Overview:

For this project I analyzed the dependency distance of all of the publicly available Perseus treebank data, including Thucydides Book 1, as well as the sections of the Plague, Perikles’ last speech, and Cleon’s speech which I treebanked. Dependency distance, acccording to a talk given by Bob Gorman at Tufts at CAMWS 2016, is a useful metric for determining the difficulty of a text. In short, the further the distance between a word and the word it depends upon, the more difficult it is for a reader to make the connection accurately. Since all of Thucydides Book 1 is available from Perseus, I treated that large treebank as a stand-in for Thucydides as a whole, like a representative sample of his work.

Methodology:

I calculated dependency distance for large amounts of treebanks quickly using a program which I wrote in R. The program is availabe at (Github link). The program is simple: it finds the value of the “id” and “head” tags for every word in a treebank, finds the absolute difference between the id and the head, finds the average of those absolute distances, and then writes the output to a csv file. The whole process is in a loop, so that it can pull data from every treebank file in a given folder and return the dependency distances for all of them. Due to a formatting difference in the publicly available trees and student produced trees, I used separate programs for the Perseus trees and my trees, but they are functionally the same.

Data:

Table 1: Dependency Distance in the Perseus Treebank

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | tlg0548.tlg001.perseus-grc1.1.1.1-1.4.1.tb.xml | 7.32332 | | tlg0007.tlg015.perseus-grc1.tb.xml | 6.428384 | | tlg0020.tlg001.perseus-grc1.tb.xml | 6.366165 | | Thuc1.1-1.146.xml | 6.234316 | | tlg0540.tlg023.perseus-grc1.tb.xml | 6.001122 | | DioSic11.1-11.92.xml | 5.686297 | | tlg0540.tlg014.perseus-grc1.tb.xml | 5.578001 | | tlg0540.tlg015.perseus-grc1.tb.xml | 5.548246 | | Polybius1.1-1.88.xml | 5.527891 | | tlg0007.tlg004.perseus-grc1.tb.xml | 5.134653 | | Hdt1.1-1.216.xml | 5.046342 | | tlg0008.tlg001.perseus-grc1.12.tb.xml | 4.9932 | | tlg0540.tlg001.perseus-grc1.tb.xml | 4.965773 | | tlg0008.tlg001.perseus-grc1.13.tb.xml | 4.808166 | | tlg0012.tlg001.perseus-grc1.tb.xml | 4.687219 | | tlg0059.tlg001.perseus-grc1.tb.xml | 4.666719 | | tlg0020.tlg003.perseus-grc1.tb.xml | 4.622267 | | tlg0012.tlg002.perseus-grc1.tb.xml | 4.595851 | | tlg0020.tlg002.perseus-grc1.tb.xml | 4.593451 | | tlg0013.tlg002.perseus-grc1.tb.xml | 4.5688 | | tlg0011.tlg001.perseus-grc2.tb.xml | 4.462774 | | tlg0085.tlg002.perseus-grc2.tb.xml | 4.418025 | | tlg0011.tlg002.perseus-grc2.tb.xml | 4.411568 | | tlg0011.tlg004.perseus-grc1.tb.xml | 4.391633 | | tlg0011.tlg003.perseus-grc1.tb.xml | 4.306738 | | tlg0085.tlg004.perseus-grc2.tb.xml | 4.296657 | | tlg0085.tlg005.perseus-grc1.tb.xml | 4.274584 | | tlg0011.tlg005.perseus-grc2.tb.xml | 4.271146 | | tlg0085.tlg007.perseus-grc1.tb.xml | 4.186867 | | tlg0096.tlg002.opp-grc2.1-53.tb.xml | 4.185597 | | tlg0085.tlg003.perseus-grc2.tb.xml | 4.122508 | | tlg0085.tlg001.perseus-grc2.tb.xml | 4.088124 | | tlg0085.tlg006.perseus-grc2.tb.xml | 4.040698 | |

As you can see, Thucydides ranks 4th in the treebank in terms of dependency distance, but first among the Classical authors. Herodotus narrowly misses the top ten for dependency distance, with Herodotus *Histories* Book 1 coming in at 5.046341.

Table 2: Differences in Dependency Distance withing Thucydides

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Thuc\_book1 | 6.234316 | | Thuc\_plague | 5.705991 | | Thuc\_perikles | 4.657703 | | Thuc\_cleon | 6.371364 | |

From this we may see that dependency distance varies tremendously within Thucydides’ writing. The speech of Perikles has a dependency distance closer to the Euthyphro (tlg0059.tlg001) than to Book 1 of the *Histories*. Similarly, the Plague has a dependency distance closer to Diodorus Siculus than to *Histories* 1, while Cleon’s speech is fairly close to the dependency distance of Book 1.

Table 3: Statistical Information about the Perseus Treebank

|  |  |
| --- | --- |
| Sample Standard Deviation, s | 0.81431349107461 |
| Variance, s2 | 0.66310646174612 |
| Population Standard Deviation, σ | 0.80188049392767 |
| Population Variance σ2 | 0.64301232654169 |
| Mean (Average) | 4.9343363876061 |

This statistical data lets us draw conclusions about Thucydides relative to other authors.

Analysis:

From the data given above, we may conclude that Thucydides’ Greek has a dependency distance which is 1.960439 standard deviations above the mean. In other words, 95% of Greek should have a lower dependency distance, and therefore be easier to read, than Thucydides. However, if we apply this method to the individual passages of Thucydides, then we find that the Last Speech of Perikles has a lower dependency distance than most of the treebank, and should be easier, while the Plague is still more difficult than most Greek, but easier than Book 1.