Ling 571 Computational Corpus Linguistics

San Diego State University
Schedule # 22330
Fall 2022
TuTh 9:30am–10:45am
NE-085

Advances in technology have revolutionized the way linguists approach their data. Using computers, extremely large bodies of text ("corpora") can be collected and analyzed at a level of detail that only a decade ago would have been unthinkable. Sources like the British National Corpus, the Corpus of Historical American English, and Google Books collection allow us access to language use across an unprecedented range of time and space. For anyone studying human communication or culture, the accelerating growth of the World Wide Web and other natural language resources have made techniques for dealing with very large texts more important than ever.

Through a combination of lectures, demonstrations, and hands-on exercises, this course will give students an introduction to the skills necessary for computer-aided text manipulation. Students will learn to search text databases using on-line tools, to write Python programs to manipulate large natural language corpora, to apply quantitative linguistic measures to existing texts, and to formulate, carry out, and describe their own corpus-based linguistic research projects.

This class has no pre-requisites.

Prof. Rob Malouf

Website:	malouf.sdsu.edu
Email:	rmalouf@sdsu.edu
Office hours:	Tu Th 1:00–2:00 or by appointment
Real office:	SHW 244
Zoom office:	SDSU.zoom.us/j/88663149131

Students are provided with an SDSU Gmail account, and this SDSU email address will be used for all communications. University Senate policy notes that students are responsible for checking their official university email once per day during the academic term. For more information, please see Student Official Email Address Use Policy here: senate.sdsu.edu/policy-file/policies/facilities#collapsed20e126_12

Class rosters are provided to the instructor with the student's legal name. Please let me know if you would prefer an alternate name and/or gender pronoun.

Student Learning Outcomes

By the end of the course, students will be able to:

- Articulate the principles of good corpus design
- Analyze patterns of word use in a corpus using concordancing tools
- Implement quantitative corpus analyses in Python
- Evaluate linguistic hypotheses on the basis of corpus data
- Apply corpus-based techniques to social media texts

Course Materials

You need to acquire two required textbooks for this course:

- Vaclav Brezina. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press.
- Danielle Barth and Stefan Schnell. 2022. *Understanding Corpus Linguistics.* Routledge.

These should be available from the campus bookstore (or the usual other places) and you should do you best to get them. Other readings will be distributed via Canvas:

- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Press. www.nltk.org/book
- Allen B. Downey. 2016. *Think Python: How to Think Like a Computer Scientist (2nd edition).* Green Tea Press. greenteapress.com/wp/think-python-2e
- Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations.* MIT Press.
- Anatol Stefanowitsch. 2020. *Corpus Linguistics: A Guide to the Methodology.* Language Science Press. langsci-press.org/catalog/book/148
- Martin Wynne (ed.). 2005. *Developing Linguistic Corpora: A Guide to Good Practice.* Oxbow Books. users.ox.ac.uk/~martinw/dlc/index.htm

Course Design

Course grades will be based on regular homeworks, a take-home mid-term project, and a final project:

- Homeworks (30%)
- Midterm project (30%)
- Final project (40%)

Tentative due date for the midterm is October 20 and for the final is December 17.

Grading Policies

The final grade will be based on homework assignments (30%), a midterm project (30%), and a final project (40%). Through the term, there will be several hands-on homework assignments in which students apply the techniques learned in class to actual corpus materials. Since it's important to not get behind on assignments, late assignments will be accepted for partial credit **for one week only** after the due date unless prior arrangements are made. Working in groups is encouraged, but please include the names of all coworkers on the assignment.

The midterm will be a take-home programming assignment, for which students will be required to replicate a published corpus analysis using Python. The final project should be a program (with documentation) to perform some substantial corpus processing task chosen by the student. Alternatively, the final project can be the collection and annotation of a new corpus, or a research project that makes crucial use of novel corpus data. More details about both projects will be given later in the term.

Schedule

Module	Topic	Readings
1	Intro	Wynne, ch. 1; Barth & Schnell, ch. 1, 2
2	Word meanings	Hanks, ch. 1; Barth & Schnell, ch. 3, 4
3	Frequency	Brezina, ch. 1, 2; Barth & Schnell, ch. 5
4	Keywords and collocations	Stefanowitsch, ch. 7; Brezina, ch. 3; Barth & Schnell, ch. 9
5	Python	Bird, ch. 1, 2; Downey
6	Corpus representation	Bird, ch. 3; Wynne, ch. 3, 4
7	Annotation	Wynne, ch. 2; Barth & Schnell, ch. 7
8	Lexical dispersion	
9	Lexical diversity	
10	Lexicogrammar	Brezina, ch. 4; Stefanowitsch, ch. 8
11	Social media	Bird, ch. 5, 11

Revised: 5 August 2022