Topic- Building a Child Language analyser.

FIT9133 Programming foundations in python S2 2018.

29968550

Ashwani Kumar Singh

Monash University

Table of Contents

Introduction

This assignment is implementing a basic language analyser to investigate the linguistic characteristics of children with some form of language disorders. The analyser is able to perform basic statistical analysis on a number of linguistic features and also to present the analysis using some form of visualisation.

The data set is a collection of narrative transcript narrative transcripts. There are two data sets the first set is form children diagnosed with Specific Language Impairment(SLI) -one form of disorders; and the second set is form children with the typical development(TD).

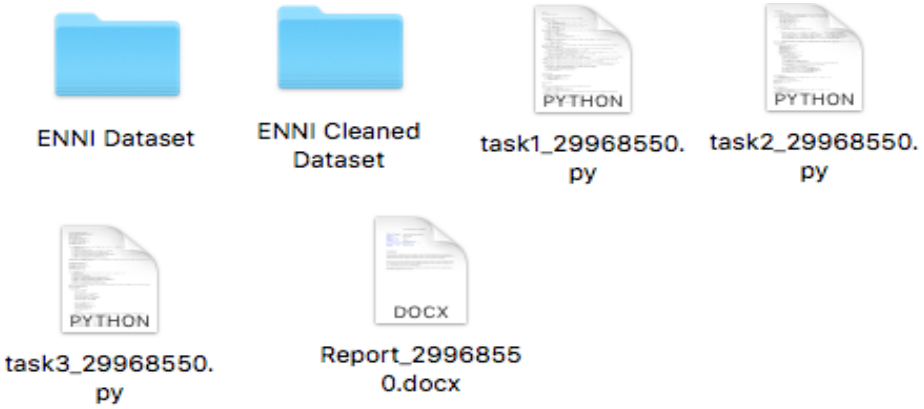
Each narrative transcripts is a record of story-telling task performed by child for the two groups(SLI and TD), under the supervision of an examiner.

Task 1

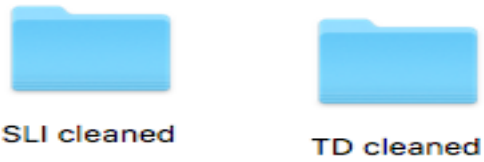
A number of pre-processing task is conducted to extract only the relevant contents or texts for analysis and visualisation. After completing pre-processing task each of the cleaned transcript is saved as an individual output file. I've created a ENNI cleaned dataset folder which has two folder inside. The first is SLI cleaned in which all the SLI cleaned files are saved. The second is TD cleaned in which all the TD cleaned files are saved.

In task 1 since the given file does not have [*] in it so I've converted [*m :+ed] to [*] and (..),(…) to (.).

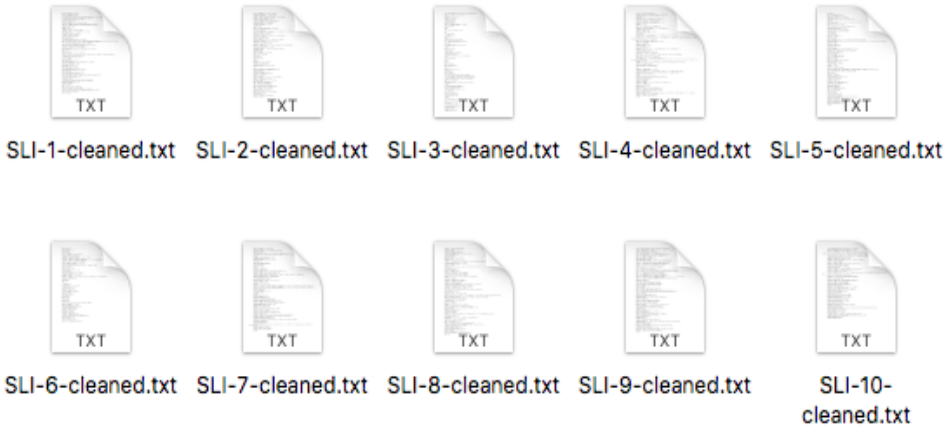
Some of the screen shot for Task 1



Cleaned ENNI Cleaned Dataset folder.



Two cleaned folder inside ENNI Cleaned Dataset



Cleaned files inside SLI cleaned

Task 2

In second task all the required data for analysis are collated.

The statistics for each of child transcript that we are interested in are:

- Length of the transcript — indicated by the number of statements
- Size of the vocabulary — indicated by the number of unique words
- Number of repetition for certain words or phrases — indicated by the CHAT symbol [/]
- Number of retracing for certain words or phrases — indicated by the CHAT symbol [//]
- Number of grammatical errors detected — indicated by the CHAT symbol [*]
- Number of pauses made — indicated by the CHAT symbol (.)

Statistics=(length of transcripts, number of repetitions, number of retraces, Unique words, number of grammatical errors , number of pauses).

Some of the screen shot for Task 2

```
Statistic of SLI transcript are:


    The length of the statement is  : 716
    The number of phrases repeated  : 230
    The number of phrases retraced  : 146
    The number of unique words used : 574
    The number of grammatical errors: 4
    The number of pauses made       : 251
```

SLI count of all the statistics.

```
Statistic of SLI transcript are:


The length of the statement is  : 865
The number of phrases repeated  : 179
The number of phrases retraced  : 163
The number of unique words used : 682
The number of grammatical errors: 1
The number of pauses made       : 359
```

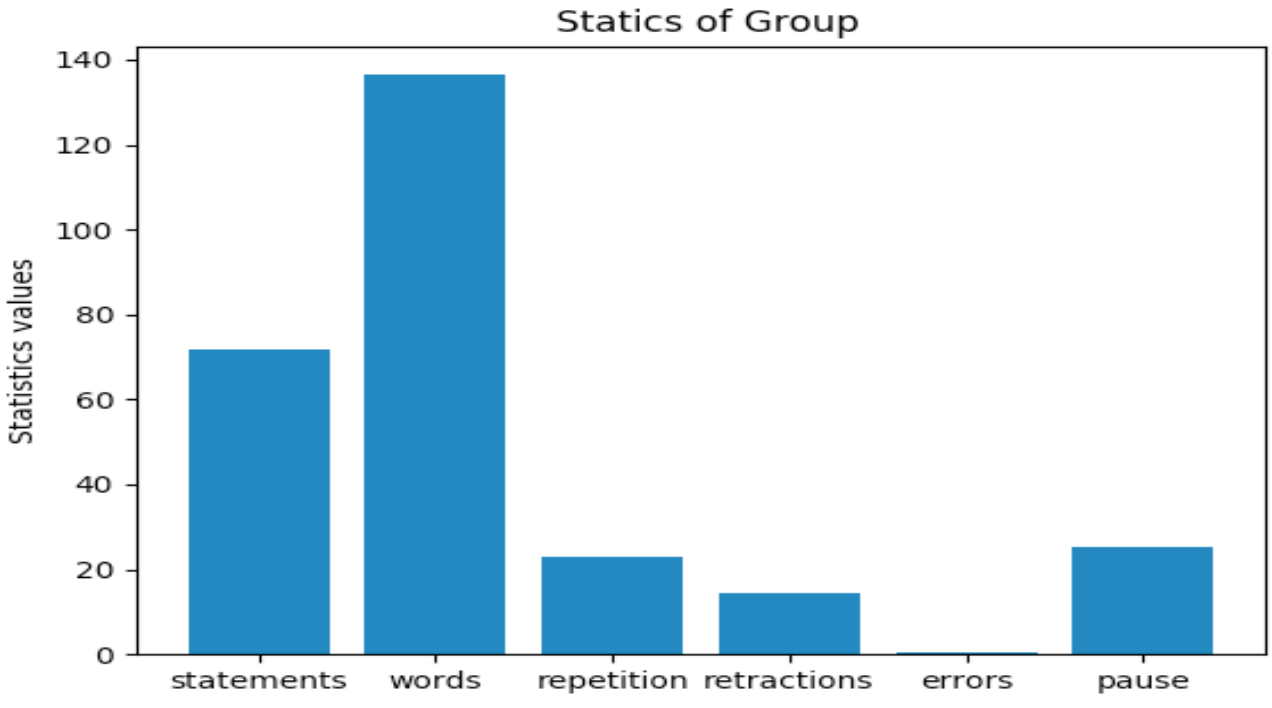
TD count of all statistics.

Task 3

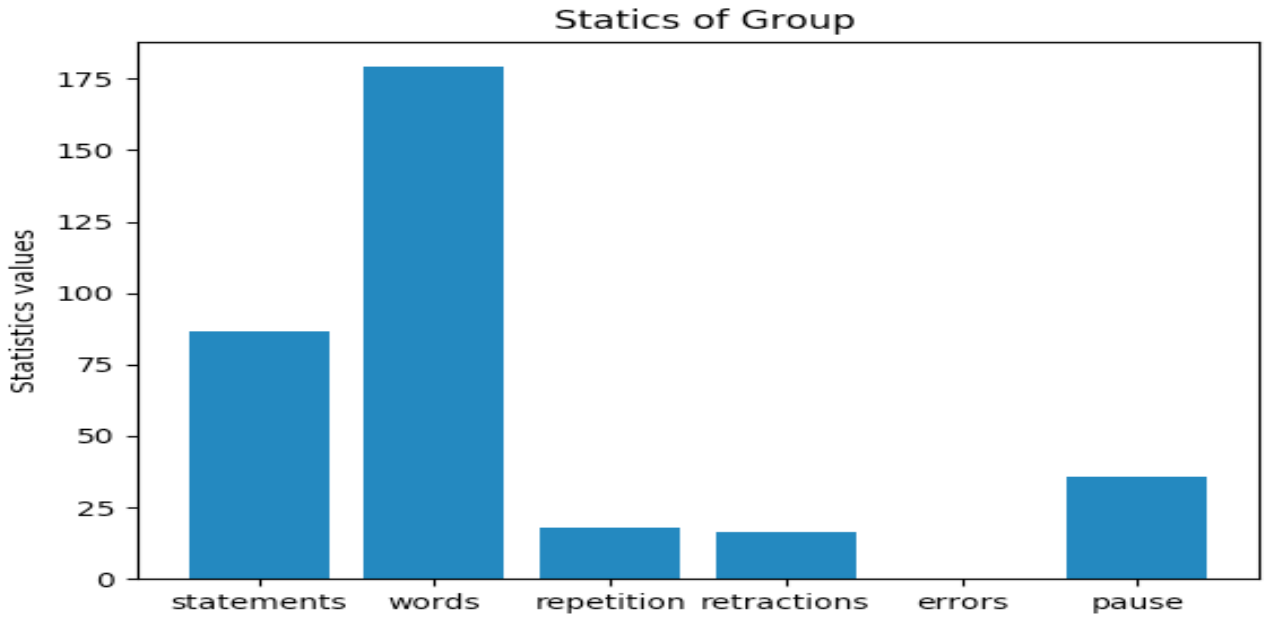
In task 3 I'm appending all the values of the six list variables from all files and the storing these list inside list. When this list is passed to Visualiser class the list of list is used to create a dictionary data frame.

The vis function is used to plot only one graph of any group at a time. Since I'm returning list from compute_average function so I'm storing two separate list for both the group. These stored list is used to compare the statistics values of both the group and the coding for drawing the graph is written outside the class.

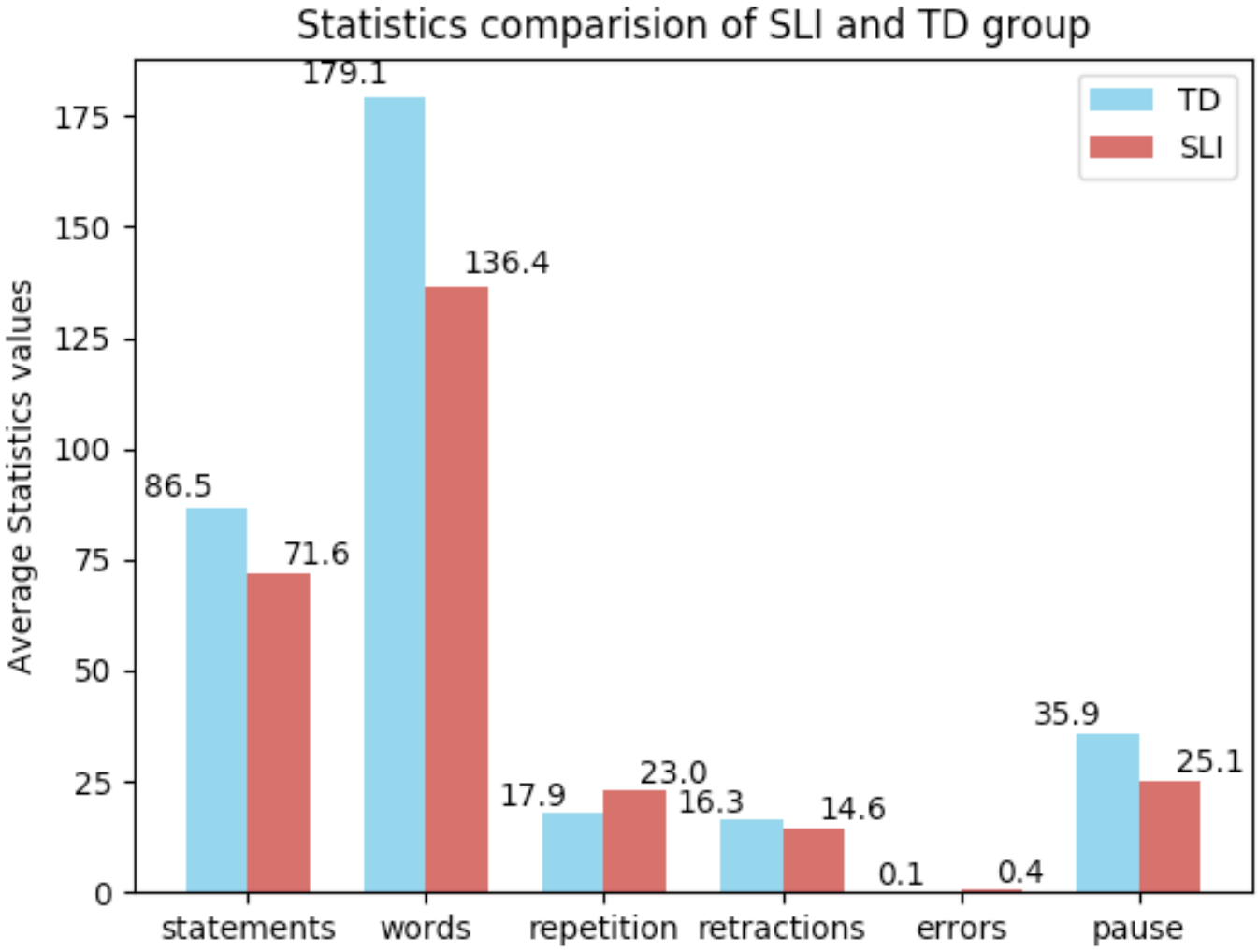
Some of the screen shot for Task 3



Statistics data values for SLI group



Statistics data values for TD group



Statistics comparison between both the groups

Some of the library used in my programming:-

1) Re
2) Matplotlib
3) Numpy
4) Panda

References

1) https://www.alexandriarepository.org :- To learn the basic of python.
2) https://www.w3schools.com/python :- To learn the basic syntax and semantics of python.
3) http://pythontutor.com:- To visualize code and identify the mistakes.
4) Numpy,matplotlib,pandas,regex and re documentation to learn about them.