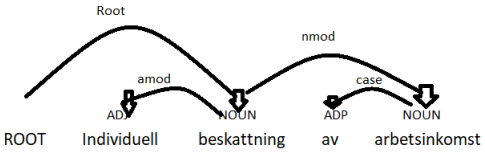
EDAN20 - Assignment 5

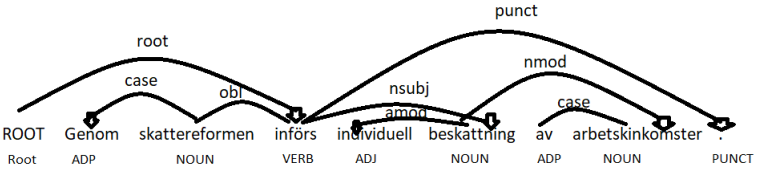
Oliver Cosic

September 2021

1 Introduction

The assignment is about the extraction of subject-verb-objects triples. The first thing we do is a graphical representation of two sentences in Swedish. My drawings are shown below. Visualizing with the conllu.js tool we get more or less the same, however the langoria pipelines use different annotations then I did so it was a bit different but the arrows are more or less the same, although pointing in the oppsite direction, unclear if I did a mistake or if it is supposed to be like that.





2 Extracting the subject-verb pairs

The next step was to extract the subject-verb pairs. This was done with the following function: And the three most frequent was: [(('har', 'som'), 45), (('får',

```
def extract_pairs(corpus):
    pairs={}
    for sentence in corpus:
        #print(sentence)
        for key in sentence.keys():
            if(sentence[key]['DEPREL'].startswith('nsubj')):
                headidx=sentence[key]['HEAD']
                headWord=sentence[headidx]['FORM'].lower()
                word=sentence[key]['FORM'].lower()
                if(headWord, word) in pairs:
                    pairs[(headWord, word)]+=1
                else:
                    pairs[(headWord, word)]=1
    return pairs
```

'du'), 19), (('har', 'vi'), 19)]

3 Extracting the subject-verb-object triples

We configured the code from the pair extraction to extract triples instead. Function is shown below: And later we compined the two in a function that returned

```
def extract_triples(corpus):
    three_pairs={}
    for sentence in corpus:
        #print(sentence)
        for key in sentence.keys():
            if(sentence[key]['DEPREL'].startswith('nsubj')):
                headidx=sentence[key]['HEAD']
                headWord=sentence[headidx]['FORM'].lower()
                word=sentence[key]['FORM'].lower()
                for key2 in sentence.keys():
                    if(sentence[key2]['DEPREL'].startswith('obj')):
                        obj=sentence[key2]['FORM'].lower()
                        objHead=sentence[key2]['HEAD']
                        if(objHead==headidx):
                            if(word, headWord, obj) in three_pairs:
                                three_pairs[(word, headWord, obj)]+=1
                            else:
                                three_pairs[(word, headWord, obj)]=1
    return three_pairs
```

both the nbest frequent pairs and triples. We did that for all the languages in the corpora. The three most frequent in French, Russian and English is shown below.

4 Resolving the entities and Extracting chunks

The last part of the assignment was to extract the relations involving named entities, where both the subject and the object are proper nouns. We did this with a function and got the expected results (suggested by the notebook). Extracting the chunks manually from the corpora we got the following:

```
freq_triples_fr
```

```
[(('il', 'fait', 'partie'), 16),
 (('elle', 'fait', 'partie'), 7),
 (('il', 'comptait', 'habitants'), 7)]
```

Russian

```
# Write your code here
```

```
#freq_pairs_ru, freq_triples_ru = extract_pairs_and_tr
freq_triples_ru
```

```
[(('мы', 'имеем', 'дело'), 6),
 (('мы', 'имеем', 'что'), 4),
 (('мы', 'сделаем', 'все'), 4)]
```

English

```
# Write your code here
```

```
#freq_pairs_en, freq_triples_en = extract_pairs_and_tr
freq_triples_en
```

```
[(('you', 'have', 'questions'), 22),
 (('you', 'think', 'what'), 12),
 (('i', 'do', 'what'), 7)]
```

The complete sentence: 'Goodwyn dutifully notes that Baba Groom didn't remember George telling drunk stories.'

Baba is a complete noun group by itself and so is George. We have to add didn't to remember to get the complete verb group.

So (Baba, didn't remember, George).

5 Article

The article *PRISMATIC: Inducing Knowledge from a Large Scale Lexicalized Relation Resource by Fan and al* (2010) presents a large scale lexicalized relation resource. Comparing the article to this assignment the fourth part of their process, called frame-cut-extraction, is more or less the same thing that we have been doing in this lab. The big difference is that they extract a variation of frame cuts. Where we extracted 'S-V-O' they also have the option to extract 'S-V-O-IO' and 'S-V-P-O'.