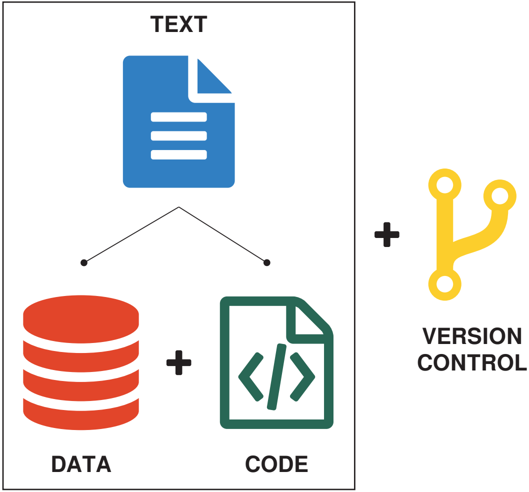
Implementing reproducibility in phonetic research: a computational workflow

Stefano Coretta

26/03/2017

Reproducible research



Why should we care?

The **problem** (Sandve et al. 2013):

- difficulty of reproduction
- difficulty of replication
- retracted papers

The “Yokuts vowels” case (Weigel 2002):

- about **75%** of the data is contrieved (Weigel 2005:149)
- some of the generalisations are **wrong** (Blevins 2004)

The **solution**:

- **Reproducible Research** (RR)

Reproducible Research in linguistics

- linked data (Bird & Simons 2003, Thieberger 2004)
- computational grammar (Maxwell & Amith 2005)
- RR in the Speech Sciences (Abari 2012)
    - lack of scientific culture
    - inefficiency of infrastucture

The workflow of phonetic research

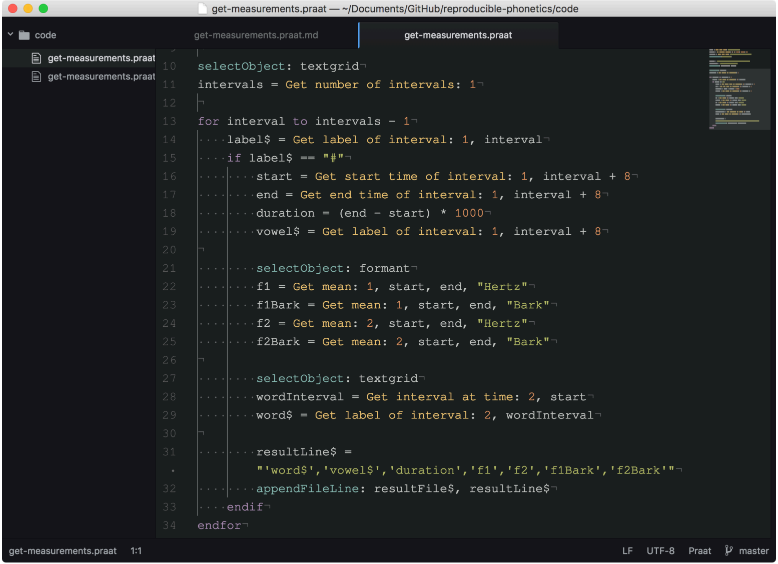
- **Phase A**: scripting (Praat)
- **Phase B**: results and analysis
- **Phase C**: disseminasion

Phase A: source code and documentation

Praat scripting:

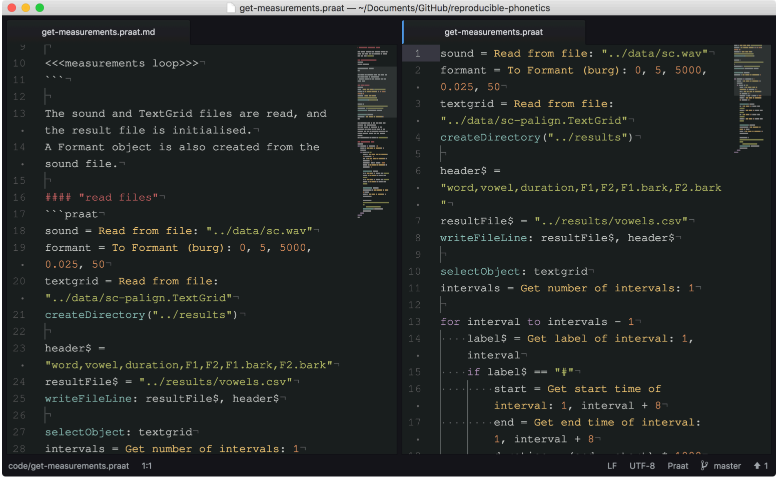
- Atom editor (for syntax highlighting and snippets)
- Literate Markdown
  - tangle: `lmt`
  - weaving: `pandoc`

Atom



```
selectObject: textgrid
intervals = Get number of intervals: 1

for interval to intervals - 1
    label$ = Get label of interval: 1, interval
    if label$ == "#"
        start = Get start time of interval: 1, interval + 8
        end = Get end time of interval: 1, interval + 8
        duration = (end - start) * 1000
        vowel$ = Get label of interval: 1, interval + 8

        selectObject: formant
        f1 = Get mean: 1, start, end, "Hertz"
        f1Bark = Get mean: 1, start, end, "Bark"
        f2 = Get mean: 2, start, end, "Hertz"
        f2Bark = Get mean: 2, start, end, "Bark"

        selectObject: textgrid
        wordInterval = Get interval at time: 2, start
        word$ = Get label of interval: 2, wordInterval

        resultLine$ = "'word$','vowel$','duration','f1','f2','f1Bark','f2Bark'"
        appendFileLine: resultFile$, resultLine$
    endif
endfor
```

lmt

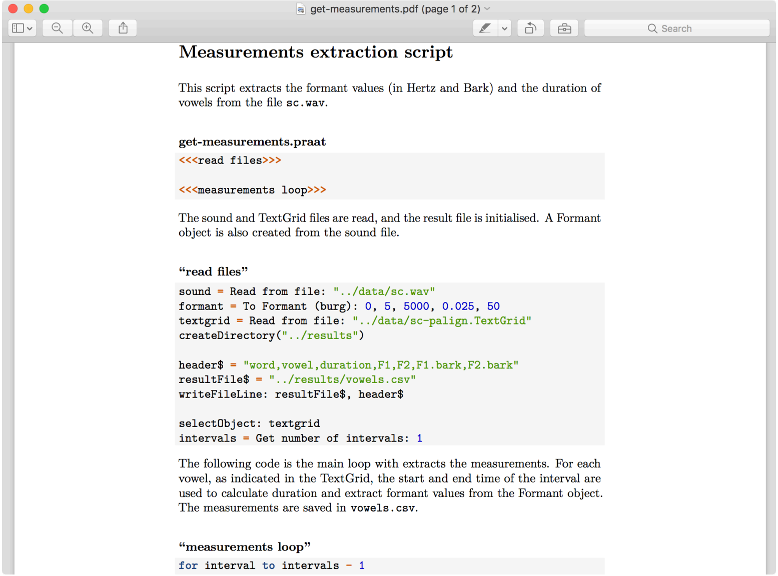

get-measurements.praat — ~/Documents/GitHub/reproducible-phonetics
get-measurements.praat.md
get-measurements.praat

~~~

<<<measurements loop>>>
```

The sound and TextGrid files are read, and the result file is initialised.
A Formant object is also created from the sound file.

#### "read files"
```praat
sound = Read from file: "../data/sc.wav"
formant = To Formant (burg): 0, 5, 5000, 0.025, 50
textgrid = Read from file: "../data/sc-palign.TextGrid"
createDirectory("../results")

header$ = "word,vowel,duration,F1,F2,F1.bark,F2.bark"
resultFile$ = "../results/vowels.csv"
writeFileLine: resultFile$, header$

selectObject: textgrid
intervals = Get number of intervals: 1
~~~

```
sound = Read from file: "../data/sc.wav"
formant = To Formant (burg): 0, 5, 5000, 0.025, 50
textgrid = Read from file: "../data/sc-palign.TextGrid"
createDirectory("../results")

header$ = "word,vowel,duration,F1,F2,F1.bark,F2.bark"
resultFile$ = "../results/vowels.csv"
writeFileLine: resultFile$, header$

selectObject: textgrid
intervals = Get number of intervals: 1

for interval to intervals - 1
    label$ = Get label of interval: 1, interval
    if label$ == "#"
        start = Get start time of interval: 1, interval + 8
        end = Get end time of interval: 1, interval + 8
```

code/get-measurements.praat 1:1 LF UTF-8 Praat master ⬆ 1

pandoc



Measurements extraction script

This script extracts the formant values (in Hertz and Bark) and the duration of vowels from the file `sc.wav`.

get-measurements.praat

```
<<<read files>>>

<<<measurements loop>>>
```

The sound and TextGrid files are read, and the result file is initialised. A Formant object is also created from the sound file.

"read files"

```
sound = Read from file: "../data/sc.wav"
formant = To Formant (burg): 0, 5, 5000, 0.025, 50
textgrid = Read from file: "../data/sc-palign.TextGrid"
createDirectory("../results")

header$ = "word,vowel,duration,F1,F2,F1.bark,F2.bark"
resultFile$ = "../results/vowels.csv"
writeFileLine: resultFile$, header$

selectObject: textgrid
intervals = Get number of intervals: 1
```

The following code is the main loop with extracts the measurements. For each vowel, as indicated in the TextGrid, the start and end time of the interval are used to calculate duration and extract formant values from the Formant object. The measurements are saved in `vowels.csv`.

"measurements loop"

```
for interval to intervals - 1
```

Phase B: the speakr package

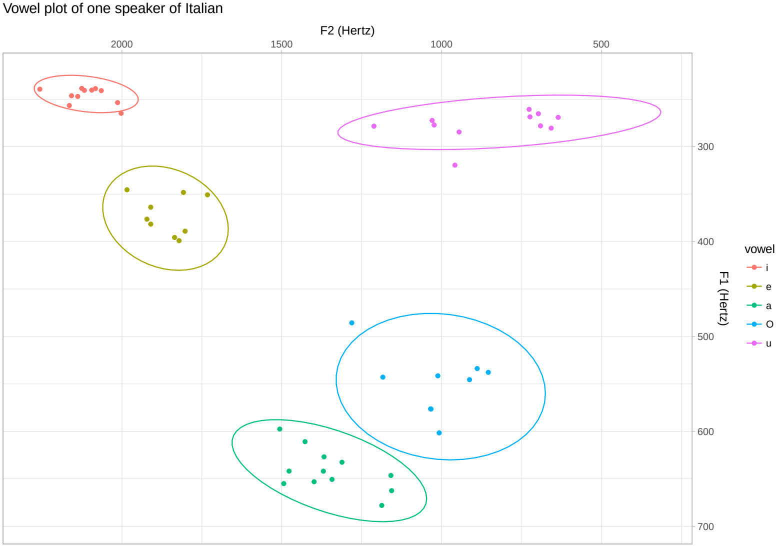
speakr is an R package to aid Praat users:

- aim: tangle and run Praat scripts from within R
- two main functions
  - lmt(): tangle a Praat script
  - praatRun(): run a Praat script

Phase B: the speakr package

```r
# Tangle a Praat script
lmt("code/get-measurements.praat.md")

# Run the script
praatRun("code/get-measurements.praat")

# Read the results of the script
vowels <- read_csv("results/vowels.csv") %>%
    mutate_if(is.character, as.factor) %>%
    mutate(vowel = factor(vowel, c("i", "e", "a",
                                   "O", "u")))
```

Phase B: the speakr package



Phase C: dissemination

- GitHub
- Open Science Framework

Abari, Kálmán. 2012. Reproducible research in speech sciences. *International Journal of Computer Science Issues* 9(6). 43–52.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 557–582.

Blevins, Juliette. 2004. A reconsideration of Yokuts vowels. *International Journal of American Linguistics* 70(1). 33–51.

Maxwell, Michael & Jonathan D. Amith. 2005. Language documentation: the Nahuatl grammar. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, 474–485. Berlin Heidelberg: Springer-Verlag.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor & Eivind Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology* 9(10). 1–4.

Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In Peter K. Austin (ed.), *Language documenta and description*, vol. 2, Hans Rausing Endangered Languages Project, School of Oriental and African Studies, University of London.

Weigel, William. 2005. *Yowlumne in the Twentieth century*: University of California, Berkley dissertation.

Weigel, William F. 2002. The Yokuts canon: A case study in the interaction of theory and description. Paper presented at the annual meeting of the Linguistics Society of America, January 2002, San Francisco.