Comp 6751 # Project 2 **Fall 2024**

Due date November 30, 2024

Goal: Develop a feature grammar for English that

- enforces number agreement in the noun phrase
- enforces person, number agreement between subject and verb
- enforces grammatical constraints for named entity detection

Data: you should continue with the pil corpus in NLTK, but in order to develop more general Named Entity detection, you should find test sentences from other sources. Important: do not limit yourself to Person, Location, Organization.

Description:

1. You are to develop a feature grammar for NLTK's feature based Earley's Chart parser (parser class "FeatureEarleyChartParser").

   You are trying to develop the best grammar you can: wider coverage, less acceptance of ungrammatical information. You may of course introduce many more features aside from the minimum requirements for this project but focus on the assigned issues first.

   You may use the NLTK feature grammars feat0.fcfg, feat1.fcfg, simple-sem.fcfg for inspiration, but you may choose to start from scratch. Use the Discussion Forum regularly, no matter how you proceed.

2. You are to develop entity detection and annotation within the Earley feature grammar without use of gazetteer lists for drugs. You have to go beyond the NLTK NE module's techniques to find and properly annotate named entities of three different types:

   (a) single and multiple word organizations

   (b) pil corpus drug product names (i.e.*Cortysil*)

   (c) pil corpus drug and chemical component names (i.e. *Cortisone Acetate BR, magnesium stearate, maize starch*)

NLTK resources: NLTK Ch 9 explains how to write a feature grammar in detail. It is required reading for this project.

Deliverables:

Create a file *Good* with your training sentences that your grammar parses and labels correctly.

Create a file *False* with your training sentences that your grammar does not parse or label correctly.

1. 1 file: your well-annotated grammar that covers all aspects of this project (agreement and entity detection) (18pts, Grad Attr. 4,5,6)
2. 1 file: *Good* with annotations (1pts, Grad Attr. 4,5,6)
3. 1 file: *False* with annotations (1pt, Grad Attr. 4,5,6)
4. 1 file: A report that includes:
   (a) justification of your design (Why did you make the grammar do what it does? Why did you allow it to not do what it doesn't? Don't forget to talk about the agreement features and the entity detection!) (3pts, Grad Attr. 1, 6)
   (b) one page critiquing your design (What does your grammar not do that you think important?) (1pt, Grad Attr. 1, 6)
   (c) one page outlining ideas what kind of semantics you can do with feature grammars (and how much additional work that would require) (1pts, Grad Attr. 1, 6)