Linguistic Data Analysis for Digital Humanities

NOVA University Lisbon

School of Social Sciences and Humanities

Study Cycle(s) and Field(s)

MA and BA degrees in Digital Humanities, Language Sciences or other related fields

Course name

Linguistic Data Analysis for Digital Humanities

Learning outcomes

After taking this course students are expected to:

- get acquainted with, to understand and to evaluate the methods and tools for analyzing and extracting information from large sets of linguistic data.

- know how to organize and use large sets of linguistic data to extract useful information directed and relevant to research issues specific of the Digital Humanities field.

- know methods of analysis and detection of linguistic cues and features and to determine their relevance to perform specific information extraction and text mining tasks for non-linguistic purposes.

- develop skills to build and use textual corpora in and analytical and informed way, according to tested methodologies and using available tools for corpus treatment and analysis.

- develop skills and strategies for detecting and using linguistic cues and features for research purposes in the field of (Digital) Arts and Humanities.

Syllabus/Contents to be covered in the course

1. Corpus Linguistics

    1.1. Introduction and theoretical framework

    1.2 Corpus constitution: criteria, parameters and representativeness

    1.3 Corpus tools and procedures: overview

2. From linguistic data to specific information extraction

2.1 Linguistic units, features and cues

2.2 Textual analysis: macro vs. micro level; syntagmatic vs. paradigmatic analysis

2.3 Lexical statistics, concordances and collocations

3. Applying Corpus Linguistics and text mining strategies

3.1 Research question, data selection and corpus compilation

3.2 Determining relevant linguistic features and cues

3.3 Results extraction and analysis

Having as basis textual data, an object of analysis that includes relevant information in linguistic form, the contents covered in the course aim at allowing for an objective and informed analysis of texts, with the help of tools that are used to extract information from large amounts of digital data. To achieve this, students need to know how to select or compile relevant data sets and which methods of analysis will allow them to achieve solid results, considering linguistic, statistical, and other criteria. The theoretical introduction to the field of Corpus Linguistics provides them this perception, as well as knowledge on current and available computational tools and methods. The relation between linguistic data and information requires promoting the sensibility to linguistic data and phenomena and training the linguistic analysis. This will allow students to establish the necessary mapping between linguistic forms and structures and the specific information they intent to extract from data, motivating the second point of the course contents. The attention to the actual practical use of tools and strategies covered in point 3, besides allowing students to consolidate the theoretical skills acquired, provides them with the opportunity to get acquainted with the specific practical steps in data treatment and management, tools handling and working pipelines and dependencies, as well as to be confronted with and solve the real issues that come up in real hands-on work.

Teaching methods and activities

The course will be delivered conjugating theoretical exposure and hands-on work, favouring the bottom-up discovery of theoretical and methodological needs and of relevant Arts and Humanities-oriented results.

This means designing controlled activities that will allow students to:

- work the data, using all the necessary tools and methods
- achieve the expected results, understanding the theoretical assumptions and explanations provided by Corpus Linguistics
- reach conclusions and propose new knowledge in Arts and Humanities, based on the analysis of large amounts of linguistic digital data
- analyse and confront the new proposals with the state of the art, being aware of the potential and the limitations of using computational methods and tools for exploring linguistic data for Arts and Humanities research goals.

The teaching methods and activities will include theoretical exposure (through classes and autonomous readings), guided hands-on activities (aided by scripts, demonstration videos, hands-on classes), peer discussion and cooperation, workflow evaluation and monitoring (through questionnaires), paper writing and presentation.

Teaching language(s)

Portuguese and/or English

Number of classes

12 to 14 classes, 3 hours each (MA)

26 to 30 classes, 2 hours each (BA)

Number of ECTS

10 ECTS (MA)

6 ECTS (BA)

Basic bibliography:

Beloso, B. S. (2015). Designing, Describing and Compiling a Corpus for English Architecture. In Procedia - Social and Behavioral Sciences 198. Elsevier. 459-464.

Ebensgaard Jensen, K. (2014). Linguistics and the digital humanities: (Computational) corpus linguistics. MedieKultur: Journal of Media and Communication Research, 30, pp. 117-136.

Hinrichs, E. M. Hinrichs, S. Kübler & T. Trippel (eds.) (2019). *Language Resources and Evaluation: Language Technologies for Digital Humanities 53(4).* https://doi.org/10.1007/s10579-019-09482-4

McEnery, T. & A. Hardie (2012). Corpus Linguistics: Method, theory and practice. Cambridge University Press.

Odebrecht, C., Belz, M., Zeldes, A., Lüdeling, A. & Krause, T. (2017). RIDGES Herbology: Designing a Diachronic Multi-Layer Corpus. In: Language Resources and Evaluation 51.3, pp. 695–725.

O'Keeffe, Anne & Mc Carthy, Michael (eds) (2010). The Routledge Handbook of Corpus Linguistics (Routledge Handbooks in Applied Linguistics). London & New York: Routledge.

Sinclair, J. (2004). Trust the text: language corpus and discourse. London & New York: Routledge