Corpus Linguistics and Its Applications

6 creds.,
2nd semester

Universidad de Valladolid

Dr. Pedro A. Fuertes-Olivera
University of Valladolid and University of Stellenbosh (South Africa)
pedro@tita.emp.uva.es

Dr. Isabel Pizarro Sánchez
Universidad de Valladolid
pizarro@lia.uva.es

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". In addition to size, there are three important ways in which the various existing synchronic English corpora differ: the genre of the texts included (whether they are specialized or not), the inclusion of either whole or sampled texts, and whether the corpus is added to over time.

The most usual way of studying a corpus is by using a concordancers, search engines, Text-analysis tools, etc. (http://www.uow.edu.au/~dlee/software.htm) . For instance, concordance citations can be sorted in various ways. For example, they can be left-sorted, i.e. sorted by the word immediately preceding the node. Where the node is a noun, or predominantly a noun, left-sorting is interesting because it enables the researcher to see quickly which groups of words premodify the node. If the node is a verb, it seems interesting right-sorting, since it focuses on the object of the verb. It is generally good practice to examine the concordances sorted in several ways, because different pieces of information about the word usually emerge (Bowker & Pearson, 2002; Kilgarriff, 2012). For example, when two or more words regularly appear in each others' environment, they are known as *collocates*; the phenomenon is known as *collocation* (Sinclair, 1991). Data can be studied in the form of an overview, or "picture" of the most frequent collocates, to the left and right of the node. It is possible to focus on a word of interest in the picture, and extract citations of the collocation. Data presented in this form are a very useful starting point for an investigation of word meaning and use, particularly if they are interpreted in terms of underlying systems, be these already described or not (Biber et al., 1999).

Tognini-Bonelli (2001) differentiates between two types of corpus work: corpus-based and corpus-driven. A corpus-based research starts with existing paradigms and investigates these using the corpus. Corpus-driven research, on the other hand, starts with a clean slate, with no assumptions about what will be found: it places the corpus at the centre of the process, and allows new categories and rules to emerge from the study. When examined carefully, this distinction tends to slip. Deignan (2005: 90), for example, claims that the metaphor research she describes has features of both corpus-based and corpus-driven work:

> It is corpus-based, in the sense that it begins with categories developed in the literature and explores them, rather than starting with a clean slate theoretically and taking the research agenda from some kind of statistical overview of the corpus. However, it is corpus-driven in the sense that it does not seek to maintain existing categories at all costs, but is prepared to reclassify the data and develop new systems of description if the data are found to contradict existing ideas.

As a relatively new approach to language studies, corpus linguistics has witnessed that the number and depth of many corpus approaches to the study of the English language is constantly increasing. Three different stages can be observed in the history of corpus linguistics. The first wave in corpus linguistics starting in the 1960s focused on developing computerized general corpora consisting of different types of spoken and written texts. For example, the Brown Corpus consists of 2,000 word samples of 500 texts which are spread across 15 categories. The second wave of general corpora, which started in the 1980s, was developed in the 1990s, taking advantage of the technological advances of computing. It has produced mega-corpora such as the 450 million-word Bank of English Corpus, or the 100 million-word British National Corpus. Unlike many early corpora, they contain complete texts rather than sections of texts. The third wave starting in the 2000s focused on both developing *giga-corpora* (i.e., corpora of texts over a billion word using websites and newswire texts as data sources[1]), and small specialized corpora designed for studies of Academic and Professional English.

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, pattern dictionary, translation memory.

-- analyze corpus extracted data

-- examine how corpus data can be used.

-- explore the complexity of corpus work.

-- demonstrate the ability to think critically about the diverse ways corpus data are adequate. For instance, Fuertes-Olivera (2012), Fuertes-Olivera and Nielsen (2012), and Fuertes-Olivera and Tarp (2014) have shown that corpus-based or corpus driven specialized dictionaries (or terminological knowledge bases, ontologies, and the like) are still far from being real products or tools for human consultation.

-- generate critical ideas for analyzing contemporary Corpus Linguistics.

Course Program

1. Corpus Linguistics today and Corpus Technology.
   1.1. What is a corpus?
   1.2. Types and uses of corpora
   1.3. Corpus linguistics applications: Machine translation, Computer assisted translation, Translation Studies, Contrastive Analysis, etc.

[1] For example, the *English Gigaword Corpus*, produced by the Linguistic Data consortium See: < http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05>,

2. How to build a Specialized Corpus
   - 2.1. The Internet as a linguistic and documentation resource
   - 2.2. How to manage a corpus: CREA and BNC
   - 2.3. Information retrieval software
3. Tools:
   - 3.1. Terminology managers
   - 3.2. Concordances
   - 3.3. Translation memories
   - 3.4. Alignment software
   - 3.5. Language search engine software

Required Readings

McCarthy, M. & A. O'Keeffe. (2010). Historical perspective: what are corpora and how have they evolved? In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, pp. 3-13.

Nelson, M. (2010). Building small specialized corpora. In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, pp. 53-65.

Tognini Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, pp. 14-27.

Walter, E. (2010). Using corpora to write dictionaries. In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, pp. 428-443.

One chapter chosen by the student from: A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge

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

Methodology

Class meetings will consist of lectures, group discussions, 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 of space and mobility, 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. Students will have to prepare two assignments.

The students will be evaluated on a combination of their participation in class, and the written assignments.

First assignment

Read the chapters included in the required reading section and write an essay with a summary of what you read and the applications corpus might have.

Second assignment

1. Build a specialized corpus
2. Size: a minimum of 250.000 words
3. Students must include and attachment describing the process and explaining:
    a. The purpose for the compilation
    b. The criteria for compilation
    c. The uses and applications of their corpora.

Assignments must be:
- submitted in English and uploaded on the Moodle page
- double-spaced
- submitted in Courier 12 pt. font (or similar)

2. The following information must appear at the top left corner of the first page of the actual assignments (please do not include a title page):
- student's name and academic year
- title of the course
- professor's name
- date
- title of the assignment

N.B.: Headings should be omitted.

3. 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.
N.B. The following are examples of bibliographical references (students will not be penalized for following other conventions as long as those conventions are applied consistently):

a. A book by a single author:
NEGROPONTE, N. (1995). *Being Digital*. New York: Vintage Books.
b. A book by two different authors:
LEECH, G. & J. SVARTVIK (1985). *A Communicative Grammar of English*. London: Longman.
c. A book by three or more authors:
QUIRK, R. et al. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
d. A dictionary entry:
"Azimuthal Equidistant Projection." (1980 ed.): *Webster´s New Collegiate Dictionary*.

e. An article in a periodical:
BEGLEY, S. (1982): “A Healthy Dose of Laughter.” *Newsweek.* 4 Oct.: 74-75.
f. A World Wide Web (WWW) Site:
Callies, M. (2003) “Introduction to linguistics” <http://staff-www.unimarburg. de/~callies> (22 Jul 2005)

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

Bibliography and Resources

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

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

Biber, Douglas (2006). *University Language*. *A Corpus-based Study of Spoken and Written Registers*. Amsterdam / Philadelphia: John Benjamins.

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

Fuertes-Olivera, Pedro A. (2007). “A corpus-based view of lexical gender in written Business English”. *English for Specific Purposes, 26 (4): 219-234*.

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. & Pérez Cabello de Alba, Beatriz (2011). “A Corpus Analysis of Prototical Causation in Written Scientific and Technical English”. *Revista Española de Lingüística Aplicada, RESLA* 24: 73-94.

Fuertes-Olivera, Pedro A. & 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. Frankfurt am Main. Berlin. Bern. Bruxelles. New York. Oxford. OEIN: Peter Lang. ISBN: 978-3-631-58789-8.

McEnery, T. & A. Hardie (2011). *Corpus Linguistics: Methods, Theory and Practice*. Cambridge: CUP.

O’Keeffe, A. & M. McCarthy 2010). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge.

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 third teaching period (second semester). The schedule is posted on the Internet website for the master's program.