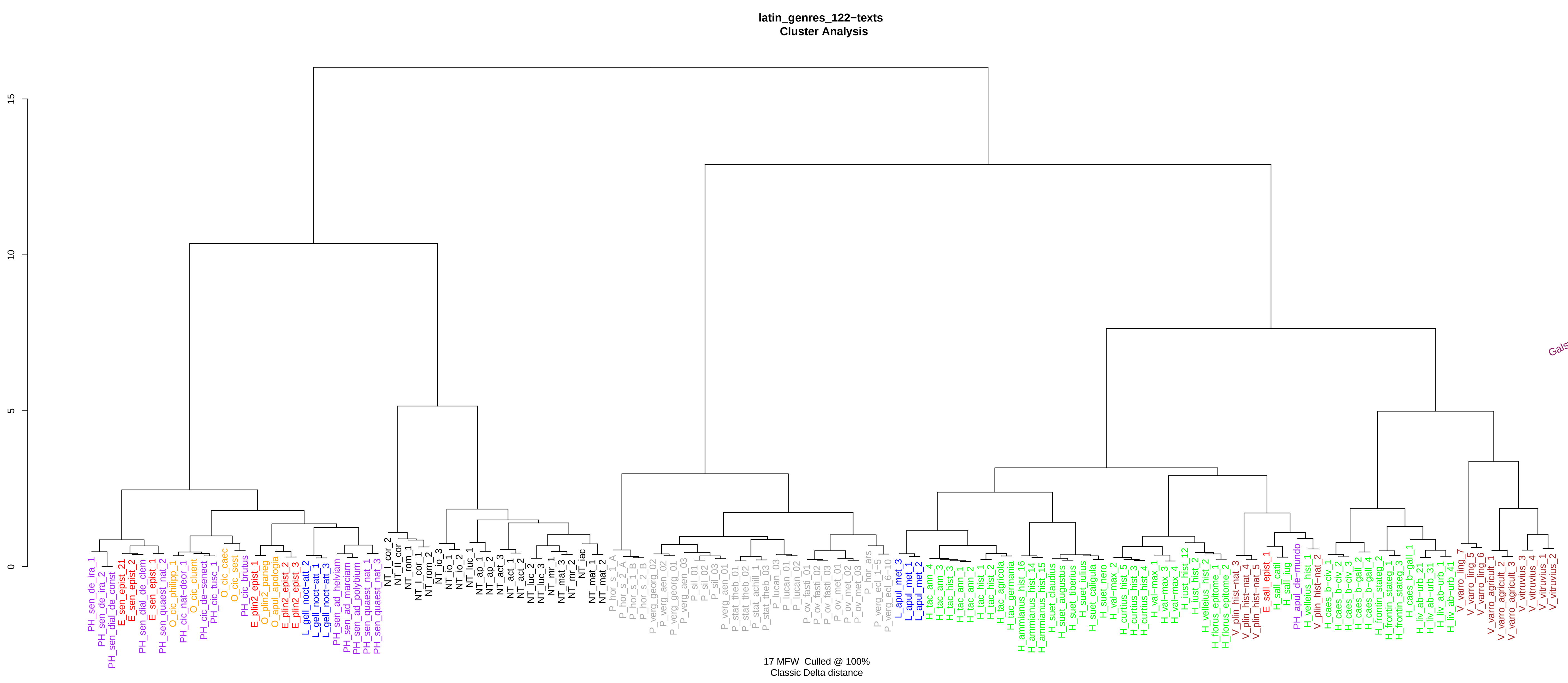
Maciej Eder [1, 4, 5] Mike Kestemont [2, 5] Jan Rybicki [3, 5]

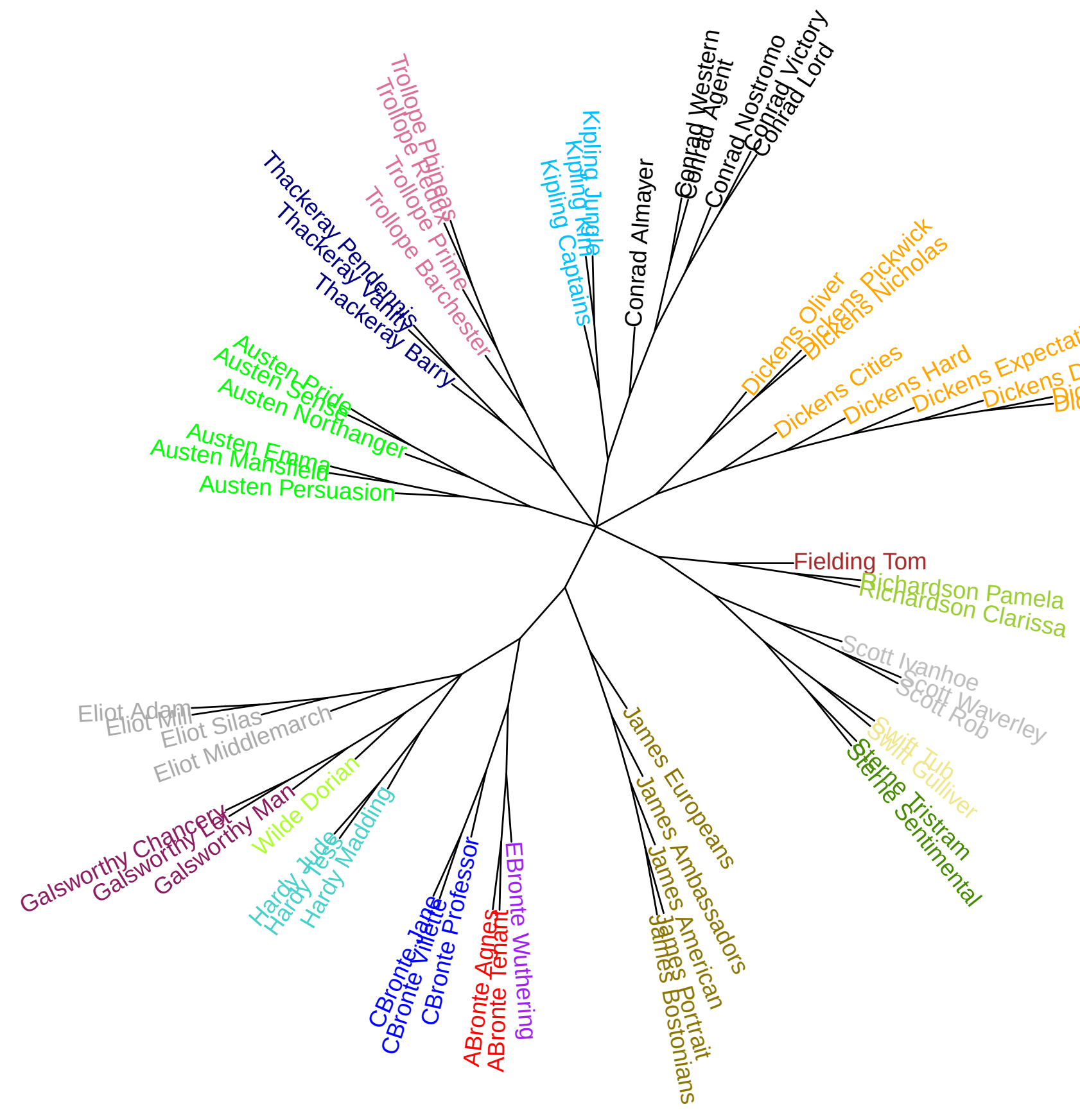
[1] Pedagogical University of Kraków [2] University of Antwerp [3] Jagiellonian University, Kraków
[4] Polish Academy of Sciences [5] Computational Stylistics Group

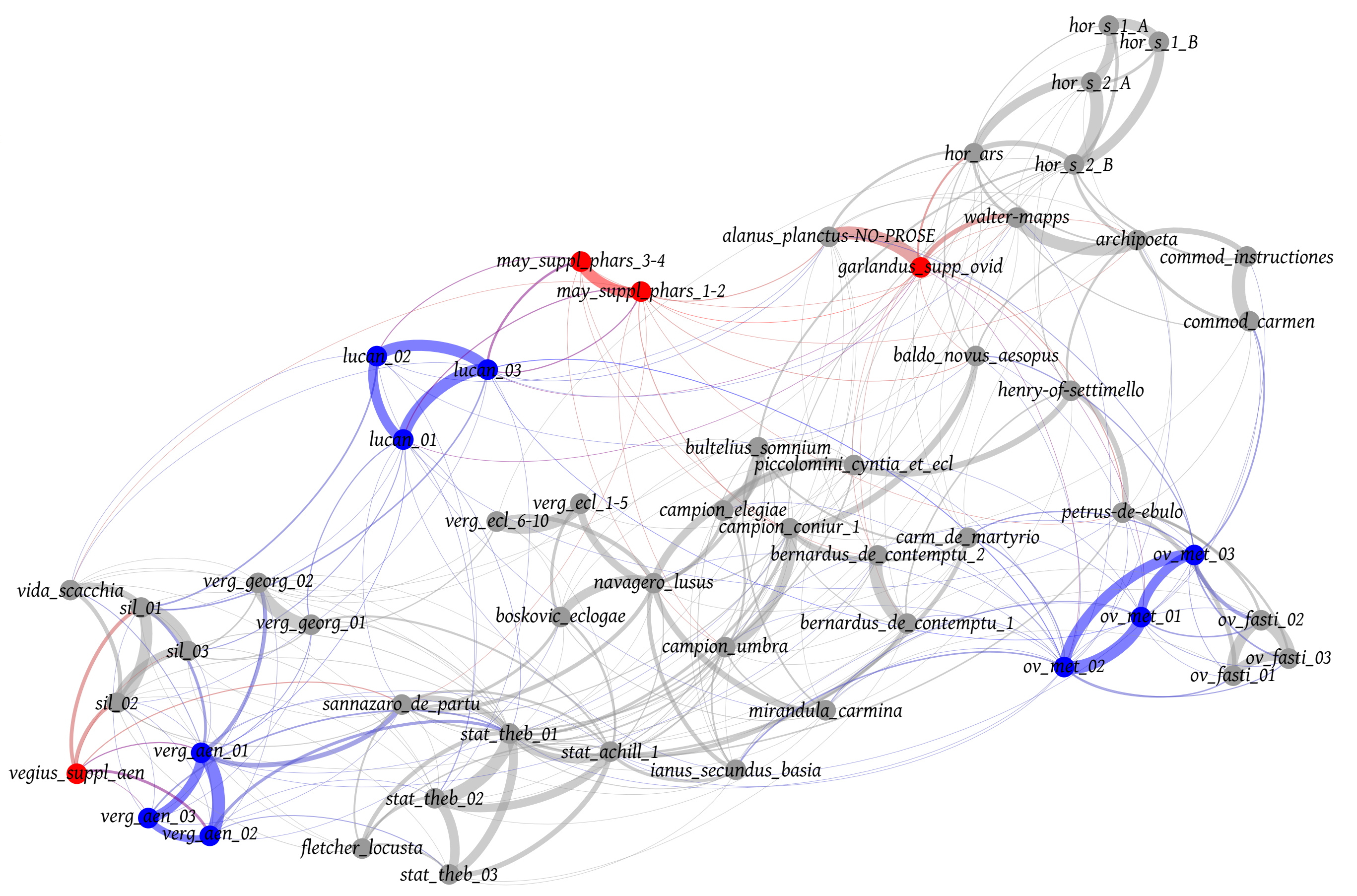
Stylometry with R
a suite of tools

Digital Humanities 2013
University of Nebraska–Lincoln, 16–19 July 3013

computational stylistics stylometry authorship attribution similarities between texts machine-learning classification Burrows's delta explanatory methods visualization dendrogram consensus tree R programming language R package







Overview

This poster describes a suite of functions written in the R programming language, for performing various analyses and/or visualizations in computational stylistics. They are provided in two formats: as separate scripts, and as a compiled library (R package 'stylo').

Features

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

Usage

Assuming you have the 'stylo' package installed, launch R and load the library in question:

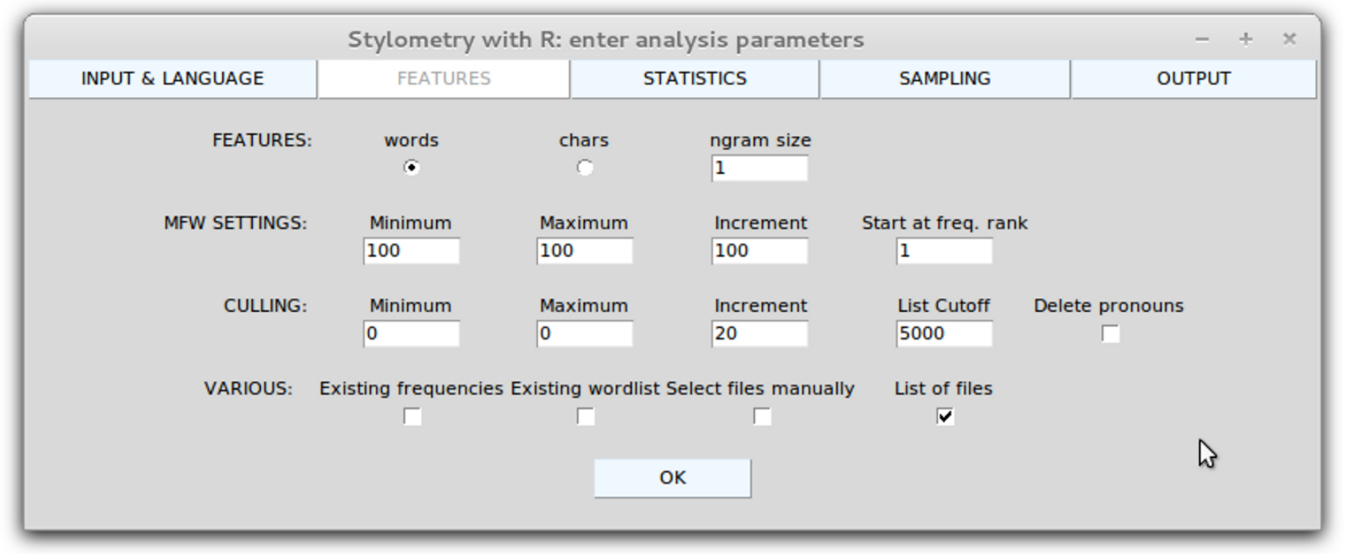
- `library(stylo)`

Next, assuming that your current working directory contains a corpus, try to invoke these functions:

- `stylo()`, `classify()`, `rolling.delta()`, ...

Graphical user interface

Since some humanists might be allergic to the raw command-line mode provided by R – an observation shared by all three authors – a simple yet effective GUI has been added:



Stylometry with R: enter analysis parameters

INPUT & LANGUAGE | FEATURES | STATISTICS | SAMPLING | OUTPUT

FEATURES: words ● chars ○ ngram size 1

MFW SETTINGS: Minimum 100 Maximum 100 Increment 100 Start at freq. rank 1

CULLING: Minimum 0 Maximum 0 Increment 20 List Cutoff 5000 Delete pronouns □

VARIOUS: Existing frequencies □ Existing wordlist □ Select files manually □ List of files ☑

OK

'stylo'

This is the main tool (R function).

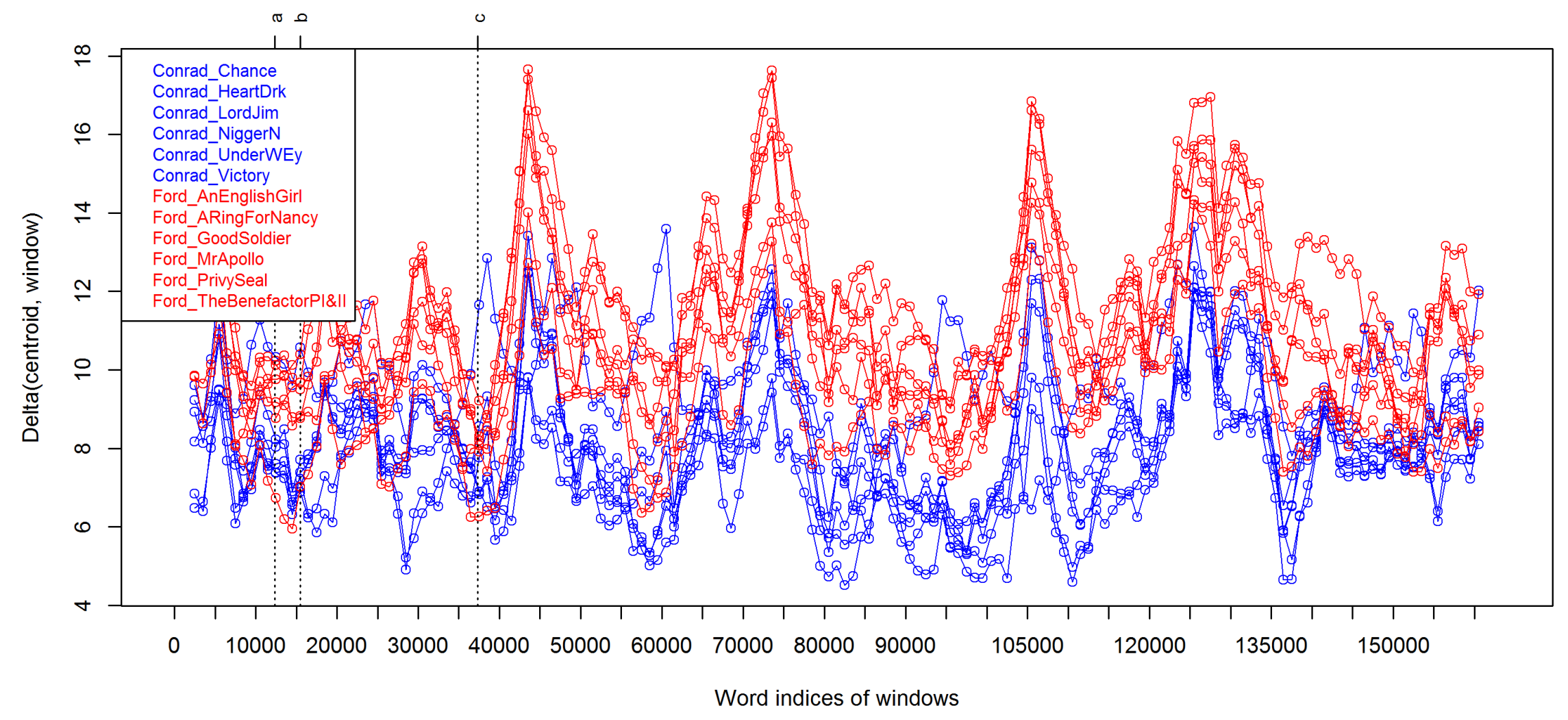
- performs PCA, MDS, Cluster Analysis, and Bootstrap Consensus Trees
- supports plain text files, XML, or HTML
- produces high-quality plots (PDF, JPEG, PNG)
- additionally-generated files (wordlists, tables of word frequencies) can be re-used in other methods
- experimental support for network analysis is available.

'classify'

It performs a number of machine-learning methods of classification: Burrows's Delta, k-NN, Support Vectors Machines, Naive Bayes, and Nearest Shrunken Centroids. Most of the options and features are derived from the 'stylo' function.

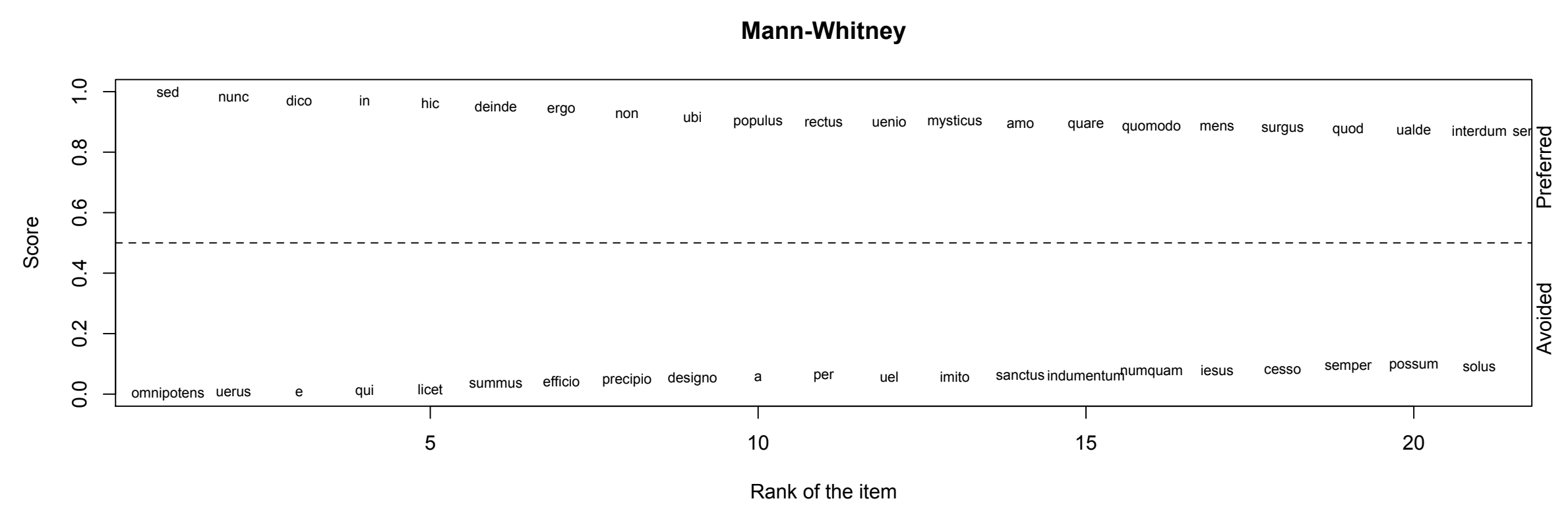
'rolling.delta'

It analyses collaborative works and tries to identify the authorship of their fragments. The first step involves a "windowing" procedure in which each reference text is segmented into consecutive samples. After "rolling" through the test text we can plot the resulting series of Deltas for each reference text in a graph:



'oppose'

It performs a contrastive analysis between two given sets of texts. It generates a list of words significantly preferred by a tested author, and another list containing the words significantly avoided. Some visualizations are available:



Contact us

The software can be downloaded from here:

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

Contact with the authors:

- Maciej Eder <maciejeder@gmail.com>
- Mike Kestemont <mike.kestemont@gmail.com>
- Jan Rybicki <jkrybicki@gmail.com>