Advanced ELAN manipulation and analysis

Niko Partanen

2017-11-01

Contents

Chapter 1

Info

1.1 Course goals

I have been teaching on courses, informal workshops and meetings the basic use of ELAN and Praat. It seems to me that instructions of how these programs are used are included in many courses and summer schools, but it is not maybe so common to move beyond from that. In principle there is no need for this: once the researcher is familiar with the GUI, basic usage and search principles, there is not necessarily that much more to cover in the program itself. I strongly believe that a course of some week and intensive use of few more is enough to master ELAN.

On the other hand, many courses and handbooks focus into statistical analysis of linguistic data. I think there is a clear need for more discussion about the ways how do we get our linguistic data most effectively into statistical or analytical software we intend to use. This may sound like a topic that is not worth that much thought, but as is often said that 80% of data analysis is data manipulation, the topic is eventually more central for our daily work than we may think.

There are numerous ways to work programmatically with ELAN files, and this can be very useful both while producing the new data or analysing existing files. Although the focus is in ELAN, I will also mention Praat from time to time. These two programs have somewhat different goals and niches, which are covered better in its own section. There is also a very different approach in these tools, since Praat can be very far manipulated through PraatScript, whereas with ELAN the means available are bit different.

This course is not an introduction to ELAN or Praat, and it is neither an introduction to programming. Basic knowledge of R or Python will help a lot, but brave mind will probably be enough. I think we have to take as starting point that the majority of us are researchers first, and programming is neither our job or best skill. However, we can all learn to pay attention to some basic programming practices that make our work easier to adopt for others. These includes code comments and version control, among some other conventions.

The main goal of this course is to help thinking about ways to automatize some parts of our workflows related to linguistic research. There are numerous tasks we do which demand hundreds and hundreds of clicks on mouse, and in all these situations we have to ask: **are we spending time with something that can be automatized, or with something that demands our expert knowledge to be solved?** We have to maximize the latter, also the time we spend when we try to get into that level of our work, instead of fighting against the cumbersome manual workflow where emphasis is easily in unnecessary parts of the task. Research necessarily is somewhat "boring", it is inevitable that we do some thing thousands of times just to find out that we didn't really learn that much. If we can find ways to speed up the manual parts of the process, we have the possibility to wander on even more unnecessary and poorly rewarding veins of though, which will eventually lead us to larger questions and their answers.

1.2 Course structure

The course materials will be divided into several parts, and which all are used depends from the length of the course. Some parts can be skipped, and some can be used only as a reference. For example, the part about tools contains brief descriptions of the R and Python packages mainly discussed here with their most commonly used functions, so that can be a good place to look for help. I have also included there a list of useful references and links about basic usage of R and Python, just to get everyone onward.

1. Introduction
   - ELAN corpora
   - Goals of this work
   - Available tools
2. Structure of the ELAN file
   - XML structure
   - How ELAN interacts with its own XML?
3. Python examples
   - Creating new files with Pympi
   - Manipulating ELAN file with Pympi
4. R examples
   - Interaction with emuR
   - Interaction between Praat and R
   - Creating new ELAN tiers in R
5. Summary

1.3 Schedule

- Thu 16.11. FRIAS
  - 09:30-11:00 **Course** Topic: ELAN corpora & Goals of programmatic workflows & ELAN file
  - 11:00-11:30 Coffee
  - 11:30-13:00 **Course** Topic: Available ecosystem (R, Python, ELAN, Praat)
  - 13:00-14:00 Lunch (on own expenses)
  - 14:00-15:30 **Course** Topic: Parsing ELAN file + metadata
  - 15:30-16:00 Coffee
  - 16:00-17:30 **Course** Topic: Further interaction with ELAN corpus in R
- Thu 16.11. University
  - 18:00–20:00 Guest lecture by Mark Davies “Examining variation and change in language and culture with large online corpora”
- Fr 17.11. FRIAS
  - 09:30-11:00 **Course** Topic: ELAN tier manipulation with Python
  - 11:00-11:30 Coffee
  - 11:30-13:00 **Course** Topic: Examples of tools in interaction
  - 13:00-14:00 Lunch (on own expenses)
  - 14:00-15:30 **Course** Topic: Summary

1.4 Resources

The whole course exists as an R package, which contains all functions discussed in the material. It can be installed with:

```
# comment: fix this later
library(devtools)
install_github('langdoc/adv_elan')
```

Besides the R package, the course repository also contains all Python code examples from the course. They are maybe also put into a module if there is time.

The course materials also contain an example ELAN corpus with associated audio files.

1.5 Practical info

I will teach with these materials, or their subset, in a workshop in Freiburg, around 16.-17. November 2017.

Chapter 2

Introduction

2.1 Elan and Praat – strengths and weaknesses

2.2 Linguistic software

2.3 ELAN corpora

2.4 Preprocessing

2.4.1 Example of preprocessing workflow

2.5 Analysis workflows

2.6 When to write back to ELAN file

2.7 Annotations as an independent dataset

Chapter 3

Tools

3.1 Git

3.2 R

3.2.1 `tidytext`

3.2.2 `xml2`

3.3 Python

3.3.1 `pympi`

3.4 Anaconda

3.5 reticulate

3.6 PraatScript

3.7 XPath

3.7.1 Examples

Chapter 4

ELAN file structure

4.1 Minimal file

4.1.1 Participant name convention

4.2 Tier type naming convention

4.3 Hierarchies

4.4 Discussion

Chapter 5

Parsing ELAN files to R

5.1 Why to read ELAN files into R?

5.2 Parsing with FRelan package

Chapter 6

Manipulating ELAN files with Pympi

Chapter 7

Pympi examples

7.1 Creating a new ELAN file

7.2 Populating the ELAN file with content

7.3 Merging ELAN files

Chapter 8

Shiny components

8.1 DT

8.2 Leaflet

8.3 ggplot2

8.4 Advantages and disadvantages of Shiny

Chapter 9

Example: Interaction with emuR

9.1 Procedure

Chapter 10

Example: Interaction with Praat

10.1 Research questions

10.2 Implementation

10.3 Shiny application

10.4 Observations

10.5 Exercise

Chapter 11

Example: Concordances and map

11.1 Use

Chapter 12

Example: Preprocessing workflow

12.1 From points to polygon

12.2 based on this:

12.3 https://gis.stackexchange.com/questions/222978/lon-lat-to-simple-fea

Chapter 13

Final Words

Chapter 14

Placeholder

Bibliography