FROM ¿QUÉ? TO CONVERSATIONAL

A Data Story on Language Learning

PROJECT OVERVIEW

This project applies NLP and quantitative analysis to track language acquisition over 112 days of Spanish-speaking practice. By analyzing speech patterns and visualizing key trends, it uncovers patterns in fluency, vocabulary growth, and filler word usage. By leveraging data analytics techniques, I deconstructed a complex skill (fluency) into insightful metrics, making language-learning progress tangible and trackable.

BACKGROUND & TOOLS

The dataset was built by recording myself speaking in Spanish over 112 days, capturing key linguistic features like words per minute (WPM), filler words, and unique vocabulary usage. I used NLP to analyze the text data, employing Python libraries like Pandas, Seaborn, and Numpy for data manipulation, visualization, and linguistic analysis.

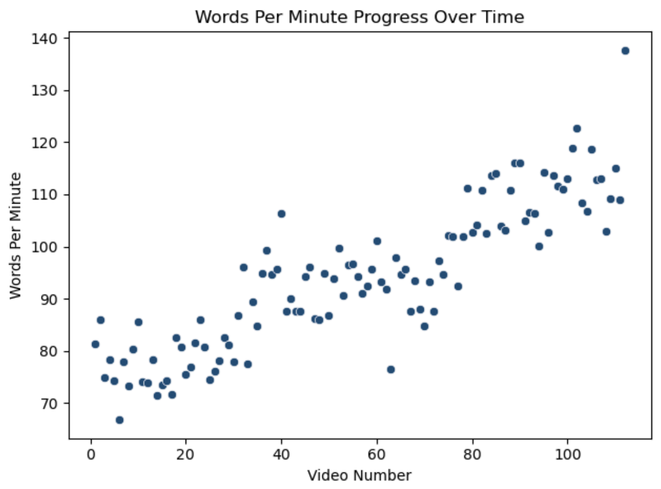
Key Insights

01

Speaking Rate is Increasing

As measured by WPM, my speaking rate increased linearly throughout the project.

Between Weeks 1 and 16, WPM increased 47.5%.

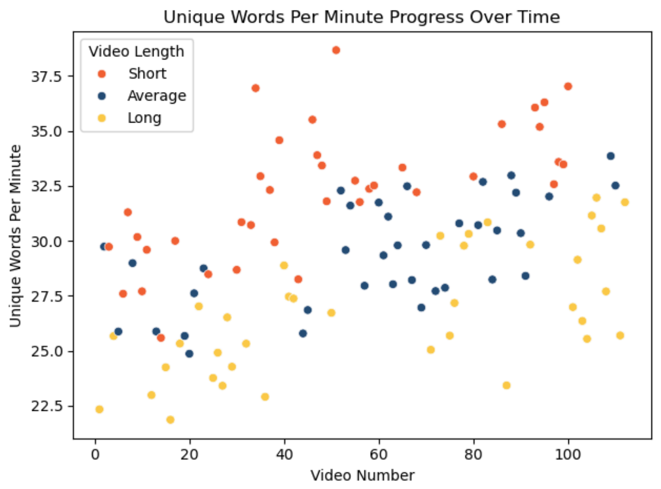


02

Vocabulary is Expanding

Unique Words per Minute has increased throughout the course of the project, but it is influenced heavily by the length of video.

The chart shows that videos within the same class of length have upward trends, with longer videos having lower Unique WPM.

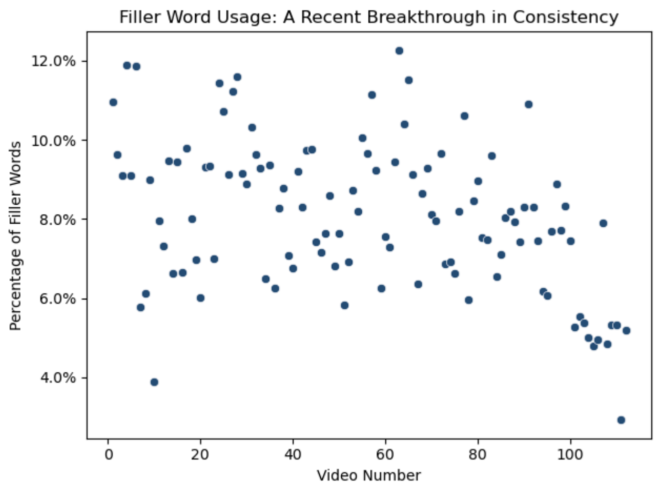


03

Recent Drop in Filler Words

Between Weeks 1 and 16, Percentage of Filler Words decreased 48.5%, with this drop materializing in the final two weeks.

This chart shows general variation throughout the project with an abrupt change appearing around Day 100.



Reflections

Limitations

- Automated transcriptions were ~95% accurate, introducing some errors and variance
- External factors (fatigue, stress) may have influenced speech patterns
- Video Length strongly influenced speech variation patterns and thus measurements

Future Work

- Benchmark metrics with native speakers to understand areas for growth
- Analyze videos from learners over time to understand factors that introduce variance

Conclusion

While Words per Minute showed a steady, linear increase over time, Unique Words per Minute was highly dependent on video length. Only after controlling for this factor did clear progress in this metric emerge. The trend of Percentage of Filler Words decreasing is a more recent development.

While I am becoming more accustomed to the pace of conversation, much of the work that lies ahead is transferring words from passive to active vocabulary. If I want to continue improving my speech, I need to practice speaking about more complex topics that elicitmore advanced vocabulary and sentence structures.