Stylometry & authorship attribution
(Juola 2006; Koppelet 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 letter? (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

- Often proxy task for identification of stylometric fingerprint

Stylometry & authorship attribution
(Juola 2006; Koppelet al. 2008; Stamatatos 2009)

- Authorship attribution as classification task
  - closed set of candidate authors for unknown 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

Outline

Literary authorship attribution

- Burrows’s Delta

Empirical evaluation

Understanding Delta

Statistical significance
A (very) brief history of authorship 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 (∆B)
(Burrows 2002)

- Frequencies of 100–5,000 most frequent words (MFW)
  form a “fingerprint” of an author’s style
  
  \[ f_i(D) \]

  relative frequency:

  \[
  f(Madding Crowd) = (0.051, 0.029, 0.026, 0.027, 0.016, 0.016, 0.014, 0.011, 0.008, 0.010, \ldots)
  
  f(Tess of the d’Urbervilles) = (0.053, 0.027, 0.027, 0.028, 0.020, 0.013, 0.015, 0.014, 0.012, 0.009, 0.018, \ldots)
  
  f(Oliver Twist) = (0.055, 0.032, 0.024, 0.023, 0.022, 0.012, 0.014, 0.011, 0.013, 0.005, \ldots)
  
- standardised z-scores:
  \[ z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i} \]

  \[
  z(Madding Crowd) = (0.53, -0.23, -0.32, 0.20, 1.66, -0.37, 1.04, 0.52, -0.44, -0.92, 0.03, \ldots)
  
  z(Tess of the d’Urbervilles) = (-0.75, -0.48, -0.08, 0.51, -0.24, -0.87, -0.60, -0.41, -0.14, -0.47, 1.39, \ldots)
  
  z(Oliver Twist) = (1.05, -1.15, -0.71, -0.56, 0.37, -1.03, -0.06, -0.74, -0.48, 0.94, \ldots)

- “It shows you didn’t care a bit about me, and were ready to desert - I only saw her for a day or two.”
  “It was like the girl had been in her grave for a few days, when the interposd 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
Distance metrics for texts

- Need to quantify similarity of “fingerprints” of two texts, i.e.
- Vectors of z-scores $z(D_1), z(D_2) \in \mathbb{R}^{n_w}$
  - $n_w =$ 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)
  \[
  \Delta_B(D, D') = \| z(D) - z(D') \|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D')|
  \]
- Quadratic Delta = Euclidean distance (Argamon 2008)
  \[
  \Delta_Q(D, D') = \| z(D) - z(D') \|_2^2 = \sum_{i=1}^{n_w} (z_i(D) - z_i(D'))^2
  \]
- Cosine Delta = angular distance (Smith and Aldridge 2011)
  \[
  \cos \Delta_C(D, D') = \frac{\sum_{i=1}^{n_w} z_i(D) \cdot z_i(D')}{{\| z(D) \|}_2 \cdot {\| z(D') \|}_2}
  \]

Hierarchical clustering based on $\Delta$

(Evert et al. 2017, Fig. 3)

Hierarchical clustering based on $\Delta$

(Büttner et al. 2017, Fig. 10)
Hierarchical clustering based on \( \Delta \)
(Rybicki and Heydel 2013, Fig. 1)

Outline

- Literary authorship attribution
- Burrows’s Delta
- Empirical evaluation
- Understanding Delta
- Statistical significance

Research questions

1. Can \( \Delta \) accurately identify the author of a text?
2. How many MFW should we use?
3. What is the effect of other parameters (e.g. \( \Delta_B \) vs. \( \Delta_Q \))?
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 \( \Delta \) to short texts?
7. How robust is \( \Delta \) against “noise” in the data (e.g. OCR errors)?
8. Why does \( \Delta \) work so well?
9. Far too rarely: Are the results statistically significant?

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 \( \Delta \), \( \approx \) purity
- Evert et al. (2017) – focus on \( n_w \) and feature scaling, ARI
- Büttner et al. (2017) – combination with further data sets
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

![Silhouette profile](image1)

![Silhouette widths](image2)

![Parameter](image3)
Parameter: Number $n_w$ of MFW

English (z-scores)

French (z-scores)

Parameter: Distance metric

▶ Generalised Minkowski Delta: $\Delta_B = \Delta_1, \Delta_Q = \Delta_2$
▶ $\Delta_p$ becomes less “robust” with increasing $p \in [0, \infty]$

$$\Delta_p(D_1, D_2) = \|z(D_1) - z(D_2)\|_p = \left( \sum_{i=1}^{n_w} |z_i(D_1) - z_i(D_2)|^p \right)^{1/p}$$
Parameter: Clustering method

Parameter: Feature selection

Parameter: Skipping the MFW
(Rybicki and Eder 2011, Fig. 1+2)

Parameter: Character n-grams instead of MFW

<table>
<thead>
<tr>
<th>n</th>
<th>all characters</th>
<th>only alpha-numeric</th>
<th>within words</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.9226</td>
<td>0.8510</td>
<td>0.8515</td>
</tr>
<tr>
<td>3</td>
<td>0.9534</td>
<td>0.9276</td>
<td>0.9242</td>
</tr>
<tr>
<td>4</td>
<td>0.9473</td>
<td>0.9299</td>
<td>0.9329</td>
</tr>
<tr>
<td>5</td>
<td>0.9407</td>
<td>0.9250</td>
<td>0.9320</td>
</tr>
<tr>
<td>6</td>
<td>0.9349</td>
<td>0.9078</td>
<td>0.9296</td>
</tr>
<tr>
<td>7</td>
<td>0.9432</td>
<td>0.9043</td>
<td>0.9220</td>
</tr>
</tbody>
</table>

ARI averaged over different $n_w$
(German novels)

65 English novels
32 epic poems in English

attribution accuracy of 1-NN classifier

prefer words with approximately normal distribution
**Parameter: Character n-grams instead of MFW**

Robustness against noise

(Eder 2013, Fig. 1+3)

Robustness against noise

(Büttner et al. 2017, Fig. 18+21)

**Delta and text length**

(Eder 2015, Fig. 1+4)

simulated noise in MHG poetry, $\Delta$ on 200–800 MFW

texts shortened by random sampling / excerpt, $\Delta$ on 200 MFW
Delta and text length

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)

English novels, all texts shortened

English novels, only disputed text shortened

German novels: are fewer MFW better for short texts?
A crucial riddle...

**Minkowsky** $\Delta_p$ becomes less “robust” for increasing $p \in [0, \infty]$.

**Vector length normalization** $\|z\|_p = 1$ (usually $p = 1, 2$).

The effect of vector length normalization

**Minkowsky** $\Delta_p$ becomes less “robust” for increasing $p \in [0, \infty]$

**Vector length normalization** $\|z\|_p = 1$ (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

- 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 ($|z| > 1$) H1
- ternarisation: $-1$ (underuse) / 0 (neutral) / $+1$ (overuse) H2
- optional: ternarisation only for lower-frequency words

Parameter: Transformation of z-scores

German (z-scores clamped to range $[-2, 2]$)

German (ternarized z-scores)

German (ternarized, skipping 200 MFW)

ternarisation (equal categories) H2

Parameter: Transformation of z-scores

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

ternarisation (equal categories, skipping first 200 MFW)
Transformation of z-scores and $\Delta_p$.

![Graph showing adjusted Rand index (%)](image)

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?

- Many function words, but also content words (overtraining?)
- Some text artefacts: Roman numerals ($x_l$, $xxxvii$) in novels with many chapters, graphemic variation (e.g. DE gibt / gibt)
- 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?

▶ Sampling variation
  ▶ How much statistical uncertainty in the evaluation scores?
  ▶ Often in terms of confidence interval

▶ Rarely addressed in literary authorship attribution

Results: Sampling variation

(Proisl et al. 2018)

![Box plots showing sampling variation](image)

Setting B: Sampling different sets of authors

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

▶ Setting B
  ▶ Sample different sets of authors
  ▶ 5,000 random samples of 25 authors × 3 novels

Results: Statistical significance

(Proisl et al. 2018)

![Box plots showing statistical significance](image)

Setting B: Sampling different sets of authors – paired tests
Results: Implications
(Proisl et al. 2018)

Uncertainty of MFW evaluation experiments (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

German (z-scores, 1−NN classifier w/ cross-validation)

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 $\Delta$ on character & word n-grams
- Can we integrate $\Delta$ and n-gram tracing?

References I


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


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/fqy046.

References II


References III


