LING 5381 (Fall 2010)

Corpus Linguistics

Classroom: 014 Trimble Hall, Monday and Wednesday, 1:00-2:30 p.m.
Professor: Laurel Smith Stvan
Office: 410 Hammond Hall
Office Hours: 2:00-3:50 Monday and Wednesday, and by appointment
Phone: (817) 272-3133
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 (Fall 2010)

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 will be available online shortly after class starts.

--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	10%
Exercises (5 X 10%)	50%
Vocabulary quiz	10%
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 exercises throughout the course, a final paper is required, 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 12-14 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 not later. No late assignments will be accepted without PRIOR approval. Even approved late submissions will receive a reduction in points.

Important Academic and Administrative Policies

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

LING 5381 (Fall 2010)

Corpus Linguistics

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 course 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.

Americans With Disabilities Act: The University of Texas at Arlington is on record as being committed to both the spirit and letter of federal equal opportunity legislation (Public Law 93112, The Rehabilitation Act of 1973 as amended). With the passage of new federal legislation entitled the "Americans With Disabilities Act" (ADA), pursuant to section 504 of The Rehabilitation Act, there is renewed focus on providing this population with the same opportunities enjoyed by all citizens.

All members of the UTA faculty are required by law to provide "reasonable accommodation" to students with disabilities, so as not to discriminate on the basis of that disability. As a student, your responsibility rests with informing the instructor at the beginning of the semester (you must inform me in writing (e-mail is fine) no later than Tuesday, Sept. 7, 2010) and in providing authorized documentation through designated administrative channels; for more information, contact UTA's Office of Students with Disabilities (located in the Lower Level of University Center).

According to Department of Linguistics and TESOL policy, "unofficial" or "informal" requests for accommodations (i.e., those not recorded by the Office of Students with Disabilities) cannot be honored.

Academic Dishonesty: At The University of Texas at Arlington, academic dishonesty is a completely unacceptable mode of conduct and will not be tolerated in any form. Students involved in academic dishonesty will be disciplined in accordance with University regulations and procedures. Discipline may include suspension or expulsion from UTA.

According the UT System Regents' Rules and Regulations, "Scholastic dishonesty includes but is not limited to cheating, plagiarism, collusion, the submission for credit of any work or materials that are attributable in whole or in part to another person, taking an examination for another person, any act designed to give unfair advantage to a student or the attempt to commit such acts" (Part One, Chapter VI, Section 3, Subsection 3.2, Subdivision 3.22).

While the Department of Linguistics and TESOL hopes to foster a sense of community in which students can enhance their educational experience by conferring with each other about the lectures, readings, and assignments, all work submitted must be the product of each student's own effort. Students are expected to know and honor the standards of academic integrity followed by American universities; ignorance of these standards is not an excuse for committing an act of academic dishonesty (including plagiarism). If you have questions, please speak with your instructor, your academic advisor, or the department chair.

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: The University of Texas at 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. These resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals to resources for any reason, students may contact the Maverick Resource Hotline at 817-272-6107 or visit www.uta.edu/resources for more information.

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. Contact the Financial Aid Office for more information. (Note: Students enrolled in graduate courses may not "replace" a grade by repeating a class).

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.

LING 5381 (Fall 2010)

Corpus Linguistics

Auditors: The Department of Linguistics and TESOL has a "no audit" policy, with one exception. With instructor permission, Department of Linguistics and TESOL faculty, staff, and students enrolled in a linguistics/TESOL degree program 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 linked to this page: http://ling.uta.edu/~laurel/5381-Folder/5381description.php).

Ling. 5381
Corpus Linguistics
Fall 2010

Proposed Schedule: (Last Updated: **Aug. 30, 2010**)

Background: Corpus Uses and Famous Corpora

1. Mon. Aug. 28 — Introduction to the class; Introduction to the lab
What is computational linguistics? What is corpus linguistics?

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

2. Mon. Sept. 6 — **Labor Day—No classes**

Wed. Sept. 8 — Fillmore, Charles J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics.";
"Henry Kucera (1991)
Exercise 1 Due (Web Queries)

3. Mon. Sept. 13 — Kennedy (1998). "Design and Development of Corpora."
Mukherjee, Joybrato. (2004). "Review Article: State of the Art in CL"

Compiling a Corpus; Inputting and Annotating Texts

Wed. Sept. 15 — B&P (2002). Ch. 2 "Introducing LSP." (25-41).
Clear (1992). "Corpus Sampling." (21-31).
Demo on using the scanner

4. Mon. Sept. 20 — Crystal (2010). "Languages on the Web." (216-223).
Bowker (2002). "Capturing Data in Electronic Form."
Demo on doing OCR

Wed. Sept. 22 — Douglas Hofstadter (1985) "Ch. 13: Metafont, Metamathematics, and Metaphysics."

5. Mon. Sept. 27 — 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).
Exercise 2 Due (Scanning and Using OCR)

Ling. 5381
Corpus Linguistics
Fall 2010

Wed. Sept. 29 — B&P (2002). Ch. 3. "Designing a Special Purpose Corpus."
B&P (2002). Ch. 4. "Compiling a Special Purpose Corpus."

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

6. Mon. Oct. 4 — B&P (2002) Ch. 7 "Introduction to Basic Corpus Processing Tools"
Exercise 3 Due (Tagging Using Different Tagsets)

Wed. Oct. 6 — AncConc demo

7. Mon. Oct. 11 — Working with Antconc

Wed. Oct. 13 — Unicode demo by Josh Jensen
WordSmith Tools 4 demo by Laurel Stvan

8. Mon. Oct. 18 — Working with WordSmith Tools 4
See manual selections (available in the lab)
Baker (2006) Ch. 3 "Frequency and Dispersion"

Applying Corpus Tools: Terminology, Translation, Language Teaching

Wed. Oct. 20 — B&P (2002) Ch. 6 "Bilingual and Multilingual corpora: Preprocessing, Alignment and Exploitation"

9. Mon. Oct. 25 — B&P (2002) Ch. 8 "Building Useful Glossaries"
Exercise 4 Due (Using AntConc or WordSmith with Untagged Texts)

Wed. Oct. 27 — B&P (2002) Ch. 9 "Term Extraction"

10.	Mon. No. 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"
	Wed. Nov. 3	B&P (2002) Ch. 10 "Using LSP Corpora as a Writing Guide"
11.	Mon. Nov. 8	B&P (2002) Ch. 11 "Using LSP Corpora as a Translation Resource"
	Wed. Nov. 10	B&P (2002) Ch. 12 "Other Applications and Future Directions"
12.	Mon. Nov. 15	Granger (1998). "The Computer Learner Corpus: A Testbed for Electronic EFL Tools." Pp. 175-188. **Exercise 5 Due (Concordancing with Tagged Texts)**
	Wed. Nov. 17	Biber, Conrad, and Reppen (1998). "Language Acquisition and Development." Pp. 172-201.
13.	Mon. Nov. 22	Biber, Conrad, and Reppen (1998). "Historical and Stylistic Investigations." Pp. 203-229 + Pp. 252-253.
	Wed. Nov. 24	Biber (1992). "Using Computer-based Text Corpora to Analyze the Referential Strategies of Spoken and Written Texts." Pp. 213-255.
14.	Mon. Nov. 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."
	Wed. Dec. 1	Wind-up and evaluations **Vocabulary Quiz**

Ling. 5381
Corpus Linguistics
Fall 2010

15.	Mon. Dec. 6	Right to left concordancing demo?
	Wed. Dec. 8	Work day

EXAM WEEK

Monday Dec. 13 11:00a.m. - 1:30 p.m.
(note longer meeting time)
In-class presentations on final projects
Everyone's final written paper is due

Additional Dates to Note

Mon. Sept. 13 Census date
Fri. Nov. 5 Last day to drop a course
Wed. Dec. 22 Grades available: **http://www.uta.edu/mymav**