Universitat Pompeu Fabra Barcelona

Academic Year/course: 2021/22

32288 - Corpora and Computational Tools

Syllabus Information

Academic Course: 2021/22
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

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

Presentation

The course "Programs and Tools for Corpus Analysis" is part of the "Màster en Lingüística Teòrica I Aplicada" of the Departament de Traducció I Ciències del Llenguatge of the Universitat Pompeu Fabra.

The course studies the methodology to carry out empirical, corpus-based research on linguistics and applied linguistics. It has a particular focus on specific software as basic tools used to handle large quantities of text data as well as to exhaustively analyze them to find particular information.

The aim of the course is to offer the fundamentals that empower the student to autonomously use both current and future tools to handle and exploit text data.

The course is taught in English.

Associated skills

The students will acquire the following competences:

Generic competences:

- Analysis and synthesis: The student will acquire the skills needed to ask thoughtful questions, as well as the knowledge of the steps required to obtain a well-founded answer.
- Critical reasoning: The student will develop skills to determine the meaning and significance of what is observed or expressed; or concerning a given inference or argument, to determine whether there is adequate justification to accept a conclusion as true.
- Autonomy: The students will develop skills to create their own routines for learning, as well as to search and find both information and sources of information.
- Use of computing tools for their needs.

Specific competences:

— 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.

Learning outcomes

The aim of the course is to offer the fundamentals that empower the student to autonomously use both current and future tools to handle and exploit text data. The outcome is students who are be able to successfully perform specific corpus-based related tasks using processing tools.

Prerequisites

The course is taught in English, but we will use texts in other languages.

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 in section 3 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

Introduction to corpus linguistics

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge UP.

Section 1

Definitions of Corpus:

John Sinclair (2005) "Corpus and Text - Basic Principles" in Martin Wynne (ed.)"Developing Linguistic Corpora: a Guide to Good Practice". Oxford Oxbow Books. [Available online from Available online from http://ahds.ac.uk/linguistic-corpora/, last Access September 2010]

Tools:

John Sinclair,1982. Reflections on computer corpora in English language research. In *Computer corpora in English language research*, ed. Stig Johansson: 1-6. Bergen.

Christopher D. Manning, Hinrich Schütze. 1999. Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press, Chapter 1.4. "Dirty Hands".

Church, K. i P Hanks. Word association norms, mutual information and lexicography. En Proceedings of the 27th Annual Meeting of the ACL, pg. 76-83. (1989).

Daniel Jurafsky and James H. Martin. 2000. Speech and Language Processing. Prentice Hall. Chapter 2.1. Regular Expressions.

Jorge Vivaldi (2009). "Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de corpus textuales" dins *Tradumática* (7). Bellaterra: Universitat Autònoma de Barcelona, Departament de Traducció i d'Interpretació. ISSN 1578-7559
http://webs2002.uab.es/tradumatica/revista/num7/articles/10/10art.htm

Oakes, Michael P. (1998) Statistics for corpus linguistics Edinburgh: Edinburgh University Press, cop. 1998

Section 2

Representativeness:

D. Biber. 1993. "Representativeness in corpus design". Literary and Linguistic Computing 8/3: 243-257. [a pdf copy can be found at internet and an extract of the paper is legally reprinted in McEnery, Xiao and Tono (2006): Corpus-Based Language Studies. Routledge]

William H. Fletcher (2010) Corpus Analysis of the World Wide Web, to appear in Chapelle, Carol A, (Ed.). (2011). Encyclopedia of Applied Linguistics. Wiley-Blackwell.http://www.encyclopediaofappliedlinguistics.com/ and available at: http://webascorpus.org/Corpus_Analysis_of_the_World_Wide_Web.pdf

Adam Kilgarriff. Googleology is bad science. *Computational Linguistics.* 33, 1 (Mar. 2007), 147-151. DOI= http://dx.doi.org/10.1162/coli.2007.33.1.147

Some reference corpora:

[BNC] British National Corpus (BNC):http://www.natcorp.ox.ac.uk/

[CORDE] Corpus Diacrónico del Español (CORDE):http://corpus.rae.es/cordenet.html
[CREA] Corpus de Referencia del Español Actual (CREA):http://corpus.rae.es/creanet.html

[CORPES XXI] Corpus del Español del Siglo XXI (CORPESXXI): http://www.rae.es/recursos/banco-de-datos/corpes-xxi

[ICE] International Corpus of English: http://www.ucl.ac.uk/english-usage/ice/

[CITLC] Corpus Informatitzat de la Llengua Catalana: http://ctilc.iec.cat/

[Europarl] http://www.statmt.org/europarl/

Types of corpora:

S. Atkins, J. Clear and N. Ostler (1991) "Corpus design criteria". http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf [last access September 2010]. In 1992 it was published in Literary and Linguistic Computing 7/1: 1-16.

Section 3.

Jorge Vivaldi Palatresi (2009). "Corpus and exploitation tool: IULACT and bwanaNet" dins Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.) *A survey on corpus-based research = Panorama de investigaciones basadas en corpus [Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 Mayo 2009, Universidad de Murcia]*. Murcia: Asociación Española de Lingüística del Corpus. Pàg. 224-239. ISBN 978-84-692-2198-3

Tony McEnery, Richard Xiao and Yukio Tono (2006): Corpus-Based Language Studies. Routledge. Unit A3 and A4.

Lou Burnad (2005) "Metadata for Corpus Work" in Martin Wynne (ed.) "Developing Linguistic Corpora: a Guide to Good Practice". Oxford Oxbow Books. [Available online from Available online from http://ahds.ac.uk/linguistic-corpora/, last Access September 2010]

Section 4

Gómez Guinovart, Xavier and Alberto Simões (2009): Parallel corpus-based bilingual terminology extraction. In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*, IRIT (Institut de recherche en Informatique de Toulouse), Université Paul Sabatier, Toulouse. http://webs.uvigo.es/sli/arquivos/TIA09.pdf

Bowker, Lynne; Pearson, Jennifer (2002). Working with Specialized Language : A Practical Guide to Using Corpora. London; New York: Routledge, 2002.

We will mainly use Sketchengine.

www.sketchengine.co.uk

Students will receive user and password.

http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex