Author: Ryan Timbrook
UW Net ID: timbrr
Project: Ling 473 Project 1
Date: August 4, 2016

Description:

This project is an application which counts the number of syntactic constituent types that occur in an annotated corpus. The corpus is a portion of the Treebank-3 corpus from the Linguistic Data Consortium.

Approach:

For this assignment I took a procedural programming approach using Python. To simplify the parsing of the document content I created a utility function which replaced all new line and tab spaces with a single white space. I utilized a simple in memory dictionary object to maintain the constituent counts while the program looped through all of the corpus files loaded to the specified directory we were analyzing this data from.

The simple constituent searches for, sentences, noun phrases, and verb phrases, were straight forward using a basic search pattern of the characters specific to the PTB symbol pattern listed below. The Ditransitive Verb Phrase and Intransitive Verb Phrase counts were trickier, where I utilized a numpy array object to help with specifying conditional statements. I used the search feature of the numpy array object to identify the index locations of the constituent I was searching for then using this index I was able to evaluate the constituents immediate elements determining if they were the right patterns which define a ditransitive verb phrase and intransitive verb phrase.

## Results:

| **Constituent** | **PTB symbol** | **Count** |
|---|---|---|
| Sentence | (S …) | 4669 |
| Noun Phrase | (NP …) | 13221 |
| Verb Phrase | (VP …) | 7920 |
| Ditransitive Verb Phrase | (VP verb (NP …)(NP …) ) | 372 |
| Intransitive Verb Phrase | (VP verb ) | 1260 |