# Advanced Python for Linguistics

## Prof. Flavio Massimiliano Cecchini

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

The goal of this course is to explore the possibilities and tools offered by the programming language Python with respect to (Computational) Linguistics. At its heart lies the desire to put into practice some of the concepts and techniques, or parts thereof, discussed in other courses like *Computational Linguistics 1 & 2* and *Formal methods and models for Linguistics*, to present useful Python modules/packages (like nltk, numpy, scipy), to describe implementations of (routine) operations on texts and strings, and to examine the possible integration of existing Natural Language Processing (nlp) tools into a Python script and more generally into a pipeline. Note: the designation “advanced” is meant to suggest that this course is not (or at least, not only) about programming fundamentals, but one in which the main aim is to develop simple scripts for linguistic analysis.

At the end of the course, the students will have gained some measure of confidence with textual and linguistic analysis by means of Python (and consequently other programming languages, too) that will serve them as the basis for producing simple scripts fitting their needs in the realm of (Computational) Linguistics, conceiving their own modules for linguistic analysis, or tweaking existing ones for their goals.

***COURSE CONTENT***

The exact shape of the course will be modelled according to the interest of the participants and the subjects encountered by them in other courses. Generally speaking, the course will be carried out as a workshop where problems of practical nature are posed and tried to be solved.

Among the most important subjects that will almost surely be dealt with are:

* character encodings and Unicode;
* regular expressions;
* the nltk library for the implementation and training of simple linguistic analysers;
* treatment and conversion of different formats for linguistic annotation (e.g. the CoNLL family);
* data processing for the evaluation of nlp tools.

***READING LIST***

No particular bibliography is tied to this course, and suggested readings will be provided during the course according to the topics that will be treated and the interests of the participants. The official [Python documentation](https://docs.python.org/3/) will be one of the main references during the course.

***TEACHING METHOD***

Live sessions of programming around tasks appearing in Computational Linguistics, with lecture parts (in English) to introduce and explain implementation-oriented topics.

***ASSESSMENT METHODS AND CRITERIA***

Each student will produce and present a small project related to the topics dealt with during the course, and contextually also answer more general questions of more theoretical nature. Prior to the actual presentation, the project will need to be sent to the teacher for approval and to collect feedback to be used as the basis for possible improvements.

***NOTES AND PREREQUISITES***

Familiarity with any kind of programming language will greatly help any student, as will some level of understanding of mathematical formalisations of concepts seen in Computational Linguistics (as treated e.g. in the first-year course *Formal Methods and Models for Computational Linguistics*). That said, there is no other requirement than the obligatory first-year course *Fundamentals of Computer Science for Linguistics*, and the actual contents of the course will be tailored to the competences of the participants.

*Office hours for students*

In person, usually on Thursday afternoons, from 15 to 17, in the teacher's office at the CIRCSE research centre. Additionally, it is always possible to arrange a meeting by mail.