CORPORA AND STATISTICS FOR LANGUAGE RESEARCH - LANGUAGE SCIENCES 10901
SPRING 2014
MONDAY, WEDNESDAY 8:30-10:15 / LLCC SEMINAR ROOM

TEACHING STAFF

Dr. Luka Crnič
Office: LLCC 220
Office hours: by appointment
Email: luka.crnic@mail.huji.ac.il

Dr. Aynat Rubinstein (coordinator)
Office: Rabin 1114
Office hours: by appointment
Email: aynat.rubinstein@mail.huji.ac.il

DESCRIPTION

Technological advances have produced a wide range of electronic corpora of natural language, which have become increasingly accessible with the advent of the Internet. They present a valuable resource for all areas of linguistics, ranging from historical linguistics, sociolinguistics and language documentation, to semantics and pragmatics, since they provide a rich empirical basis for developing and verifying linguistic generalizations and theories. High quality research in all these areas has become increasingly dependent on documentation and validation of proposals based on data drawn from corpora. Since facility with working with corpora is becoming a requirement for much work in linguistics and language-related research, this course is designed as an introduction to quantitative corpus methods for students of a variety of linguistic backgrounds.

AIMS

The main goal of the course is to familiarize students with the procedures of corpus work and to demonstrate their usefulness. Students will be taught basic programming skills and statistical methods required to obtain and analyze relevant linguistic data from corpora. Various applications of corpora across languages and linguistic disciplines will be discussed.

LEARNING OUTCOMES

By the end of the semester, students will be able to:

- Describe the usefulness and limitations of corpus methods in linguistics
- Classify corpora according to different parameters
- Analyze data in raw text and using data sets extracted from corpora
- Formulate hypotheses and test theoretical questions using corpus methods
- Present study results using appropriate graphics
- Build corpora to answer specific research questions
- Program in R

ATTENDANCE

You will not be graded on attendance, but participation in lectures is absolutely key to your success in this class (and it will make things more fun!). The content of the course will come primarily from class lectures, so it will not be possible simply to "do the reading".

TEACHING ARRANGEMENT AND METHOD OF INSTRUCTION

Class lectures will encompass both theoretical and practical ("hands on") components. In addition, you will need to meet with the instructors individually throughout the semester in preparation of your final project (more on that below).

CONTENT

- Introduction: the quantitative turn in the humanities, corpus methods in different domains of linguistic inquiry, limitations
- Corpus exploration and visualization:
    - Get to know your corpus
    - Zipf's law
    - Dispersion of words in a corpus (by genre, diachronic timeline)
- Methods in corpus linguistics:
    - Frequency lists
    - Concordances
    - Collocations
- Creating corpora: collection, digitization, markup, annotation
- Statistics for corpus linguistics:
    - Hypothesis formulation and significance testing
    - Collostruction strength
    - Clustering
    - Inter-annotator agreement
- Selected topics: adjectival modification, negative polarity items and scalar particles, modality

REQUIRED READING (SELECTED)

Chomsky, Noam. 1959. A Review of B. F. Skinner's Verbal Behavior. *Language* 35(1):26-58.

von Fintel, Kai. 2006. Modality and language. In Donald M. Borchert (ed.), *Encyclopedia of philosophy* (2nd edition). MacMillan.

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: A practical introduction.* Routledge.

Hacquard, Valentine and Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5: 1-29.

Kennedy, Christopher and Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language* 82: 345-381.

Kilgarriff, Adam. 2012. Getting to know your corpus. *Lecture Notes in Computer Science* 7499: 3-15.

Ladusaw, William. 1980. On the notion 'affective' in the analysis of negative-polarity items. *Journal of Linguistic Research* 1: 1-16.

de Marneffe, Marie-Catherine and Christopher Potts. 2014. Developing linguistic theories using annotated corpora. Draft under review for Ide, Nancy and James Pustejovsky (eds.), *The Handbook of Linguistic Annotation*.

Rubinstein, Aynat, Hillary Harner, Elizabeth Krawczyk, Dan Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. *Proceedings of IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pp. 38-46.

Schwarz, Bernhard. 2000. Notes on *even*. Manuscript, University of Stuttgart.

FURTHER READING

Baayen, R. H. 2008. *Analyzing linguistics data: A practical introduction to statistics using R*. Cambridge.

Fadida, Hanna, Alon Itai, and Shuly Wintner. 2013. A Hebrew verb–complement dictionary. *Language Resources and Evaluation.* Online First.

von Fintel, Kai. 1999. NPI licensing, Strawson entailment, and context dependencies. *Journal of Semantics* 16: 97–148.
Gries, Stefan Th. 2013. *Statistics for linguistics with R: A practical introduction.* 2nd edition. De Gruyter.
Hoeksema, John. 2008. There is no number effect in the licensing of negative polarity items: A reply to Guerzoni and Sharvit. *Linguistics and Philosophy* 31:397–407.
Kennedy, Christopher. 2007. Vagueness and Grammar: The semantics of relative and absolute gradable predicates. *Linguistics and Philosophy* 30: 1–45.
Lahiri, Utpal. 1998. Focus and negative polarity in Hindi. *Natural Language Semantics* 6: 57–123.
Legate, Julie, David Pesetsky, and Charles Yang. 2013. Recursive misrepresentations: A reply to Levinson (2013). Manuscript, University of Pennsylvania & MIT.
Levinson, Stephen. 2013. Recursion in pragmatics. *Language*, 89: 149-162.
Lüdeling, Anke and Merja Kytö. 2008/2009. *Corpus linguistics: An international handbook*. Volumes 1 and 2. [Electronic book version available from the library]
McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge.
Piotrowski, Michael. 2012. *Natural Language Processing for historical texts*. Morgan & Claypool.

EVALUATION

During the semester, three homework assignments will be distributed. These will cover the material discussed in class. You are permitted and encouraged to collaborate on homework assignments. However, each person must hand in their own write-up of the assignment (direct copies or jointly authored assignments are not allowed). We will accept write-ups in Hebrew and in English.

In addition to the homework assignments, students will be required to design and carry out a project on a language, corpus and theoretical question of their choice. Several students can jointly work on the same project, though individual projects are also allowed. Project topics should be approved by the instructors by **the first class after the Passover break**. They will be presented in class in the last week of the semester, and the write-up of the projects will be due on **July 1st**. More detailed instructions will be given throughout the semester.

Breakdown of the final grade

Homework assignments	30%
Project presentation (in class)	20%
Write-up of the project	50%