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

- Facility with manipulating natural language data in a Unix environment from the command line.
- Practical experience with statistical methods for analyzing naturalistic language data.
- Practical experience with tools and resources for automatically enriching corpus annotations.

This course will involve several in-class labs and two short projects:

**Labs:** Every Friday, we will have an in-class lab in small groups together. Each lab provides several short questions. If you complete the questions by the end of the Friday meeting, you may turn them in at that time; otherwise, turn them in at the beginning of the following Monday class.

**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 may address the same underlying question and use the same corpus – in fact, this is encouraged.

**Grading breakdown**

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labs</td>
<td>60%</td>
</tr>
<tr>
<td>Short Project I</td>
<td>15%</td>
</tr>
<tr>
<td>Short Project II</td>
<td>15%</td>
</tr>
<tr>
<td>Short Project Presentation</td>
<td>10%</td>
</tr>
</tbody>
</table>

**Polices**

1. **Incompletes.** These generally create more problems than they solve. If you are considering requesting an incomplete, please familiarize yourself first with the Graduate School’s official policy at:

   [http://grad.buffalo.edu/Academics/Policies-Procedures/Grading-Procedures.html#incomplete](http://grad.buffalo.edu/Academics/Policies-Procedures/Grading-Procedures.html#incomplete)

2. **Academic Integrity:** All students should be sure that they understand the Graduate School’s Academic Integrity policy before completing any assignments. If you have any questions or concerns about the policy, please discuss them with me. You will find the policy at:

   [http://grad.buffalo.edu/Academics/Policies-Procedures/Academic-Integrity.html](http://grad.buffalo.edu/Academics/Policies-Procedures/Academic-Integrity.html)
3. **Accessibility Resources:** If you require classroom or testing accommodations due to a disability, please contact Accessibility Resources, located at 25 Capen Hall. AR can be reached by phone at (716) 645-2608 or by email at stu-accessibility@buffalo.edu. Please inform me as soon as possible about your needs so that we can coordinate your accommodations.

---

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

- Aug 29: Intro to the course.
- Aug 31: Intro to Unix environments.
- Sep 02 Lab: Working with ssh and Unix command line environments.

**Week 2:** Introduction to corpus linguistics and Unix Bootcamp (cont’d)

- Sep 05: Labor day observed – no lecture.
- Sep 07: Available corpora and language resources.
- Sep 09 Lab: bash and perl.

**Week 3:** Annotation.

- Sep 12: Annotation.
- Sep 14: Searching corpora.
- Sep 16 Lab: Searching corpora using Unix tools and perl.

**Week 4:** Annotation (cont’d)

- Sep 19: Workshopping project proposals (bring draft proposals for both project I and II).
- Sep 21: Probability theory.
- Sep 23 Lab: The NXT query language (with Switchboard in NXT)

**Week 5:** Probability theory

- Sep 26: Intro to probability theory and n-gram language models.
- Sep 28: Graphical models.
- Sep 30 Lab: Intro to R.

**Short Project I and II Proposals due Sep 29th in class**

**Week 6:** Statistical Analysis.

- Oct 03: Regression analysis.
  - (Jaeger, 2008)
- Oct 05: Interpreting regressions.
- Oct 07 Lab: Regression analysis lab with lme4.

**Week 7:** Enriching corpus annotations: NLP tools.

- Oct 10: Overview of NLP tools.
- Oct 12: Evaluating silver-standard annotations
  - (Passoneau and Carpenter, 2014)
- Oct 14 Lab: Parsing lab.

**Week 8:** Information-theoretic perspectives.
Oct 17: Information theory in corpus linguistics.
Oct 19: Basics of coding theory.
   • More information: (Mackay, 2003, Ch. 1, 2, 4-6, 8-11).
Oct 21 Lab: n-grams, smoothing, and collocations

Week 9: Enriching corpus annotations: Language resources.
   Oct 26: Open Multilingual WordNet

Week 10: Distributed representations.
      • More info: Blei et al. (2003)
      **Short Project I due Oct 31st in class**
   Nov 02: Distributed representations of meaning.
      • More info: Mikolov et al. (2013), Pennington et al. (2014)
   Nov 04 Lab: Distributed representations lab.

Week 11: Enriching corpus annotations: Speech
   Nov 07: More on VerbNet
   Nov 09: Speech recognition systems and forced alignment.
   Nov 11 Lab: Forced alignment lab with Prosodylab-Aligner.

Week 12: Summarizing and visualizing data.
   Nov 14: Manipulating data in R and basic plots.
   Nov 16: Beyond barcharts and scatterplots.
   Nov 18 Lab: Visualizing data using R with ggplot2.

Week 13: Very large datasets.
   Nov 21: Handling very large datasets with Bloom filters and Count-Min Sketch
      • (Cormode and Muthukrishnan, 2005).
      **Short Project II due Nov 21st in class**
   Nov 23: Fall recess – no lecture.
   Nov 25 Lab: Fall recess – no lab.

Week 14: Data Privacy and model-based corpus analysis.
   Nov 28: Overview of data privacy and Netflix challenge paper.
      • (Narayanan and Shmatikov, 2008), (Dwork and Roth, 2014, Ch. 1)
   Nov 30: Model-based corpus analysis I
      • Kwiatkowski et al. (2012)
   Dec 02: Model-based corpus analysis II
      • Bouchard-Côté et al. (2012)

Week 15: Short Project Presentations
   Dec 05: 3 project presentations (your choice of Project I or Project II).
   Dec 07: 3 project presentations (your choice of Project I or Project II).
   Dec 09: 3 project presentations (your choice of Project I or Project II).
      ** No Final **
References


