Crowdsourcing Dialect Characterization through Twitter

Bruno Gonçalves [1,2,†], David Sánchez[3]

1 Aix-Marseille Université, CNRS, CPT, UMR 7332, 13288 Marseille, France
2 Université de Toulon, CNRS, CPT, UMR 7332, 83957 La Garde, France
3 Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (UIB–CSIC), E-07122 Palma de Mallorca, Spain
† E-mail: Corresponding bgoncalves@gmail.com

Abstract

We perform a large-scale analysis of language diatopic variation using geotagged microblogging datasets. By collecting all Twitter messages written in Spanish over more than two years, we build a corpus from which a carefully selected list of concepts allows us to characterize Spanish varieties on a global scale. A cluster analysis proves the existence of well defined macroregions sharing common lexical properties. Remarkably enough, we find that Spanish language is split into two superdialects, namely, an urban speech used across major American and Spanish citites and a diverse form that encompasses rural areas and small towns. The latter can be further clustered into smaller varieties with a stronger regional character.

1 Introduction

Language is the most characteristic trait of human communication but takes on many heterogeneous forms. Dialects, in particular, are linguistic varieties which differ phonologically, gramatically or lexically in geographically separated regions [1]. However, despite its fundamental importance and many recent developments, the way language varies spatially is still poorly understood.

Traditional methodological approaches in the study of regional dialects are based on interviews and questionnaires administered by a researcher to a small number (typically, a few hundred) of selected speakers known as informants [2]. Based on the answers provided, linguistic atlases are generated that are naturally limited in scope and subject to the particular choice of locations and informants and perhaps not completely free of unwanted influences from the dialectologist. Another approach is the use of mass media corpora which provide a wealth of information on language usage but suffer from the tendency of media and newspapers to use standard norms (the "BBC English" for example) [3] that limits their usefulness for the study of informal local variations.

On the other hand, the recent rise of online social tools has resulted in an unprecedented avalanche of content that is naturally and organically generated by millions or tens of millions of geographically distributed individuals that are likely to speak in vernacular and do not feel constrained to use standard linguistic norms. This, combined with the widespread usage of GPS enabled smartphones to access social media tools provides a unique opportunity to observe how languages are used in everyday life and across vast regions of space.

In this work, we use a large dataset of geolocated Tweets to study local language variations across the world. Similar datasets have recently been used to map public opinion and social behavior [4–11] and to analyze planetary language diversity [12].

Preliminary results demonstrating the feasibility of this approach have thus far been limited to considering only few words or just a few geographical areas [13, 14]. Here, we move beyond the mere proof of concept and provide a detailed global picture of spatial variants for a specific language. For definiteness, we choose Spanish as it is not only one of the most spoken in the world but it has the added advantage of being spatially distributed across several continents [15, 16]. Several other languages such as Mandarin or English have more native speakers or higher supra-regional status but their use is hindered by the limited local availability of Twitter (Mandarin) or a high abundance of homographs that percludes a detailed lexicographic analysis (English).

2 Methods

We used the Twitter gardenhose to gather an unbiased sample of all tweets written in Spanish that contained GPS information over the course of over two years. Language detection was performed using the state of the art Chromium Compact Language Detector [17] software library.

The resulting dataset contained over 5×10^7 geolocated tweets written in Spanish distributed across the world (see Fig. 1). As expected, most tweets are localized in Spain, Spanish America and extensive areas of the United States. These results are consistent with recent sociolinguistic data [18, 19], providing an initial level of validation to our approach. Interestingly, we also find significant contributions from major non-Spanish-speaking cities in Latin America and Western Europe, likely due to considerable population of temporary settlers and tourists. See Ref. [12] for further details and results on this dataset.

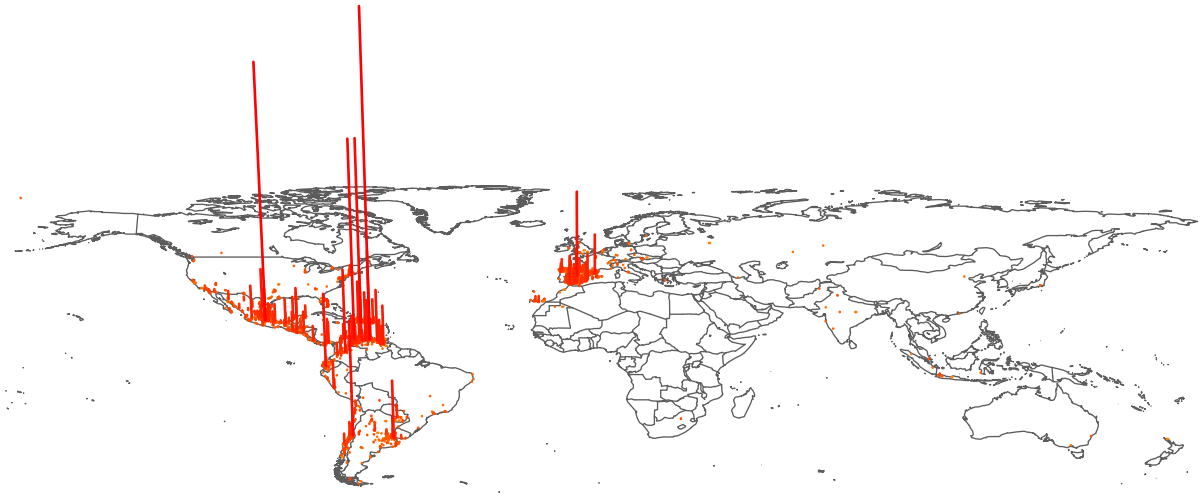

Figure 1. Spanish tweet locations. The overwhelming majority of Spanish tweets are located in Spain and Spanish America but significant contributions arise in certain US states and major Western European and Brazilian cities.

Traditional approaches in dialectology have preferred rural, male informants while modern analyses include interactions with urban speakers regardless of age and gender. On average, Twitter users are young, urban [20] and more likely to be technologically savvy thus providing

more modern perspective on the use of language.

To be able to determine exactly what the major local varieties of Spanish are, we use a list of concepts and utterances selected from an exhaustive study of lexical variants in major Spanish-speaking cities. Reference [21] provides a comprehensive list of possible words representing several concepts, such as "popcorn", "car", "bus", etc. We selected a subset of concepts that minimized possible semantic ambiguities by ensuring that they contained no common words[1].

In our initial set of Tweets we observed 7.5×10^5 geolocated instances where words from our catalogue were used. Individual instances were then agregated geographically into cells of $0.25" \times 0.25"$, which corresponds to an approximate area of 25×25 km^2 in the equator.

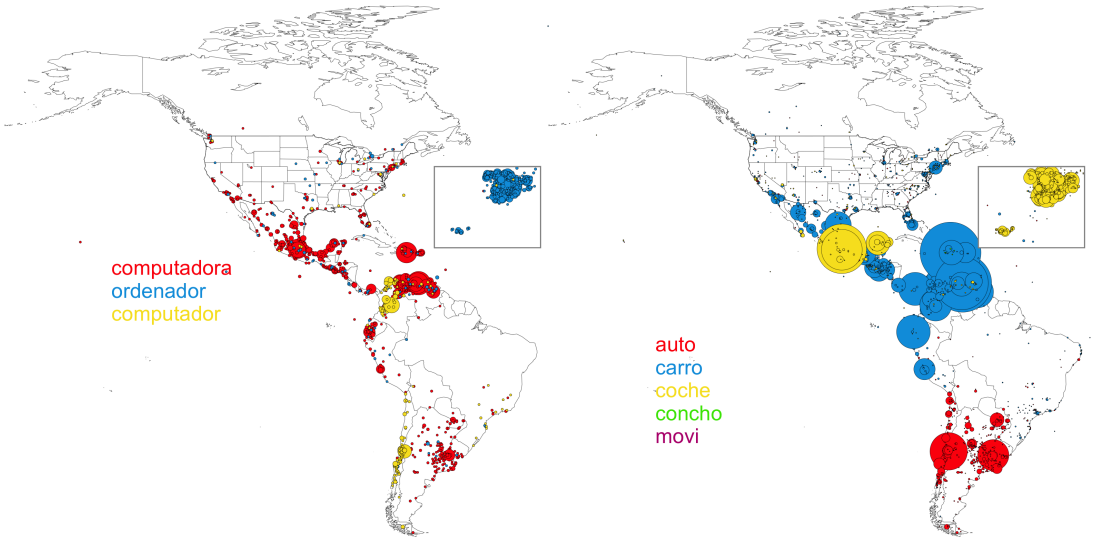

Figure 2. Geographical distribution of the dominant word for the concepts Computer (left) and Car (right). Map locations are colored according to the most common expression found in the corresponding cell. The area of the circle is proportional to the number of tweets

Finally, we define the dominant word for each concept in each geographical cell by a simple majority rule and generate a $M = N_{cells} \times N_{words}$ matrix where element M_{ij} is 1 when word j is the dominant for a given concept in cell i and 0 otherwise. The resulting matrix has $N_{cells} = 1135$ rows and $N_{words} = 131$ columns and constitutes the dataset used for the analysis presented in the remainder of this paper.

3 Results and discussion

Figure 2 illustrates two illustrative concepts ('computer' and 'car') that are both associated to multiple utterances. Each utterance is represented with a different color. We draw a circle centered on each cell with an area proportional to the number of tweets that use the corresponding

[1]The complete list of words for each concept studied can be accessed at `http://www.bgoncalves.com/languages/spanish.html`

expression[2] It is clear from the map that some expressions (*computadora*, *ordernador*, *computador*) are strongly clustered in space, allowing us to easily define regional dialects characterized by the set of dominant words used to express the concepts in our list. Due to the unique resolution of our data we could limit the isoglosses (boundaries) of the regions corresponding to each concept-word with a high degree of precision. However, the isoglosses corresponding to different concepts can overlap and bundle rendering any simple arrangement of dialect areas almost impossible.

The natural way to overcome this difficulty and characterize the various regional dialects present in modern day Spanish is to apply machine learning (ML) approaches to automatically cluster the M matrix and identify which cells are closely related to one another. We start by applying Principal Component Analysis to reduce the dimensionality of the matrix M. PCA determines the linear combinations of the columns (features in ML literature) of the matrix that explain most of the variance observed in the rows (observations). We find that by projecting the data onto the 40 principal components (see Fig. 3) we are able to maintain over 94% of the variance in the data while reducing by 2/3 the dimension of the matrix with clear numerical advantages.

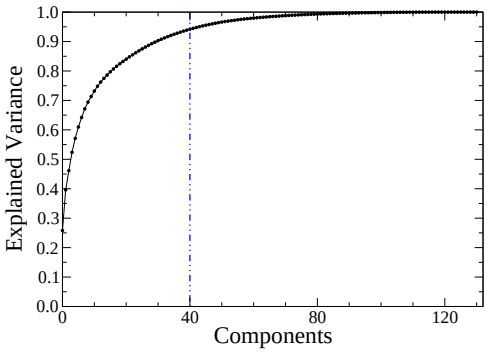

Figure 3. Cumulative variance explained as a function of the number of components With 40 components (vertical blue line) we are able to maintain over 94% of the variance present in the data while significantly reducing the matrix size.

3.1 Superdialects

The task of identifying meaningful clusters in this matrix is now simplified. We proceed by applying the well known K-means [22] algorithm that iteratively refines the position of the centers of K clusters until it finds a stable set of locations. The main dificulty of utilizing this algorithm lies in identifying the correct number K of clusters to utilize. Here, we apply the $f(k)$ metric introduced by Pham *et al.* to establish the best value for K. We run K-means with values of K up to 20 using 100 different random initializations and depict the results in Fig. 4

[2]The corresponding maps for the other concepts in our catalog can be seen at: `http://www.bgoncalves.com/languages/spanish.html`

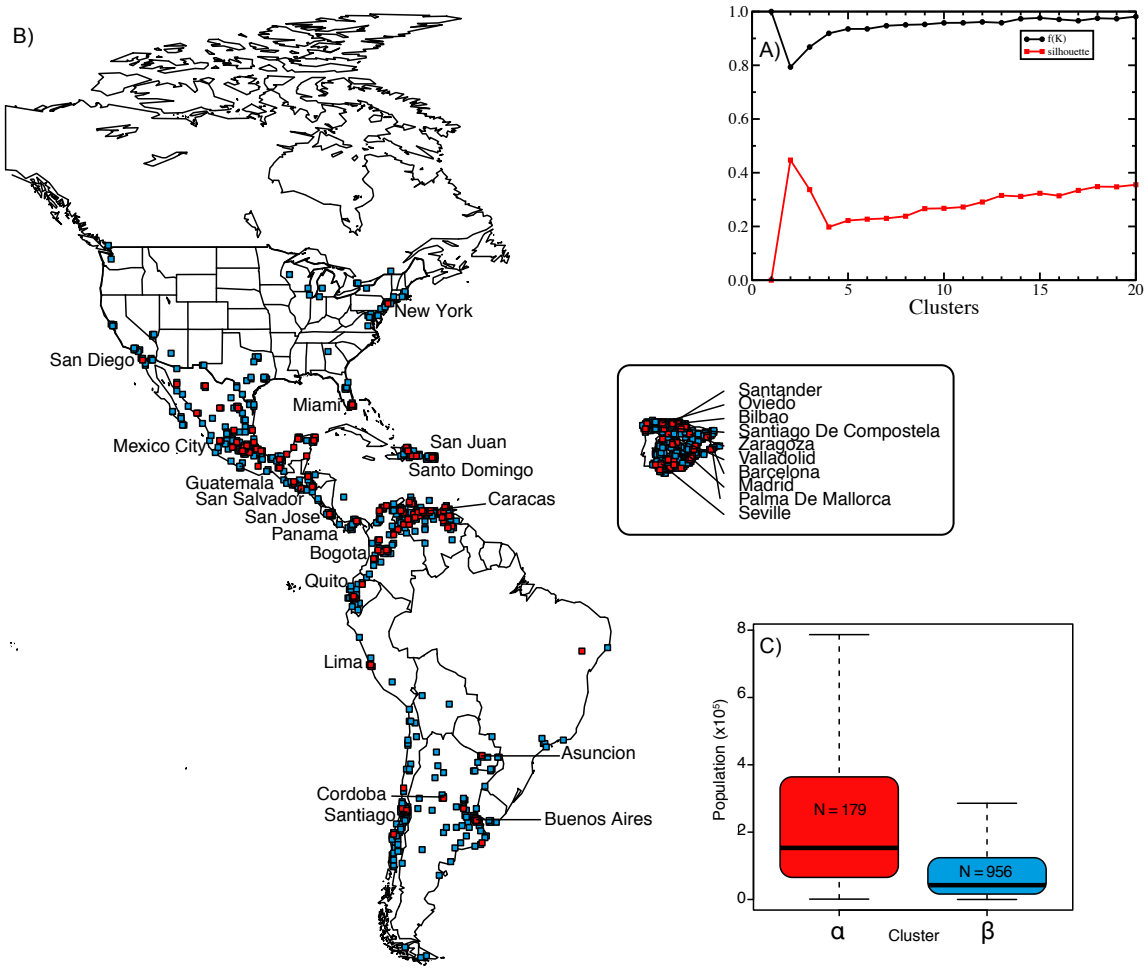

Figure 4. Characterization of the two superdialects A) $f\,(K)$ and silhouette statistics as a function of K. B) Geographical representation of the two clusters, α (red) and β (blue). For visualization purposes we increased the size of each cell. The name of main cities corresponding to superdialect α are shown for clarity. C) Population distribution of the cells corresponding to each cluster.

A). For verification purposes, we also plot the value of the Silhouette [23] of the clusters found with each value of K. Both metrics agree that 2 is the correct number of clusters (both curves show a extremum at that point), leading to two clusters of size 179 (cluster α) and 956 (cluster β), respectively.

A geographic plot of the location of the cells belonging to each clusters (α and β) provides a fundamental clue to their meaning (see Fig. 4 B)). Strikingly, we find a profound correlation between location of cells belonging to cluster α (red dots) and areas of high population density. We validate this idea using estimates of the population living within each cell provided by the LandScan dataset. Hence, we plot the population distribution boxplot for each cluster in Fig. 4 C). The results clearly confirm our intuition. Cluster α corresponds to cells with a typical

population that is significantly larger than cluster β. This suggests a natural lexical bipartition of Spanish into two superdialects. Superdialect α is utilized by speakers in main American and Spanish cities and corresponds to an international variety with a strongly urban component while superdialect β is comprised mostly of rural areas and small towns.

Our result provides some evidence that the increasing globalization of major languages leads to an homogenization that is especially apparent for the active lexicon [25]. Cities (our superdialect α) naturally exert an intrinsic linguistic centripetal force that favors dialect unification, smoothing possible lexical differences. This leveling process present in all countries (thereby its international denomination) is reinforced by the rapid increase of worldwide social ties and the powerful influence of mass media precisely located in important metropolitan areas (Madrid, Mexico City, Miami) [26].Several other sociolinguistic aspects (prestige, higher educational status) also have a role that is more visible in urban environments.

In contrast, rural areas (superdialect β) are generally more conservative and keep a larger number of characteristic lexical items and native words. As a result, the dialectal area corresponding to superdialect β is much more geographically diverse and can be further split, as discussed below.

3.2 Regional dialects

The size imbalance between the two clusters when combined with our intuition suggest that we can also employ the statistical procedure discussed above to further divide the largest cluster (β). We apply K-means recursively until the remaining cluster has a similar size to the previous ones. In the end, we obtain six well defined clusters that we display in Fig. 5. Clearly, three regions can be distinguished. Yellow dots span a wide area covering Mexico, Central America, the Caribbean and north-western areas of South America. Green dots correspond to the Southern Cone while blue dots are almost exclusively accumulated within Spain. The first region is quite diverse. In fact, smaller cells can be aggregated into two additional clusters (depicted with magenta and orange dots in Fig. 5). Interestingly, the magenta and orange dots seem to be localized in the Mexican plateau, the interior of Central America and Andean Colombia, in contrast with the speech of Venezuela, the Antilles and coastal areas represented with yellow dots. This division between highland and lowland varieties agrees with classifications discussed previously in the linguistics literature [27].

The two regions marked in Fig. 5 partly reflect the settlement patterns and the formal colonial Spanish administration within the Empire. Conquerors and settlers occupied first the territories of Mexico, Peru and the Caribbean, and only much later colonists established permanent residence in the Southern Cone, which stayed away from prestigious linguistic norms. This strong cultural heritage that can still be observed, centuries later, in our datasets deserves to be further analysed in future works.

4 Conclusions

Using a large dataset of user generated content in vernacular Spanish, we analyse the diatopic structure of modern day Spanish language at the lexical level. By applying standard machine learning techniques, we find, for the first time, two large Spanish varieties which are related

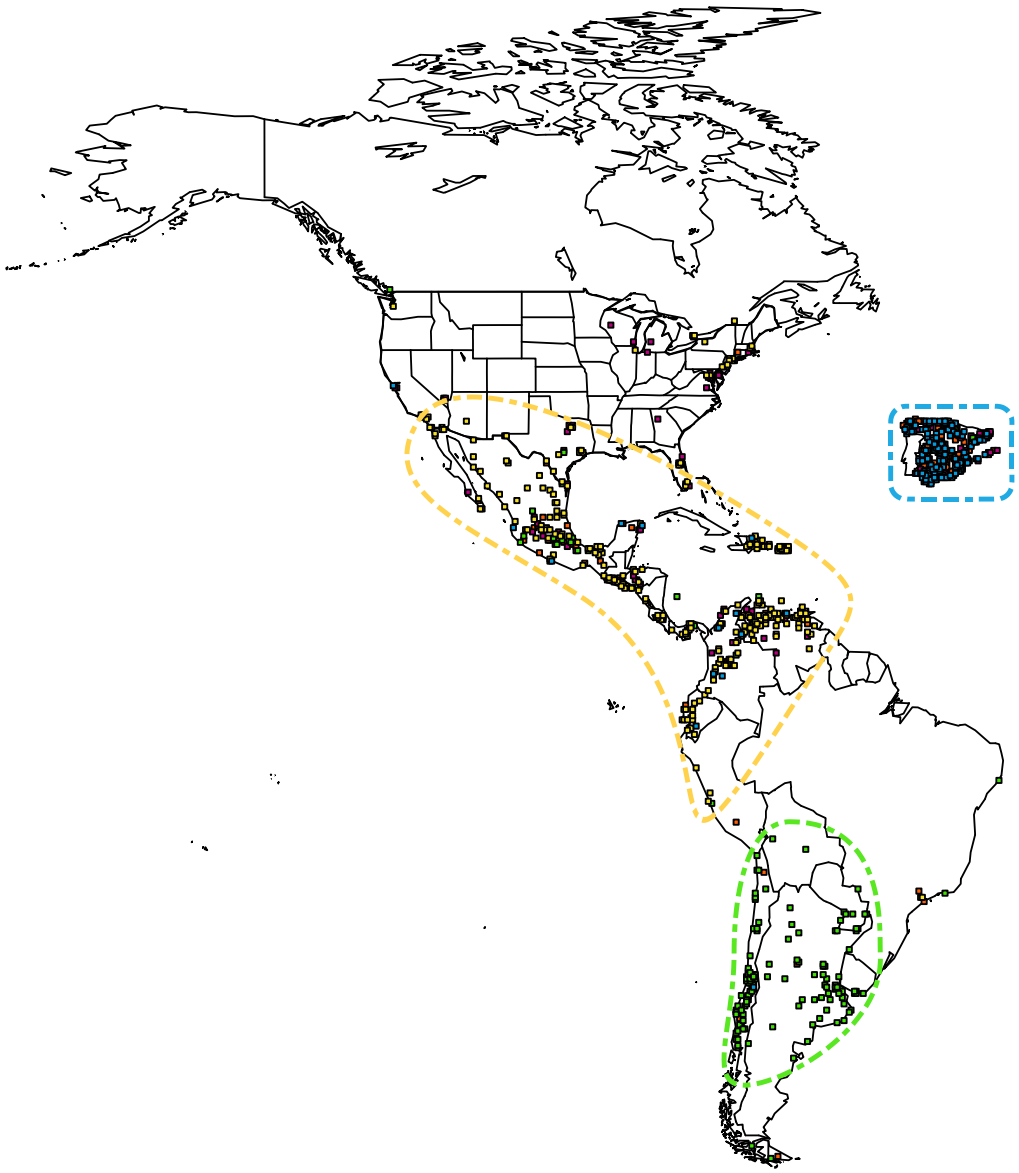

Figure 5. Characterization of major cluster β Geographical representation of regional dialects. For visualization purposes we increased the size of each cell. Three well separated regions are indicated with dashed lines.

to, respectively, international and local speeches. We can also identify regional dialects and their approximate isoglosses. Our results are relevant to empirically understand how languages are used in real life across vastly different geographical regions. We believe that our work has considerable latitude for further applications in the computational study of linguistics, a field full of rewarding opportunities. One can envisage much deeper analyses pointing the way towards new developments in sociolinguistic studies (bilingualism, creole varieties). Our work is based on

a synchronous approach to language. However, the possibilities presented by the combination of large scale online social networks with easily affordable GPS enabled devices are so remarkable that might permit us to observe, for the first time, how diatopic differences arise and develop in time.

5 Acknowledgments

We thank I. Fernández-Ordóñez for useful discussions. This product was made utilizing the LandScan 2007^{TM} High Resolution global Population Data Set copyrighted by UT-Battelle, LLC, operator of Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government has certain rights in this Data Set. Neither UT-BATTELLE, LLC NOR THE UNITED STATES DEPARTMENT OF ENERGY, NOR ANY OF THEIR EMPLOYEES, MAKES ANY WARRANTY, EXPRESS OR IMPLIED, OR ASSUMES ANY LEGAL LIABILITY OR RESPONSIBILITY FOR THE ACCURACY, COMPLETENESS, OR USEFULNESS OF THE DATA SET.

References

1. Chambers J and Trudgill P (1998) Dialectology. Cambridge University Press

2. Labov W, Sharon A, Boberg C (2005) Atlas of North American English. Phonetics, Phonology and Sound Change. De Gruyter

3. Bauer, L (2004) Inferring variation and change from public corpora. In *The handbook of language variation and change*, ed. by Chambers J K, Trudgill P, and Schilling-Estes N, Backwell Publishing, 97:114

4. Borge-Holthoefer J, Rivero A, García I, Cauhé E, Ferrer A, et al. (2011) Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. PLoS One 6: e23883.

5. Tumasjan A, Sprenger T, Sandner P, Welpe I (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media:. 178–185.

6. Culotta A (2010) Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the First Workshop on Social Media Analytics. ACM: 115–122.

7. Salathe M, Khandelwal S (2011) Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. PLoS Computational Biology 7: e1002199.

8. Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, et al. (2012) Digital Epidemiology. PLoS Comput Biol 8: E1002616.

9. Kulshrestha J, Kooti F, Nikravesh A, Gummadi K (2012) Geographic dissection of the twitter network. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM).

10. Mislove A, Lehmann S, Ahn Y, Onnela J, Rosenquist J (2011) Understanding the demographics of twitter users. In: Fifth International AAAI Conference on Weblogs and Social Media.

11. Hong L, Convertino G, Chi E (2011) Language matters in twitter: A large scale study. In: International AAAI Conference on Weblogs and Social Media: 518–521.

12. Mocanu D, Baronchelli A, Perra N, Goncalves B, Zhang Q, Vespignani A (2013) The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLoS ONE 8: e61981

13. Eisenstein J, O'Connor B, Smith N and Xing, E. (2010) A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing: 1277-1287.

14. Russ, B (2012). Examining large-scale regional variation through online geotagged corpora. ADS Annual Meeting

15. (Retrieved Dec. 2012). Languages of the world. Summary by language size. URL `http://www.ethnologue.org/ethno_docs/distribution.asp?by=size`.

16. Penny R (2000) Variation and change in Spanish. Cambridge University Press

17. Candless MM (2012). URL `http://code.google.com/p/chromium-compact-language-detector`.

18. Stewart, M (1999) The Spanish language today. Routledge

19. Moreno Fernández, A and Otero Roth, J (2007) Atlas de la lengua española. Ariel

20. Smith A and Rainie L (2010, December 8) Overview: The people who use Twitter. Retrieved from http://pewinternet.org/Reports/2010/Twitter-Update- 2010/Findings/Overview.aspx.

21. Ueda H and Takagaki T (1993) VARILEX, Variación léxica del español del mundo. University of Tokyo. Available at http://lecture.ecc.u-tokyo.ac.jp/∼cueda/varilex/

22. Pham D T and Dimov S S and Nguyen C D (2005) Selection of K in K-means clustering. Journal of Mechanical Engineering Science 219: 103

23. Rousseeuw P J (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53

24. Landscan 2007 Dataset Retrieved from http://web.ornl.gov/sci/landscan/index.shtml

25. López Morales, H (2005) Últimas investigaciones sobre léxico hispanoamericano: unidad y variedad. In: *Homenaje a José Joaquín Montes Giraldo: estudios de dialectología, lexicografía, lingüística general, etnolingüística e historia cultural*: 333–358.

26. Trudgill, P and Hannah Jean (2002) International English. A guide to the varieties of Standard English. Arnold

27. Cotton, E G and Sharp, J M (1988) Spanish in the Americas. Georgetown University Press