Syllabus: LING-479 Language Corpora and Software

Damir Cavar
English Department
Eastern Michigan University

Jan. 2013

1 Details

- **Instructor:** Dr. Damir Cavar
- **Class meets:** MW 11:00 A – 12:15 P, PRAY-H 327
- **Office hours:** MW 10:00–11:00 AM, 12:30–1:30 AM, TH 4:00–6:30 PM in PRAY-H 613D, other hours TUE 9 AM – 1:30 PM and by arrangement in Cooper at ILIT or LINGUIST List, Suite 104
- **Communication:** use email, Google Hangout, Skype, come to the office hours
- **Class email:** 25007.201320@emich.edu

2 Content

This course will introduce corpora as collections of text of different origin, in various formats, with and without specific annotations. We will discuss different types of corpora, their use and purpose, how they were created and annotated, and how they can be used for research, analysis, study of language and language technologies.

We will develop our own corpora and analyze them, and this way hands-on learn about different encoding standards, annotation strategies, and technologies for processing of large quantities of such text. The technologies we will look at will include raw text in different encodings, XML annotated text, various tag sets for part-of-speech annotation, and treebank-annotation. We will make use of software tools for the storage and analysis (indexing, search and retrieval) of these corpora.

Our focus will be manyfold. On the one hand we will study and discuss the linguistic relevance of corpora. This will cover topics in language change and linguistic variation, or language learning and acquisition studies. On the other hand, we will get used to various technologies to create corpora and process them, as it is of relevance for statistical analyses, language technologies, machine translation and other computational linguistic domains.

We will make use of editors to process text, XML files, and annotations like Notepad++ and various XML editors. Further, we will make use of Concordance, Word Cruncher, Philologic, WordSmith, AntConc, Unix commands and database tools for the storage, indexing and analysis of corpora.

Our goals are to:

- understand corpora, their purpose and uses,
- understand and practically experience the creation of corpora
- be able to use corpora as a database to study a wide range of linguistic phenomena using common software tools

3 Equipment recommended

We will be able to use the computer pool and a server for corpus analysis. We recommend that you make use of your own computer for the assignments and projects. You can even bring your own laptop to class, if you wish. The instructor will help you setting up software tools for the course, language data and corpora.

4 Grading

Grading will be based on three parts:

amount	**type**	**%**
7	assignments	30%
2	projects	35%
1	presentation	35%

- **assignments:**
  - Assignments will involve corpus preparation, annotation and analysis tasks, with two weeks preparation time.
- **projects**
  - Corpus preparation task: A collection of text-data has to be digitally prepared and annotated.
  - Corpus analysis: Various corpus analysis tasks will be preformed and documents, including the counting of types and tokens, identification of specific properties, and integration in a common corpus processing tool.
  - The project is documented in written form and presented in class in the final session in 10 to 15 minutes each.
- **presentation:**
  Everybody will prepare a 10 to 15 minutes presentation that will cover some well-known corpus source, processing and analysis system, or annotation standard and tool.

The grades will be related or scaled based on your attendance.

4.1 Attendance

Regular class attendance is expected of all students. Missing more than two sessions unexcused might seriously harm your grade. Reasons for missing class have to be discussed with the instructor.

5 Literature

We will read selected sections and articles from various textbooks and online sources. All articles are listed on the course web-page and the reading is announced on a weekly basis.

6 Schedule

The following topics represent the sequence of sessions and topics (subject to change on the basis of individual preferences and interests):

1. Introduction
2. Overview
3. Text structure and XML-annotation
4. Frequency profiles
5. Concordances
6. Collocations
7. Tools introduction
8. Lexicography
9. Lexicology
10. Treebanks
11. Document classes and their properties
12. Authorship study
13. Multilingual or parallel corpora (for translation studies and models)
14. Presentations