ESL Syntactic Analysis

Tianyi Zheng

Background

ESL, PELIC, and Project Goals

What is ESL?

- English as a Second Language (ESL) is the instruction of English to native speakers of other languages
  - Classes may be comprised of speakers from a variety of native languages (L1s) and proficiency levels
  - There is a lot of research delving into second language acquisition, and more specifically ESL

What is PELIC?

PELIC is a corpus of writing and speech samples from students at Pitt's English Language Institute

- Managed by Drs. Alan Juffs, Na-Rae Han, and Ben Naismith
- Consists of 7 years' worth of data, specifically data from students over time (longitudinal data)

Project Goals

My goal for this project was to analyze variations in syntactic complexity in PELIC

- How does the students' syntax vary between proficiency levels?
- Are there significant differences in syntactic complexity between students with different L1s?

I tried to investigate these questions by performing statistical analyses on the PELIC written data

Measures of Syntax

I decided to investigate ~~7~~ 6 syntactic measures:

1. Number of T-units per sentence
2. Mean length of T-units
3. Number of clauses per T-unit
4. Mean length of clauses
5. Number of prepositions per clause
6. Number of subordinating conjunctions per clause
7. ~~Number of discourse markers per clause~~

What is a T-Unit?

A T-unit is defined to be an independent clause and all of its associated dependent clauses

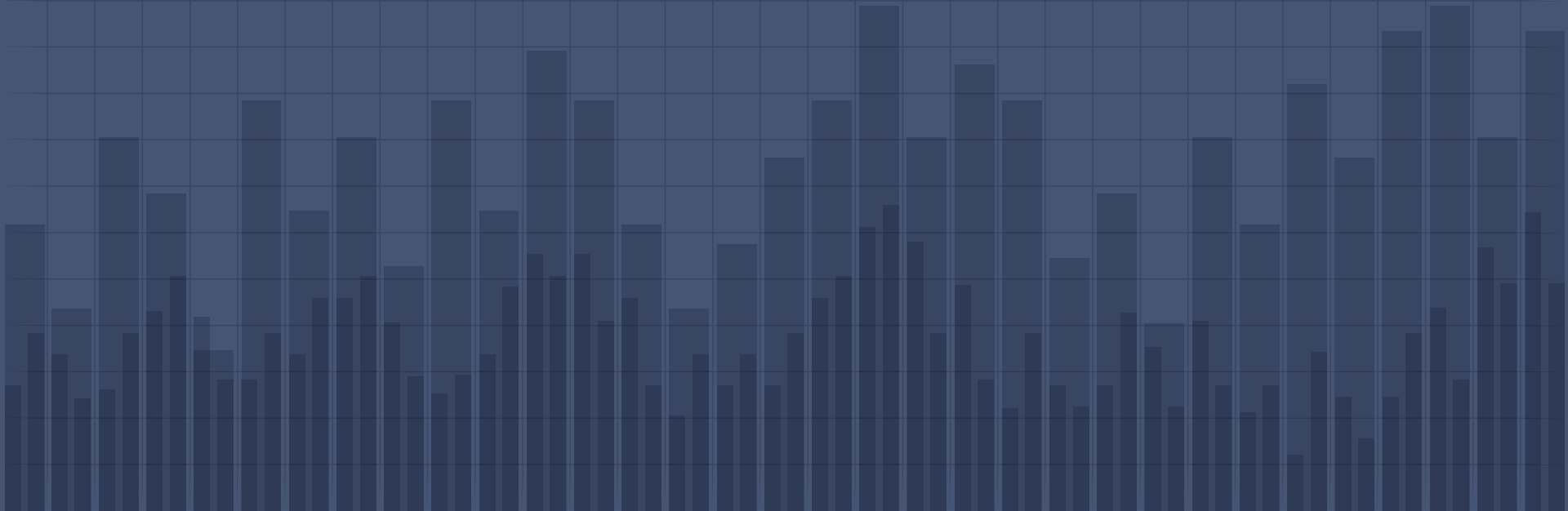
Examples:

Because this sentence has one independent clause, it has 1 T-unit

This is a compound sentence with 2 independent clauses, so it has 2 T-units

The Data

Exploration and Visualization



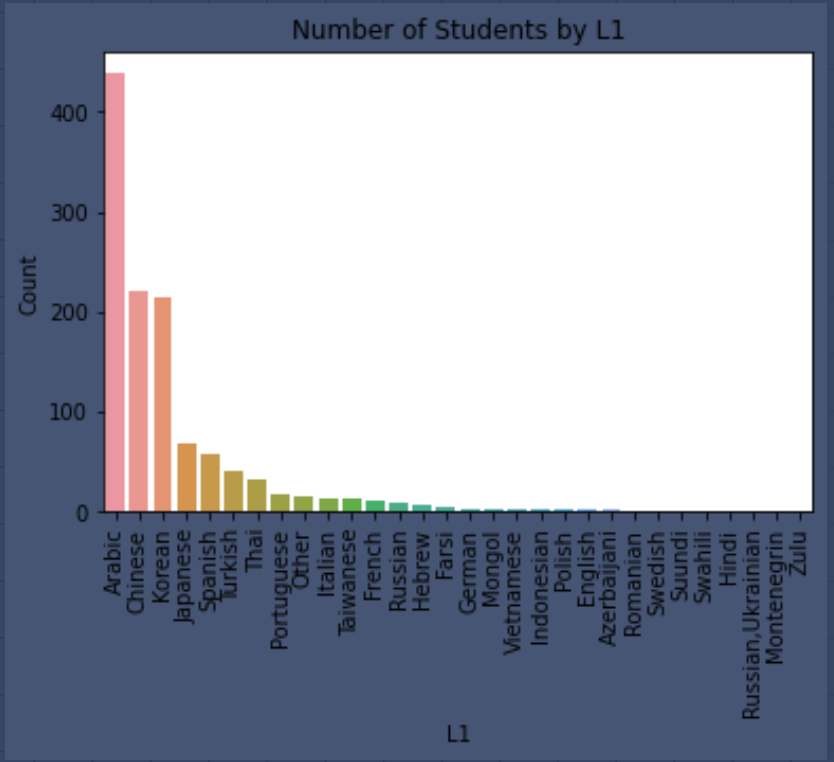
Overview of PELIC Data

The written corpus consists of 46,204 student essays

Over 30 L1s represented

- L1 counts vary drastically

Arabic and CJK students dominate the dataset



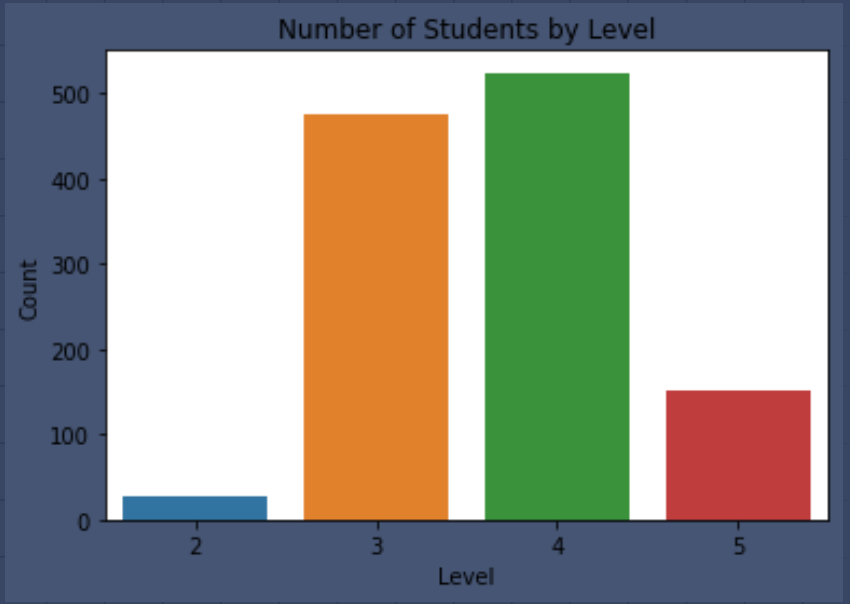
Overview of PELIC Data

The students' proficiency levels ranged from low-intermediate to advanced

Proficiency levels:

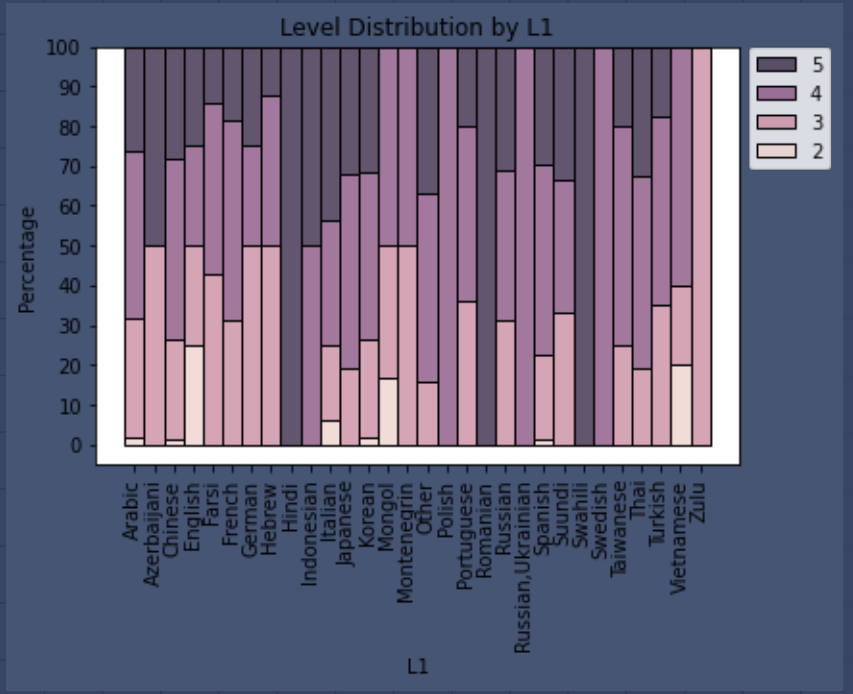
- 2: CEFR A2 – B1
- 3: CEFR B1
- 4: CEFR B1 – B2
- 5: CEFR B2 – C1

Most students were at a mid-to-high-intermediate level



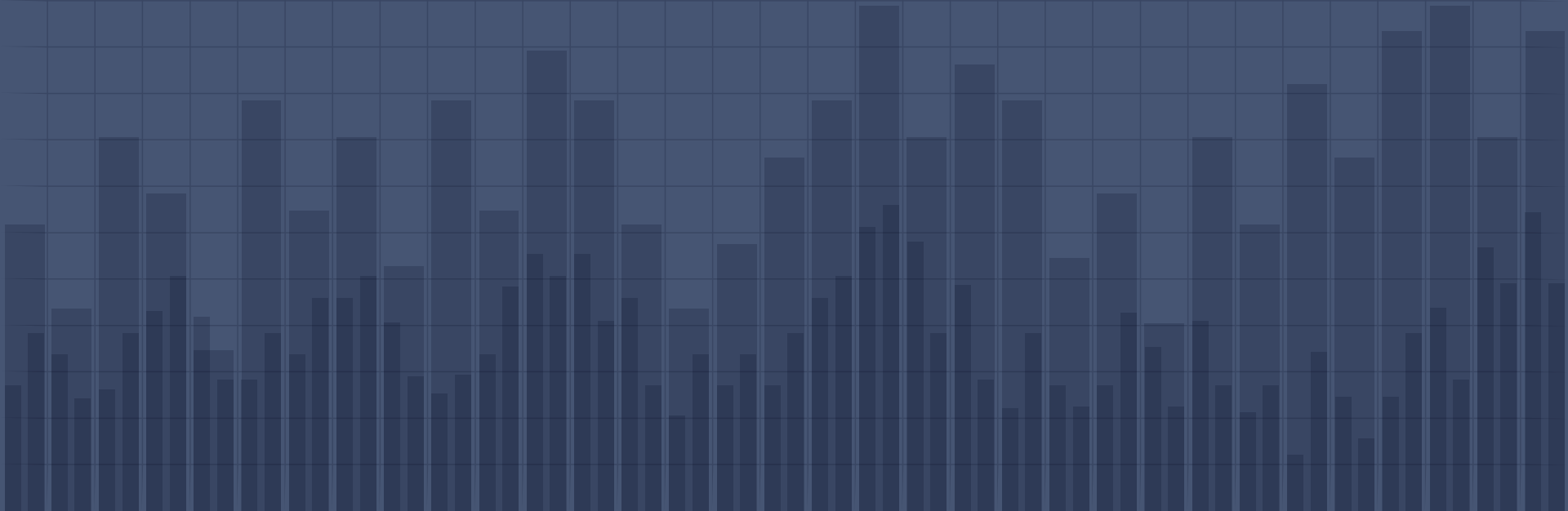
Overview of PELIC Data

When broken down by L1, the distribution of proficiency levels vary drastically



Generating the Measures

Where the pain *really* begins



What is TAASSC?

The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) is a program that can calculate many numerical measures of syntactic complexity for writing samples

- It takes in individual text files and outputs their measures in a CSV file

The Bottleneck

I wrote every essay to a separate text file and ran TAASSC on those text files (or at least I tried)

It turned out that TAASSC is *ludicrously inefficient*

- When I had tried to process the essays for my 2nd progress report, TAASSC took *two days* to get through about half to two-thirds of the essays
- It also stored temporary files for every essay that it doesn't delete until it's completely done running, which bloated the size of the program over time

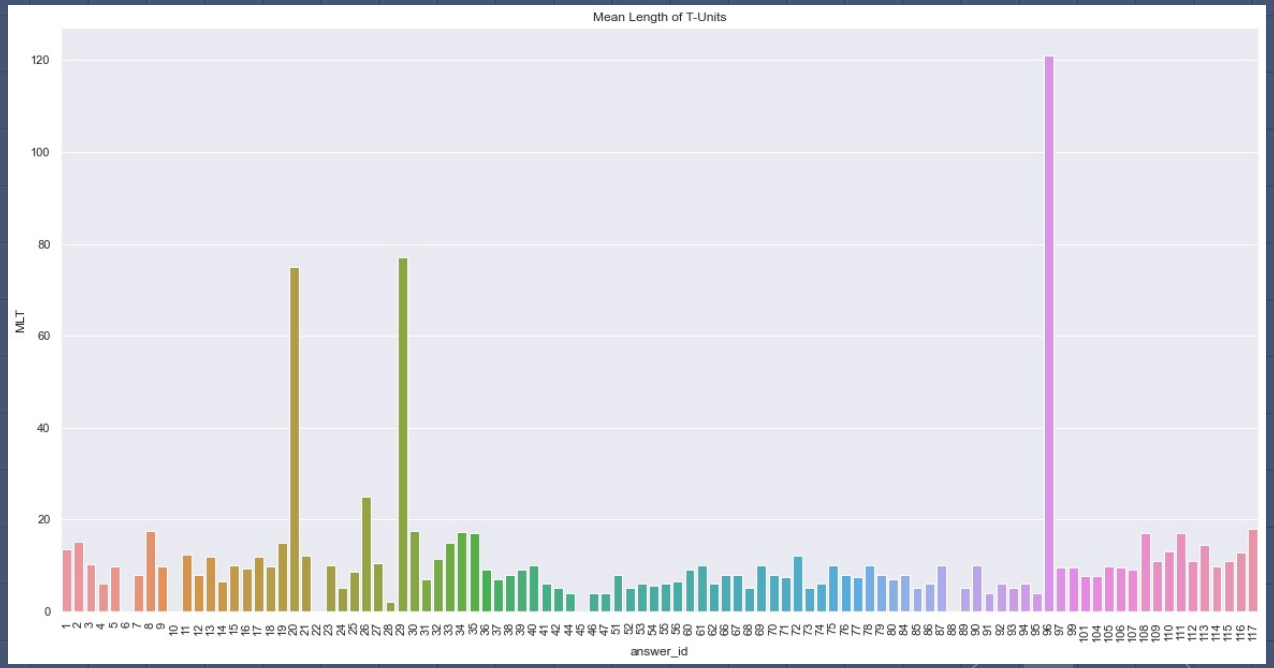
The Pain

While working on my 2nd progress report, I discovered that *none* of the essays had any discourse markers

- Did TAASSC only count certain kinds of markers? I'm not sure and I couldn't find any info about it
- I ended up dropping discourse markers from my list of syntactic measures (hence why it was crossed out earlier)

The Pain 2: Electric Boogaloo

- TAASSC was also *incredibly sensitive* to imperfect writing
  - Things like non-standard grammar, punctuation, and whitespace could easily trip it up
  - As a result, the generated syntactic data contained a lot of (incorrect) outliers
  - This was obviously a massive problem since I'm working with ESL writing in particular



Essay 96:
Failing a test is not easy successful ,but if you want to do this,there are many method can reach this goal.For example,sometimes i always forget what the teacher say in the class,and i never write down any important thing in my notebook.So i never get a high score in my test.Then, you needn't to do your homework and don't go to class on time.When you have a test,don't write anything on your test paper if you know the answer.On the other hand,you needn't to review your classes everyday and don't highlight any important vocabulary or senternces on your book.Finally,if your follow tese steps, you'll esay to fail a test.

The Compromise

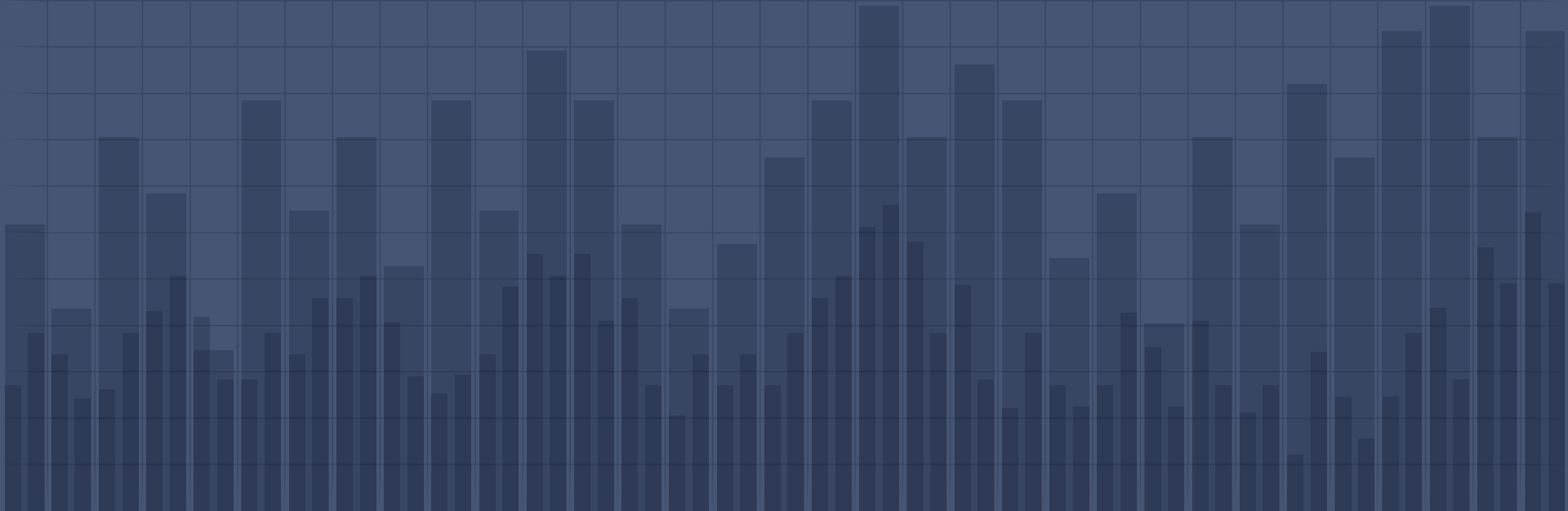
I decided to massively scale down my sample size

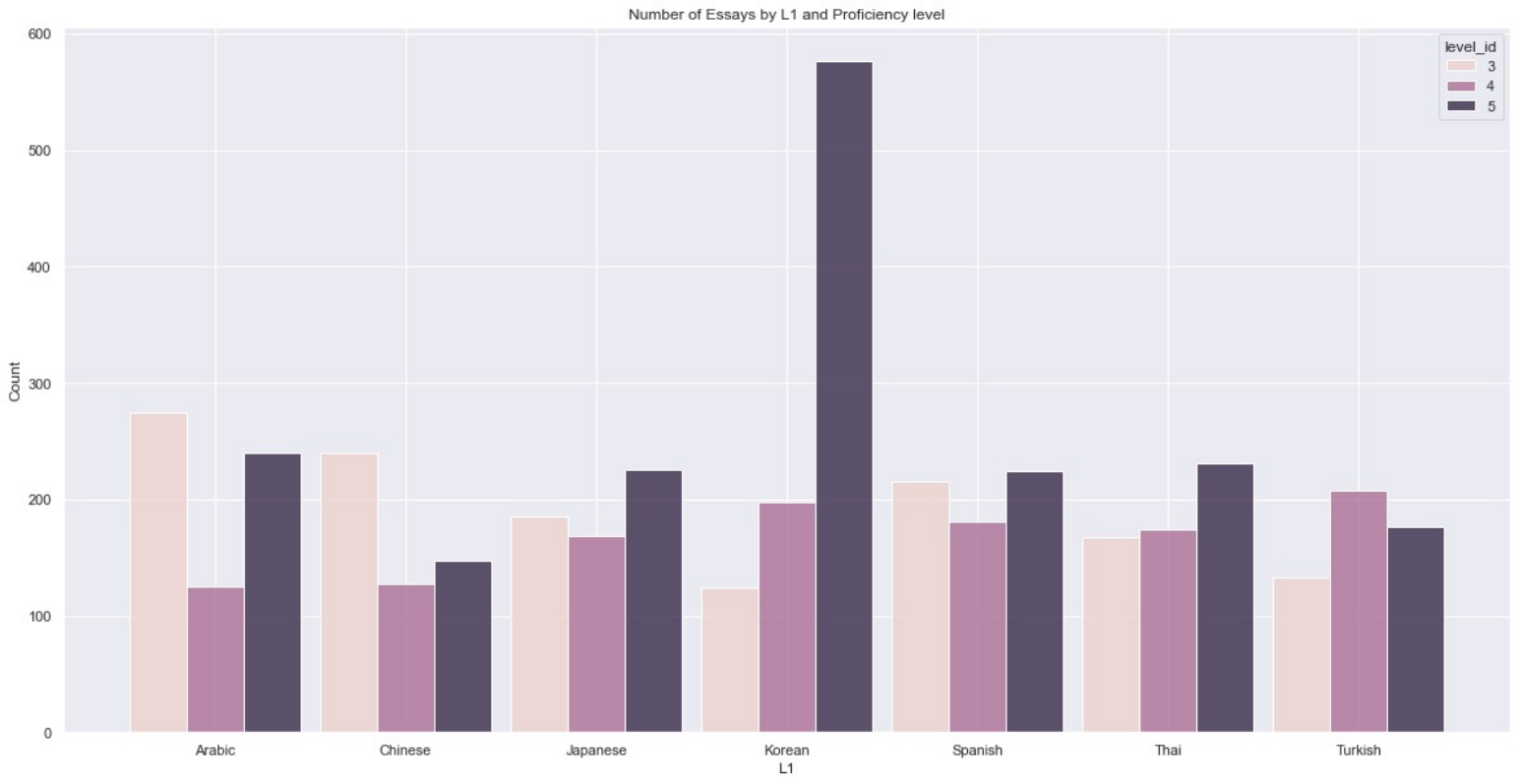
- I stuck to only L1s with more than 30 students
  - Arabic, Chinese, Korean, Japanese, Spanish, Turkish, and Thai
  - At least I got good variation in language families
- For each L1, I randomly sampled 10 students each for proficiency levels 3, 4, and 5

This cut down my dataset to only 4,341 essays (but it still took TAASSC hours to run)

Final Data Analysis

(Or Lack Thereof)





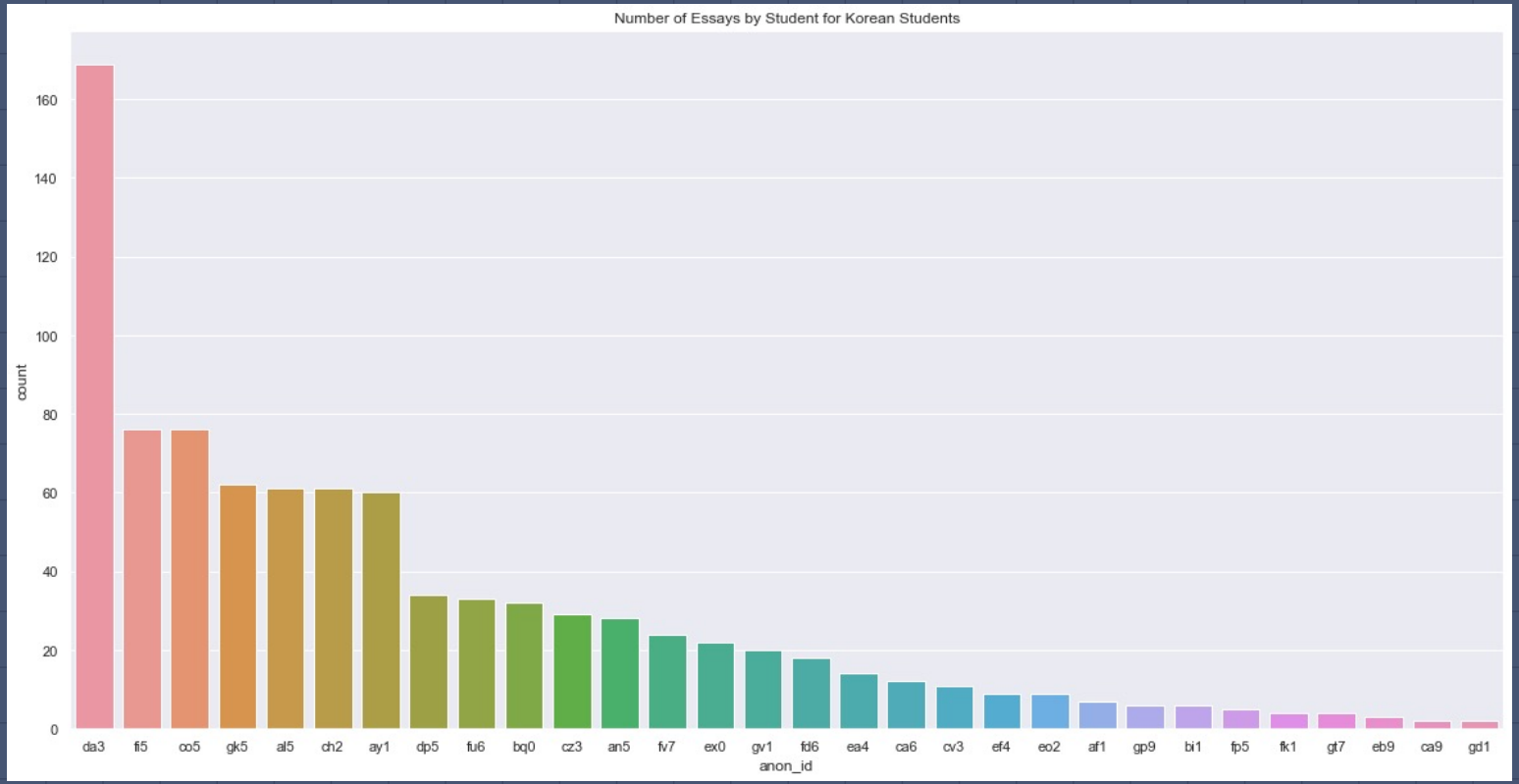
The Final Dataset

Okay Something Went Horribly Wrong

There were *way* more essays by advanced Korean students than any other group in my sample

It turned out that advanced Korean speakers had more essays on average than most groups in the dataset, but it wasn't nearly as extreme

I decided to investigate the Korean students specifically



For some reason this one student had over twice as many essays as everyone else...

Bad Luck

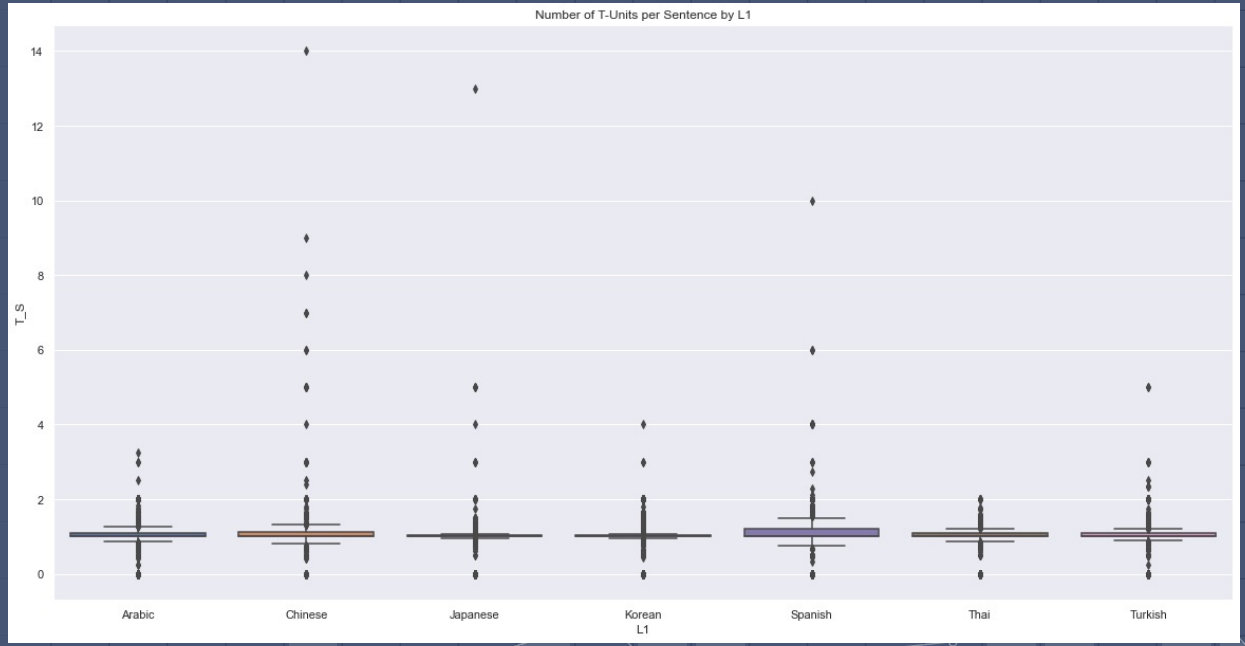
Apparently, I just had really bad luck and ended up selecting that one Korean student, which skewed my distribution

- Since there was no way that I'd have enough time to try to fix this (using the whole dataset, sampling multiple times, etc.), I had to just stick with it and hope for the best

Hitting a Wall

Remember how I said that TAASSC was really sensitive to non-standard writing? Yeah, about that…

- Those incorrect calculations/outliers ended up being a serious problem



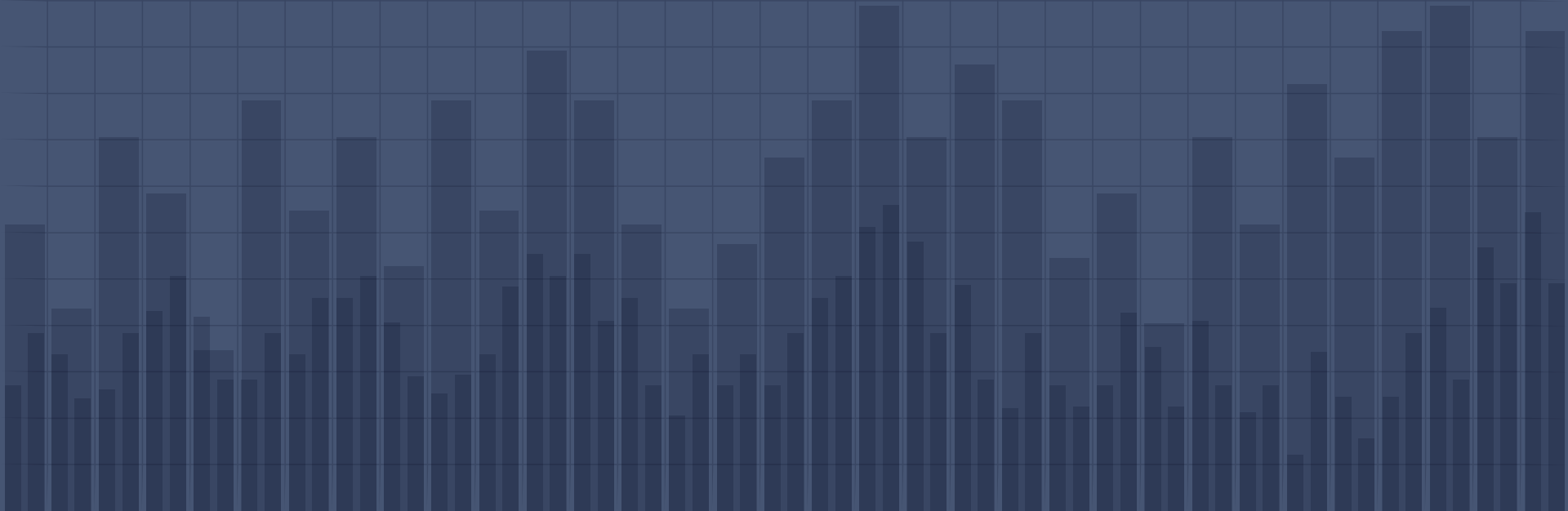
The most extreme “outlier”: Pon-ANON_NAME_0 is an island near Taiwan,I went there with my classmates.That was my graduation travel.According to my experience,I think it was the best place to go on a vacation.First of all,we can do many water activiies there,like snorkling and taking a boat,they are very exciting.Second,we can try some dishes which are their traditional foods,the foods were made by local.Third,there are also many good places in Pon-ANON_NAME_0,we can see nature and strange stones,if we go there,we can feel relaxing.The most importantly,there are firework fastival in summer vacation,if we go there on July,we can see fireworks there,they are very wonderful because many pictures like animal and colors in the sky.According to these things,I think the best vacation place is Pon-ANON_NAME_0.

Unreliable Data

- “Outliers” like this were scattered throughout my dataset
  - Although I could only examine the most extreme outliers, it’s likely that the measures for the other essays, both non-outliers and less extreme outliers, were still miscalculated to some degree
  - It’s likely that these inaccurate calculations are seriously skewing the distributions of the syntactic measures

Conclusion

It’s Almost Finally Over



What Now?

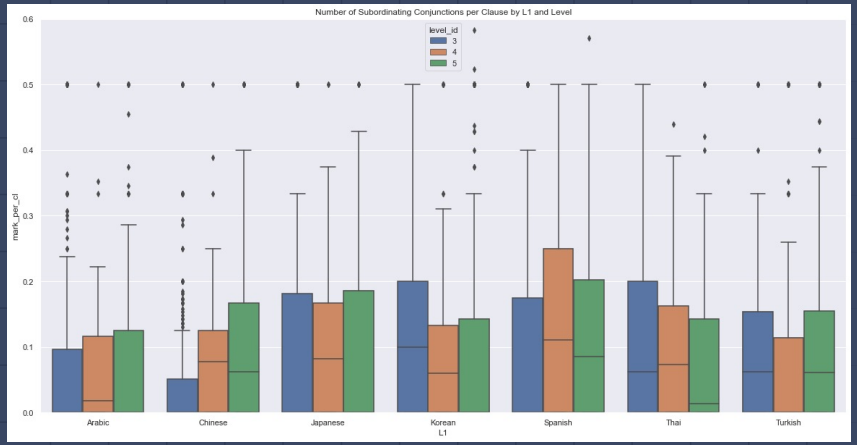
I’m torn about how I should proceed from here, if at all

- I don’t think I can in good conscience perform statistic analyses on this data and draw conclusions when I know that the data is this inaccurate

Glimmers of Hope

Exploratory data analysis revealed that there may still be trends worth exploring (ignoring the outliers)

- Could be tested with t-tests or ANOVA tests
- Likely not reliable until better methods are found



Some Ideas

Find a substitute for TAASSC?

Manual tagging and parsing?

Other ideas that I haven't thought of?

- Open to suggestions

Questions? Ideas?