R-based strategies for DH in English Linguistics: a case study

Nicolas Ballier & Paula Lissón
Université Paris Diderot (USPC) – EA 3967, CLILLAC-ARP

Université Sorbonne Paris Cité

CLILLAC-ARP

Université Paris Diderot Paris 7

Abstract

Position statement advocating the development of the programming language R in a curriculum of English linguistics. This is an illustration of a possible strategy for the requirements of NLP for DH studies in a classic curriculum. R plays the role of a Trojan Horse for NLP and statistics, while promoting the acquisition of a programming language. We report an overview of existing practices implemented in an MA and PhD programme at the University of Paris Diderot in the recent years. We emphasize completed aspects of the curriculum and detail existing teaching strategies . Our last section alludes to work still under way, such as getting PhD students to write their own R packages.

Context

- Option in an MA in English Linguistics
- Little background in Maths/NLP
- Establishing better practices for future PhD candidates
- Competing curricula in Linguistics and NLP

Why R?

- Free, multiplatform
- More versatile
- Richness of packages
- Tailored to suit students' research questions
- Interface with the Maths department
- Interdisciplinary bridges with other faculties (Geography, Physics)
- Common culture for Humanities
- **Command line** (RStudio as a compromise)
- Benefits of data visualisation
- Huge reactive community
- Initiation to forums and github
- Building brick for the future Paris Institute of Data Science
- Better employability

Modules and Packages

UNDERGRADUATE

R among tools for linguists following *Language and Computers* (Dickinson *et al.* 2012). Using *spam-ham* classifiers to introduce confusion matrixes.

MA

Year	First semester	Second semester
M1	Corpus methodology: descriptive statistics	Phonetic Analysis 1 (normalisation, plots, visualisation)
M2	Language and Variation (inferential statistics)	Computational phonology (classifiers)/ Phonetic Analysis 2

PHD

- 12 sessions of 90 min covering advanced statistical techniques for linguists. This seminar covers probability distribution, linear regression models, ANOVAs, linear discriminant analysis, principal component analysis…
- data sessions: discussions with a statistician based on PhD datasets

The way you 'R'

A typology of teaching profiles addressing various priorities:

a) Scaring students for their greatest benefit

Insisting on pluridisciplinarity and on several packages (for future PhD students)

b) R for statistics in disguise

(colleague from the Statistics department)

c) Motivating Students

Insisting on data visualisation (i.e. boxplots)

d) Intermediate point

Including how to report and how to interpret results in papers

Issues and limitations

More motivating tasks :

- adapting and completing scripts
- write your scripts for your RQ
- reverse teaching : present the package
- anatomy of a package (MA on lexical complexity with KoRpus)
- co-designing your own package or writing your R package to promote your research (under way)

ISSUES

- Working memory
- Floating point calculation
- Quirky syntax
- Code fetishism

'the encyclopaedic ignorance of self-taught linguists' : becoming experts at secondary details but missing the basic maths behind them

Contact

Professor Nicolas Ballier ; Paula Lissón
Department of English Linguistics,
Université Paris Diderot (USPC), CLILLAC-ARP (EA 3967)
nicolas.ballier@univ-paris-diderot.fr
paula.lisson@etu.univ-paris-diderot.fr

References

Arnold, T., & Tilton, L. 2015. *Humanities Data in R*. Springer.
Ballier, N. forthcoming. R, pour un écosystème du traitement des données? L'exemple de la linguistique. Ph. Caron (ed.) *Données, Métadonnées des corpus et catalogage des objets en sciences humaines et sociales*. Rennes: Presses universitaires de Rennes.
Gries, S. T. 2009. *Quantitative corpus linguistics with R. A practical introduction*. New York-London: Routledge.
Gries, S. T. 2013. *Statistics for linguistics with R: a practical introduction*. *2nd edition.* Walter de Gruyter.
Zaki, M. J., Meira Jr, W., & Meira, W. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.
Silge, J. & Robinson, D. 2017. *Text Mining with R: A Tidy Approach.* O'Reilly.
R Core Team. 2016. R: A language and environment for statistical computing. (Version 3.3.1). Vienna, Austria.: R Foundation for Statistical Computing.
Levshina, N. 2015. *How to do Linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.