Exploration of COREFL: Integrating Linguistic Analysis, Statistical Modeling, and Streamlit Application Development

Introduction

Our project explores the Corpus of English as a Foreign Language (COREFL) to identify statistical relationships between various variables such as gender and first language. The primary goal is to integrate linguistic and statistical analyses to create a statistical model that predicts individuals' English proficiency based solely on their utterances. This approach is driven by the hypothesis that significant differences in speech patterns exist between speakers at different proficiency levels.

Additionally, we aim to enhance the corpus's accessibility for non-technical researchers by developing a Streamlit application. This application will facilitate the exploration of this second language corpus, contributing significantly to both academic and educational fields.

For statistical analysis, we have chosen to focus exclusively on the Charlie Chaplin description task, as it provides a uniform input for each subject. This uniformity enables a focused analysis of linguistic differences between subjects, thereby minimizing outcome variance due to task-related factors.

The project is divided into three parts: The first involves a statistical exploration of the corpus, reporting both significant and non-significant results, and investigating relationships between various variables. The second part is dedicated to constructing a multilinear regression model to predict English test scores using only textual information, exploring which textual cues are relevant for predicting subjects' English proficiency. The final part entails developing a Streamlit application to enhance the corpus's accessibility, enabling exploration without requiring coding skills.

Mikhail

Influence of Proficiency Level, Gender, and Native Language on Third-Person Verb Agreement among English Learners

For L2 learners of English language the ability to correctly conjugate verbs, particularly in the third-person singular form, is a crucial skill that might present a challenge. This segment focuses on how different factors such as English proficiency level, gender, and native language (L1) influence learners' ability to conjugate third-person singular verbs accurately.

The dataset for this research included learners across various English proficiency levels, ranging from A1 to C2. It featured a diverse group of participants in terms of gender, with a significant number of native German and Spanish speakers. In this analysis we employed the Spacy library for defining the following functions: is_verb_correct_for_third_person_singular and analyze_third_person_singular_verbs. The first function determines the correctness of verb conjugations, and the second function identifies and evaluates verb-subject pairs in the learners' responses.

To understand the patterns and correlations in verb conjugation accuracy, Chi-squared tests and Simple Linear Regression were utilized. The Chi-squared tests examined the error rates in verb conjugations across different proficiency levels, genders, and native languages. In contrast, the Simple Linear Regression analysis explored the relationship between the learners' score percentages and their correct total ratios of verb conjugation.

The findings revealed a statistically significant influence of proficiency level on verb conjugation accuracy. Notably, error rates showed a marked decrease at higher proficiency levels, specifically at the C1 level, indicating that higher proficiency correlates with better conjugation accuracy. This was statistically significant, with a Chi-squared value of 813.933681 and a p-value $p < 0.0001$.

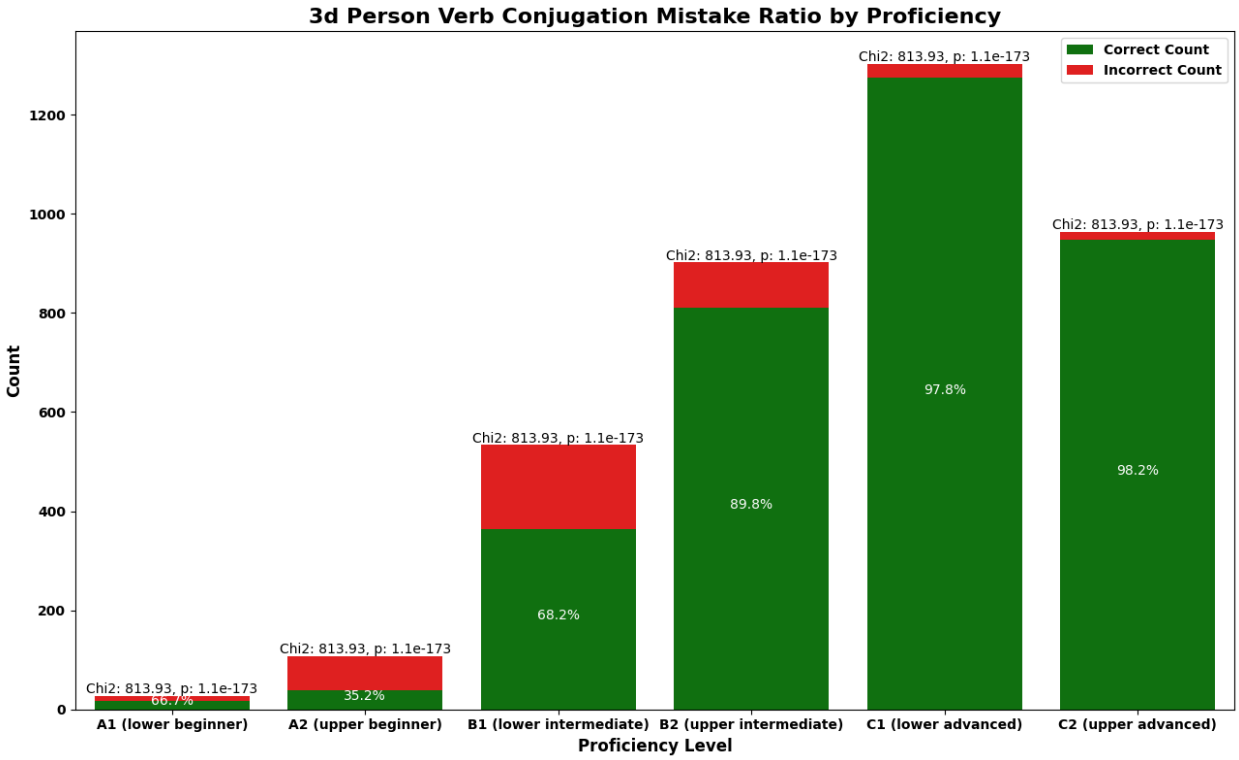


Fig.1 A bar chart showing the influence of Proficiency Level on Third-Person Verb Agreement among English Learners.

Gender differences in verb conjugation accuracy were also observed. At the A2 level, female learners outperformed their male counterparts, with a Chi-squared value of 4.57 and a p-value of 0.03. However, this trend reversed at the B1 and B2 levels, where male learners demonstrated higher accuracy in verb conjugation than females, as indicated by Chi-squared values of 24.966532 ($p < 0.0001$) and 13.404987 ($p=0.00025$), respectively.

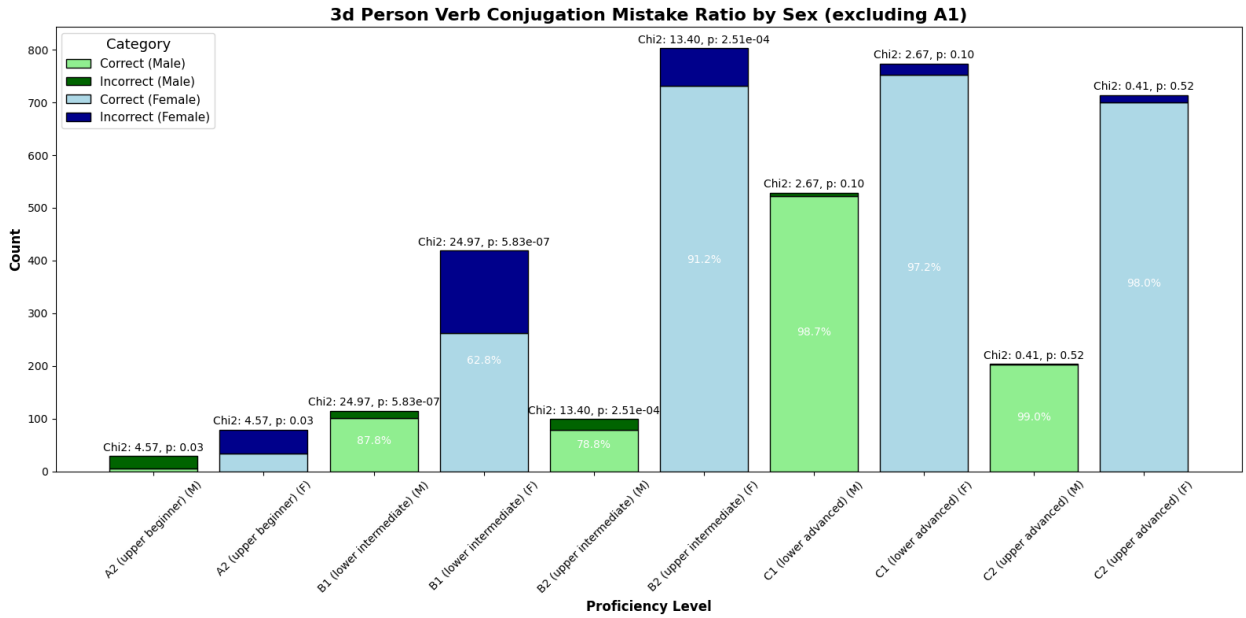


Fig.2 A bar chart showing the influence of Gender across different proficiency levels on Third-Person Verb Agreement among English Learners.

Another significant finding was the consistently superior performance of native German speakers compared to native Spanish speakers in verb conjugation accuracy from the B1 to C2 proficiency levels, with the results indicating statistical significance ($p \leq 0.01$). This suggests that native language may play a role in the ease of acquiring English verb conjugation patterns.

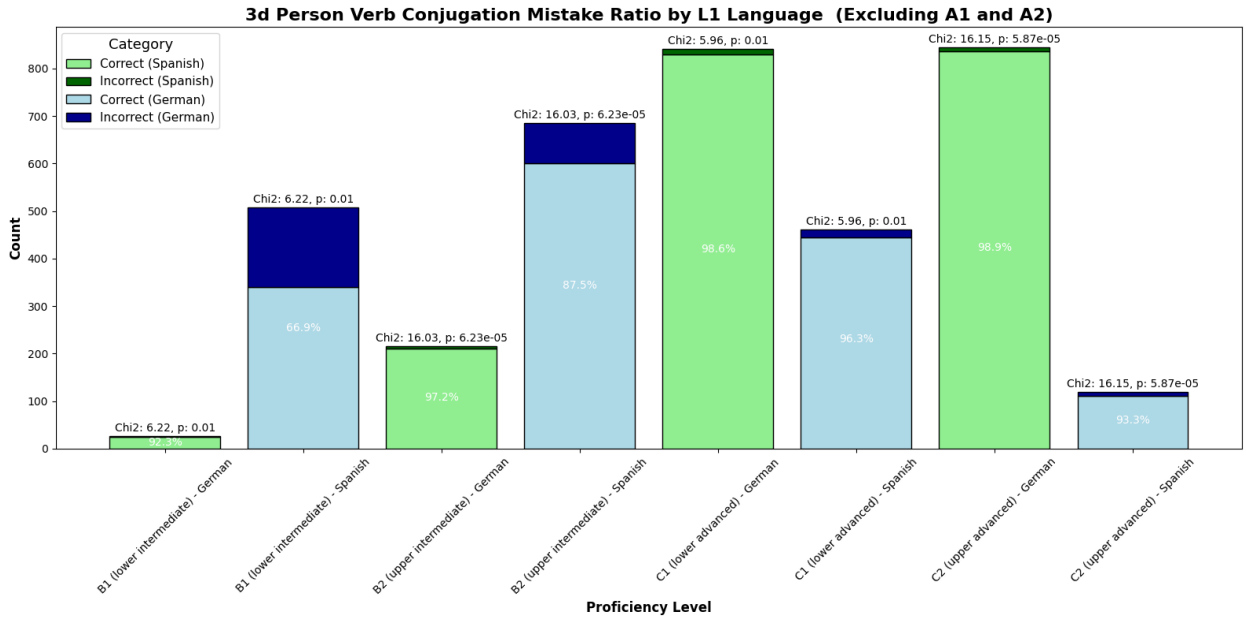


Fig.3 A bar chart showing the influence of Native Language across different proficiency levels on Third-Person Verb Agreement among English Learners.

Additionally, the Simple Linear Regression analysis showed a moderate correlation (R=0.61) between the learners' score percentages and their correct total ratio in verb conjugation, indicating a significant relationship between overall language proficiency and conjugation accuracy. The moderate correlation found in the regression analysis suggests a relationship between scores and conjugation accuracy, pointing to the potential of predictive models in estimating proficiency level.

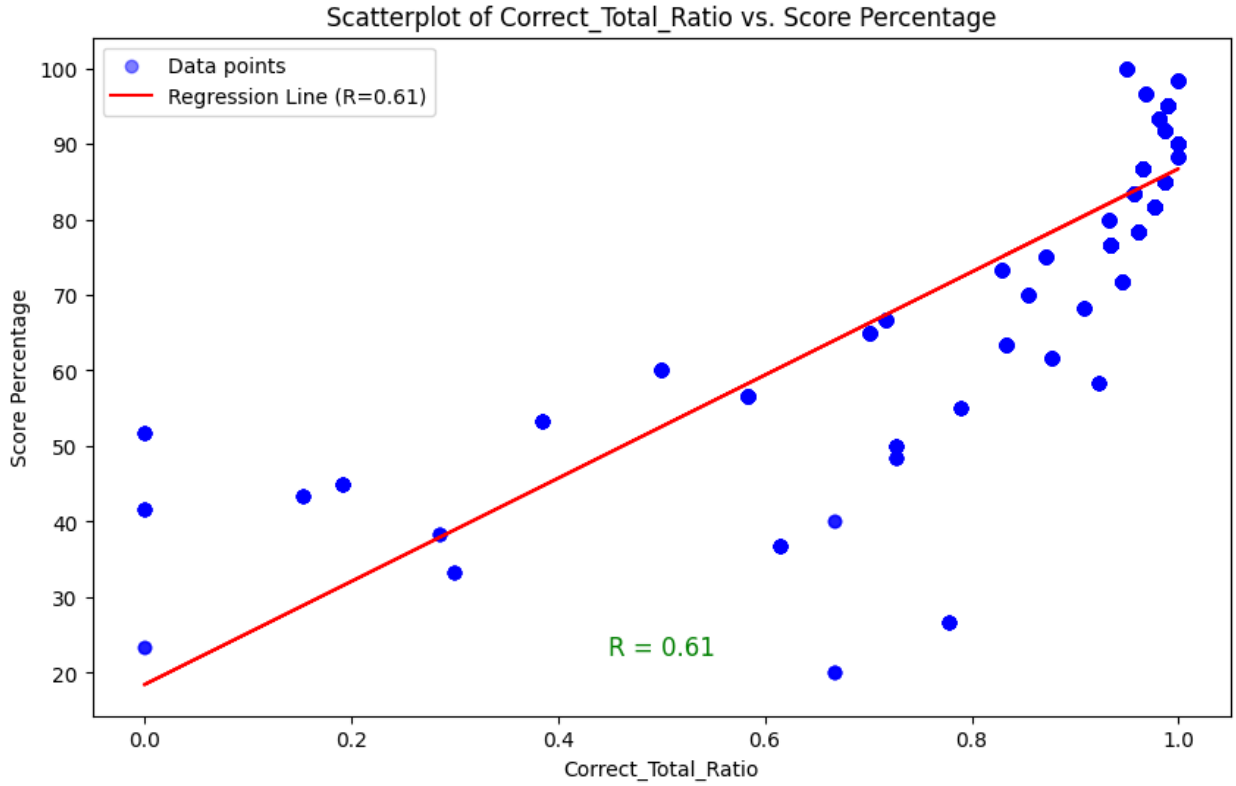


Fig.4 A scatter plot of English test scores against the correct ratio of 3-d person verb conjugation instances. Blue dots correspond to data points that represent unique scores.

Conclusions

The analysis results reveal that proficiency level, gender, and native language are significantly correlated with verb conjugation accuracy in English learners. However, it's crucial to note that correlation does not imply causation. The observed differences, such as the superior performance of German speakers, might not be directly caused by their native language alone. Instead, these differences could be influenced by other factors, including educational methodologies or learners' attitudes towards language learning. While these aspects correlate with learners' L1, it's possible that they independently contribute to the variations in verb conjugation accuracy.

While the results of the L2 corpus analysis are revealing, they should be considered exploratory due to limitations such as sample size and diversity. Further research with a larger and more varied dataset would be necessary to generalize these findings.

In summary, this comprehensive analysis combining Chi-squared tests and Simple Linear Regression potentially reveals the factors influencing third-person singular verb conjugation among English learners. While proficiency level, gender, and native language play significant roles, the relationship between learners' scores and conjugation accuracy could be utilized for developing a predictive model in proficiency level assessment.

Yusuke Taira

Linguistic Features and English Proficiency: A Multi Linear Regression Approach

This section explored the textual information of the corpus to find linguistic features that can be used to predict English proficiency test scores of participants. We first explored whether the filler words such as “uh” or “um” and the pausing marker “/” are more frequent in beginners due to the less fluency and narrow vocabulary range of the speakers. Thus, we conducted a correlation analysis using the scipy.stats module. Firstly, we extracted the pausing and filler words “uh” and “/” with regex. But since each subject produced a different length of speech, this count value was divided by the speech length so that we can calculate the pausing and filler words rate (%) which denotes how often each subject pauses in every 100 word token. Passing this value as an independent variable while English proficiency test scores as the dependent variable, Pearson correlation coefficient was calculated. The result revealed a significant negative correlation between the two variables [r = -0.59, p < .01], indicating less frequently pausing subjects tend to have higher English scores (Fig 5).

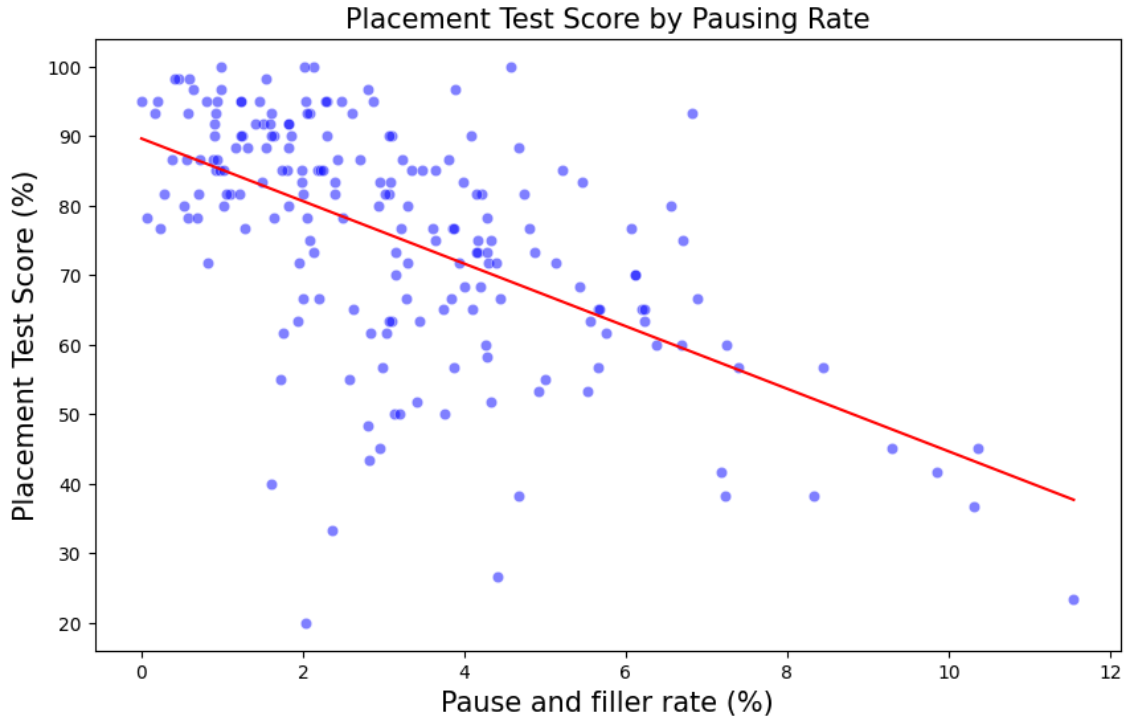


Fig 5. A scatter plot of English test scores against pausing rate with a regression line. Data points with a higher pausing rate correspond to lower English test scores.

The second feature we explored was part of speech (POS) tags. We hypothesize that frequency of adverb usage might differ across different levels of English as this POS does not significantly affect the overall meaning of a sentence, hence less prioritized in second language learning. The result revealed a positive correlation [$r = 0.58$, $p < .01$], indicating a tendency - the higher English level, the more often adverbs are used (Fig 6). Notably, other POS tags (noun, verb, adjective) did not show any significant correlations with English test scores (see Table 1).

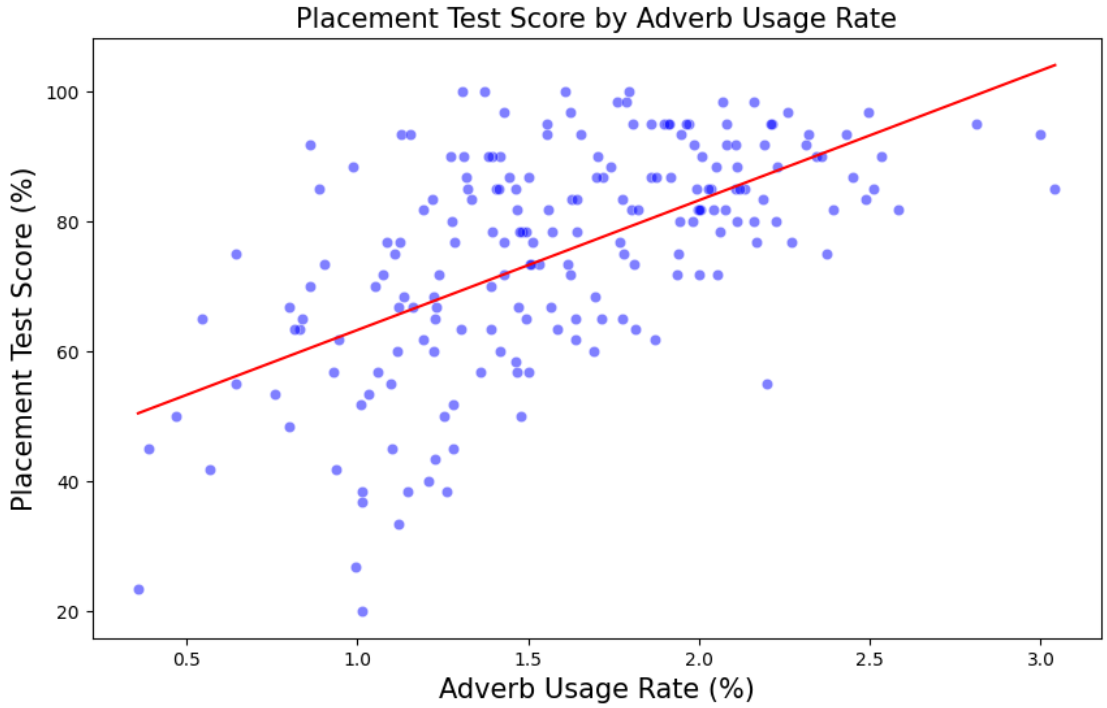


Fig 6. A scatter plot of English test scores against adverb usage rate. Data points with higher adverb usage rates correspond to higher English test scores.

Table 1. Pearson correlation coefficients (r) with p-values for each Part-of-Speech (POS) tag. Among these POS tags, the adverb usage rate shows the most significant correlation with English test scores.

	noun	verb	pronoun	adjective	adverb
R	0.22	0.19	0.24	0.16	0.58
p-value	< .01	< .01	< .01	< .05	< .01

Constructing a Multi Linear Regression Model

With the correct usage rate of 3rd person verb agreement, discussed in the previous section, these three variables (pausing rate, adverb usage rate and correct usage rate of 3rd person verb agreement) were selected as independent variables and included in the multi linear regression model. Using *sklearn*, 80% of the corpus data were used as the train set while the other 20% were used to test the predictability of the model. The test set contained 40 data points randomly selected from the corpus.The variance of the test data was 258.53, and our goal was to get Mean Squared Error (MSE) lower than this value to ensure the predictability of the model. The following figure shows the test results (Fig 7).

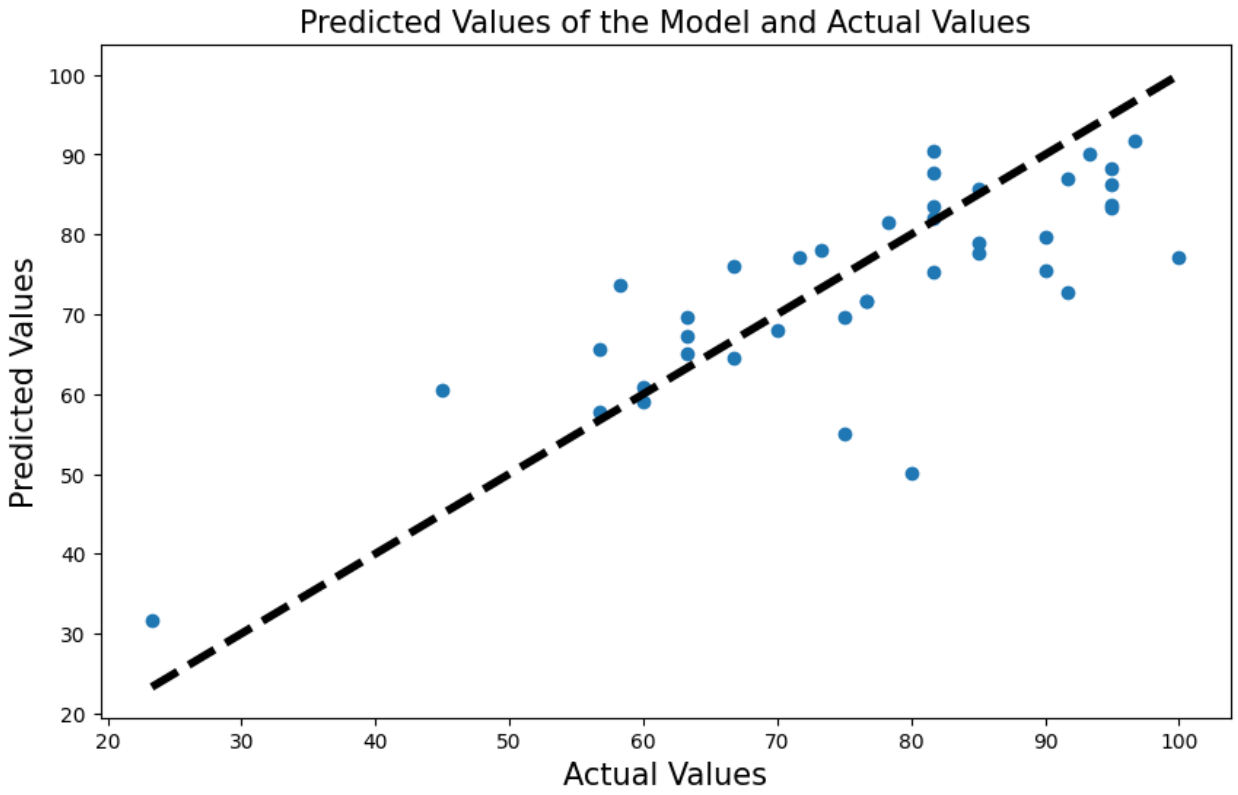


Fig 7. A scatter plot of the actual values against the predicted values of the test data. Blue dots represent the actual values from the corpus, while the black line indicates the predicted values based on the selected linguistic feature variables.

The predicted values of the model did not deviate much from the actual values (MSE = 102.93, R^2 = 0.59). The mean squared error was way below the variance of the actual values (258.53), which indicates the good fitness of the model.

Conclusions

We explored the Corpus of English as a Foreign Language (COREFL) to identify statistical relationships between various variables and attempted to construct a regression model for predicting English test scores based solely on the textual information in the corpus.

In terms of linguistic features relevant to English test scores, we found that third-person verb agreement, adverb usage rate, and pause rate were highly correlated with the scores. A multiple linear regression model incorporating these variables was able to predict English scores with minimal error.

Notably, past studies (Grant et al., 2000; Paredes et al., 2014) reported that adverb usage tends to increase with the proficiency of written language. Our findings suggest that this increase in adverb usage is also observable in spoken language. The discovery of the correlation between language proficiency and adverb usage rate presents an intriguing question about the relationship between language use and cognitive development, which future cognitive science studies are expected to address.

In addition, as the next section introduces, we developed a Streamlit application driven by the hope that corpus exploration can be accessible to a wider audience, especially those with little knowledge of programming. The intuitive design of the app enables users to access and explore the corpus easily. Our greatest aspiration is that this application will contribute to linguistic research, particularly in unraveling the mysteries of the relationship between language and cognition.

References

Grant, L. & Ginther, A. 2000. “Using computer-tagged linguistic features to describe L2 writing differences”. Journal of Second Language Writing, 9 (2), 123–145. DOI: 10.1016/S1060-3743(00)00019-9

Pérez-Paredes, Pascual & Sánchez-Tornel, María. (2014). Adverb use and language proficiency in young learners’ writing. International Journal of Corpus Linguistics. 19. 10.1075/ijcl.19.2.02per.

Maika

Overview of the Application

This application replicates the features of COEFL using Python. Additionally, it includes useful functions not found in COEFL, such as the ability to easily visualize the result of analysis.

Purpose

To apply the knowledge learned in the class by creating a corpus analysis tool using Python techniques. The tool can be used without writing any code. It’s user-friendly and allows for easy input of information. Users can easily choose the values they want to search for and select their preferred display method.

About our Application

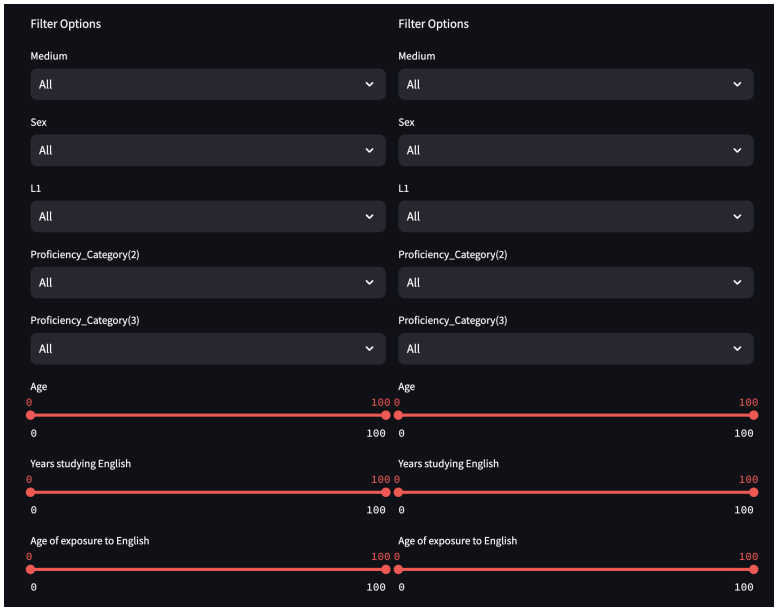
This section provides a brief overview of the application.

Preprocessing

In this application, preprocessing has been applied to the entire corpus data of learners, making it ready for analysis. The raw data consists of English sentences spoken or written by users in response to a particular task, stored alongside the data of the speaker. Essentially, the app manipulates data in the form of pandas dataframes, processing the Text (the text generated from the task, either written or transcribed from speech.) and saving the results as new columns.

Filtering Feature

In this section, data sets are generated artificially, primarily for the purpose of comparing two variables. The items selected from the options are used as keys for filtering, with the mechanism designed so that no filter is applied if 'All' is selected. This approach is used to divide the data frame into two data sets. The following are the types of filter categories. For more details, please refer to the 'Corpus Design' section on the official COEFL website.



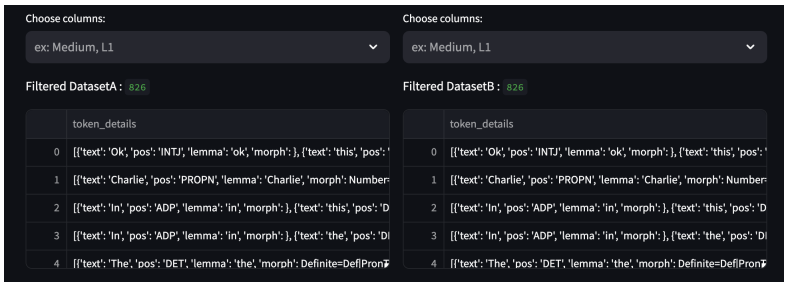
Selecting the Task Types

There are four types of tasks to choose from. Currently, Chaplin's data is the most abundant, making it the most suitable for analysis.

Task Type	Task contents
Famous Person	Talk about a famous person.
Film	Summarize a film you have seen recently.
Frog	Tell the story shown in the pictures. Text should start One day...
Chaplin	Watch the Chaplin video clip (4 minutes) and summarize the story.

1. **Task Type**: Please refer to the table above.
2. **Medium**: Options - 'All', 'Written', 'Spoken'. Represents the method of the test.
3. **Sex**: Options - 'All', 'Male', 'Female'. Indicates the gender of the subject.
4. **L1**: Options - 'All', 'German', 'Spanish'. Denotes the subject's native language.
5. **Proficiency_Category**: Options - 'All', 'Intermediate', 'Advanced'. Reflects the subject's English proficiency level, derived from test scores.
6. **Age**: Slider - Select the subject's age.
7. **Years Studying English**: Slider - Choose the number of years the subject has studied English.
8. **Age of Exposure to English**: Slider - Select the number of years the subject has been exposed to English.

For numerical values, Streamlit's slider is used, for others the selectbox is used. The results show the total number of data entries after selection and the dataframe itself. Users can choose which columns to display, allowing them to verify the proper application of filters. The contents of the filters can be viewed in the left sidebar. Later, differences in the data sets can be observed by viewing graphs or similar visualizations.



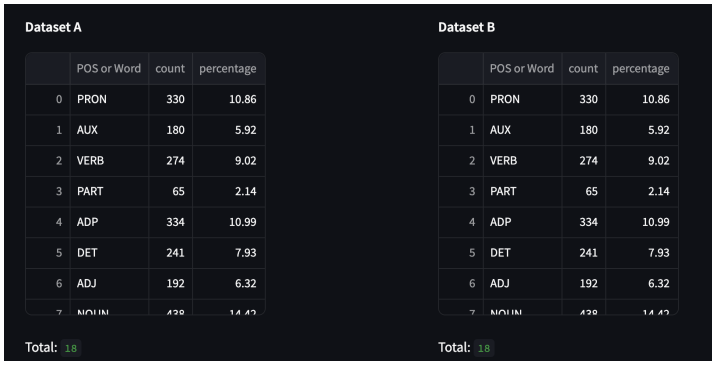
Comparison of Results

In the following sections, the data sets that have been filtered as described above will be visualized.

First, select the type of analysis as below:

- POS Counts: Counts and categorizes words by their parts of speech

- With Lemmatization: Counts words in their base form, combining variations like “runs,” “running,” and “ran” into “run.” It uses the results obtained from lemmatization with SpaCy.
- Without Lemmatization: Counts words as they appear, treating different forms of the same word separately.

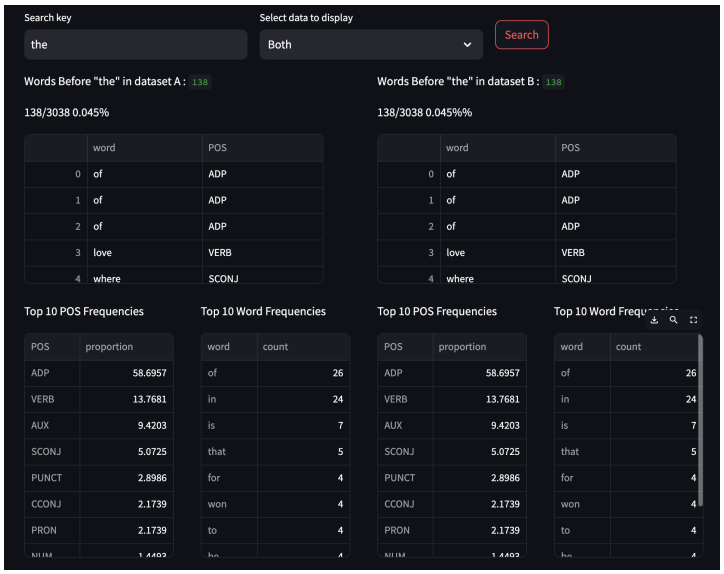


	POS or Word	count	percentage
0	PRON	330	10.86
1	AUX	180	5.92
2	VERB	274	9.02
3	PART	65	2.14
4	ADP	334	10.99
5	DET	241	7.93
6	ADJ	192	6.32

Subsequently, calculate and display the total frequency of occurrences and their respective proportions using the groupby method, presented in the form of a dataframe.

Word search

Here, detailed searches and analyses related to words can be performed.



	word	POS
0	of	ADP
1	of	ADP
2	of	ADP
3	love	VERB
4	where	SCONJ

POS	proportion
ADP	58.6957
VERB	13.7681
AUX	9.4203
SCONJ	5.0725
PUNCT	2.8986
CCONJ	2.1739
PRON	2.1739

word	count
of	26
in	24
is	7
that	5
for	4
won	4
to	4

By entering a search word and selecting the type of data display, the following information can be obtained:

- The frequency of the word's occurrence.
- The proportion of the word in relation to the total number of occurrences, with adjustable decimal point representation.

The context of occurrence.

- Comparison of the above with two datasets, A and B, as mentioned earlier.
- The word preceding the searched word.
- The word following the searched word.
- Search the word based on 'word_pos_pairs' while maintaining word order, retrieve the preceding and following words, along with their POS (Part of Speech) information. Calculate their proportions from the total count, sort, and display the top 10 results in a dataframe.

Visualization

The datasets from earlier or the results of the word search can be graphed, enabling a visual comparison of the data.

Selection for Visualization

- Choose whether to graph datasets A and B for comparison, or the results of the word search.

Options to Select

- POS or Word: Decide which type of graph to display.
- Ranking Top X: Choose how many elements to display in the graph. Graphs are arranged in order of highest values. Including too many elements can make the graph cluttered.
- Ignore: Select words to be excluded, such as fillers not desired in the rankings. The selected words are searched and removed from the original dataframe one by one, and the subsequent dataframe is graphed to avoid displaying these words.

Visualization Types

- Word Cloud: Displayed only when 'word' analysis is selected. Created using the wordcloud package. Compiles all English text from the filtered dataset's [text], displaying more frequently occurring words in larger fonts. Words to be excluded can be chosen from the ignore list.
- Bar Graph: Facilitates easy identification of significant differences as it displays the same words or POS values of datasets A and B side by side. Remains readable even when displaying a lot of data like the top 20.
- Pie Chart: Clearly shows the proportion each element occupies. However, including too much data like the top 20 can make it difficult to read.