**JANNIDIS — Improving**

<5 figures; 1 table, equations>

**Improving Burrows’ Delta—An Empirical Evaluation of Text Distance Measures**

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T.

Since John Burrows first proposed Delta as a new stylometric measure (Burrows, 2002), it has become one of the most robust distance measures for authorship attribution (Juola, 2006; Stamatatos, 2009; Koppel et al., 2009). It has been shown to render very useful results in different text genres (Hoover, 2004a) and languages (Eder and Rybicki, 2013). Nowadays, Delta is widely used not the least because there is the free *stylo* package in R (Eder et al., 2013). There have been several proposals to improve Delta (Hoover, 2004b; Argamon, 2008; Eder et al., 2013; Smith and Aldridge, 2011; Kestemont and Rybicki, 2013). In the following, we report on a series of experiments to test these proposals using collections of novels in three languages. Our results will show that one of Hoover’s and one of Argamon’s measures show good results, but are outperformed in general by Burrows’ Delta and by Eder’s Delta. Most notably, the modification of Delta proposed by Smith and Aldridge shows a remarkable improvement of the results in all languages and has the advantage of providing a stable increase of performance up to a specific point, unlike the other measures, which are very sensitive to the amount of most frequent words (mfw) used. These results also allow discussion of some of the theoretical assumptions for the success of these measures, even if we are still far away from providing a ‘compelling theoretical justification’ (Hoover, 2005) for their success.

**Material and Methods**

*Text Collections*

We built three collections of novels (English, French, German), each consisting of 75 texts: 25 authors, each represented with three novels. The collection of British novels contains texts published between 1838 and 1921 coming from Project Gutenberg (www.gutenberg.org). The collection of French novels contains texts published between 1827 and 1934 originating mainly from Ebooks libres et gratuits (www.ebooksgratuits.com). The collection of German novels consists of texts from the 19th and the first half of the 20th centuries, which come from the TextGrid collection (http://www.textgrid.de/Digitale-Bibliothek).

*Text Distance Measures*

Argamon (2008) proposed three variants of Delta, each one improving on one aspect of Burrows’ Delta.

He argues that Burrows’ Delta inherently assumes that the distribution of a particular word across multiple texts follows the multivariate Laplace distribution, but it uses the standard deviation as a normalization factor, which only makes sense for a Gaussian distribution. Quadratic Delta assumes a multivariate Gaussian distribution of the words. Linear Delta, on the other hand, adjusts the normalization to take into account the calculation of spread for a Laplace distribution. Like Burrows’ Delta, both variants are based on the (implausible) assumption that the frequencies of the words are independent. The third variant, Rotated Delta, rotates the frequency differences into a space where they are maximally independent. It also assumes a Gaussian distribution. An analysis of the word frequency distribution in our English test set shows indeed that the normal distribution represents the data much better than the Laplace distribution (Figure 1a). The same is true for German high-frequency words (Figure 1b). We should therefore expect Rotated Delta to perform best, and Quadratic Delta to yield better results than Linear Delta or Burrows’ Delta. Eder’s Delta slightly increases the weights of frequent words and is meant to perform better with highly inflected languages. It should perform better for German and French than for English. Smith and Aldridge replace the Manhattan distance used by Burrows with cosine based on findings in text mining that the latter shows greater reliability with large word vectors (Smith and Aldridge, 2011). Their experiments show an impressive improvement for English texts.

We implemented these measures in Python, using the output of *stylo*, where a parallel implementation existed, as a benchmark (see the appendix for the formulas).

*Evaluating Distance Measures*

In order to provide useful performance indicators for these measures for authorship attribution, we concentrated on the question of how well a particular distance measure allows distinguishing between a situation where two compared texts are written by the same author from a situation where two texts are from different authors. Besides relying on the Adjusted Rand Index as a well-known but rather abstract measure for clustering quality (Everitt et al., 2011, 264f.), we also established a simple algorithm to count clustering errors representing the researcher’s intuition of correct clustering. In order to obtain another more subtle performance indicator independent of any clustering algorithm, we grouped the calculated distance scores into two sets: ingroup comparisons containing distances between texts actually written by the same author, and outgroup comparisons. The larger the difference between the ingroup distances and the outgroup distances, the better a distance measure is assumed to perform. After evaluating several potential performance indicators (for example, t-values, or using proportions of distribution overlap) in terms of how well they correlate with the number of clustering errors, we settled on the simple difference of z-transformed means because it showed the best correlation with the clustering error measure.

**Results and Discussion**

Most interestingly, we could not only confirm the findings of Smith and Aldridge on our English texts, but also show that Cosine Delta outperforms all other measures on our three collections (Figures 2–4). Equally important, it proves to be more stable with increasing mfw size. While Burrows’ and Eder’s Delta usually show a peak around 1,000–1,500 mfw, and then behave a bit erratically on longer word vectors, Cosine Delta reaches a plateau at 2,000 and stays there (Figure 5). As Smith and Aldridge argue, based on their very different data, that using word vectors longer than 500 words doesn’t improve the performance, we assume that this number is a function of the corpus size.

Our empirical tests did not substantiate Argamon’s theoretical arguments. Eder’s variant didn’t show consistently better results with more highly inflected languages like German and French and performed similarly to Burrows’ Delta. Both Quadratic Delta and Rotated Delta perform much worse than should be expected on theoretical grounds. And Linear Delta, although being among the top-five distance measures, seems to be an improvement over Burrows’ Delta only under special circumstances. Argamon’s modifications were based on correct assumptions about the kind of distributions Delta is working on, but nevertheless those algorithms did not perform better, something that points to the operation of factors not yet understood. The fact that those algorithms consistently perform differently in different languages and that these differences cannot, or at least only partially, be explained by the degree of inflection (Eder and Rybicki, 2013), adds to this enigma at the moment. There is almost no other algorithm in stylometry we know as much about as Delta, and yet there is still no theoretical framework that can explain its success.

**Future Work**

We tried to assemble corpora similar in genre, time, etc., but there is the real possibility that the variation we attribute to the languages is really only one between the specifics of the corpora. So it is important to test the robustness of our findings using different corpora. Another line of future investigations is the testing of more variations of Delta, using Cosine Delta as a starting point. Also, a systematic study of how the length of the mfw list determines Cosine Delta’s performance in relation to the size of the texts and the corpora could allow the automatic choice of the best parameters. And last but not least we have to analyze the connection between the performances of some variants of Delta and specifics of different languages in order to gain a deeper theoretical insight into the working of Delta in general.

\_\_\_\_\_

The python script implementing these distance measures can be found at

https://github.com/fotis007/pydelta.

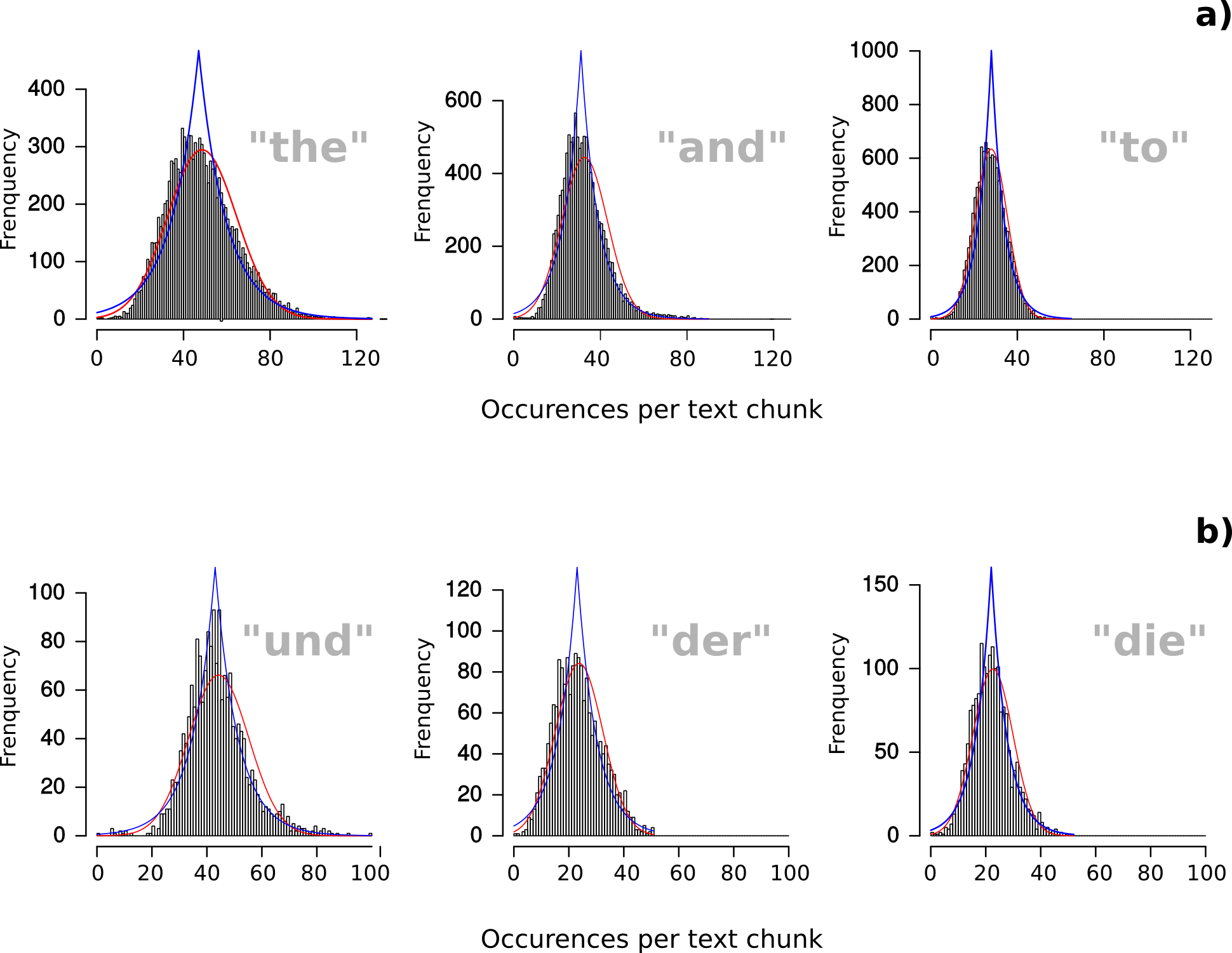


Figure 1. Empirical distribution of word frequencies compared to idealised Gauss and Laplace distributions. Indicated for the three most frequent words in the English (a) and German (b) dataset.

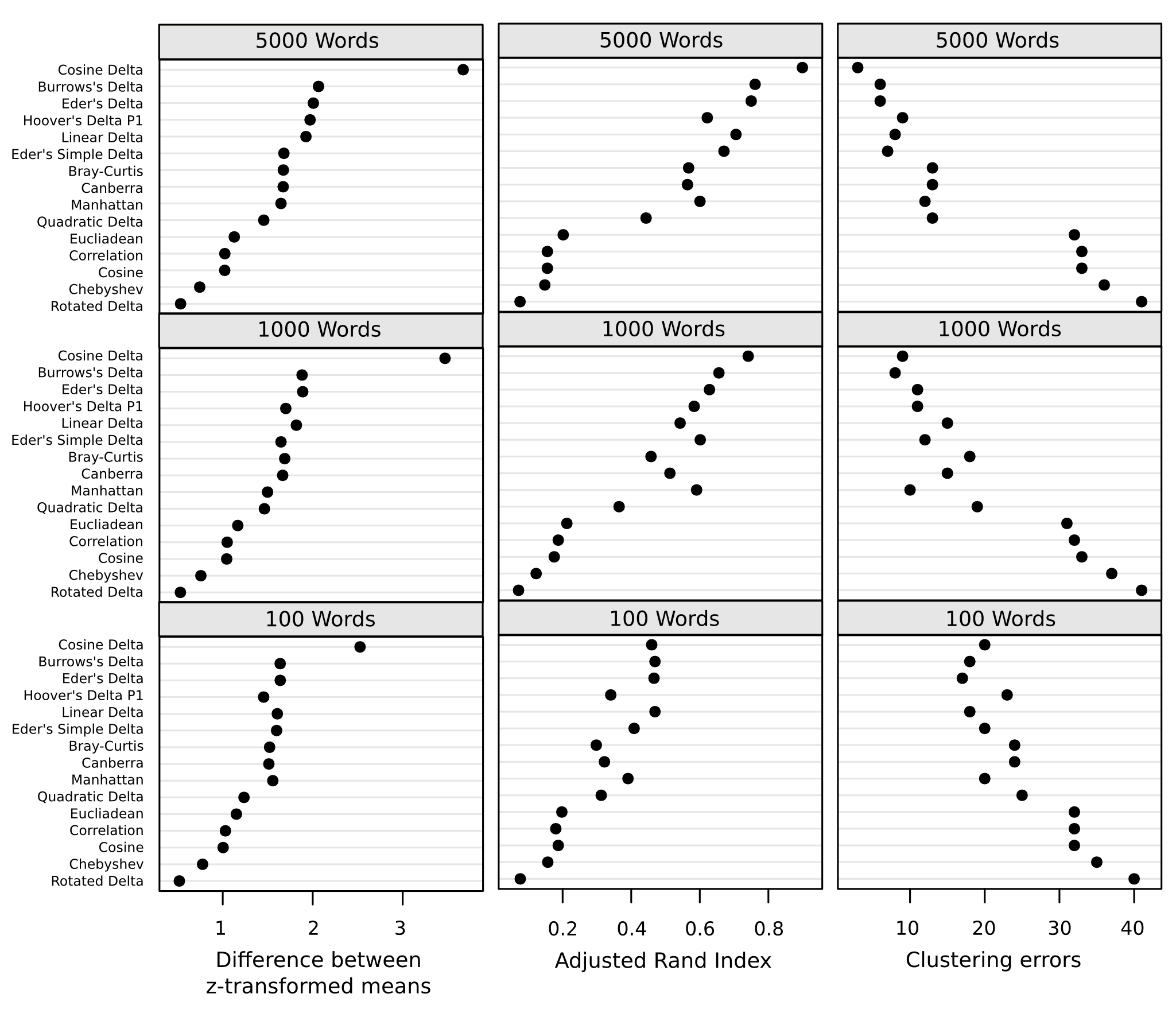


Figure 2. Performance of distance measures on English texts. Indicated in terms of both the difference between z-transformed means of ingroup (same author) and outgroup distances, as Adjusted Rand Index (higher values indicate better differentiation), and in terms of clustering errors (lower values indicate better differentiation). Distance measures are sorted according to their maximum performance in all test conditions.

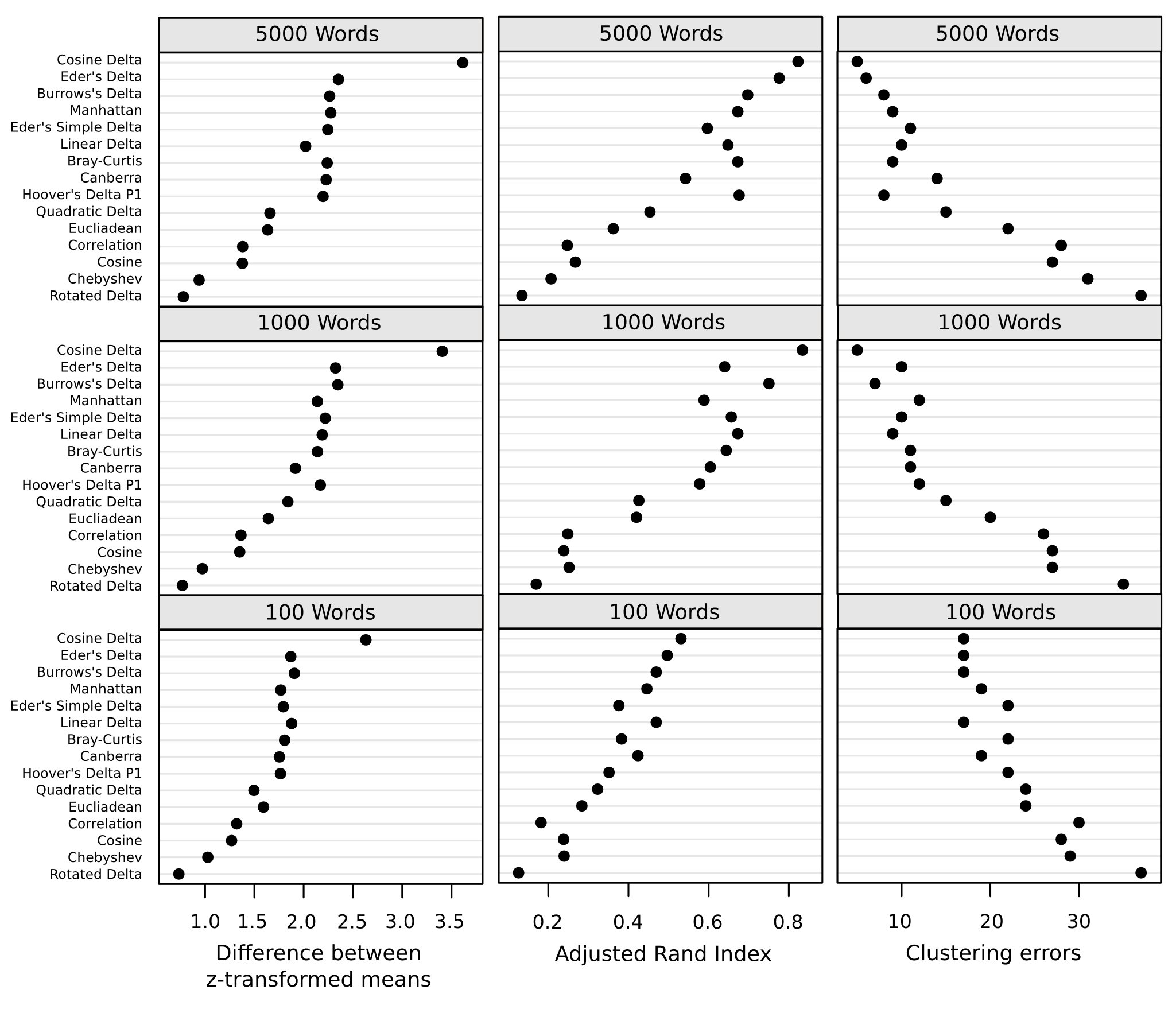


Figure 3. Performance of distance measures on French texts.For a detailed explanation, see Figure 2.

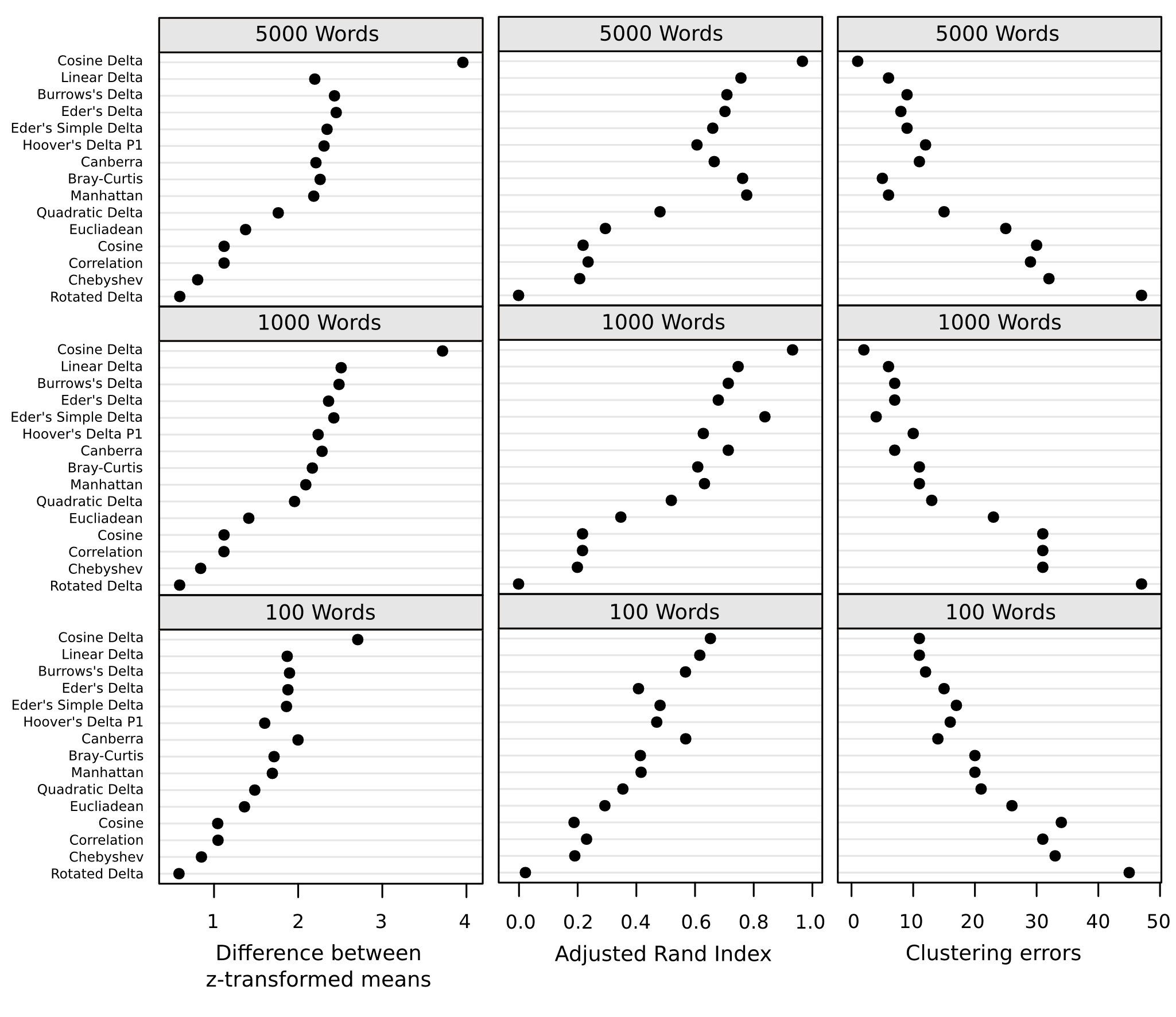


Figure 4. Performance of distance measures on German texts. For a detailed explanation, see Figure 2.

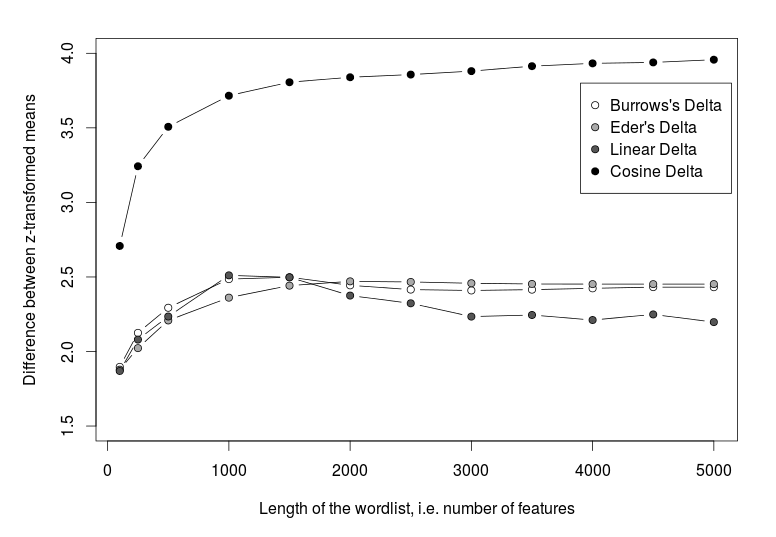


Figure 5. Difference between z-transformed means of ingroup and outgroup distances as a function of word list length. Indicated for selected delta measures on the German text collection.

**Appendix: Text Distance Measures**

Most delta measures normalize features first and then apply a basic distance function.

|  |  |  |  |
| --- | --- | --- | --- |
| **Distance Measure** | **Basic Distance Function** | **Feature Normalisation** |  |
| Burrows’ Delta | Manhattan Distance | z-score |
| Argamon’s Linear Delta | Manhattan Distance | diversity |
| Eder's Delta | Manhattan Distance | z-score · Eder's Ranking Factor: |
| Eder’s Simple Delta | Manhattan Distance | square root |
| Manhattan Distance | Manhattan Distance | – |
| Argamon’s Quadratic Delta | Euclidean Distance | z-score |
| Argamon’s Rotated Delta | Euclidean Distance | Stripped-down eigenvectors of covariance matrix |
| Euclidean Distance | Euclidean Distance | – |
| Cosine Delta | Cosine Distance | z-score |
| Cosine Distance | Cosine Distance | – |
| Correlation Distance | Cosine Distance | center |
| Hoover's Delta-P1 | (own measure) | z-score |
| Canberra distance | Canberra Distance | – |
| Bray-Curtis distance | Bray-Curtis Distance | – |
| Chebishev Distance | max abs. distance | – |
| **Basic Distance Function** | **Definition** | |
| Manhattan Distance |  | |
| Euclidean Distance |  | |
| Cosine Distance |  | |
| Canberra Distance |  | |
| Bray-Curtis Distance |  | |
| Hoover’s Delta-P1 |  | |
| z-score |  | |
| Diversity |  | |
| Eder’s Ranking Factor |  | |

**References**

**Argamon, S.** (2008). Interpreting Burrows’ Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing,* **23**(2): 131–47.

**Burrows, J.** (2002). ‘Delta’—A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing,* **17**(3): 267–87.

**Eder, M., Kestemont, M. and Rybicki, J.** (2013). Stylometry with R. *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, pp. 487–89.

**Eder, M. and Rybicki, J.** (2013). Do Birds of a Feather Really Flock Together, Or How to Choose Training Samples for Authorship Attribution. *Literary and Linguistic Computing,* **28**(2): 229–36.

**Everitt, B., et al.** (2011). *Cluster Analysis*. 5th ed. Chichester.

**Hoover, D.** (2004a). Testing Burrows’ Delta. *Literary and Linguistic Computing,* **19**(4): 453–75.

**Hoover, D.** (2004b). Delta Prime? *Literary and Linguistic Computing,* **19**(4): 477–95.

**Hoover, D.** (2005). Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method. *Proceedings of the 2005 ALLC/ACH Conference,* http://tomcat-stable.hcmc.uvic.ca:8080/ach/site/xhtml.xq?id=73.

**Juola, P.** (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval****,*****1**(3): 233–334.

**Koppel, M., Schler, J. and Argamon, S.** (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology,* **60**(1): 9–26.

**Rybicki, J. and Eder, M.** (2011). Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words? *Literary and Linguistic Computing,* **26**(3): 315–21.

**Smith, P. and Aldridge, W.** (2011). Improving Authorship Attribution: Optimizing Burrows’ Delta Method. *Journal of Quantitative Linguistics,* **18**(1): 63–88.

**Stamatatos, E.** (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology,* **60**(3): 538–56.