Capstone 1

Exploration of Coronavirus data and data from Duolingo

The spread of Coronavirus

In 2019 a novel coronavirus was discovered in China

The novel virus has been named SARS-CoV-2

The disease it causes has been named Covid-19

The virus is fairly contagious and relatively lethal

It has disrupted markets on a global scale and has been declared a pandemic by WHO

Predicting the spread of Coronavirus may help save lives

Coronavirus and temperature

Coronaviruses are typically less stable at higher temperatures

I wanted to understand if the ambient air temperature in a region was a predictor determining the growth rate of the novel coronavirus

I downloaded data on March 17th from the European Center for Disease Control

https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

I analyzed Covid-19 case growth rates among countries that had 50 or more recorded cases - Amounted to 45 countries

Coronavirus and temperature

Cases of COVID-19 show an exponential growth rate in the population

- I fitted exponential models to the portions of the curves that exhibited exponential growth

I then examined the relationship between the ambient temperature in a country and the growth rate of COVID-19

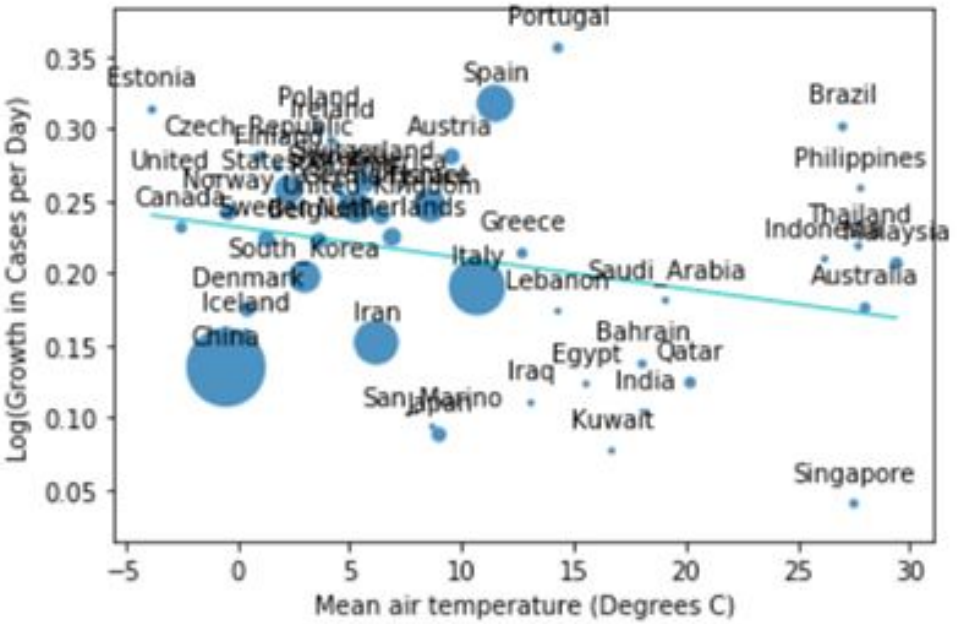
Coronavirus and temperature: Results

R2 = 0.08

P = 0.06

No strong relationship between growth rate and temperature

Notice that many countries that fall below the curve put in place strong social distancing orders



Duolingo and second language acquisition

Duolingo is a website that promotes second language acquisition and learning

I have been interested in examining if it is possible to predict how intensely a user will interact with the website Duolingo

The following slides show some plots, and statistical summaries I have generated as part of my exploratory data analysis

Duolingo data wrangling

The data I downloaded from duolingo was available as .csv file from the following website

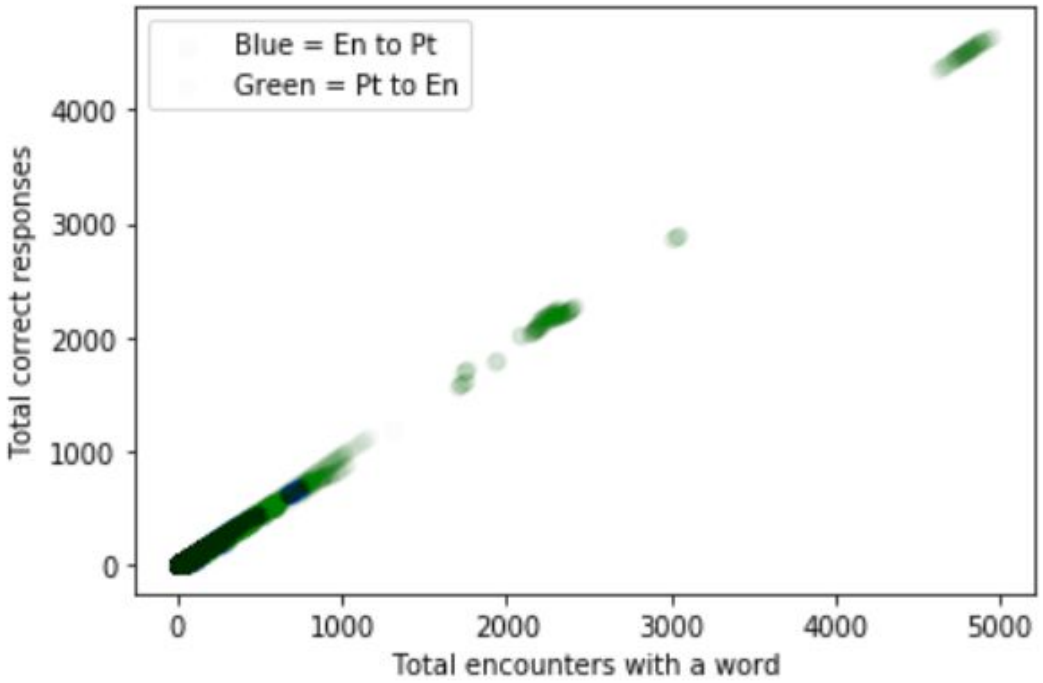
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N8XJME

The file itself was well taken care of, and required very little reshaping for the exploratory analysis phase

I began the exploratory phase of analysis by examining how many times a user correctly identified a word and how many times a user had encountered a word

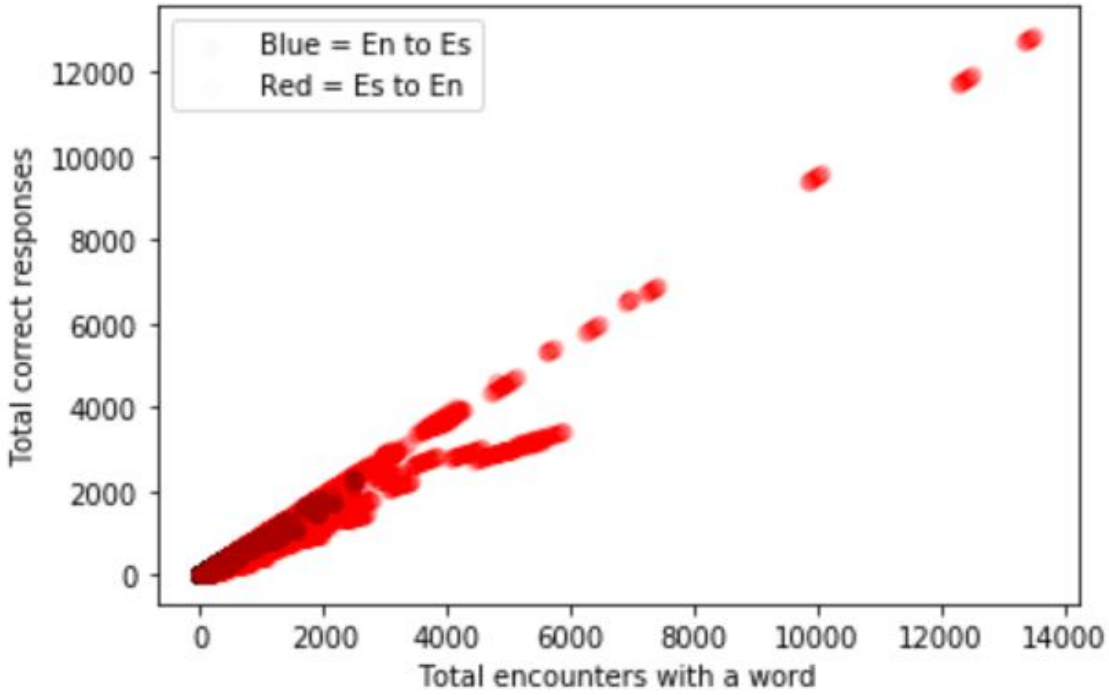
Duolingo: En to Pt and Pt to En

English learners from Portuguese appear to show more persistence (as measured by total encounters)



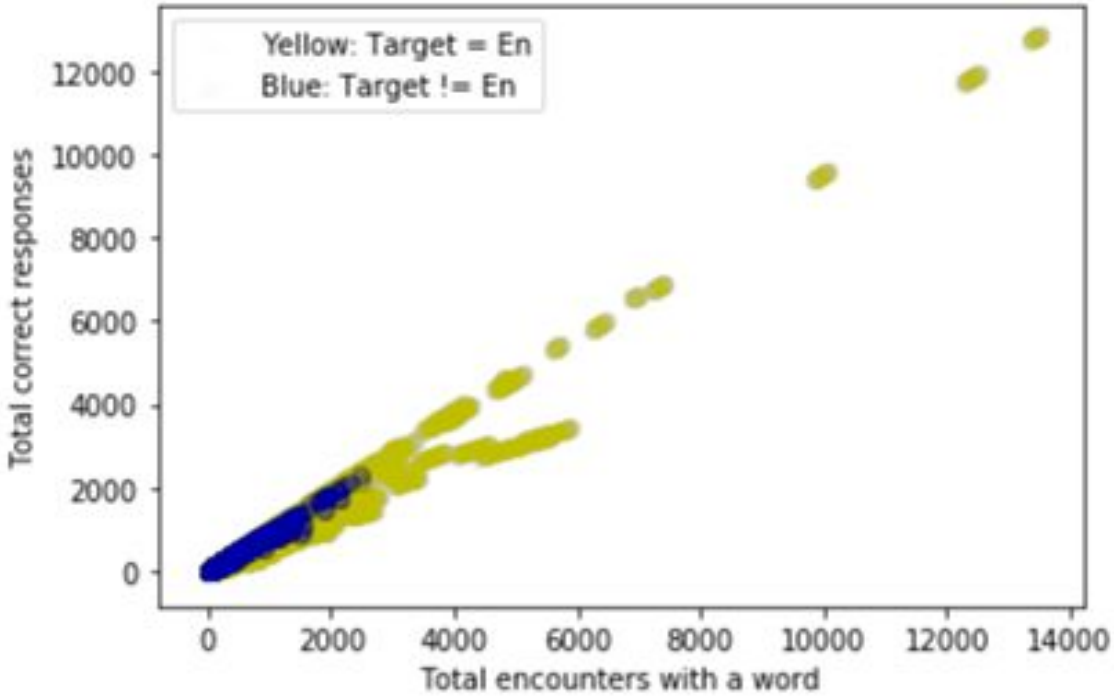
Duolingo: En to Es and Es to En

English learners from Spanish appear to show more persistence (as measured by total encounters)



Duolingo: En to Target and Target to En

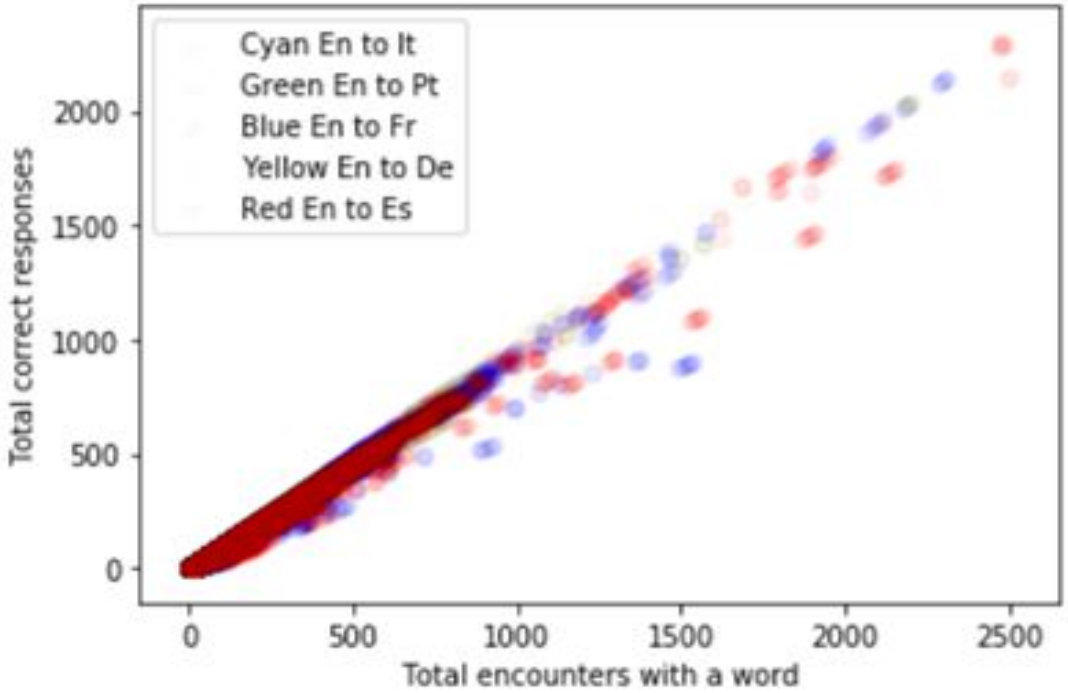
English learners from any target language appear to show more persistence (as measured by total encounters)



Duolingo: Native English Users

Do native English speakers show differences with different target languages?

The resulting plot does not obviously show any differences. Let's use an ANOVA and Tukey HSD



Duolingo: Native English Users ANOVA, Tukey HSD

ANOVA - English learners show differences with different target languages

The only target languages that did not show differences with Portuguese and Spanish

Multiple Comparison of Means - Tukey HSD, FWER=0.01

group1	group2	meandiff	p-adj	lower	upper	reject
de	es	1.1043	0.001	0.9833	1.2252	True
de	fr	7.4926	0.001	7.3576	7.6275	True
de	it	-1.7342	0.001	-1.9046	-1.5638	True
de	pt	0.8976	0.001	0.6566	1.1386	True
es	fr	6.3883	0.001	6.2773	6.4993	True
es	it	-2.8385	0.001	-2.9906	-2.6864	True
es	pt	-0.2067	0.0269	-0.4352	0.0218	False
fr	it	-9.2268	0.001	-9.3902	-9.0633	True
fr	pt	-6.595	0.001	-6.8311	-6.3588	True
it	pt	2.6318	0.001	2.3737	2.8899	True

Duolingo: Time between practice sessions

Time between sessions may also be important for predicting user intensity

The resulting plot shows parabolic distribution

Time between sessions likely important future modeling