Finished goals in green; IP in yellow.

Stanford Data Extraction

1. Write class to extract coreference data from XML (StanfordDocumentInformation)
   1. Write so that it will work in conjunction with OrganizedBigtable
   2. Ensure that the data in StanfordDocumentInformation is ordered in the right way
2. Write class to correlate the results (OrganizedBigtable)
   1. Make option for hyphenating or not hyphenating results
3. Refactor savePOS.py so that it is integrated into StanfordDocumentInformation and OrganizedBigtable.
   1. Add support for extracting PoS to OrganizedBigtable.
   2. Add ingestion support to the DocumentInformation.
4. Clean up BigTabletoStanford.py and integrate into OrganizedBigTable.py
5. Extract syntactic information from Stanford XML (in StanfordDocumentInformation).
6. Integrate syntactic information into StanfordBigtableCorrelator.

Column Addition

1. Order the columns
2. Write script to run all Stanford and add current coreference/PoS info to bigtable (Run\_Stanford.py).
   1. Make sure to include options both for cross-speaker and non-cross speaker settings.
3. Clean up and make a fully automated Run\_Stanford (assuming that the Java code for PoS is either run independently or irrelevant).
4. Add syntactic information column into Run\_Stanford.
   1. Dump parse tree into the csv
   2. Figure out how to represent syntactic complexity and function in sentence
5. Create script to add all “previous mention” categories (e.g. previous explicit mention part of speech, non-explicit mention PoS) to bigtable.
6. Add function to that script to calculate how many times explicitly mentioned/non-explicitly/both and add as a column.