GOALS

• Apply a natural language Context-Free Grammar (CFG) grammar
• Gain hands-on experience with an NLP constituency parser
• Examine variation in syntax structure among English language texts

PROBLEM 2 – Constituency parsing

In this problem, you will use the Stanza Constituency parser to check your work from Problem 1. Note
that by default, Stanza uses the Penn Treebank model for the English language LINK.
• Check your work for Problem 1 by applying the Stanza constituency parser to the three sentences of Problem 1.
• Show your output in your notebook. Your output should show a sentence parse with constituent
labels for each of these three sentences.
• You may go back to Problem 1 and revise your answers if you made mistakes, but you should
first try the problems by hand, to become familiar with CFG production rules as used in NLP.
• Note that the Stanza parser outputs one parse for 1(c), "She buys a gift with gold". You will need to identify the other possible parse for this ambiguous sentence in Problem

PROBLEM 3 – Reading the data

• Read in data the following three files in the folder Texts-Together-OneCSVperFile: (Since parsing
is computationally expensive, we will use a very small dataset for this problem)
- climate change.csv
- Gangs.csv
- Thatcher.csv

• Remove rows that do not have an "Elementary" parse, and then merge all 3 datasets in a single
combined dataset.
• To show that you have loaded the data correctly, print the number of rows in your combined
dataset. Show this number in your notebook. You should see **35 rows,** after removing rows with
no Elementary text.
• For the first row in your dataset, print the Elementary and Advanced texts. Show the output in
your notebook. Does the Advanced text seem to use more complex language than the Elementary text? (You do not have to answer this question in writing)

PROBLEM 4 – Analyzing the data

In this problem, you will compare Elementary and Advanced texts that you read in the previous
problem, to consider how texts that express the same ideas can vary syntactically by reading level.
• Write a function that takes a list of texts as input, applies the Stanza constituency parser to each

multi-sentence text, and then uses the output to create a ***data summary*** of these texts. Your
output should include these attributes:

- The average number of sentences in each text
- The average number of ***prepositional phrases*** in each text [You can compute this by scanning the tree recursively, or by searching the output of stanza's "pretty_print()" function for "PP", the Penn Treebank symbol for prepositional phrases]
- ***One other attribute of your choice*** that is based on output of the stanza pipeline

• Apply your function twice, to the data you created in Problem 3:

- The set of 35 Elementary texts
- The set of 35 Advanced texts (after dropping rows with no analogous Elementary text)

• Show your results in your notebook. Check that you are showing all 3 attributes on both the
Elementary and Advanced datasets.