Feedback Prize – English Language Learning

Robert Fajardo William Locke Nikita Lokhmachev

Brayan Murillo Arefa Patwary Mdakbar Sarkar

ABSTRACT

Automated writing evaluation (AWE) is an invaluable tool for students and teachers to help improve students' writing abilities, especially for English Language Learners (ELL). This project, part of the Feedback Prize – English Language Learning competition hosted by the Learning Agency Lab, Georgia State University, and Vanderbilt University [1], evaluates student-written essays and outputs scores for Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Convention. We extract six independent numerical features from each essay and use them to predict the six dependent variables (scores) listed above. We carry out statistical analyses on both the six independent variables and also the six dependent variables.

1 Overview

The task of the Feedback Prize – English Language Learning competition is to take an essay written by an ELL student and assign it scores in Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Convention. The scores should match those given by human raters reading the same essay.

To accomplish this, we extract from the essays six independent numerical features that we hope will be predictive of the scores those essays receive. These features are:

1. Grammar Errors – the number of grammatical errors found in an essay.
2. Spelling Errors – the number of spelling errors found in an essay.
3. Sum of Errors – the total number of errors found in an essay, including but not limited to both grammar and spelling errors.
4. Vocabulary Size – the total number of unique words in an essay.
5. Unknown Words – the total number of words in an essay that are not recognized from a large English language corpus.
6. Perplexity – a measure of the predictability of a text, based on bigram word probabilities.

More information on these features can be found in Section 4.2 *Independent Variables*. In addition to these, we also have six dependent variables, the Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Convention scores mentioned above. These are the scores we are attempting to predict using our independent variables or features. However, these dependent variables or scores can also be analyzed on their own to better understand their underlying distributions and parameters. We carry out this statistical analysis in Section 4.1 *Dependent Variables*.

Our goal is to use statistical analyses of these six independent and six dependent variables in order to design a more efficient machine learning algorithm that can take the text of an essay and predict its scores in these six different fields.

The remainder of this proposal is structured as follows: Section 2 covers the goals of the project in more detailed and concrete terms; Section 3 covers further specifications of the project; Section 4 covers the exploratory data analysis (EDA) we carried out on the independent and dependent variables; Section 5 covers achieved and expected milestones; and Section 6 covers the responsibilities of different members of the group.

2 Goals

As stated above, the ultimate goal of the project is to use statistical analyses of six independent and six dependent variables in order to design a more efficient machine learning algorithm for predicting scores of essays written by ELL students.

By "more efficient", we are comparing ourselves to other machine learning algorithms already being used in the Kaggle competition to try to solve the problem of AWE [2]. We go into these other algorithms in more detail in Section 3, but for now the important points are that they rely on massive pretrained language models, they do not take advantage of certain statistical relationships between scores and texts, and the features they extract from the text are not human-readable. Our goal therefore is to carry out EDA on these scores and features that will allow us to exploit statistical patterns within the data and outperform those other algorithms with a lower computational cost.

The best score on the competition leaderboard is currently 0.43 MCRMSE (mean columnwise root mean squared error) [3], and it is shared by 683 contestants (these contestants' scores are ordered by precision beyond the hundredths, but that level of detail is not shown on the public leaderboards). Our goal therefore is to improve upon 0.43 and get either a Bronze, Silver, or Gold medal in the competition. Our deadline is November 29, the final submission deadline of the competition. The metric used to determine success, MCRMSE, is the mean of the root mean-

squared error against all six scores [4]. The equation is shown in (1).

$$\text{MCRMSE} = \frac{1}{N_t}\sum_{j=1}^{N_t}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\hat{y}_{ij})^2} \quad (1)$$

3 Specifications

3.1 Tools

The whole project will be implemented in Python. For visualization, we have decided to use the matplotlib and plotly libraries as they are flexible, fast, and easy to use. We will also use Scikit-Learn, catboost, nltk, numpy, pandas, and language_tool_python for EDA, data processing, and regression.

The best solution on Kaggle so far is the solution based on Distributional Neural Networks (DNNs) that uses huggingface implementations of transformer BERT-like models, pytorch to train/fine-tune the models, and RAPIDS SVR to perform multi-label regression. The features extracted by this model are uninterpretable feature vectors extracted by the pretrained deBERTa transformers and containing 600 features each. They are then used as independent features for regression.

3.2 Dataset

The ELLIPSE dataset contains over 7000 essays written by 8th-12th grade ELL students [1]. We are given ~4000 of these to train a model on. Each essay is rated by two human experts according to the six score categories listed in Section 1, with scores ranging from 1.0 to 5.0 in increments of 0.5.

3.3 Implementation

3.3.1 Objectives

Our research has two objectives:

1. Perform EDA and hypothesis testing in order to discover patterns in data that will be insightful for making a reliable machine learning model for score estimation.
2. Use the kaggle solutions and insights in combination with the introduced approach to potentially outperform the best model.

3.3.2 Constraints

The main constraints in this project are:

1. Time required to perform extensive dataset analysis.
2. Lack of budget / computational power needed to fine-tune a large deBERTa model to boost the score.
3. The dataset is fairly small (~4k essays) which might not be enough to make a generalized regression model.

3.3.3 Needs

Given the aforementioned constraints and objectives, we need:

1. A working Python environment.
2. A GPU capable of performing training of deBERTa models in a reasonable amount of time (4-5 hours/experiment max).
3. The tools and libraries mentioned above.
4. 6 laptops/PCs (one computer per team member).

3.3.4 Budget

This project has no funding.

3.3.5 Deadline

The deadline for the competition is November 29th.

3.3.6 Implementation management plan

Kaggle implementation

So far, the best solution presented on kaggle takes the train data and uses multiple pretrained BERT, roBERTa, and deBERTa models to extract text features from the essays. Those extracted essay features are then used to train a multi-label support vector regression algorithm designed to put all the computations on GPU, which makes it very fast and easy to conduct experiments. After training, the whole pipeline is used to generate predictions: extract text features with the models and pass the feature vectors to RAPIDS SVR to output 6 scores.

Our experiments and implementation

The biggest flaw of the kaggle approach is it completely relies on the accuracy of the pretrained deep learning models that have not even seen the essay data. Moreover, that approach does not use any knowledge about score distributions or correlation between text errors and labels. We are planning to bridge that gap by performing extensive data exploration. Since the goal of this research project is mostly EDA and hypothesis testing, we used different text processing and correction libraries to extract text features for analysis. By doing so, our plan is to find interesting data patterns and correlations and extract the most useful independent features that can be used for model training.
Among the experiments that can be conducted are:

1. Incorporating the extracted features into kaggle models.
2. Exploiting score distribution to boost model performance.
3. Using data from external sources for more accurate model training
4. Fine-tuning the transformer models on the given essays and essays from external sources.
5. Training a separate model based on the extracted features.
6. Using regression approaches other than SVR (such as gradient boosting regression)

4 Exploratory Data Analysis

We run EDA on both our independent and dependent variables, because both of these have their own distributions and parameters that can help us better understand the data and build our own machine learning models.

4.1 Dependent Variables

In our dataset, we have six dependent variables, which are the scores for Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Convention. These scores form a distribution over a range from 1 to 5, with a discrete step size of 0.5. Our six histogram plots (**Fig 1**) suggest that all of the variables are similarly and normally distributed: the data is unimodal, roughly symmetric about the mean, and approximately traces out the bell-shaped curve of a normal distribution. The distributions' means are all near 3.0 and their standard deviations are near 0.5 (**Table 1**).

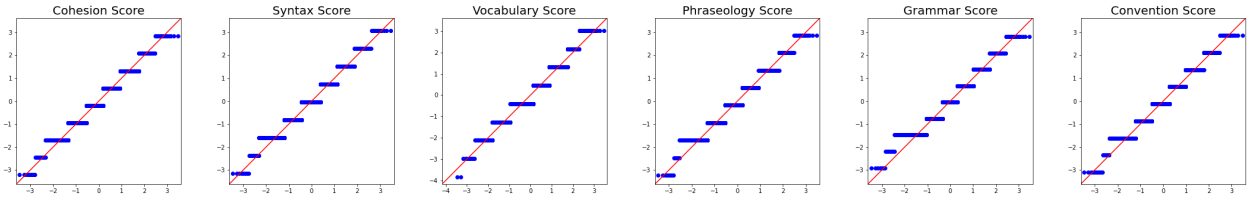
We also have six QQ plots (**Fig 2**) where we can see that all the datapoints mostly follow a straight line with little deviation or outliers. However, when we plug the datapoints into the Ryan-Joiner Normality Test on Statdisk, it rejects the hypothesis that this is a Normal Distribution with both p = 0.05 and p = 0.01. The reason for this is likely due to the fact that the underlying population is discrete, while most normality tests assume continuous populations [5]. However, based on the histograms and QQ plots, we can still argue that the distribution is approximately normal.

As a result of this normality, we can use the empirical rule to estimate that roughly 95% of our scores in all categories should fall between 2 and 4, or within roughly two standard deviations of the mean. Calculating this on our dataset, we find that it holds true for all dependent variables. This also bears on our analysis of the independent variables, as we might expect features in the texts to follow this same normal distribution (though as we will see in section 4.2, this is not necessarily the case).

Another question is the degree of independence between these variables. Some of the scores seem like they would be very closely correlated, e.g. "Grammar" and "Syntax", while others don't have any apparent connection, e.g. "Vocabulary" and "Cohesion." Looking at the scatter plot in **Appendix 1** of dependent vs dependent variables, we see that there is actually a high degree of covariance between all the dependent variables, with a positive trend between them, meaning that as one score goes up, the others tend to go up along with it. Though it is not apparent from the categories themselves that this would be the case, it does match the intuition that an essay which scores well in one domain is likely to score well in others. It also raises the possibility of using some scores to impute the values of others.

Fig 1: Dependent Variable Histograms

Fig 2: Dependent Variable QQ Plots



	Features	$\bar{x}$	s^2	s	μ	σ^2	σ
0	Cohesion Score	3.1	0.4	0.7	(3.1, 3.1)	(0.4, 0.5)	(0.6, 0.7)
1	Syntax Score	3.0	0.4	0.6	(3.0, 3.0)	(0.4, 0.4)	(0.6, 0.7)
2	Vocabulary Score	3.2	0.3	0.6	(3.2, 3.3)	(0.3, 0.4)	(0.6, 0.6)
3	Phraseology Score	3.1	0.4	0.7	(3.1, 3.1)	(0.4, 0.5)	(0.6, 0.7)
4	Grammar Score	3.0	0.5	0.7	(3.0, 3.1)	(0.5, 0.5)	(0.7, 0.7)
5	Convention Score	3.1	0.5	0.7	(3.1, 3.1)	(0.4, 0.5)	(0.7, 0.7)

Table 1: Dependent Variable Statistics and Parameters
Parameter Confidence Intervals calculated with Confidence Level 95%

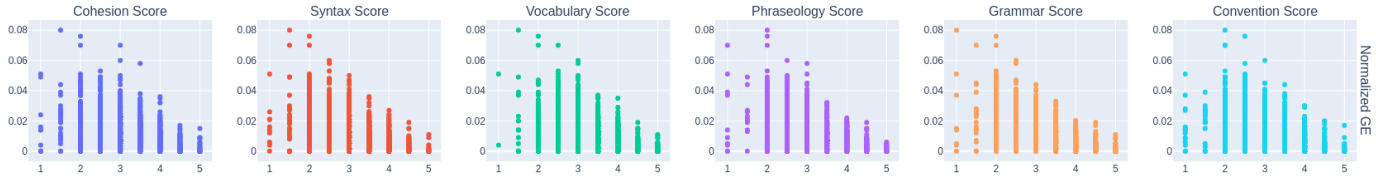
4.2 Independent Variables

The independent variables we extracted from the text in order to analyze and predict essay scores are: Grammar Errors, Spelling Errors, Sum of Errors, Vocabulary Size, Unknown Words, and Perplexity. In each of these variables, there is some possibility of error or noise due to how they were extracted. For example, the first three variables, Grammar Errors, Spelling Errors, and Sum of Errors, were all extracted using an open-source tool, Language Tool Python [6]. While we manually checked through and confirmed some of the errors this tool identified, there is always the possibility of False Positives or False Negatives mixed in with the results. The last three variables, Vocabulary Size, Unknown Words, and Perplexity, we calculated ourselves, but here too there are sources of possible error or noise due to details of our implementation and ambiguities of the text. Regardless, we hope that these variables are informative enough that we can find useful statistical links between them and our dependent variables.

Fig 3a: Grammar Errors vs Scores

Note that scores (dependent variables) are on the x-axis, while grammar errors (independent variables) are on the y-axis.

Fig 3b: Grammar Errors vs Dependent Variables Normalized by Essay Length

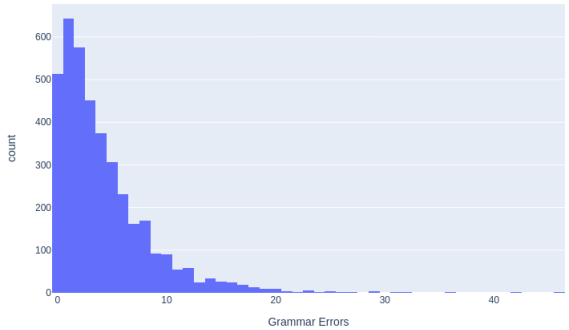


4.2.1 Grammar Errors

Scatter Plots

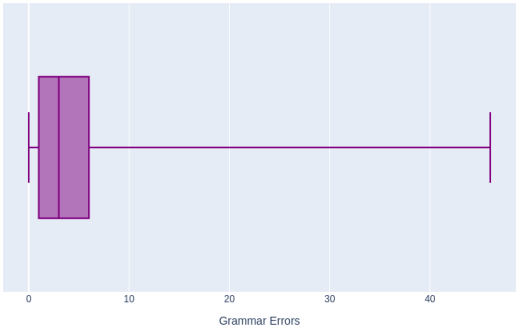
The scatter plots of **Fig 3a** show the independent variable "Grammar Errors" (GE) plotted against each of the six dependent variables. At first impression one could say that the GE variable affects all of these scores in a similar way, since the dispersion of the points tends to be similar, in that those essays that have high scores tend to have few grammar errors; however, essays with low scores also tend to have few grammar errors. Personally, I expected the plots to be a straight line with negative slope (and with some noise), which does not show up very well in the results. One possible reason is that the probability of making grammatical errors increases with the length of the text, i.e. Grammar Errors and Essay Length are positively correlated. This can be seen in the scatter plot of independent vs independent variables in **Appendix 1**, specifically in the graph showing Grammar Errors vs Length. In this case, a very short text with few or no errors should still have a low score because your writing is not long enough to conclude that you have mastered the topic. Knowing this, if we normalize GE by dividing by the length of the essay, we can see in the normalized graph (**Fig 3b)** a slightly clearer negative trend between GE and the essay scores. However, the fact that there are essays with 0 GE that still score 3 or 4 on the dependent variables shows that GE must be combined with other independent variables to determine scores.

Histogram



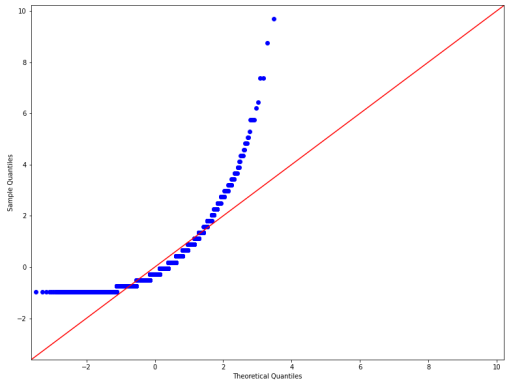
The histogram shows a right-skewed distribution, with a maximum peak at 1 GE which may indicate that most of the students in the sample tend to have few grammatical errors (below 6 or so). It is also possible to observe some gaps in the histogram which may refer to outliers in the sample.

Box plot:



The plot confirms the right-skewed distribution of the sample, it shows that 50% of the samples are between 1 and 6 GE with median of 3 GE. It is also noticeable how the presence of outlier data extends the right whisker of the plot.

QQ plot:



The QQ plot indicates that the sample is not following a normal distribution, as the points do not follow the straight red line.

Parameter Estimation:

The first row of **Table 2** lists the mean, variance, and standard deviation of GE. The fact that the mean is non-resistant makes this value differ considerably from the median, and the value of the variance and standard deviation may not be very reliable if outliers are not removed. Confidence intervals are given for a 95% confidence level to determine the mean, variance, and standard deviation parameters of the population. However, because the sample does not follow a normal distribution, it is recommended not to use the intervals given for the variance and standard deviation.

	Features	$\bar{x}$	s^2	s	μ	σ^2	σ
0	Grammar Errors	4.2	18.6	4.3	(4.0, 4.3)	(17.8, 19.5)	(4.2, 4.4)
1	Misspelling Errors	9.9	154.1	12.4	(9.5, 10.3)	(147.4, 161.1)	(12.1, 12.7)
2	Sum of Errors	21.9	325.8	18.1	(21.3, 22.4)	(311.8, 340.8)	(17.7, 18.5)
3	Vocabulary Size	150.8	2510.0	50.1	(149.3, 152.4)	(2402.3, 2625.1)	(49.0, 51.2)
4	Unknown	11.4	139.7	11.8	(11.1, 11.8)	(133.7, 146.2)	(11.6, 12.1)
5	Perplexity	129.1	544.1	23.3	(128.4, 129.9)	(520.7, 569.0)	(22.8, 23.9)

Table 2: Independent Variable Statistics and Parameters

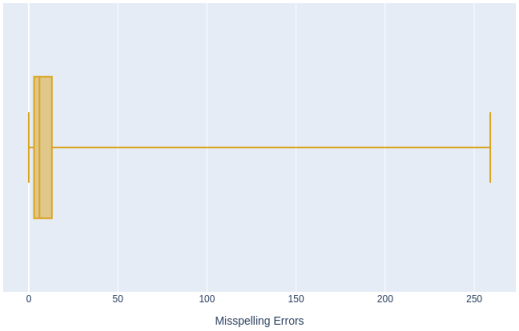
4.2.2 Spelling Errors

Spelling errors is one of independent features that affects the overall score of our analysis in Automated Writing Evaluation.

Scatter Plots

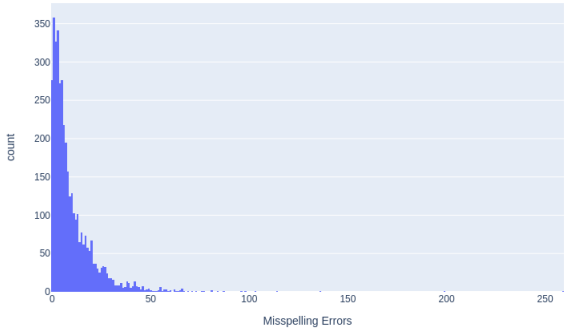
We generate scatter plots (**Fig 4**) plotting the independent variables against the dependent variables. We can visualize the impact of misspellings on overall score with the scatter plots. The plots have different characteristics for different score types. But the plots do not recommend going with a specific pattern, rather in most cases it is rightly skewed. It is also not evident from the plot that students get low scores with high spelling errors. This could be an issue of the combined effects of other independent features or an error with the feature extraction from the dataset. However, we do see higher scores with lower spelling errors, which is a reasonable pattern.

Box Plot



The box plot reflects the how the data is distributed throughout the samples. Spelling errors in the range of 0-13 make up 75% of the data. The IQR (13-3=10) claims that data points beyond 1.5 x IQR = 15 are significantly high and could be outliers in our case.

Histogram



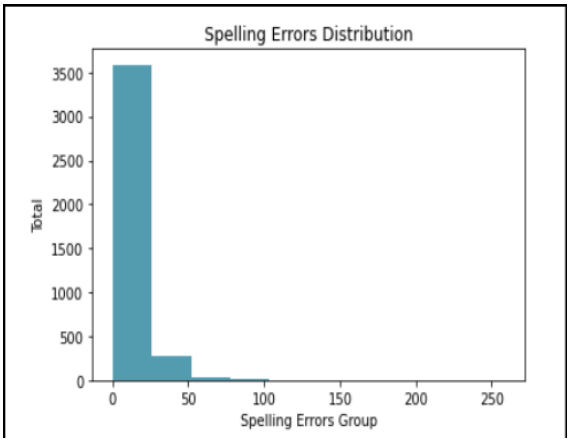
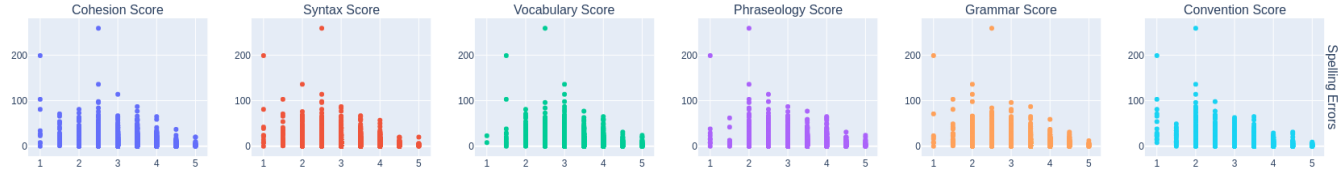
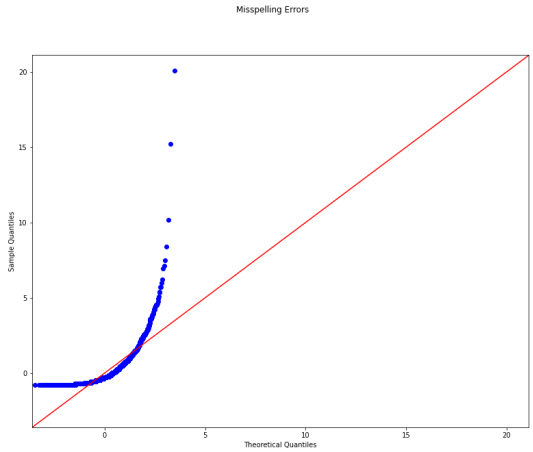


Fig 4. Spelling Errors vs Scores



Both histograms and the QQ plot below do not go with a normal distribution. The histogram follows a right skewed pattern and is unimodal. Also, the QQ plot reflects the same scenario where most data points do not fall on a straight line or the pattern differs from a straight line.



Normality Assessment

Our observation of the data distribution shows us it is not normal, and the fact that students with a low score also have low spelling errors makes the data questionable. Histogram and QQ plots reject the normality of the data on spelling errors.

Parameter Estimation

We estimate the parameters based on the assumption that our sample collection is more than 30 and randomly selected. We measure the confidence intervals at 95% confidence for the mean, variance, and standard deviation. Since the population standard deviation is unknown, we use the t distribution. From **Table 1** we get a sample mean of 9.9 and sample standard deviation of 12.4. With 95% confidence, we put the population mean between 9.5 and 10.3, and population standard deviation between 12.1 and 12.7.

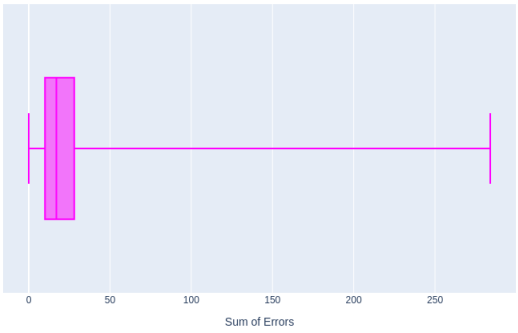
4.2.3 Sum of Errors

Sum of Errors is the total number of errors of all types in an essay. This includes grammar and spelling errors, so it is obviously not independent of those earlier two features; however, it also accounts for other errors not included in this project, so it is not fully determined by the previous two features. Because of this, it might be a good candidate for missing value imputation as a possible extension of this project.

Scatter Plots

In the scatter plots in **Fig 4**, we can see the relationship between Sum of Errors and the essay scores across different criteria. Generally, the graphs are all skewed to the right, which makes sense as the higher the score is, the fewer errors there are supposed to be. However, it is still surprising that the essays with most of the errors generally have scores between 2 and 3 and not 1. This may be due to software-based error detection inaccuracies. For the same reason, even the essays that have a score of 5 still have some errors. Another reason for that might be that some people get 5 for one grading criterion and 3 or 4 for another.

Box Plot



The box plot suggests that the minimum number of errors in an essay is 0 and in most cases, the number of errors is not large (median is ~15 errors). Q1 (25th percentile) is ~10 errors, Q3 (75th percentile) is ~30 errors, whereas the max number of errors is ~280 (outliers are not shown).

Histogram and Normality Assessment

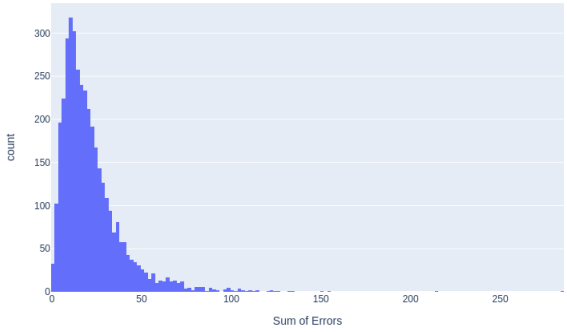
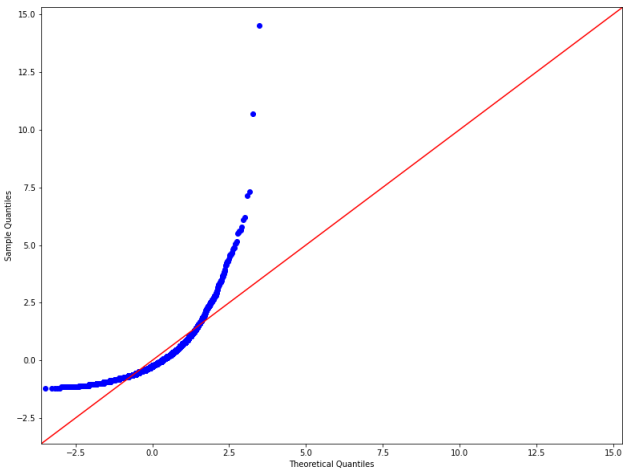


Fig 4: *Sum of Errors vs Scores*



In order to perform a normality assessment, we can either plot a histogram of the variable or plot a QQ graph. The histogram clearly does not look symmetrical and is skewed to the right, which suggests that the distribution is not normal. It can also be seen from the histogram that most of the writers make from 0 to 100 errors in total in their essays. The numbers are not necessarily accurate as the errors were discovered by software and not by a group of humans. For that reason, the graph shows that some people make up to ~350 errors in their essays, which cannot be true given the average essay length and the number of people with decent essay scores. Nevertheless, the trend should still be similar.

From the QQ-plot, we can see that the graph is not following the line pattern which also suggests that the variable is not normally distributed.

Parameter Estimation

We assumed the dataset is a simple random sample. In the parameter estimation table (**Table 1**), we have estimated the grammar score to have the sample mean (= point estimate of the population mean) of 21.9 and the sample standard deviation of 18.1. Moreover, we estimated the population mean with the population standard deviation unknown and the confidence interval of 95% to be between 21.3 and 22.4 by using the t-Distribution. We also estimated the population standard deviation by employing the Chi-Squared Distribution. It is between 17.7 and 18.5. Based on the population standard deviation estimation, it is not difficult to estimate the population variance: $311.8 < \sigma^2 < 340.8$. Unsurprisingly, the Sum of Errors variable has a fairly large mean and std (compared to the other features) as it incorporates all the errors in the essays.

4.3.4 Vocabulary Size

This is the number of unique words in an essay. We can measure it as an absolute number or as a percent of the total terms used.

Scatter Plots

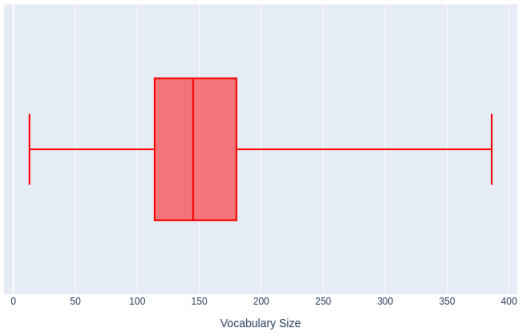
The scatter plots of **Fig 5a** show the relationship between the independent variable "vocabulary size" and each of the dependent variables. In the scatter plots in **Fig 5a**, we can see the relationship between the vocabulary size and the essay scores across different criteria. The graphs show a gradual positive correlation between the independent variable and each of the dependent variables. But for each criteria, we can see that, most of the scores of these dependent variables falls within scores of 3 to 4 and this is regardless of the vocabulary size. This means, both an essay of vocabulary size 10 and size 300 has the same score. On the other hand, we can also see that an essay with a vocabulary size of 300 or 100 can have a score anywhere from 1 to 5. This implies that vocabulary size and scores have no direct relationship, and the majority of the essay scores for each category falls into scores 3 to 4 regardless of vocabulary size. This is further clarified if we look into the plot normalized by the essay length. In the normalized plots in **Fig 5b** we can see that the vocabulary size no longer has any direct relationship with the dependent variable scores. This is because the score was actually changing because the essay length was increasing, hence there were more unknown words, more spelling mistakes, more grammatical errors and so on, not directly because of the vocabulary size.

Fig 5: Vocabulary Size vs Scores

Fig 5b: Vocabulary Size vs Scores Normalized by Essay Length

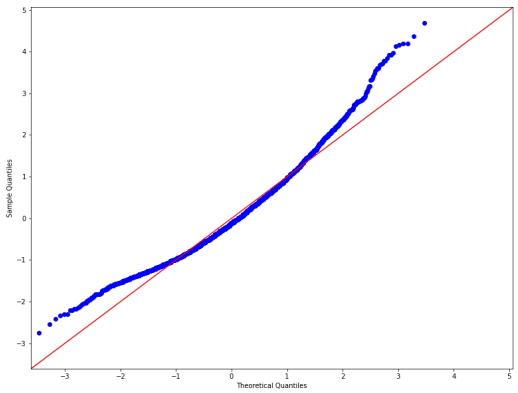
If we look at the inter-correlations of all the independent variables, we see even more confounding factors. The scatter plots of Independent vs Independent Variables in **Appendix 1** show the relationship between the independent variable "vocabulary size" and each of the other independent variables. For example, looking at Vocabulary Size vs Grammar Error, we can see that, as the vocabulary size increases the grammatical errors also increases, with a few outliers in the middle. This would make sense as, if a sentence has too many unique words or almost all the words are unique, it would not grammatically make sense. Now, for Vocabulary Size vs Spelling Errors, we can see that as the vocabulary size increases the spelling errors slightly increase, but the plot shows an almost normal distribution. This makes sense in a way that, as the vocab size increases, the tendency of spelling mistakes also increases but not as high as the grammatical errors. Again, for the Vocabulary Size vs Unknown Words plot, we can see that it shows a very similar plot as the Vocabulary Size vs Spelling Errors, which means as the number of vocabulary increases the number of unknown words would also increase. But the strongest correlation that the vocabulary size has is with the length of the essay, which is a strong positive correlation. This makes sense because, as we saw from the plots above, the vocabulary size and the scores were correlated with essay length but had little correlation among themselves.

Box Plot



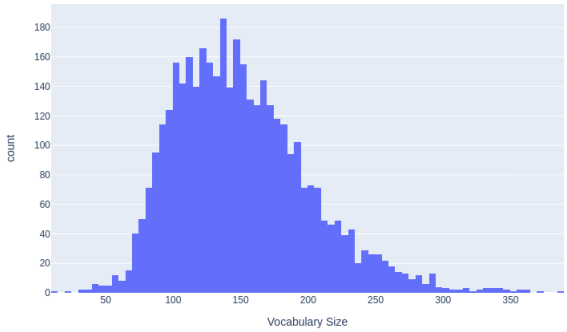
From the box plot we can see that the minimum value of vocabulary size in essays lies around around 15. The mean of the vocabulary size is around 145 and the highest value lies around 375. From the plot we can also infer that, the 25th percentile of the unique words are somewhere around 115 and the 75th percentile is around 180.

QQ Plot



From the Q-Q plot, we can see that it shows a systematic deviation from the normal distribution. But since almost all the data falls around the straight line or slightly above it, we could conclude that the vocabulary size distribution is approximately normal. The line passes through the point defined by the lower quartiles and the point defined by the upper quartiles.

Histogram



From the histogram plot, we can see that the plot is slightly skewed to the right. Here the mode is around the vocabulary size of 140 and the mean value is somewhere on the right of the mode. From the boxplot, we know that the mean is around 145 which further supports the results derived from the histogram. Again, from the Q-Q plot we can see that the data are slightly deviated from the normal distribution but do somewhat follow it. In fact, from the histogram we can also see that the right skewness is very slight and it almost looks like a normal distribution as well. Finally, by comparing all 3 individual plots, we can conclude that the data was not entirely normally distributed and was slightly skewed to the right.

Parameter Estimation

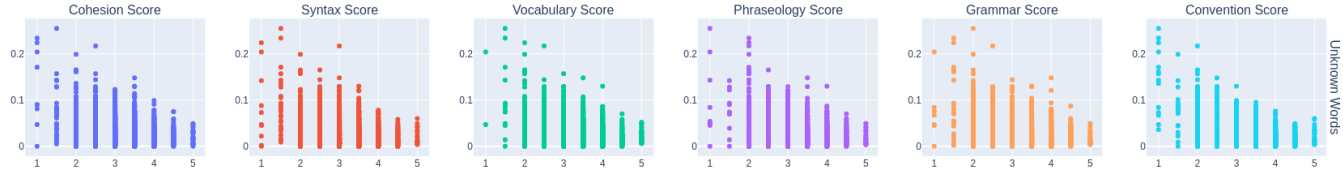
Looking at **Table 3**, we see the sample mean and standard deviation are 150.8 and 50.1 respectively; this agrees with our earlier plots. Given the sample size and relative normality of the data, we can estimate the population mean and standard deviation to be between 149.3 to 152.4 and 49.0 to 51.2 respectively, with a confidence level of 95%.

4.3.5 Unknown Words

To calculate the number of unknown words in an essay, we must first create a set of "known" words to compare it to. We do this by collecting a vocabulary of the 27,000 most common words from the Brown Corpus, an English language corpus of over 1 million words [5]. Any word appearing in an essay that is not in this list is considered "unknown" and assumed to be either misspelled or otherwise non-standard. In most cases this will be true, since the most common 27,000 words account for over 98% of the over 1 million word occurrences in the Brown Corpus. However, there are exceptions, such as the word "online," which, due to the time period in which the Brown Corpus was compiled, is not present in the corpus and therefore considered "unknown" even though it is a very common word today. Other similar examples may exist, creating noise in this variable.

Fig 6a: Unknown Words vs Scores

Fig 6b: Unknown Words vs Scores with Outlier Removed and Normalized by Essay Length

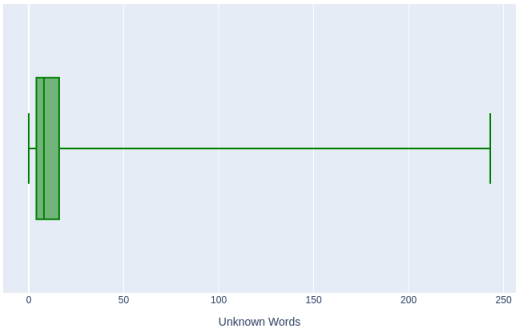


Scatter Plots

As seen with many other independent variables, the scatter plot of Unknown Words against the different scores in **Fig 6a** does not show any pronounced trends. The expectation is that a higher number of unknown words would correlate to a lower score, at least in word-centric categories such as Vocabulary, but this does not seem to be borne out by the data. Essays with very few unknown words appear at all score levels for all dependent variables, and the essays with the most unknown words seem to mostly receive scores between 2 to 4. Going back to our analysis of Dependent Variables, this could be because this is where the majority of essays are in the dataset, and so we should expect to see a larger range of unknown words appearing there. However, it still seems counter-intuitive that the essays with the highest numbers of unknown words wouldn't receive scores of 1.

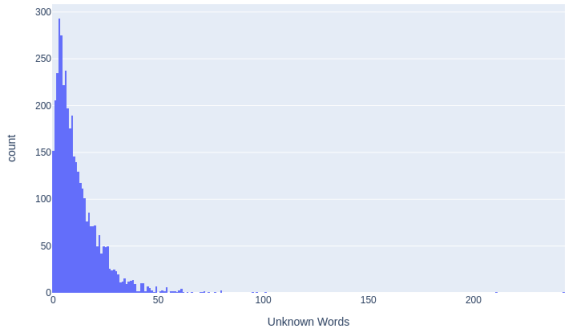
There are two factors that might be making our scatter plots less informative. First, there are two prominent outliers with over 200 unknown words, twice as many as the next highest value. Interestingly, these two outliers appear to correspond with two outliers in the Spelling Errors and Sum of Errors scatter plots, which supports our assumption that most unknown words are in fact misspellings. Another factor is that the number of Unknown Words, like Grammar Errors and Vocabulary Size, might be correlated with the length of the essay, so longer essays will have more unknown words. To correct for both of these, we can remove the two outlier cases and normalize the number of unknown words by the length of the essay, thus producing the scatter plots in **Fig 6b**. These scatter plots don't show a deterministic relationship between the number of unknown words and an essay's score, e.g. there are still essays with few or no unknown words at all score-levels, but it does show a much more pronounced downward trend in the maximum number of unknown words an essay might have and its score in any domain.

Box Plot



The box plot shows very clearly the presence of the outliers relative to the rest of the distribution. If we had used the 1.5 x IQR rule for identifying outliers, the two highest values over 200 would be identified with stars, and the highest whisker would extend to just over 100. As it is, you can still see clearly that the vast majority of cases sit very low, with three-fourths of the data points having fewer than 16 unknown words and half having fewer than 8. This likely contributes to the low-informativeness of the initial scatter plot—the majority of essays at all score-levels simply don't have that many unknown words.

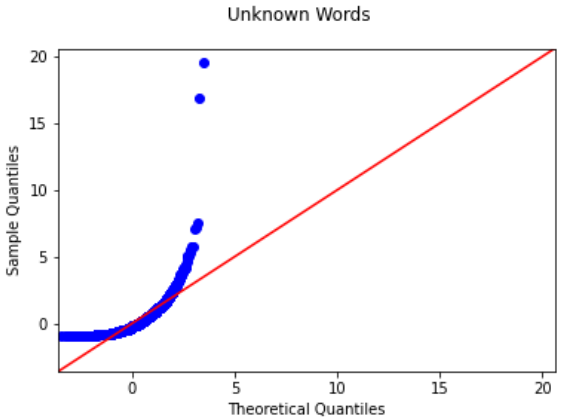
Histogram



Looking at the histogram, we see again the information given by the box plot, which is that the vast majority of essays have a

small (<50) number of unknown words, with a very thin tail creeping out to 100 unknown words in an essay and then the two outliers (here not really visible) above 200. The distribution does not look Normal, skewing heavily right, but it might be approximated by a Poisson distribution with a low value for λ. This would make sense for the variable, if we interpret the distribution as the number of unknown words encountered within a set length of an essay.

QQ Plot and Normality Assessment



As already stated in the analysis of the histogram, the Unknown Words variable does not seem to follow the Normal distribution, and this QQ plot reaffirms that, as the data points clearly curve upwards from the line even without taking into account the two outliers.

Parameter Estimation

The sample mean and standard deviation are given in **Table 2** as 11.4 and 11.8 respectively. The mean, not being resistant to outliers, is certainly pulled up by the two extreme outliers in the data, though as we saw the median (which is more resistant) was not too much lower at 8 unknown words. 11.4 would also work as a point estimate of the population mean, given the number of samples it is based on is well over 30. However, since the sample does not appear to be Normal and there is no reason to assume the underlying population is Normally distributed (indeed, it seems more likely to follow a Poisson distribution), we can't apply normal estimation techniques to the standard deviation. We did calculate a 95% confidence interval for the population mean, which is between 11.1 and 11.8.

4.3.6 Perplexity

We measure the perplexity of an essay to calculate its predictability based on bigram word probabilities. Another way to view the perplexity is that it represents a "naturalness" score, and the lower the score, the more natural the essay should sound. We hope this will capture syntactic agreement (e.g. "She walks" should have lower perplexity than "She walk"). When implemented with subword units, it measures correct spelling too (e.g. "library" should have lower perplexity than incorrectly spelled "liberary").

This is a sample implementation of Perplexity using Bigram Models. We first split up a sentence as follows:

Sentence: "My name is Robert"
Split into bigrams:
Bigrams: (My, name) , (name, is) , (is, Robert)

We then calculate the probability of this sentence by calculating the conditional probability of each bigram:

*P(sentence) = P(name | My) * P(is | name) * P(Robert | is)*

Where P (name | My) means the probability of the word "name" following the word "My". We assign the probabilities according to their occurrence in a large corpus of English language text, in this case the Brown Corpus from Brown University.

We then calculate the Perplexity of the essay, using the nth root of the inverse of the probability, where n is the number of words in the essay. The higher the probability of the sentence, the lower the perplexity, and the more natural the sentence sounds (at least in theory).

Equation: PP is the Perplexity and P is the Probability

$$PP(w_1...w_N) = P(w_1...w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1...w_N)}}$$

In the case where we have words that don't exist in our training data, we replace it with an "UNK" token to capture the word that is not recognized in the student essays (this could include rare English words, but also misspellings and incorrect word-forms). The "UNK" token actually introduces some problems into the calculation of Perplexity, because if the "UNK" token is assigned too much probability in the bigram language model, essays with more "UNK" tokens will receive lower perplexity values than those with fewer "UNK" tokens. In the extreme case, essays composed almost exclusively of "UNK" tokens could receive much lower perplexity values than well-written English essays. Thus we will need to assess the value of perplexity and determine the score based on the number "UNK" tokens in the sentences.

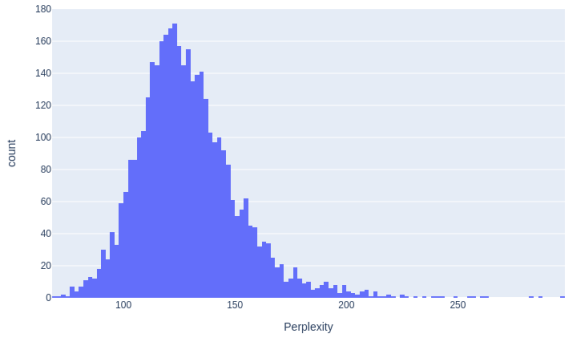
Fortunately, as we saw in the previous section, "UNK" tokens are relatively rare in the dataset. However, we can still work to decrease this noise in the variable by pushing down the probability of the "UNK" token in our bigram language model or by incorporating subword units into Perplexity and removing the "UNK" token altogether.

Scatter Plots

As seen in **Fig 7**, the scores range from 1 (low) to 5 (high) and bins of 0.5 in width. Prior to making predictions, we need to analyze the sentences and their current score using the 6 criteria. We observe perplexity by assigning each sentence to their graded score bin. Scores of 1 to 3 generally have a higher maximum perplexity score than the criteria scores at 3.5 and higher. Based on these scores and perplexity values, lower perplexity correlates with higher essay scores.

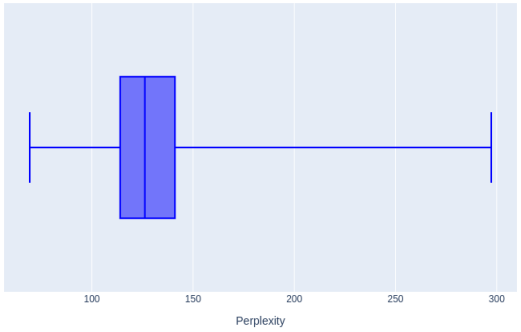
Fig 7: Perplexity vs Scores

Histogram



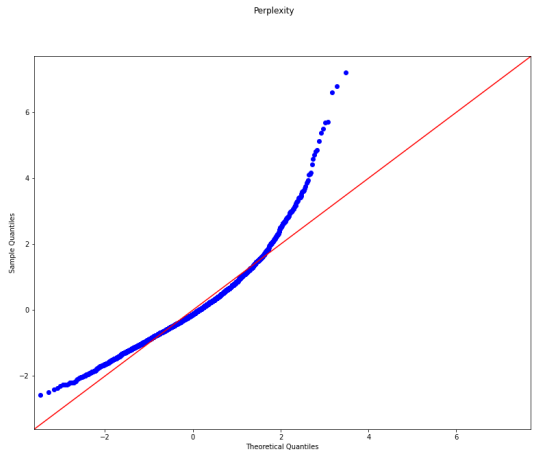
This histogram includes the count of essays with a given perplexity score. Most of our essays will lie within the range of 110-145 and the distribution of perplexity scores skews to right.

Box Plot



This box plot confirms the skewness to the right whereby a major of the perplexity scores lie within the 110-145 range. Most of our outliers are above 145 showing a skewness to the right.

QQ Plot



The QQ plot shows that the observations are sloping exponentially in a positive direction away from the normal line plot. This proves that the scores for perplexity are not normally distributed.

Parameter Estimation

Table 2 shows that the perplexity values have a mean of 129.1 and a standard deviation of 23.3. We calculated the population mean to be between 128.4 and 129.9, and the population standard deviation to be between 22.8 and 23.9, both with 95% confidence.

4.3.7 Length

Though it was not analyzed as an independent variable, we did consider length as a confounding factor in our features. Interestingly, as can be seen in **Appendix 1** it does seem to be positively correlated with essay scores, moreso even than many of the independent variables we examined more closely. For several of the independent variables, we normalized their values by dividing by the length of the essay in order to remove its confounding influence. Sometimes, as in Grammar Score and Unknown Words, this produced a more apparent relationship between the independent variable and dependent variables. Other times, as in Vocabulary Size, this weakened or removed an apparent relationship between the two. Length is also correlated directly with several of the independent variables, most strongly with Vocabular Size, as can again be seen in **Appendix 1**. Moving forward, we will need to decide on the best ways to either minimize or take advantage of this confounding factor.

5 Milestones

Our first two milestones have already been reached upon the completion of this proposal. These are:

1. Extract six numerical independent variables from the texts of the essays.
2. Perform EDA on both these independent variables and the dependent variables of the essay scores.

Our remaining milestones are:

Project Update 1 (no date): expand EDA from single variable to multivariable analysis with three independent variables.

Project Update 2 (no date): conduct hypothesis testing to see what statistical effects can be shown to hold between the independent and dependent variables.

Competition Submission (Nov 29): finish training and submit the results of a machine learning model incorporating the statistical information gained above.

Project Presentation (no date): present our project to the class.

6 Group Responsibilities

All six members of our group analyzed a single independent variable: Brayan Murillo analyzed Grammar Errors, Mdakbar Sarkar analyzed Spelling Errors, Nikita Lokhmachev analyzed Sum of Errors, Arefa Patwary analyzed Vocabulary Size, William Locke analyzed Unknown Words, and Robert Fajardo analyzed Perplexity.

In addition, William Locke wrote the paper, Nikita Lokhmachev visualized the data, Brayan Gutierrez calculated the parameters. Arefa, William, and Nikita all worked together to extract features from the text. Mdakbar analyzed the Dependent Variables.

REFERENCES

[1] Learning Agency Lab. 2022. Feedback Prize Competition Series. Retrieved Nov 4, 2022 from https://www.the-learning-agency-lab.com/the-feedback-prize-overview/

[2] Kaggle. 2022. Feedback Prize – English Language Learning. *Code*. Retrieved Nov 4, 2022 from https://www.kaggle.com/competitions/feedback-prize-english-language-learning/code

[3] Kaggle. 2022. Feedback Prize – English Language Learning. *Leaderboard*. Retrieved Nov 4, 2022 from https://www.kaggle.com/competitions/feedback-prize-english-language-learning/leaderboard

[4] Kaggle. 2022. Feedback Prize – English Language Learning. *Evaluation*. Retrieved Nov 4, 2022 from https://www.kaggle.com/competitions/feedback-prize-english-language-learning/overview/evaluation

[5] Minitab Blog Editor. July 1, 2011. Assumptions and Normality. *Minitab Blog*. Retrieved Nov. 6 from https://blog.minitab.com/en/quality-data-analysis-and-statistics/assumptions-and-normality

[6] jxmorris12. Jan 12, 2022. LanguageToolPython 2.7.0. Retrieved Nov 6, 2022 at https://github.com/jxmorris12/language_tool_python

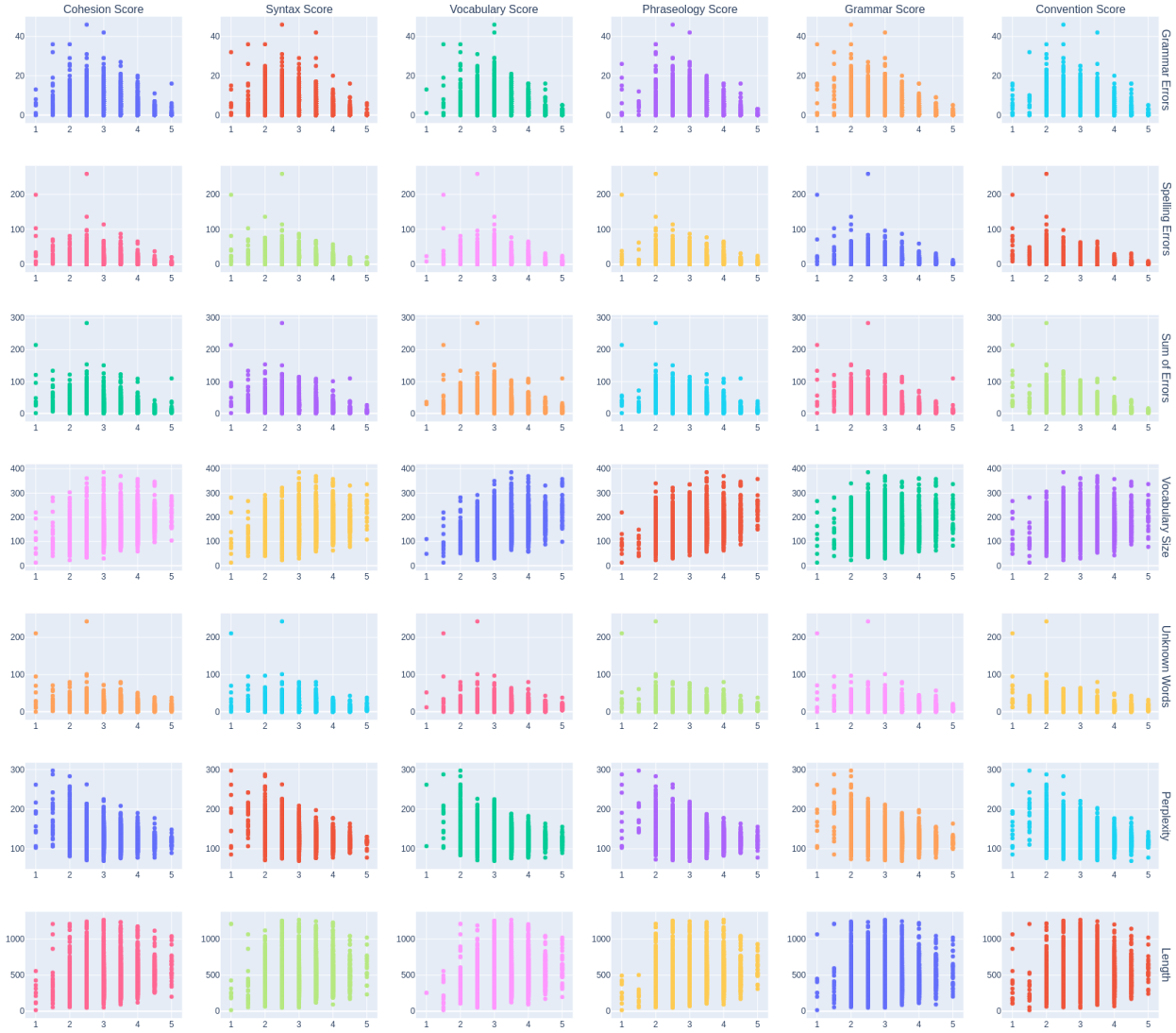
[7] Francis, W. Nelson & Henry Kucera. 1967. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press.

[8] Martin, S., J. Liermann, and H. Ney. 1998. "Algorithms for Bigram and Trigram Word Clustering." Speech Communication 24 (1): 19–37. Retrieved from https://search-ebscohost-com.libproxy.library.unt.edu/login.aspx?direct=true&db=inh&AN=6006417&scope=site

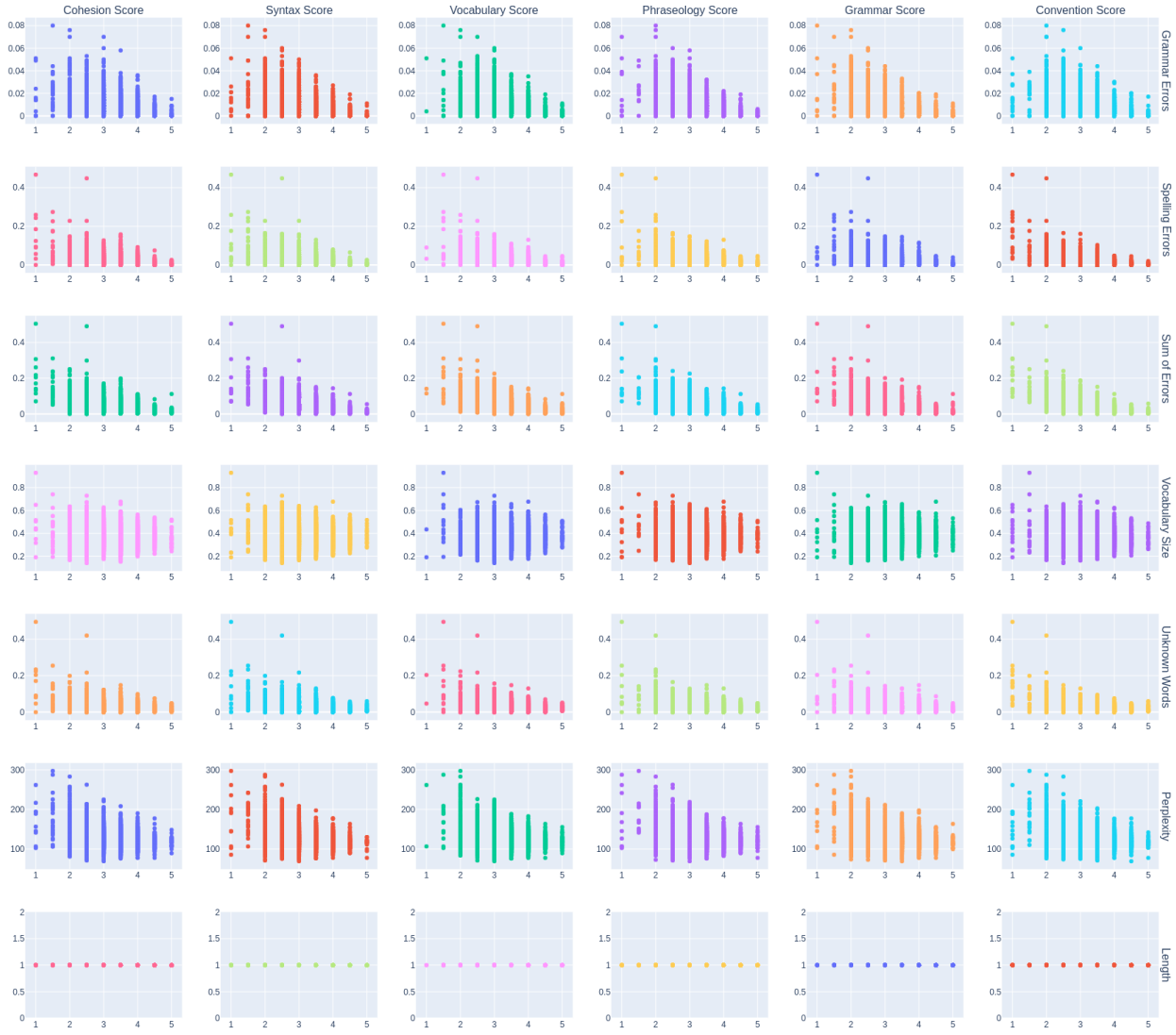
[9] Hockenmaier, Julia. "Lecture 3: Language Models." *Natural Language Processing*, n.d., 50. Retrieved from https://courses.engr.illinois.edu/cs447/fa2018/Slides/Lecture03.pdf

Appendix 1

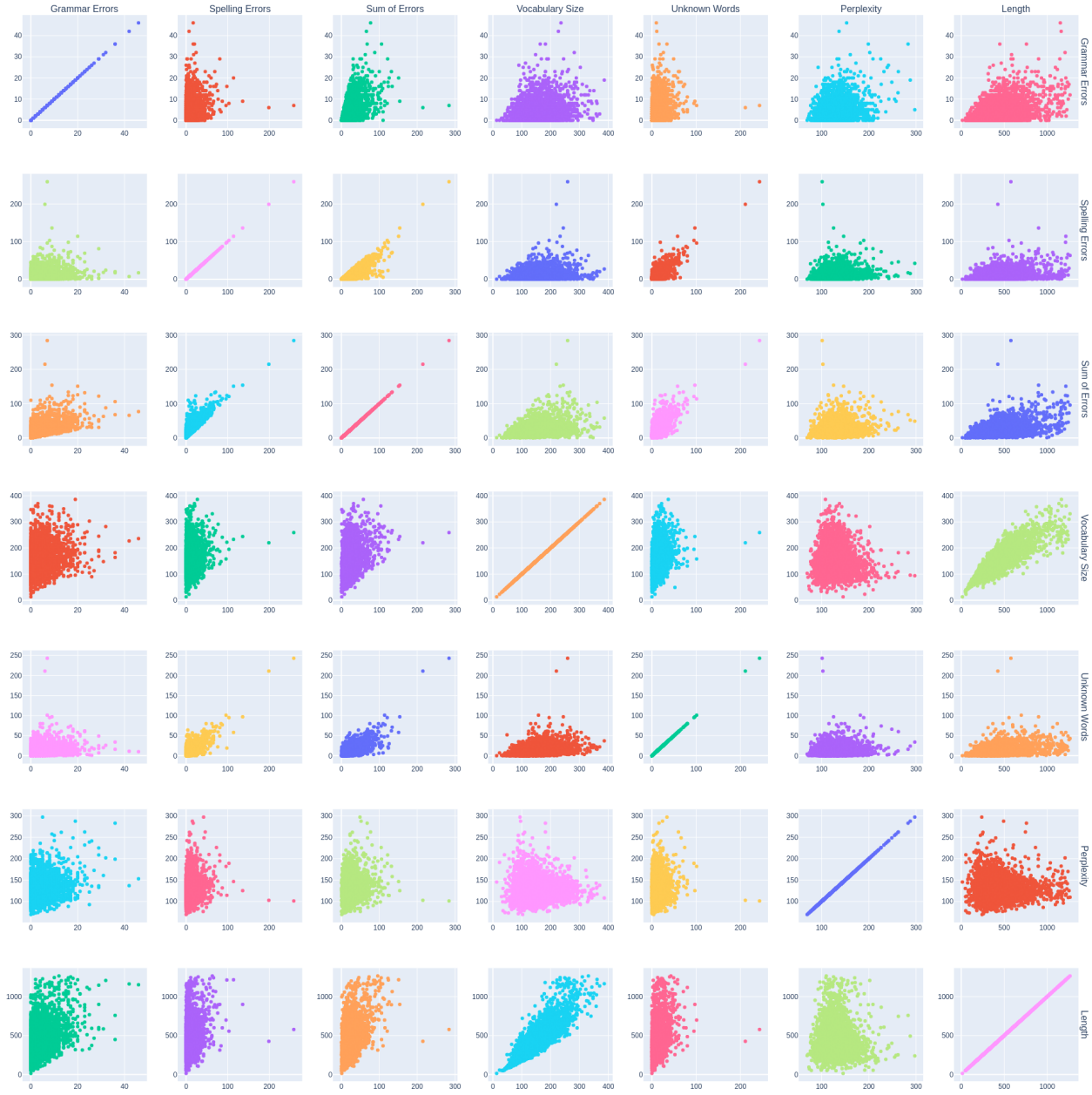
All Independent vs All Dependent Variables



All Independent vs All Dependent Variables Normalized by Length of Essay



All Independent vs All Independent Variables



All Dependent vs All Dependent Variables