# Computational linguistics 1

## Prof. Marco Carlo Passarotti

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

The course aims to introduce the fundamental concepts and methodologies of the discipline, specifically focusing on linguistic resources (corpora, lexica) and methods and tools for the automatic processing of morphology and syntax.

At the end of the course, students will be provided with (a) expertise about the theoretical linguistic framework(s) most used in linguistic resources and (b) skills to use different kinds of linguistic resources as well as to analyze automatically linguistic data with Natural Language Processing tools at the morphological and syntactic level.

Furthermore, students will be able to run evaluation campaigns of the results of tools for lemmatization, Part of Speech tagging and dependency parsing. Students will be provided with the ability to choose the resources and tools that best fit their specific purposes.

***COURSE CONTENT***

After a brief overview on the history of the discipline, the course is divided into two sections:

1. presentation of two kinds of collections of linguistic empirical data ("linguistic resources"):

 a. textual resources: description of the layers of linguistic annotation of a textual corpus (including data formats); introduction to both the main theoretical linguistic frameworks and a number of query languages used for syntactically annotated corpora, particularly for dependency treebanks (with practical exercises);

 b. lexical resources: introduction to WordNet, ProbBank, NomBank, VerbNet and FrameNet, and to their data formats.

2. Methods and tools for Natural Language Processing. The course will provide the students with the basics for using the Command Line Interface for text processing purposes. Both the rule-based and the data-driven paradigms used in Natural Language Processing will be presented. A number of lemmatizers, Part of Speech taggers and dependency parsers will be presented and applied to textual data in various languages. Different evaluation procedures will be introduced.

***READING LIST***

[All the readings reported below are not mandatory for the exam]

AA.VV., Le due culture. Almanacco Letterario 1962, pp. 143-144 & 313-318.

A. Abeillé, Introduction. In Treebanks. Building and Using Parsed Corpora, Dordrecht (The Netherlands), Kluwer Academic Publisher, 2003, pp. xiii-xxvi.

E.M. Bender, Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax, Synthesis lectures on human language technologies 6.3, 2013, pp. 1-184.

r. Busa, L'Analisi linguistica nell'evoluzione mondiale dei mezzi d'informazione, Almanacco Letterario 1962, pp. 103-108.

J. Carroll, Parsing. In M. Ruslan, The Oxford handbook of computational linguistics, Oxford University Press, 2005, pp. 233-248.

M.C. De Marneffe, et al., Universal dependencies, Computational linguistics 47.2, 2021, pp. 255-308. <https://nlp.stanford.edu/pubs/demarneffe2021universal.pdf>

J. Egbert, T. Larsson, D. Biber, Doing linguistics with a corpus: Methodological considerations for the everyday user, Cambridge University Press, 2020.

C.J. Fillmore, C.R. Johnson, M.R.L. Petruck, Background to framenet. International journal of lexicography 16.3, 2003, pp. 235-250.

J. Hutchins, ALPAC: the (in)famous report. MT News International, no. 14, June 1996, pp. 9-12. <https://pangeanic.co.uk/wp-content/uploads/sites/2/2014/04/ALPAC-1996.pdf>

J. Nyhan, M. Passarotti (eds.), One Origin of Digital Humanities: Fr Roberto Busa in His Own Words, 2019, Springer Nature.

M. Palmer, D. Gildea, P. Kingsbury, The proposition bank: a corpus annotated with semantic roles. Computational Linguistics Journal 31, 1, 2005, pp. 71-105. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.439.6122&rep=rep1&type=pdf>

W.K. Percival, Reflections on the history of dependency notions in linguistics, Historiographia linguistica 17.1-2, 1990, pp. 29-47. <https://www.ingentaconnect.com/content/jbp/hl/1990/00000017/f0020001/art00004?crawler=true&casa_token=pvZP95DThSAAAAAA:zmr1L0QXNlcvFSp_iPmONxKLdsB5EBTNYV2Q69hzAqT9bdYS2Alvfube9XPEH5--1ogxZDXV5rPvFbU>

C.P. Snow, The Two Cultures. In C. P. Snow, The Two Cultures and the Scientific Revolution. The Rede Lecture. New York: Cambridge University Press, 1959, pp. 1-22. <http://www.jstor.org/stable/pdf/1578601.pdf>

L. Tesnière, Elements of structural syntax, John Benjamins Publishing Company, 2015. <https://library.oapen.org/bitstream/handle/20.500.12657/30722/643257.pdf?sequence=1>

M. Wynne (ed.), Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books, 2005. Available online from <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/> [Chapters 1, 2 and 3]

***TEACHING METHOD***

Lectures (in English) with exercises.

***ASSESSMENT METHOD AND CRITERIA***

Oral exam to verify the degree of acquisition of the course contents. The exam includes also one or more practical exercises about the use of linguistic resources and/or Natural Language Processing tools.

In particular, questions are about two main kinds of contents: (a) theoretical aspects of linguistic resources and Natural Language Processing tools, like for instance the different theoretical linguistic frameworks supporting syntactic annotation in treebanks and the rule-based vs. data-driven paradigm in Natural Language Processing; (b) practical issues, like for instance using regular expressions for querying linguistic data and running a Part of Speech tagger via Command Line Interface.

In terms of assesment criteria, there is no difference between the questions about the theoretical questions of the course and those on the practical aspects.

***NOTES AND PREREQUISITES***

Given the introductory nature of the course, no prior specific expertise in computer science is required, apart from basic competence (e.g. using a web browser).

*Place and time of consultation hours*

On appointment, by sending an e-mail to marco.passarotti@unicatt.it.

Place: CIRCSE Research Center, Franciscanum building, second floor, room n. 210.