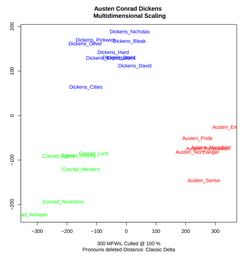
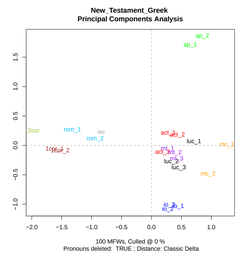
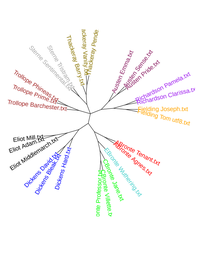
Maciej Eder Jan Rybicki
Pedagogical University of Kraków

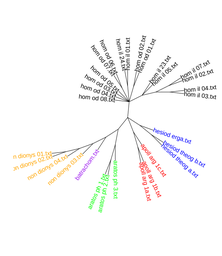
Stylometry with R

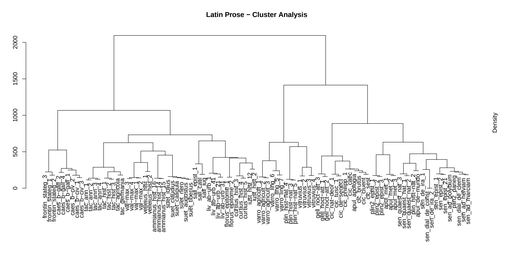
Digital Humanities 2011
Stanford, CA, June 19–22

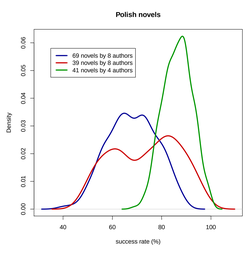


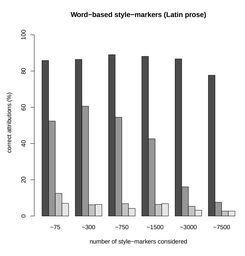


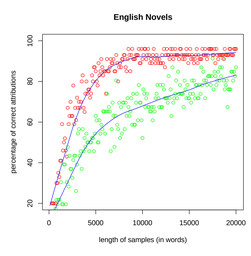


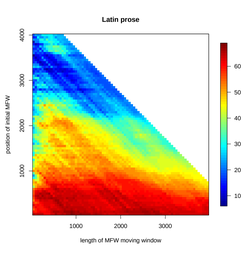












Abstract

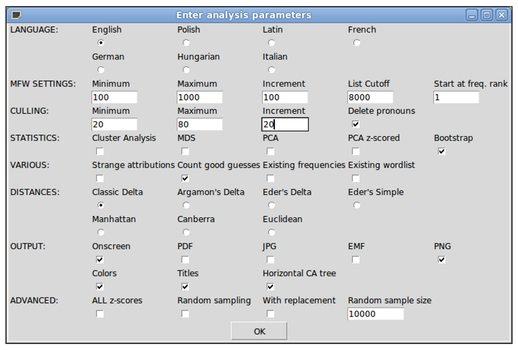
This poster describes an all-in-one script for performing various analyses in computational stylistics. The script is written in R programing language, and supports a number of nearest neighbor classification methods used in stylometry.

Features

- free, open-source (GPL licensed), and cross-platform
- supplemented with a Tcl/Tk graphic user interface
- adjustable to particular purposes
- suitable for large-scale experiments (thousands of iterations)
- fast!

How it works

1. uploads texts from `primary_set` and `secondary_set`
2. creates a list of all words used in all texts studied
   - optionally, a custom list of words can be used
3. calculates frequencies of words in the individual texts...
4. ... and places them in a huge matrix
   - optionally, external matrices can be used (e.g., from Excel)
5. performs normalization, such as z-scores (if applicable)
6. reduces the size of the matrix to the desired number of most-frequent words (MFWs); optionally, performs a further reduction:
   - deletion of personal pronouns (available for 7 languages)
   - "culling", i.e. removal of too-characteristic words
7. calculates a multidimensional distance for each pair of texts, using a specified distance measure (one of 7 available)
8. builds a matrix of all the distances between the texts
9. performs a chosen nearest neighbor classification:
   - Delta (and writes output to file)
   - Cluster Analysis
   - Multidimensional Scaling
   - Principal Components Analysis
10. generates graphs (if applicable)
11. optionally, performs the above calculations for other sets of parameters (different ranges of MFWs and culling)
12. optionally, plots a bootstrap consensus tree



Corpus preparation

Colors on graphs are assigned according to filenames: the sequence of letters before "_" (underscore) is assumed to be the label of the author (genre, etc.). This is case sensitive. A sample working directory could contain:

1. `primary_set` (subdirectory)
   - ABronte_Tenant.txt
   - Austen_Northanger.txt
   - Conrad_Nostromo.txt
   - ...
2. `secondary_set` (subdirectory)
   - ABronte_Agnes.txt
   - Austen_Emma.txt
   - Austen_Pride.txt
   - Conrad_Lord.txt
   - Dickens_Pickwick.txt
   - ...
3. `delta_test_0-4-0.r` (R script)

Usage

In an active R shell, type the following code:

- `setwd("/path/to/your/corpus/")`
- `source("delta_test_0-4-0.r")`

Alternatively, you can use a batch file and just double-click the script file icon.

Distance measures

- Classic Delta:
$$\delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i(A) - f_i(B)}{\sigma_i}\right|$$
- Argamon's Delta:
$$\delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\sqrt{f_i(A)^2 - f_i(B)^2}}{\sigma_i}\right|$$
- Eder's Delta:
$$\delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{f_i(A) - f_i(B)}{\sigma_i}\right| \times \frac{n - n_i + 1}{n}\right)$$
- Eder's Simple:
$$\delta_{(AB)} = \sum_{i=1}^{n}\left|\sqrt{f_i(A)} - \sqrt{f_i(B)}\right|$$
- Manhattan:
$$\delta_{(AB)} = \sum_{i=1}^{n}|f_i(A) - f_i(B)|$$
- Canberra:
$$\delta_{(AB)} = \sum_{i=1}^{n}\frac{|f_i(A) - f_i(B)|}{|f_i(A)| + |f_i(B)|}$$
- Euclidean:
$$\delta_{(AB)} = \sum_{i=1}^{n}\sqrt{|f_i(A)^2 - f_i(B)^2|}$$

Applications so far

- an attempt to measure the behaviour of Delta at a variety of intervals of the word frequency rank lists in a variety of languages (Rybicki and Eder 2011)
- a study of attribution accuracy dependence on studied texts' sizes (Eder 2010)
- a study of a reliable choice of training samples (Eder and Rybicki 2011)
- a multi-language study of translational style (Rybicki 2011)
- an experiment measuring the effectiveness of different word- and letter-based style-markers (Eder 2011)
- a series of MA projects in authorial and/or translational attribution (Pedagogical University of Kraków, 2010–2011)

Documentation

Opening the script in any text editor (e.g. Notepad++), one has an insight into the source code, but also to the authors' comments. Especially the initial options and configurable variables are commented in detail. A more comprehensive manual is pending.

Contact us

The script can be downloaded from here:

- `https://sites.google.com/site/computationalstylistics/`

Contact with the authors:

- Maciej Eder <maciejeder@gmail.com>
- Jan Rybicki <jkrybicki@gmail.com>

References

- Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing* 23 (2), 131–47.
- Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Burrows, J. (2002). "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17 (3), 267–87.
- Gries, S. Th. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York and London: Routledge.
- Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing* 19 (4), 453–71.