General information

This course is an R programming-based introduction to corpus-linguistic 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. It is broadly based on the second 2016 edition of my textbook *Quantitative corpus linguistics with R: a practical introduction* and McEnery & Hardie's (2012) *Corpus Linguistics*, supplemented with a variety of research articles.

The first four sessions are devoted to learning enough R programming and text processing to be able to implement the four main corpus-linguistic methods: frequency lists, dispersion measures, key words, and co-occurrence phenomena. After that, we will read a selection of corpus-linguistic papers that involve these methods and (i) will briefly discuss them and (ii) then try to write R scripts that allow to replicate them with corpora of different formats.

Thus, the course aims at enabling you (i) to understand and replicate corpus-linguistic work, (ii) to pursue your own corpus-linguistic studies on a wide variety of data, and (iii) to acquire basic skills in programming and regular expressions, which are extremely useful within and outside academia.

Course requirements

i. A review (≈6,000 characters) of an empirical corpus-linguistic paper that was published in *Corpus Linguistics and Linguistic Theory*; the review must conform to *CLLT*'s style sheet (30% of your final grade);
ii. a small paper (≈15,000 characters) with a small empirical corpus-linguistic study that uses an R script you wrote (also to be submitted). The paper must have the structure 'introduction-methods-results-discussion', and must conform to *CLLT*'s style sheet (70% of your final grade);
iii. attendance, participation in class, and doing the homework readings/assignments.

The review and the final paper are due in electronic format on 13 June at 07.00 (that is A.M. and PDT) in my inbox. I strongly encourage you to discuss ideas concerning the reviews as well as the topic, technicalities, and other issues regarding the research paper with me *as early as possible*.

Contact
Office hours: Wed 14:30-15:30 (make an appointment for Zoom)
Web: <http://www.stgries.info>
Email: <stgries@linguistics.ucsb.edu>

Course plan

(1) 03/31: Processing (corpus) data in R 1
Read for next time: Gries (2016: Chapter 1, 3 as far as we got plus exercises), McEnery & Hardie (2012: Ch. 1)
Suggested reading(s): Atkins, Clear, & Ostler (1992); Biber (1993)

(2) 04/07: Processing (corpus) data in R 2
Read for next time: Gries (2016: Chapter 3 as far as we got plus exercises), McEnery & Hardie (2012: Ch. 2)

(3) 04/14: Processing (corpus) data in R 3
Read for next time: Gries (2016: Chapter 3 as far as we got plus exercises), McEnery & Hardie (2012: Ch. 3, 4)

(4) 04/21: Wrap-up and practice
Read for next time: Miller (to appear), Brezina & Gablasova (2015)

(5) 04/28: Frequencies
Read for next time: Adelman et al. (2006), Gries (to appear a)
Suggested reading(s): Kuperman & Van Dyke (2013), Gries (to appear c)

(6) 05/05: Measures of dispersion
Read for next time: Leech & Fallon (1992), Rayson & Potts (to appear)
Suggested reading(s): McEnery & Hardie (2012: Ch. 5, 6)

(7) 05/12: Key words
Read for next time: Bybee & Scheibman (1999), Jurafsky et al. (2001)
Suggested reading(s): Bell et al. (2003)

(8) 05/19: Concordancing and collocations
Read for next time: McEnery & Hardie (2012: Ch. 8)
Suggested reading(s): Colleman & Bernolet (2012)

(9) 05/26: Anonymous functions and parallelizing
Read for next time: McEnery & Hardie (2012: Ch. 7), Baayen (2010), revisit session 6
Suggested reading(s): Stefanowitsch (2005, 2006)

(10) 06/02: Brute-force stupid tagging
Read for 'next time': McEnery & Hardie (2012: Ch. 9), Gries (to appear b), Gries & Paquot (to appear)

References

Adelman, James S., Gordon D.A. Brown, & José F. Quesada. 2006. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science* 17(9). 814-823.

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

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436-461.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2). 1001-1024.

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

Brezina, Vaclav & Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the new General Service List. *Applied Linguistics* 36(1). 1-22.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4). 575-596.

Colleman, Timothy & Sarah Bernolet. 2012. Alternation biases in corpora vs. picture-description experiments: DO-biased and PD-biased verbs in the Dutch dative alternation. In Dagmar S. Divjak & Stefan Th. Gries (eds.), *Frequency effects in language representation*, 87-125. Berlin & New York: De Gruyter Mouton.

Gries, Stefan Th. 2016. *Quantitative corpus linguistics with R: a practical introduction*. 2nd ed. London & New York: Routledge.

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

Gries, Stefan Th. to appear b. Managing synchronic corpus data with the British National Corpus (BNC). In Andrea L. Berez-Kroeker, Brad McDonnell, Eve Koller, & Lauren Collister (eds.), *MIT Open Handbook of Linguistic Data Management*. Cambridge, MA: The MIT Press.

Gries, Stefan Th. to appear c. On, or against?, (just) frequency. In Hans C. Boas (ed.), *Applications of cognitive linguistics*. Boston & Berlin: De Gruyter Mouton.

Gries, Stefan Th. & Magali Paquot. to appear. Writing up a corpus-linguistic paper. In Magali Paquot & Stefan Th. Gries (eds.). *A practical handbook of corpus linguistics*. Berlin & New York: Springer.

Jurafsky, Daniel, Alan Bell, Michelle Gregory, & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229-254. Amsterdam and Philadelphia: John Benjamins.

Kuperman, Victor & Julie A. Van Dyke. 2013. Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance* 39(3). 802-823.

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

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics*. Cambridge: Cambridge University Press.

Miller, Don. to appear. Analysing frequency lists. In Magali Paquot & Stefan Th. Gries (eds.). *A practical handbook of corpus linguistics*. Berlin & New York: Springer.

R Development Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Rayson, Paul & Amanda Potts. to appear a. Analysing keyword lists. In Magali Paquot & Stefan Th. Gries (eds.). *Practical handbook of corpus linguistics*. Berlin & New York: Springer.

Stefanowitsch, Anatol. 2005. New York, Dayton (Ohio), and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 1(2). 295-301.

Stefanowitsch, Anatol. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2(1). 61-77.

Tao, Hongyin. 2003. A usage-based approach to argument structure: 'remember' and 'forget' in spoken English. *International Journal of Corpus Linguistics* 8(1). 75-95.