

Ling 571 Computational Corpus Linguistics

San Diego State University

Schedule # 22199

Fall 2020

TuTh 9:30am–10:45am

On-line: SDSU.zoom.us/j/98341874085

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: Tue & Wed 1:00pm–2:00pm, or by appointment

Zoom office: SDSU.zoom.us/j/91271627039

Google Chat room: chat.google.com/room/AAA9IJKDUo

Communications

We’ll be using several different modes of communication this semester. My goal is to get as little email as possible, so I’d prefer to use other communication channels as appropriate. Official class communications, like announcements, assignments, videos, etc., will be posted on **Canvas**. To access your course log in at canvas.sdsu.edu and sign in using your SDSUid. You can also contact me or other students using the Canvas ‘Conversations’ function: click on the Inbox icon and go from there.

Note: You are responsible for adjusting your notification settings in such a way that you receive all announcements regarding this class. All Canvas email notifications will be delivered to your SDSU email address. You can add additional email addresses and sign up for text/mobile app notifications via the settings in your Canvas Profile, and then adjust your notifications in the Notifications Tab. If you have technical issues with Canvas, please contact the SDSU Canvas 24/7 support line at 619.483.0632.

Synchronous communications, like lectures and office hours, will be run using **Zoom** for video and in-class chat. All Zoom class sessions (but not office hours) will be recorded and posted to Canvas afterwards, so even if you aren't able to attend you'll still be able to see what you missed. For details about Zoom meetings, see Canvas.

For informal conversations, we'll use **Google Chat**. This is where you can ask questions about readings, set up study groups, chat with other students, and so on. I'll be available on Google Chat, but nothing that happens there will be graded. I will send out invitations for the Google Chat room to all registered students the first week of class (let me know if you didn't get one).

Finally, if you need to you can always send me **email** at rmalouf@sdsu.edu. Having all classes fully online is going to be a new experience for a lot of us, and my hope is that having multiple channels will make things easier and help support a sense of community. It's not supposed to make things harder, and there's **no wrong way** to get in touch with me. If you can't figure out which mode is best, just send me an email: that will always work, and I won't hold it against you!

No matter what channels you use, please be respectful in your interactions with your fellow students (see the 'Classroom conduct' and 'Prohibited harassment' sections below) and let someone know about any problems that come up (see 'Student affairs').

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.
- Martin Weisser. 2016. *Practical Corpus Linguistics: An Introduction To Corpus-Based Language Analysis*. Wiley.

These should be available from the campus bookstore (or the usual other places) and you should do you best to get them. Weisser is also available to check out from the SDSU library in digital form from the library.

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

Due to the ongoing pandemic, this class will be entirely online. Lectures, assignments, exams, and readings will either be posted to Canvas or held via Zoom. Any group Zoom meetings will be recorded and posted to Canvas: synchronous attendance is optional (though highly recommended).

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	Stefanowitsch, ch. 2&3; Wynne, ch. 1
2	Word meanings	Hanks, ch. 1; Weisser, ch. 5&8
3	Frequency	Brezina, ch. 1&2; Weisser, ch. 9
4	Keywords	Brezina, ch. 3
5	Collocations	Stefanowitsch, ch. 7; Weisser, ch. 10
6	Annotation	Weisser, ch. 7; Wynne, ch. 2
7	Python	Bird, ch. 1&2; Downey
8	Corpus representation	Bird, ch. 3; Weisser, ch. 4&11; Wynne, ch. 3&4
9	Lexical dispersion	
10	Lexical diversity	
11	Lexicogrammar	Brezina, ch. 4; Stefanowitsch, ch. 8
12	Social media	Bird, ch. 5&11

University Policies

Accommodations

If you are a student with a disability and are in need of accommodations for this class, please contact Student Ability Success Center at 619-594-6473 as soon as possible. Please know accom-

modations are not retroactive, and I cannot provide accommodations based upon disability until I have received an accommodation letter from Student Ability Success Center.

Student privacy and intellectual property

The Family Educational Rights and Privacy Act (FERPA) mandates the protection of student information, including contact information, grades, and graded assignments. I will use Canvas to communicate with you, and I will not post grades or leave graded assignments in public places. Students will be notified at the time of an assignment if copies of student work will be retained beyond the end of the semester or used as examples for future students or the wider public. Students maintain intellectual property rights to work products they create as part of this course unless they are formally notified otherwise.

Religious observances

According to the University Policy File, students should notify the instructors of affected courses of planned absences for religious observances by the end of the second week of classes.

Deferred Action for Childhood Arrivals (DACA)

All students with questions or concerns regarding DACA are encouraged to see the College of Health and Human Services Assistant Dean for Student Affairs, Jessica Robinson (jmrobinson@sdsu.edu).

Academic honesty

The University adheres to a strict policy prohibiting cheating and plagiarism. Examples of academic dishonesty include but are not limited to:

- copying, in part or in whole, from another's test or other examination;
- obtaining copies of a test, an examination, or other course material without the permission of the instructor;
- collaborating with another or others in work to be presented without the permission of the instructor;
- falsifying records, laboratory work, or other course data;
- submitting work previously presented in another course, if contrary to the rules of the course;

- altering or interfering with grading procedures;
- assisting another student in any of the above;
- using sources verbatim or paraphrasing without giving proper attribution (this can include phrases, sentences, paragraphs and/or pages of work);
- copying and pasting work from an online or offline source directly and calling it your own;
- using information you find from an online or offline source without giving the author credit;
- replacing words or phrases from another source and inserting your own words or phrases.

The California State University system requires instructors to report all instances of academic misconduct to the Center for Student Rights and Responsibilities. Academic dishonesty will result in disciplinary review by the University and may lead to probation, suspension, or expulsion. Instructors may also, at their discretion, penalize student grades on any assignment or assessment discovered to have been produced in an academically dishonest manner.

Resources for students

A complete list of all academic support services—including the Writing Center and Math Learning Center—is available on the Student Affairs' Academic Success website. Counseling and Psychological Services (619-594-5220) offers confidential counseling services by licensed therapists; you can Live Chat with a counselor at go.sdsu.edu/student_affairs/cps/therapist-consultation.aspx between 4:00pm and 10:00pm, or call San Diego Access and Crisis 24-hour Hotline at (888) 724-7240.

Sexual violence / Title IX mandated reporting

As an instructor, one of my responsibilities is to help create a safe learning environment on our campus. I am a mandated reporter in my role as an SDSU employee. It is my goal that you feel able to share information related to your life experiences in classroom discussions, in your written work, and in our one-on-one meetings. I will seek to keep the information you share private to the greatest extent possible. However, I am required to share information regarding sexual violence on SDSU's campus with the Title IX coordinator, Jessica Rentto 619-594-6017. She (or her designee) will contact you to let you know about accommodations and support services at SDSU and possibilities for holding accountable the person who harmed you. Know that you will not be forced to share information you do not wish to disclose and your level of involvement will be your choice. If you do not want the Title IX Officer notified, instead of disclosing this

information to your instructor, you can speak confidentially with the following people on campus and in the community. They can connect you with support services and discuss options for pursuing a University or criminal investigation. Sexual Violence Victim Advocate 619-594-0210 or Counseling and Psychological Services 619-594-5220, psycserv@sdsu.edu. For more information regarding your university rights and options as a survivor of sexual misconduct or sexual violence, please visit titleix.sdsu.edu or sdsutalks.sdsu.edu.

Classroom conduct standards

SDSU students are expected to abide by the terms of the Student Conduct Code in classrooms and other instructional settings. Prohibited conduct includes:

- Willful, material and substantial disruption or obstruction of a University-related activity, or any on-campus activity.
- Participating in an activity that substantially and materially disrupts the normal operations of the University or infringes on the rights of members of the University community.
- Unauthorized recording, dissemination, or publication (including on websites or social media) of lectures or other course materials. This includes posting of exam problems or questions to on-line platforms.
- Conduct that threatens or endangers the health or safety of any person within or related to the University community, including physical abuse, threats, intimidation, or harassment; or sexual misconduct.

Violation of these standards will result in referral to appropriate campus authorities.

Medical-related absences

Students are instructed to contact their professor/instructor/coach in the event they need to miss class, etc. due to an illness, injury or emergency. All decisions about the impact of an absence, as well as any arrangements for making up work, rest with the instructors. Student Health Services (SHS) does not provide medical excuses for short-term absences due to illness or injury. When a medical-related absence persists beyond five days, SHS will work with students to provide appropriate documentation. When a student is hospitalized or has a serious, ongoing illness or injury, SHS will, at the student's request and with the student's consent, communicate with the student's instructors via the Vice President for Student Affairs and may communicate with the student's Assistant Dean and/or the Student Ability Success Center.

SDSU Economic Crisis Response Team

If you or a friend are experiencing food or housing insecurity, or any unforeseen financial crisis, visit sdsu.edu/ecrt, email ecrt@sdsu.edu, or walk-in to Well-being & Health Promotion on the 3rd floor of Calpulli Center.

Concerns regarding classroom activity, grades, or other student affairs matters

Though students have the option of contacting the San Diego State University Ombudsman or the Assistant Dean for Student Affairs at any time regarding classroom activity, grades, or other student affairs matters, students are encouraged to meet with their professor first to discuss the situation. If the issue is not resolved at this level, the student should contact their undergraduate or graduate advisor. If the problem is not resolved at this level, contact should be made to their department director and finally their Assistant Dean for Student Affairs.

Prohibited harassment

The CSU has affirmed its commitment to “protecting access, affordability, intellectual freedom, inclusivity, and diversity for all students ... including supporting DACA students.” Discrimination, harassment, or retaliation against students, faculty, and staff on the basis of race, religion, gender, sexuality, disability, nationality, immigration status and other categories of identity is prohibited. If you have concerns about your status at the university, please visit studentaffairs.sdsu.edu/EOP for information or contact the Dean of Students or the Assistant Dean for Student Affairs in your College.