# Computational Linguistics

## Prof. Francesco Mambrini

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

The course aims to provide students with the basic notions of Computational Linguistics, as well as to offer a first introduction to the world of language resources, especially textual. It will be divided into three parts. After a first introduction on the definition and history of Computational Linguistics, the course will focus on the concept of *corpus* and the digital representation of textual information. In the second part of the course, we illustrate some key techniques for the computational treatment and processing of textual data. Finally, we discuss some of the basic methods of textual exploration (concordances, collocations) and introduce the main Natural Language Processing techniques.

At the end of the course, students will be able to:

* know the main features of digital textual information;
* carry out simple operations for the automatic processing of texts through regular expressions and command-line tools (*wc*, *grep*, *cut, tr,* and more);
* identify the most appropriate computational methodologies to carry out the different types of linguistic analysis.

***COURSE CONTENT***

1. An introduction to computational linguistics: definition, goals, and a brief history.

2. The concept of *corpus* between Corpus Linguistics and Computational Linguistics.

3. The digital text: the encoding of textual information.

4. How to process the digital text: regular expressions and some command-line applications.

5. The quantitative analysis of texts: word frequency and distributions.

6. Exploring a corpus: concordances and collocations.

7. Annotating a text: an introduction to Natural Language Processing.

***READING LIST***

Lenci A.-Montemagni S. & Pirrelli V. (2016). *Testo e computer. Introduzione alla linguistica computazionale*. Rome, Carocci (second edition). **Compulsory:** chapters 1, 2. **Optional:** chapters 4, 5, 7, 8.

***TEACHING METHOD***

The lectures will be based on both the theory and the practice. The students who have a laptop are invited to use it during classes.

***ASSESSMENT METHOD AND CRITERIA***

The final assessment will consist in an oral exam. The aim of the examination is to assess the acquisition of the contents presented in class and the ability to use the newly-acquired concepts to solve some practical exercises related to the computational analysis of textual data.

In particular, the questions of the exam will be focused on: a) theoretical aspects related to digital texts, textual corpora, the tools and the methods for their analysis; b) practical scenarios, such as: 1. the choice of the best tool to extract a certain type of information from a text, 2. the planning of a hypothetical scenario proposed by the lecturer (data collection, selection of tools, data processing) to carry out a linguistic research activity on corpora.

Both types of questions will have the same weight in the determination of the final mark.

## ***NOTES AND PREREQUISITES***

As this is an introductory course, there are no prerequisites in terms of computer skills. A general familiarity with the use of computers will be sufficient.

Further information can be found on the lecturer's webpage at http://docenti.unicatt.it/web/searchByName.do?language=ENG, or on the Faculty notice board.