"¿CHE, DE QUÉ PARTE SOS VOS?": EXAMINING LEXICAL DIFFERENCES IN SOUTH AMERICA

VENEZUELA
SURINAME
COLOMBIA
ECUADOR
PERU
BRAZIL
BOLIVIA
PARAGUAY
CHILE
URUGUAY
ARGENTINA

Alexander S Chase

Thesis submitted in Partial Fulfilment of the Requirements for the Degree of Bachelor of Arts in Latin American and Caribbean Studies, with Honors, at Brown University

Thesis Primary Advisor: Prof. Mercedes Vaquero
Second Reader: Dr. Dan Potter

April 15th, 2019

“¿CHE, DE QUÉ PARTE SOS VOS?”: EXAMINING LEXICAL DIFFERENCES IN SOUTH AMERICA

“If you talk to a man in a language he understands, that goes to his head. If you talk to him in his own language, that goes to his heart.”

-Nelson Mandela

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

My interest in Latin American and Caribbean Studies and a senior thesis stems from my sophomore year of high school when I studied in Salamanca, Spain for the year. I had been taught Spanish in the classroom for a few years before then, but Salamanca was where I truly learned the language for the first time. I was taught proper Castilian Spanish at a language school just minutes from the oldest university in Hispanic world. When I returned to the U.S. I continued to study the Spanish language, literature, and history throughout high school and college. In these years, I interacted with many different classmates and acquaintances in Spanish from all over the world, and was amazed by the wide variety of words they used and accents they had. No type of Spanish stuck out to me more than that which I experienced studying abroad Junior year of college in Buenos Aires, Argentina. I thought I knew Spanish well until I arrived in Buenos Aires and realized they practically spoke a different language; different lexicon (‘che boludo’), different phonetics (pronunciation of ‘ll’ and ‘y’ as the English phoneme [ʃ]), and different morphological characteristics (‘vos’ instead of ‘tú’).

This senior thesis is a culmination of my past studies and experiences relating to the Hispanic world. I would like to thank my parents, Thomas and Sarah, for always supporting me and encouraging me to explore my interests. I would like to thank Pablo Montoro Alonso for introducing me to Salamanca and being a mentor throughout high school. I would like to thank María José and Magda who taught me ‘Castellano Rioplatense’. Most importantly, I would like to thank Mercedes Vaquero, Erica Durante, and Dan Potter for their help and guidance on this project over the past semester.

INTRODUCTION

I collapsed into my chair, slamming down a book with the title, "¡Che Boludo! A Gringo's Guide to the Argentines." It was the second day of my semester abroad in Buenos Aires, Argentina and I was already stumped by the complexities of the Argentine language. I thought I was proficient in Spanish, but had spent the previous day bewildered by words I did not recognize; what was a 'palta', what was a 'quilombo', and what was a 'boludo'? I thought this book would be the solution to my struggles, little did I know it was just the beginning of my exploration into the linguistic intricacies of the Argentine language.

For my senior thesis, I aim to analyze in depth the variety of differences in the Spanish language. As a joint computer science concentrator, I am drawn towards the potential of using data science techniques to gain more insight into how language is being used throughout the Spanish speaking world. This led me to my research question of, how can Twitter data be used to analyze and visualize lexical differences in South America?

Linguistics is a vast field that is defined as the scientific study of human language.[1] Because of its complex and wide encompassing nature, linguistics is frequently separated into several more specific disciplines. Some of the major linguistic subfields include phonetics – the study of human speech sounds, phonology – the study of how human speech sounds are utilized in a language, morphology – the study of word structure in a language, and lexicology – the study of words and the relation between words.[2] For my thesis I will be focusing on the subfield in linguistics of lexicology. I am interested in studying the formation, meaning, and use of words that make up the vocabulary of Spanish speakers in South America.

[1] Akmajian, Adrian, Richard A. Demers, Ann K. Farmer, and Robert M. Harnish. *Linguistics: An Introduction to Language and Communication*. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2001.
[2] *Ibid*.

For my project, much of the research I rely on to guide my focus is based on Latin American dialectology. I am using Merriam-Webster's definition of a dialect as "a regional variety of language distinguished by features of vocabulary, grammar, and pronunciation."[3] Past authors such as Lipski[4] and Resnick[5] have explored the landscape of Spanish dialectology referring to dialects as linguistic variations between regions. I will use this notion of dialects in my project to analyze and visualize the differences that exist throughout South America in order to gain a better understanding of the complexities of the Spanish language.

In the first chapter I focus on the background research that I conducted in developing this project. I first talk about other attempts to use data in analyzing variations in English and Spanish dialects. Next, I examine the nine Spanish speaking countries in South America and provide historical context for how Spanish evolved overtime in certain regions. The second chapter goes into depth about the construction and findings of my project. I first talk about the data I collected and the various roadblocks I faced. Next, I discuss the insights gained from the data and dive into the significance of what it is showing. Lastly, I describe the website I created and give context to the visualizations I made.

With the rise of massive amounts of easily accessible data, linguistics is shifting from a social science to a data science. In this project, I aim to apply a data science lens to build upon past research done in Latin American linguistics. There are a few limitations with regards to my data as I am using online methods and not surveying individuals in person. These online methods, however, allow me to collect massive amounts of data in a short period of time, from which meaningful linguistic insights can be gained. The main dataset I will be using is one that I created

[3] Webster, Noah. *The Merriam-Webster Dictionary*. New York: Pocket Books, 1977.
[4] Lipski, John M. *Latin American Spanish*. London: Longman, 2001.
[5] Resnick, Melvyn C. *Phonological Variants and Dialect Identification in Latin American Spanish*. The Hague: Mouton, 1980.

myself and consists of words and phrases from approximately one million Twitter users in South America.

I chose to focus on the nine Spanish speaking countries in South America instead of the entire Spanish speaking world for the scope of this project. Narrowing the focus will allow me to do detailed background research on all nine countries. I plan to discuss the dialect zones of each country in detail and use my data to further explore this prior research. Only looking at the Spanish speaking countries of South America will enable me to focus in on specific cities to conduct in-depth analysis. This project will serve as a credible proof of concept of how Twitter data can be used to analyze and visualize lexical differences.

CHAPTER ONE

HISTORICAL ANALYSIS OF LINGUISTIC VARIATION

Researching linguistic variation, I consulted several past publications that guided the investigation and development of my project. I was inspired by the work that had been done in parallel fields and drew ideas from them. I was also able to make use of resources on how data science methods could be used in the field of linguistics. This research provided a solid foundation from which I built my project.

One of my main inspirations for this project was the *New York Times* dialect quiz by Josh Katz and Wilson Andrews.[6] This quiz is a series of 25 questions about the way people in the U.S. pronounce and refer to things using certain words. At the end of the quiz it displays a map showing what regions of the U.S. the test taker speaks most similar to. This project started in 2002 as a part of a Harvard linguistic project by Bert Vaux.[7] Vaux created and conducted an online survey to analyze the differences in dialects in the U.S. with the goal of eventually forming an Atlas of English Dialects.[8] He was driven by a belief that human beings have a strong need to identify with their own groups and that linguistic variations were a tool used to distinguish from other groups. The survey focused on specific questions that would reveal variations in dialects across different regions. Overtime 350,000 survey responses were collected and in 2013 Josh Katz, the graphics editor for the *New York Times*, put together this data to form a quiz and graphic maps.

The resources of this *New York Times* quiz outweigh those disposable to me, yet I seek to recreate a few aspects that caught my attention. For this project, the *New York Times* collected hundreds of thousands of survey responses targeting specific areas of dialectic interest. Given my time and resource restraints this approach would not be possible to do. With the proliferation of

[6] Katz, Josh, and Wilson Andrews. "How Y'all, Youse and You Guys Talk." *The New York Times*. December 21, 2013. Accessed April 04, 2019. https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html.

[7] Gewertz, Ken. "Standing on Line at the Bubbler with a Hoagie in My Hand: Bert Vaux Maps America's Dialects." *The Harvard Gazette*, December 12, 2002. Accessed April 04, 2019. https://news.harvard.edu/gazette/story/2002/12/standing-on-line-at-the-bubbler-with-a-hoagie-in-my-hand/.

[8] Ibid.

the internet in recent years, however, it is possible to use social media as a tool for capturing insights about the Spanish language.

Looking at how data is being used in linguistics, I found that many online research groups and companies are beginning to analyze users' language in a variety of ways. Mikael Brunila is a researcher who has done past work on extracting geocoded data from Twitter.[9] Brice Russ is another investigator who has used geotagged corpora data to gain insights about regional variations.[10] This past data science work in the field of linguistics helped give focus to the development of my project.

Most of the research that has been done using social media data for linguistic exploration is focused on the English language. Many of these concepts can be applied towards examining the Spanish language through Twitter, and there are a few researchers that have delved into analyzing Spanish Twitter data. One article that I found particularly helpful and relevant to my project was an article by Bruno Gonçalves and David Sánchez titled, "Learning About Spanish Dialects Through Twitter."[11] In this article, they used geographically tagged Twitter messages to extract Spanish lexical variation. By analyzing tweets they were able to visualize broad linguistic differences in the Spanish speaking world.

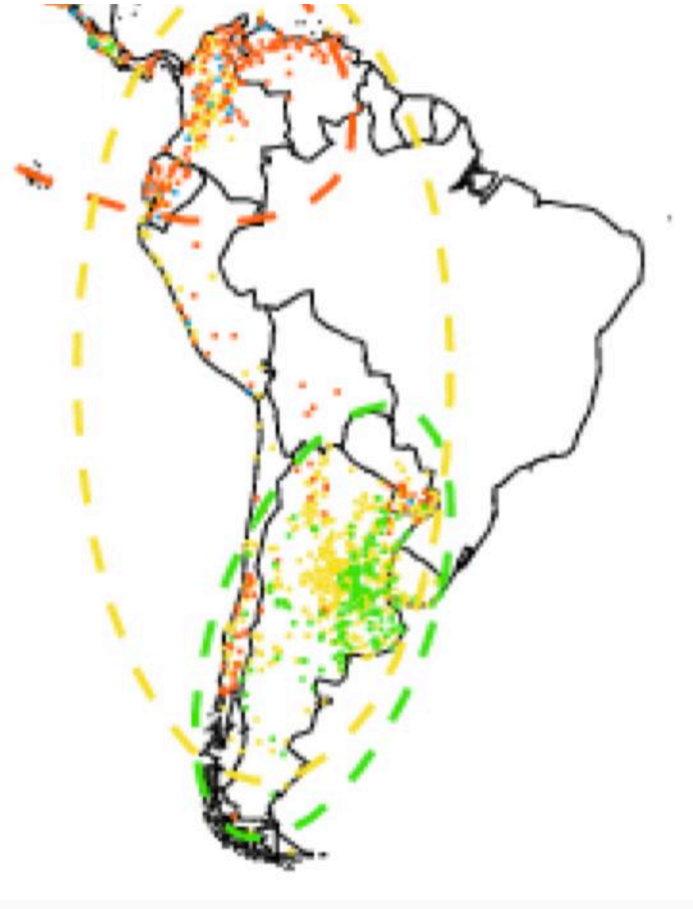
The work of Gonçalves and Sánchez resulted in various interesting insights into the complex relationship between different regions of the Spanish speaking world. They noted three distinct regional dialects which corresponded to 1) the Iberian Peninsula, 2) North America, Central America, and the Northern part of South America, and 3) the Southern Cone. There are

[9] Brunila, Mikael. "Scraping, Extracting and Mapping Geodata from Twitter." April 22, 2017. Accessed April 04, 2019. http://www.mikaelbrunila.fi/2017/03/27/scraping-extracting-mapping-geodata-twitter/.

[10] Russ, Brice. "Examining Large-Scale Regional Variation Through Online Geotagged Corpora." Lecture, 2012 ADS Annual Meeting, The Ohio State University. Accessed April 04, 2019. http://www.briceruss.com/ADStalk.pdf

[11] Gonçalves, Bruno, and David Sánchez. "Learning About Spanish Dialects Through Twitter." *Revista Internacional de Lingüística Iberoamericana* 16 (2016): 1-16. Accessed April 04, 2019. https://arxiv.org/pdf/1511.04970.pdf.

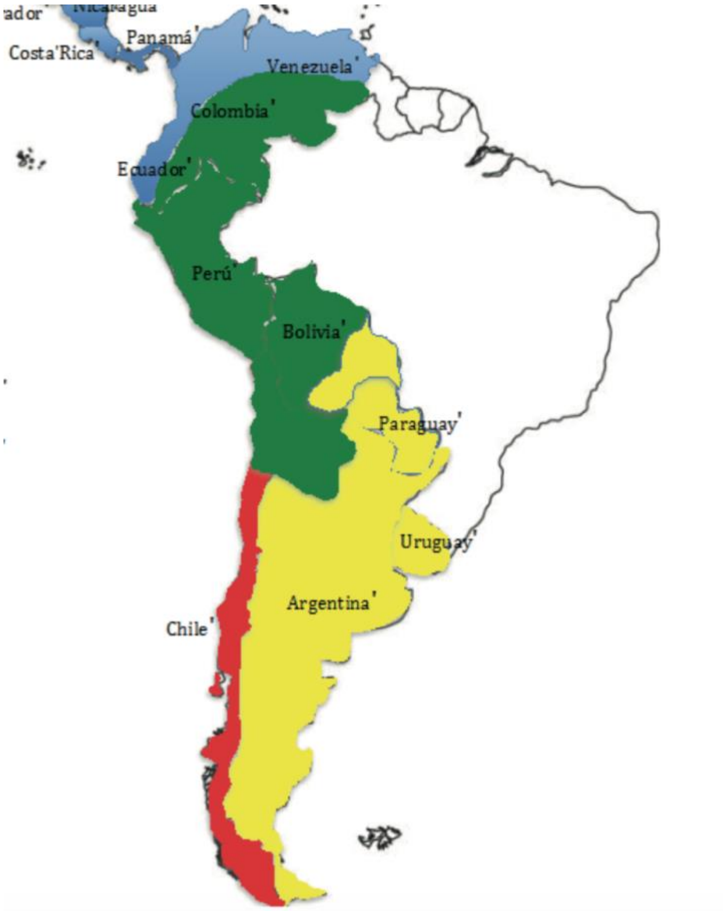
also fainter traces of another regional dialect that is scattered throughout South America with predominance along the Andes mountain range and the interior of Argentina. Furthermore, Chile is one country that does not clearly fit into any of the defined regions suggesting it could be a separate region. The map below shows their findings:[12]



These linguistic discoveries parallel those found by past analysts who have studied South America. This is very interesting and shows that data can be used to analyze and visualize lexical differences. I aim to expand upon the discoveries of Gonçalves and Sánchez to examine more

[12] Gonçalves, Bruno, and David Sánchez. "Learning About Spanish Dialects Through Twitter." *Revista Internacional de Lingüística Iberoamericana* 16 (2016): 1-16. Accessed April 04, 2019. https://arxiv.org/pdf/1511.04970.pdf.

specific regional dialects within South America. Below is one example of a typical dialectic division map:[13]



In thinking more specifically about how different words are used throughout South America I explored the *Corpus del Español NOW* (News on the Web). This is an online corpus developed by Mark Davies from Brigham Young University. His research group is responsible for

[13] Quesada Pacheco, Miguel Ángel. "División dialectal del español de América según sus hablantes: Análisis dialectológico perceptual." *Boletín de Filología* 49 (2014): 257-309. Accessed April 04, 2019. https://scielo.conicyt.cl/pdf/bfilol/v49n2/art_12.pdf.

creating a number of trusted corpora that are widely used in the academic community.[14] The *Corpus del Español NOW* collects information from web-based newspapers and magazines from 2010 to the present and currently contains over 6.2 billion words of data.[15] The corpus collects words from all the Spanish speaking countries in the world by web scraping a large unbiased sample of news articles.[16] In other words, they collect words from online news articles for all Spanish speaking countries and do not rely too heavily on any one news source. This corpus has been used in several academic articles that examine the linguistic makeup of the Spanish language, such as the work of Paloma Sánchez Hernández.[17] For my project this established collection of Spanish language data was important in guiding the trajectory of my work.

Individual Country Dialects

It is necessary to examine in close detail the differences in lexicon between countries in South America in order to create an accurate and engaging project. Spanish is the official language of nine out of the twelve independent countries in South America. The origins of Spanish can be traced back to the expansion of the Latin-speaking Roman Empire.[18] Starting around 1500, the Spanish language expanded globally due to colonial expansion giving rise to a variety of new dialects throughout the Americas. Spanish in Latin America has evolved overtime through

[14] Lindquist, Hans, and Magnus Levin. "Corpus Linguistics." In *Corpus Linguistics and the Description of English*, 1-24. Edinburgh: Edinburgh University Press, 2018. Accessed April 04, 2019. http://www.jstor.org/stable/10.3366/j.ctv7h0vxk.7.

[15] Davies, Mark. *Corpus del Español: NOW*. Accessed April 04, 2019. https://www.corpusdelespanol.org/now/.

[16] Davies, Mark. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English." *Literary and Linguistic Computing* 25 (2011): 447-465. Accessed April 04, 2019. https://corpus.byu.edu/coha/files/davies_llc_2011.pdf

[17] Sánchez Hernández, Paloma. "Sobre la estructura sintagmática de algunos verbos del subcampo semántico: Frecuencia de uso y repercusiones lexicográficas." *Neuphilologische Mitteilungen* 116, no. 2 (2015): 329-352. Accessed April 04, 2019. https://www.jstor.org/stable/26372478.

[18] Boberg, Charles, John A. Nerbonne, and Dominic James Landon Watt. *The Handbook of Dialectology*. Hoboken, NJ: John Wiley & Sons, 2018.

geographical, ethnic, political, and social division to include a wide variation in pronunciation, vocabulary, and syntax. While national boundaries do not always coincide with dialect divisions most linguistic research has focused on a country by country analysis. In this section I will analyze the lexical characteristics that are common in certain countries in South America and the regional divisions within the countries.

Argentina

Argentina is the largest Spanish-speaking nation in terms of surface area and is home to a variety of different regional dialects. The lexicon of Argentina is region dependent, but as the region with the greatest population the Buenos Aires dialect has become the most widely used. Argentine Spanish carries influence from Peninsular Spanish, the large number of Italian immigrants to the region, and 'lunfardo' (a slang developed by the marginalized classes of Buenos Aires and adopted throughout the country).[19] The Spanish of Argentina can be best categorized into seven distinct regions, which have been the subject of extensive linguistic research.[20] 1) The Autonomous City of Buenos Aires and parts of the surrounding province that has developed its own unique dialect, 'porteño'. 2) The coastal region extending from Buenos Aires, Entre Ríos, and Santa Fe to southern Argentina where the influence of 'porteño' has spread and evolved. 3) The western part of Argentina including the cities of Mendoza and San Juan, which share dialect characteristics with Chile. 4) The northwestern portion of Argentina including Tucumán, Salta, and Jujuy that were influenced by Quechua. 5) The northeast part of Argentina including Corrientes, Misiones, and Formosa that were influenced by Guaraní. 6) The central region including Córdoba, which includes aspects of the other linguistic zones. 7) Small enclaves, most notably that of Santiago del Estero and the area bordering Bolivia.

[19] Lipski, John M. *Latin American Spanish*. London: Longman, 2001, p. 175.
[20] *Ibid*. p. 162.

Bolivia

Bolivia is a country that has various dialects, mainly influenced by geography and indigenous communities. The Andes mountain range coupled with the Amazonian lowlands creates a rich diversity of natural geographical barriers that influenced the spread of populations throughout the country. The existence of indigenous populations in the area have also had a large effect on the dialect of Bolivia. Much of the variation in lexical differences stem from the influence of Quechua and Aymara in the highlands and Guaraní and Chiquitana in the eastern lowlands. Research into Bolivian linguistics divide the country into four main regions. 1) The Altiplano highlands including the cities of La Paz, El Alto, and Oruro. 2) The Cochabamba Valley which holds the city of Cochabamba. 3) The 'Llano' lowland region that holds the country's largest city Santa Cruz. 4) The Hispanic Tarija region in southeastern Bolivia bordering Argentina.

Chile

Chile is a country that has large linguistic differences from neighboring countries mainly due to the geographic barrier of the Andes Mountains. Most of the unique features of the Chilean lexicon have their origins with the indigenous populations of the country, of which the Mapuches were the principal group. There is little regional variation within Chile, but is most commonly divided into four linguistic zones. 1) The northern regions of the country with a strong influence from the dialects of the Bolivian highlands and southern Peru. 2) The central region including the largest city, Santiago, and Valparaiso. 3) The southern region which was settled later, is sparsely populated, and brought down much of the linguistic influence of the central region. 4) The Chiloé Archipelago group of islands located off the coast of Chile which has remained isolated linguistically and economically from the remainder of the country, and has a strong influence of the Huilliche dialect of the Mapuche language.[21]

Colombia

Colombia has a wide variety of rich dialects and have been the subject of many linguistic studies. There exist remote areas whose dialects are unknown, as well as the educated speech of Bogotá that has the reputation of being the 'purest' Spanish in South America. Influence in the Spanish of Colombia stems mainly from the large African and indigenous populations. During the colonial period Cartagena de Indias was the major Spanish American port for the importation of slaves and has remained a powerful linguistic and cultural force.[22] There most extreme case of influence is in Palenque de San Basilio where a distinct Afro-Hispanic creole language is spoken.[23] Colombia is most commonly divided into two large dialect zones. 1) The interior highlands with a predominantly Spanish-derived lexicon including large cities such as Bogotá and Medellín. 2) The coastal Caribbean region including the country's principal ports, Cartagena and Barranquilla. A large part of Colombia is made up of the Amazonian region which does not qualify as a unified linguistic region due to the large indigenous populations that have diverse languages.

Ecuador

The Spanish spoken in Ecuador is a combination of the Native American languages, Afro-Hispanic influence, and colonial Spanish heritage. Quechua is the dominant indigenous language in contemporary Ecuador mainly found in the rural regions of the Andes highlands and Amazonian basin. African descendants make up approximately 25% of the population and are concentrated

[21] Cárdenas, Renato, Dante Montiel, and Catherine Hall. *Los chono y los veliche de Chiloé*. Santiago, Chile: Olimpho, 1991.
[22] Lipski, John M. *Latin American Spanish*. London: Longman, 2001, p. 207.
[23] *Ibid*. p. 204.

primarily in the northwestern coastal province of Esmeraldas.[24] Throughout the country there is much regional and social variation making distinct dialects difficult to pin down, but five zones are most predominant. 1) The coastal region including the provinces of Esmeraldas, Guayas, and Los Ríos. 2) Carchi, a province in the extreme north-central part of the country 3) The central highlands including the city of Quito. 4) Loja, a province in the southern part of Ecuador bordering Peru. 5) The Amazonian region with heavy influence from indigenous languages. Lexical differences in Ecuador have arisen primarily by geographical boundaries between distinct groups within the country.

Paraguay

Paraguay is a fascinating linguistic case study as it is the only country in South America where a Native American language is more predominant than Spanish. Guaraní is spoken by most residents of Paraguay and the country is frequently cited as a bilingual nation. Influence from the Guaraní language has been shown to be directly correlated to educational levels.[25] This means that the Spanish spoken by wealthy and well educated Paraguayans is similar to that of the rest of the southern cone, but the speech of those of lower socioeconomic and educational status is largely influenced by Guaraní. Regional variation is also present with three primary zones. 1) The capital district including the major city of Asunción. 2) The central region of the country. 3) the Alto Paraná region bordering Argentina. Overall, the lexicon of Paraguayan Spanish is very similar to that of the Southern Cone with slight differences arising from the influence of Guaraní, socioeconomic factors, and regional location.

[24] *Ibid*. p. 246.

[25] *Ibid*. p. 310.

Peru

Peru is an interesting country to study in terms of South American Spanish. Its capital, Lima, was one of the richest cities in the Spanish Colonies and a coastal city with the port of El Callao just miles away. The 'highland' speech that is often associated with wealthy capitals and 'lowland' speech that is found near port cities combines in Lima. As the Spaniards settled in Peru, Lima quickly became the center of all cultural and economic activity. Peru used to be the center of the Inca empire and a strong mark of the Quechua language still remains. Quechua has brought numerous lexical items to Peruvian Spanish and outside of the major cities the population has retained the indigenous languages. Peru is often examined in four distinct dialect zones. 1) Lima and the central coast which is the most densely populated part of the country. 2) Andean highlands which has a heavy influence from Quechua. 3) The southern coast and southwestern Andean region bordering Chile and Bolivia. 4) The Amazonian lowlands which is inhabited by mostly indigenous and non-native Spanish speakers. The lexical variations are very regionalized by the geography of Peru and the Quechua language.

Uruguay

Uruguay is the smallest Spanish speaking country in South America and over two thirds of the country's population live in Montevideo. Being of proximity to Buenos Aires, many linguists have deemed the speech of Montevideo to simply be an extension of the 'porteño' speech of Argentina. While the dialect of Montevideo is most representative of Uruguayan Spanish, there exist other linguistic zones mainly divided by urban-rural separation. Uruguay can be studied as having three main dialect zones. 1) Montevideo which is the most populated and economically influential part of the country. 2) The interior of the country which is very rural and sparsely populated. 3) The area bordering Brazil where there is a fluid Spanish/Portuguese speech known as ‘fronterizo’. On a general level, Uruguay has many of the linguistic characteristics common in the River Plate region. It uses the ‘voseo’, however, the use of ‘tú’ is also present, mainly in the northern and southeastern areas. The lexicon of Uruguay is nearly identical to that of Buenos Aires with the only slight variations arising from regional slang.

Venezuela

The Spanish of Venezuela has been influenced overtime by demographic shifts of people across the country. The major influences were the import of African slaves, the Native American groups, its proximity to the Caribbean, and the immigration that resulted from the oil industry. Venezuela is most commonly divided into three distinct dialect zones. 1) Caracas, which is the country’s capital and dominates the social and economic scene. 2) The Andean region that share attributes similar to those of highland Colombia. 3) The province of Zulia where the ‘maracucho’ dialect containing ‘voseo’ and some different vocabulary is used. Overall, the Venezuelan lexicon is swayed by the trends of Caracas and has absorbed various indigenisms and a few Africanisms.

CHAPTER TWO
DATA DRIVEN ANALYSIS

The analysis of this project is primarily based on the dataset I collected of approximately one million tweets from Spanish speaking Twitter users in South America. Using Tweepy, a library that allows you to connect to the Twitter API, I opened a stream that collected an unbiased sample of tweets filtered by language and geolocation. In other words, by using Twitter I gathered an inclusive and representative sample of Spanish language tweets from South America. Gathering the location was done through geocoded instances where the latitude and longitude of a tweet were attached to it. Using the geocoded tweets, I could limit my results to South American countries. I also put a language limitation on the data I gathered in order to only collect tweets that were in Spanish. Similar data collection methods have been used before by Gonçalves and Sánchez,[26] and have been shown to be an effective way of examining differences in dialects. Gathering such a large amount of tweets took a lot of time and compute power so I uploaded my code to the Google Cloud Platform in order to collect a large amount of data. I ran into several problems with rate limiting due to the quantity of tweets I was gathering, but worked around it with multiple access keys and sleeping the server. Once I gathered my data I uploaded it to a relational database management system (MySQL) so that I could easily access it for analysis.

Creating a reliable dataset of my own was the first step in better understanding the lexical differences between South American countries. The *Corpus del Español NOW* dataset was a helpful benchmark, but it also came with limitations, mainly geographical categorization and accessibility. The categorization of location in the corpus was limited to country boundaries. As linguists have pointed out, political boundaries between nations do not always directly correspond with dialect zones.[27] With my own dataset I can look more into depth at the geographical location

[26] Gonçalves, Bruno, and David Sánchez. "Learning About Spanish Dialects Through Twitter." *Revista Internacional de Lingüística Iberoamericana* 16 (2016): 1-16. Accessed April 04, 2019. https://arxiv.org/pdf/1511.04970.pdf.
[27] Lipski, John M. *Latin American Spanish*. London: Longman, 2001, p. 153.

to identify patterns in cities. Additionally, having my own dataset allows me to have access to the information necessary to create visualizations that plot the data on a map. Using my Twitter dataset, I first took a deep dive into exploring the lexicon of individual cities to make sense of certain regional variations.

Analysis Of Cities

After researching the unique regional dialects that exist in the Spanish of South America I used my dataset to explore these variations in more detail. The Twitter data is not all encompassing of a regions' dialect, however, it provides an interesting insight into the linguistic trends in a certain location. Looking at past research done with social media data and geolocation identification, I was interested by the concept of Location Indicative Words (LIW). This is the study of words that "implicitly or explicitly encode an association with a particular location."[28] In this section I will be looking at some of the largest cities in South America and exploring the linguistic particularities that they have and how they relate to other cities. My choice of city comparisons is based upon analyzing in more depth the linguistic observations talked about in the previous chapter. The data I will be using for this section was gathered by first identifying users in my dataset that are from a certain city. Since only large cities had a significant number of users, my analysis in this portion is focused around more populous cities. After I collected a list of users from a certain city I looked at their past 400 tweets. By doing so I was able to compile a large database for each city of approximately one million tweets (2500 *users* * 400 *tweets per user* = 1,000,000 tweets). After I collected and loaded these tweets into my database I began exploring various hypotheses.

[28] Han, Bo, Paul Cook, and Timothy Baldwin. "Geolocation Prediction in Social Media Data by Finding Location Indicative Words." Proceedings of 24th International Conference on Computational Linguistics, India, Mumbai. COLING 2012 Organizing Committee. 1045-1062. December 2012. Accessed April 04, 2019. https://www.aclweb.org/anthology/C12-1064.

1) Differences between Bogotá and Buenos Aires

Being two of the largest cities in South America, I am interested in examining the variations that are present between the place that speaks the 'purest' Spanish in South America versus the center of the 'Río de la Plata' region. First, I look at words that appear often in one city, but not in the other. By doing so I can calculate a percentage, which is the difference between the frequencies that a word appears in the cities normalized on the frequency of the word *((freq1 – freq2) / (freq1+freq2)).* I will call this percentage the 'variation percentage' of a word. A higher absolute value of the variation percentage means the word appears much more commonly in one area than another. Note that all words were converted to lower case and had their accents and punctuation removed to make comparisons more accurate.

The words with the greatest variation percentage for Bogotá, Colombia are: ['colombia', 'pues', 'santos', 'quieres', 'vaina', 'millonarios', 'tienes', 'usted', 'metro', …]. There are many more words, but in this section I will focus on the ones with the starkest regional variation. By looking at these words that are used more frequently in Bogotá we can see that they fall into several distinct categories:

- First, we have regional words that refer to features of that region ('colombia' – the country that Bogotá is in, 'santos' – president of Colombia from 2010 to 2018, 'millonarios' – Millonarios F.C., a professional soccer team from Bogotá).
- Second, we see differences in grammar ('quieres', 'tienes', and 'usted' used in Bogotá versus 'querés', 'tenés', and 'vos' used in Buenos Aires).
- Third, we have local words that have different meanings or are much more popular in an area ('pues', 'vaina', 'metro').

The first two categories, regional features and grammar, are important to acknowledge as they highlight aspects of the language that are different. The third category, local words, is the most interesting to examine in detail and from which the subtleties of the Spanish language can be seen. For the remainder of this section I will focus on local words and explore their context. This does not pretend to be an in-depth analysis of every word, rather point out the lexical discrepancies present among two cities in South America. The local words more common in Bogotá are:

- ‘pues’: an informal ending to sentences.
- ‘vaina’: a versatile slang word that is similar to ‘thing’.
- ‘metro’: referring to subway train system as opposed to ‘subte’ in Buenos Aires.
- ‘parce’: Colombian slang for ‘friend’, ‘buddy’, ‘bro’.
- ‘hijueputa’: the most popular Colombian curse word, which means ‘son of a bitch’.

For Buenos Aires, Argentina the words with the greatest variation percentage fell into similar categories of regional words, grammatical differences, and local words. The local words more common in Buenos Aires are:

- ‘onda’ – used in many Argentine expressions, similar to the English word ‘vibe’.
- ‘re’ – used as a substitute for ‘muy’, means ‘very’ or ‘really’.
- ‘lpm’: an acronym for ‘la puta madre’ which directly translates to ‘a whore mother’.
- ‘ahre’: a slang word to let people know that previous phrase was sarcastic.
- ‘che’: an interjection, similar to ‘hey’, used in the River Plate region.
- ‘pibe’: a word for ‘boy’ or ‘youngster’.

Bogotá and Buenos Aires are two very distinct cities with a large amount of geographical separation making the lexical differences very apparent in the Twitter data. Next, I want to examine two cities that are close to each other in geographic proximity to see if the data will show similar differences in language.

2) Differences between Santiago and Mendoza

I chose Santiago, Chile and Mendoza, Argentina as two cities that are geographically very close, but separated by the Andes Mountains and a political border. Looking at the most common used words in each region, the lexicon of Mendoza is more aligned with that of the rest of Argentina whereas Santiago has several distinct words that stick out. Mendoza, being the wine capital of Argentina, has several words that relate to the wine industry such as 'vino' and 'bodega'. The presence of 'voseo'[29] in Mendoza is also apparent, whereas 'tuteo'[30] is more common in Santiago. Regarding local words, the ones more used in Mendoza are:

- 'tranqui': an abbreviation of 'tranquilo', used to describe when things are calm.
- 'novio': a word for boyfriend, in Chile they say 'pololo' instead.
- 'orto': a rude word that stands for a persons' behind.
- 'che': an interjection, similar to 'hey', used in the River Plate region.
- 'pelotudo': used to describe someone who does something stupid or without reason.
- 'inflacion': means 'inflation', this is a big problem in Argentina at the moment.

The words used in Mendoza are predominantly reflective of the rest of the lexicon of Argentina. The main words used in Santiago, however, differ substantially highlighting the differences of the Chilean dialect. They are:

- 'oye': an interjection similar to 'hey'.
- 'pega': colloquial word for 'work' or 'job'.

[29] The use of 'vos' as the second person singular pronoun. (Lipski, John M. *Latin American Spanish*. London: Longman, 2001, p. 13.)

[30] The use of 'tú' as the second person singular pronoun. (*Ibid*.)

- ‘hogar’: a home or residence.
- ‘raja’: versatile word that means ‘really cool’.
- ‘pucha’: implied sentiment of disappointment.
- ‘weon’: slang word for friend, similar to the use of ‘boludo’ in Argentina.
- ‘bacan’: common Chilean expression that means ‘awesome’.
- ‘piscina’: means ‘pool’, ‘pileta’ is used in Argentina.
- ‘fome’: refers to something that you don’t like or is dumb

The richness of the Chilean lexicon can be seen in the local words that are much more commonly used in Santiago. Even though the two cities are very close in geographical proximity the regional dialects are quite distinct.

3) Lexical exploration of Lima

The next city I want to look into is Lima, Peru. This is a city that I know very little about and am interested in seeing how the most common lexicon varies from that of other cities in different countries. There are several political figures and soccer clubs among the most different words, but like above I will focus on the local jargon of the city of Lima which can be found below:

- ‘trafico’: means ‘traffic’, a big problem in the city of Lima.
- ‘pata’: a slang replacement for the word ‘friend’.
- ‘pisco’: a very popular brandy drink made of grapes.
- ‘chamba’: an informal way to say ‘work’ or ‘job’.
- ‘pues’: informal ending to sentences.
- ‘reto’: a challenge or obstacle.
- ‘flaca’: refers to a girlfriend.
- ‘ceviche’: a dish consisting of raw seafood very popular in Peru.

- ‘chela’: a local word for beer.

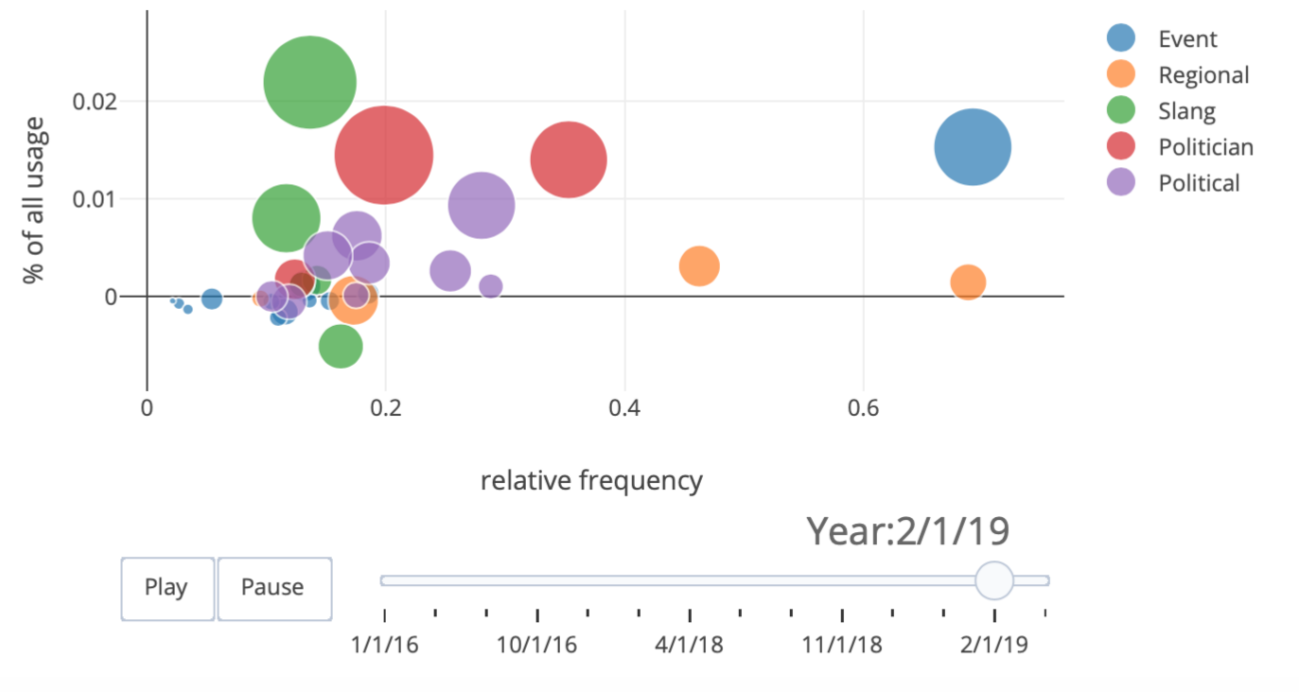
This quick look into the most common words used in Lima gives a glimpse into the slang and the popular cultural factors of the city. The data of all the cities I have looked into contains many interesting differences in the lexicon, but there are also underlying social patterns present. Out of all the cities I analyzed, Caracas, Venezuela has the most interesting Twitter trends.

4) Lexical exploration of Caracas

Looking at the data from Caracas, Venezuela I am fascinated not by the differences in slang, but by the sociopolitical words relating to the current situation in the country. The ongoing humanitarian crisis that is occurring in Venezuela is reflected in the Twitter data I collected, which mostly is from the past several months. I am going to coin the language being used in this situation ‘disaster language’. Disaster language can tell us a great deal about the sociopolitical situation of a country. Some of the most common words being used in Caracas, Venezuela are [‘humanitaria’, ‘guaido’ (Juan Guaidó – President of the National Assembly of Venezuela), ‘regimen’, ‘electrico’, ‘apagon’, ‘militares’, ‘onu’ (United Nations), ‘maduro’ (Nicolás Maduro – disputed President of Venezuela), ‘sinluz’, ‘oposicion’, ‘frontera’, ‘envia’, ‘protesta’, ‘cucuta’ (Cúcuta – Colombian city on Venezuelan border), ‘dictadura’, ‘miseria’, ‘hospitales’, ‘emergencia, ‘medicamentos’].

To gain even more insight into the time horizon of each event I plotted a graph showing the amount a word is used (y-axis) against its relative frequency (x-axis). More specifically, the y-axis represents the number of times a word appears divided by the number of all words used in Caracas. The x-axis measures how much of a certain words’ usage occurs in the given time period. The size of the dots is proportional to the number of uses of that specific word and the colors of the dots correspond to the category the word falls under.

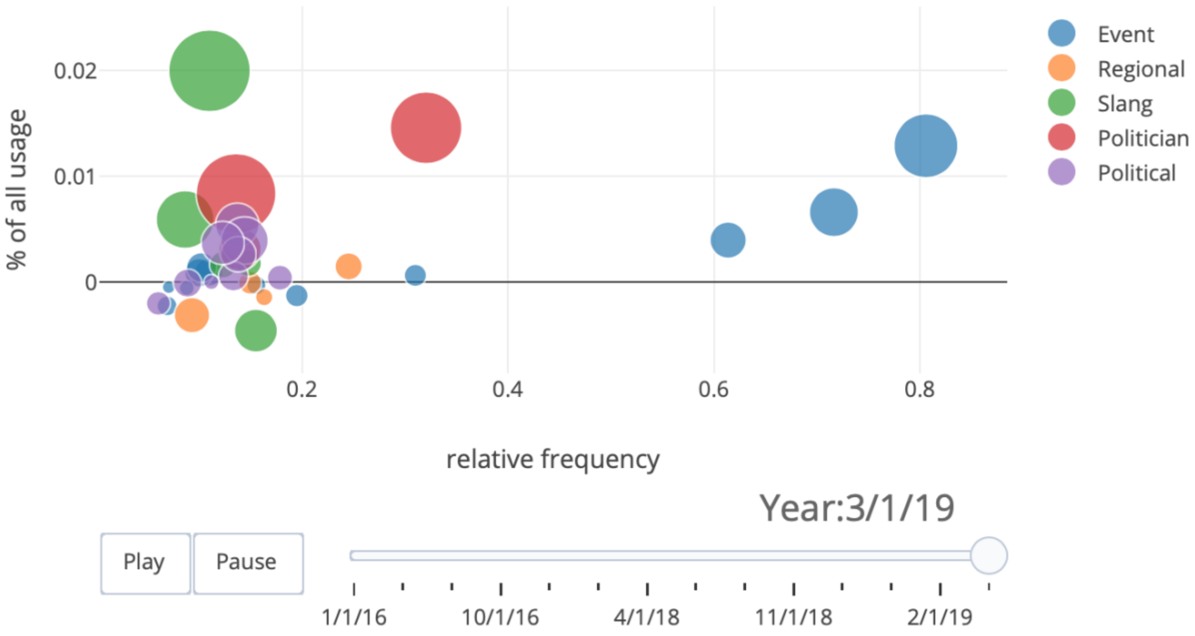
For February 2019, the following graph is shown:



The blue dot and two orange dots furthest to the right on the x-axis are ‘humanitaria’, ‘cucuta’, and ‘frontera’ respectively. Other words that have high relative frequency include ‘guaido’ and ‘transicion’. These words give insight into the situation in February where the leader of the transitionary government, Juan Guaidó, was attempting to bring humanitarian aid through the border city of Cúcuta, Colombia.[31]

[31] Rueda, Manuel. "Some 3 Million Venezuelans Have Fled Their Country - Here's What It's like at Ground Zero for Their Exodus." *Business Insider*. February 05, 2019. Accessed April 04, 2019. https://www.businessinsider.com/ap-ap-explains-cucuta-colombia-the-gateway-into-venezuela-2019-2.

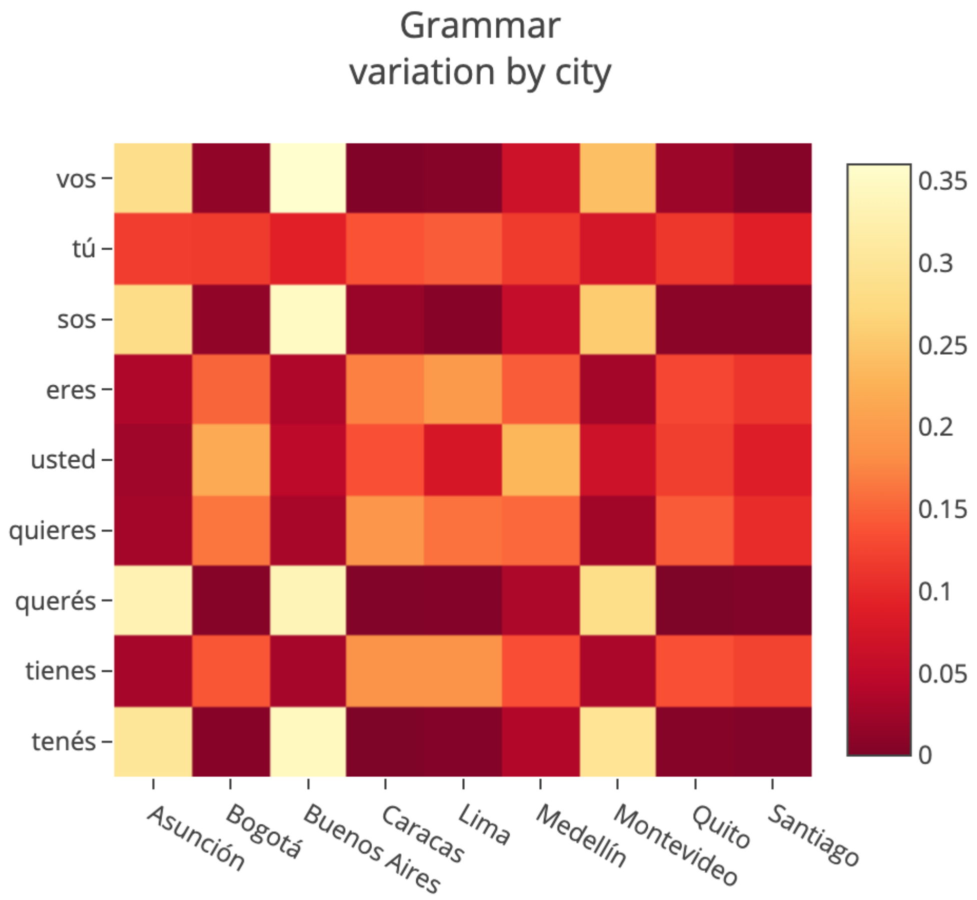
For March 2019, the following graph is shown:



The blue dots furthest to the right on the x-axis are ‘apagon’, ‘electrico’, ‘sinluz’, and ‘hospitales’ respectively. These words describe the power outages that took place across the country on March 7, 2019 and have occurred several more times in the following weeks. The blackouts across the country are causing living conditions to become dire, especially in hospitals where doctors are struggling to obtain the resources necessary to provide treatment.

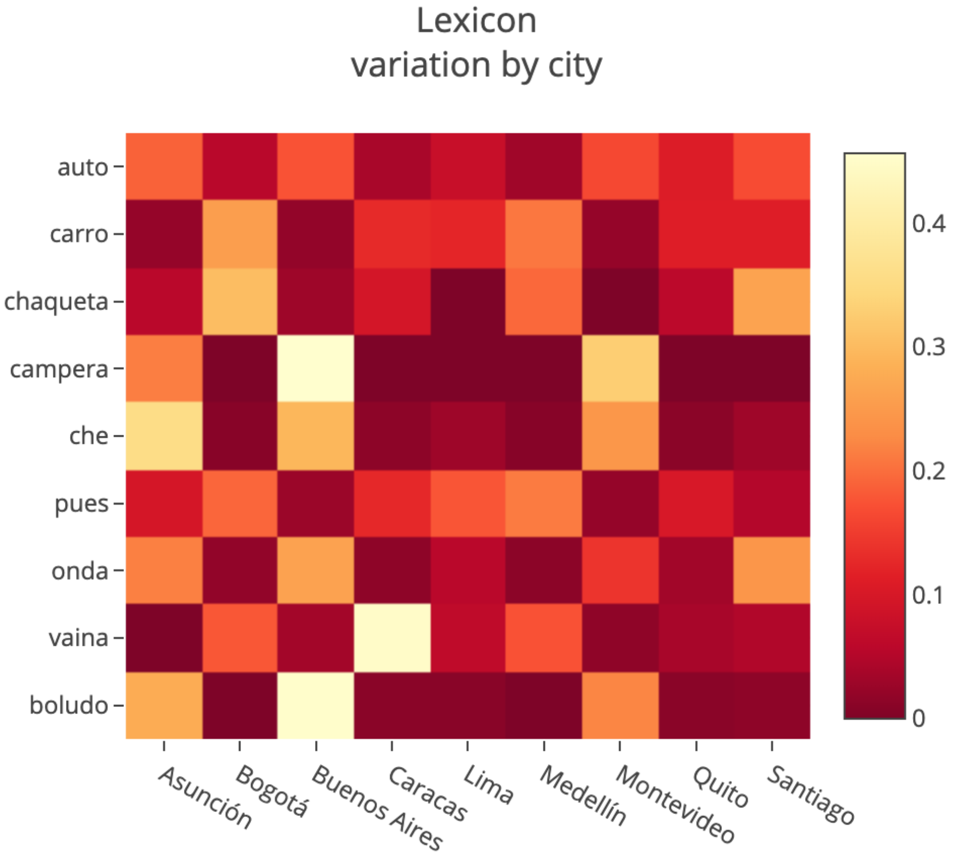
5) Multiple city comparison

Looking at various cities it is interesting to see the biggest lexical variations between them. I am able to explore each city in depth and compile lists of words that are used more frequently in certain cities. With this list of words, I next wanted to use the variation percentage metric I created in order to visualize the largest differences in words. I decided to apply this metric to nine of the top cities in the region to compare the frequency that a word is used. First, I took the top grammatical differences I saw in my initial exploration and plotted them to create a heatmap. A heatmap is a visual depiction of data that uses color-coding to represent different values. In this graphic, a lighter yellow color means the word is used more frequently in a city while a darker red color means the word is used less frequently. In this image, I want to call attention to the impact of 'voseo' present in Argentina, Uruguay, and Paraguay.



I also wanted to visualize the differences in regional words across the major cities in South America. The words I chose stuck out during my research as having significant regional variation.

The heatmap below represents these differences with a lighter yellow color meaning the word is used more frequently in a city and a darker red color meaning the word is used less frequently.



Description Of Website

The project I created is a close examination of the lexical characteristics of regions in South America. The final project is displayed in the form of a website that can be found at https://zanderchase.github.io/linguistics/. This website contains a paragraph of introduction about the project, an interactive word map, visualizations of linguistic differences, and a lexical identification quiz. The two main aspects I want to highlight in this next section, that I have not touched on before, are the interactive map and the lexicon quiz.

The first main part of my website is a visualization that displays a map of South America with the data I collected overlaid on it. Using Carto, an online data mapping library, I was able to align the geocoded tweets with their corresponding region. There are various filters on the map that allow users to analyze the data in more depth. The first dropdown filter is dots vs a heatmap. Each dot on the map represents one user who used the word that is selected in a tweet. The heatmap selection visualizes the data in a slightly different way highlighting the areas where certain words are used more often. The next dropdown allows filtering by selected country for inspection of specific areas. The third dropdown is the word that is being displayed on graph by number of instances it is used. For default I set this word to 'boludo', a common expression used in the River Plate region. I set it to one word as to not overload the map with the data points of every single tweet. I included in this list words that I found distinctly associated with certain regions throughout my research. By selecting different words, you can interact with the map and visually see how certain lexicon is used throughout South America.

The second part of my website I want to make note of is the lexical identification quiz. This quiz takes users through a series of questions and depending on their answers tells them what region of South America their lexicon most closely represents. Right now, the quiz consists of six questions and narrows down a user's location to a single country. I came up with the questions for this quiz from looking at how word usage varied between countries. I picked questions that didn't have too many distinct answers, but enough so that I could slowly narrow down the region. Each of the answers to the first five questions had an answer vector of length nine with weights corresponding to the likelihood that it would be used in that country. The weights were determined from a combination of the frequency that a word appeared in a certain country in the *Corpus del Español NOW* and in my Twitter dataset. For the last question, I had narrowed it down to a certain country and was therefore able to ask a country specific question. The future goal would be to build upon this to eventually be able to pinpoint users to a city.

This website brings together the research that I did for my thesis and presents lexical differences between South American countries in a visual way. By combining computer science with linguistics, I was able to analyze the Spanish language in a unique way. This research is best viewed concurrently with the website to bring together the technological and linguistic aspects of my project.

CONCLUSION

The previous two chapters described the background research that I conducted and the project that I created about linguistic variation in South America. I gathered my own dataset of approximately one million tweets from Twitter users in the continent. Using this original dataset, I delved into the intricacies of the lexicon and explored the particularities of the regional variations. My background investigation guided my inquiries and culminated with a data driven visual project displayed in the form of a website. This thesis fulfilled my goal of enabling users to learn more about the lexicon of South America in a unique and interactive way.

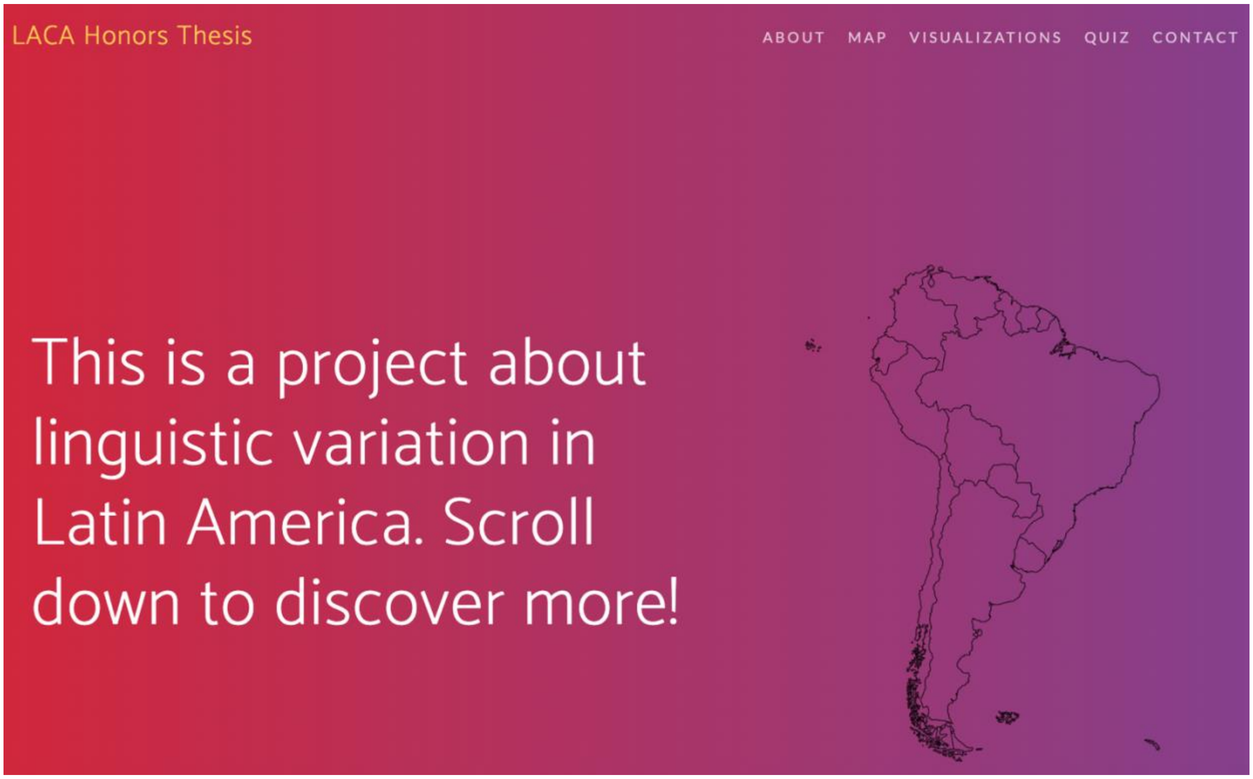
My research built upon that of other linguists who have started to use data to better understand the Spanish language. It showed how Twitter can be used to quickly gather large amounts of natural language data of users in certain locations. By attaching coordinates to where tweets are from, meaningful investigations into the lexicon of a region can be conducted. In my research, I examined the differences between various cities in South America. The data showed clear regional variations in lexicon among cities in geographically distant locations as well as neighboring cities. These insights reflected much of the research I did into the local linguistic characteristics of the country as well as revealed additional findings.

Supplementary to the analysis I did on lexical differences between cities, I also created various visualizations to present the data in a unique way. The data visualization techniques that I used are novel in the field of linguistics and provided an interesting perspective of how to look at the words used in certain regions. By overlaying tweets on a map, viewers could easily explore the location and density of certain words. Furthermore, I used the differences in the frequency of regional words to create visualizations charting the contrasting usage. With the Twitter data I was also able to associate a date with when words were used and examined how language shifted overtime with sociopolitical events. This type of analysis was particularly fascinating in the case of Venezuela where political discourse has dominated the recent headlines.

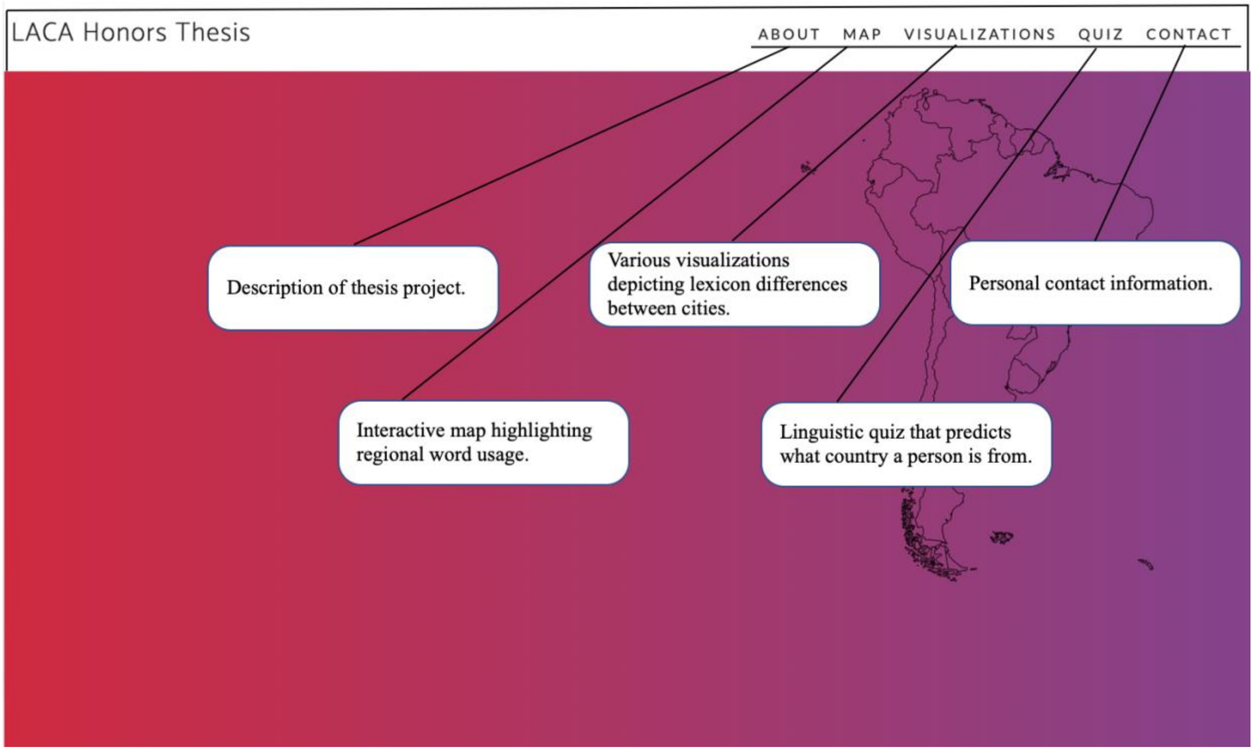
The opportunities for using social media data and computer science techniques in the field of linguistics are vast. Future research in the realm of Spanish linguistic variation could push this project even further. My work serves as a roadmap for the possibilities of using social media data for analyzing Spanish linguistic differences. Future work could build upon Twitter data to encompass other forms of social media such as Facebook, Instagram, and YouTube. By exploring these platforms conclusions could be made about how language is being used and evolving on the internet. It would also be interesting to look more closely at the intricacies of lexicon between cities in the same country and build upon the foundation of my dialect quiz. There are many possibilities for using data to explore language and I hope this project serves as inspiration for future Spanish linguistics research.

APPENDIX

A large part of my thesis is the website that I created. This website can be found at the following URL: https://zanderchase.github.io/linguistics/. This project is interactive and displayed digitally so it cannot be fully contained within this written thesis. In the appendix, however, I wanted to include some images from my website that highlight the design and the data visualized on a map.



This is a project about linguistic variation in Latin America. Scroll down to discover more!



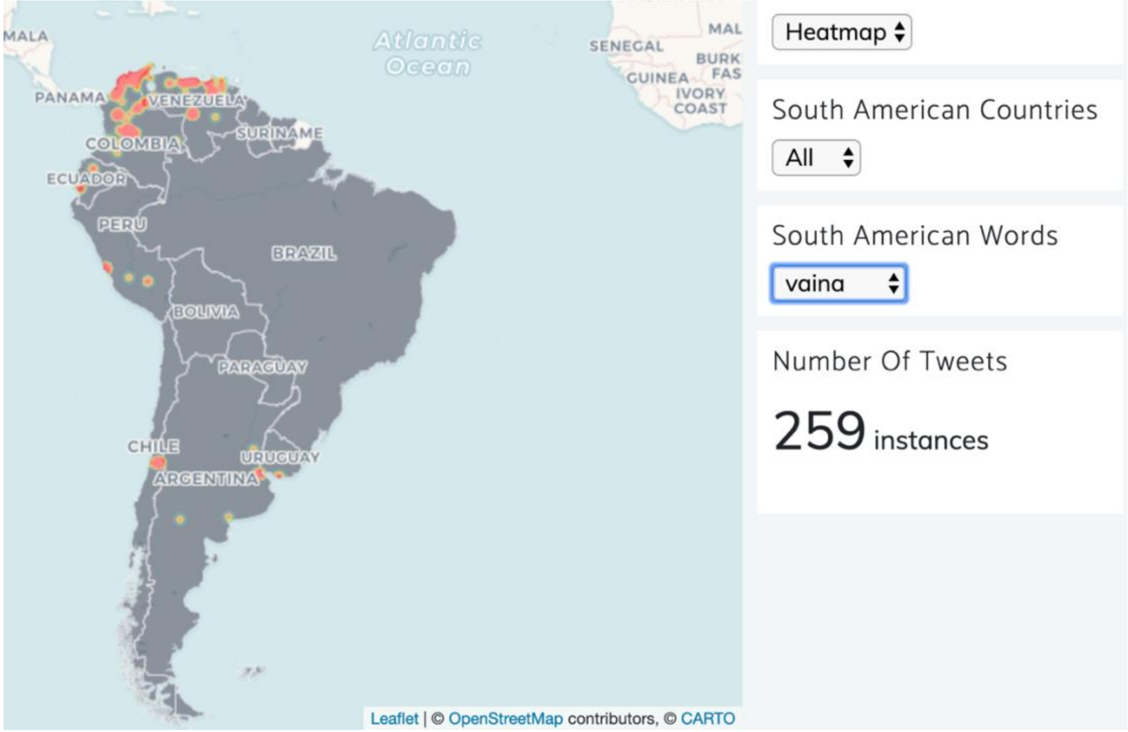
Description of thesis project.

Interactive map highlighting regional word usage.

Various visualizations depicting lexicon differences between cities.

Linguistic quiz that predicts what country a person is from.

Personal contact information.



Heatmap

South American Countries

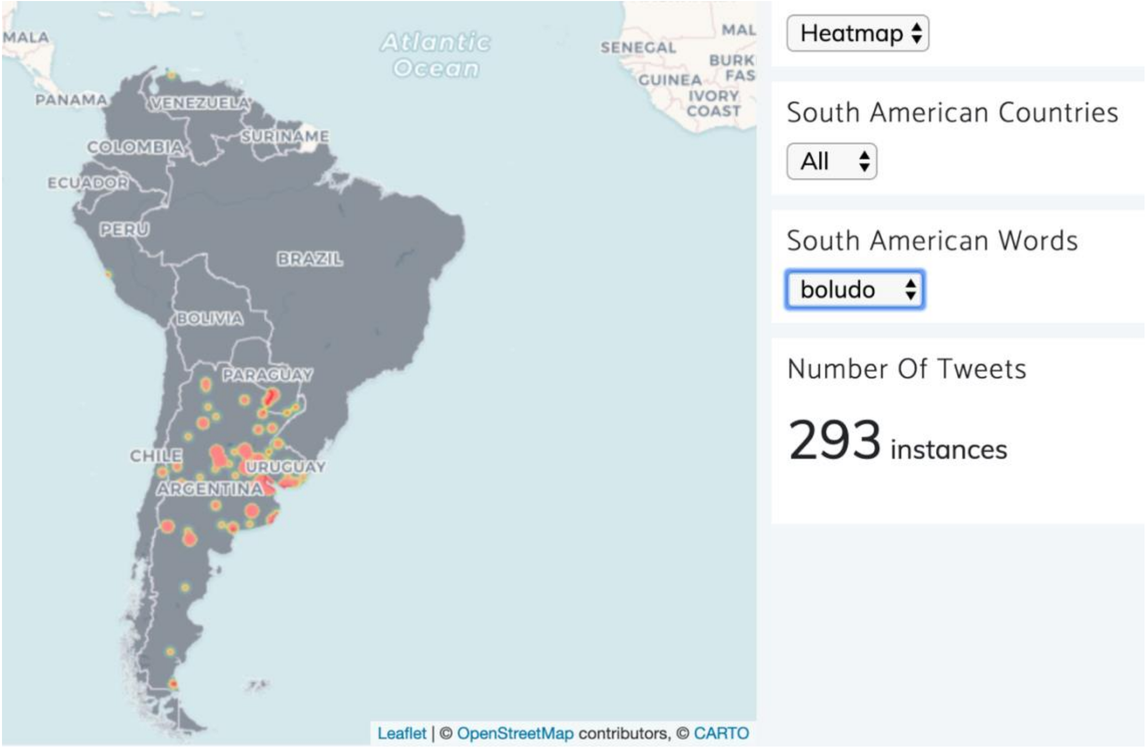
All

South American Words

vaina

Number Of Tweets

259 instances



Heatmap

South American Countries

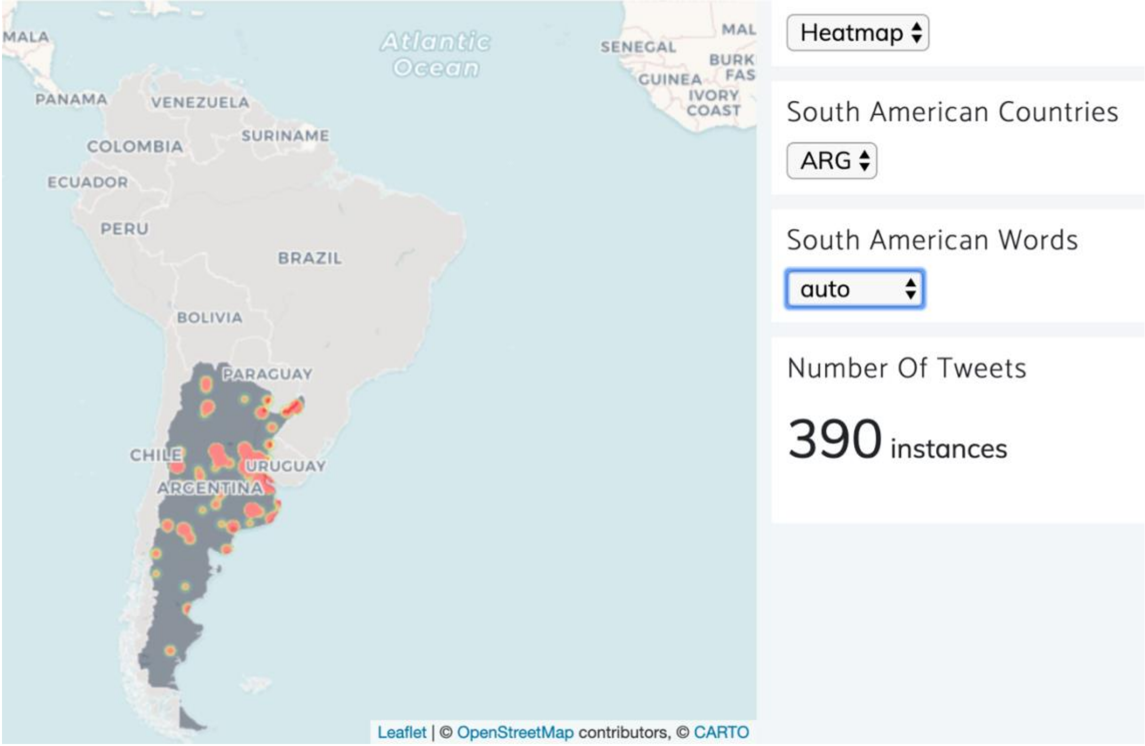
All

South American Words

boludo

Number Of Tweets

293 instances



Heatmap

South American Countries

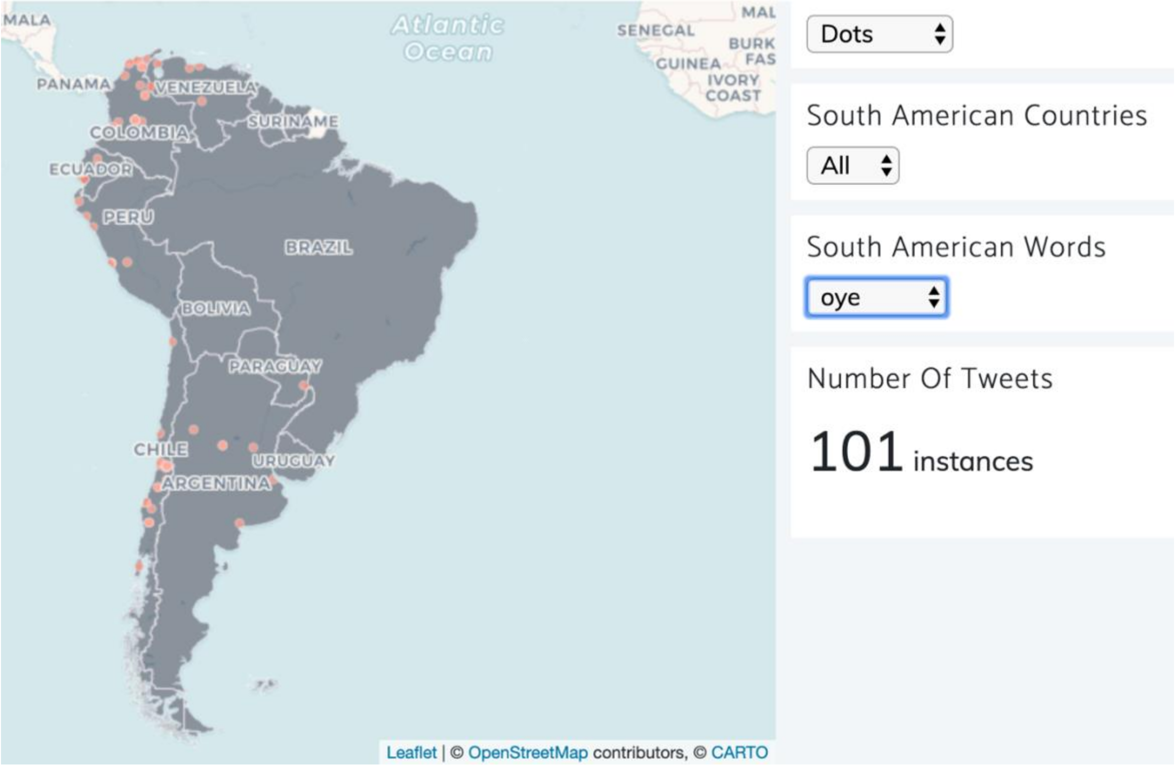
ARG

South American Words

auto

Number Of Tweets

390 instances



Dots

South American Countries

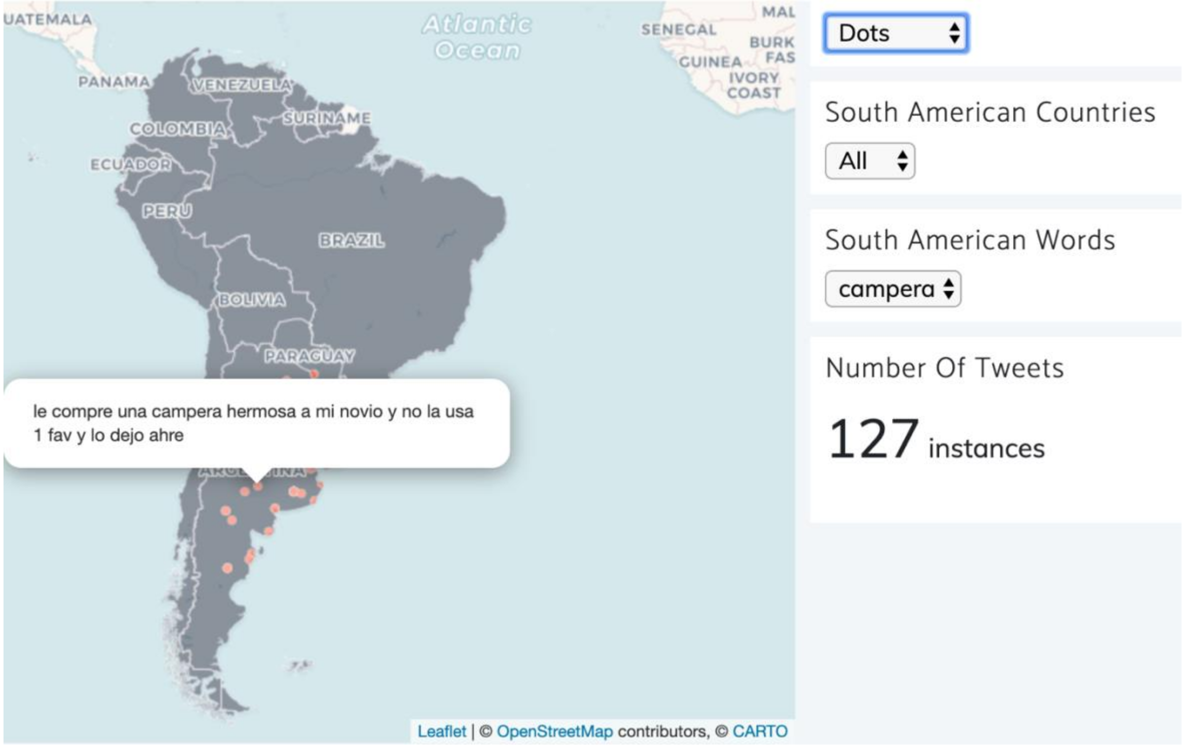
All

South American Words

oye

Number Of Tweets

101 instances



le compre una campera hermosa a mi novio y no la usa
1 fav y lo dejo ahre

Dots

South American Countries

All

South American Words

campera

Number Of Tweets

127 instances

BIBLIOGRAPHY

General References

Akmajian, Adrian, Richard A. Demers, Ann K. Farmer, and Robert M. Harnish. *Linguistics: An Introduction to Language and Communication*. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2001.

Boberg, Charles, John A. Nerbonne, and Dominic James Landon Watt. *The Handbook of Dialectology*. Hoboken, NJ: John Wiley & Sons, 2018.

Cárdenas, Renato, Dante Montiel, and Catherine Hall. *Los chono y los veliche de Chiloé*. Santiago, Chile: Olimpho, 1991.

Coloma, Germán. "La importancia de diez características fonéticas para definir áreas dialectales en español." *Dialectologia* 9 (February 1, 2012): 1-26. Accessed April 04, 2019. https://core.ac.uk/download/pdf/39122643.pdf.

Davies, Mark. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English." *Literary and Linguistic Computing* 25 (2011): 447-465. Accessed April 04, 2019. https://corpus.byu.edu/coha/files/davies_llc_2011.pdf

Gewertz, Ken. "Standing on Line at the Bubbler with a Hoagie in My Hand: Bert Vaux Maps America's Dialects." *The Harvard Gazette*, December 12, 2002. Accessed April 04, 2019. https://news.harvard.edu/gazette/story/2002/12/standing-on-line-at-the-bubbler-with-a-hoagie-in-my-hand/.

Lindquist, Hans, and Magnus Levin. "Corpus Linguistics." In *Corpus Linguistics and the Description of English*, 1-24. Edinburgh: Edinburgh University Press, 2018. Accessed April 04, 2019. http://www.jstor.org/stable/10.3366/j.ctv7h0vxk.7.

Lipski, John M. *Latin American Spanish*. London: Longman, 2001.

Quesada Pacheco, Miguel Ángel. "División dialectal del español de América según sus hablantes: Análisis dialectológico perceptual." *Boletín de Filología* 49 (2014): 257-309. Accessed April 04, 2019. https://scielo.conicyt.cl/pdf/bfilol/v49n2/art_12.pdf.

Resnick, Melvyn C. *Phonological Variants and Dialect Identification in Latin American Spanish*. The Hague: Mouton, 1980.

Rueda, Manuel. "Some 3 Million Venezuelans Have Fled Their Country - Here's What It's like at Ground Zero for Their Exodus." *Business Insider*. February 05, 2019. Accessed April 04, 2019. https://www.businessinsider.com/ap-ap-explains-cucuta-colombia-the-gateway-into-venezuela-2019-2.

Sánchez Hernández, Paloma. "Sobre la estructura sintagmática de algunos verbos del subcampo semántico: Frecuencia de uso y repercusiones lexicográficas." *Neuphilologische Mitteilungen* 116, no. 2 (2015): 329-352. Accessed April 04, 2019. https://www.jstor.org/stable/26372478.

Webster, Noah. *The Merriam-Webster Dictionary*. New York: Pocket Books, 1977.

Data References

Brunila, Mikael. "Scraping, Extracting and Mapping Geodata from Twitter." April 22, 2017. Accessed April 05, 2019. http://www.mikaelbrunila.fi/2017/03/27/scraping-extracting-mapping-geodata-twitter/.

Davies, Mark. *Corpus del Español: NOW*. Accessed April 04, 2019. https://www.corpusdelespanol.org/now/.

Gonçalves, Bruno, and David Sánchez. "Learning About Spanish Dialects Through Twitter." *Revista Internacional de Lingüística Iberoamericana* 16 (2016): 1-16. Accessed April 4, 2019. https://arxiv.org/pdf/1511.04970.pdf.

Han, Bo, Paul Cook, and Timothy Baldwin. "Geolocation Prediction in Social Media Data by Finding Location Indicative Words." Proceedings of 24th International Conference on Computational Linguistics, India, Mumbai. COLING 2012 Organizing Committee. 1045-1062. December 2012. Accessed April 04, 2019. https://www.aclweb.org/anthology/C12-1064.

Katz, Josh, and Wilson Andrews. "How Y'all, Youse and You Guys Talk." *The New York Times*. December 21, 2013. Accessed April 04, 2019. https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html.

Russ, Brice. "Examining Large-Scale Regional Variation Through Online Geotagged Corpora." Lecture, 2012 ADS Annual Meeting, The Ohio State University. Accessed April 04, 2019. http://www.briceruss.com/ADStalk.pdf