# Syllabus: LING 592 Statistics in Language Technologies

Damir Cavar
English Department
Eastern Michigan University

Jan. 2012

## 1 Details

- **Instructor:** Dr. Damir Cavar
- **Class meets:** TUE in PRAY-H 313 6:30–9:10 PM
- **Office hours:** TUE and THU 1PM-3:30PM

## 2 Content

In this course we will discuss descriptive statistical methods that are useful for any type of linguistic research, documentation or production activity. We will extend the statistical methods to inductive statistical approaches for the analysis of languages and linguistic properties, or computational methods for natural language processing. This will include basic probability theory, and also concepts from Information Theory, like Entropy and derived measures.

Our goals are to understand:

- statistical description of linguistic data
- significance tests, anova and multivariate analysis of language data and experimental results

- probability theory and its application in linguistic analysis, natural language processing and language technologies
- information theoretic concepts like entropy and its application
- statistical clustering methods and multivariate text, word, distribution analysis
- language technologies that make use of such statistical methods

## 3 Equipment recommended

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

- We will make use of the R programming language and environment.

## 4 Grading

There will be regular assignments that will be discussed in class and help you understand the topic, in particular, reading assignments.

Grading will be based on two parts:

| **amount** | **type** | **%** |
|---|---|---|
| 2 | projects | 60% |
| 1 | presentation | 40% |

- **projects**
  - Processing of large scale language data or corpora: analysis of collocations, concurrence patterns, distributional regularities and
  - Preparation of material (for example corpora, data, language models from distribution patterns), and analysis of properties using clustering techniques, multivariate statistical analysis, information theoretic measures etc.

  - The projects are documented in written form and presented in class in the final session in 10 to 15 minutes time slots.

- **presentation:**
  Everybody will prepare a 10 to 15 minutes presentation that will cover some aspects of statistical linguistic properties, information-theoretic measures, distributional properties and how they are used in some concrete application or technology.

### 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. Our main textbooks will be:

- Shravan Vasishth and Michael Broe (2010) *The Foundations of Statistics: A Simulation-based Approach*. Springer. ISBN: 978-3-642-16312-8
- Chris Manning and Hinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA. ISBN: 978-0262133609
- Daniel Jurafsky and James H. Martin (2008) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Second Edition. Pearson Prentice Hall. ISBN: 978-0131873216

For an introduction to statistics the following books could be used:

- Lloyd R. Jaisingh (2006) *Statistics for the utterly confused*. McGraw-Hill. ISBN: 978-0071461931
- Deborah Rumsey (2011) *Statistics for Dummies*. John Wiley & Sons. ISBN: 978-0470911082

- Larry J. Stephens (2006) *Schaum's outline of theory and problems of beginning statistics*. Second edition. McGraw Hill Professional. ISBN: 978-0071459327

For material on R and statistical methods and research in linguistics, see R. Harald Baayen's home page, links and textbook.

- R. Harald Baayen (2008) *Analyzing Linguistic Data: A practical introduction to statistics*. Cambridge University Press. ISBN: 978-0521709187

## 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. Significance
3. Probability Theory
4. Information Theory
5. N-Gram models and their application
6. Probabilistic grammars and automata
7. Multivariate analysis
8. Applied topics and examples:
   - Document similarity
   - Authorship study
   - Lexical classification and clustering (distributional properties and part-of-speech)
9. Presentations