Hands-on work with data annotation, extraction, and exploitation for linguistic analysis

Malvina Nissim
malvina.nissim@unibo.it

--Topics and activities--

Students will be presented with different levels of annotation, intended for different kinds of linguistic analysis, and with various tools necessary for data processing.

The course will be nearly exclusively hands-on so as to equip PhD students with skills that will enable them to choose and/or create the appropriate corpus for a given study and the appropriate tools to extract and analyse relevant data. Theoretical issues will be touched on, but discussed in relation to the practical activities we will engage in only. Specifically, students will be asked to use specific annotation software, and customize it whenever necessary to match the requirements of a case study. They will also use regular expressions to identify specific patterns in the data and see how information extracted from the text can be used to model specific linguistic phenomena in a machine learning setting.

--Day 1--

1. Part-of-speech tagging with TreeTagger
2. Using regular expressions on annotated data using AntConc
3. Combined searches using CQP

--Day 2--

1. Customization of the GATE annotation tool for specific phenomena (including writing XSLT stylesheets)
2. Manual annotation of a specific phenomenon using GATE; excursus on inter-annotator agreement
3. Extracting relevant information for machine learning experiments with WEKA

The activities of the two days can be conceived as sequential, in an increasing degree of complexity (also of the linguistic phenomena dealt with), but can also be carried out more or less independently of one another.

References

**Software (all freely downloadable):

- TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
- AntConc, http://www.antlab.sci.waseda.ac.jp/software.html
- GATE, http://gate.ac.uk/
- WEKA, http://www.cs.waikato.ac.nz/ml/weka/

References for the software, such as relevant papers and manuals, can be found at the software webpage directly.

**General:

Manning C. and Schuetze H. (1999). Foundations of Statistical Natural Language Processing, MIT Press.

Jurafsky D. and Martin J.H. (2008). Speech and Language Processing, Prentice Hall.

Clark A., Fox c. and Lippin S. (2010). The Handbook of Computational Linguistics and Natural Language Processing, Blackwell.

Other pointers will be provided at a date closer to the course and during the lectures.