FURMAN
COMPUTER SCIENCE

# Dialect Dive

## The Intersection Between Linguistics and Technology

Ainsley Yoshizumi

## ABSTRACT

Following Mandarin, Spanish is one of the most spoken languages in the world. Over time Spanish has slowly diverged into many different regional variations. There is Catalan and Andalusian, Caribbean, and Andean, to name a few. There is a plethora of information and research surrounding the language's diversity. However, studying Spanish and its variations can be difficult because this information is stratified across the internet. In this project we aimed to create an online resource that provides information on these regional variations, helping L2 learners understand key distinguishing factors of these variations, ultimately supporting the growth of their listening and speaking skills.

## INTRODUCTION

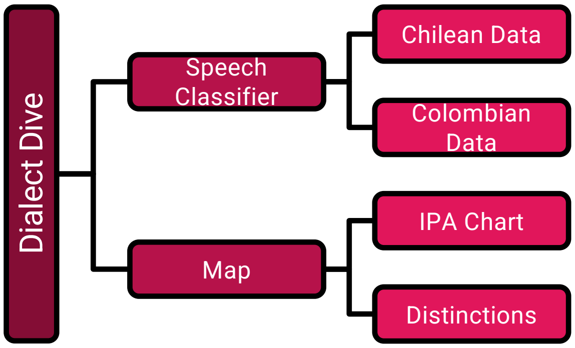
Dialect Dive is a website created by an L2 learner, for L2 learners. What started as a genuine interest in the many regional variations in Spanish slowly turned into a tool used to create new mediums for education regarding Spanish linguistics. Dialect Dive feature two sections, a space to practice speech, and another to explore phonemic differences across multiple languages.

## DATA

| Audio Filename | $V_0$ | $V_1$ | $V_2$ | $V_3$ | $V_4$ |
|---|---|---|---|---|---|
| File 1 | -0.000106 | -0.000219 | -0.000209 | -0.000246 | -0.000185 |
| File 2 | 0.00097 | 0.001404 | 0.001253 | 0.001405 | 0.001204 |
| File 3 | -0.000174 | -0.000079 | 0.000005 | -0.000186 | -0.000086 |
| File 4 | 0.000069 | 0.000033 | 0.000045 | 0.000058 | 0.000038 |
| File 5 | 0.000002 | 0.000018 | -0.000014 | 0.000015 | 0.000021 |

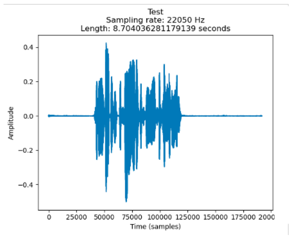
We created a classifier derived from a crowdsourced dataset created by Google language resources found on https://www.openslr.org/71 and https://www.openslr.org/72 The classifier has been trained on Chilean female and Columbian female speech audio and classifies the audio files using the K-Nearest Neighbor algorithm. We have two excel files, one for Chilean and Colombian that look like the figure above. The dataset contains roughly two gigabytes of data for each set, this figure is just a small snippet of the CSV file created with the data. There are 1,738 rows and over 16,000 columns in the file, each filled with the vector value at each given second.

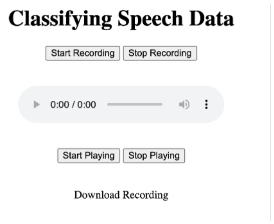
## PROCESS



The process of creating our dialect resource website began with aggregating the data obtained from audio file graphs and training a classifier on the dataset. The next step was to build a map showing the different locations which different phonemes in various languages can be found. Another feature of Dialect Dive is the speech recorder.

## MODEL




Classifying Speech Data

Using the language resource dataset, we created graphs (as shown in the above left hand figure) from the speech audio based on amplitude over time. Each of these audio samples have Y-vector values that we aggregated into CSV file for both Chilean and Colombian audio. These amplitude based vectors are essentially what we used to train our classifier to distinguish between the two dialects. On the right, is a screenshot of the audio recorder featured on Dialect Dive. This takes input from the user and allows them to playback the audio, this is ideal for L2 learners wanting to practice their pronunciation of different Spanish phonemes.

## RESULTS

| Regional Variety | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Colombian | 0.66 | 0.96 | 0.78 | 472 |
| Chilean | 0.86 | 0.33 | 0.47 | 350 |
| Accuracy | | | 0.69 | 822 |
| Macro Avg. | 0.76 | 0.64 | 0.63 | 822 |
| Weighted Avg. | 0.74 | 0.69 | 0.65 | 822 |

In the above figure are results from the K-NN classifier we built on the Google language resources data set we found online. This classifier has a projected accuracy of 69% at classifying data as either Colombian or Chilean. It is important to note the imbalance in audio files, as we have more data for Colombian than Chilean. Regardless, the accuracy is still an ideal number for the amount of data which the classifier was trained on.

In addition to this, we also created a map which displays what phonemes are character to that specific regional variation indicated on the map.

## CONCLUSION

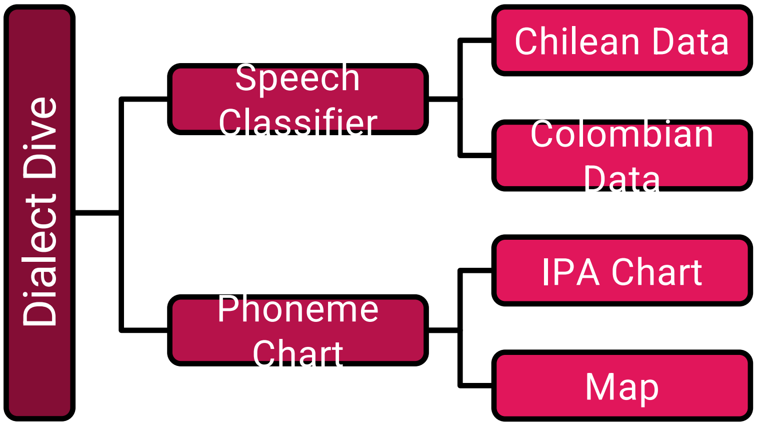
In short, Dialect Dive can be used as an important resource for L2 learners interested in deepening their understanding of the language. Spanish regional variation has evolved over the course of time, reflecting historical, cultural, and sociolinguistic changes within the language. To have such important information regarding regional variation all in one place encourages users to explore these topics at the click of a button

## FUTURE WORK

Dialect Dive still requires work, more specifically on the audio classifier. We hope to add multiple other dialects to train the classifier on recognizing. Additionally, when we started our work, we were interested in a website created by North Carolina State that allows users to compare various audios of Spanish speakers from different countries. While Dialect Dive allows users to gain insight and a deeper understanding of the exact phonetic variation regionally, there exists no such website that combines both projects.

http://github.com/ainsleyyoshizumi/CSC475Dialects

https://ainsleyyoshizumi.github.io/CSC475Dialects/demo.html


Dialect Dive
Speech Classifier
Chilean Data
Colombian Data
Phoneme Chart
IPA Chart
Map