Towards a better understanding of Burrows's Delta in literary authorship attribution

Stefan Evert and **Thomas Proisl**
FAU Erlangen-Nürnberg
Bismarckstr. 6
91054 Erlangen, Germany
stefan.evert@fau.de
thomas.proisl@fau.de

Fotis Jannidis, Steffen Pielström,
Christof Schöch and **Thorsten Vitt**
Universität Würzburg
Am Hubland
97074 Würzburg, Germany
fotis.jannidis@uni-wuerzburg.de

Abstract

Burrows's Delta is the most established measure for stylometric difference in literary authorship attribution. Several improvements on the original Delta have been proposed. However, a recent empirical study showed that none of the proposed variants constitute a major improvement in terms of authorship attribution performance. With this paper, we try to improve our understanding of how and why these text distance measures work for authorship attribution. We evaluate the effects of standardization and vector normalization on the statistical distributions of features and the resulting text clustering quality. Furthermore, we explore supervised selection of discriminant words as a procedure for further improving authorship attribution.

1 Introduction

Authorship Attribution is a research area in quantitative text analysis concerned with attributing texts of unknown or disputed authorship to their actual author based on quantitatively measured linguistic evidence (Juola, 2006; Stamatatos, 2009; Koppel et al., 2008). Authorship attribution has applications e.g. in literary studies, history, and forensics, and uses methods from Natural Language Processing, Text Mining, and Corpus Stylistics. The fundamental assumption in authorship attribution is that individuals have idiosyncratic habits of language use, leading to a stylistic similarity of texts written by the same person. Many of these stylistic habits can be measured by assessing the relative frequencies of function words or parts of speech, vocabulary richness, and other linguistic features. This, in turn, allows using the relative similarity of the texts to each other in clustering or classification tasks and to attribute a text of unknown authorship to the most similar of a (usually closed) set of candidate authors.

One of the most crucial elements in quantitative authorship attribution methods is the distance measure used to quantify the degree of similarity between texts. A major advance in this area has been Delta, as proposed by Burrows (2002), which has proven to be a very robust measure in different genres and languages (Hoover, 2004b; Eder and Rybicki, 2013). Since 2002, a number of variants of Burrows's Delta have been proposed (Hoover, 2004a; Argamon, 2008; Smith and Aldridge, 2011; Eder et al., 2013). In a recent publication, empirical tests of authorship attribution performance for Delta as well as 13 precursors and/or variants of it have been reported (Jannidis et al., 2015). That study, using three test corpora in English, German and French, has shown that Burrows's Delta remains a strong contender, but is outperformed quite clearly by Cosine Delta as proposed by Smith and Aldridge (2011). The study has also shown that some of the theoretical arguments by Argamon (2008) do not find empirical confirmation. This means that, intriguingly, there is still no clear theoretical model which is able to explain why these various distance measures yield varying performance; we don't have a clear understanding why Burrows's Delta and Cosine Delta are so robust and reliable.

In the absence of compelling theoretical arguments, systematic empirical testing becomes paramount, and this paper proposes to continue such

Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, pages 79–88,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

investigations. Previous work has focused on feature selection either in the sense of deciding what type of feature (e.g. character, word or part-of-speech n-grams) has the best discriminatory power for authorship attribution (Forsyth and Holmes, 1996; Rogati and Yang, 2002), or in the sense of deciding which part of the list of most frequent words yields the best results (Rybicki and Eder, 2011). Other publications explored strategies of deliberately picking a very small numbers of particularly disciminative features (Cartright and Bendersky, 2008; Marsden et al., 2013). Our strategy builds on such approaches but differs from them in that we focus on word unigrams only and examine how the treatment of the input feature vector (i.e., the list of word tokens used and their frequencies) interacts with the performance of distance measures. Each distance measure implements a specific combination of standardization and/or normalization of the feature vector. In addition, the feature vector can be preprocessed in several ways before submitting it to the distance measure.

In the following, we report on a series of experiments which assess the effects of standardization and normalization, as well as of feature vector manipulation, on the performance of distance measures for authorship attribution.

Although we use attribution success as our performance indicator, our ultimate goal is not so much to optimize the results, but rather to gain a deeper understanding of the mechanisms behind distance measures. We hope that a deeper theoretical understanding will help choose the right parameters in authorship attribution cases.

2 Notation

All measures in the Delta family share the same basic procedure for measuring dissimilarities between the text documents D in a collection $\mathcal{D}$ of size $n_{\mathcal{D}}$.[1]

- Each text $D \in \mathcal{D}$ is represented by a profile of the relative frequencies $f_i(D)$ of the n_w most frequent words (mfw) $w_1, w_2, \ldots w_{n_w}$.
- The complete profile of D is given by the feature vector $\mathbf{f}(D) = (f_1(D), \ldots, f_{n_w}(D))$.
- Features are re-scaled, usually with a linear transformation, in order to adapt the weight given to each of the mfw. The most common choice is to standardize features using a z-transformation
$$z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i}$$
where μ_i is the mean of the distribution of f_i across the collection $\mathcal{D}$ and σ_i its standard deviation (s.d.). After the transformation, each feature z_i has mean $\mu = 0$ and s.d. $\sigma = 1$.
- Dissimilarities between the scaled feature vectors are computed according to some distance metric. Optionally, feature vectors may first be normalized to have length 1 under the same metric.

Different choices of a distance metric lead to various well-known variants of Delta. The original Burrows's Delta Δ_B (Burrows, 2002) corresponds to the Manhattan distance between feature vectors:

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

Quadratic Delta Δ_Q (Argamon, 2008) corresponds to the squared Euclidean distance:

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

and is fully equivalent to Euclidean distance $\sqrt{\Delta_Q}$.

Cosine Delta $\Delta_\angle$ (Smith and Aldridge, 2011) measures the angle α between two profile vectors

$$\Delta_\angle(D, D') = \alpha$$

which can be computed from the cosine similarity of $\mathbf{x} = \mathbf{z}(D)$ and $\mathbf{y} = \mathbf{z}(D')$:

$$\cos\alpha = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2}$$

where $\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n_w} x_i y_i$ is the dot product and $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n_w} x_i^2}$ denotes the length of the vector $\mathbf{x}$ according to the Euclidean norm. All three variants of Delta agree in using standardized frequencies of the most frequent words as their underlying features.

[1] The notation introduced here follows Argamon (2008) and Jannidis et al. (2015).

3 Understanding the parameters of Delta

Different versions of Delta can be obtained by setting the parameters of the general procedure outlined in Sec. 2, in particular:

- n_w, i.e. the number of words used as features in the frequency profiles;
- how these words are selected (e.g. taking the most frequent words, choosing words based on the number df of texts they occur in, etc.);
- how frequency profiles $\mathbf{f}(D)$ are scaled to feature vectors $\mathbf{z}(D)$
- whether feature vectors are normalized to unit length $\|\mathbf{z}(D)\| = 1$ (and according to which norm); and
- which distance metric is used to measure dissimilarities between feature vectors.

We focus here on three key variants of the Delta measure:

(i) the original Burrows's Delta Δ_B because it is consistently one of the best-performing Delta variants despite its simplicity and lack of a convincing mathematical motivation (Argamon, 2008);

(ii) Quadratic Delta Δ_Q because it can be derived from a probabilistic interpretation of the standardized frequency profiles (Argamon, 2008); and

(iii) Cosine Delta $\Delta_\angle$ because it achieved the best results in the evaluation study of Jannidis et al. (2015).

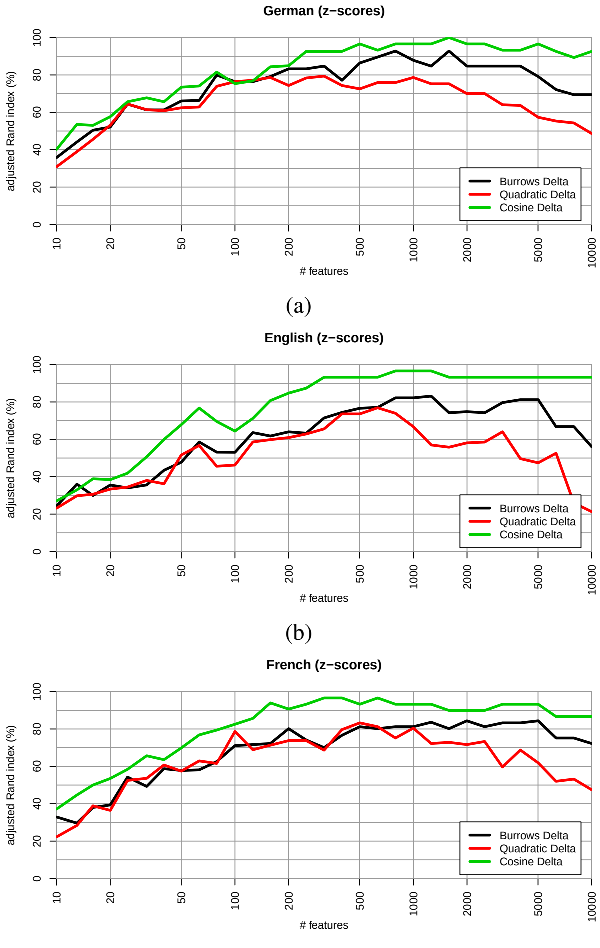
All three variants use some number n_w of mfw as features and scale them by standardization (z-transformation). At first sight, they appear to differ only with respect to the distance metric used: Manhattan distance (Δ_B), Euclidean distance (Δ_Q), or angular distance ($\Delta_\angle$).

There is a close connection between angular distance and Euclidean distance because the (squared) Euclidean norm can be expressed as a dot product $\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x}$. Therefore,

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|_2^2 &= (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \\ &= \mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y} - 2\mathbf{x}^T\mathbf{y} \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 \cos\alpha\end{aligned}$$

If the profile vectors are normalized wrt. the Euclidean norm, i.e. $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$, Euclidean distance is a monotonic function of the angle α:

$$\|\mathbf{x}-\mathbf{y}\|_2^2 = 2 - 2\cos\alpha$$



(a)

(b)

(c)

Figure 1: Clustering quality for German, English and French texts in our replication of Jannidis et al. (2015)

As a result, Δ_Q and $\Delta_\angle$ are equivalent for normalized feature vectors. The difference between Quadratic and Cosine Delta is a matter of the normalization parameter at heart; they are not based on genuinely different distance metrics.

3.1 The number of features

As a first step, we replicate the findings of Jannidis et al. (2015). Their data set is composed of three collections of novels written in English, French and German. Each collection contains 3 novels each from 25 different authors, i.e. a total of 75 texts. The collection of British novels contains texts published between 1838 and 1921 coming from Project Gutenberg.[2] The collection of French novels con-

[2] www.gutenberg.org

tains texts published between 1827 and 1934 originating mainly from Ebooks libres et gratuits.[3] The collection of German novels consists of texts from the 19th and the first half of the 20th Century which come from the TextGrid collection.[4]

Our experiments extend the previous study in three respects:

1. We use a different clustering algorithm, partitioning around medoids (Kaufman and Rousseeuw, 1990), which has proven to be very robust especially on linguistic data (Lapesa and Evert, 2014). The number of clusters is set to 25, corresponding to the number of different authors in each of the collections.
2. We evaluate clustering quality using a well-established criterion, the chance-adjusted Rand index (Hubert and Arabie, 1985), rather than cluster purity. This improves comparability with other evaluation studies.
3. Jannidis et al. (2015) consider only three arbitrarily chosen values $n_w = 100, 1000, 5\,000$. Since clustering quality does not always improve if a larger number of mfw is used, this approach draws an incomplete picture and does not show whether there is a clearly defined optimal value n_w or whether the Delta measures are robust wrt. the choice of n_w. Our evaluation systematically varies n_w from 10 to 10 000.

Fig. 1(a) shows evaluation results for the German texts; Fig. 1(b) and 1(c) show the corresponding results on English and French data. Our experiments confirm the observations of Jannidis et al. (2015):

- For a small number of mfw as features (roughly $n_w \leq 500$), Δ_B and Δ_Q achieve the same clustering quality. However, Δ_Q proves less robust if the number of features is further increased ($n_w > 500$), despite the convincing probabilistic motivation given by Argamon (2008).
- $\Delta_\angle$ consistently outperforms the other Delta measures, regardless of the choice of n_w. It is robust for values up to $n_w = 10\,000$, degrading much more slowly than Δ_B and Δ_Q.
- The clustering quality achieved by $\Delta_\angle$ is very impressive. With an adjusted Rand index above 90% for a wide range of n_w, most of the texts in the collections are correctly grouped by author.

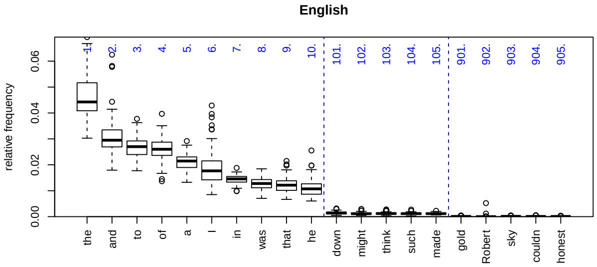


Figure 2: Distribution of relative frequencies for selected English words (numbers at the top show mfw rank)

It is obvious from these evaluation graphs that an appropriate choice of n_w plays a crucial role for Δ_Q, and to a somewhat lesser extent also for Δ_B. In all three languages, clustering quality is substantially diminished for $n_w > 5\,000$. Since there are already noticeable differences between the three collections, it has to be assumed that the optimal n_w depends on many factors – language, text type, length of the texts, quality and preprocessing of the text files (e.g. spelling normalization), etc. – and cannot be known *a priori*. It would be desirable either to re-scale the relative frequencies in a way that gives less weight to “noisy” features, or to re-rank the most frequent words by a different criterion for which a clear cut-off point can be determined.

Alternatively, more robust variants of Delta such as $\Delta_\angle$ might be used, although there is still a gradual decline, especially for the French data in Fig. 1(c). Since $\Delta_\angle$ differs from the least robust measure Δ_Q only in its implicit normalization of the feature vectors, vector normalization appears to be the key to robust authorship attribution.

3.2 Feature scaling

Burrows applied a z-transformation to the frequency profiles with the explicit intention to “treat all of these words as markers of potentially equal power” (Burrows, 2002, p. 271). Fig. 2 and 3 illustrate this intuition for some of the most frequent words in the English collection.

Without standardization, words with mfw ranks above 100 make a negligible contribution to the frequency profiles (Fig. 2). The evaluation graph in Fig. 4 confirms that Delta measures are hardly affected at all by words above mfw rank 100 if no z-

[3] www.ebooksgratuits.com

[4] www.textgrid.de/Digitale-Bibliothek

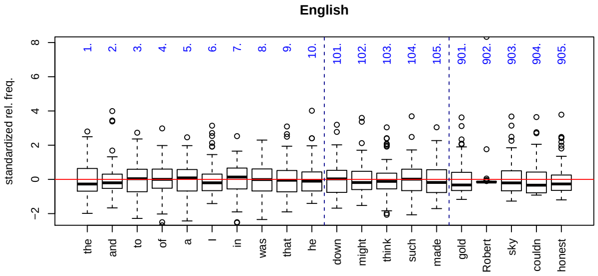


Figure 3: Distribution of standardized z-scores

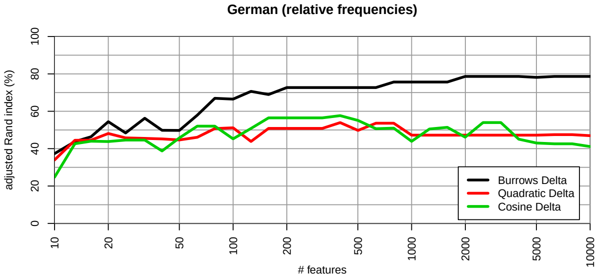


Figure 4: Clustering quality of German texts based on unscaled relative frequencies (without standardization)

transformation is applied. While results are robust with respect to n_w, the few mfw that make a noticeable contribution are not sufficient to achieve a reasonable clustering quality. After standardization, the z-scores show a similar distribution for all features (Fig. 3).

Argamon (2008) argued that standardization is only meaningful if the relative frequencies roughly follow a Gaussian distribution across the texts in a collection $\mathcal{D}$, which is indeed the case for high-frequency words (Jannidis et al., 2015). With some further assumptions, Argamon showed that $\Delta_Q(D, D')$ can be used as a test statistic for authorship attribution, with an asymptotic $\chi^2_{n_w}$ distribution under the null hypothesis that both texts are from the same author. It can also be shown that standardization gives all features equal weight in Δ_Q in a strict sense, i.e. each feature makes exactly the same average contribution to the pairwise squared Euclidean distances (and analogously for $\Delta_\angle$).

This strict interpretation of equal weight does not hold for Δ_B, so Burrows's original intention has not been fully realized. Fig. 5 displays the actual contribution made to each feature to $\Delta_B(D, D')$, i.e. to the pairwise Manhattan distances between z-transformed feature vectors $\mathbf{z}(D)$ and $\mathbf{z}(D')$. It shows that less frequent words have a moderately smaller weight than the mfw up to rank 5 000. Words that occur just in a small number of texts (their document frequency df, indicated by point colour in Fig. 5) carry a low weight regardless of their overall frequency.

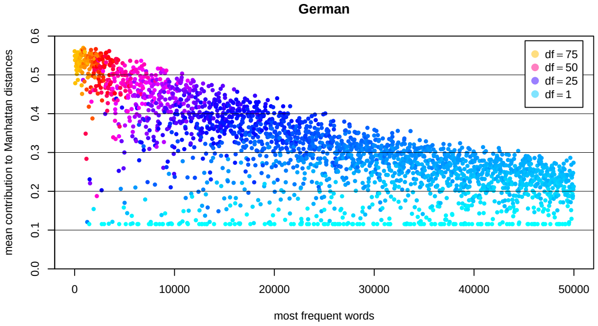


Figure 5: Average contribution of features to pairwise Δ_B distances; colour indicates document frequency (df)

Our conclusion is that Δ_B appears to be more robust than Δ_Q precisely because it gives less weight to the standardized frequencies of "noisy" words above mfw rank 5 000 in contrast to the claim made by Burrows. Moreover, it strongly demotes words that concentrate in a small number of texts, which are likely idiosyncratic expressions from a particular novel (e.g. character names) or a narrow sub-genre. It is plausible that such words are of little use for the purpose of authorship attribution.

Surprisingly, ranking the mfw by their contribution to Δ_B (so that e.g. words with $df < 10$ are never included as features) is less effective than ranking by overall frequency (not shown for space reasons). We also experimented with a number of alternative scaling methods – including the scaling suggested by Argamon (2008) for a probabilistic interpretation of Δ_B – obtaining consistently worse clustering quality than with standardization.

3.3 Vector normalization

As shown at the beginning of Sec. 3, the main difference between $\Delta_\angle$ (the best and most robust measure in our evaluation) and Δ_Q (the worst and least robust measure) lies in the normalization of feature vectors. This observation suggests that other Delta measures such as Δ_B might also benefit from vector normalization. We test this hypothesis with the evaluation

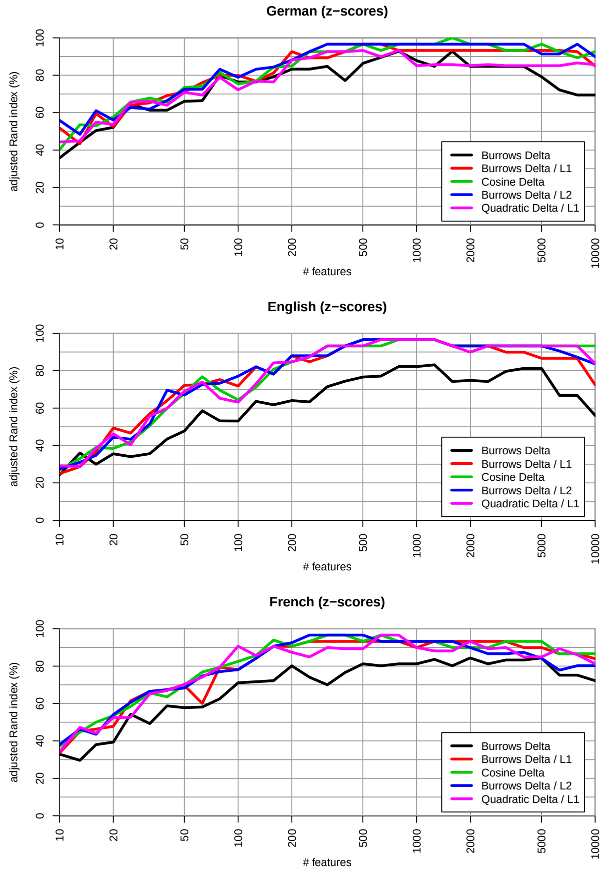


Figure 6: The effect of vector normalization on clustering quality (L2 = Euclidean norm, L1 = Manhattan norm)

shown in Fig. 6.

The quality curves for Δ_Q with Euclidean normalization are in fact identical to the curves for Cosine Delta ($\Delta_\angle$) and are not shown separately here. Δ_B is also improved substantially by vector normalization (contrast the black curve with the red and blue ones), resulting in clustering quality equal to $\Delta_\angle$, although Δ_B might be slightly less robust for $n_w > 5\,000$. Interestingly, it seems to make little difference whether an appropriate normalization is used (L1 for Δ_B and L2 for Δ_Q) or not (vice versa).

Our tentative explanation for these findings is as follows. We conjecture that authorial style is primarily reflected by the pattern of positive and negative deviations z_i of word frequencies from the "norm", i.e. the average frequency across the text collection. This characteristic pattern is not expressed to the same degree in all texts by a given author, leading to differences in the average magnitude of the values z_i and hence the length $\|\mathbf{z}(D)\|$ of the feature vectors. If this is indeed the case, vector normalization makes the stylistic pattern of each author stand out more clearly because it equalizes the average magnitude of the z_i.

Fig. 7 visualizes the Euclidean length of feature vectors for texts written by different German authors. In the situation depicted by the left panel ($n_w = 150$) normalization has no substantial effect, whereas in the situation depicted by the right panel ($n_w = 5\,000$) unnormalized Δ_Q performs much worse than normalized $\Delta_\angle$ (cf. Fig. 1(a)).

Each point in the plots represents the feature vector $\mathbf{z}(D)$ of one text. The distance from the origin indicates its Euclidean (L2) norm $\|\mathbf{z}(D)\|_2$, relative to the average vector length $\sqrt{n_w}$. All points on a circle thus correspond to feature vectors of the same Euclidean length. The angular position of a text shows the relative contribution of positive features ($z_i > 0$, i.e. words used with above-average frequency) and negative features ($z_i < 0$, words used with below-average frequency) as a rough indicator of its stylistic pattern. Texts below the dashed diagonal thus have more (or larger) positive deviations z_i, texts above the diagonal have more (or larger) negative deviations. In both panels, some authors are characterized quite well by vector length and the balance of positive vs. negative deviations. For other authors, however, one of the texts shows a much larger deviation from the norm than the other two, i.e. larger Euclidean length (Freytag and Spielhagen in the right panel). Similar patterns can be observed for the Manhattan (L1) norm as well as among the English and French novels. In such cases, normalization reduces the distances between texts from the same author and thus improves clustering quality.

Fig. 7 also reveals a plausible explanation for the poor evaluation results of Δ_Q as n_w is increased. Because of the skewed distribution of z_i for lower-frequency words (see Fig. 3), the contribution of positive values to the Euclidean norm outweighs the negative values (but this is not the case for the Manhattan norm and Δ_B). Therefore, all points in the right panel are below the diagonal and their stylistic profiles become increasingly similar. Differences in vector length between texts from the same author have a stronger influence on Δ_Q distances in this situation, resulting in many clustering errors.

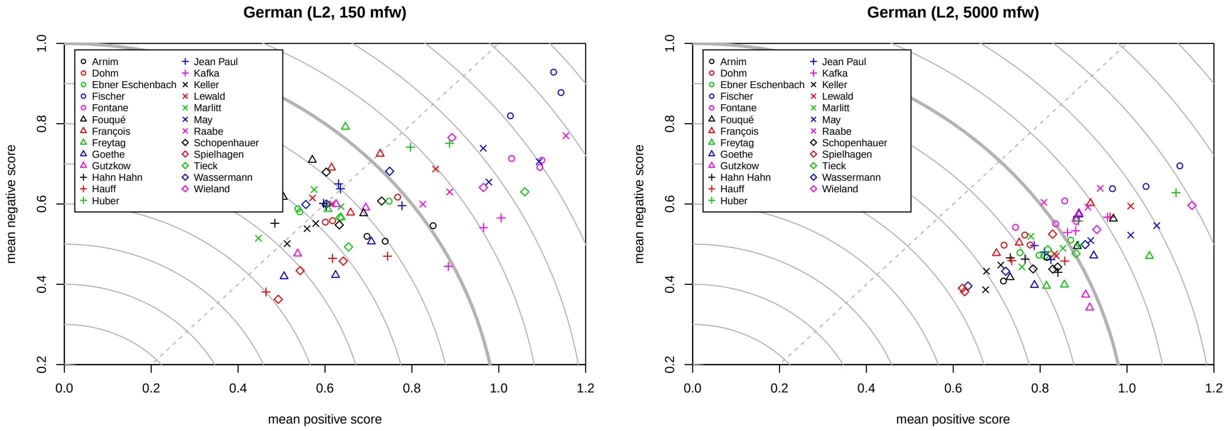


Figure 7: Euclidean length of unnormalized vectors for texts in the German corpus. Coordinates of the points indicate the average contribution of positive and negative features to the total length of the vector.

	English	French	German
nr. of features	246	381	234
SVC accuracy	0.99 (±0.04)	1.00 (±0.00)	1.00 (±0.00)
MaxEnt accuracy	1.00 (±0.00)	1.00 (±0.00)	1.00 (±0.00)
Cosine Delta ARI	0.966	1.000	1.000

Table 1: Results for cross-validation and clustering experiments

4 Feature selection as a contribution to explanation

In this section, we explore another strategy for obtaining an optimal set of features. Instead of using a threshold on (document) frequencies of words for feature selection, we systematically identify a set of discriminant words by using the method of recursive feature elimination. The resulting feature set is much smaller and not only works well in a machine learning setting, but also outperforms the most-frequent-words approach when clustering a test corpus of mainly unseen authors.

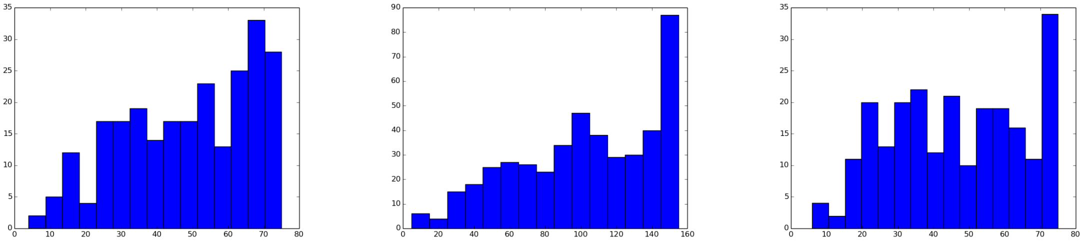
4.1 Recursive feature elimination

Recursive feature elimination is a greedy algorithm that relies on a ranking of features and on each step selects only the top ranked features, pruning the remaining features. For our feature elimination experiments we rely on a Support Vector Classifier (SVC) with linear kernel for feature ranking.[5] During training, an SVC assigns weights to the individual features, with greater absolute weights indicating more important features. We can use the weight magnitude as ranking criterion and perform recursive feature elimination by repeating the following three steps:

1. Train the classifier, i. e. assign weights to the features.
2. Rank the features according to the absolute value of their weights.
3. Prune the n lowest ranking features.

Since it is more efficient to remove several features at a time (with the possible disadvantage of introducing a slight degradation in classification performance) and we are starting with a few hundreds of thousands of features and are aiming for a much smaller set, we first reduce the number of features to 500 in three stages. First we reduce the number of features to 50 000 by recursively pruning the 10 000 lowest ranking features, then we reduce those 50 000 features to 5 000 features by pruning 1 000 features at a time and finally we reduce the number of fea-

[5] The scikit-learn implementation of Support Vector Machines we used is based on libsvm and supports multiclass classification via a one-vs.-one scheme. We used it with default parameter settings.



(a) English

(b) French

(c) German

Figure 8: Distributions of document frequencies for selected features

	unscaled full fs	rescaled full fs	selected fs
SVC accuracy	0.91 (±0.03)	0.57 (±0.13)	0.84 (±0.14)
MaxEnt accuracy	0.95 (±0.03)	0.95 (±0.03)	0.90 (±0.08)
Cosine Delta ARI	0.835	0.835	0.871

Table 2: Evaluation results for the selected features on the second additional test set compared to all features (for Cosine Delta clustering, 2 000 mfw are used in this case)

tures to 500 by pruning the 100 lowest ranking features at a time.

Once we have reduced the number of features to 500, we try to find an optimal number of features by pruning only one feature at a time, doing a stratified threefold cross-validation on the data after each pruning step to test classification accuracy.

Since Support Vector Machines are not scale-invariant, we rescaled each feature to [0, 1] during preprocessing. Simply rescaling the data should work better than standardization because it preserves sparsity (the standardized feature matrix is dense, replacing every zero by a small negative value). As an additional preprocessing step we removed all words with a document frequency of 1.

4.2 Examination and validation

The recursive feature elimination process can choose from an abundance of features, and therefore it is no surprise that it is able to find a subset of features that yields perfect results for both classification (accuracy was determined using stratified threefold cross-validation and is given as the mean plus/minus two standard deviations) and clustering using $\Delta_{\angle}$.[6] Cf. Table 1 for an overview of the results.

Figures 8(a)-8(c) show the distribution of the document frequencies of those features. For all corpora there are some highly specific features that occur only in a fraction of the texts, but most selected features have a rather high document frequency.

The features which turn out to be maximally distinctive of authors show a number of interesting patterns. For example, they are not limited to function words. This is a relevant finding, because it is often assumed that function words are the best indicators of authorship. However, content words may be more prone to overfitting than function words. Also, in the English and French collections, a small number of roman numerals are included (such as “XL” or “XXXVVII”), which may be characteristic of novels with an unusually high number of chapters. This, in turn, may in fact be characteristic of certain authors. Finally, in the German collection, a certain number of words show historical orthographic variants (such as “Heimath” or “giebt”). These are most likely artifacts of the corpus rather than actual stylistic characteristics of certain authors.

Perfect cross-validation and clustering results suggest that there may be severe overfitting. In order to verify how well the set of selected features performs on unseen data, we used two additional evaluation data sets:

1. An unbalanced set of 71 additional unseen nov-

[6] We used agglomerative clustering with complete linkage.

els by 19 authors from the German collection;

2. An unbalanced set of 155 unseen novels by 34 authors, with at least 3 novels per author (6 authors are also in the original collection).

For the first test set, we trained an SVC and a Maximum Entropy classifier (MaxEnt) on the original German corpus using the set of 234 selected features and evaluated classifier accuracy on the test set. Both SVC and MaxEnt achieved 0.97 accuracy on that test set, indicating that the selected features are not overfit to the specific novels in the training corpus but generalize very well to other works from the same authors. Since this test set includes singletons (a single novel by an author), cross-validation and clustering experiments cannot sensibly be conducted here.

For the second test set, we evaluated classification accuracy with stratified threefold cross-validation using only the set of 234 selected features. We also clustered the texts using $\Delta_\angle$ based on the same features. To have a point of reference, we furthermore evaluated classification and clustering using the full feature set, once using relative frequencies and once using rescaled relative frequencies. For the clustering we used the 2 000 mfw as features, which our experiments in Section 3.1 showed to be a robust and nearly optimal number. The results are summarized in Table 2.[7]

Comparing evaluation results for the 234 selected features from the original corpus with the full rescaled feature set, we see a considerable increase in SVC accuracy (due to the smaller number of features),[8] a small decrease in MaxEnt accuracy and an increase in clustering quality, indicating that the selected features are not overfitted to the training data and generalize fairly well to texts from other authors. Nevertheless, the difference in clustering accuracy between the first and the second test set indicates that these features are author-dependent to some extent.

5 Conclusion

The results presented here shed some light on the properties of Burrows's Delta and related text distance measures, as well as the contributions of the underlying features and their statistical distribution. Using the most frequent words as features and standardizing them with a z-transformation (Burrows, 2002) proves to be better than many alternative strategies (Sec. 3.2). However, the number n_w of features remains a critical factor (Sec. 3.1) for which no good strategy is available. Vector normalization is revealed as the key factor behind the success of Cosine Delta. It also improves Burrows's Delta and makes all measures robust wrt. the choice of n_w (Sec. 3.3). In Sec. 4 we showed that supervised feature selection may be a viable approach to further improve authorship attribution and determine a suitable value for n_w in a principled manner.

Although we are still not able to explain in full how Burrows's Delta and its variants are able to distinguish so well between texts of different authorship, why the choices made by Burrows (use of mfw, z-scores, and Manhattan distance) are better than many alternatives with better mathematical justification, and why vector normalization yields excellent and robust authorship attribution regardless of the distance metric used, the present results constitute an important step towards answering these questions.

References

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

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

Marc-Allen Cartright and Michael Bendersky. 2008. Towards scalable data-driven authorship attribution. *Center for Intelligent Information Retrieval*.

Maciej Eder and Jan Rybicki. 2013. Do birds of a feather really flock together, or how to choose training sam-

[7] Clustering results on the unscaled and rescaled full feature sets are identical because of the z-transformation involved in cosine delta.

[8] While Support Vector Machines are supposed to be effective for data sets with more features than training samples, they don't deal very well with huge feature sets that exceed the training samples by several orders of magnitude. For this reason, the poor performance of SVC on the full features set was to be expected. It is surprising, however, that SVC performs much better on unscaled features. We believe this to be a lucky coincidence: The SVC optimization criterion prefers high-frequency words that require smaller feature weights; they also happen to be the most informative and robust features in this case.

ples for authorship attribution. *Literary and Linguistic Computing*, 28(2):229–236, June.

Maciej Eder, Mike Kestemont, and Jan Rybicki. 2013. Stylometry with R: a suite of tools. In *Digital Humanities 2013: Conference Abstracts*, pages 487–489, Lincoln. University of Nebraska.

R. S. Forsyth and D. I. Holmes. 1996. Feature-finding for text classification. *Literary and Linguistic Computing*, 11(4):163–174, December.

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

David L. Hoover. 2004b. Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4):453 –475, November.

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

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

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

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

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

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

John Marsden, David Budden, Hugh Craig, and Pablo Moscato. 2013. Language individuation and marker words: Shakespeare and his maxwell's demon. *PloS one*, 8(6):e66813.

Monica Rogati and Yiming Yang. 2002. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 659–661, New York, NY, USA. ACM.

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

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

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