Ling 571 Computational Corpus Linguistics
Fall 2024
Schedule # 5278

SDSU

COURSE INFORMATION

Time:	TTh 9:30am–10:45am	Email:	rmalouf@sdsu.edu
Place:	NE-085	Office hours:	TTh 1:00pm–2:00pm
Mode:	In person	Real office:	SHW 201
Instructor:	Prof. Rob Malouf	Zoom office:	see Canvas

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.

ESSENTIAL STUDENT INFORMATION

For essential information about student academic success, please see the SDSU Student Academic Success Handbook.

- SDSU provides disability-related accommodations via Student Disability Services (sds@sdsu.edu |https://sds.sdsu.edu/). Please allow 10-14 business days for this process.
- 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.

COURSE MATERIALS

You need to acquire two required textbooks for this course:

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

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 linked via Canvas:

- Catherine Anderson, Ai Taniguchi, Bronwyn Bjorkman, Derek Denis, and Nathan Sanders. 2021. *Essentials of Linguistics* (2nd edition). McMaster University.
- André Block. 2012. "From the blackhand side: Twitter as a cultural conversation." *Journal of broadcasting & electronic media* 56(4): 529–549.
- Allen B. Downey. 2024. *Think Python: How to Think Like a Computer Scientist* (3rd edition). O'Reilly Media, Inc.
- Jeroen Janssens and Thijs Nieuwdorp. 2025. *Python Polars: The Definitive Guide* [early release edition]. O'Reilly Media, Inc.
- Mark Mets, Andres Karjus, Indrek Ibrus, Maximilian Schich. 2024. Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. *PLoS ONE* 19(4): e0302380.
- Sara Može. 2017 “Norms and exploitations in lexicography.” In P. Hanks, G.-M. de Schryver (eds.), *International Handbook of Modern Lexis and Lexicography.* Springer.
- Anatol Stefanowitsch. 2020. *Corpus Linguistics: A Guide to the Methodology.* Language Science Press.
- Martin Wynne (ed.). 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books.

All materials other than the two required textbooks will be made available on Canvas.

COURSE DESIGN: MAJOR ASSIGNMENTS AND ASSESSMENTS

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

- In class assignments (20%)
- Homeworks (20%)
- Midterm exam (30%)
- Final project (30%)

For in class assignments, you will need to bring a device with a keyboard. Tentative due date for the midterm is October 25 and for the final is December 18.

COURSE SCHEDULE

Due to scheduling conflicts, class will not be held at the usual time on **August 29, September 12, September 26, October 10, and November 21**. Check Canvas for alternate activities on those days. UPDATE: More dates have been added: **September 19, October 3, and November 14**.

Proposed course outline:

Intro
Frequency
Keywords and collocations
Word meanings
Corpus construction
Annotation
Lexical diversity
Lexicogrammar
Sociolinguistic variation
LLMs

GRADING POLICIES

- In class assignments will be graded on a scale from 0 (didn't do it) to 5 (perfect). The lowest two assignment grades will be dropped.
- Late homework will be accepted with a grade penalty
- Work in groups on assignments and homework, but not exams or projects

AI SYLLABUS STATEMENT

Students should not use generative AI applications in this course except as approved by the instructor. Any use of generative AI outside of instructor-approved guidelines constitutes misuse. Misuse of generative AI is a violation of the course policy on

academic honesty and will be reported to the Center for Student Rights and Responsibilities.

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

Last revised: 29 August 2024