Linguistic Data 2: Collection & Modeling

Assignment 1

24 January, 2020

Each one of the course participants have been allotted a corpus consisting of 100 sentences each. The task is as follows:

- Tokenise your respective corpus. Make sure each token has a *token_id*.
- For every token in the resultant corpus, run the shallow parser to get the morph analysis and the POS tag for every token.
- For every sentence, mark the dependency label for each token in the sentence.

After performing the above steps, each line in your corpus should look like:
token_id $<$ *token* $>$ *POS_tag* *morph_analysis* *dependency_label*

Resources

- Your respective corpus can be found here.
- Shallow parsers for all languages can be found here. Set up the shallow parser or use the online interface, whichever suits you. In using the online interface, in order to get the POS tags and morphology analysis for every token in the sentence, click on "Intermediate Outputs" and take the required information from that page.
- The dependency relations tag-set can be found here.

By the next class on 28 January, each one of you should be done with a couple of files and be ready to show your output files to the whole class.

Post any and all doubts on the Moodle thread for this assignment.

DEADLINE: 3 February 2020, 11:55 PM