General information

This course is an introduction to computerized research methods, which are applied to large data bases of language used in natural communicative settings to supplement more traditional ways of linguistic analysis in all linguistic sub-disciplines. In the first part of this particular class, we will begin with a theoretical introduction: what is a corpus / what are corpora, what kinds of corpora are there and how are they created/compiled, and why would one use corpora in the first place? In the second part, we will familiarize ourselves with the open source programming language and environment R. In the third part, we will read a variety of simple but published corpus-linguistic studies as well as replicate, modify, or extend them. The topics to be covered include syntax (patterns and alternations), lexis/semantics (key words in different cultures and near synonymy), psycholinguistics (disfluencies), and others.

Note$_1$: This course is based on the second edition of my textbook *Quantitative corpus linguistics with R: a practical introduction*. New York: Routledge, Taylor & Francis Group, which you will need to have: it will teach you most fundamentals of R programming for text analysis (and can therefore be useful way beyond this course) and contains all readings for the 1st half of the course as well as additional answer keys and exercises for parts of the 2nd half.

Note$_2$ and this is very important: We will be using a programming language, which means that the course absolutely requires computer literacy beyond swiping, pinching, long-tapping, and uploading/sending something to/via Facebook, Instagram, Pinterest, Snapchat, or whatever: If you cannot install software, or if you can install software but then don't know 'where the program is', and/or if you download a file on your own personal computer but will then ask me where it went, and/or if you do not know what *unzipping a file* means (not just opening it, *unzipping*!), you will not be happy in this course!

Course requirements

i. submission of four to-be-graded assignments (that will also serve as your preparation for one of the programming sessions):
   - <120_07_cult-word-freq.pdf> with a deadline of: 13 May 2021, 18:00 PST;
   - <120_08_ic-ical-adj..pdf> with a deadline of: 20 May 2021, 18:00 PST;
   - <120_09_dispersion.pdf> with a deadline of: 27 May 2021, 18:00 PST;
   - <120_10_prenominal-adj.pdf> with a deadline of: 03 June 2021, 18:00 PST;

   (each of these accounts for 15% of your grade)

ii. a functioning and properly organized R script with code and other answers in comments that address the problem/questions of one (!) of the three assignments, which accounts for the remaining 40% of your grade; the deadline for this is 12 June 2021, 18:00 PST.

Contact

Office hours Prof. Gries: by appointment
Web: <http://www.stgries.info>
Email: <stgries@linguistics.ucsb.edu>

Office hours: Chadi Ben Youssef: TR,14:00-15:00, with an appointment made via Nectir <https://ucsb.nectir.io/invite/N5usA3>
Web: <https://cbenyoussef.wordpress.com>
Email: <chadi@ucsb.edu>

Course plan

(1) 04/02: introduction to the course: what are corpora? what kinds of corpora exist? web as corpus; examples of corpora, compiling corpora, why use corpora
Review: <120_01_intro-corp.pdf>, QCLWR2: pp. 1-20
Prep.: QCLWR2: pp. 21-49 (exercises boxes are not obligatory!)
Install R and RStudio and download the files from the book's companion website at <http://www.stgries.info/research/qclwr/qclwr.html>

(2) 04/09: introduction to R (part 1)
Review: QCLWR2: Chapter 3, as far as we got

(3) 04/16: no class
Review: QCLWR2: Chapter 3, as far as we got

(4) 04/23: introduction to R (part 2)
Review: QCLWR2: Chapter 3, as far as we got

(5) 04/30: introduction to R (part 3)
Review: QCLWR2: Chapter 3

(6) 05/07: application(s): psycholinguistics: disfluencies
Prep.: Leech & Fallon (1992) and <120_07_cult-word-freq.pdf>

(7) 05/14: application(s): cultural word frequency comparisons
Optional review: QCLWR2: Section 5.2.6
Oblig. prep: Kaunisto (1999), <120_08_ic-ical-adj.pdf>

(8) 05/21: application(s): concordancing *-ic*/*-ical* adjectives
Optional review: QCLWR2: Sections 5.2.7 & 5.3.2, Gries (2001 or 2003)
Prep: Gries (to appear), <120_09_dispersion.pdf>

(9) 05/28: application(s): dispersion
Optional review: QCLWR2: Section 5.1
Prep: Wulff (2003) and <120_10_prenominal-adj.pdf>

(10) 06/04: application(s): prenominal adjective order
Optional review: QCLWR2: Section 5.4.5

Articles to read as well as assignment files can be found on the course website. The reference section below also lists some additional and optional interesting readings.

Main course book

Gries, Stefan Th. 2016. *Quantitative corpus linguistics with R: A practical introduction*. 2nd rev. & ext. ed. London, New York: Taylor and Francis.

References for the case study sessions

Gries, Stefan Th. 2001. A corpus-linguistic analysis of -*ic* and -*ical* adjectives. *ICAME Journal* 25. 65-108.

Gries, Stefan Th. 2003. Testing the sub-test: a collocational-overlap analysis of English -*ic* and -*ical* adjectives. *International Journal of Corpus Linguistics* 8(1). 31-61.

Gries, Stefan Th. to appear. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.). *Practical handbook of corpus linguistics*. Berlin & New York: Springer.

Kaunisto, Mark. 1999. *Electric*/*electrical* and *classic*/*classical*: variation between the suffixes -*ic* and -*ical*. English Studies 4. 343-370.

Leech, Geoffrey & Roger Fallon. 1992. Computer corpora: What do they tell us about culture? *ICAME Journal* 16:29-50.

Wulff, Stefanie. 2003. A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics* 8(2). 245-282.

Optional references

Atkins, Beryl T. Sue, Jeremy Clear, & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1-16.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5(4). 257-269.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243-257.

Church, Kenneth W. & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22-29.

McEnery, Tony & Andrew Hardie. 2011. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, Anthony, Richard Ziao, & Yukio Tono. 2006. *Corpus-based language studies: an advanced resource book*. London, New York: Routledge.

Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Stefanowitsch, Anatol. 2005. New York, Dayton (Ohio), and the Raw Frequency Fallacy. *Corpus Linguistics and Linguistic Theory* 1(2). 295-301.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press.