"Corpora of text and speech & databases in research" 5. –9. October 2015

Venue: Quality Hotel Leangkollen, Asker

Organizer: Janne Bondi Johannessen (UiO)
Area: Corpora in linguistic research
Course credits: 5 ECTS
Contact: jannebj@iln.uio.no

Contents of the course

Short description
Any research task within the field of linguistics can benefit from some empirical grounding. With strict deadlines and many requirements, any help in this respect will mean time saved. At the same time, some empirical research material can also be used as an anchor and a focus point for a thesis – or any kind of research work. Existing empirical data in the form of linguistic corpora can also be beneficial for comparison with other data, or for checking against a gold standard.

In this course PhD candidates will learn how to use specific corpora to get the information they require, and also how to apply statistical measures in order to be sure that the research results are statistically significant.

The course covers corpora with different modalities (spoken/written), different languages (a vast array of languages), multilingual translations, many kinds of metadata (grammatical tags, home of informants, age, gender, and many more). Knowledge of these resources will be advantageous for PhD candidates, both for their own specific research project, but also to get first-hand information of what exists and what can be done. For candidates pursuing a career in academia it will be valuable to have a good overview of digital resources for use both for themselves and for future colleagues and students.

The course offers talks by leading experts in the field of linguistics and corpus linguistics, and hands-on exercises with real-life research questions from various fields of linguistics.

The course, which includes reading the course literature (a reading list will be provided), lectures and a final presentation, covers 5 CSTS.

Preliminary program
Monday
Introduction to corpora

- History
- The corpora at the Text Laboratory, UiO
- Copora and the web
- Corpus annotation: tagging and transcription
- Metadata

- fields of linguistics where corpora can be useful (morphology, syntax, dialectology, sociolinguistics...)
- Introduction to statistics: Learning to use statistical methods for calculating significance in data retrieved from corpus, correlation tests, t-tests and anova.

Tuesday
Introduction to the RUN-Euro corpus (Russian, Norwegian, English, Swedish, Bosnian, Croation, Serbian, Bulgarian, German, Italian, Polish)

Spoken language corpora
    - Nordic Dialect Corpus (with Norwegian, Swedish, Danish, Icelandic and Faroese)
    - NoTa (Corpus of Oslo Speech)
    - TAUS (Corpus of older Oslo Speech)
    - Big Brother Corpus
    - Doctor-Patient Corpus
    - Norwegian in America
    - Ruija-Corpus (Finnish and Kven)
- Hands-on practice with linguistic tasks
- Wrapping-up discussion

Wednesday
Introduction to Oslo Multilingual Corpus (English, Norwegian, German, French, Portuguese)
- Hands-on practice with linguistic tasks

Introduction to databases
    - Nordic Syntax Database
    - Repertory of Conjectures on Horace
    - Kelly (Keywords for Language Learning for Young and adults alike) word pairs from 9 languages: Arabic, English, Greek, Italian, Chinese, Norwegian, Polish, Russian, Swedish)
    - Maid Chinese spoken dictionary
- Hands-on practice with linguistic tasks
- Wrapping-up discussion

Thursday
Introduction to PROIEL (old Indo-European languages: Latin, Gothic, Armenian and Old Church Slavonic)

Monolingual corpora
    - French newspaper corpus
    - Amharic corpus
    - Norwegian Web as a Corpus
    - Lexicographic Bokmål Corpus
    - British National Corpus
- Hands-on practice with linguistic tasks

- Wrapping-up discussion
- Practicing statistics with linguistic tasks

Friday

- Student presentations with feedback (from statistician and other selected instructors)

Concrete dates and number of classes

5. –9. October 2015
6 hours per day for five days

List of instructors

Prof. Atle Grønn (ILOS, UiO)
Prof. Dag Trygve Truslew Haug (IFIKK, UiO)
Prof. Hilde Hasselgård (ILOS, UiO)
Prof. Janne Bondi Johannessen (Text Lab+MultiLing, ILN, UiO)
Høgskolelektor Bård Uri Jensen: (Høgskolen i Hedmark)
Senior engineer Kristin Hagen (Text Lab, ILN, UiO)
Senior engineer Anders Nøklestad (Text Lab, ILN, UiO)
Senior engineer Joel Priestley (Text Lab, ILN, UiO)

Who is it suitable for?

This course is suitable for researchers with some kind of empirical basis in their research. The project should still have some unexplored empirical issues left for the course to be maximally useful. Linguists within all areas (phonology, morphology, syntax, semantics, sociolinguistics, dialectology, discourse analysis etc.) will find the course beneficial.