Assignment #4: Extraction of subject-verb-object triples

## **Objectives**

The objectives of this assignment are to:

* Extract the subject–verb pairs from a parsed corpus
* Extend the extraction to subject–verb–object triples
* Understand how dependency parsing can help create a knowledge base

## **Organization and location**

The fourth lab session will take place on

* Group 1: Wednesday, October 1 from 10:15 to 12:00 in the Alpha room
* Group 2: Wednesday, October 1 from 10:15 to 12:00 in the Gamma room
* Group 3: Wednesday, October 1 from 13:15 to 15:00 in the Gamma room

You can work alone or collaborate with another student.

Each group will have to:

* Extract all the subject–verb pairs and subject–verb–object triples from a parsed corpus
* Rank the pairs and triples by frequency
* Create a simple cofererence solver for the *som* ‘who’ pronoun

## **Programming**

This assignment is inspired by the Prismatic knowledge base used in the IBM Watson system. See [this paper](http://www.aclweb.org/anthology/W/W10/W10-0915.pdf) for details.

In this session, you will use a parsed corpus of Swedish to extract the pairs and triples.

### **Choosing the parsed corpus**

1. In this assignment, you will use the CONLL-X Swedish corpus. Download the tar archives containing the training and test sets for Swedish and uncompress them: [[data sets](http://ilk.uvt.nl/conll/free_data.html)]. Local copies: [[training set](http://fileadmin.cs.lth.se/cs/Education/EDAN20/corpus/conllx/sv/swedish_talbanken05_train.conll)] [[test set](http://fileadmin.cs.lth.se/cs/Education/EDAN20/corpus/conllx/sv/swedish_talbanken05_test_blind.conll)] [[test set with answers](http://fileadmin.cs.lth.se/cs/Education/EDAN20/corpus/conllx/sv/swedish_talbanken05_test.conll)]. You will use the training set only.
2. Draw a graphical representation of the two first sentences of the training set.
3. Download [What's wrong with my NLP](http://code.google.com/p/whatswrong/) and use it to check your representations.

### **Extracting the subject–verb pairs and subject–verb–object triples**

You will extract all the subject–verb pairs and the subject–verb–object triples from the training corpus. To help you start, download this program that reads the CoNLL corpus format [[code](http://fileadmin.cs.lth.se/cs/Education/EDAN20/programs/parser/format.zip)].

You can design the program you want. However, here is a suggestion on how to organize your program:

* Create a new package in the Java program that you will call store. This package will hold the extraction classes.
* Create the class Triple to represent the triples and a class TripleStore to hold all the triples
* You will probably need to use the Java collections: maps, lists, and sets. Review them before you start this lab.
* Extract all the subject-verb pairs in the corpus. The subject function uses the SS code in the Swedish corpus of CoNLL-X.
* Compute the total number of pairs. You should find 18885 pairs.
* Sort your pairs by frequency order and give the five most frequent pairs
* Extract all the subject–verb–object triples of the corpus. The object function uses the OO code. Compute the total number of pairs. You should find 5844 triples.
* Sort your triples by frequency order and give the five most frequent pairs
* Should you be interested, you can try your program on multiple languages using corpora annotated by Google. [[link to the corpora](http://code.google.com/p/uni-dep-tb/)].

### **Solving coreferences (Optional)**

In this part, you will resolve a simple anaphor involving the *som* ‘who’ pronoun.

* Use *What's wrong with my NLP* to derive a simple rule to find the antecedent of *som*
* Replace all the occurrences of *som* as a subject with its antecedent in your pairs and triples.