# Formal methods and models for Computational Linguistics

## Prof. Flavio Massimiliano Cecchini

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

The aim of this course is to lay the foundations for a sound understanding of the most computational and mathematical aspects of Computational Linguistics. There will be a special focus on formal and methodological elements (such as correct definitions and motivations for the use of given instruments) and their operative implementation rather than on demonstrations. The topics will range over a variety of mathematical areas, among which most relevance will be given to set theory, statistics & probability, graph theory, linear algebra (vector spaces). All covered subjects will be put into relation with problems and themes of Computational Linguistics/Natural Language Processing, and possibly accompanied by the reading of related papers and works.

At the end of the course, the students will have gained some form of familiarity with the notions and methods underlying the most technical passages in works of literature belonging to the field of Computational Linguistics (and beyond). This will also allow the students to more easily implement concepts of a mathematical nature into their computer programming scripts, and to use existing modules of mathematical/statistical kind for such programming languages, helping them further develop their skills in this area. Moreover, the insights gained during the course will give the students some more confidence in redefining and tweaking already existing mathematical instruments towards their goals in tasks involving language analysis.

***COURSE CONTENT***

The course will include an introductive part about basic concepts and notational formalism. Each subsequent section will begin with the necessary basic definitions for that subject, and will then proceed to illustrate methods and techniques with particular regard to Computational Linguistics.

The most relevant subjects will be:

* words and texts from the statistical point of view, such as frequency and probability distributions, entropy, power (“Zipf's”) law, small-world networks;
* word and text representations, such as bags of words, vector spaces, embeddings, word graphs and (syntactic) trees;
* standard methods for automated analysis, such as probabilistic and especially (hidden) Markov models, maximum entropy, neural networks;
* comparison of strings and sets: editing distances, evaluation scores (precision, recall, F-measure), similarity measures between clusters.

***READING LIST***

Various sources (papers, books...) will be proposed as readings during the course, ranging from closely related to the discussed topics to lectures for the curious or interested.

Some general texts which will serve as references for the main topics of the course are:

* Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Second edition. Englewood Cliffs, nj, usa: Prentice Hall, 2008. Drafts of third, updated edition freely accessible [online](https://web.stanford.edu/~jurafsky/slp3/).
* Manning, Chris, and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, ma, usa: mit Press, May 1999. It has a companion [website](https://nlp.stanford.edu/fsnlp/).
* Havelka, Jiří. “[Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax](https://dspace.cuni.cz/handle/20.500.11956/12614?locale-attribute=en).” PhD diss., Univerzita Karlova – Matematicko-fyzikální fakulta, June 2007.

In addition to these, more general resources and handbooks of mathematical nature will be pointed out during the course.

***TEACHING METHOD***

Lectures (in English) with very occasional use of software in class.

***ASSESSMENT METHODS AND CRITERIA***

The exam will be divided into two parts: a written and an oral one. Passing the written part is required to access the oral one. The written part can also be sustained in the form of two midterm tests, one in the middle and the other one at the end of the semester (corresponding to half of the final exam).

The written part will consist mostly of exercises tied to the arguments discussed in class (e.g. standard exercises of basic probability, computing precision, recall and F1-score for a simulated automated linguistic annotation...), with some related (possibly open) questions of more theorical nature.

The oral part will range over the more theoretical parts of the course, and it will include more open-ended questions to stimulate reasoning, and possibly the reading and comment of an excerpt of an article or other work in the field of Computational Linguistics..

The two parts will be considered of equivalent weight when determining the final grade.

***NOTES AND PREREQUISITES***

This course does not require any particular previous mathematical nor programming knowledge, even if some familiarity in these areas could help. Basic notions of Computational Linguistics (especially on the Linguistics side) are required to understand the general reference framework of this course. Overall, an inquisitive and active attitude is encouraged.

*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.