# Course Advanced Language Technologies

### Module "Knowledge Technologies" Jožef Stefan International Postgraduate School Winter 2013 / Spring 2014

#### URL <http://nl.ijs.si/et/teach/mps13-hlt/>

### Topics for seminar work

### 2014-02-18

The course assessment is through seminar work on a topic connected with HLT, where ½ of the grade is on the basis of the quality of work, and ½ on the quality of the report, comprising its structure, motivation, review of related work including literature, presentation of the data-set and experiment, conclusions and language. For more substantial projects, students also have the option of performing the work in pairs (they will both get the same grade).

In 2014 there is also the biennial Slovenian conference on language technologies (c.f. <http://www.sdjt.si/konference.html>) with the paper deadline in June. Students are encouraged to submit their seminar work to the conference.

Some possible topics are given below, but students have the option of choosing the topic themselves, in agreement with the lecturer.

Each student can have consultations with the lecturer regarding their project – send email to arrange for the time of the consultation.

The seminar should ideally be finished by the exam date in late May / early June. The seminar work should be sent to the lecturer one week before the exam date.

# Suggested topics

Most of the topics take some (Slovene) language resources as the starting point, and the task is to develop (either with heuristics or with some machine learning approach) a model for a particular aspect of Slovene. Most of the resources mentioned below are available from <http://eng.slovenscina.eu/>

For related work, in addition to Google, there is also <http://saffron.deri.ie/> for searching in HLT proceedings.

## Derivational relations between Slovene lemmas

The Sloleks inflectional lexicon of Slovene contains 100.000 lemmas together with their inflectional paradigms, and is a very important resources for processing Slovene. However, many lemmas are still missing. One way to extend the lexicon is to include derivationally related words which are currently missing. For example, the lexicon might contain the lemma “pisati” but is missing “podpisati”, “prepisati” or “pisariti”. On the basis of the lemmas already in Sloleks, see if it is, and by how much, possible to extend the lexicon using such derivational relations. The prefixes / suffixes might be chosen by hand or induced on the basis of the lemmas in Sloleks. A corpus (e.g. ccKRES) can be used to identify missing lemmas and assess the extension of the lexical coverage. The seminar should include a lexicon of prefixes & suffixes for word-formation.

## Developing a Slovene-English machine translation system

While statistical machine translation has quite complex underlying mathematics, it has now become relatively simple to install and use SMT thanks to the Moses toolkit. The project will take an already prepared English-Slovene parallel corpus and train Moses on this data and make an evaluation of the results. Knowledge of Linux is needed.

## Using NooJ for derivations, inflections, local grammars

NooJ, <http://www.nooj4nlp.net/pages/nooj.html> is a finite-state environment for writing grammars, from inflection to named entities. A few resources have been already written for Slovene (<http://www.nooj4nlp.net/pages/slovene.html> ) but there is much opportunity to expand them.

## Adding processing for Slovene to a OS HLT library

Many resources for Slovene are now available under permissive licences, so it is possible to add processing for Slovene to various open source libraries. Two in particular, are interesting:

* FreeLing, <http://nlp.lsi.upc.edu/freeling/> where Slovene is lacking most modules. Tokenisation or sentence splitting could be first additions. Linux experience is needed
* NLTK, <http://www.nltk.org/>

## WordNet and sense similarity

The task is to adapt Ted Pedersen's WordNet::Similarity package for Princeton WN to work with Slovene WN. The task includes a small evaluation of the results.

WordNet::Similarity takes two words and computes their semantic relatedness given wordnet; it can use various similarity measures, not all of them possible for Slovene.

Prerequisites: knowledge of Perl / Linux; knowledge of Slovene for evaluation.

<http://nl.ijs.si/slownet/>   
<http://www.d.umn.edu/~tpederse/similarity.html>

## Tagging Slovene with RFTagger

The task is to train and evaluation the RFTagger on Slovene, using the JOS corpora. The tagger has already been trained on Slovene data, but not very well. You should convert the JOS resources into the format needed by the tagger, train it and evaluate the results.

Prerequisites: some knowledge of programming.

<http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/>  
<http://nl.ijs.si/jos/>

## Writing HLT Wikipedia articles in Slovene

This might be appropriate for non-programmers and native Slovene speakers. The article(s) should also give pointers to specifically Slovene resources. Care should be taken with terminology and, possibly, related articles. On possibility would be to write an article on SSJ project.

# Chosen topics

--