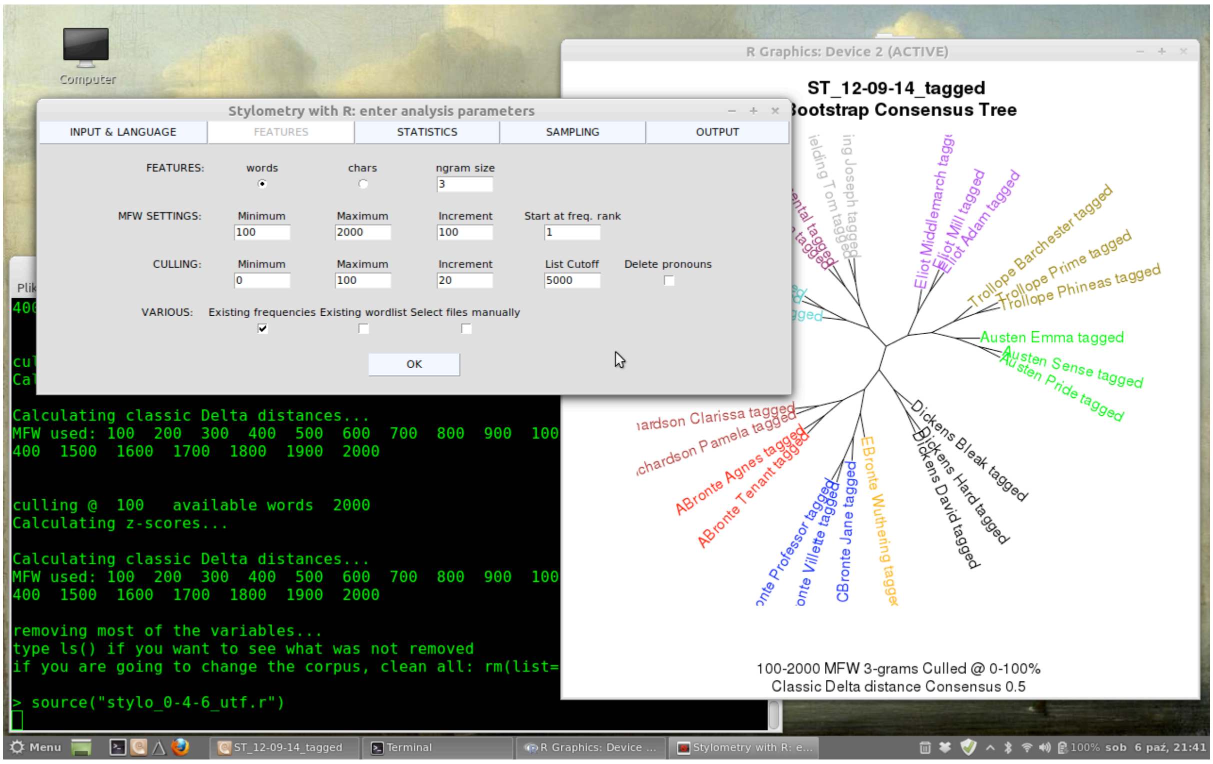
Stylometry with R: a Suite of Tools

Maciej EDER, Mike KESTEMONT, Jan RYBICKI

Pedagogical University in Krakow, Poland and University of Antwerp, Belgium

Stylometry today uses either stand-alone dedicated programs, custom-made by stylometrists, or applies existing software, often one for each stage of the analysis. Stylometry with R can be placed somewhere in-between, as the powerful open-source statistical programming environment provides, on the one hand, the opportunity of building statistical applications from scratch, and, on the other, allows less advanced researchers to use ready-made scripts and libraries. In our own stylometric adventure with R, one of the aims was to build a tool (or a set of tools) that would combine sophisticated state-of-the-art algorithms of classification and/or clustering with a user-friendly interface. In particular, we wanted to implement a number of multidimensional methods that could be used by scholars without programming skills. And more: it soon became evident that once our R scripts are made, provided with a graphic user interface and more or less documented, they are highly usable in class; experience shows that this is an excellent way to work around R's normally steep learning curve without losing anything of the environment's considerable computing power and speed.

The crucial point in building the interface was to keep all the stages of the entire analysis – from loading texts to final results in numeric and graphic form – in a single script. To exemplify, our Stylo script does all the work: it processes electronic texts to create a list of all the words used in all texts studied, with their frequencies in the individual texts; normalizes the frequencies with z-scores (if applicable); selects words from stated frequency ranges for analysis; performs additional procedures that (usually) improve attribution, such as Hoover's (2004a, 2004b) automatic deletion of personal pronouns and culling (automatic removal of words too characteristic for individual texts); compares the results for individual texts; performs a variety of multivariate analyses; presents the similarities/distances obtained in tree diagrams; finally, produces a bootstrap consensus tree – a new graph that combines many tree diagrams for a variety of parameter values. It was our aim to develop a general platform for multi-iteration attribution tests; for instance, an alternate script produced heatmaps to show the degree of Delta's success in attribution at various intervals of the word frequency ranking list (Rybicki and Eder, 2011). The last stage of the interface design was to add a GUI, since some humanists might be allergic to the raw command-line mode provided by R – an observation shared by all three authors – and a host of various small improvements, like saving (and loading) the parameters for the most recent analysis, a wide choice of graphic output formats, etc.



FEATURES: words ● chars ○ ngram size 3

MFW SETTINGS: Minimum 100 Maximum 2000 Increment 100 Start at freq. rank 1

CULLING: Minimum 0 Maximum 100 Increment 20 List Cutoff 5000 Delete pronouns □

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

```
Calculating classic Delta distances...
MFW used: 100  200  300  400  500  600  700  800  900  100
400  1500  1600  1700  1800  1900  2000


culling @  100   available words  2000
Calculating z-scores...

Calculating classic Delta distances...
MFW used: 100  200  300  400  500  600  700  800  900  100
400  1500  1600  1700  1800  1900  2000

removing most of the variables...
type ls() if you want to see what was not removed
if you are going to change the corpus, clean all: rm(list=

> source("stylo_0-4-6_utf.r")
```

Fig. 1 The Stylo script with a bootstrap consensus plot.

The authors believe that at some point the suite of out-of-the-box scripts will cover a wide range of methods used in stylometry. So far, we offer the following tools:

(1) the **Stylo** script, now in version 0.4.8. This is the main tool, thoroughly tested and (partially) documented. It performs Principal Components Analysis, Cluster Analysis, Multidimensional Scaling, and Bootstrap Consensus Trees. The script reads plain text files, XML, or HTML; it supports explicitly nine languages, and implicitly many more (e.g. preliminary tests with a Chinese corpus were quite promising). Publication-quality plots can be exported in PDF, JPEG, PNG, or EMF formats. Additionally-generated files, such as a wordlist used and a table of word frequencies, can be re-used in other scripts or other statistical tools.

(2) the **Classify** script. It performs Delta (Burrows, 2002), k-Nearest Neighbors classification, Support Vectors Machines, Naive Bayes, and Nearest Shrunken Centroids (Jockers et al., 2008). Most of the options are derived from the above-mentioned Stylo script.

(3) the **Rolling Delta** script. It analyses collaborative works and tries to identify the authorship of their fragments. The first step involves a "windowing" procedure (Dalen-Oskam and Zundert, 2007) in which each reference text is segmented into consecutive, equal-sized samples or windows. After "rolling" through the test text we can plot the resulting series of Deltas for each reference text in a graph.

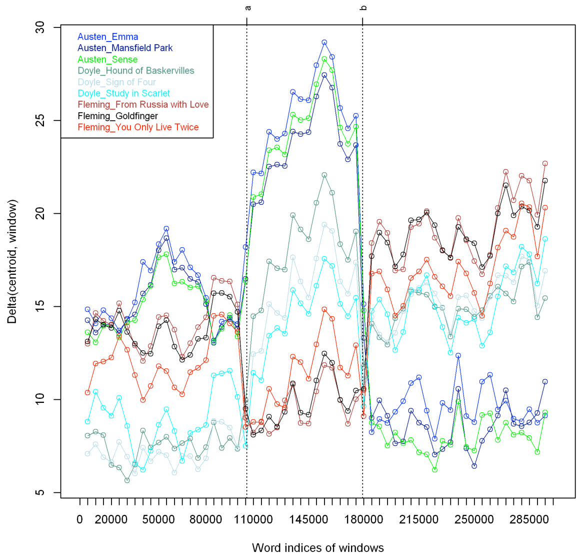


Fig. 2 Sample plot generated by the Rolling Delta script.

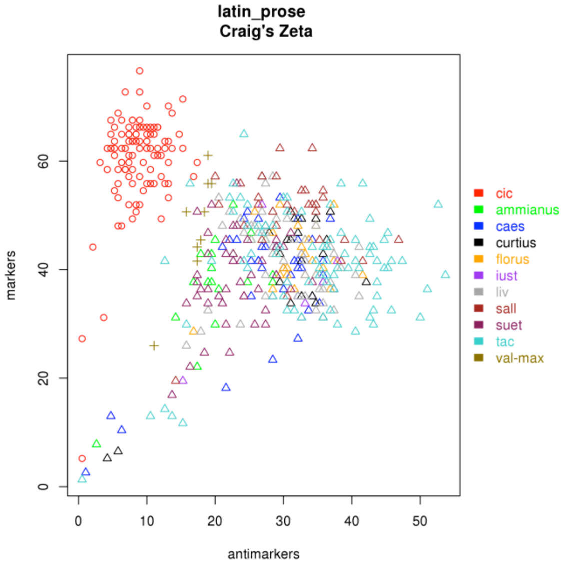


Fig. 3 Sample plot generated by the Oppose Test script.

(4) the **Oppose Test** script. It performs a contrastive analysis between two given sets of texts, using Burrows's Zeta (2006) in its different flavours, including Craig's extensions (Craig and Kinney, 2009). The script generates a list of words significantly preferred by a tested author, and another list containing the words significantly avoided.

(5) the **Keywords** script. This considerably simple tool is an implementation of the concept of "keywords", i.e. words appearing with a statistically significantly higher frequency in one text or collection of texts in comparison to another text or collection.

The scripts are available on https://sites.google.com/site/computationalstylistics/

References

Burrows, J. F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. F. (2006). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, **22**(1): 27–48.

Craig, H. and Kinney, A. F. (eds) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Dalen-Oskam, K. van and Zundert, J. van (2007). Delta for Middle Dutch – Author and Copyist Distinction in 'Walewein'. *Literary and Linguistic Computing,* **22**(4): 345–62.

Hoover, D. L. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Hoover, D. L. (2004b) Delta Prime? *Literary and Linguistic Computing*, **19**(4): 477–95.

Jockers, M. L., Witten, D. M. and Criddle, C. S. (2008). Reassessing authorship of the 'Book of Mormon' using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, **23**(4): 465–91.

Rybicki, J. and Eder, M. (2011). Deeper delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, **26**(3): 315–21.