# Student projects for course "Annotating language data"

**Tomaž Erjavec**
**2006-11-15**

The project counts for 70% of the final grade, and is composed of the practical work + written report. The project work is to be presented and discussed at the last lecture (1.12.2006) and the report handed in by the end of the term, (1.2.2007) at the latest.

Projects can be done individually or in groups. If students prepare the work together, the end result should be proportionately more substantial.

The written report should be written as a standard ACL paper; the stylesheet with instructions is available at http://www.aclweb.org/acl2005/index.php?stylefiles The report should contain an introduction and explain the task undertaken. Esp. interesting is the discussion of problems encountered, and their solutions.

Students are free to come up with their project proposals, which should, however, be first discussed with me. Follow some suggestions:

## Corpus survey

Make a compressive study of available corpora according to some specific criterion (e.g. all corpora of German language, all parallel corpora, all treebanks...) and compare and contrast them according to various criteria (size, conditions of use, encoding, annotation …). The report should have as an Appendix a table giving the corpora / criteria considered. Discuss how you found the corpora, and how much various finding aids turned out to be useful, e.g. Google, corpus overview Web sites, OLAC, catalogues of LDC, ELRA …

## Web corpus

Construct a 100.000 word corpus of English, German, or some other language, concentrating on a particular topic (e.g. nuclear safety, smell related texts, book reviews…). The corpus can be (probably in several steps) gathered with on line BootCat, http://corpora.fi.muni.cz/bootcat/, or, for the more computer savvy, http://sslmit.unibo.it/~baroni/bootcat.html
Analyze the corpus in terms of how well it represents the chosen domain. Perform a lexical analysis of the corpus using a concordancing program (Wordsmith, if available at Graz Uni). You can also use the concordancers available at the on-line BootCat, although there you are limited to a fairly small corpus. Discuss the results.

## Corpus tagging

Construct a 2.000 word corpus of German (or other language), and PoS tag it. You can use the on-line BootCat with the tagging option, or, for the more adventurous, install TreeTagger. Some languages not covered by TreeTagger (such as Slovene) can be tagged by the http://nl2.ijs.si/analyze/ service, by choosing CLOG as the lemmatiser.

Import the data into Excel or other spreadsheet program, and correct the tagging mistakes. In Excel make a quantitative analysis of the tagging errors. Discuss the results.
Reference: Tomaž Erjavec, Bence Sárossy: Morphosyntactic Tagging of Slovene Legal Language. Informatica, in print.

## Semantic lexicon

Get acquainted with Princeton WordNet and the on-line interface to it, and make study of kinship terms (parent, mother, cousin, etc.) How many such words does WordNet distinguish? How are they related? Special attention should be given to the word "mother" – in which synsets (concepts) does it appear in, and what are the links (esp. direct and indirect hyper- and hyponyms) of its meaning as "female parent". Make a translation of some of these concepts into German and discuss the problems that emerge.

A similar thing (just concentrating on mother), using the Omega4 ontology: http://omega.isi.edu/

## Treebank

Register and download the TIGERsearch and TIGERcorpus http://www.ims.uni-stuttgart.de/projekte/TIGER/ for to get access to the TIGER treebank of German. Choose one (small) syntactic phenomenon to study, e.g. expletives. The report should describe the corpus and software, and how it was used to study the syntactic phenomenon.