Background

McDonald et al. (2013) introduced a collection of homogeneous treebanks in six languages (German, English, Swedish, Spanish, French, and Korean). Later, version 1.0 of the universal dependency treebanks was released which included more languages (Czech, Finnish, Irish, Italian, and Hungarian). The promise is that syntactic annotations are consistent across languages when possible.

Goals

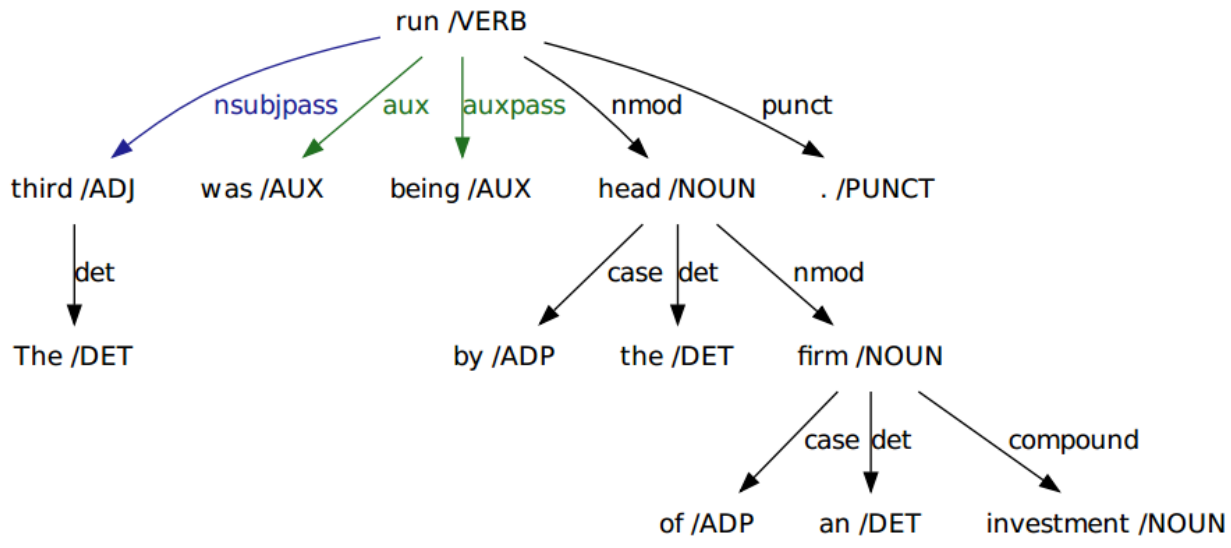
In this project, we propose to analyze (both quantitatively and qualitatively) the (dis)similarities between treebank annotations in different languages. The project goals are as follows:

1. Develop a useful tool for exploring the universal dependency treebanks.
2. Verify the consistency of annotations across languages in the universal dependency treebanks.
3. Quantify the degree to which different languages exhibit certain syntactic patterns.

Details

Consider the following dependency parse tree in the universal dependency treebank for English:

The third was being run by the head of an investment firm .



Following are some syntactic patterns we can extract from this parse tree (in GFL notation):

ADJ >$_{nsubjpass}$ VERB	from “third … run”
DET >$_{det}$ ADJ >$_{nsubjpass}$ VERB	from “The third … run”
AUX >$_{aux}$ VERB	from “was … run”
AUX >$_{auxpass}$ VERB	from “being run”
VERB <$_{nmod}$ NOUN	from “run … head”
VERB <$_{nmod}$ (ADP >$_{case}$ NOUN)	from “run by … head”
VERB <$_{nmod}$ (DET >$_{det}$ NOUN)	from “run … the head”
VERB <$_{nmod}$ NOUN <$_{nmod}$ NOUN	from “run … head … firm”
VERB <$_{nmod}$ NOUN <$_{nmod}$ (ADP >$_{case}$ NOUN)	from “run … head of … firm”
VERB <$_{nmod}$ NOUN <$_{nmod}$ (DET >$_{det}$ NOUN)	from “run … head … an … firm”
VERB <$_{nmod}$ NOUN <$_{nmod}$ (NOUN >$_{compound}$ NOUN)	from “run … head … investment firm”

Our analysis will be primarily based on quantifying how salient different syntactic pattern are across languages. We can, for instance, compute the percentage of each pattern in each language and display as shown below in a tabular format[1]. Note how some syntactic constructions are rare/missing in some languages.

syntactic pattern	**en**	**fr**	**ar**	**.....**
ADJ >$_{amod}$ NOUN	2.1%	0.0%	0.0%	
NOUN <$_{amod}$ ADJ	0.0%	2.4%	2.2%	
ADJ >$_{nsubjpass}$ VERB	0.4%	0.0%	0.0%	
DET >$_{det}$ ADJ >$_{nsubjpass}$ VERB	0.2%	0.0%	0.0%	
AUX >$_{auxpass}$ VERB	0.8%	0.9%	0.0%	
VERB <$_{nmod}$ (DET >$_{det}$ NOUN)	2.0%	2.1%	0.0%	
VERB <$_{nmod}$ NOUN <$_{nmod}$ (DET >$_{det}$ NOUN)	1.0%	0.9%	0.0%	
VERB <$_{nmod}$ NOUN	2.3%	2.8%	3.1%	
…..				

We plan to achieve our first goal by implementing a tool which extracts this data from the universal dependency treebanks, in addition to enabling simple commands to filter patterns based on various criteria.

Our second goal will be achieved by indepently identifying some syntactic patterns which are expected to be salient in certain languages but not in others (using our knowledge about language typologies), and then evaluate the degree to which our expectation matches the empirical frequency of these patterns in the universal dependency treebanks.

As for our third goal, we plan to use statistical tools to identify language pairs/clusters which are most/least similar, and contrast our findings to typological features of studied languages, as documented by previous work.

[1] The frequencies mentioned here are set arbitrarily for illustration purposes only.