Lign 245: Computational Corpus Linguistics

Prof. Andrew Kehler
UCSD Department of Linguistics
kehler@ling.ucsd.edu
(858) 534-1159

Fall, 2009
Tuesdays and Thursdays, 12-2
Office Hours: Tuesdays 2-3, or by appt.

Overview

This course will provide an introduction to computational corpus tools for performing empirically-grounded linguistic investigations. You should leave this class knowing: (i) about available corpora to support your research, both raw and annotated, (ii) how to use existing software packages for manipulating these corpora (e.g., TGREP), (iii) how to use the UNIX operating system and write simple programs in the PYTHON scripting language, and (iv) how corpora have been used in the literature to influence linguistic research.

Prerequisites

You do not need to know how to program, but you do need to be willing to learn the basics.

Readings

Required: Bird, Steven; Klein, Ewan; and Loper, Edward. Natural Language Processing with Python. O'Reilly, 2009. Available at Amazon.com ($38.77).

Recommended: Hetland, Magnus Lie. Beginning Python: From Novice to Professional. Apress, 2005. Available at Amazon.com. Or borrow it from someone who took the class in 2006.

Recommended: Pilgrim, Mark. Dive into Python. Download from http://diveintopython.org/.

Requirements

There will be regular homework assignments, and a final research project. For the research project, you need to have your own research idea that would benefit by being informed by naturally-occurring data. This might be your a topic for course paper, comps paper, qualifying paper, or dissertation, or just something you've read in the literature that you doubt would hold up in an examination of actual language use. The deliverable for this project will be the results of your empirical investigation and any code written.

Highly Provisional Schedule

9/25: Overview of course. Introduction to UNIX. Using the EMACS editor.

9/30: Regular Expressions. Searching corpora with UNIX GREP.

10/1: Parsed Corpora I: The Penn Treebank. The TGREP search program.

10/3: Doing it Yourself: Introduction to the PYTHON scripting language. Introduction to NLTK.

10/8: Python: Lists, Tuples, Variables, etc.

10/13: Working with NLTK, corpora.

10/15: Python: Working with Strings

10/20: Python: Structuring Programs with Conditionals, Loops, etc.

10/22: Python: Working through some examples

10/27: Python: Dictionaries

10/29: Python: File Input/Output

11/3: Python: Functions

11/5: Python: Working through more examples

11/10: Using naturally-occurring data in your analyses: examples from the literature

11/12: Using naturally-occurring data in your analyses: examples from the literature

11/17: Other Corpora: What's out there, and what's here. Annotation schemes.

11/19: Survey of useful CL tools: taggers, parsers, and machine learning programs

11/24: TBD.

12/1: TBD.

12/3: TBD.