

CSL1.204 : Language Typology and Universals

Spring 2021

Assignment 3

Date of Issue: 16th February, 2021

Date of Submission: 23rd February, 2021, 11:55pm

Take both Hindi and English sentences from the [IIT Bombay English-Hindi parallel corpus](#), take 100 sentences from each language and:

- POS tag them
- Parse them for Universal Dependencies

Using available libraries (recommendation: Spacy for English and Stanford CoreNLP for Hindi, or Stanford CoreNLP for both languages. A tutorial will be held on the usage of Stanford CoreNLP soon.)

And identify:

- Repeated patterns in word order and clause: subject object, s-v-o, part of speech pattern, prepositions, nouns, adjectives, etc..
- The difference in expressions of similar deep structures (noun-adjective order, etc) between English- a rigid word order SVO language- and Hindi- a SOV free word order language.
- Bonus marks: find patterns relating to interdependency of clauses (using the parse tree)

Submit:

- The sentences analyzed- return both POS tag parses and dependency parses.
- The code for the *analysis* - your code for searching for patterns, and analysis- will be checked for plagiarism.

Marking:

This assignment will be marked out of 50, for finding patterns using the POS tagged data. 10 bonus marks will be given for any form of analysis of the dependency parsed data.