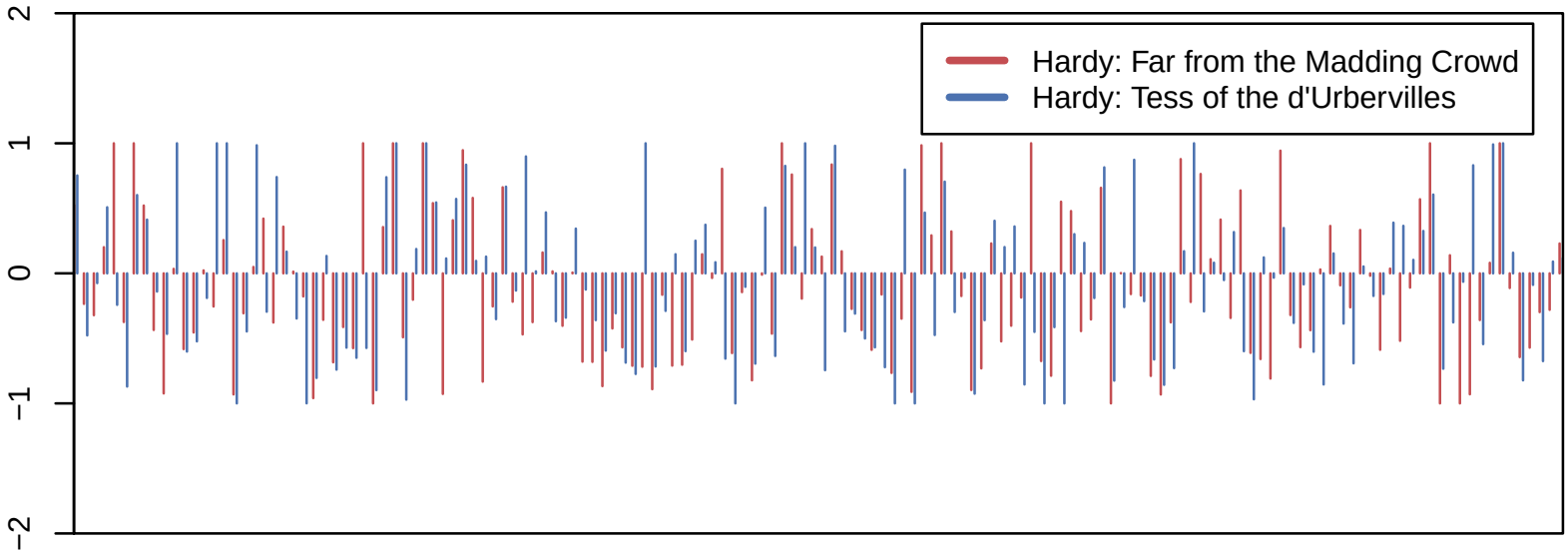
- normalisation of vector length



Approach: compare different transformations of z-scores

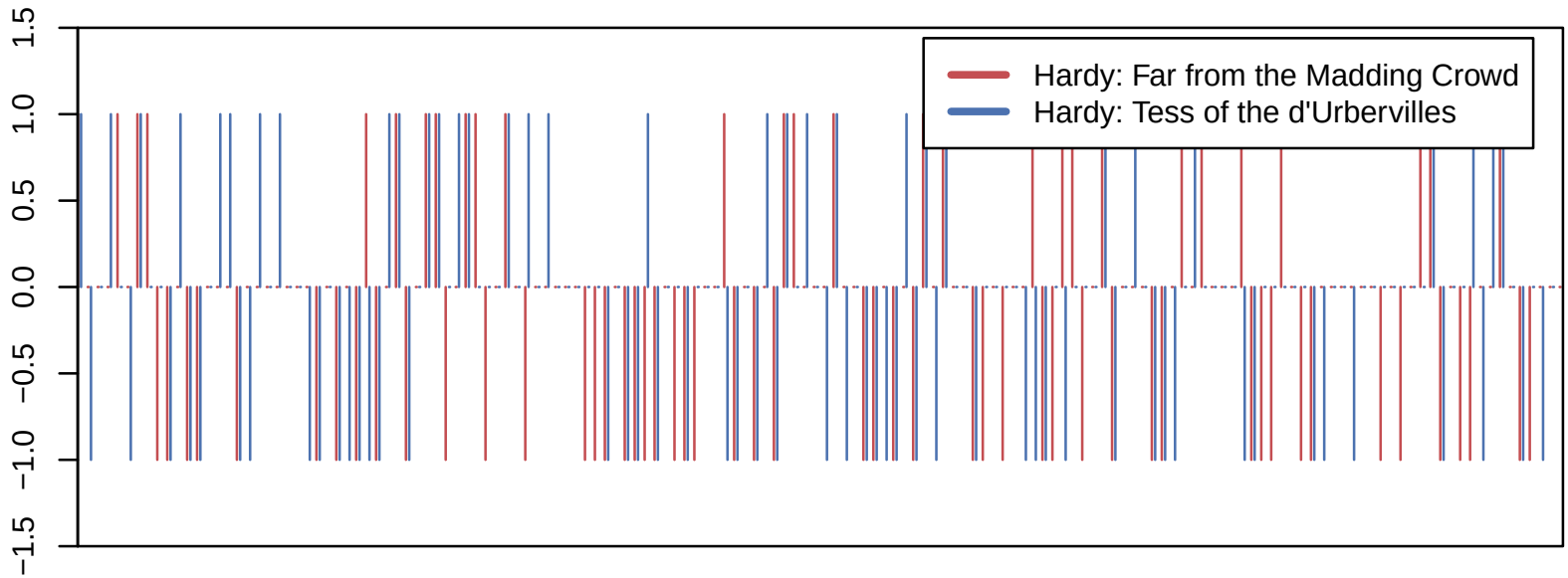
- normalisation of vector length
- clamping outliers (e.g. $|z| > 1$)

H1



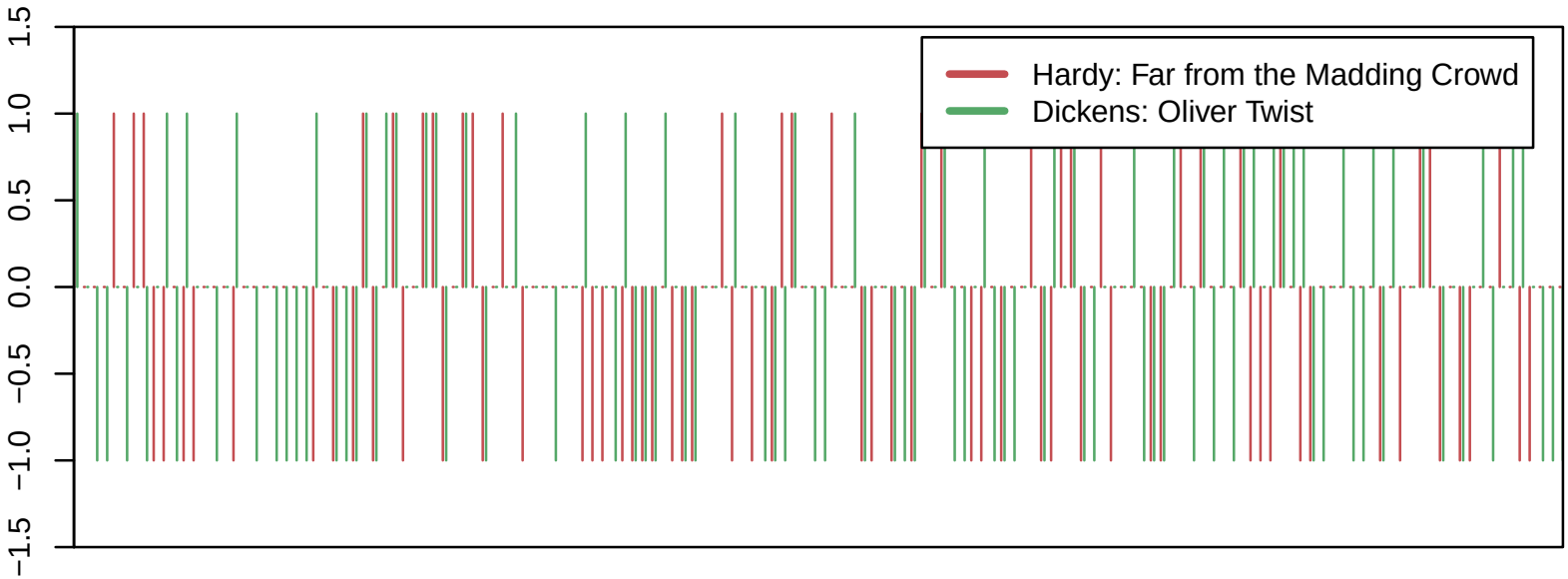
Approach: compare different transformations of z-scores

- normalisation of vector length
- clamping outliers (e.g. $|z| > 1$) H1
- ternarisation: -1 (underuse) / 0 (neutral) / $+1$ (overuse) H2



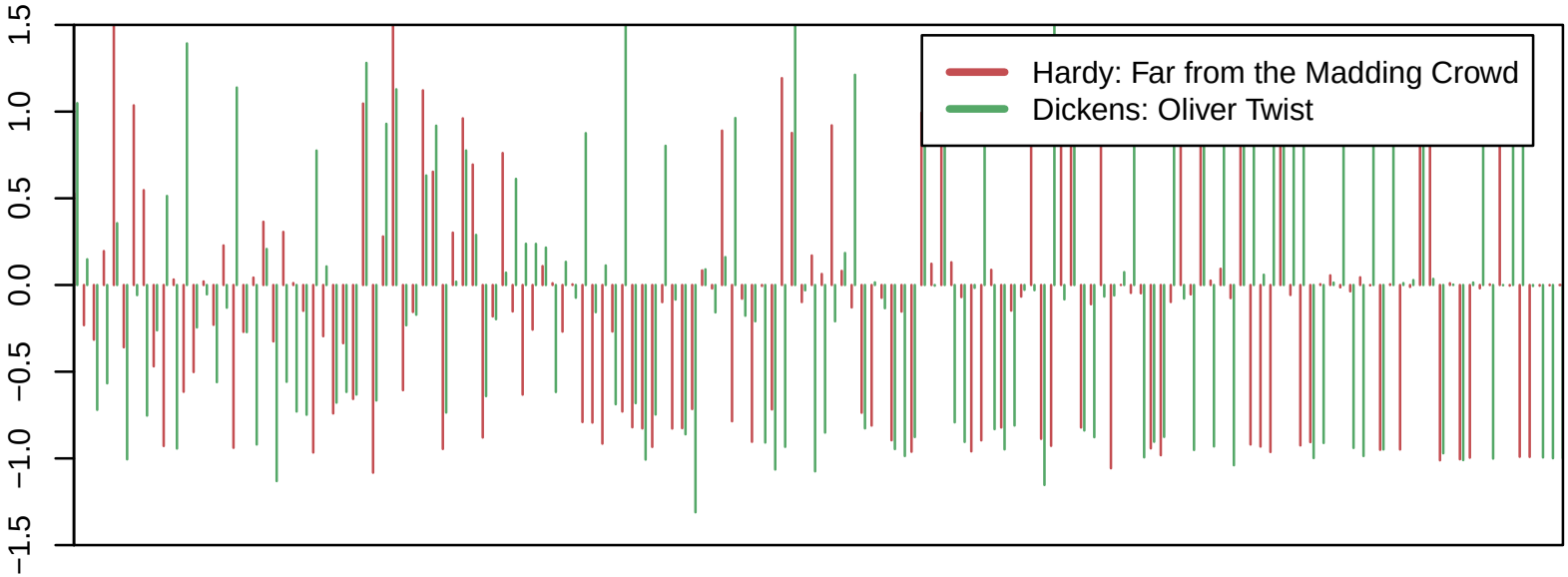
Approach: compare different transformations of z-scores

- normalisation of vector length
- clamping outliers (e.g. $|z| > 1$) H1
- ternarisation: -1 (underuse) / 0 (neutral) / $+1$ (overuse) H2

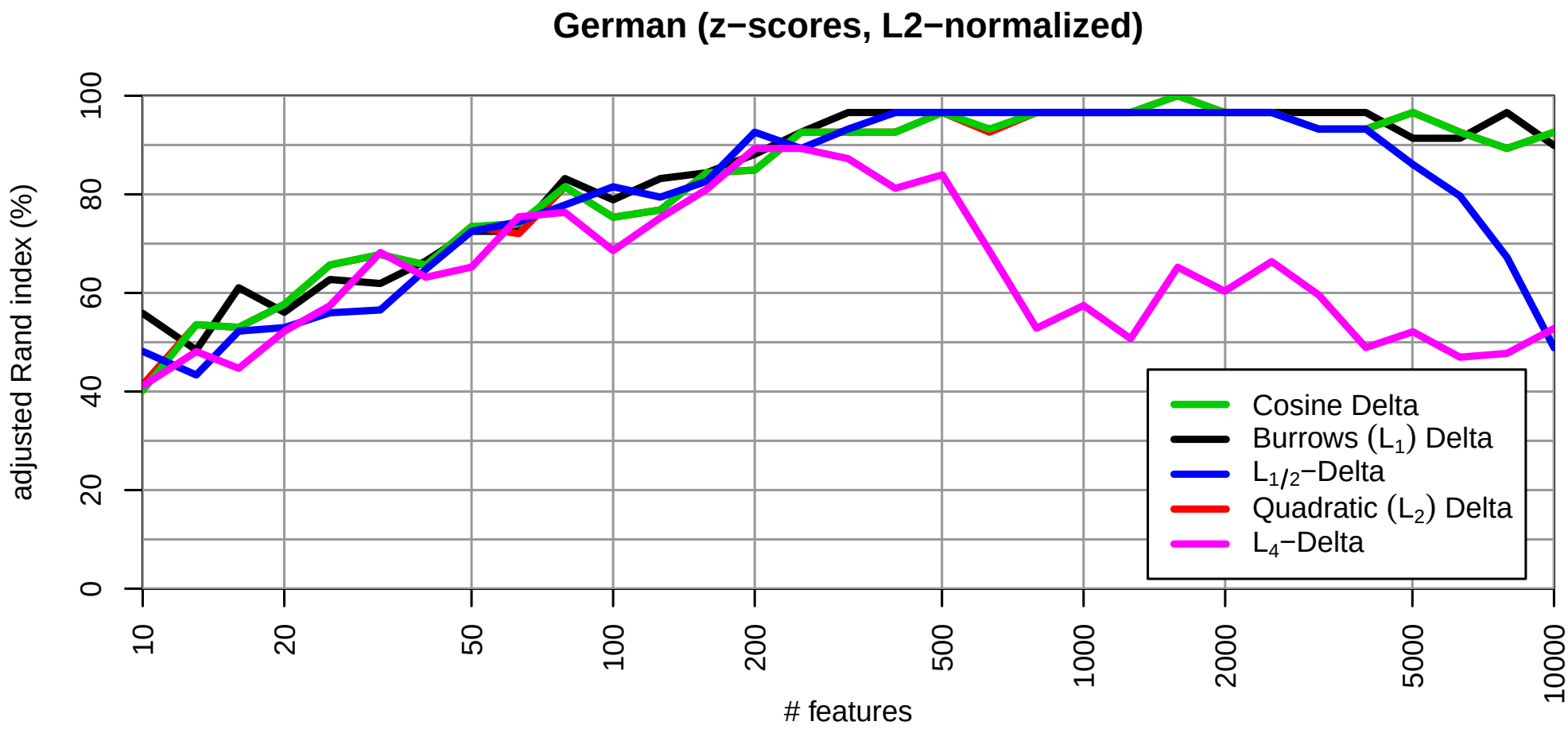


Approach: compare different transformations of z-scores

- normalisation of vector length
- clamping outliers (e.g. $|z| > 1$) H1
- ternarisation: -1 (underuse) / 0 (neutral) / $+1$ (overuse) H2
- optional: ternarization only for lower-frequency words

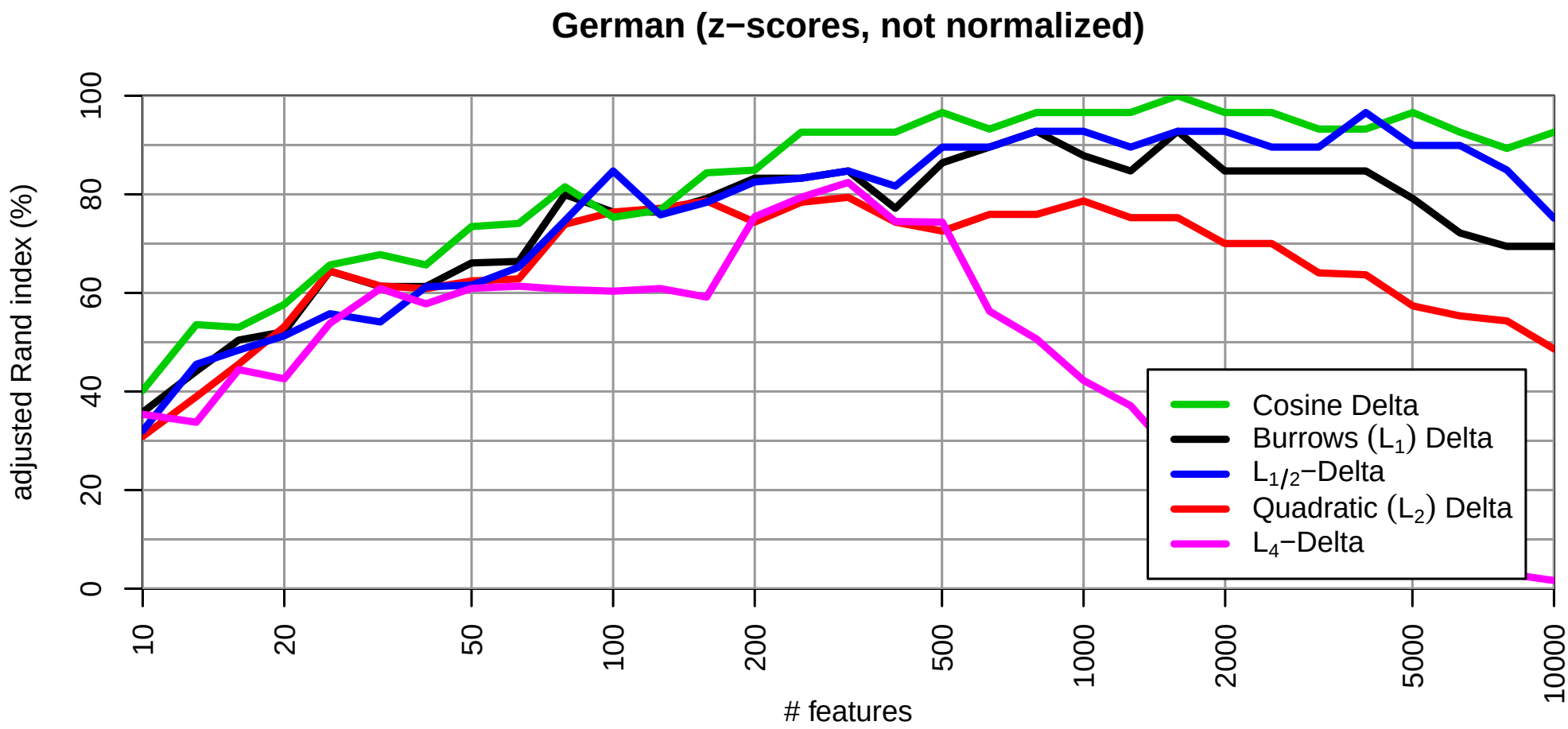


Parameter: Transformation of z-scores

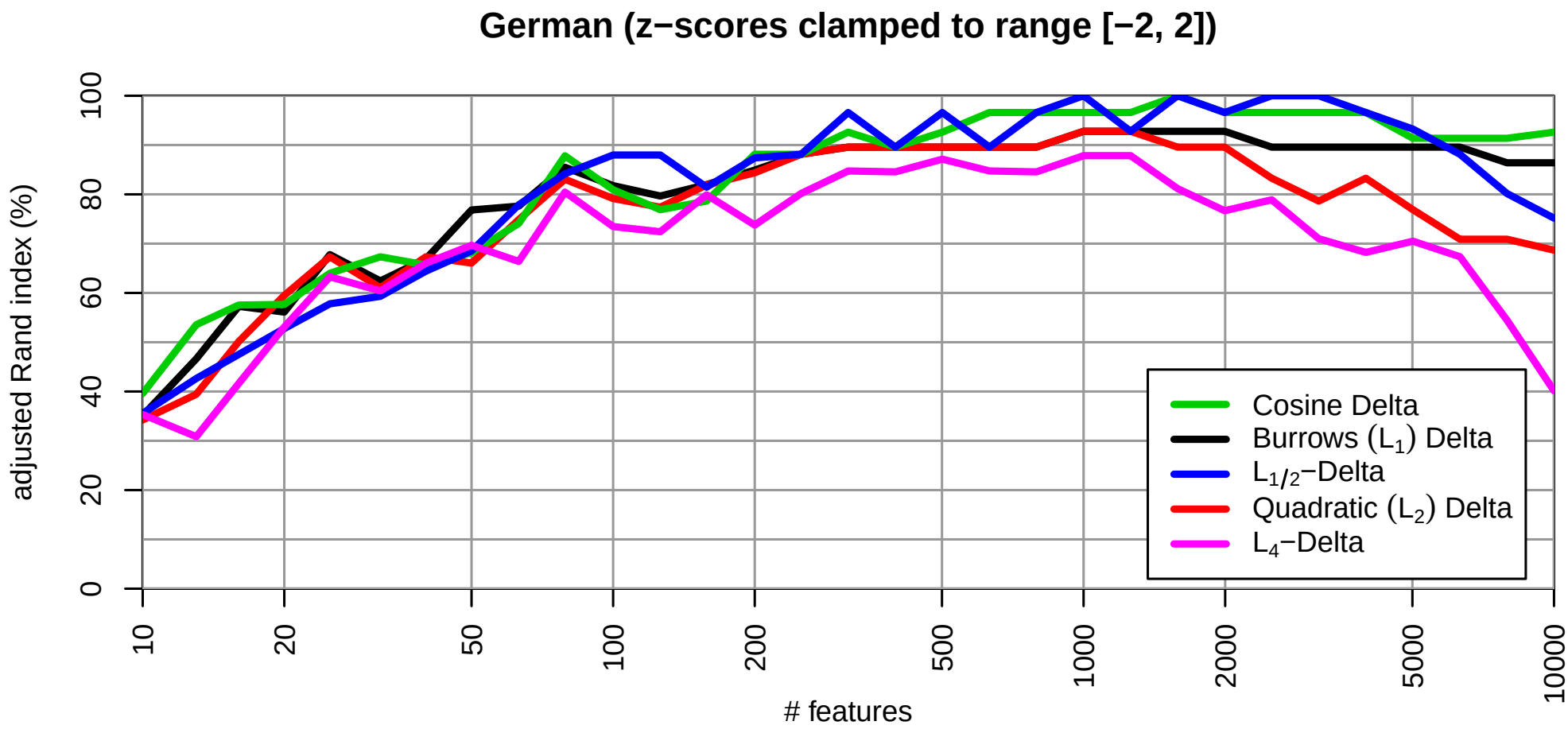


normalisation of vector lengths

Parameter: Transformation of z-scores

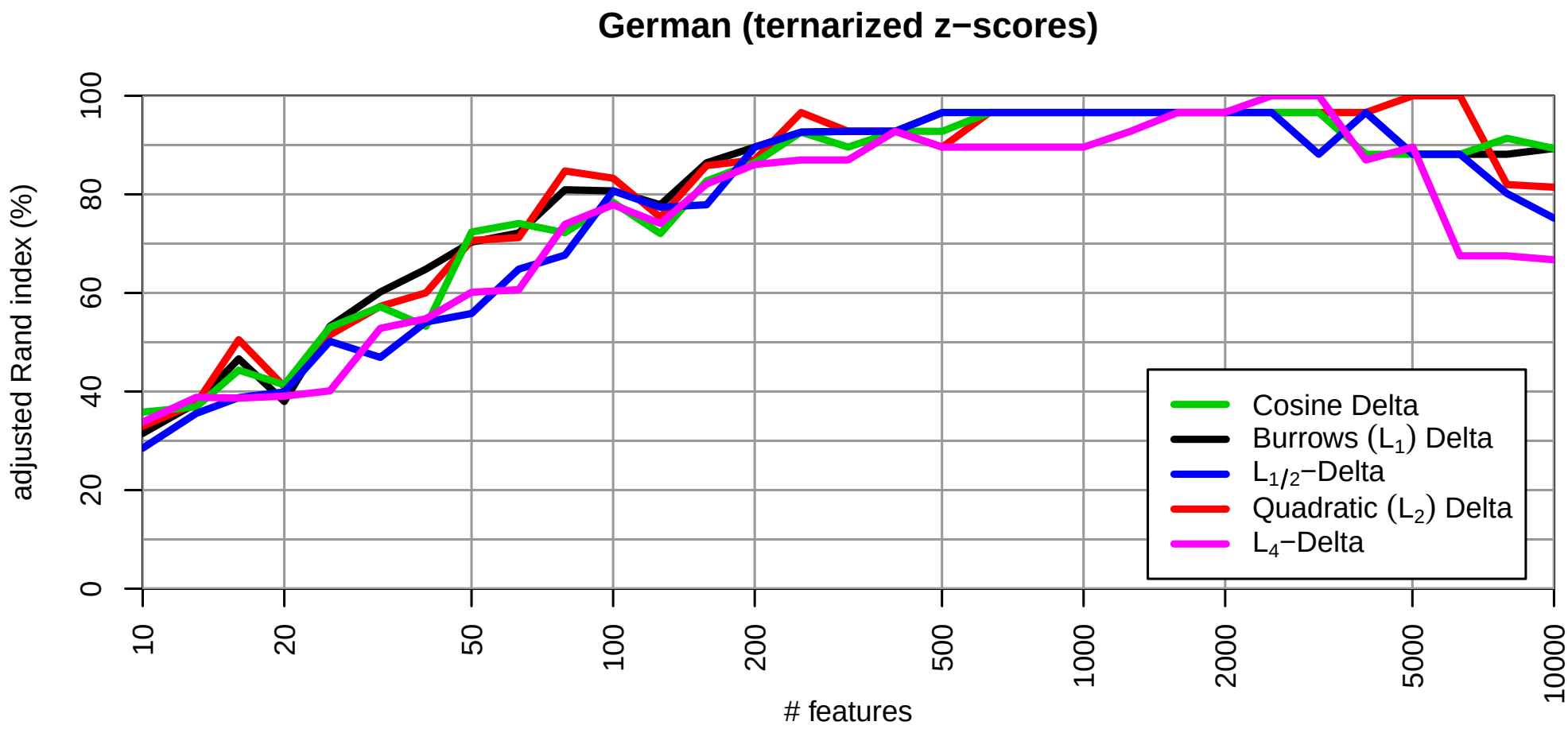


Parameter: Transformation of z-scores



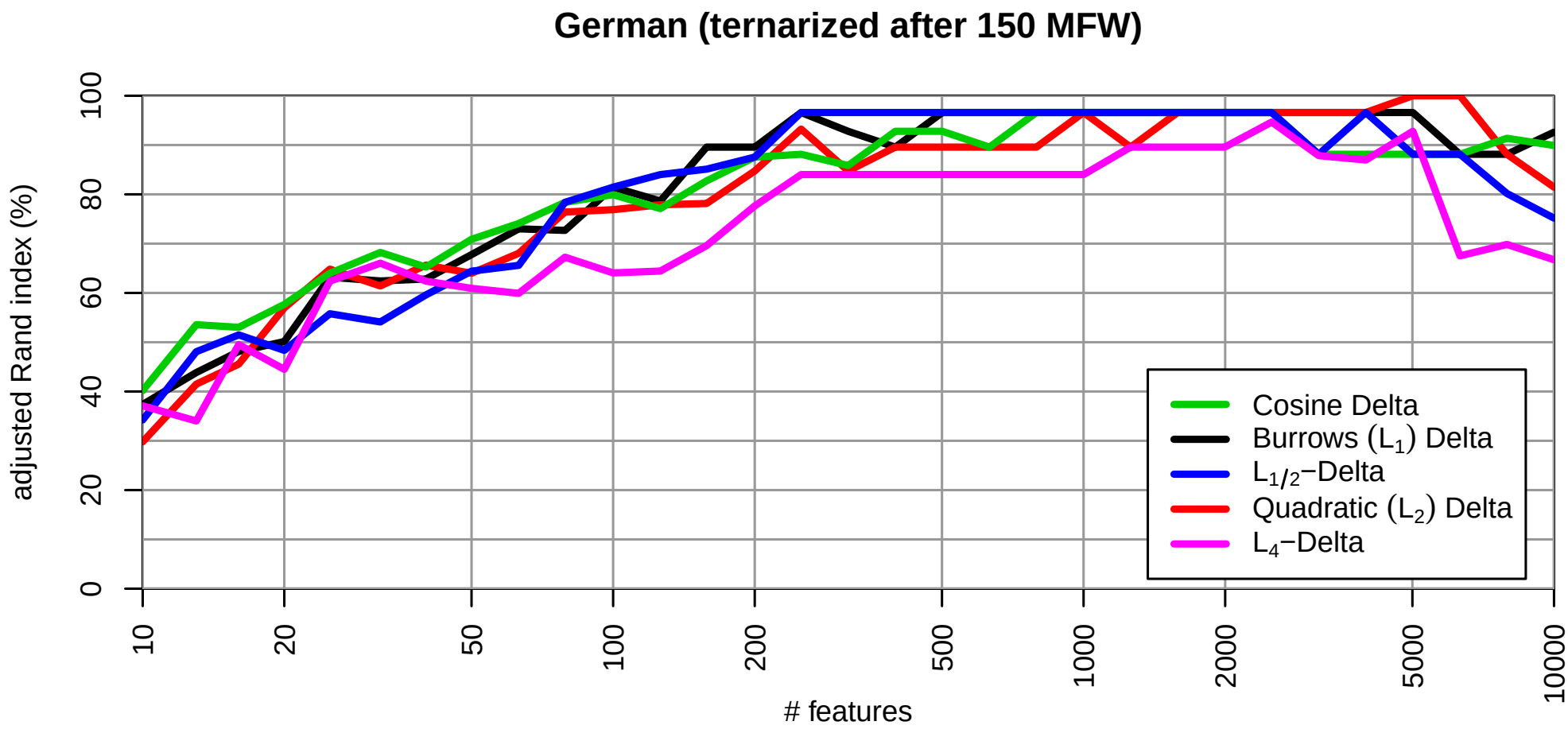
clamping outliers ($|z| > 2$) H1

Parameter: Transformation of z-scores



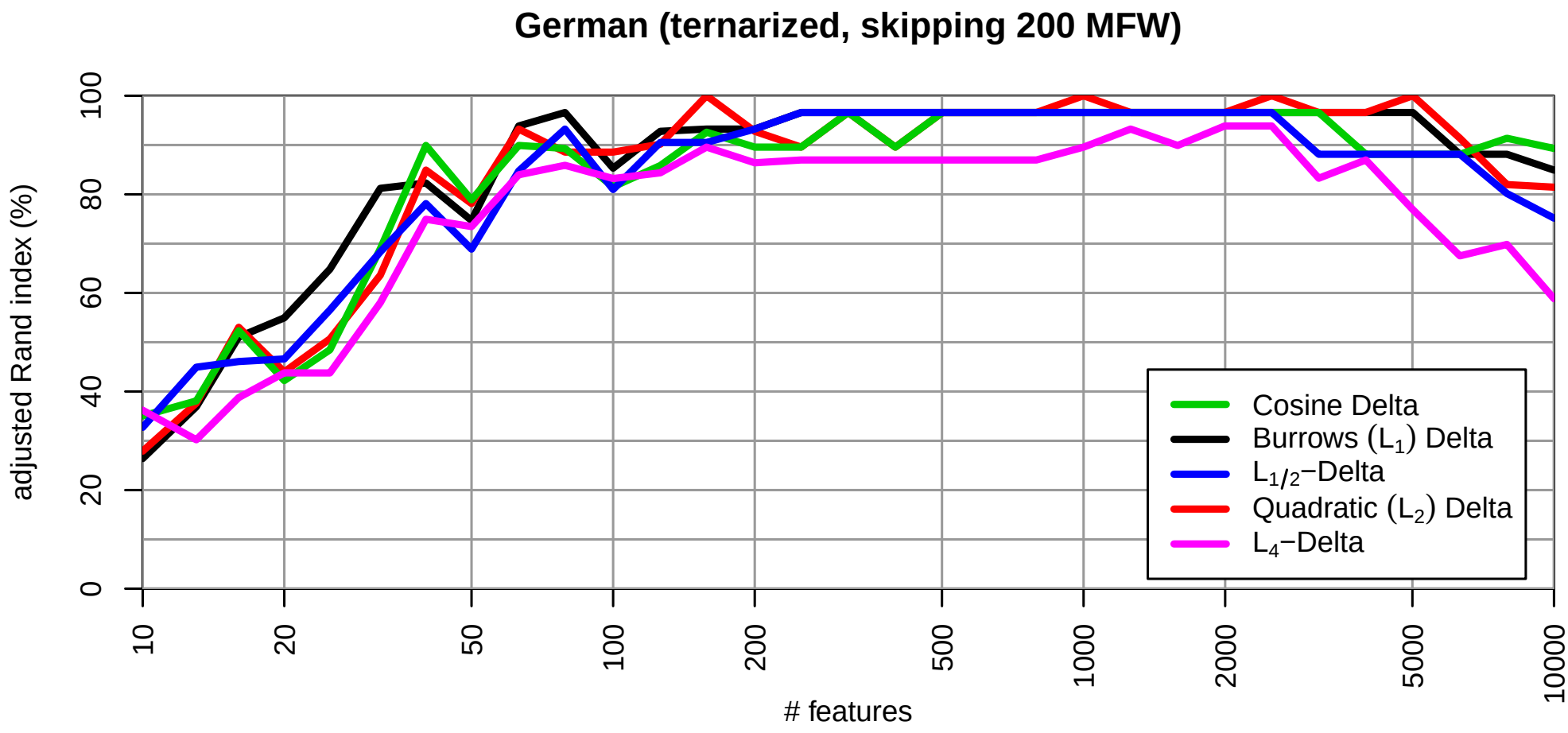
ternarisation (equal categories) H2

Parameter: Transformation of z-scores



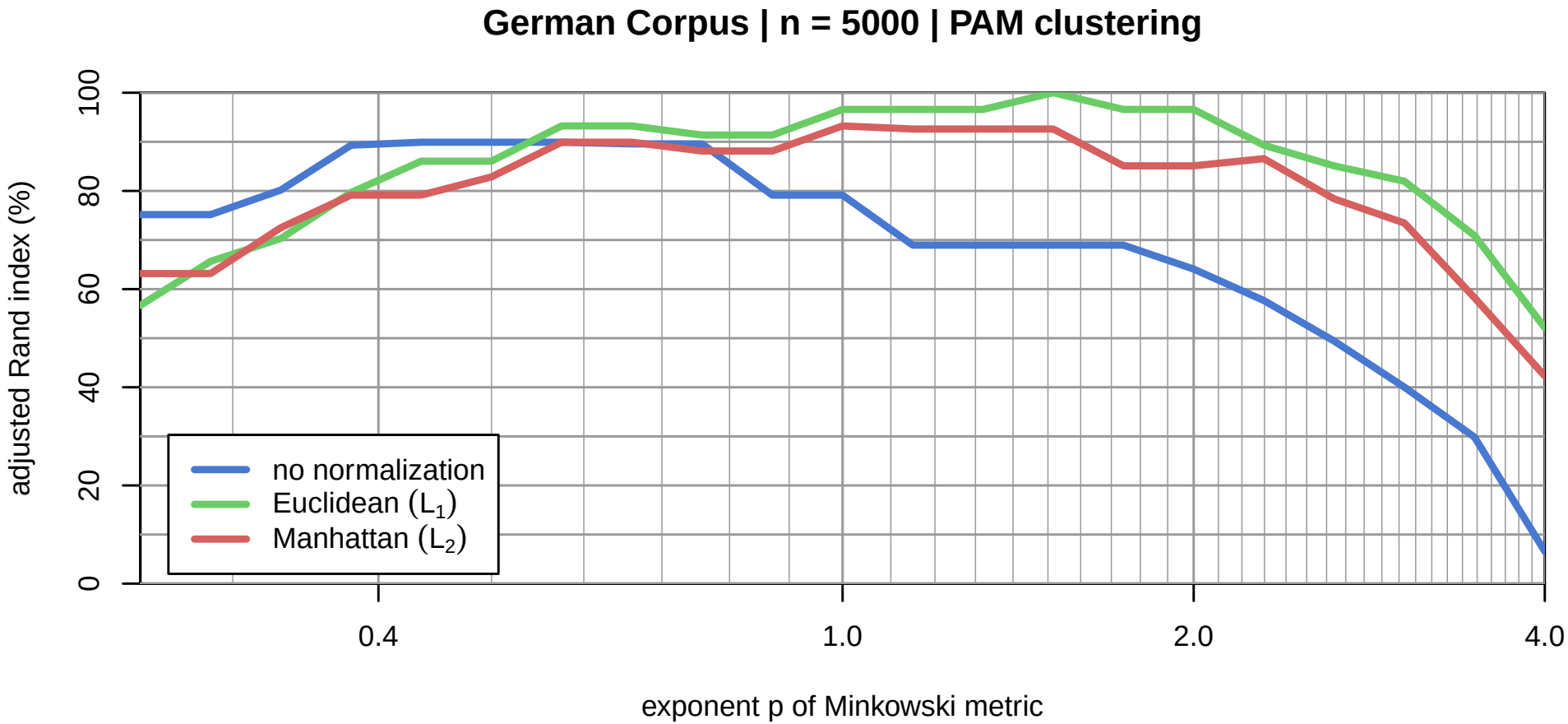
ternarisation (equal categories, after 150 MFW)

Parameter: Transformation of z-scores

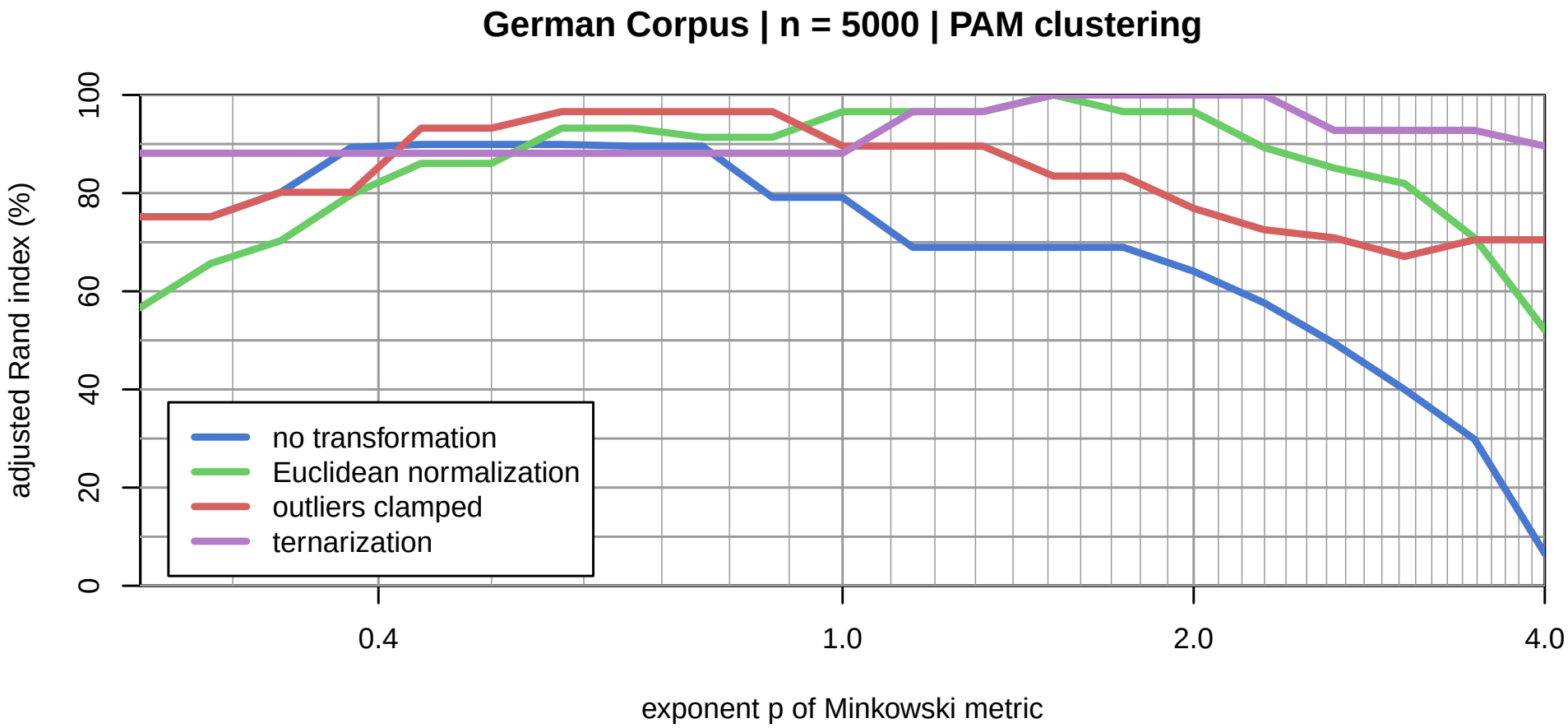


ternarisation (equal categories, *skipping* first 200 MFW)

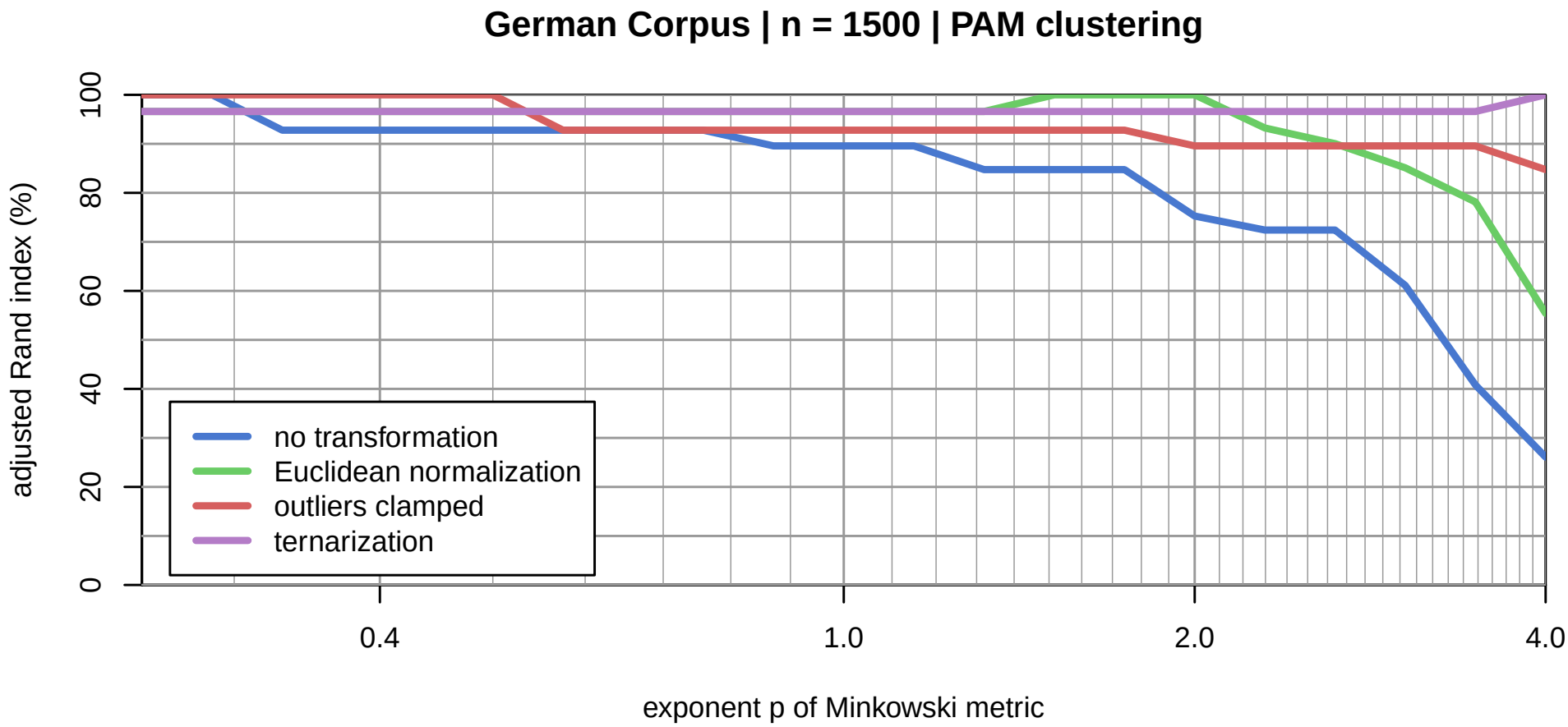
Transformation of z-scores and Δ_p



Transformation of z-scores and Δ_p



Transformation of z-scores and Δ_p



What is in the fingerprint?

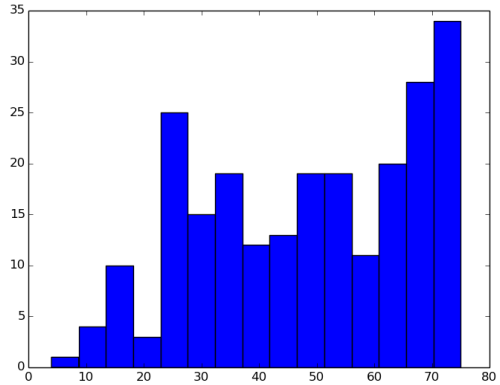
- Greedy algorithm for selection of an optimal set of features: recursive feature elimination

- Procedure:
  - train linear support vector machine (SVM)
  - based on $[0, 1]$-scaled relative frequencies (not on z-scores)
  - discard k features with lowest SVM weights

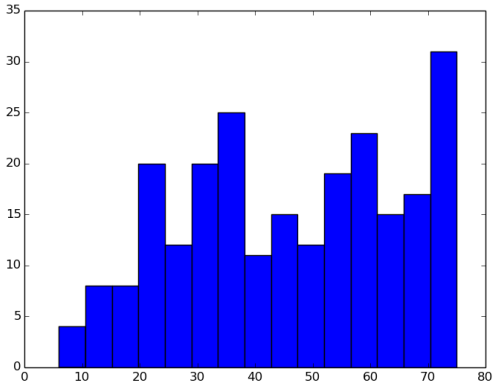
- Iterative reduction of feature set
  1. all recurrent words ($df > 1$)
  2. down to $n_w = 50{,}000$ ($k = 10{,}000$)
  3. down to $n_w = 5{,}000$ ($k = 1{,}000$)
  4. down to $n_w = 500$ ($k = 100$)
  5. find minimal feature set by cross-validation ($k = 1$)

What is in the fingerprint?

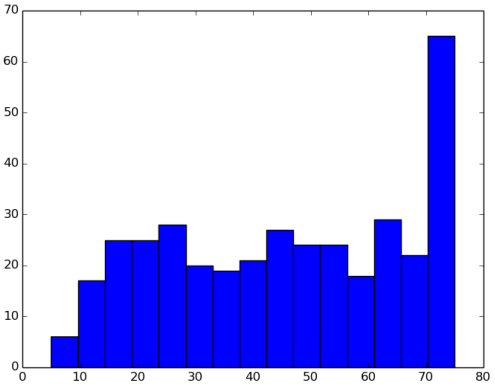
Document frequencies (df) of selected features



English ($n_w = 233$)



German ($n_w = 240$)

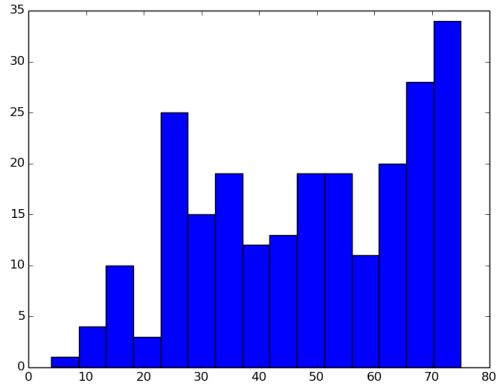


French ($n_w = 370$)

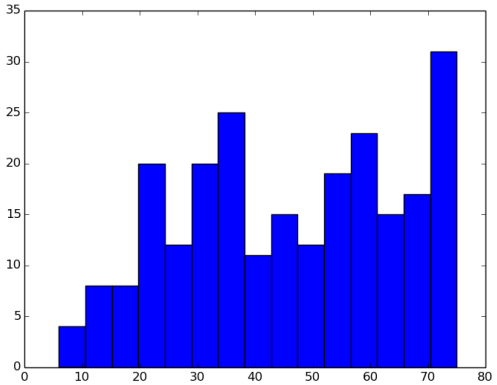
- Many function words, but also content words (➜ overtraining?)

What is in the fingerprint?

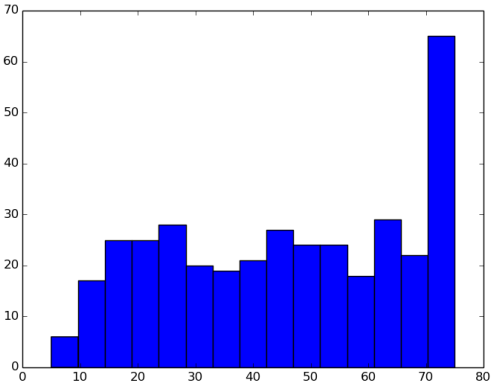
Document frequencies (df) of selected features



English ($n_w = 233$)



German ($n_w = 240$)

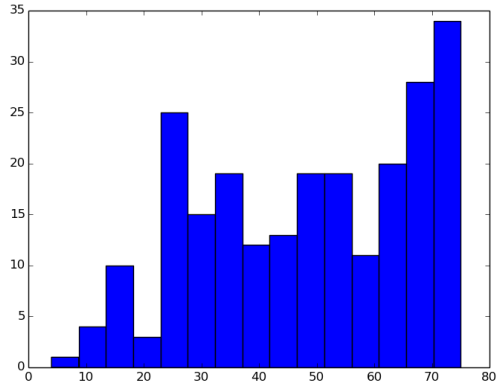


French ($n_w = 370$)

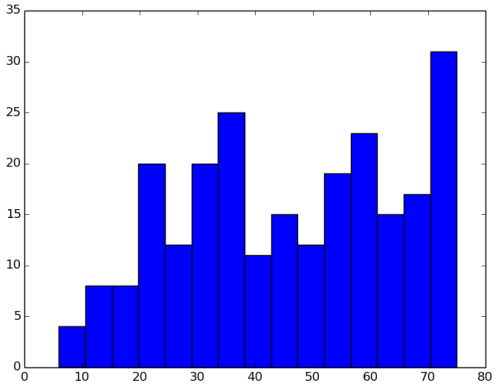
- Many function words, but also content words (➜ overtraining?)
- Some text artefacts: Roman numerals (XL, XXXVII) in novels with many chapters, graphemic variation (e.g. DE *gibt* / *giebt*)

What is in the fingerprint?

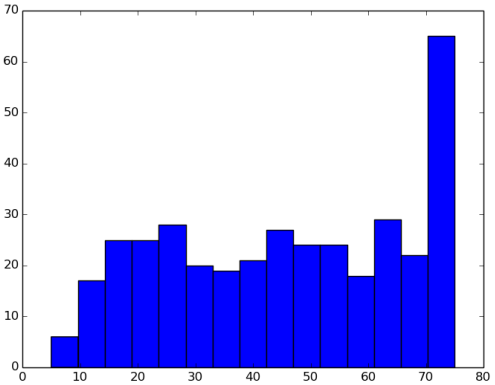
Document frequencies (df) of selected features



English ($n_w = 233$)



German ($n_w = 240$)



French ($n_w = 370$)

- Many function words, but also content words (➜ overtraining?)
- Some text artefacts: Roman numerals (XL, XXXVII) in novels with many chapters, graphemic variation (e.g. DE *gibt* / *giebt*)
- Key words for English novels: *with*, *so*, *t*, *But*, *And*, *upon*, *don*, *head*, *Then*, *looking*, *almost*, *indeed*, *nor*, *London*, *feel*, *cannot*, ..., *XXXVII* ($df = 34$), *XLI* ($df = 29$), *XLIII* ($df = 26$), *hereabout* ($df = 11$), *vilest* ($df = 15$), *contours* ($df = 9$), *Ecod* ($df = 4$)

Outline

Literary authorship attribution

Burrows's Delta

Empirical evaluation

Understanding Delta

Statistical significance

But is any of this significant?

- Statistical significance
  - If the experiment is repeated with a new data set, will method A again perform better than method B?

But is any of this significant?

- Statistical significance
  - If the experiment is repeated with a new data set, will method A again perform better than method B?
- Sampling variation
  - How much statistical uncertainty in the evaluation scores?
  - often in terms of confidence interval

But is any of this significant?

- Statistical significance
  - If the experiment is repeated with a new data set, will method A again perform better than method B?
- Sampling variation
  - How much statistical uncertainty in the evaluation scores?
  - often in terms of confidence interval
- Rarely addressed in literary authorship attribution

Empirical evidence

(Proisl *et al.* 2018)

- Simulate repeated experiments
  - by sampling texts from a large collection (e.g. Gutenberg)
  - 973 German novels from 131 different authors (1789–1914)
  - all text shortened to first 30k tokens

Empirical evidence

(Proisl *et al.* 2018)

- Simulate repeated experiments
  - by sampling texts from a large collection (e.g. Gutenberg)
  - 973 German novels from 131 different authors (1789–1914)
  - all text shortened to first 30k tokens
- Setting A
  - sample different novels from the same 25 authors
  - 5,000 random samples of 25×3 novels

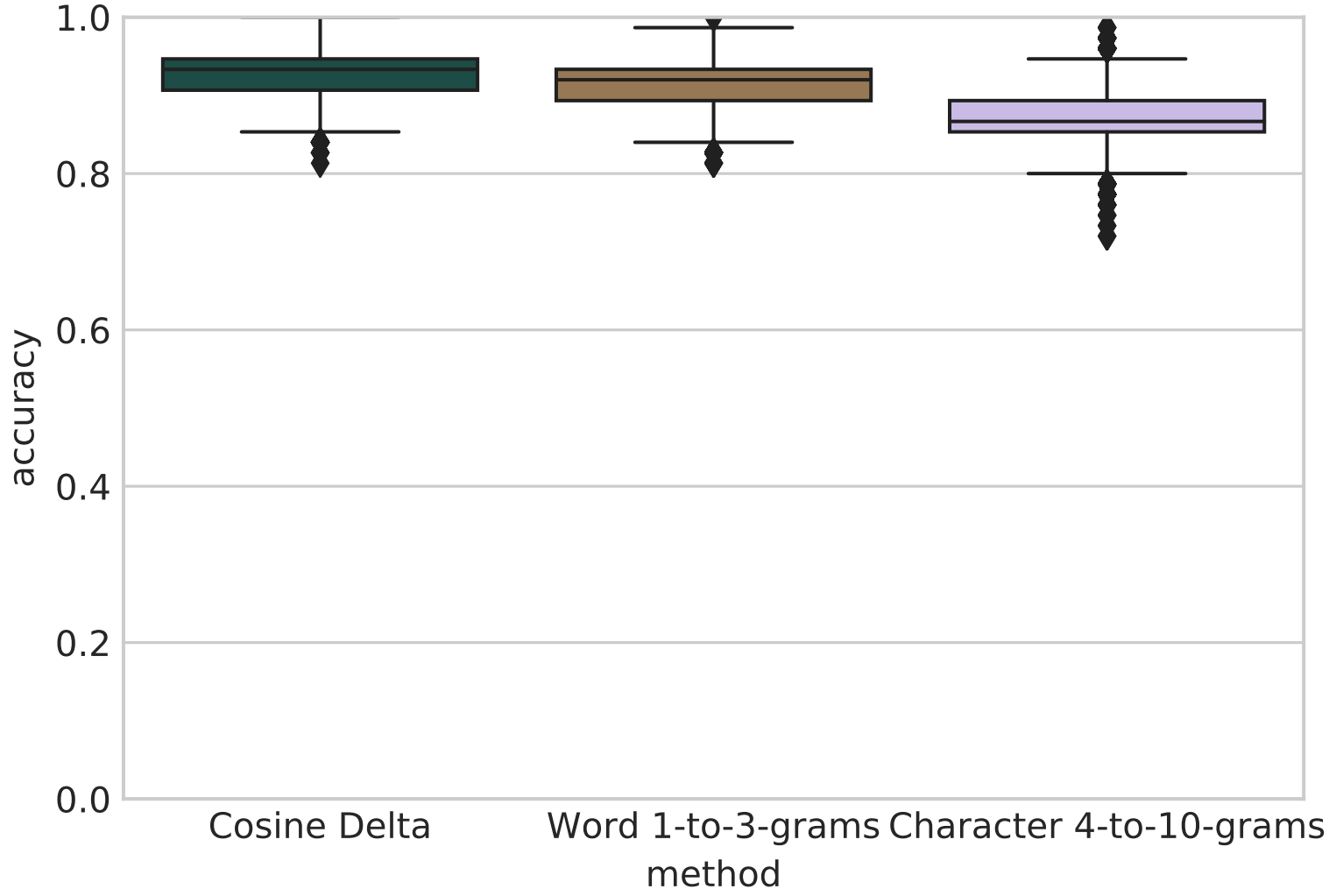
Empirical evidence

(Proisl *et al.* 2018)

- Simulate repeated experiments
  - by sampling texts from a large collection (e.g. Gutenberg)
  - 973 German novels from 131 different authors (1789–1914)
  - all text shortened to first 30k tokens
- Setting A
  - sample different novels from the same 25 authors
  - 5,000 random samples of 25 $\times$ 3 novels
- Setting B
  - sample different sets of authors
  - 5,000 random samples of 25 authors $\times$ 3 novels

Results: Sampling variation

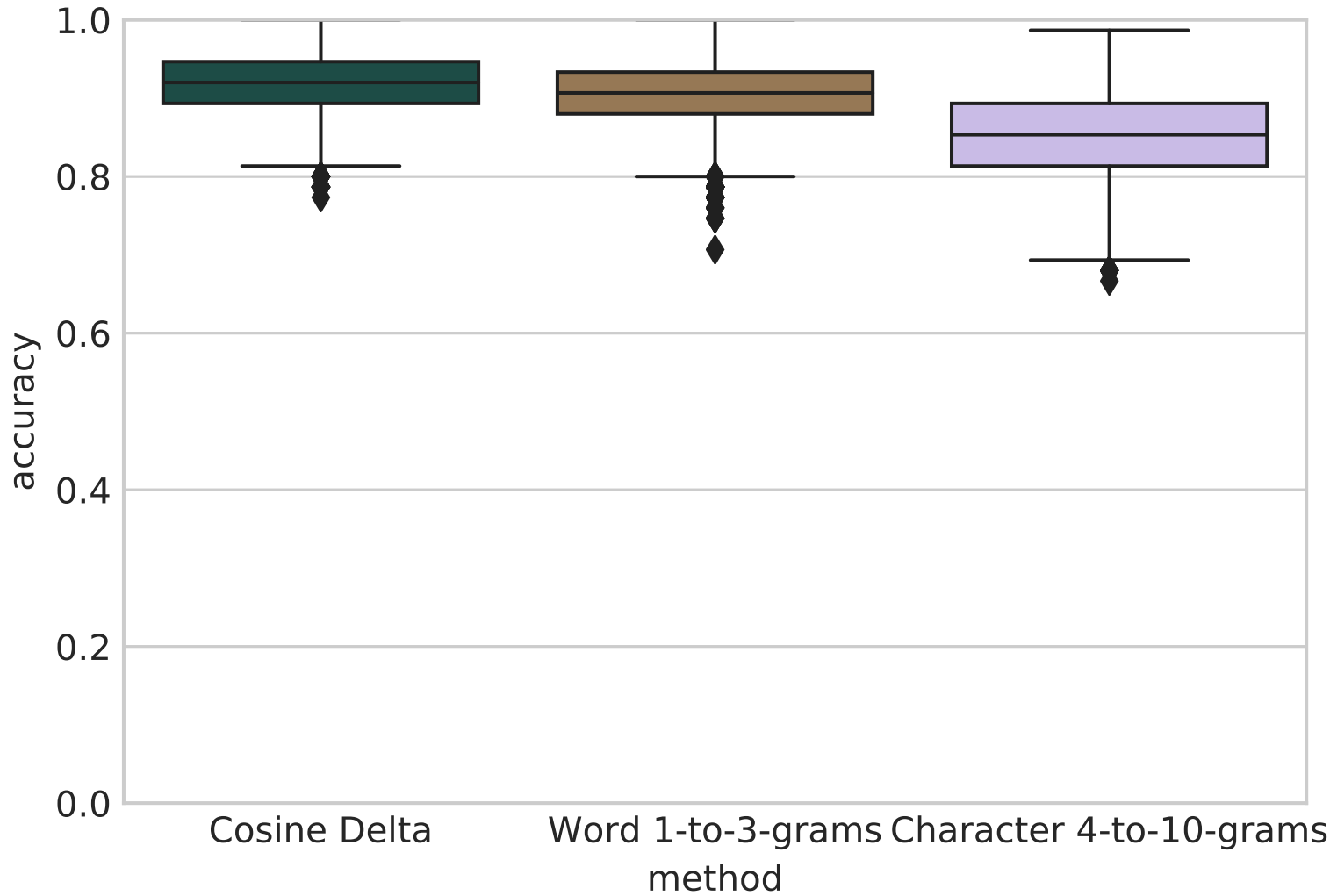
(Proisl *et al.* 2018)



Setting A: Sampling different novels from the same authors

Results: Sampling variation

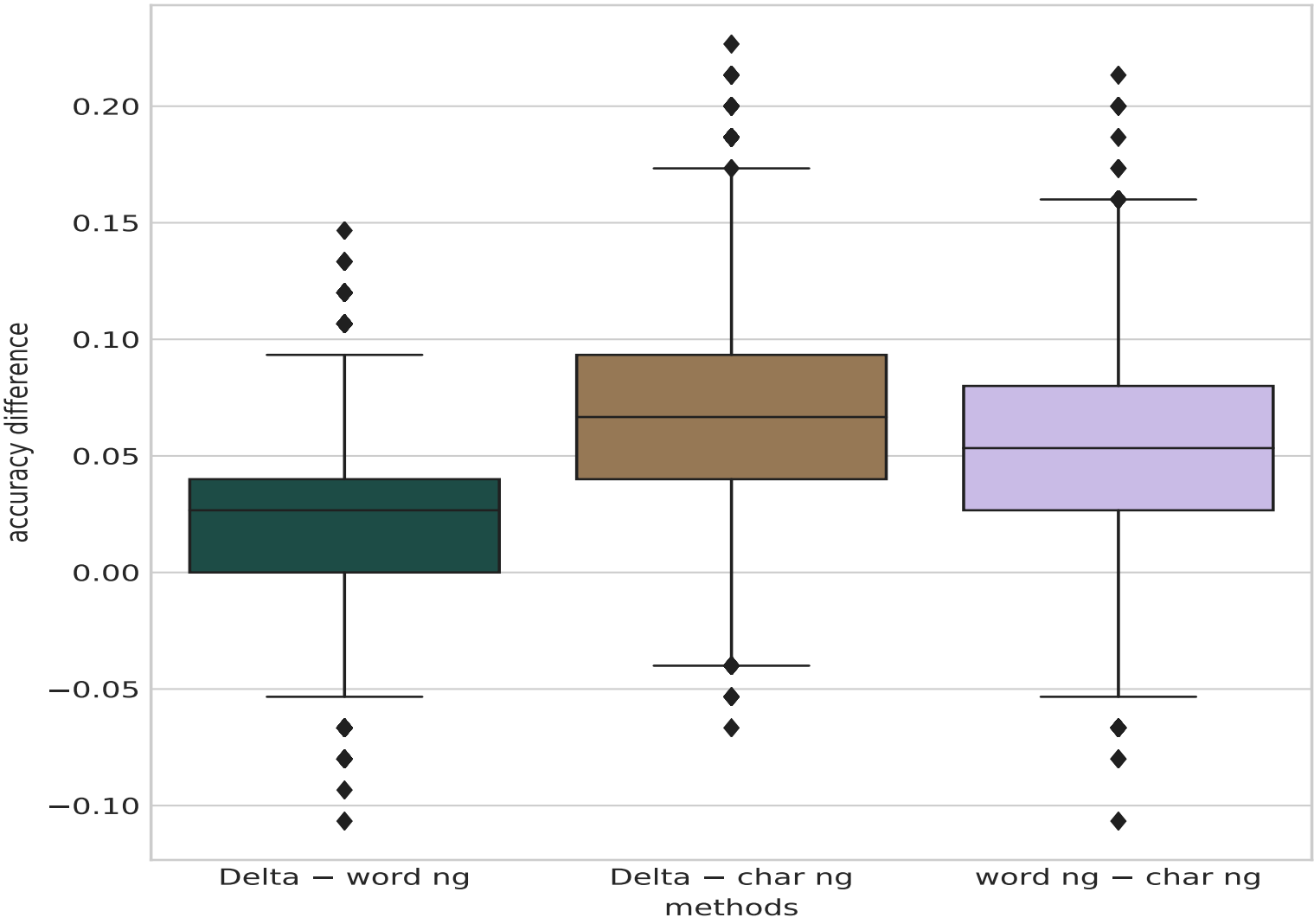
(Proisl *et al.* 2018)



Setting B: Sampling different sets of authors

Results: Statistical significance

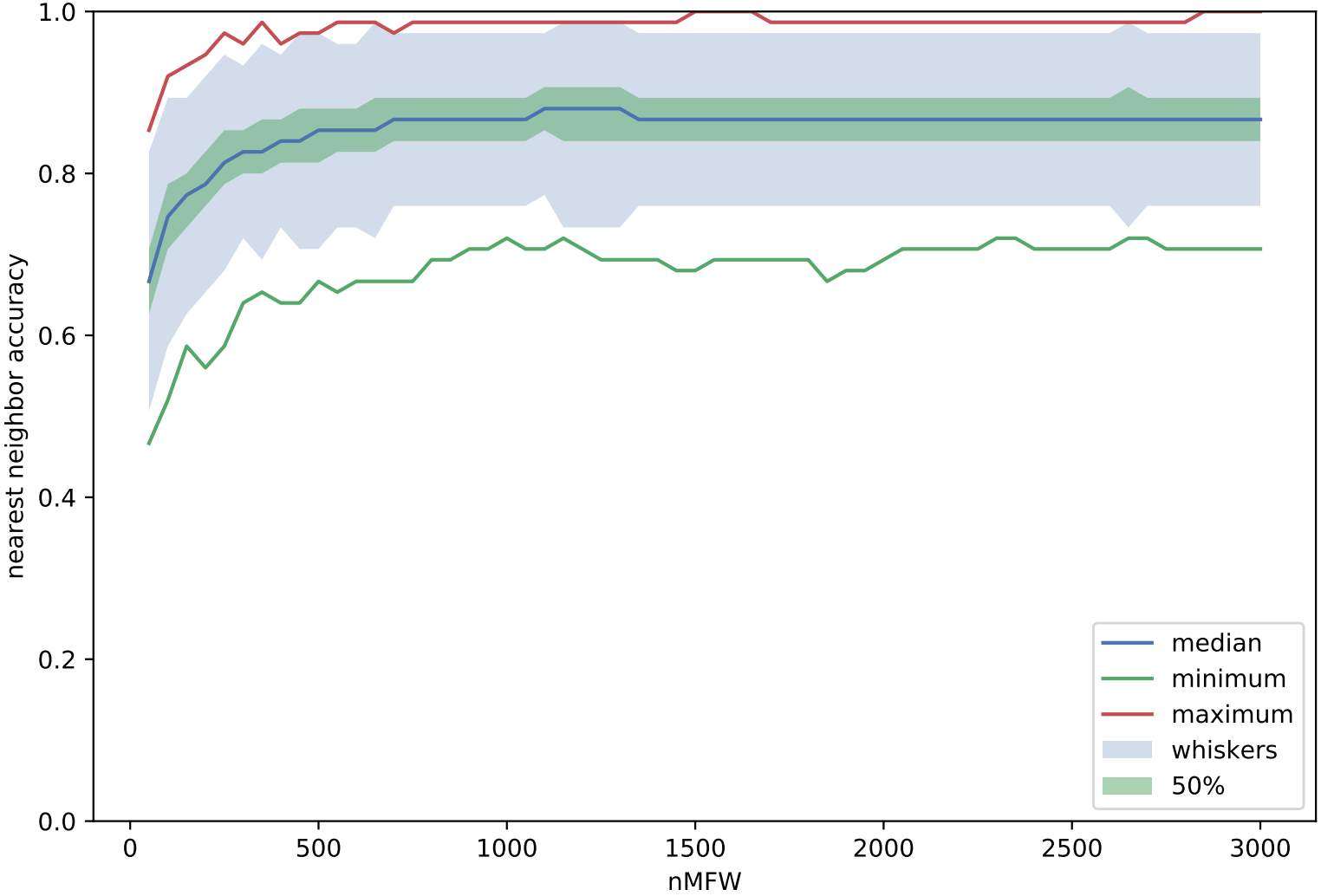
(Proisl *et al.* 2018)



Setting B: Sampling different sets of authors – paired tests

Results: Implications

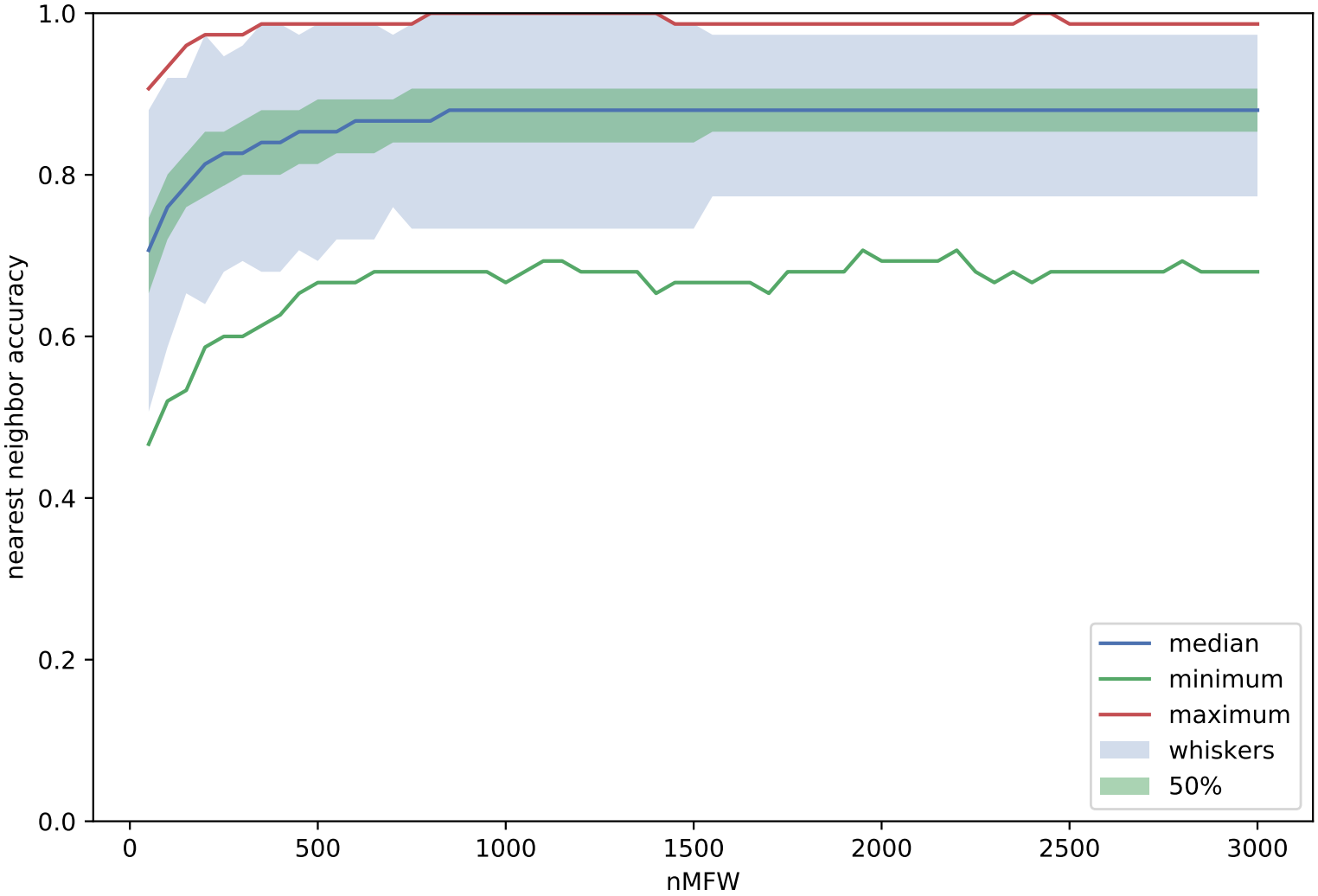
(Proisl *et al.* 2018)



Uncertainty of MFW evaluation experiments (Setting A)

Results: Implications

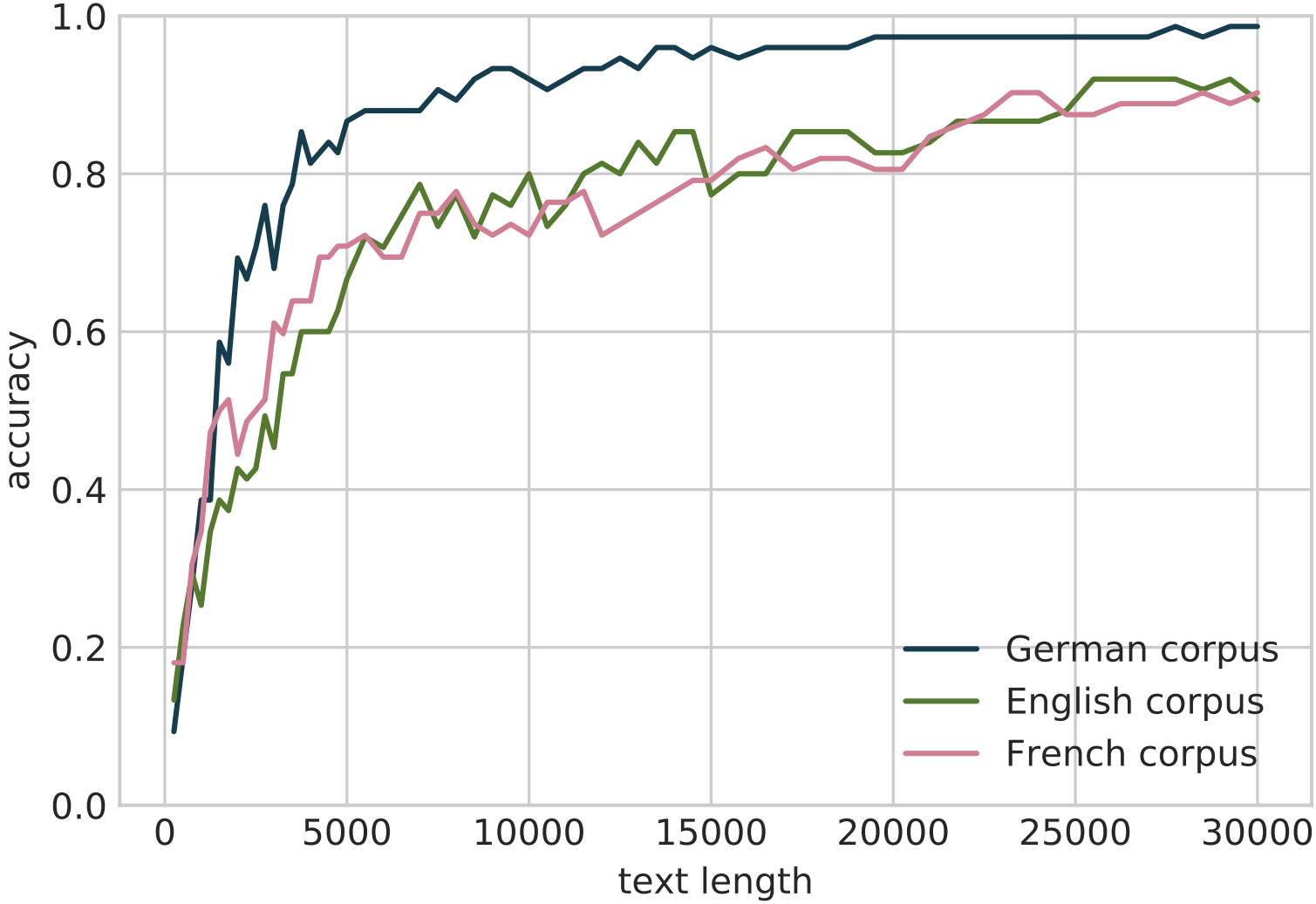
(Proisl *et al.* 2018)



Uncertainty of MFW evaluation experiments (Setting B)

Results: Implications

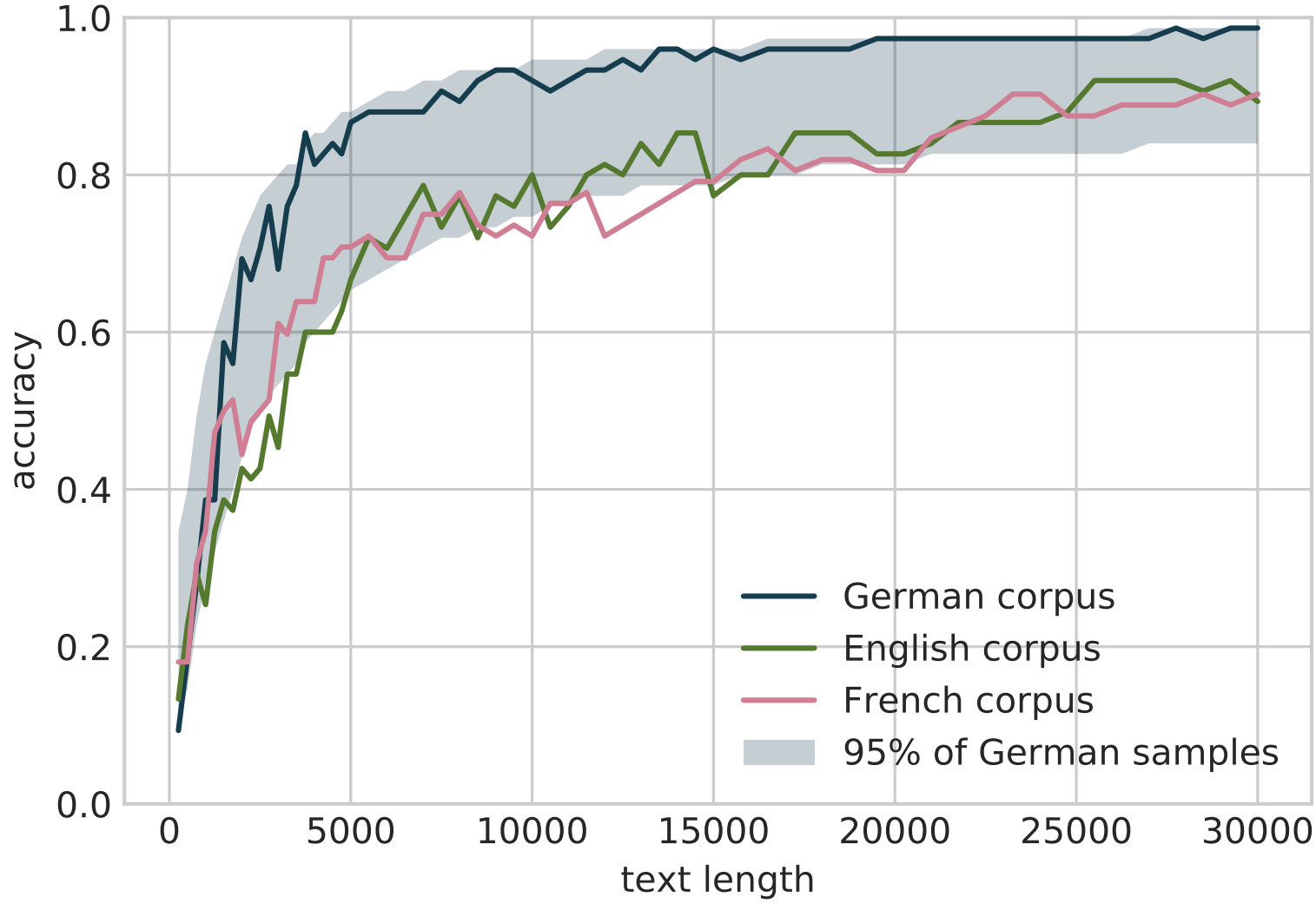
(Proisl *et al.* 2018)



Are differences between the languages significant?

Results: Implications

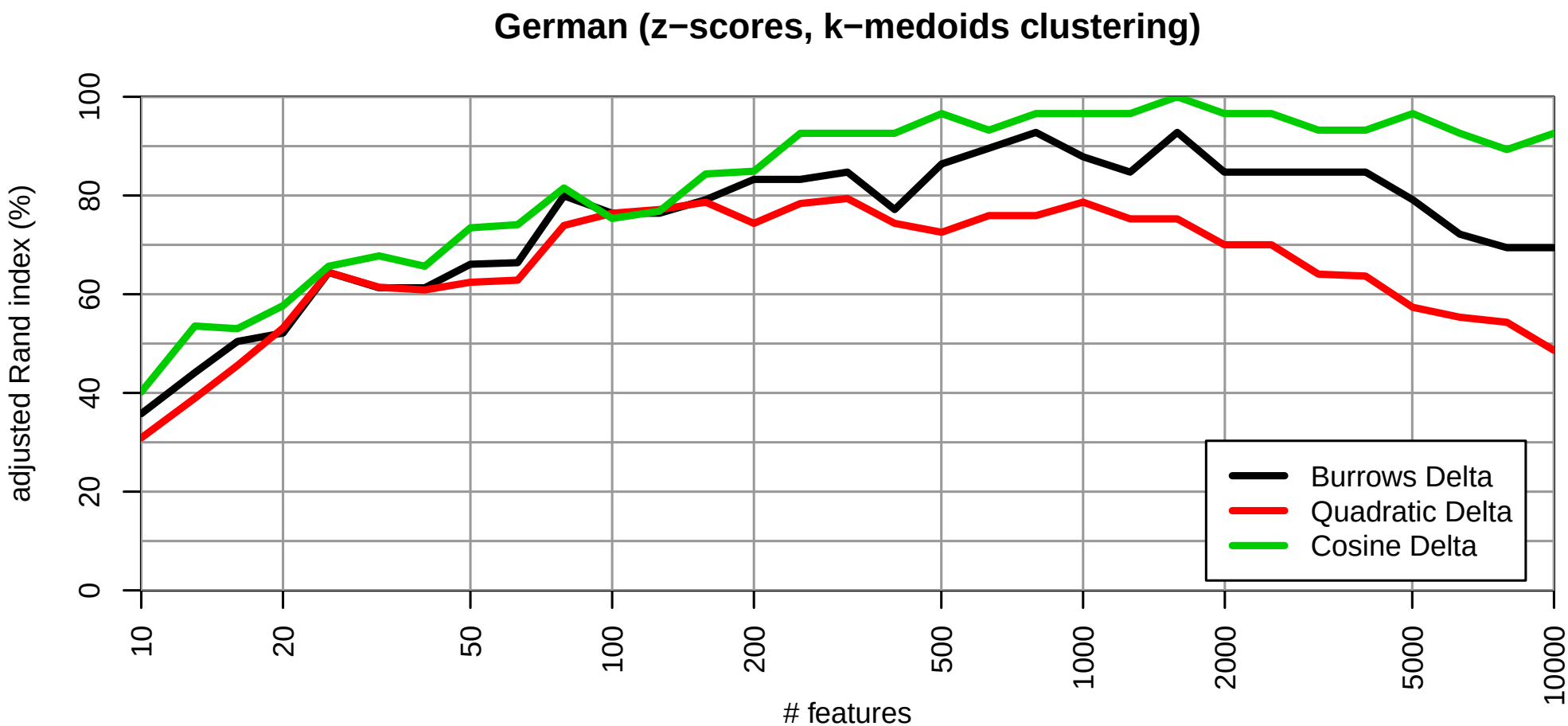
(Proisl *et al.* 2018)



Are differences between the languages significant? (Setting B)

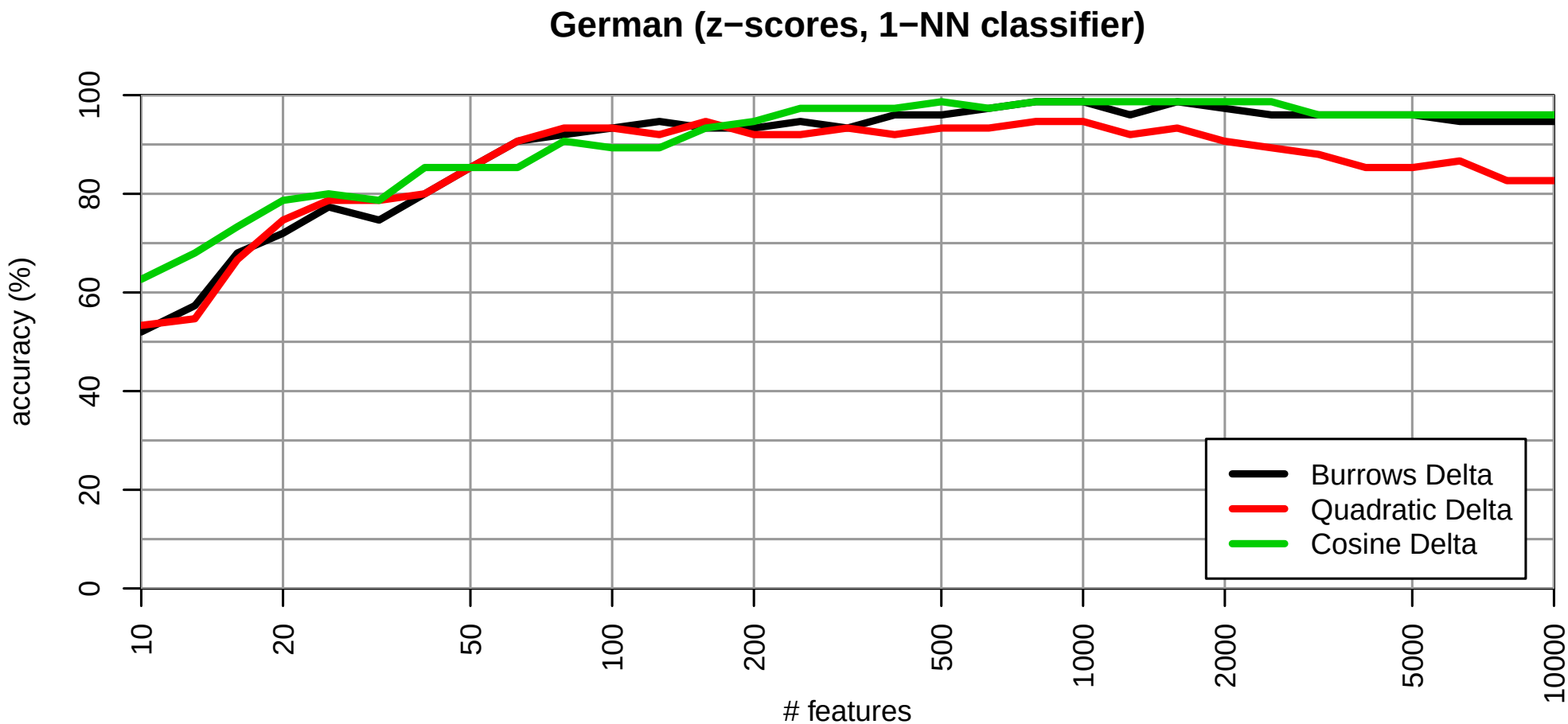
Could / should we have known?

- Standard approach for supervised classification tasks: cross-validation
  - assume test items are random sample ➜ binomial distribution
  - ignores variability due to test set and choice of authors



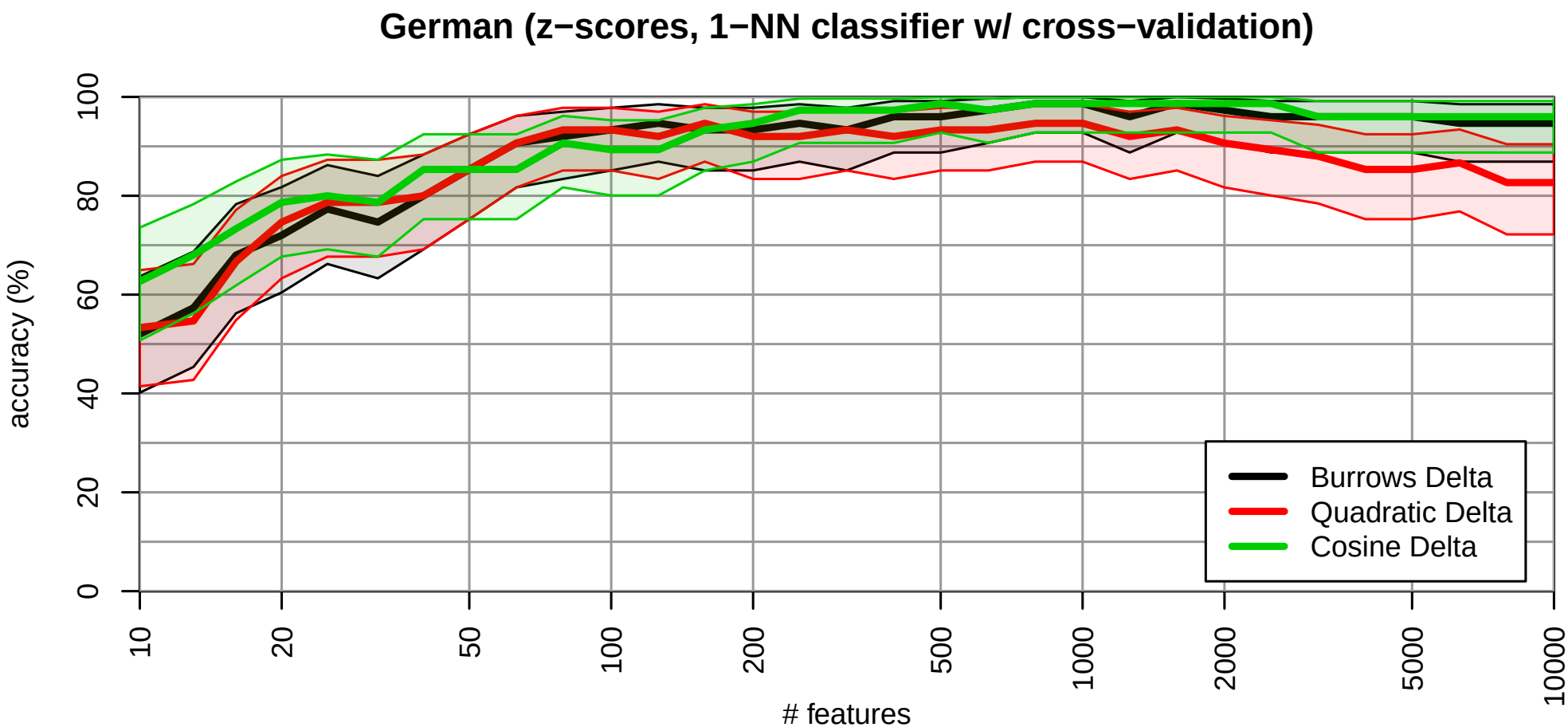
Could / should we have known?

- Standard approach for supervised classification tasks: cross-validation
  - assume test items are random sample ➜ binomial distribution
  - ignores variability due to test set and choice of authors



Could / should we have known?

- Standard approach for supervised classification tasks: cross-validation
  - assume test items are random sample ➜ binomial distribution
  - ignores variability due to test set and choice of authors



Could / should we have known?

- Significance testing for clustering task is difficult :-(

Could / should we have known?

- Significance testing for clustering task is difficult :-(

- Bootstrapping doesn't make sense for clustering
  - re-sampling with replacement ➜ duplicate texts
  - theory of bootstrapping focused on single numerical variable

Could / should we have known?

- Significance testing for clustering task is difficult :-(
- Bootstrapping doesn't make sense for clustering
  - re-sampling with replacement ➜ duplicate texts
  - theory of bootstrapping focused on single numerical variable
- Data sets usually too small for sub-sampling
  - our data set: 3-fold cross validation (each 25×2 novels)

Could / should we have known?

- Significance testing for clustering task is difficult :-(
- Bootstrapping doesn't make sense for clustering
  - re-sampling with replacement ➜ duplicate texts
  - theory of bootstrapping focused on single numerical variable
- Data sets usually too small for sub-sampling
  - our data set: 3-fold cross validation (each 25×2 novels)
- P-value clustering does something entirely different (Suzuki and Shimodaira 2006)
  - MFW as random sample from population of possible features

Could / should we have known?

- Significance testing for clustering task is difficult :-(
- Bootstrapping doesn't make sense for clustering
  - re-sampling with replacement ➜ duplicate texts
  - theory of bootstrapping focused on single numerical variable
- Data sets usually too small for sub-sampling
  - our data set: 3-fold cross validation (each 25×2 novels)
- P-value clustering does something entirely different (Suzuki and Shimodaira 2006)
  - MFW as random sample from population of possible features
- Possible approach: bootstrapping re-samples of texts
  - but ignores variability due to choice of authors & novels

Thank you!

Next steps:

- ☞ Replicate all findings with sampling experiments
- ☞ Implement text-level bootstrapping for clustering task
- ☞ Consensus clustering for Δ on character & word n-grams
- ☞ Can we integrate Δ and n-gram tracing?

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.

Büttner, Andreas; Dimpel, Friedrich Michael; Evert, Stefan; Jannidis, Fotis; Pielström, Steffen; Proisl, Thomas; Reger, Isabella; Schöch, Christof; Vitt, Thorsten (2017). »Delta« in der stilometrischen Autorschaftsattribution. *Zeitschrift für digitale Geisteswissenschaften*. `https://dx.doi.org/10.17175/2017_006`.

Dimpel, Friedrich Michael (2018). Die guten ins Töpfchen: Zur Anwendbarkeit von Burrows' Delta bei kurzen mittelhochdeutschen Texten nebst eines Attributionstests zu Konrads ‚Halber Birne'. In *Proceedings of DHd 2018*, pages 168–173, Cologne, Germany.

Eder, Maciej (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, **28**(4), 603–614.

Eder, Maciej (2015). Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, **30**(2), 167–182.

Evert, Stefan; Proisl, Thomas; Jannidis, Fotis; Reger, Isabella; Pielström, Steffen; Schöch, Christof; Vitt, Thorsten (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*. Advance access `https://doi.org/10.1093/llc/fqw046`.

References II

Grieve, Jack; Clarke, Isobelle; Chiang, Emily; Gideon, Hannah; Heini, Annina; Nini, Andrea; Waibel, Emily (2018). Attributing the Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqy042.

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

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

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.

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

Proisl, Thomas; Evert, Stefan; Jannidis, Fotis; Schöch, Christof; Konle, Leonard; Pielström, Steffen (2018). Delta vs. n-gram tracing: Evaluating the robustness of authorship attribution methods. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3309–3314, Miyazaki, Japan.

References III

Rybicki, Jan and Eder, Maciej (2011). Deeper delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, **26**(3), 315–321.

Rybicki, Jan and Heydel, Magda (2013). The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing*, **28**(4), 708–717.

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.

Suzuki, Ryota and Shimodaira, Hidetoshi (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**(12), 1540–1542.

Thisted, Ronald and Efron, Bradley (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, **74**(3), 445–455.