Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

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

Explaining Delta

How do distance measures for authorship attribution work?

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

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

Lancaster, 24 July 2015

Outline

Authorship attribution

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

- Identify unknown author or settle case of disputed authorship
  - Federalist papers: Hamilton vs. Madison (Mosteller and Wallace 1963)
  - Did Shakespeare really exist?
  - 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/

Authorship attribution

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

- Identify unknown author or settle case of disputed authorship
  - Federalist papers: Hamilton vs. Madison (Mosteller and Wallace 1963)
  - Did Shakespeare really exist?
  - 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/

- Which stylometric features determine the characteristic style of a literary author?
  - authorship attribution as a proxy task
  - "successful" features ➜ particularly characteristic for author

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
  - evaluation: classification accuracy

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
  - evaluation: classification accuracy
- Authorship attribution as clustering task
  - given set of unknown texts
  - group texts written by same author into cluster
  - evaluation: adjusted Rand index (ARI)

  ☞ more general approach

Authorship attribution

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

- Popular approach: supervised machine learning
  - wide range of stylometric features
  - ML trained on texts with known authorship
  - feature selection & weighting

Authorship attribution

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

- Popular approach: supervised machine learning
  - wide range of stylometric features
  - ML trained on texts with known authorship
  - feature selection & weighting
- But not suitable for clustering task
  - no supervised training data available
  - clustering based on stylometric distance between texts (metric)
  - no easy way to determine feature weights for metric

Authorship attribution

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

- Popular approach: supervised machine learning
  - wide range of stylometric features
  - ML trained on texts with known authorship
  - feature selection & weighting
- But not suitable for clustering task
  - no supervised training data available
  - clustering based on stylometric distance between texts (metric)
  - no easy way to determine feature weights for metric
- Simple Delta measure (Burrows 2002) very successful

Burrows's Delta (Δ_B)

(Burrows 2002)

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

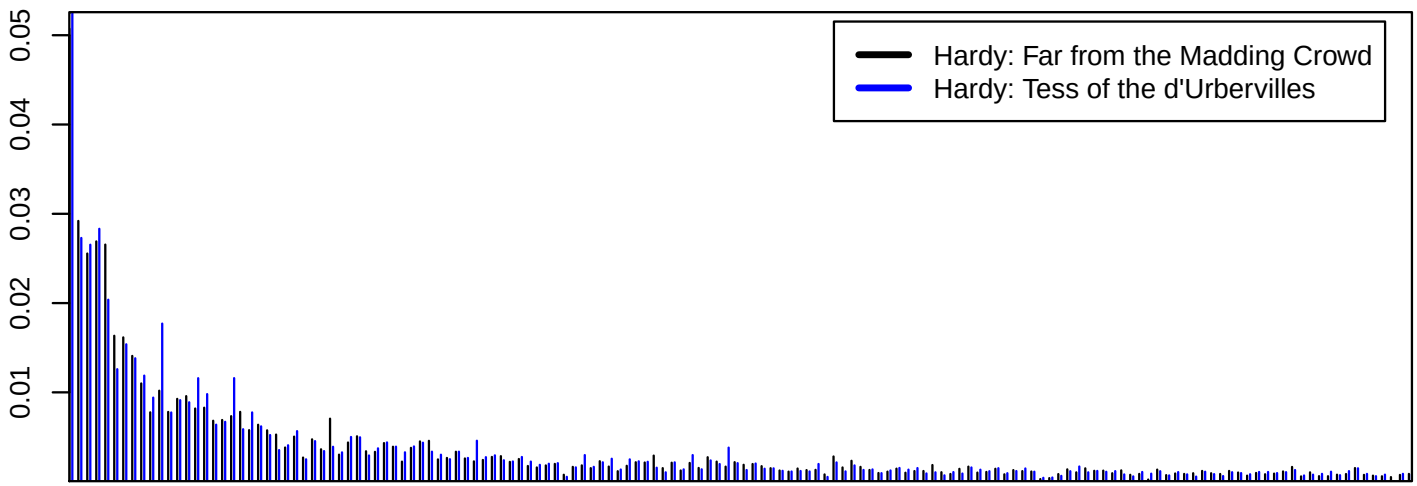
	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*	
f(Madding Crowd) =	(.051,	.029,	.026,	.027,	.027,	.016,	.016,	.014,	.011,	.008,	.010,	...)
f(Tess of the d'U.) =	(.053,	.027,	.027,	.028,	.020,	.013,	.015,	.014,	.012,	.009,	.018,	...)
f(Oliver Twist) =	(.055,	.032,	.024,	.023,	.022,	.012,	.014,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

$$
\begin{array}{rl}
 & \quad \textit{the} \quad \textit{and} \quad \textit{to} \quad \textit{of} \quad \textit{a} \quad \textit{I} \quad \textit{in} \quad \textit{was} \quad \textit{that} \quad \textit{he} \quad \textit{her} \\
\mathbf{f}(\text{Madding Crowd}) = & (.051, .029, .026, .027, .027, .016, .016, .014, .011, .008, .010, \ldots) \\
\mathbf{f}(\text{Tess of the d'U.}) = & (.053, .027, .027, .028, .020, .013, .015, .014, .012, .009, .018, \ldots) \\
\mathbf{f}(\text{Oliver Twist}) = & (.055, .032, .024, .023, .022, .012, .014, .011, .011, .013, .005, \ldots)
\end{array}
$$

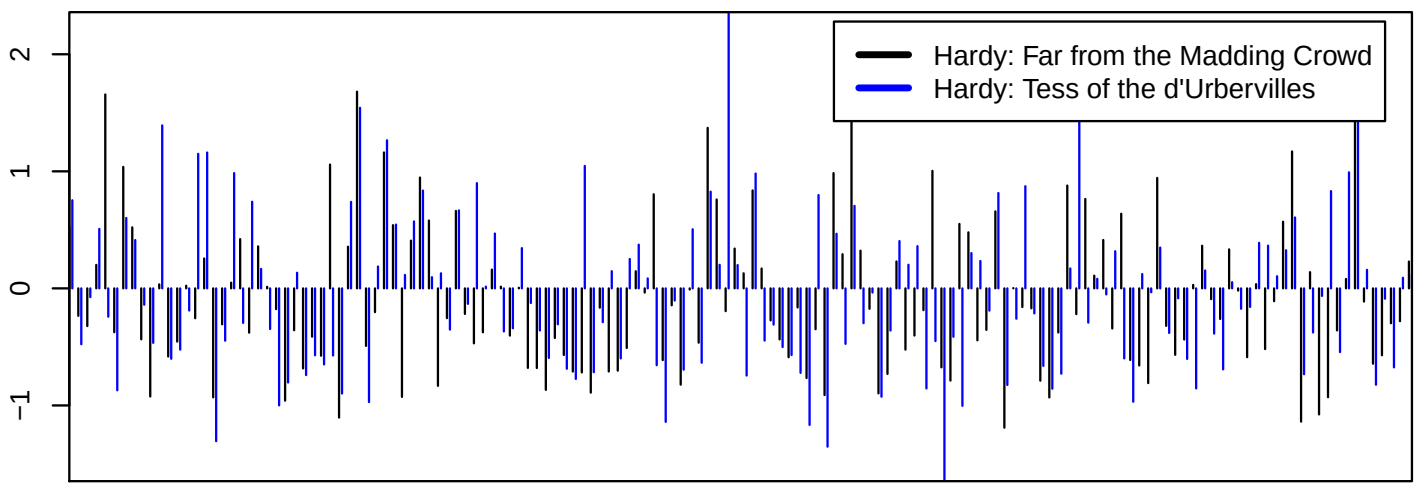


Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100 – 5,000 most frequent words (MFW) form a "fingerprint" of an author's style
- Standardized to z-scores to give each word equal weight

	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,	...)

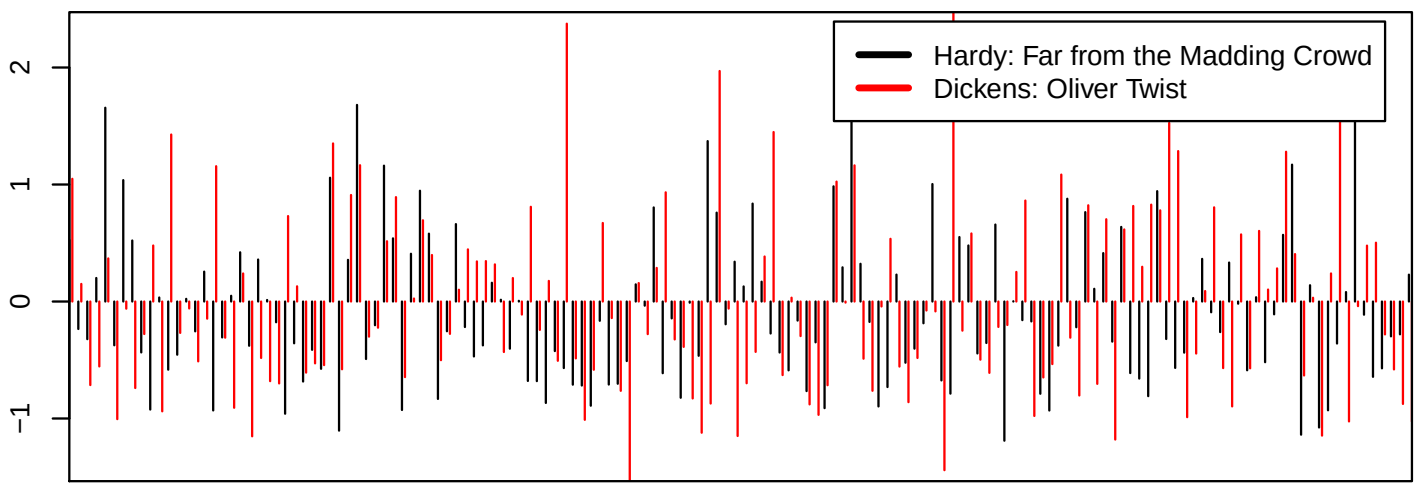


Burrows's Delta (Δ_B)

(Burrows 2002)

- Frequencies of 100 – 5,000 most frequent words (MFW) form a "fingerprint" of an author's style
- Standardized to z-scores to give each word equal weight

	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,	...)



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)

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

Experiments

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

- Empirical study based on data of Jannidis *et al.* (2015)
  - corpora of English, German and French novels
  - 75 novels per corpus: 3 novels each from 75 authors
  - early 19th C. to middle of 20th C.
- Exp. 1: Understanding the parameters of Delta measures
- Exp. 2: How much data are needed?
- Exp. 3: Supervised feature selection

Understanding the parametern of Delta

Prior work by Jannidis *et al.* (2015)

- Novels grouped into 25 clusters based on Delta distances
- All known Delta measures for $n_w = 100, 1000, 5000$ MFW
- Evaluation: within/between distances, cluster purity
- Best results: Cosine Delta $\Delta_\angle$ (Smith and Aldridge 2011) and the original Burrows Delta Δ_B (Burrows 2002)
- Mathematically sensible variants of Delta (Argamon 2008) are much worse than Δ_B

Understanding the parametern of Delta

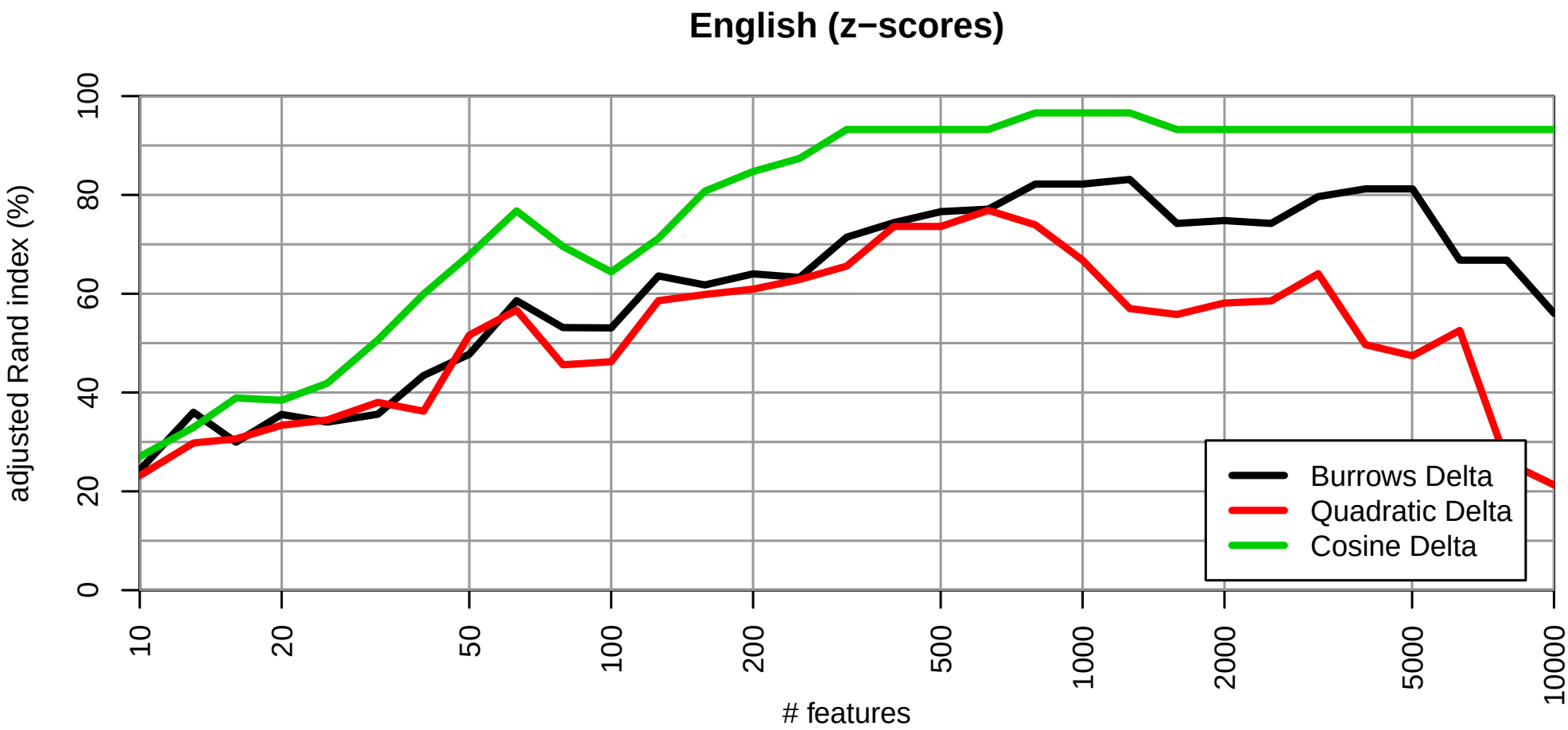
Prior work by Jannidis *et al.* (2015)

- Novels grouped into 25 clusters based on Delta distances
- All known Delta measures for $n_w = 100, 1000, 5000$ MFW
- Evaluation: within/between distances, cluster purity
- Best results: Cosine Delta $\Delta_\angle$ (Smith and Aldridge 2011) and the original Burrows Delta Δ_B (Burrows 2002)
- Mathematically sensible variants of Delta (Argamon 2008) are much worse than Δ_B

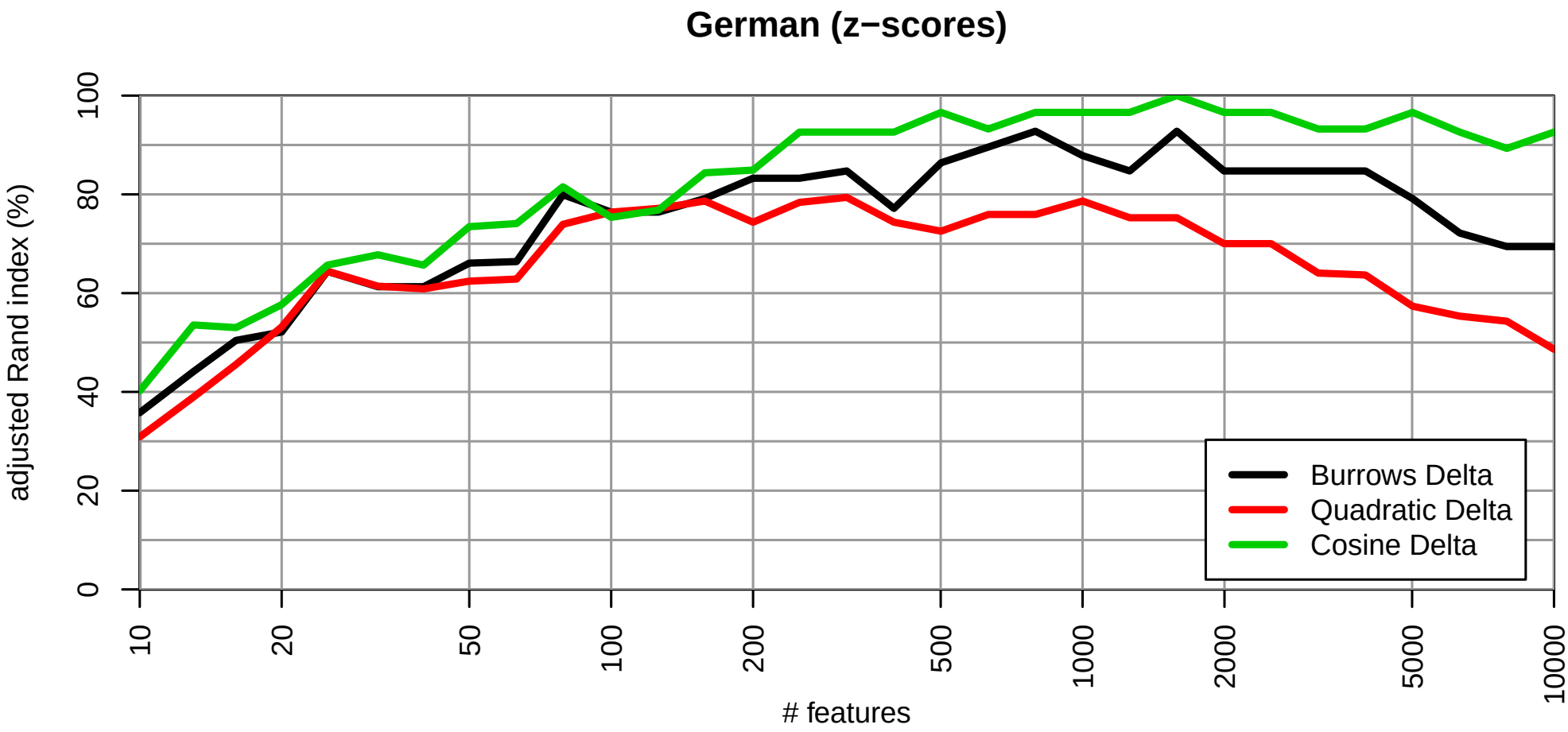
New results (Evert *et al.* 2015)

- Detailed plots of n_w for Δ_B, Δ_Q and $\Delta_\angle$
- Systematic experiments with different parameters of Delta
- Evaluation: adjusted Rand index (Hubert and Arabie 1985)

Parameter: Number n_w of MFW

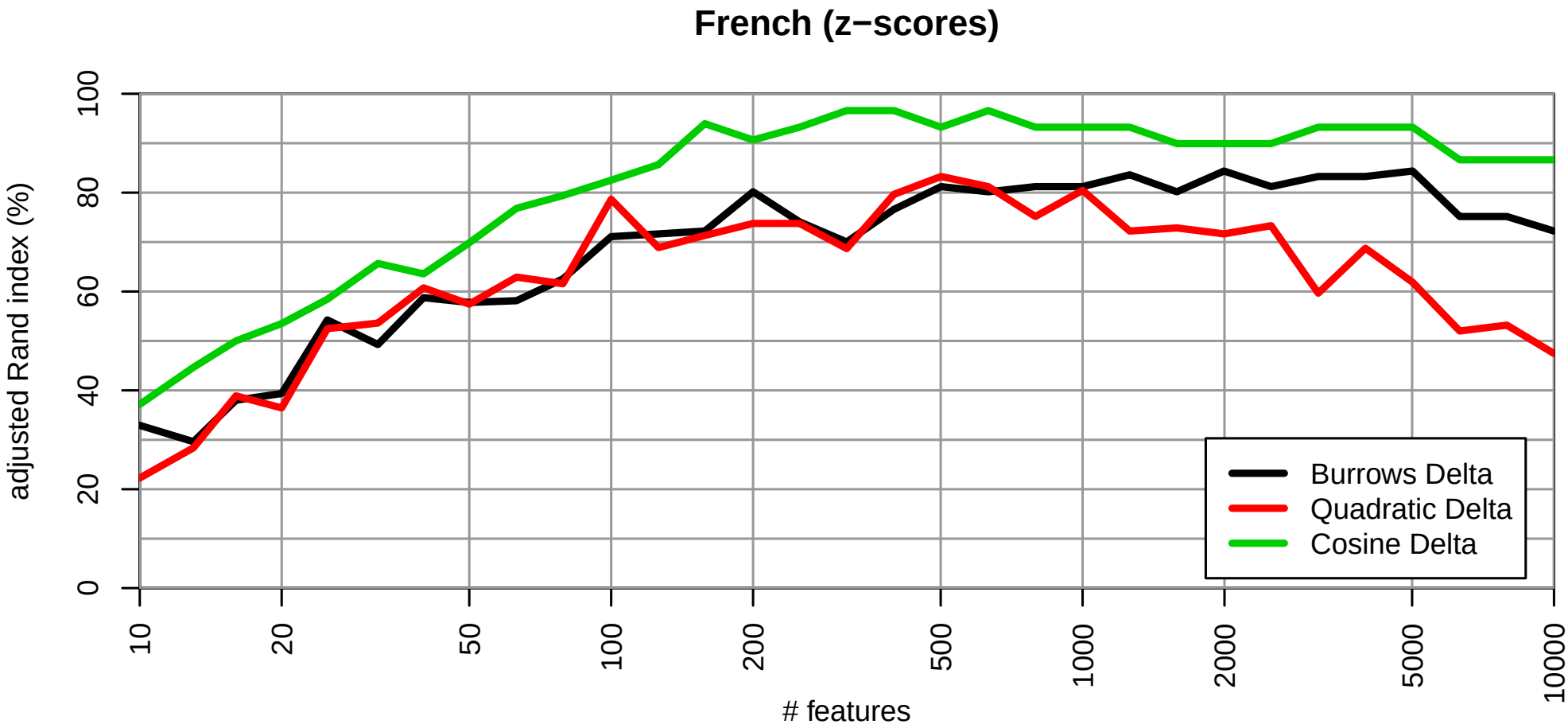


Parameter: Number n_w of MFW



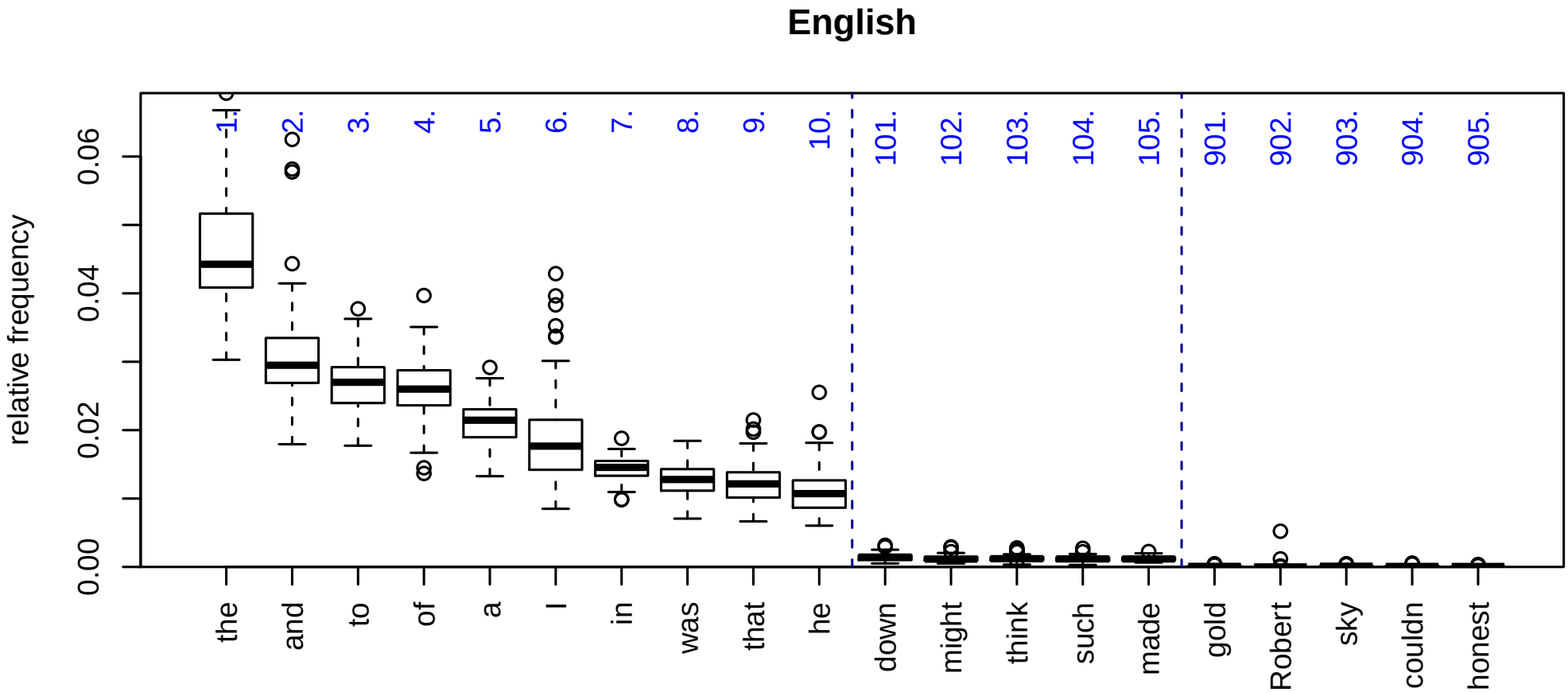
(following slides will show results for English corpus only)

Parameter: Number n_w of MFW



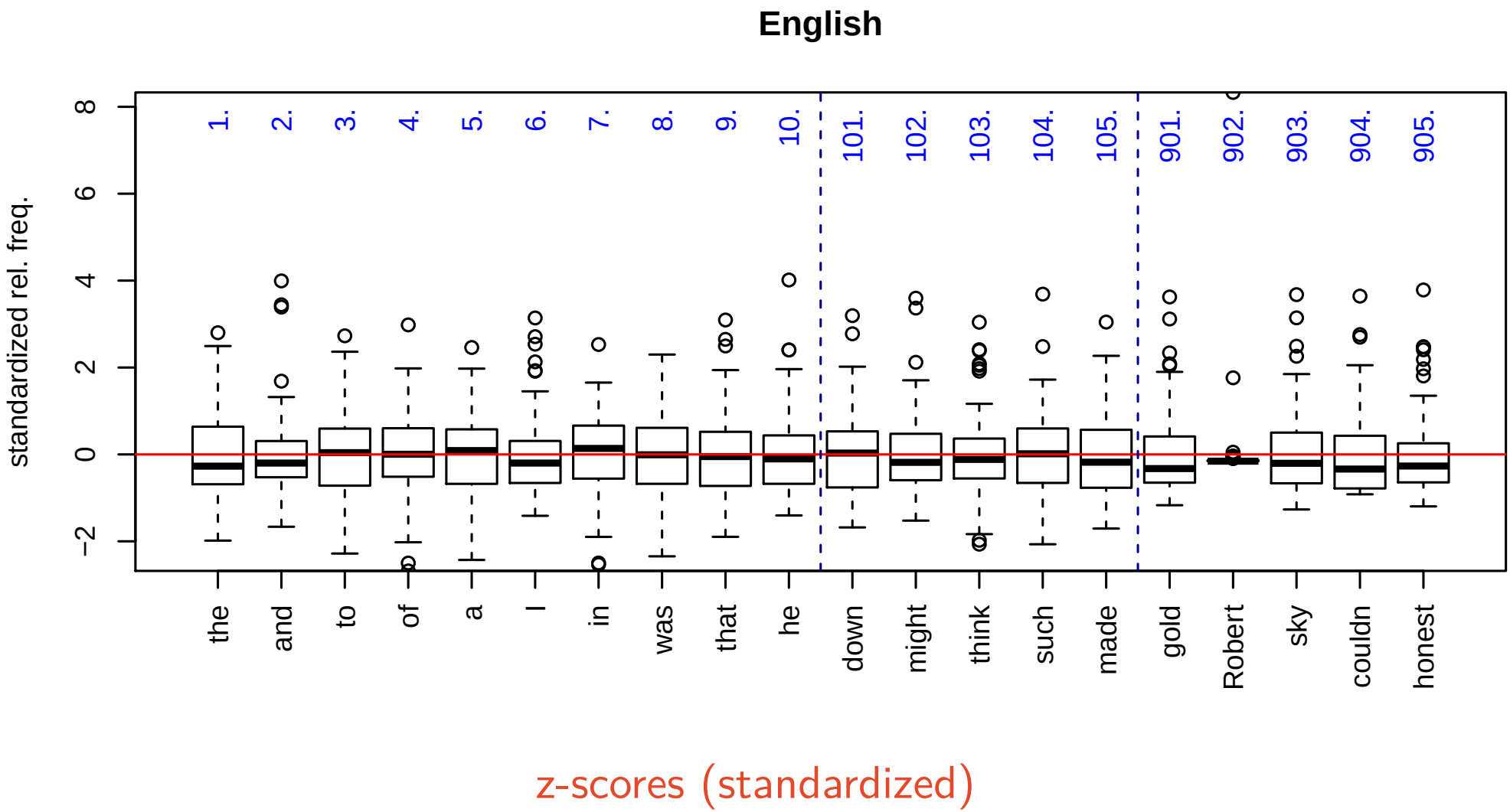
(following slides will show results for English corpus only)

Parameter: Standardization of relative frequencies



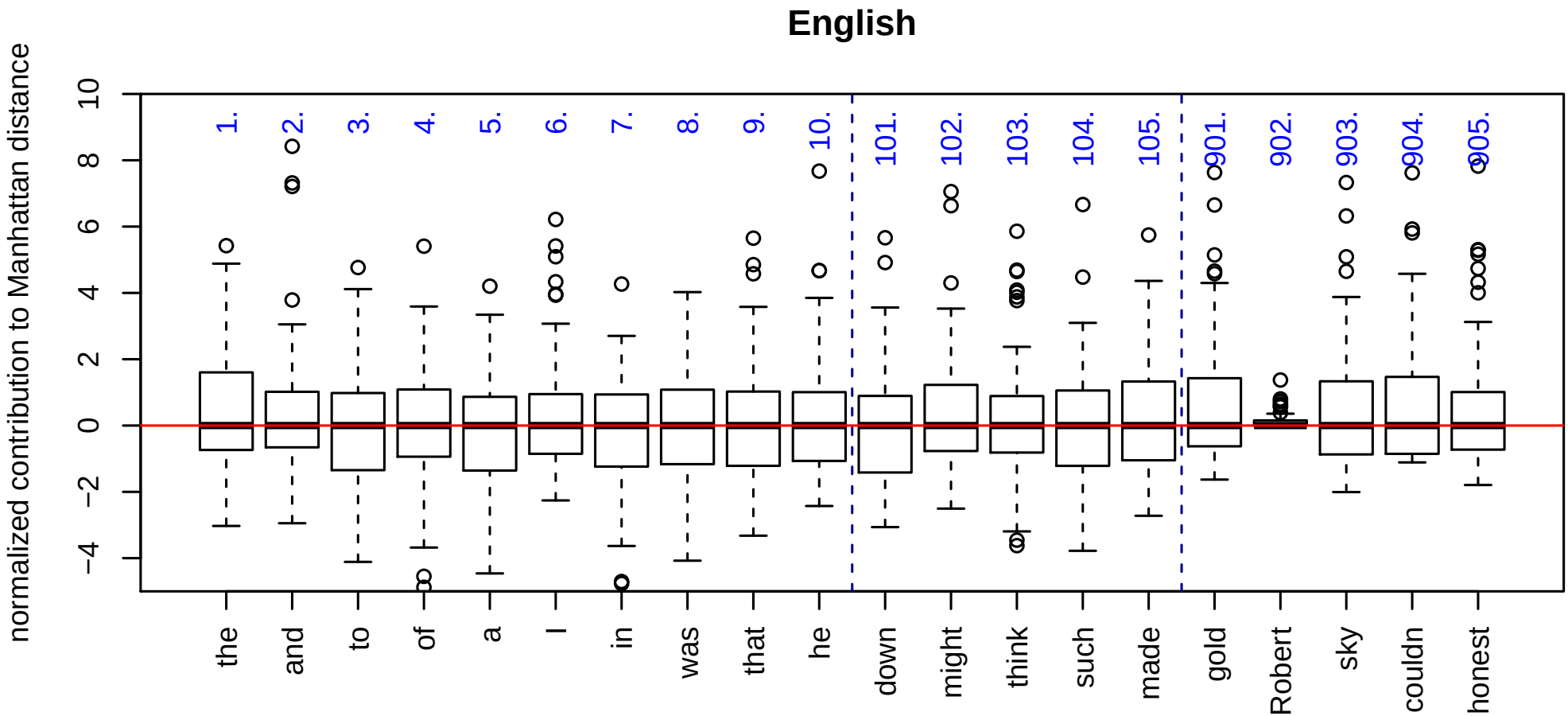
relative frequencies (unscaled)

Parameter: Standardization of relative frequencies



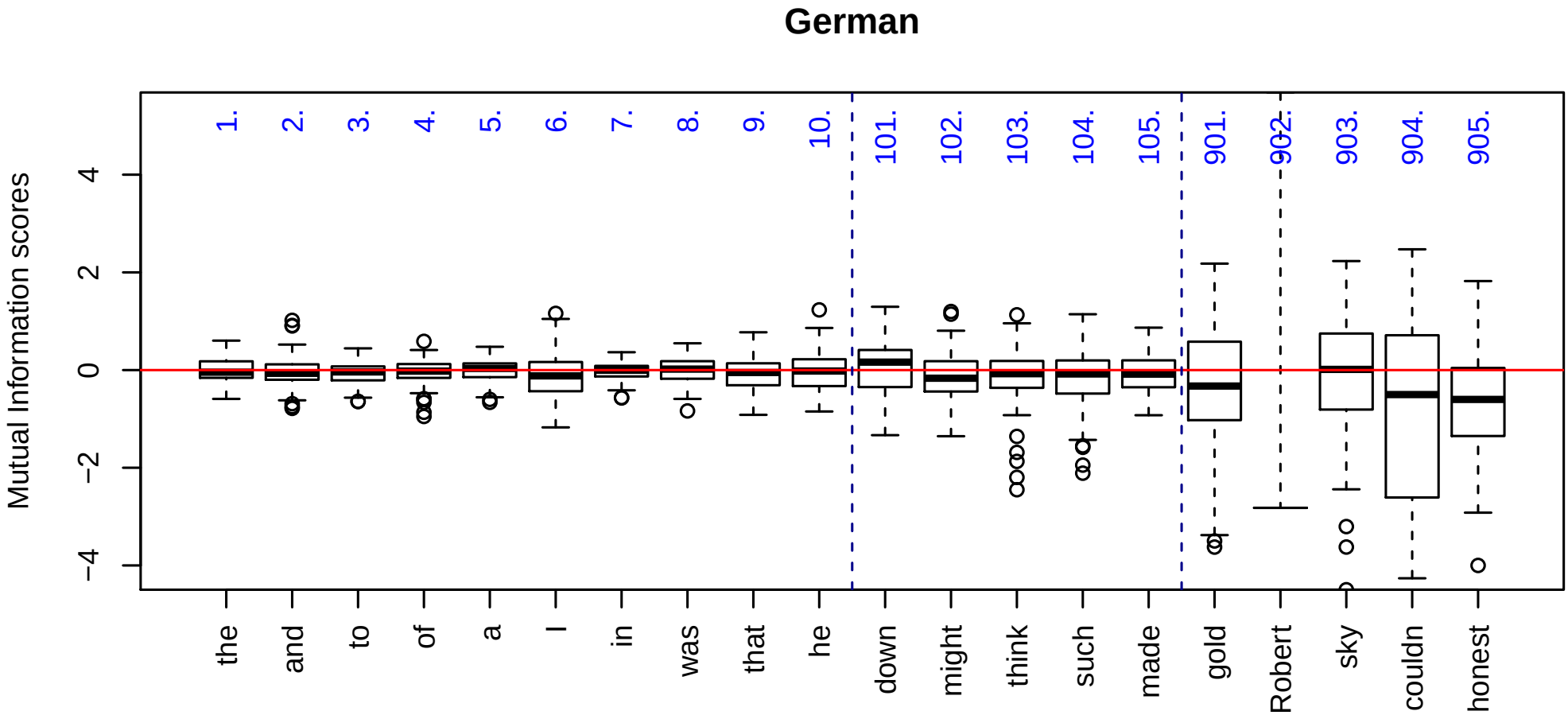
z-scores (standardized)

Parameter: Standardization of relative frequencies



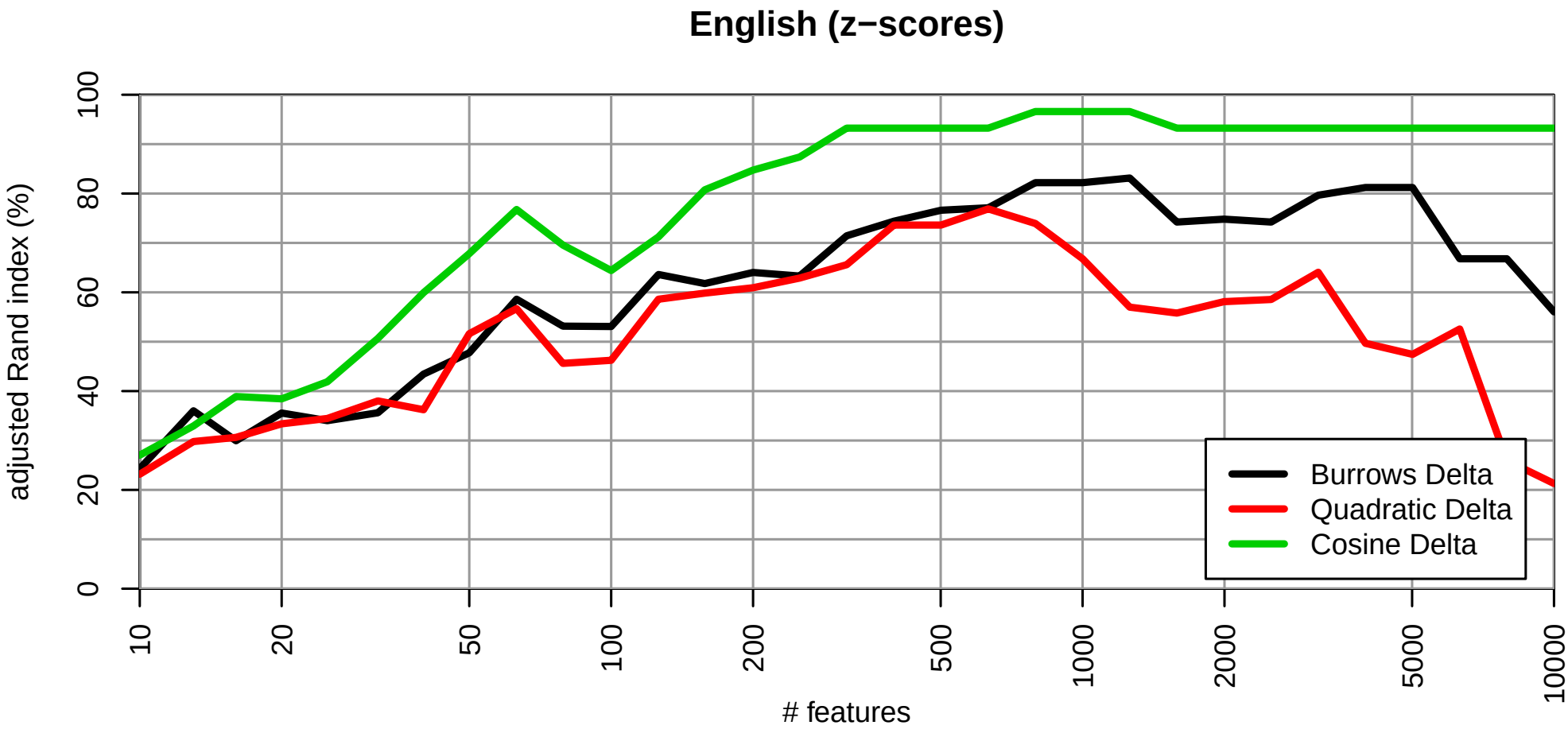
L_1-scaling $\rightarrow$ worse

Parameter: Standardization of relative frequencies

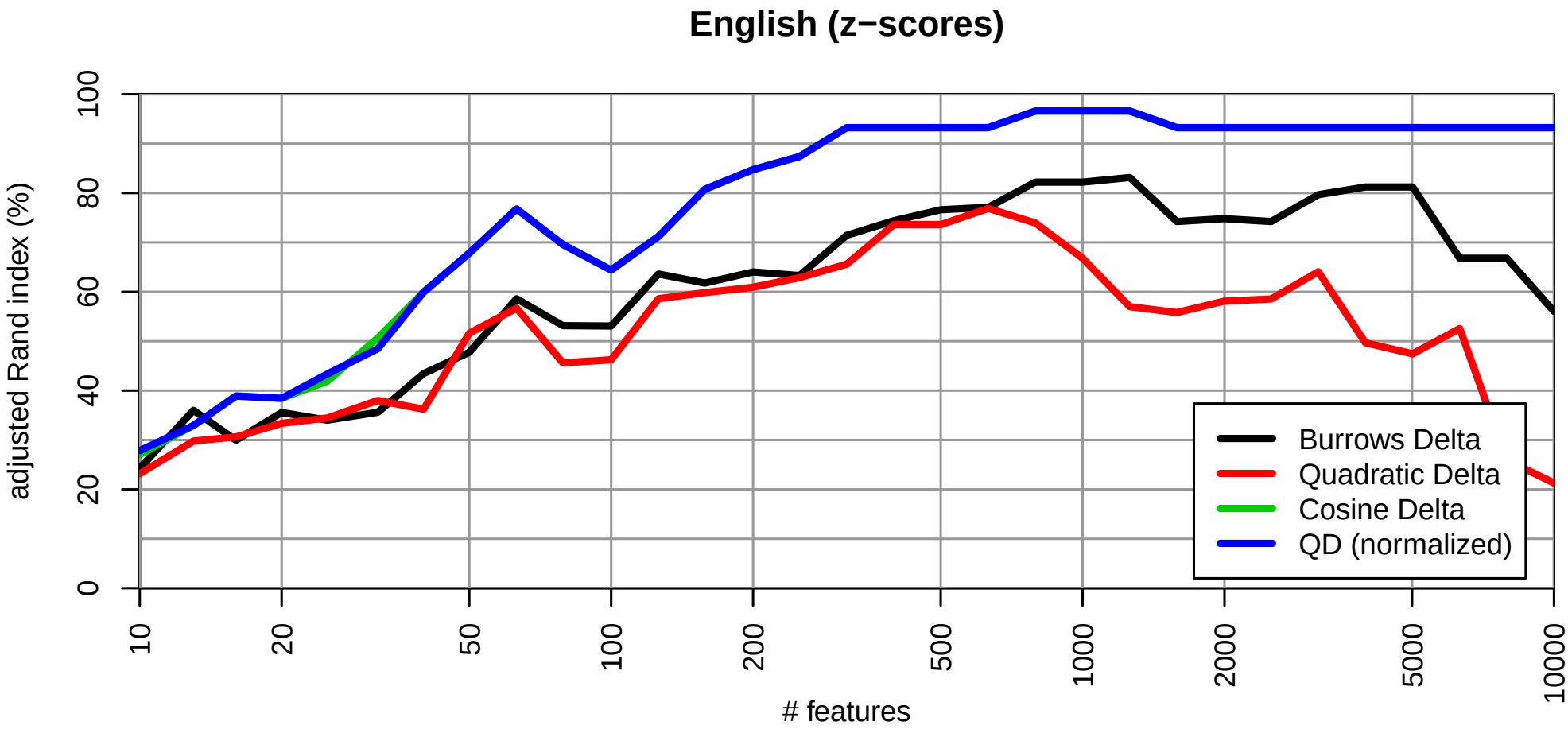


association measure: Mutual Information → much worse

Parameter: Normalization of vector length

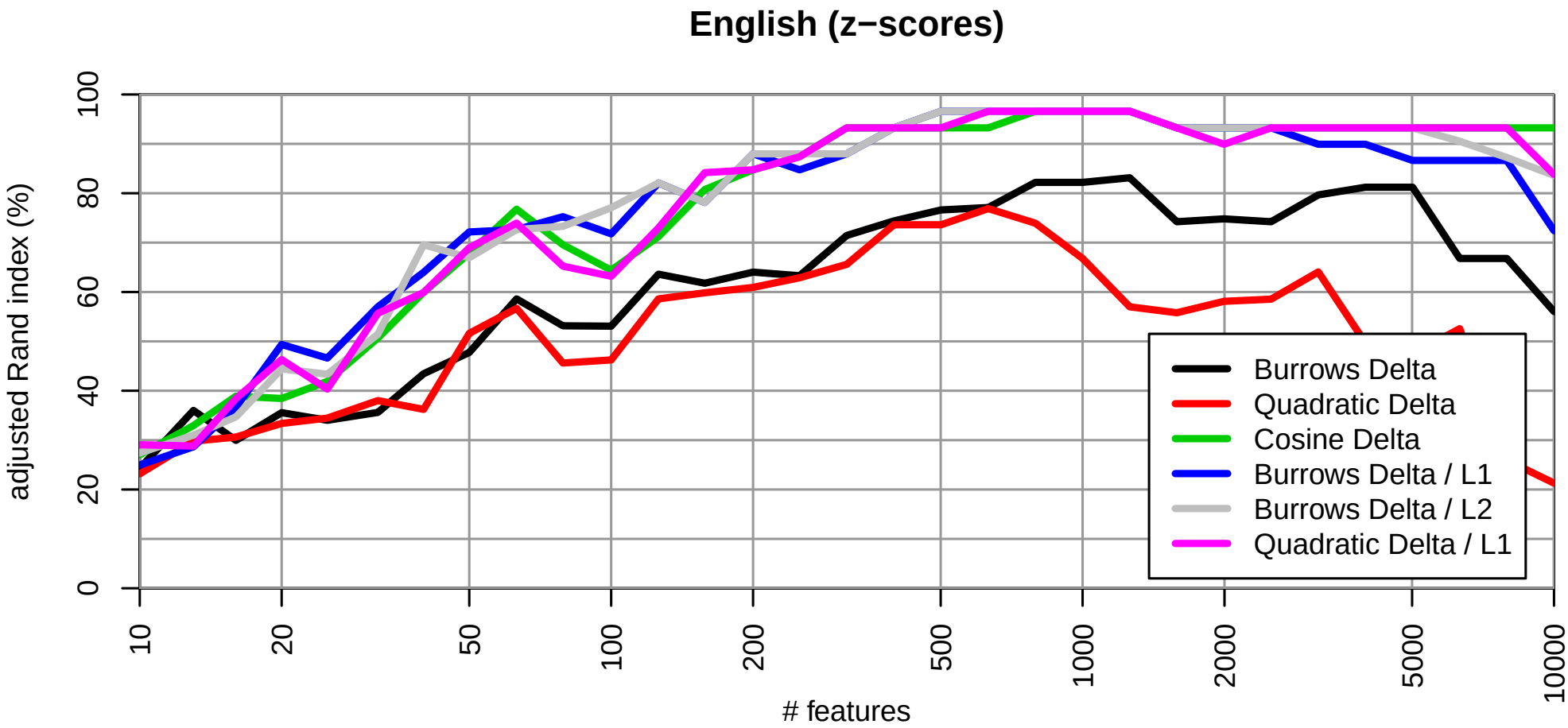


Parameter: Normalization of vector length



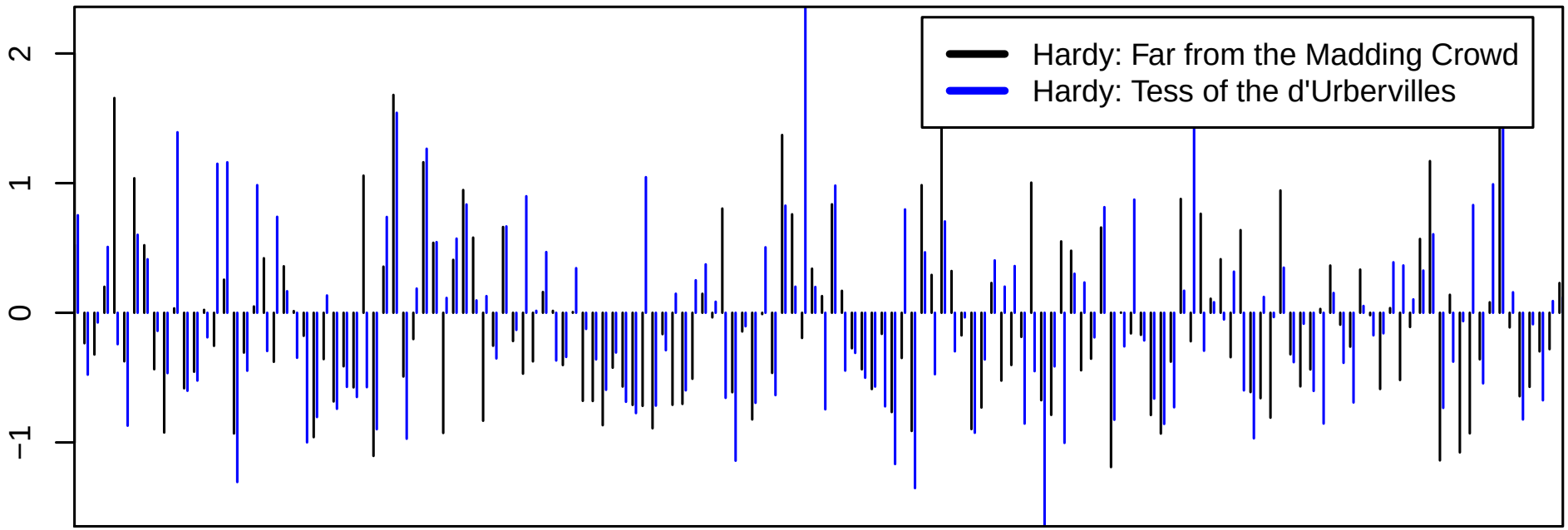
Euclidean distance (QD) + normalization = Cosine Delta

Parameter: Normalization of vector length



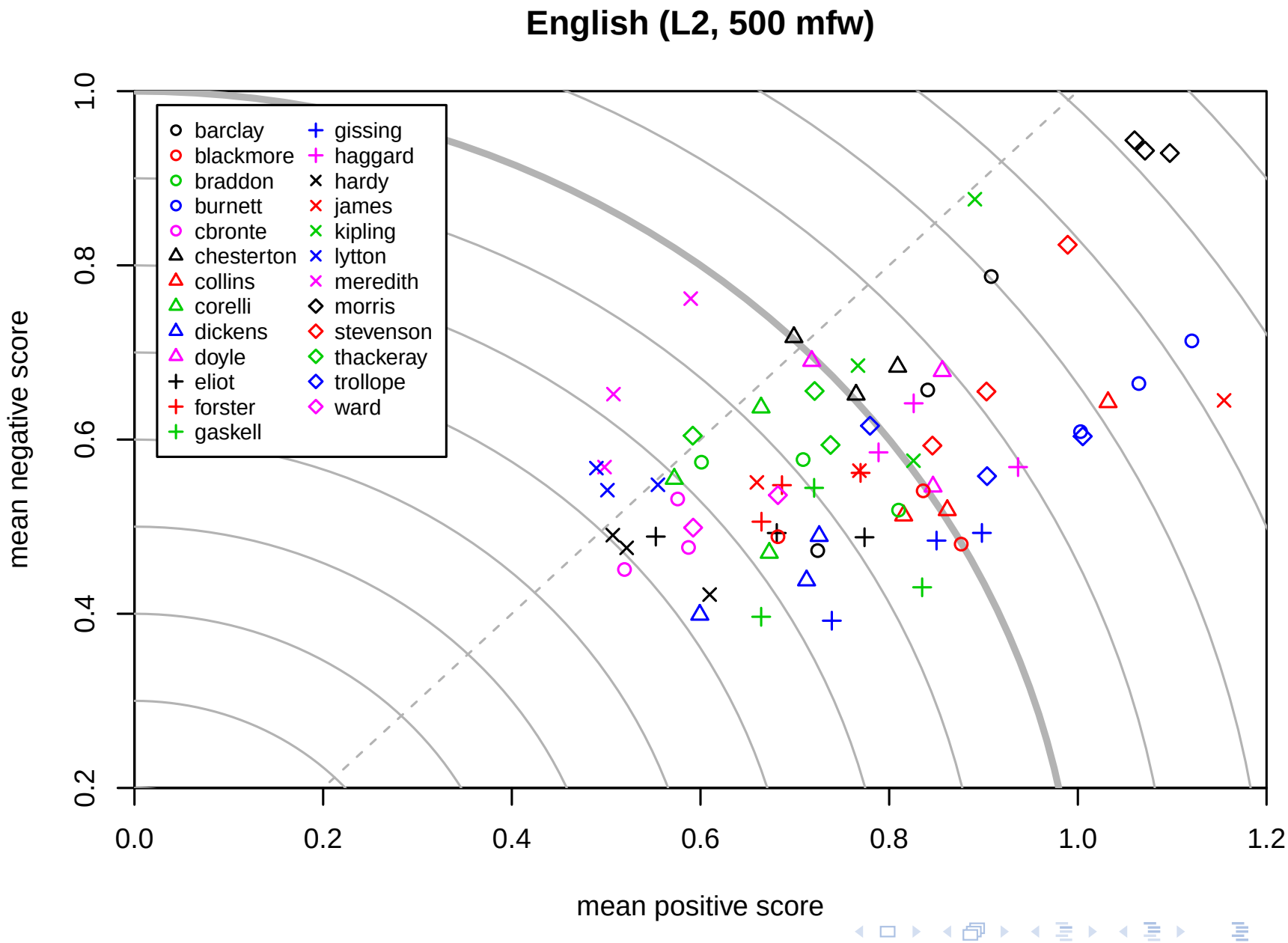
Normalization has crucial effect on clustering quality!

Why is vector normalization so important?

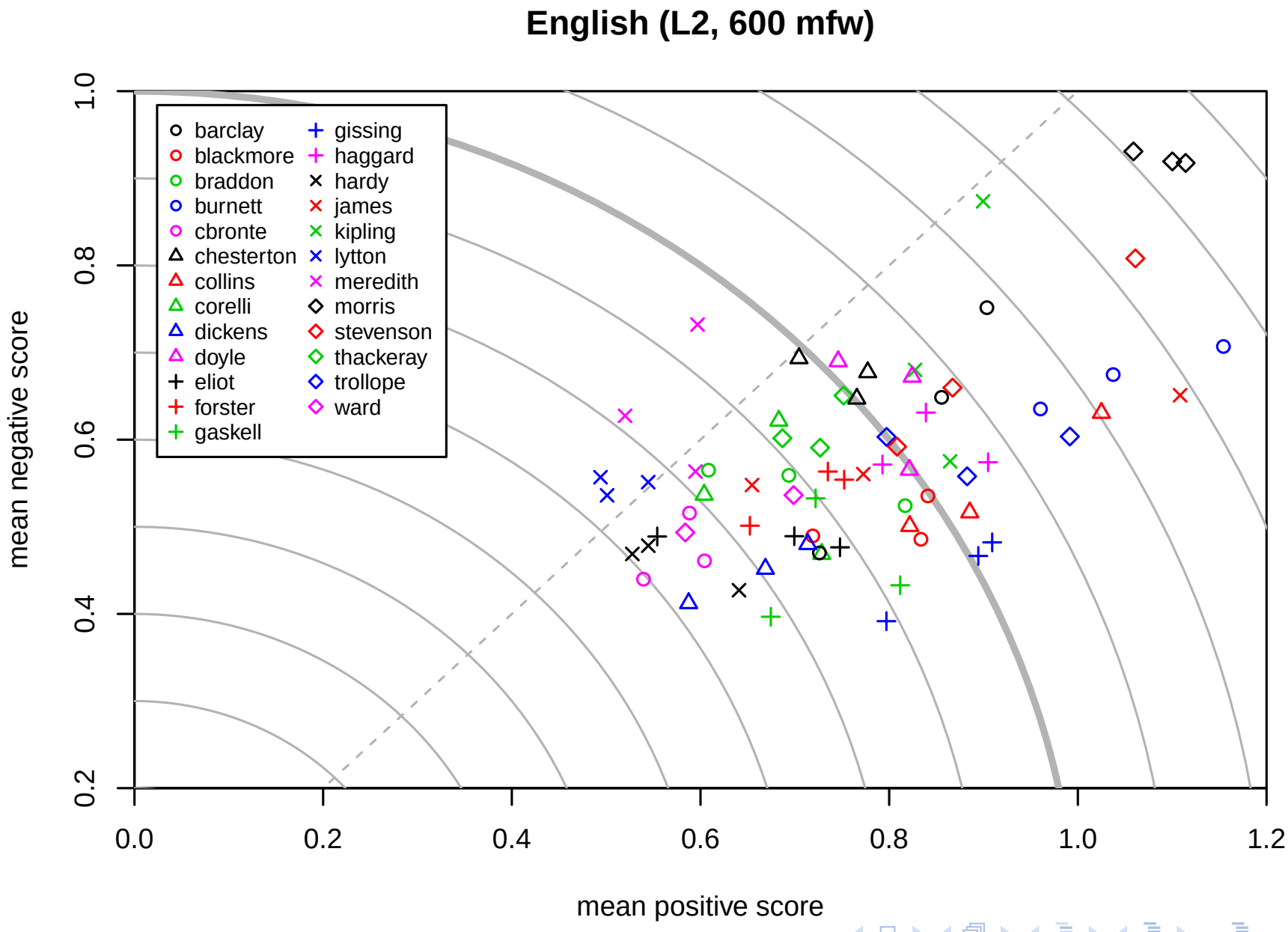


- Feature vector $\mathbf{z}(D)$ = stylistic "fingerprint" of author
- Our conjecture: pattern of positive/negative deviations from norm reflects individual stylistic profile of an author
- Vector length = degree to which individual style is expressed

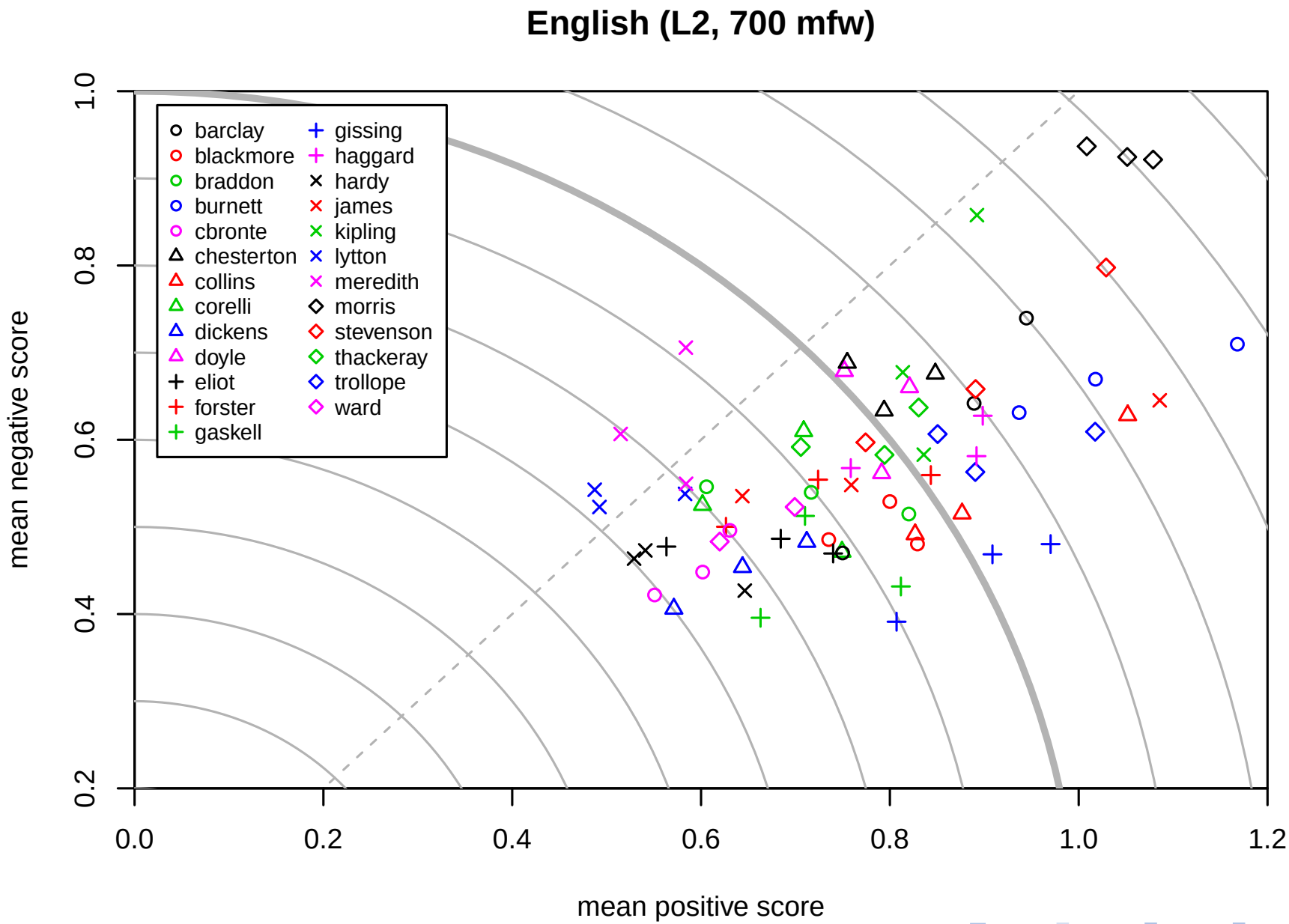
Why is vector normalization so important?



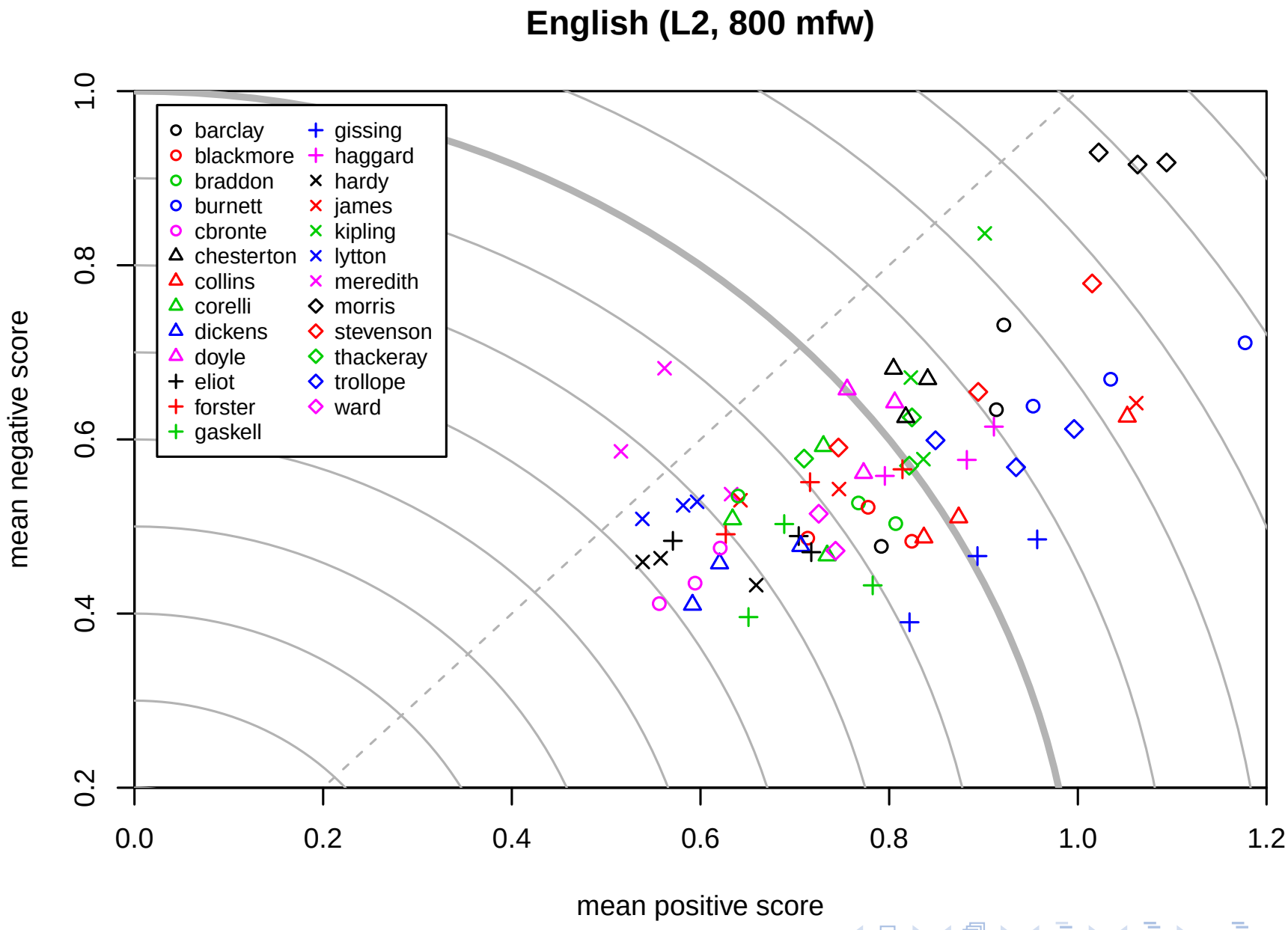
Why is vector normalization so important?



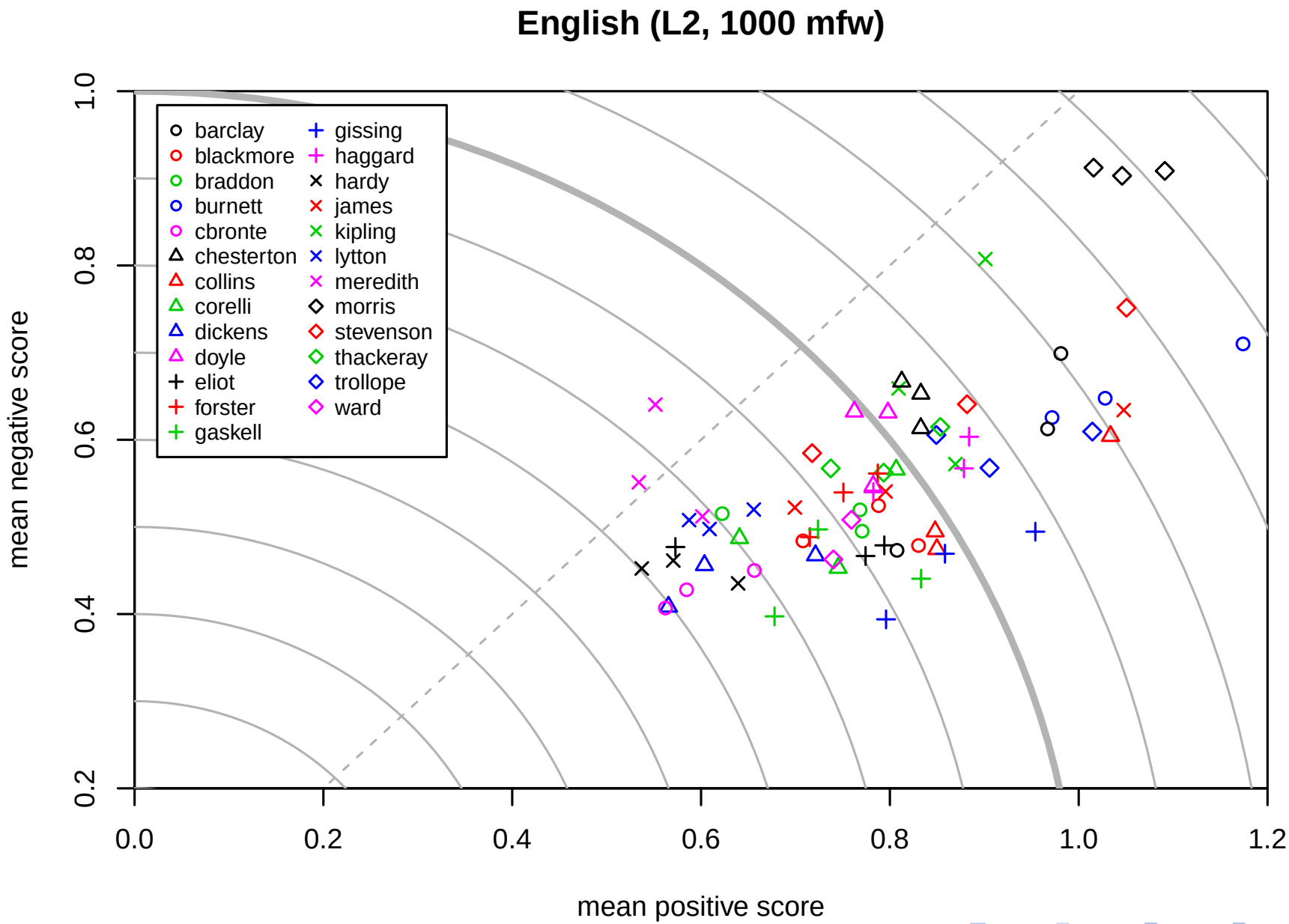
Why is vector normalization so important?



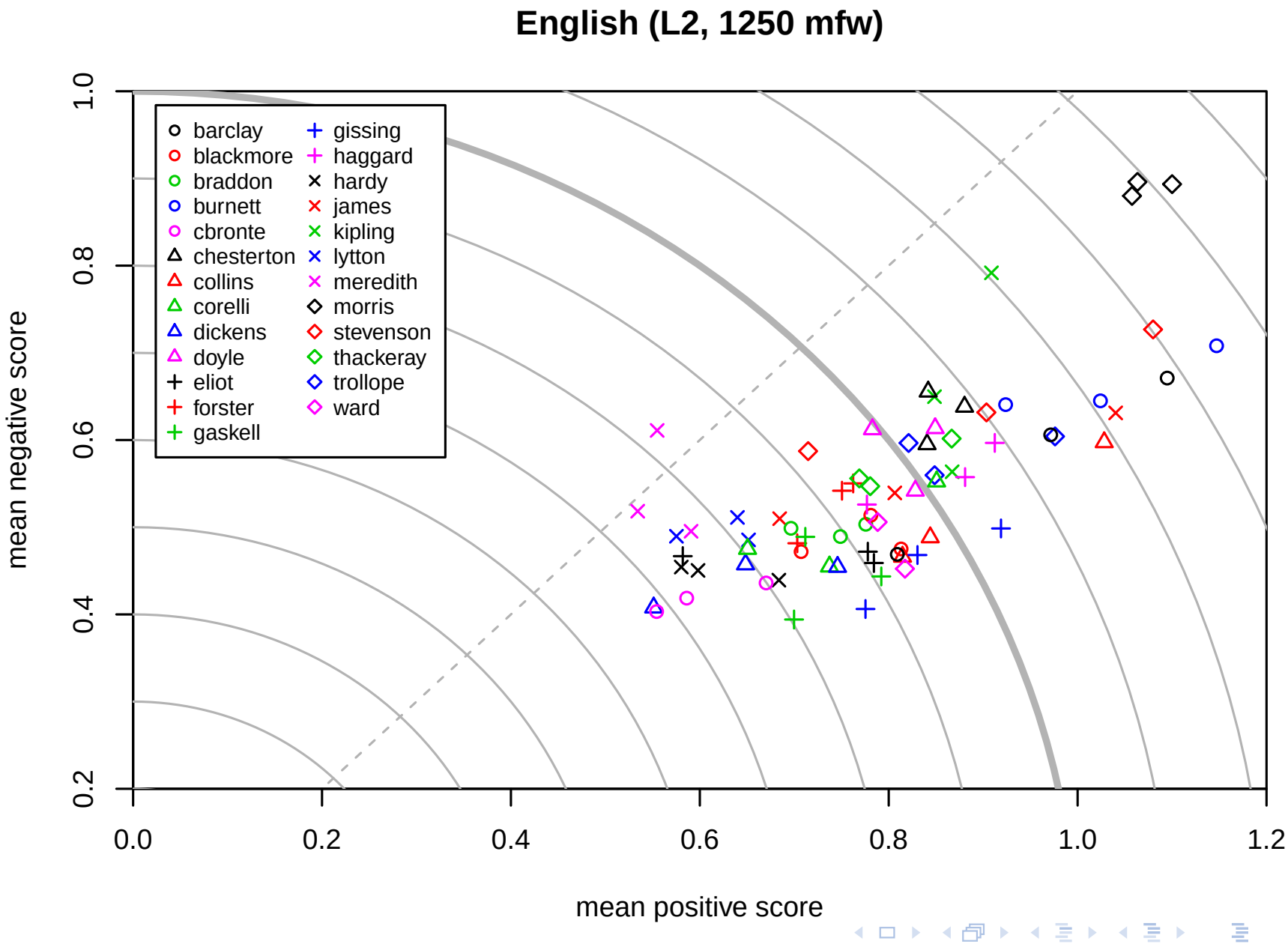
Why is vector normalization so important?



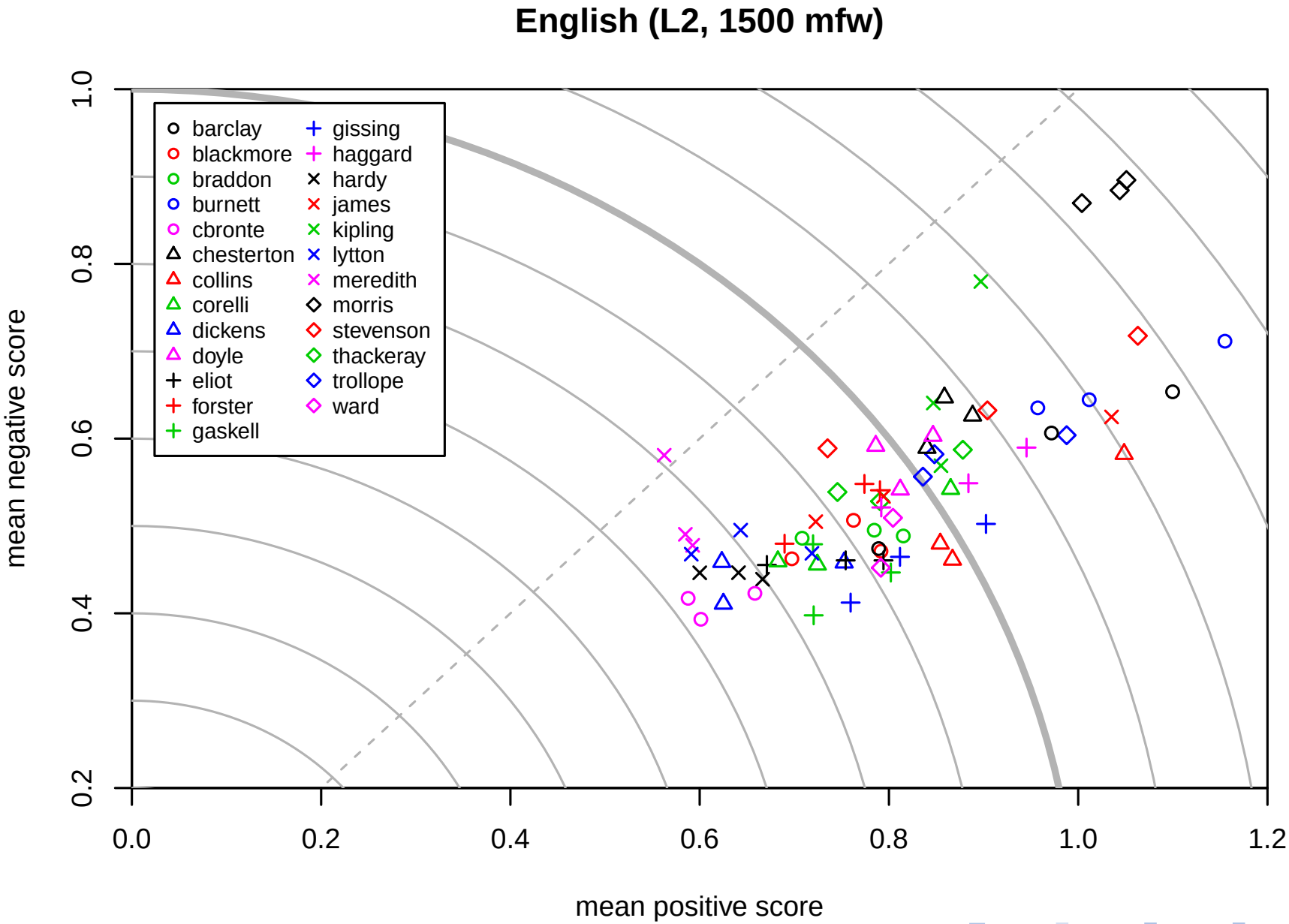
Why is vector normalization so important?



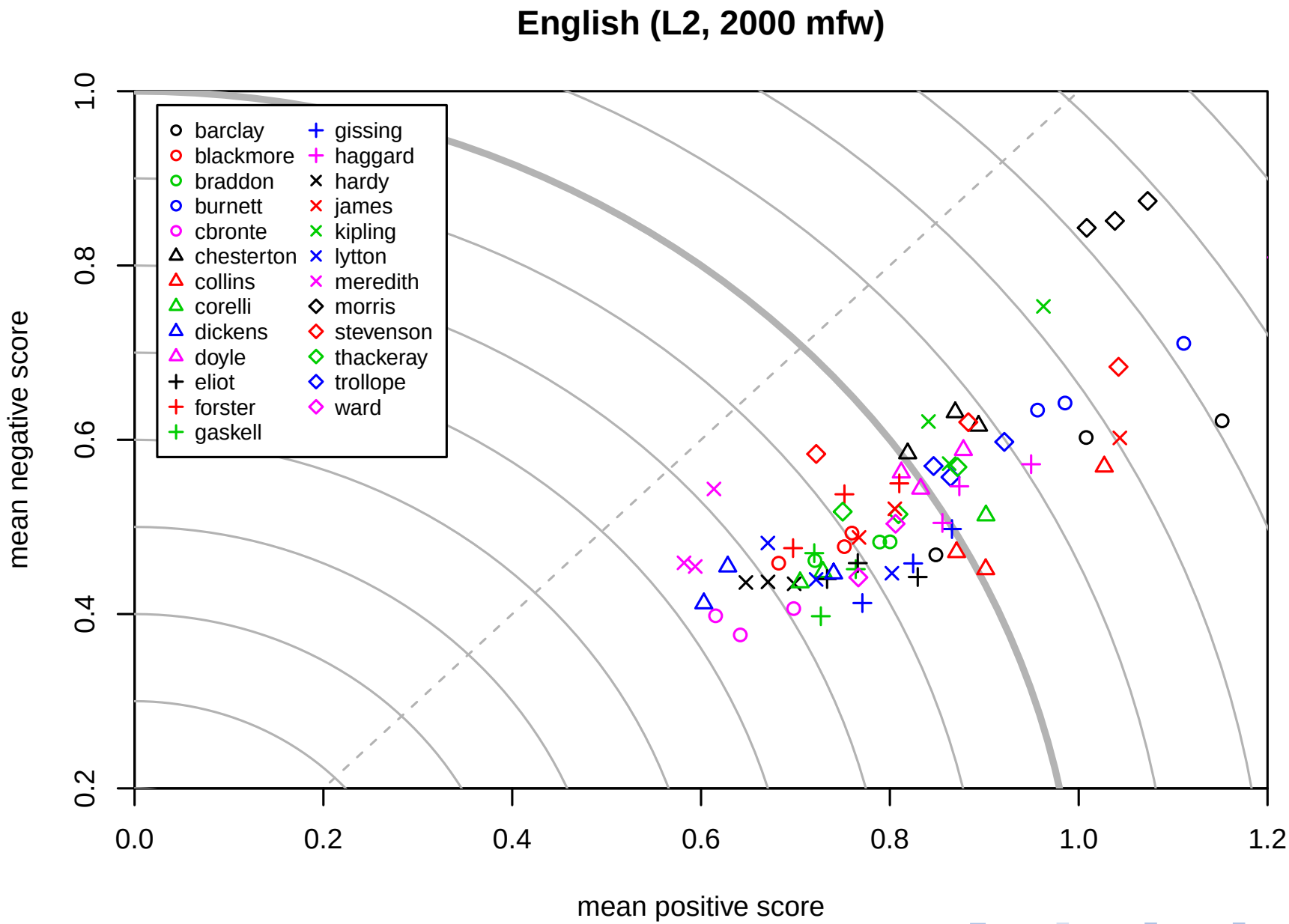
Why is vector normalization so important?



Why is vector normalization so important?

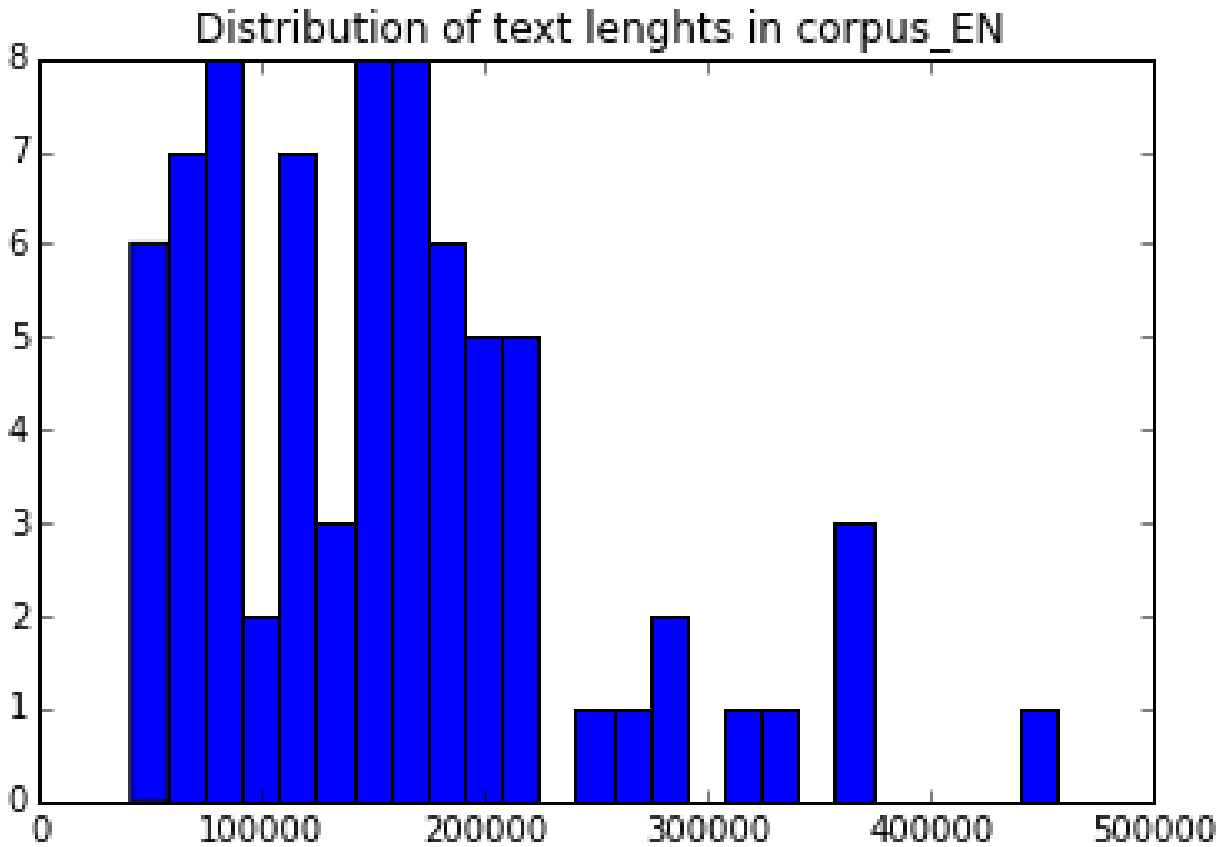


Why is vector normalization so important?



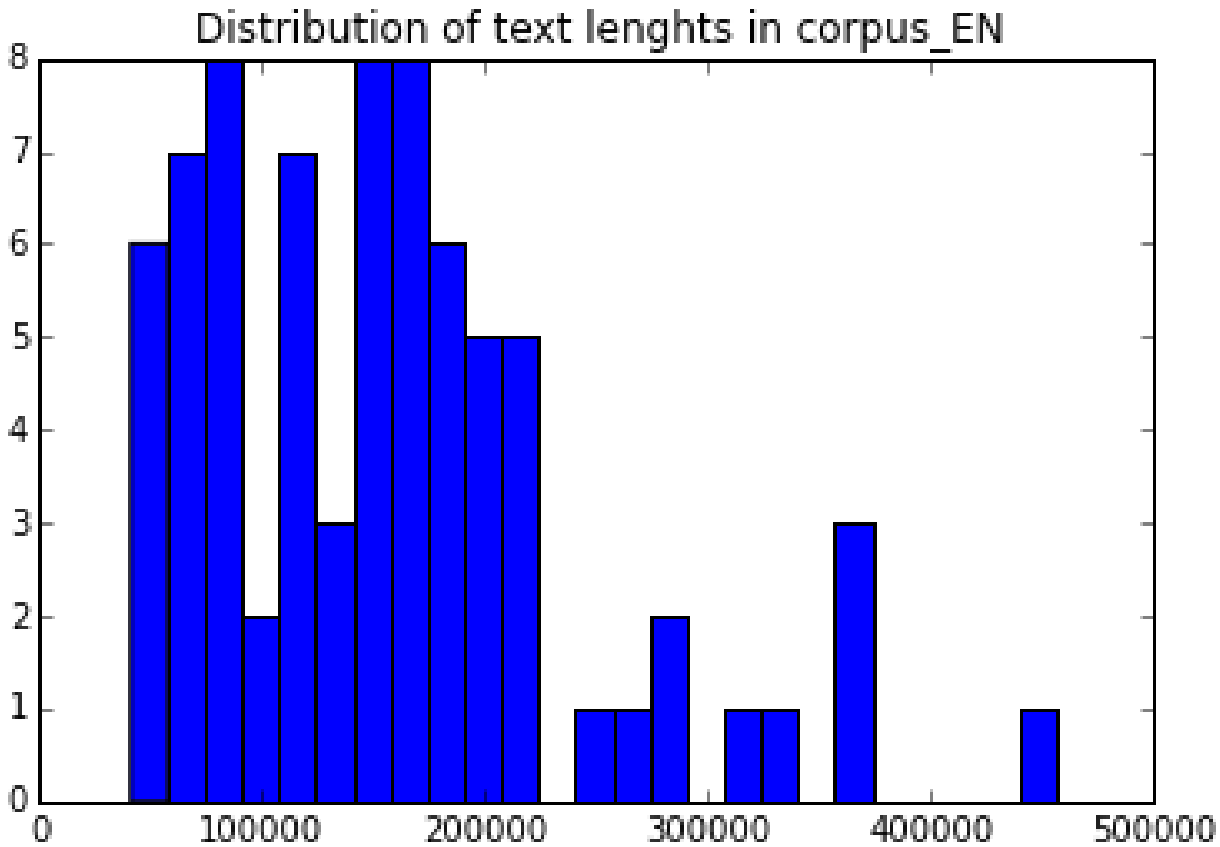
Learning curves

- All experiments carried out on complete novels so far

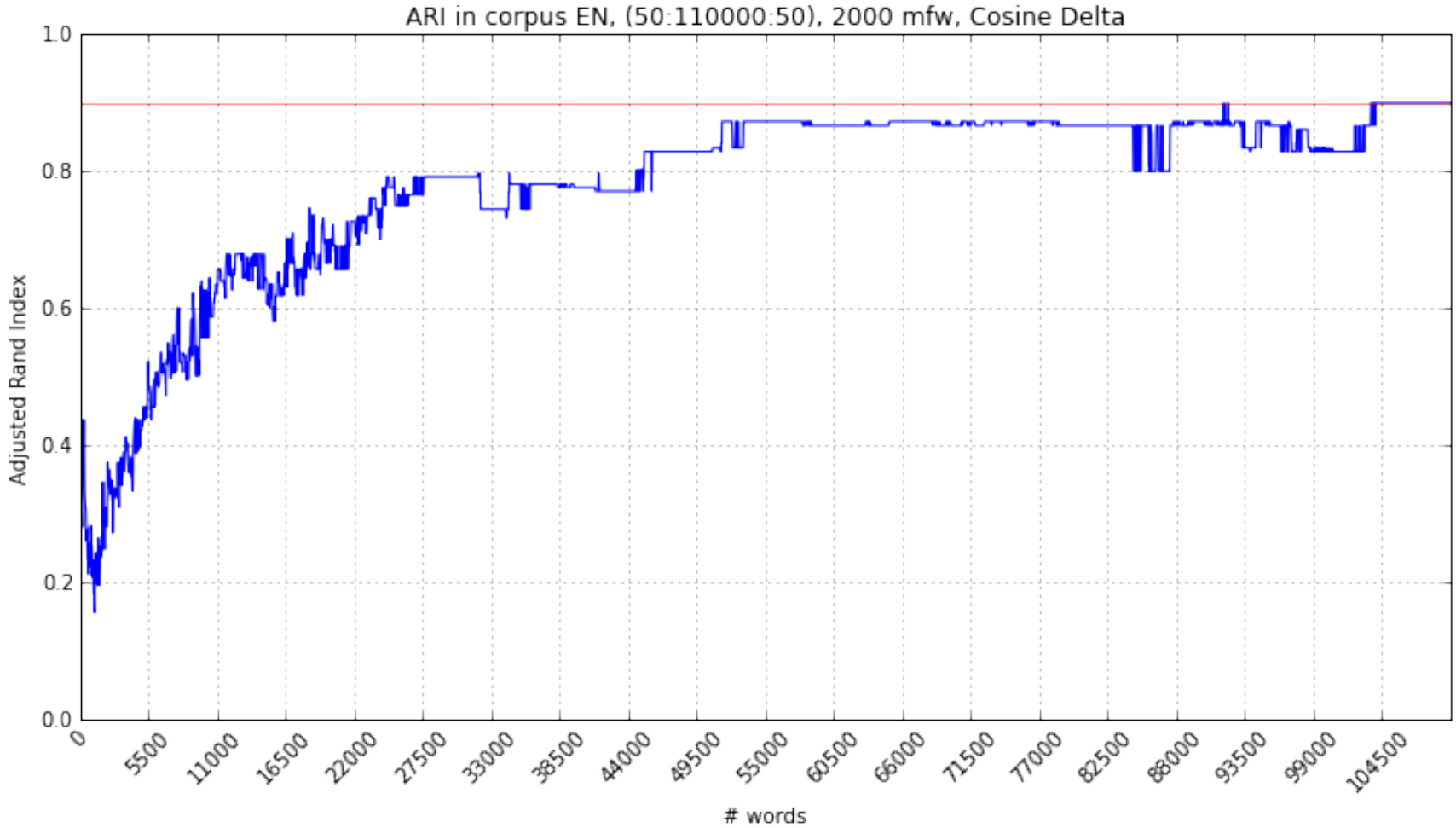


Learning curves

- All experiments carried out on complete novels so far
- Does authorship attribution also work for shorter texts?
  - ☞ experiments with Cosine Delta $\Delta_{\angle}$ and $n_w = 2000$ MFW

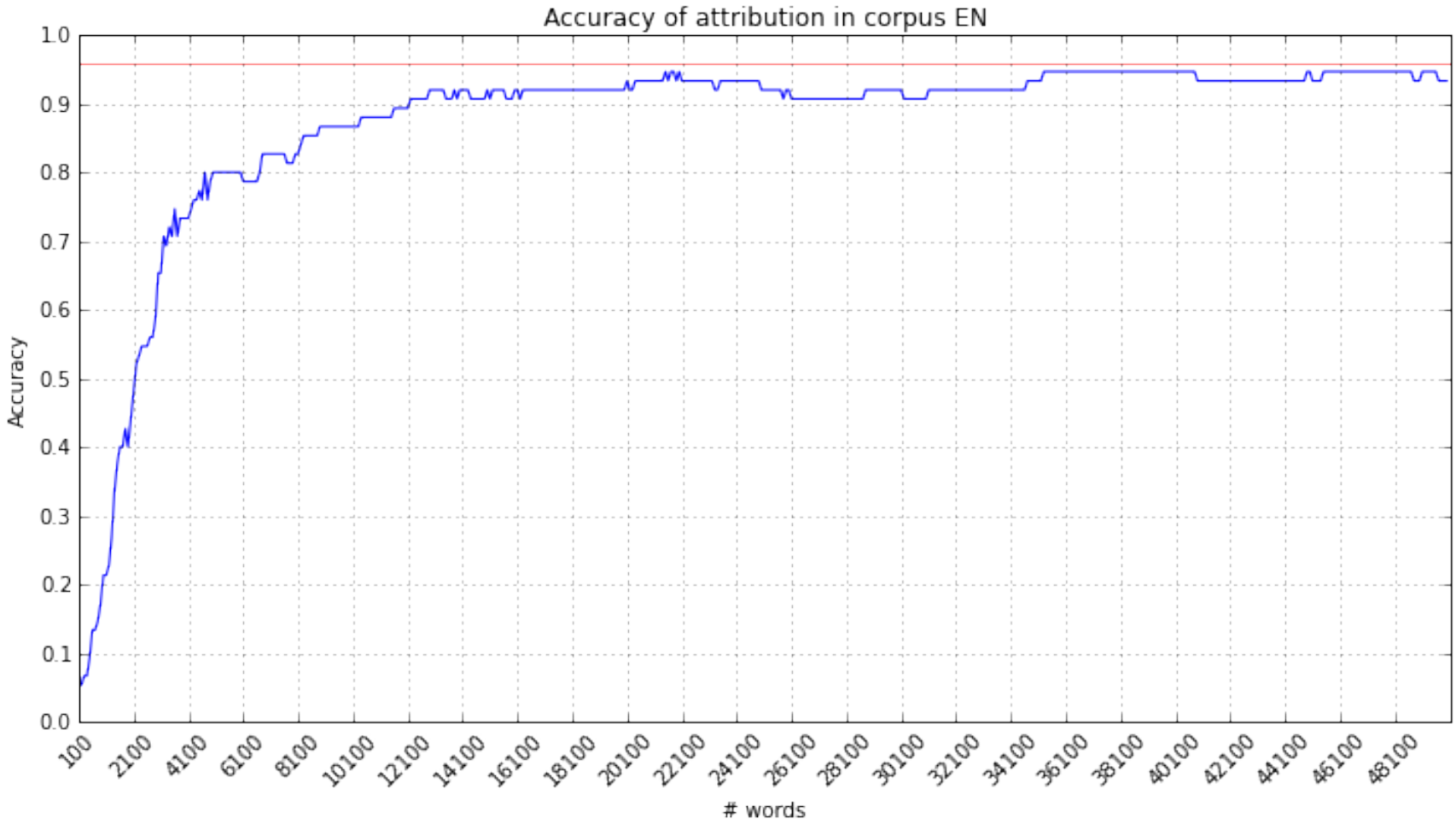


Learning curves: Clustering task



all texts shortened to specified number of word tokens

Learning curves: Classification task



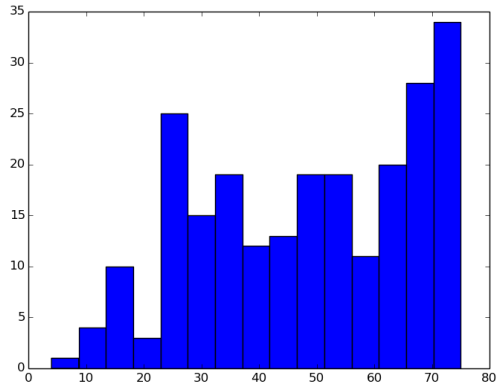
one text shortened, other texts used as training data

Finding the key words: recursive feature elimination

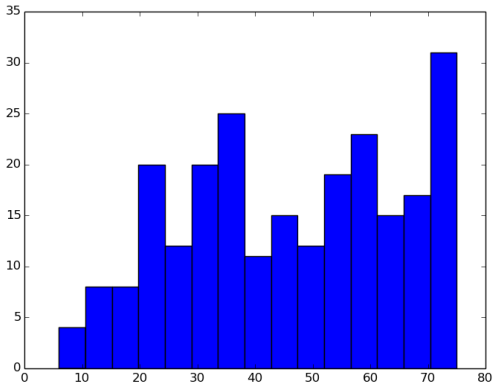
- Greedy algorithm for selection of an optimal set of features
- Procedure:
  - train linear support vector machine (SVM)
  - based on $[0, 1]$-scaled relative frequencies (not on z-scores)
  - discard k features with lowest SVM weights
- Iterative reduction of feature set
  1. all recurrent words ($df > 1$)
  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 ($k = 1$)

Automatically selected features

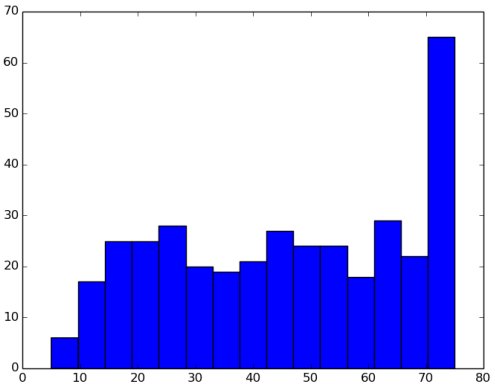
Document frequencies (df) of selected features



English ($n_w = 233$)



German ($n_w = 240$)

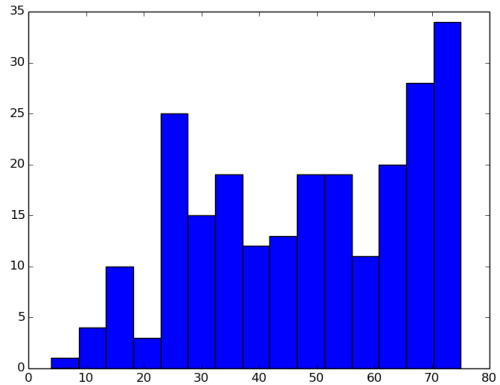


French ($n_w = 370$)

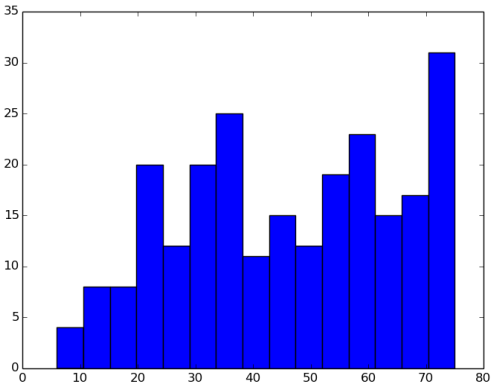
- Many function words, but also content words (➜ overtraining)

Automatically selected features

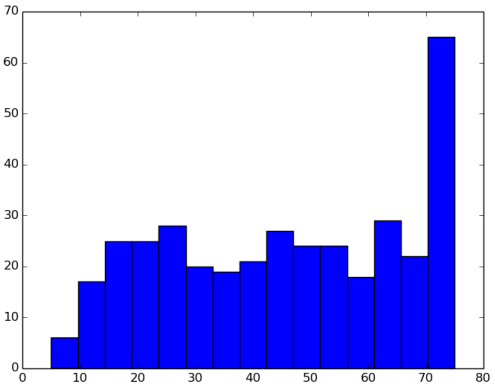
Document frequencies (df) of selected features



English ($n_w = 233$)



German ($n_w = 240$)

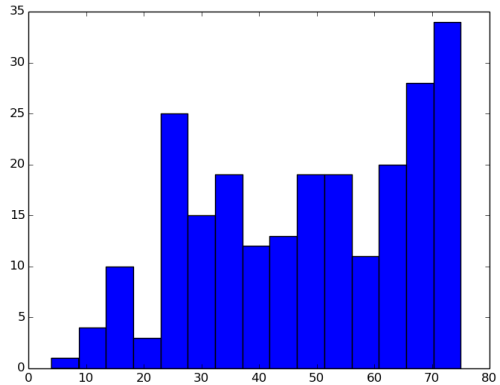


French ($n_w = 370$)

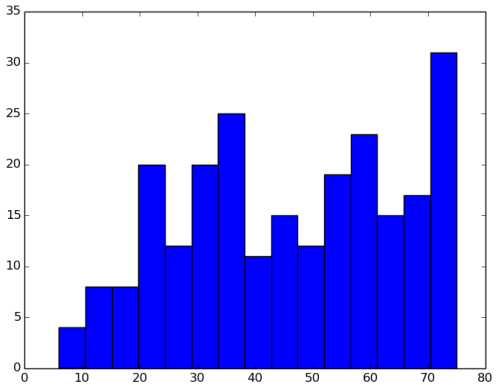
- Many function words, but also content words (➜ overtraining)
- Some text artefacts: Roman numerals (XL, XXXVII) in novels with many chapters, graphemic variation (e.g. DE *gibt* / *giebt*)

Automatically selected features

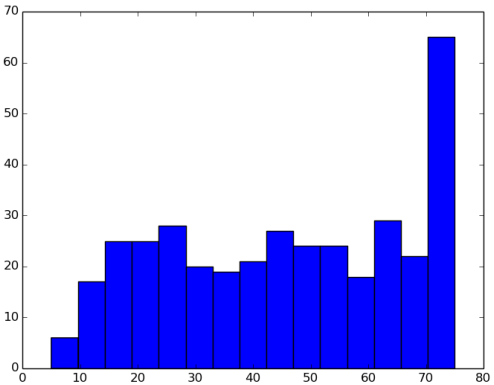
Document frequencies (df) of selected features



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* ($df = 4$)

Validation

- Validation of German features on unseen test sets

Validation

- Validation of German features on unseen test sets

- Test set A: classification
  - 71 unseen novels from 19 of the 25 authors
  - unbalanced, with singleton authors
  - Maximum Entropy classifier trained on German corpus
  - result: 97% classification accuracy

Validation

- Validation of German features on unseen test sets

- Test set A: classification
  - 71 unseen novels from 19 of the 25 authors
  - unbalanced, with singleton authors
  - Maximum Entropy classifier trained on German corpus
  - result: 97% classification accuracy

- Test set B: clustering
  - 155 unseen novels from 34 authors (6 seen, 28 unseen)
  - clustering based on Cosine Delta into 34 groups
  - result:

features	ARI
240 selected	87%
2000 MFW	83%

Conclusion & outlook

Further reading

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.

Next steps

- Consistency: do fragments of the same text cluster?
- Performance on selected parts of speech (e.g. function words)
- Pre-processing: lemmatization, stem + suffix, ...

Intepretation of Delta

- Identify features with largest contribution to Δ clustering
- Delta measures based on general stylometric features

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.