LING 5381 (Spring 2013)

Corpus Linguistics

Classroom: 014 Trimble Hall, Monday 4:00-6:50 p.m.
Professor: Laurel Smith Stvan
Office: 132 Hammond Hall
Office Hours: 2:00-3:00 Monday and Wednesday, and by appointment
Phone: (817) 272-94xx
Email: stvan@uta.edu (preferred method of contact!)

Course Description

This course will explore some of the ways that computer science and linguistics can inform each other. We will be concerned in particular with the means by which computers can be used to both obtain the data we examine (a corpus of texts) and to provide the tools we use for analysis (concordance tools). A range of linguistic issues and problems that can benefit from computational approaches will be surveyed. These issues will be illustrated through readings and practical experience with several different software programs as well as sources of online corpora. No programming experience is required.

This course fulfills a requirement for the PhD. in Linguistics, but is open as an elective to any graduate student. There is no prerequisite.

This course is intended to help you in achieving the following four objectives:

• Acquiring the knowledge and vocabulary to discuss (both orally and in writing) current and past approaches to corpus linguistics in particular, and computational linguistics in general.

• Practice in reading corpus linguistics literature in order to gain insight into both the kinds of questions asked in this field and typical ways of researching the answers.

• Practice in encountering and evaluating different software to find out how computers can automate many of the tasks you do that use language as data.

• Learning to construct and evaluate investigations whose goal is to discover fuller ways to describe and manipulate a body of naturally occurring language data.

LING 5381 (Spring 2013)

Corpus Linguistics

Student Learning Outcomes:
Upon successfully completing this course, students should be able to:

- illustrate that you can open a text file in a corpus program and produce concordance lines
- illustrate that you can scan an image of text, run OCR on it and save it as editable text.
- create a frequency list and describe some of the distinctive aspects the list reveals about a language's vocabulary
- identify linguistic benefits of working with a corpus that is annotated with POS tags.
- describe and illustrate some of the factors that are useful to consider in compiling text samples for a corpus
- describe and illustrate how querying a corpus can offer information to linguistic description beyond what is available via intuition about a language
- describe a corpus search that would be useful in classroom lesson for second language learners.

Required Course Materials

There are two required texts for this class:

--The book *Working with Specialized Language: A Practical Guide to Using Corpora* by Bowker and Pearson (Routledge, 2002) is available at the campus bookstore, or through any other bookseller of your choice (ISBN: 0-415-23699-1). (There is also a copy of the textbook on 2-hour reserve in the UTA Library.)

--We will also use a set of required articles that are available online in the class Blackboard folder (https://elearn.uta.edu).

--You will also find it useful to have a USB flash drive and/or familiarity with using the school J-drive to save the work you do in the lab during the semester.

Course Requirements

Your course grade will be determined according to the following grading key:

Attendance, preparation, and participation	5%
Applied exercises (5 X 10%)	50%
Reading and analysis assignment	10%
Vocabulary quiz	5%
Final Paper	30%
	100%

Grading Scale

The grades for each component will be determined as follows:

A-	90-92 %	B-	80-82 %	C-	70-72	D-	60-62%	F	59 or lower
A	93-96 %	B	83-86 %	C	73-76	D	63-66		
A+	97-100 %	B+	87-89	C+	77-79	D+	67-69		

Graded Assignments

In addition to written exercises will be required throughout the term, paper is required of graduate students, as an opportunity for you to produce a carefully crafted, extended piece of writing showing an application of computer analysis to data of your choice. Here you will demonstrate how the techniques we have discussed in class might assist you in analyzing your own material. The final paper should be 12-18 typewritten pages. No final exam will be given.

Course Policies

Class attendance is **required**. You are responsible for the material presented in class lectures and for any handouts passed out in class as well as for any group work done in class; for your own benefit, come to class. But if you must miss a lecture, do the reading and homework, get notes and information from another student, and then make an appointment to talk to me as soon as possible.

Assignments are due at the beginning of class on the day listed in the schedule, and no later. No late assignments will be accepted without PRIOR approval. Even approved late submissions will receive a reduction in points.

LING 5381 (Spring 2013)
Corpus Linguistics

Important Academic and Administrative Policies

Electronic Communication: UT Arlington has adopted MavMail as its official means to communicate with students about important deadlines and events, as well as to transact university-related business regarding financial aid, tuition, grades, graduation, etc. All students are assigned a MavMail account and are responsible for checking the inbox regularly. There is no additional charge to students for using this account, which remains active even after graduation. Information about activating and using MavMail is available at http://www.uta.edu/oit/cs/email/mavmail.php.

Student Feedback Survey: At the end of each term, students enrolled in classes categorized as lecture, seminar, or laboratory shall be directed to complete a Student Feedback Survey (SFS). Instructions on how to access the SFS for this course will be sent directly to each student through MavMail approximately 10 days before the end of the term. Each student's feedback enters the SFS database anonymously and is aggregated with that of other students enrolled in the course. UT Arlington's effort to solicit, gather, tabulate, and publish student feedback is required by state law; students are strongly urged to participate. For more information, visit http://www.uta.edu/sfs.

Final Review Week: A period of five class days prior to the first day of final examinations in the long sessions shall be designated as Final Review Week. The purpose of this week is to allow students sufficient time to prepare for final examinations. During this week, there shall be no scheduled activities such as required field trips or performances; and no instructor shall assign any themes, research problems or exercises of similar scope that have a completion date during or following this week *unless specified in the class syllabus*. During Final Review Week, an instructor shall not give any examinations constituting 10% or more of the final grade, except makeup tests and laboratory examinations. In addition, no instructor shall give any portion of the final examination during Final Review Week. During this week, classes are held as scheduled. In addition, instructors are not required to limit content to topics that have been previously covered; they may introduce new concepts as appropriate.

Americans With Disabilities Act: The University of Texas at Arlington is on record as being committed to both the spirit and letter of all federal equal opportunity legislation, including the *Americans with Disabilities Act (ADA)*. All instructors at UT Arlington are required by law to provide "reasonable accommodations" to students with disabilities, so as not to discriminate on the basis of that disability. Any student requiring an accommodation for this course must provide the instructor with official

documentation in the form of a letter certified by the staff in the Office for Students with Disabilities, University Hall 102. Only those students who have officially documented a need for an accommodation will have their request honored. Information regarding diagnostic criteria and policies for obtaining disability-based academic accommodations can be found at www.uta.edu/disability or by calling the Office for Students with Disabilities at (817) 272-3364.

Academic Integrity: All students enrolled in this course are expected to adhere to the UT Arlington Honor Code:

I pledge, on my honor, to uphold UT Arlington's tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence.

I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.

Instructors may employ the Honor Code as they see fit in their courses, including (but not limited to) having students acknowledge the honor code as part of an examination or requiring students to incorporate the honor code into any work submitted. Per UT System *Regents' Rule* 50101, §2.2, suspected violations of university's standards for academic integrity (including the Honor Code) will be referred to the Office of Student Conduct. Violators will be disciplined in accordance with University policy, which may result in the student's suspension or expulsion from the University.

Please be advised that departmental policy requires instructors to formally file charges with the Office of Student Conduct, following procedures laid out for faculty there (http://www.uta.edu/studentaffairs/conduct/faculty.html), as well as notify the department chair of the filing of the charges.

Student Support Services Available **Student Support Services**: UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals, students may visit the reception desk at University College (Ransom Hall), call the Maverick Resource Hotline at 817-272-6107, send a message to resources@uta.edu, or view the information at www.uta.edu/resources.

LING 5381 (Spring 2013)

Corpus Linguistics

Drop Policy: Students may drop or swap (adding and dropping a class concurrently) classes through self-service in MyMav from the beginning of the registration period through the late registration period. After the late registration period, students must see their academic advisor to drop a class or withdraw. Undeclared students must see an advisor in the University Advising Center. Drops can continue through a point two-thirds of the way through the term or session. It is the student's responsibility to officially withdraw if they do not plan to attend after registering. **Students will not be automatically dropped for non-attendance**. Repayment of certain types of financial aid administered through the University may be required as the result of dropping classes or withdrawing. For more information, contact the Office of Financial Aid and Scholarships (http://wweb.uta.edu/ses/fao). **(Note: Students enrolled in graduate courses may not repeat a class to "replace" a grade).**

A student dropping his/her last (only) course cannot withdraw as above. Rather, s/he must go in person to the UTA Registrar's Office (Davis Hall, First Floor) and complete a request to resign from the university.

Auditors: Department of Linguistics and TESOL faculty, staff, and students currently enrolled in a linguistics/TESOL programs may be able to audit a course (with the permission of the professor). Audited courses cannot be used to satisfy any degree or program requirements/electives, nor will any credit (including retroactive) be granted for audited courses.

Schedule

If there are any changes from the paper copy given out on the first day of class, the most current course schedule of readings and assignments and any additional links to citations and readings that come up in class will be uploaded in Blackboard. You are responsible for checking the site regularly.

Ling. 5381 /4330
Corpus Linguistics
Spring 2013

Proposed Schedule: (Last Updated: **Jan 14, 2013**)

Background: Corpus Uses and Famous Corpora

1. Mon. Jan. 14 — Introduction to the class; Introduction to the lab
What is computational linguistics? What is corpus linguistics?

The web as corpus; B&P Ch. 1 "Introducing Corpora and Corpus Analysis Tools" (9-24).

2. Mon. Jan. 21 — **Labor Day—No classes**

Fillmore, Charles J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics"; "Henry Kucera (1991)
Boyd& Crawford (2012) Critical Question sof Big Data
Exercise 1 Due (Web Queries)

3. Mon. Jan. 28 — Kennedy (1998). "Design and Development of Corpora."
Mukherjee, Joybrato. (2004). "Review Article: State of the Art in CL"
B&P (2002). Ch. 2 "Introducing LSP." (25-41).
Clear (1992). "Corpus Sampling." (21-31).
Demo on using the scanner

Compiling a Corpus; Inputting and Annotating Texts

4. Mon. Feb. 4 — Crystal (2010). "Languages on the Web." (216-223).
Bowker (2002). "Capturing Data in Electronic Form."
Demo on doing OCR
Douglas Hofstadter (1985) "Ch. 13: Metafont, Metamathematics, and Metaphysics."

5. Mon. Feb. 11 — B&P (2002). Ch. 5 "Markup and Annotation." (75-91).
Leech (1997). "Grammatical Tagging." (19-33)
Leech (1997) Appendix III (The C7 and C5 Tagsets) (256-260).
B&P (2002). Ch. 3. "Designing a Special Purpose Corpus."
B&P (2002). Ch. 4. "Compiling a Special Purpose Corpus."
Exercise 2 Due (Scanning and Using OCR)

Ling. 5381 /4330
Corpus Linguistics
Spring 2013

Tools to Use on a Corpus: Concordance Software and Indexing Software

6. Mon. Feb. 25 — B&P (2002) Ch. 7 "Introduction to Basic Corpus Processing Tools"
AncConc demo
Exercise 3 Due (Tagging Using Different Tagsets)

7. Mon. March 4 — Working with Antconc
Unicode demo

March 11-15 ---Spring Break ---

Applying Corpus Tools: Terminology, Translation, Language Teaching

8. Mon. March 18 — Baker (2006) Ch. 3 "Frequency and Dispersion"
B&P (2002) Ch. 6 "Bilingual and Multilingual corpora: Preprocessing, Alignment and Exploitation"
Exercise 4 Due (Using AntConc with Untagged Texts)

9. Mon. March 25 — B&P (2002) Ch. 8 "Building Useful Glossaries"
B&P (2002) Ch. 9 "Term Extraction"

10. Mon. April 1 — McEnery, Tony, Jean-Marc Langé, Michael Oakes, and Jean Véronis. (1997). (CA Ch. 15) "The exploitation of multilingual annotated corpora for term extraction"
Stvan (2005). "Inferring New Vocabulary Using Online Texts"
B&P (2002) Ch. 10 "Using LSP Corpora as a Writing Guide"
Exercise 5 Due (Concordancing with Tagged Texts)

11. Mon. April 8 — B&P (2002) Ch. 11 "Using LSP Corpora as a Translation Resource"
B&P (2002) Ch. 12 "Other Applications and Future Directions"

12.	Mon. April 15	Granger (1998). "The Computer Learner Corpus: A Testbed for Electronic EFL Tools." Pp. 175-188. **Reading and Analysis Exercise 1 Due** Biber, Conrad, and Reppen (1998). "Language Acquisition and Development." Pp. 172-201.
13.	Mon. April 22	Biber, Conrad, and Reppen (1998). "Historical and Stylistic Investigations." Pp. 203-229 + Pp. 252-253. Biber (1992). "Using Computer-based Text Corpora to Analyze the Referential Strategies of Spoken and Written Texts." Pp. 213-255.
14.	Mon. April 29	Lindquist (2000) "Livelier or More Lively? Syntactic and Contextual Factors Influencing the Comparison of Disyllabic Adjectives." Stvan (2006). "Diachronic Change in The Discourse Markers Why and Say in American English." **Undergrad Reading and Analysis Exercise 2 Du** Wind-up and evaluations **Vocabulary Quiz**

EXAM WEEK

Wed. May 8	2:00 -4:30 pm. (note different meeting time and day)
	In-class presentations on final projects
	Undergrad Reading and Analysis Exercise 3 Due
	Grad students' final written paper Due

Additional Dates to Note

Wed. Jan. 30 Census date (last day to add a class)
Fri. March 29 Last day to drop a course
Wed. May 15 Grades available: **http://www.uta.edu/mymav**