This course is a practical introduction to corpus linguistics. The learning objectives are:

• How to manage data analysis and manipulation packages in a Unix environment from the command line.

• Practical experience with statistical methods for analyzing naturalistic data.

• Practical experience with tools and resources for automatically enriching corpus annotations.

• Basic knowledge of data privacy issues and differential privacy.

This course will involve two short projects:

**Short Project I**: Analysis of existing annotations. Select a corpus and answer a linguistic question by performing a statistical analysis on the annotations of that corpus.

**Short Project II**: Extending corpus annotations. Use NLP tools to enrich the annotations of a corpus, evaluate the quality of the enriched annotations, and use the annotations to answer a linguistic question. For example, you might:

• Force-align and parse a speech corpus that has only utterance-level transcript and see if there is a relationship between vowel durations and syntactic environment.

• Parse text from Project Gutenberg with multiple syntactic parsers, and use the trees to measure changes in the selectional preference of verbs over time.

Both projects can address the same question and use the same corpus – in fact, this is encouraged. There will also be six homeworks distributed on Fridays and due the following Friday.

**Grading breakdown**

- Participation: 8%
- Homeworks: 42% (7 at 6 each)
- Short Project I: 20%
- Short Project II: 20%
- Short Project Presentation: 10%

**Week 1**: Introduction to corpus linguistics and Unix Bootcamp

Aug 31: Intro to the course.
Sep 02: Intro to Unix environments.
Sep 04 Lab: Working with the Unix command line and package management in Unix environments.
Homework 1: vim or emacs.

**Week 2**: Introduction to corpus linguistics and Unix Bootcamp (cont’d)
Sep 07: Labor day observed – no lecture.
Sep 09: Available corpora and language resources.
Sep 11 Lab: bash and perl.
Homework 2: Simple bash and perl scripting – scalars, arrays, and hashes.

**Week 3:** Annotation.
   Sep 14: Annotation.
   Sep 16: Searching corpora.
   Sep 18 Lab: Searching corpora using Unix tools and perl.

**Week 4:** Annotation (cont’d)
   Sep 21: Workshopping project proposals (bring draft proposals for both project I and II).
   Sep 23: Probability theory.
   Sep 25 Lab: The NXT query language (with Switchboard in NXT)

**Week 5:** Probability theory
   Sep 28: Intro to probability theory and n-gram language models.
   Sep 30: Graphical models.
   Oct 02 Lab: R scripting.

**Short Project I and II Proposals due Oct 5th at 22:00**

**Week 6:** Statistical Analysis.
   Oct 05: Regression analysis.
   • (Jaeger, 2008)
   Oct 07: Bayesian analysis
   • (Kwiatkowski et al., 2012; Bouchard-Côté et al., 2012)
   Oct 09 Lab: Regression analysis lab with lme4 and stan.

**Week 7:** Enriching corpus annotations: NLP tools
   Oct 12: Overview of NLP tools.
   Oct 14: Evaluating silver-standard annotations
   • (Passoneau and Carpenter, 2014)
   Oct 16 Lab: Parsing lab.

**Week 8:** Information-theoretic perspectives
   Oct 21: Basics of coding theory.
   Oct 23 Lab: n-grams, smoothing, and collocations

**Week 9:** Enriching corpus annotations: Language resources
   Oct 28: Open Multilingual WordNet

**Week 10:** Interpreting regression and language resources.
Nov 02: More on regression and interpretation.

**Short Project I due Nov 3rd at 17:00**

Nov 04: Distributed representations of meaning.
- (Mikolov et al., 2013; Pennington et al., 2014)

Nov 06 Lab: distributed representations.

**Week 11:** Enriching corpus annotations: Speech

Nov 09: More on VerbNet
Nov 11: Speech recognition systems and forced alignment.
- SailAlign (Katsamanis et al., 2011).

Nov 13 Lab: Forced alignment lab with Prosodylab-Aligner.
Homework 4 subroutines in perl

**Week 12:** Summarizing and visualizing data.

Nov 16: Manipulating data in R and basic plots.
Nov 18: Beyond barcharts and scatterplots.

Nov 20 Lab: Visualizing data using R with ggplot2.
Homework 5 functions in R

**Week 13:** Handling web-scale data.


**Short Project II due Nov 24th at 17:00**

Nov 25: Fall recess – no lecture.

Nov 27 Lab: Fall recess – no lab.

**Week 14:** Data Privacy

Nov 30: Overview of data privacy and Netflix challenge paper.
- (Narayanan and Shmatikov, 2008), (Dwork and Roth, 2014, Ch. 1)

Dec 02: Light introduction to differential privacy.
- (Dwork and Roth, 2014, Ch. 3)

Dec 04 Lab: Topic modeling

**Week 15:** Short Project Presentations

Homework 6 distributed Plotting in R

Dec 07: 4 project presentations (your choice of Project I or Project II).
- ebelling
- abeerali
- alimoham
- mengyang

Dec 09: 5 project presentations (your choice of Project I or Project II).
- rchepbur
- hayk
- rjhatche
- juliaren
- jcarota
Dec 11: 5 project presentations (your choice of Project I or Project II).
- jaschech
- cslees
- tsdiaz
- shong23
- yli57

Week 16: Finals week

Dec 14: Homework 6 due Plotting in R

References


