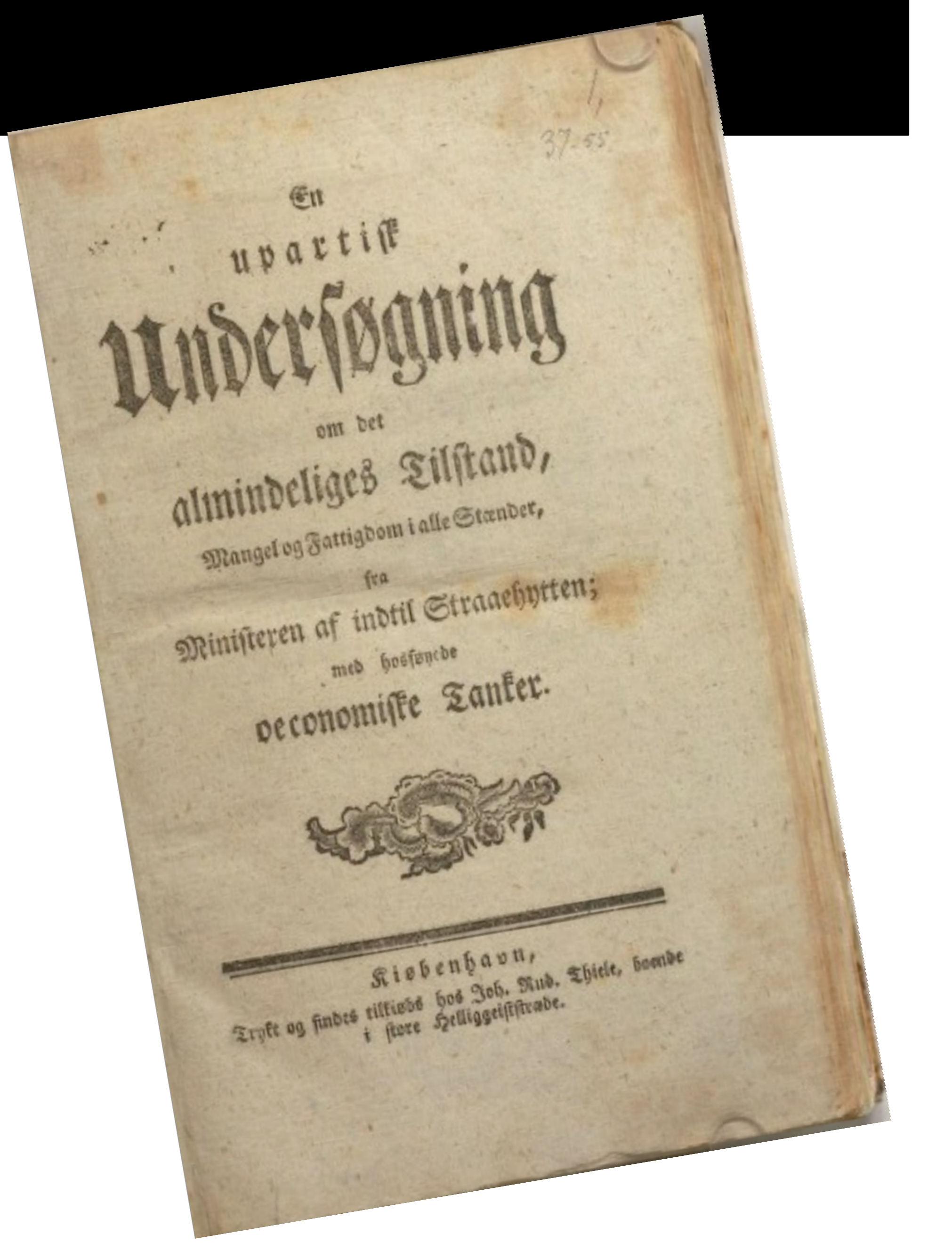
# Towards Authorship Attribution in the *Trykkefrihedsskrifter*: A Stylistic Analysis of the Danish Freedom of the Press Writings' Main Writers

Florian Meier



En
upartiſk
Underſøgning
om det
almindeliges Tilſtand,
Mangel og Fattigdom i alle Stænder,
fra
Miniſteren af indtil Straaehytten;
med hosføiede
oeconomiſke Tanker.

Kiøbenhavn,
Trykt og findes tilkiøbs hos Joh. Rud. Thiele, boende
i ſtore Helliggeiſtſtræde.

37-55

## 1. Context and problem

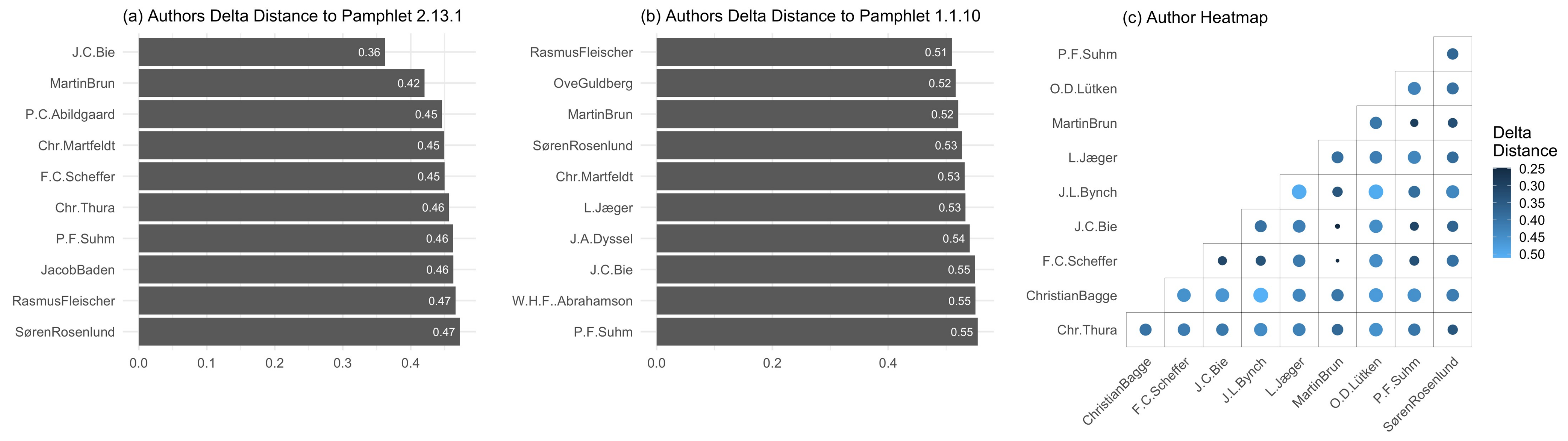
- The Danish Freedom of the Press writings (*Trykkefrihedsskrifter*), is a collection of pamphlets published and collected during the 1770s in the kingdom of Denmark-Norway.
- To better understand the collection and study idea generation and knowledge diffusion in that period, knowing who wrote the texts is of high importance.
- Authorship attribution is usually done on a comparative basis and performed as (multinominal) classification problem
- Close to 50% of the pamphlets are of unknown authorship which makes selecting candidates for machine-learning-based experiments challenging
- This poster presents three steps that can be considered in the candidate selection process

## 2. Measures of vocabulary richness

- Lexical measures of vocabulary richness can give good insights into a corpus and it's authors
- Brun wrote 54 books which – compared to others – are quite short. Jæger has the highest avg. word count per book.
- None of the other measures (share of big words, TTR, Herdan C) show a striking difference between authors.

| Author | Number of Books | Avg. Book Word Count | Avg. Token Length | Avg. Type Token Ratio | Avg. Herdan C |
|---|---|---|---|---|---|
| MartinBrun | 54 | 2323.98 | 4.53 | 0.50 | 0.91 |
| J.C.Bie | 16 | 4213.31 | 4.69 | 0.47 | 0.91 |
| J.L.Bynch | 16 | 6113.00 | 4.68 | 0.45 | 0.90 |
| P.F.Suhm | 14 | 7794.50 | 4.68 | 0.43 | 0.90 |
| SørenRosenlund | 14 | 6202.64 | 4.48 | 0.39 | 0.89 |
| ChristianBagge | 11 | 2596.64 | 4.95 | 0.48 | 0.90 |
| F.C.Scheffer | 9 | 3620.22 | 4.67 | 0.46 | 0.90 |
| Chr.Thura | 6 | 11856.83 | 4.76 | 0.38 | 0.89 |
| L.Jæger | 6 | 12752.00 | 4.71 | 0.31 | 0.88 |
| O.D.Lütken | 6 | 10584.33 | 5.05 | 0.36 | 0.89 |

## 3. Delta distance

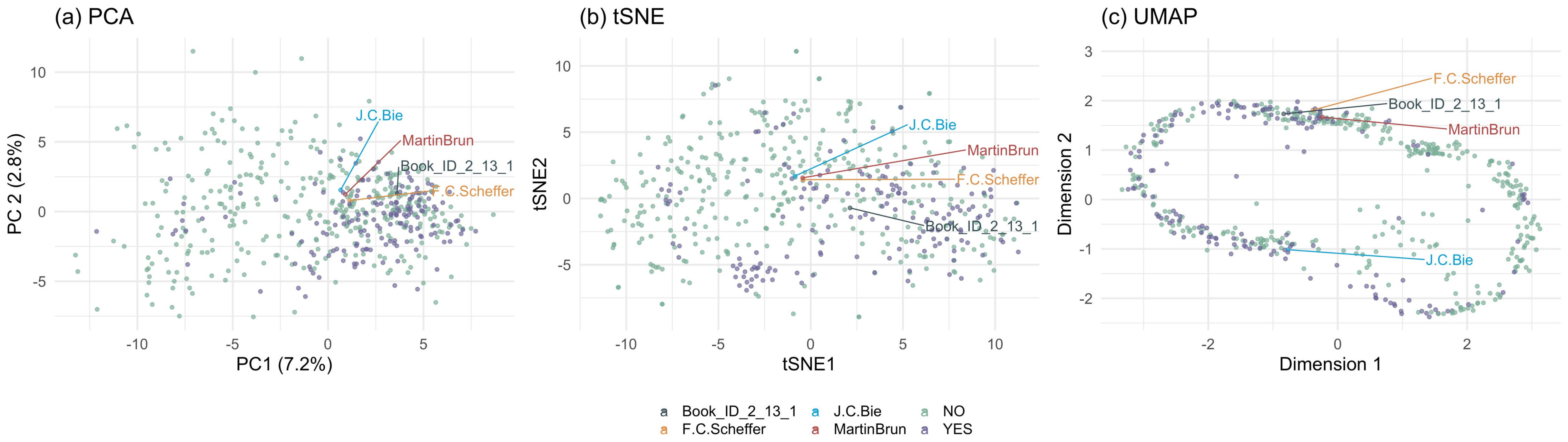


Low distance between author profiles can give insights into which authors might be difficult to distinguish Bie, Brun and Scheffer are stylistically close

- Burrow's Delta distance is a measure that builds on the assumption that authors use (function) words unconsciously, thus their normalized frequency inherently reflects their writing style (300 MFW)
- Low delta distance between author profiles and texts can be an indicator of authorship
- Multiple authors can be close, which makes the selection process difficult

## 4. Dimensionality reduction

UMAP is not a good choice as it does not support the intuition that Bie, Brun and Scheffer have similar styles



- Dimensionality reduction techniques and their possibilities for creating 2D visualisations of n-dimensional feature vectors can support the process and give further evidence for picking appropriate candidates in ML-based AA experiments
- tSNE is to be preferred over PCA due to the lack of variance explained by PC1/PC2

Dr. Florian Meier
Aalborg University Copenhagen
Department of Communication and Psychology
A.C. Meyers Vænge 15
2450 Copenhagen SV, Denmark

fmeier@ikp.aau.dk
@meier_flo
http://personprofil.aau.dk/142274

AALBORG UNIVERSITY
DENMARK