Corpus Linguistics and its Applications

6 credits,
2nd semester

Dra. Belén López Arroyo (mariabelen.lopez@uva.es)
Universidad de Valladolid

Dra. Isabel Pizarro-Sánchez (isabel.pizarro@uva.es)
Universidad de Valladolid

Aims and Objectives

The term "corpus" is mostly used to refer to a relatively large collection of naturally-occurring texts, which have been stored in machine-readable form (McEnery & Hardie, 2011). In this form, the texts are then studied using various computer programs, in the branch of linguistics known as "corpus linguistics". The course focuses on two main theoretical aspects, approaches to corpus linguistics and main corpus linguistics applications and uses, and a practical approach to corpus compilation and exploitation (tools).

When students have successfully completed this course, they should be able to:

- make informed and critical use of central terms like *concordance, n-grams, type, token, type/token ratio, annotation, abstraction, statistics, collocation, constructional analysis, keyword, keyness, pattern grammar, LLM*, and so on;
- analyze corpus extracted data;
- examine how corpus data can be used:
- explore the complexity of corpus work;
- use and analyze LLM data as a corpus;
- demonstrate the ability to think critically about the diverse ways corpus data are adequate;
- generate critical ideas for analyzing contemporary Corpus Linguistics.

Course Program

1) Corpus Linguistics today and Corpus Technology.
   a) What is a corpus?
   b) Types and uses of corpora
   c) History of Corpus Linguistics
   d) The Use of Corpora in Linguistics: Grammar; Lexicography; Language Teaching; language Learning; Sociolinguistics; Pragmatics; Discourse; and so on.
   e) Corpus linguistics applications: Machine translation, Computer assisted translation, Translation Studies, Contrastive Analysis, etc.
2) How to build a Specialized Corpus: The Internet as a linguistic and documentation resource
3) How to manage existing corpora: COCA-BNC, CREA, etc.
4) Artificial Intelligence and Corpus Linguistics
5) Tools for corpus linguistics analysis
6) Data analysis

Methodology

Class meetings will consist of lectures, group discussions, individual and group work, and oral presentations. The course will put strong emphasis not only on oral discussions, but also on activities designed to stimulate the students' critical thinking and writing skills. Since regularly we have an international mix of students in the course, there will be ample opportunity for participants to share their own ideas and experiences, and to bring these to bear on the analysis of corpus-based and corpus-driven data.

Assessment

A selection of texts (including but not limited to those indicated above) will be specified at the beginning of the course for class discussion. Students are required to read each assigned text before class and to come prepared to discuss it.

The students will be evaluated on a combination of their participation in class and written assignments. Specifically, students will be assessed based on the following assignments:

Dra. Belén López Arroyo, Universidad de Valladolid
E-mail: mariabelen.lopez@uva.es
Dra. Isabel Pizarro-Sánchez, Universidad de Valladolid
E-mail: isabel.pizarro@uva.es

First assignment (10%)

- Reading the chapters included in the required reading section and answer the corresponding tasks described in Moodle (Campus Virtual, UVa).
- Answering short tasks (corpus queries)

Second assignment (30%)

- Corpus research project proposal, including its compulsory oral presentation. Your project should include at least the following:
  - Research question(s)
  - Motivation of your research (Why)
  - Corpus / corpora for this research (including a description of your own corpus)
  - Methods, including the description of any software used in the process (How, steps)
  - Expected results / conclusion
  - Implication (applications)
- Compilation of a small, specialized corpus for the project proposal. Aim: to learn the process.

Third assignment (60%)

Students must write an essay (final paper) describing the process of compiling and exploiting a specialized corpus (second assignment), explaining:

- The purpose for the corpus compilation (research question that leads to compile your corpus)
- The criteria for compilation (connected to the research question)
- How corpus files are organized (names, folders…)
- Any other significant issue (tagging…) worth commenting.
- Answer their research question (or explain how to do it if it were not possible)
- Other possible uses and applications of their corpus and how to extract information from it

Important:

Assignments must be submitted in English and uploaded on the Campus Virtual-Moodle page (University of Valladolid, https://campusvirtual.uva.es/), no e-mail works will be accepted. Detailed instructions will be provided in class and via the Moodle platform. All assignments must include a bibliography, and wherever necessary, appendices and tables should be provided. However, none of these additional pieces of information should figure into the page count (see section I). The bibliography should include all consulted references.

All information appearing in the assignments, including those examples taken from actual sources (e.g. books, articles, newspapers, etc.) and course notes must be properly documented. More specifically, sources, authors, and other bibliographical information must be cited in order to avoid

Dra. Belén López Arroyo, Universidad de Valladolid
E-mail: mariabelen.lopez@uva.es
Dra. Isabel Pizarro-Sánchez, Universidad de Valladolid
E-mail: isabel.pizarro@uva.es

plagiarism. Any student who is discovered to have plagiarized either all or part of his/her assignment will be severely reprimanded.

Bibliography and Resources

Anderson, W., and J. Corbett. 2009. *Exploring English with On-line Corpora: An Introduction*. Palgrave Macmillan.

Baker, Paul, Andrew Hardie and Tony McEnery. 2006. *A Glossary of Corpus Linguistics*. Edinburgh University Press.

Barnbrook, Geoff, Pernilla Danielsson and Michaela Mahlberg (eds.). 2005. *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Continuum.

Biber, Douglas. 2006. *University Language. A Corpus-based Study of Spoken and Written Registers*. John Benjamins.

Biber, Douglas, et al. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language. A Practical Guide to Using Corpora*. Routledge.

Brown, David West, and D. Zawodny Wetzel (eds.). 2023. *Rhetorically Informed Text Analysis*. John Benjamins.

Corpas Pastor, Gloria, and Miriam Seghiri Domínguez. 2010. "Size Matters: A Quantitative Approach to Corpus Representativeness." *Lengua, traducción, recepción. En honor de Julio César Santoyo*, edited by Rosa Rabadán, pp. 112-146.

Crawford, William J., and Eniko Csomay. 2024. *Doing Corpus Linguistics*. Routledge.

Curry, Niall, Paul Baker, and Gavin Brookes. 2024. "Generative AI for Corpus Approaches to Discourse Studies: A Critical Evaluation of ChatGPT." *Applied Corpus Linguistics*, vol. 4, no. 1, https://doi.org/10.1016/j.acorp.2023.100082.

Dunn, Jonathan. 2022. *Natural Language Processing for Corpus Linguistics*. Cambridge University Press, https://doi.org/10.1017/9781009070447.

Fuertes-Olivera, Pedro A. 2008. "Pedagogical Application of Corpora in ESP Teaching: The Case of the UVaSTECorpus". *Scripta Manent. Journal of the Slovene Association of LSP Teachers 3*(2): 2-15. Editor: Slovene Association of LSP Teachers.

Fuertes-Olivera, Pedro A. and Rodrigues Rodrigues, José María. 2010. "Working with English Specialized Corpora: lexical bundles in Written Scientific and Technical English" In Sánchez, Aquilino & Almela, Moisés (eds), *A Mosaic of Corpus Linguistics. Selected Approaches*, 121-136. Peter Lang.

Gavioli, Laura. 2005. *Exploring Corpora for ESP Learning*. John Benjamins.

Jones, Christian, and Daniel Waller. 2015. *Corpus Linguistics for Grammar: A Guide for Research*. Routledge.

McEnery, T. & A. Hardie. 2011. *Corpus Linguistics: Methods, Theory and Practice*. Cambridge University Press.

Meyer, Charles. F. 2023. *English Corpus Linguistics*. Cambridge University Press.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Analysis*. Routledge.

Zufferey, Sandrine. 2020. *Introduction to Corpus Linguistics*. Willey.

Schedule

The course meets twice a week, in the seminar room of the English Department. We will have 10 three-hour sessions, to be distributed in the five weeks of the second teaching period (second semester). The schedule is posted on the Internet website for the master's program.

Dra. Belén López Arroyo, Universidad de Valladolid
E-mail: mariabelen.lopez@uva.es

Dra. Isabel Pizarro-Sánchez, Universidad de Valladolid
E-mail: isabel.pizarro@uva.es