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

Jan Hajic
hajic@ufal.mff.cuni.cz

--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. We will use and manipulate data from more than one language (including English, Czech, Italian).

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 (Jan Hajic) --

1. Intro to treebanking

2. Treebank formats - phrase-based ("bracketed"), dependency

3. Levels and layers of annotation: morphology, syntax, semantics

4. Tha Prague Dependency Treebank style of annotation for English and Czech

-- Day 2 (Jan Hajic) --

1. Hands-on annotation of simple sentences (surface dependecny syntax) using TrEd annotation tool

2. Hands-on annotation of simple sentences (semantic annotation, valency)

3. Interannotator agreement and annotation evaluation, principles and basic math

-- Day 3 (Jan Hajic) --

1. Treebank Search using PML-TQ

2. Searching simple dependency and phrase-based trees using simple graphical interface (over the web)

3. Complex searches for cross-layer annotated syntactic and semantic treebanks

-- References --

**Software (all freely downloadable):

- PDT family of treebanks: samples and all documentation at:

  http://ufal.mff.cuni.cz/pdt2.0 Prague Dependecy Treebank (Czech)

  http://ufal.mff.cuni.cz/pedt2.0 Prague English Dep. Treeabnk (Eng., WSJ)

  http://ufal.mff.cuni.cz/pcedt2.0 Prague parallel Eng-Cz Treebank

- PML-TQ search, http://ufal.mff.cuni.cz/pmltq (no need to download, web-based)

- TrEd, http://ufal.mff.cuni.cz/tred

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 Lappin 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.