The TextometrieR package: textual data analysis for social sciences and humanities[†]

Sylvain Loiseau[1,*], **Jean-Philippe Magué**[1], **Serge Heiden**[1]

[1] UMR 5191 ICAR, Université de Lyon
[*] sloiseau@ens-lsh.fr

[†] This work was funded by ANR Grant ANR-06-CORP-029.

Keywords: Statistics in the Social and Political Sciences, Corpus linguistics, Textometry, Textual data analysis

We present the **TextometrieR** package which aims at providing tools for texts and corpus analysis. The package originates in the french tradition of textometry, born after the Benzécri's seminal works on multidimensionnal analysis (Lebart *et al.* 1998). This tradition has developped numerous methods for exploring and visualizing textual data. In the context of a growing interest for R in corpus linguistics (Gries 2008), this package add a new milestone after the packages zipfR (Hevert and Baroni, 2006) and languageR (Baayen, 2008).

The **textometrieR** package is developped as part of a research project gathering several representants of this french tradition, including statisticians, linguists, scolars working on the historical, political or literary discourses, etc. This project aims at summing up the work done in textometry in the past decade. It will provide methods for exploring collocation between words (allowing to identify various phenomenon, from lexicalized multi-word units to stylistic/ideological association between words), building concordance, observing lexical distribution (growing of vocabulary, zipfian distribution) performing multifactorial analysis (for instance in texts typology, comparing diachronical or genre difference between texts, etc.).

From an architectural point of view, the **textometrieR** package is the statistical component of a platform which associates R and the IMS Open Corpus Workbench, a powerfull full text indexer and search engine, under a single Java API. The R/Java communication is based on rJava. This plateform provides a uniform environment to query plain text corpora or annotated corpora (*i.e.* containing metadata or linguistic annotations), in order to build quantitative structures such as frequency lists or contingency tables, and to benefit from R power to analyze and visualize those structures.

References

Baayen R. H. (2008). Analyzing Linguistic Data: A practical introduction to statistics. Cambridge: Cambridge University Press.
`http://cran.r-project.org/web/packages/languageR/index.html`

Evert, S. and Baroni, M. (2006). The zipfR library: Words and other rare events in R, useR!2006 (Vienna, Austria).
`http://zipfR.R-Forge.R-project.org/`.

Gries S. Th. (2008). Quantitative Corpus Linguistics With R: A Practical Introduction. New York: Routledge.

Lebart L., Salem A., Berry L. (1998). Exploring textual data. Dordrecht : Kluwer.