Universitat Pompeu Fabra Barcelona

Academic Year/course: 2022/23

32288 - Corpora and Computational Tools

Syllabus Information

Academic Course: 2022/23
Academic Center: 803 - Masters Centre of the Department of Translation and Language Sciences
Study: 8039 - Master in Discourse Studies: Communication, Society and Learning
Subject: 32288 - Corpora and Computational Tools
Credits: 5.0
Course: 1
Teaching languages:

Theory:	Group 1: English
Seminar:	Group 101: English
	Group 102: English

Teachers: Maria Nuria Bel Rafecas
Teaching Period: First Quarter
Schedule:

Presentation

The course Corpus and software tools is about methodology for carrying out empirical, corpus-based research on linguistics and applied linguistics. In particular is about the use of software programs as basic tools to handle large quantities of text data.

The objective of the course is to get students acquainted with the rationale behind the programs and tools that assist researchers when handling linguistic data coming from an object of study called a corpus, which is a collection of texts.

Associated skills

CGS1. Creativity for postgraduate research and professional practice

CT1.6. Ability to design and review processes systematically

CE. Develop with the methodology of argumentative and empirical linguistic analysis.

Learning outcomes

— Application of state-of-the-art criteria for corpus design and use of tools to compile a corpus for specific purposes

— Use of concepts, such as representativeness and significance, for empirical research in linguistics

— Definition of requirements needed to find tools (and information sources) and functionalities to utilize them.

— Discovery, installation and use of tools that perform typical Corpus Linguistics functions, including the understanding of pattern matching with Regular Expressions and corpus annotation tools

— Familiarity with terminology of NLP and Text Processing.

Sustainable Development Goals

#ODS4 Educació de qualitat

4.3 Per al 2030, assegurar l'accés en condicions d'igualtat per a tots els homes i les dones a una formació tècnica, professional i superior, inclosa la universitat, accessible i de qualitat.

4.4 Per al 2030, augmentar substancialment el nombre de joves i adults que tenen les competències necessàries, en particular tècniques i professionals, per accedir a l'ocupació, el treball decent i l'emprenedoria

#ODS5 Igualtat de gènere

5.b Millorar l'ús de la tecnologia instrumental, en particular la tecnologia de la informació i les comunicacions, per promoure l'apoderament de la dona

#ODS9 Indústria, innovació i infraestructura

9.5 Augmentar la recerca científica i millorar la capacitat tecnològica dels sectors industrials de tots els països, en particular els països en desenvolupament, entre altres coses fomentant, per al 2030, la innovació i l'augment substancial del nombre de persones que treballen en el camp de la investigació i el desenvolupament per cada milió de persones, així com augmentant les despeses en recerca i desenvolupament dels sectors públic i privat

Prerequisites

No prerequisits

Contents

Section 1

— What is a corpus? Why to use computers?

— Tools for basic functions. Keyword in Context, KWIC and concordances. Count frequencies of words. Significance of frequency related to contexts. Count frequencies of sequences of words. Count frequencies of sequences of words that are specially related, i.e. collocations. Assessment of the strength of a relation, i.e. Mutual Information. Pattern search and Regular Expressions.

Section 2

— Representativeness, balance and sampling. Reference corpus. Most well-known reference corpus and other sources of texts.

— Types of corpora. General corpora. Specialized corpora. Written corpora. Spoken corpora. Synchronic corpora. Diachronic corpora. Learner corpora. Monitor corpora. Copyrights and other legal issues.

Section 3

— Corpus mark-up. From character encoding to Corpus mark-up languages. Metadata for describing corpus.

— Corpus annotation. Levels of Linguistic Annotation. Tools for the annotation of corpora.

Section 4

— Parallel corpora and specific tools. Tools for finding parallel texts. Alignment of parallel texts. Exploitation of parallel corpora.

Teaching Methods

The main characteristics of the course are the following: The course is based mainly on practical exercises for the student to acquire the competences listed in section 3 of this document. Since *competence* is defined as a learned ability to adequately perform a task and encompasses knowledge, skills and attitudes, the goal of this course is for the students to be able to successfully perform specific corpus-based related tasks using processing tools. The class time will be devoted to the introduction of contents regarding these tools. The seminar time will be devoted to discussions and exercises.

The course will be organized in two main blocks that roughly correspond to 5 weeks each. In the first half, we will follow two selected studies (and publications), which compose a quick introduction to prototypical tools following experiments/studies made by others. Thus, the students will work on exercises following the content introduced in classes.

The second half of the course requires the students to define an experiment that involves the definition and creation of a corpus and its exploitation by means of the tools they have learned about in the first part of the course. In order to evaluate his or her project, each student will be required to write a paper about it. The paper will be peer-reviewed by other students of the course (using guidelines based on current peer review process for outstanding conferences).

Evaluation

Main evaluation will be based on getting evidence on the acquisition of the competences mentioned before and the final mark will be assessed from the following ratios:

a. Homework assignments: 45%

b. Final project, the paper: 45%

c. Participation in the final project peer-evaluation process: 10%

In case of failure after the main evaluation, the student must deliver a (revised) final project in two months time.

Bibliography and information resources

Nice short introduction to corpus and use of tools by Tony McEnery:

https://www.youtube.com/watch?v=YJTM3i5HxsQ

Anatol Stefanowitsch, 2019, Corpus Linguistics: A guide to the methodology. Language Science Press. Available http://langsci-press.org/catalog/book/148

We will mainly use SketchEngine. https://www.sketchengine.eu/guide/single-sign-on-sso/
Students will be provided with a licence to use it.

Anthony, Laurence. (2013). “A critical look at software tools in corpus linguistics.” Linguistic Research 30(2), 141-161. http://www.laurenceanthony.net/research/20130827_linguistic_research_paper/linguistic_research_paper_final.pdf

O’Keeffe A, McCarthy M. The Routledge handbook of corpus linguistics. Second edition. O’Keeffe A, McCarthy M, editors. London ;: Routledge; 2022.