ISSN 2385-4138 (digital)
https://doi.org/10.5565/rev/isogloss.207

Isogloss 2022, 8(1)/9
1-30

Grouping Spanish-speaking countries by dialect: An exploratory corpus dialectometric approach

David Ellingson Eddington
Brigham Young University
eddington@byu.edu

Received: 12-02-2022
Accepted: 28-05-2022
Published: 04-06-2022

How to cite: Eddington, David E. 2022. Grouping Spanish-speaking countries by dialect: An exploratory corpus dialectometric approach. *Isogloss. Open Journal of Romance Linguistics* 8(1)/9, 1-30.
DOI: https://doi.org/10.5565/rev/isogloss.207

Abstract

The present study attempts to cluster Spanish-speaking countries into dialect regions by computational means. The frequencies of 592 lexical and grammatical features for 21 countries were obtained the from Corpus del Español-Web Dialects. Principal components analysis and hierarchical clustering analyses used the resulting data to group countries into dialect regions. A number of algorithms were used to rank features in terms of how much they aided in dialect classification, which allowed grouping based on a smaller set of features.

Six dialect zones were identified: European (Spain), Southern Cone (Uruguay, Argentina), Southern Central America (Costa Rica, Panama), Caribbean (Puerto Rico, Dominican Republic), Northern Central America (Nicaragua, El Salvador, Guatemala, Honduras), Andean South America (Bolivia, Paraguay, Chile, Peru). However, different subsets of features, and different clustering algorithms produced groupings that varied somewhat. The bulk of the variation dealt with where Cuba, Ecuador, Mexico, Venezuela, Colombia, and the US fit into the dialect regions.

The difficulties of the computational approach to dialect classification are discussed. Allowing computer algorithms to determine dialect boundaries appears objective. However, interpreting a principal components analysis entails a degree of subjectivity. Furthermore, the plethora of different classification algorithms allows the researcher to choose the one that produces the desired outcome.

Keywords: dialectometry, Spanish dialects, corpus approach, statistical analysis.

1. Introduction

Early studies in the field of dialectology were carried out by interviewing speakers, extracting features from their speech, and then placing isoglosses on a map to delineate where the boundaries between features existed spatially. Since isoglosses for different features are notorious for not coinciding with each other, except where topographic features such as oceans and mountain ranges are found, determining the exact boundary between dialects was difficult. The use of isoglosses is also problematic in another way since linguistic features are rarely binary, as isoglosses suggest, but scalar in nature. What is more, sociolinguistic studies have uncovered the wealth of variation that exists within the bounds of what may be considered a single dialect.

Delineating dialect areas in the Spanish-speaking world has been the focus of many linguistic investigations (see Rodríguez Vázquez 2019 for a review). The early studies carried out by Armas y Céspedes (1882) and Wagner (1920) were followed by more substantial investigations that divide the Spanish-speaking world into different dialect areas in a number of different ways. Canfield (1962) makes a tripartite division, while Henríquez Ureña (1921) posits five regions. Zamora & Guitart's (1988) division includes nine and Rona (1964) suggests 16. The differences between the dialect boundaries that have been proposed is principally the result of different criteria employed by each researcher. For example, Resnick (1975) bases his on eight phonetic differences. On the one hand, we hope that precise boundaries will be found once enough features have been considered. On the other hand, reducing the complexities of language and language variation to a series of lines on a map can sometimes seem like a futile endeavor (Alba 1992).

In any event, the invention of the internet, the widespread availability of powerful computers, and the existence of large corpora have led to innovative approaches to dialect studies. The most notable characteristic of contemporary approaches is that they do not depend on small numbers of features, but follow the advice of researchers who argue that dialectology must aggregate large number of features to obtain maximally reliable results (e.g. Nerbonne 2009, Séguy 1971). Among these is the use of data from Twitter. Every second 6,000 short messages are broadcast to the world as tweets, and many of these contain geotags that allows their authors to be mapped in space. The sheer amount of language data produced in tweets makes many a linguist feel like a kid in a candy store. Some have taken advantage of the data to delineate dialect boundaries in the US (Huang et al 2016), while others have examined dialect boundaries within a single Spanish-speaking country such as Columbia (Rodríguez-Díaz et al. 2018) and Spain (Aliaga Jiménez 2003, Donoso & Sánchez 2017, García Mouton 1991, Moreno Fernández 1991). More germane to the present paper are studies of tweets in the Spanish-speaking world (e.g. Brown 2016, Gonçalves & Sánchez 2014).

The copious amount of data produced by tweeters requires a systematic way to examine them. One approach is that of Tellez et al. (2021) who take an include-almost-everything approach to their analysis of 800 million tweets, in which they only exclude the 100 most frequent words, and very infrequent words, but retain everything else. Their give their results in terms of scalar similarities rather than setting firm dialect boundaries. Gonçalves & Sánchez (2016) take a more manageable sample of 331 words that represent 46 different concepts taken from the VARILEX project (Tinoco & Ueda 2007). For example, a 'merry-go-round' is known as a *caballitos*, *calesita*, *carrusel*, *tiovivo*, or *machina* in different regions. Their analysis combines countries into three groups: 1) Spain, 2) Uruguay, Argentina, Paraguay, 3) all other countries (see also Moreno Fernández & Ueda 2018).

Another approach to determining dialect boundaries involves data from surveys. For example, Burridge et al. (2019) used the results of the Cambridge Online Survey of World Englishes to map dialect areas. That survey was principally based on vocabulary differences such as different words for traffic circle, tennis shoes, and pill bugs. In a similar vein, the VARILEX database (Tinoco & Ueda 2007) contains 2382 words representing 206 different concepts in Spanish. Among these are words for closet, ring, and suspenders. Speakers from 47 Spanish-speaking cities, principally capital cities, were asked to choose which word they use for each concept. Based on the results of the survey, Ueda (2009) suggests six major dialect areas: 1) Spain, 2) Caribbean: Puerto Rico, Cuba, Dominican Republic 3) Mexico, 4) Central America: Guatemala, Costa Rica, Panama, Honduras, Nicaragua, Colombia, Venezuela, 5) Andean countries: Peru, Bolivia, Ecuador, 6) Southern Cone: Chile, Uruguay, Paraguay, Argentina.

An innovative approach to grouping countries is that of Quesada Pacheco (2014), which involved asking speakers from each country which countries sounded most similar to their own. According to speakers' perceptions eight divisions exist: 1) Cuba, Puerto Rico, Panama, Northern Venezuela, Northern Colombia, 2) Ecuador, Peru, Bolivia, Southern Venezuela, Southern Colombia, 3) Uruguay, Chile, Paraguay, Argentina, 4) Mexico, 5) Guatemala, 6) Honduras, 7) Costa Rica, 8) El Salvador. What makes his findings unusual is the fine-grained differentiation between Central America countries, rather than their conglomeration.

All of these approaches fall into what has been called dialectometry which can be defined as using computational means to study dialectal differences (see Wieling & Nerbonne 2015 for an overview). One subdomain of dialectometry is corpus-based dialectometry, which involves computational analyses of corpora (Szmrecsanyi 2011). One example of using corpora to delineate dialect boundaries is Grieve's study (2012) of letters written to the editor in a number of major cities in the US. In those letters he examined variation in adverb placement (e.g. *often repeated* versus *repeated often*). In another corpus Grieve (2011) studied the variable use of contractions (e.g. *don't* versus *do not*, 2011) in American English. In both studies, and used the results to computationally determine dialect areas.

The first attempts at grouping Spanish-speaking countries according to their linguistic similarities relied on the author's personal experience, reports of dialect features reported by other researchers, and small dialect surveys. However, recent advances in technology have made it possible to expand on previous work by examining much more extensive data sets. For example, a number of large-scale surveys have been carried out and linguistic features have been examined in large collections of tweets. One method that has yet to be applied to the task of associating Spanish-speaking countries into dialect groups is using extant corpora. In the present study, the Corpus del Español-Web Dialects (Davies 2017) is used to this end. In addition to establishing country clusterings another goal of the study is highlighting the features that are most helpful in making those clusterings. In the following sections, a number of statistical and computational methods are described which are applied to the task. The novel use of these methods to investigate the question at hand means that the results can be viewed as exploratory in nature. The challenges that this corpus-based dialectometry approach presents are discussed as well.

2. Data Set

One way subjectivity may creep into an analysis is in the choice of the features used. In order to address this issue a large number of features should be included computational algorithms should be used to choose the most important ones. The analysis described below is based on 592 features and their corpus frequencies (DOI 10.17605/OSF.IO/892MW). These features were chosen since they have been used in previous variationist studies. Of these, 45 come from Eddington (2021). These include six grammatical differences such as the frequency of the use of present perfect versus preterite (e.g. *Esta mañana he comido huevos* 'This morning I've eaten eggs'), the use of present subjunctive in embedded clauses with present tense matrix clauses (e.g. *Le pedí que no lo haga/hiciera* 'I asked him not to do it'), five nominal gender variations (*el/la sarten* 'the pan'), and 34 vocabulary differences (*valija / maleta* 'suitcase'). Also included were the items from VARILEX used by Gonçalves & Sánchez (2014). This consisted of 454 lexical items for 43 concepts (e.g. 'sidewalk' *acera, andén, badén, calzada, contén, escarpa, vereda*). The per million frequencies of these lexical items in each country was obtained from the Corpus de Español-Web Dialects corpus (Davies 2017, extracted data: DOI 10.17605/OSF.IO/892MW). This 2 billion word corpus comprises from 24 to 440 million words per country, 78% of which are from Latin American sources. Some of the lexical variants from VARILEX had a frequency of zero in the Corpus del Español. This meant that they were not helpful for the purposes of the present study and were not included the analysis. In addition to these words, a handful of other lexical items were added as well (e.g. *zumo, coger, damasco, giz* 'juice, to get, apricot, chalk') whose per million frequency also came from Davies (2017).

In many ways using the Corpus de Español-Web Dialects serves the purposes of the study well. It contains a sizable amount of data from each country. What is more, it was designed to represent less formal levels of language since 60% of it derives from blogs. There are, however, serious limitations. For example, information about the writers (e.g. age, gender, social class) is not available. The possibility also exists that a writer from one country may be included in the corpus from another country. These are issues that apply to the whole of corpus linguistics. More importantly, level of granularity is based on country boundaries. That means that different varieties in a single country are lumped together. It is hoped that the by country results reported herein may serve as a guide for future research that makes use of subtler geographic distinctions.

3.1. Evaluating different feature subsets

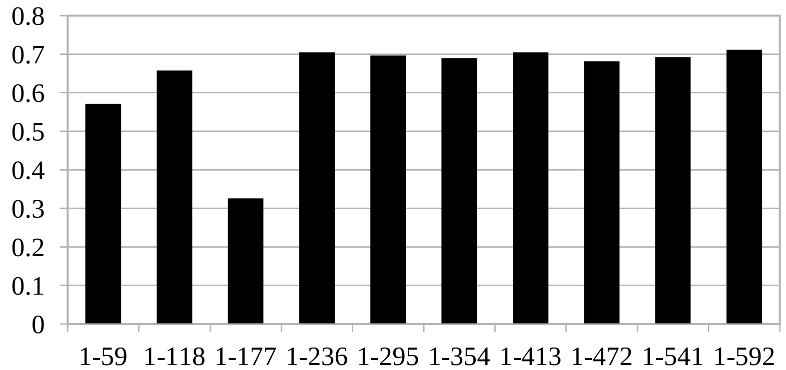
The idea behind agglomerating large numbers of features is to eliminate the influence that a handful of features may have, and also to keep the researcher from picking and choosing only those features that may align with his or her preconceived notions of what countries should be clustered together. However, an important aspect of clustering is determining how many features provide optimal results, which features those are, and and if systematically eliminating features helps the task of classification.

One method of reducing the number of variables to a more manageable size is principal components analysis (PCA). PCA is a commonly used procedure in dialect studies of the kind presented here (e.g. Huang et al. 2016, Manni et al. 2008, Moreno Fernández & Ueda 2018). For a discussion of how PCA compares to other methods see Leino et al. 2008. PCA is an exploratory procedure designed to reduce a large number of variables to a more manageable and easily interpretable format. Rather than eliminating variables altogether, PCA creates new variables from the original variables called principal components. These components retain as much of the original variation in the data as possible. The principal components are ordered so that the first one accounts for more of the variation than the second, and so on. In general, the first two components are the most important.

In order to compare analyses with different subsets of variables, there is a need for a standard of comparison. Geography itself was used in the initial analyses. This was done by calculating the distance in kilometers between each country's capital city as a point of reference. The Euclidean distance between the first two dimensions of each of the principal components analyses (PCA) described below was calculated, and a correlation between the distance between each capital city and the euclidean distance between the two PCA dimensions was performed. The resulting r^2 provides a point of comparison. Analyses with similar r^2 values place the dialects in a similar geographical position in contrast to analyses with different r^2 values.

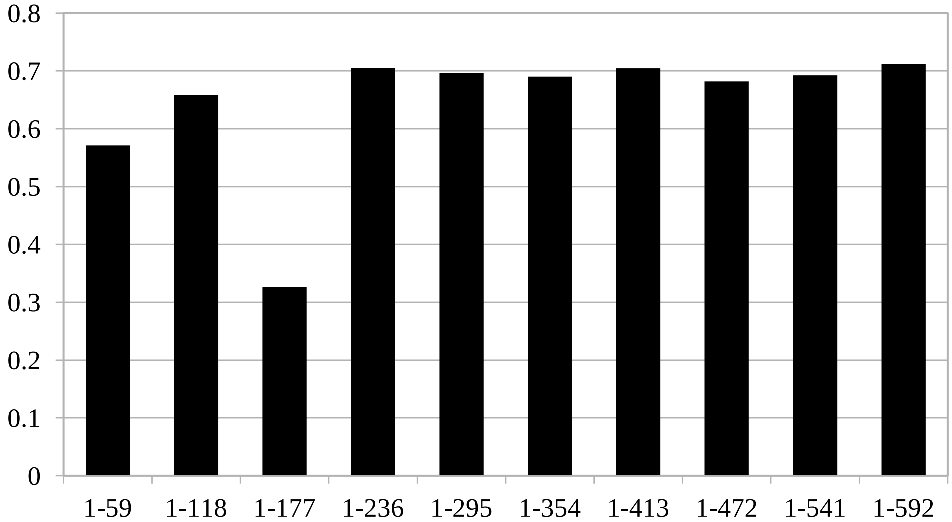
The first question to be examined is how many features are needed to make good dialect groupings. To answer this, the 592 features were ordered randomly and divided into ten groups of about 59 features each. A PCA was run on each group and the resulting r^2 values appear in Figure 1. The wide discrepancies between subsets is a clear indicator that some groups of features situate countries differently with respect to each other. Perhaps this problem is the result of agglomerating too few features.

Figure 1. R^2 between PCA and capital city distances for each subset of 59 features



To address the issue of how many features are needed, the 10 subsets were recombined in the following manner. The first subset comprises the first 1-59 features, the second adds another 59 features and comprises features 1-118. Each subset grows in this manner until the last group includes all 592 features. The results of these analyses appear in Figure 2. What is clear is that once 236 features are included an overlap between the results of the PCA and the distance between capital cities reaches about 70%, a level no subset of 59 features alone reaches. However, increasing the number of features beyond that point does little to change the spatial grouping of the countries in relation to their dialectal features. The clear takeaway is that good predictions may be made with only a subset of the features considered.

Figure 2. R^2 Between PCA and capital city distances for subsets of increasing numbers of features

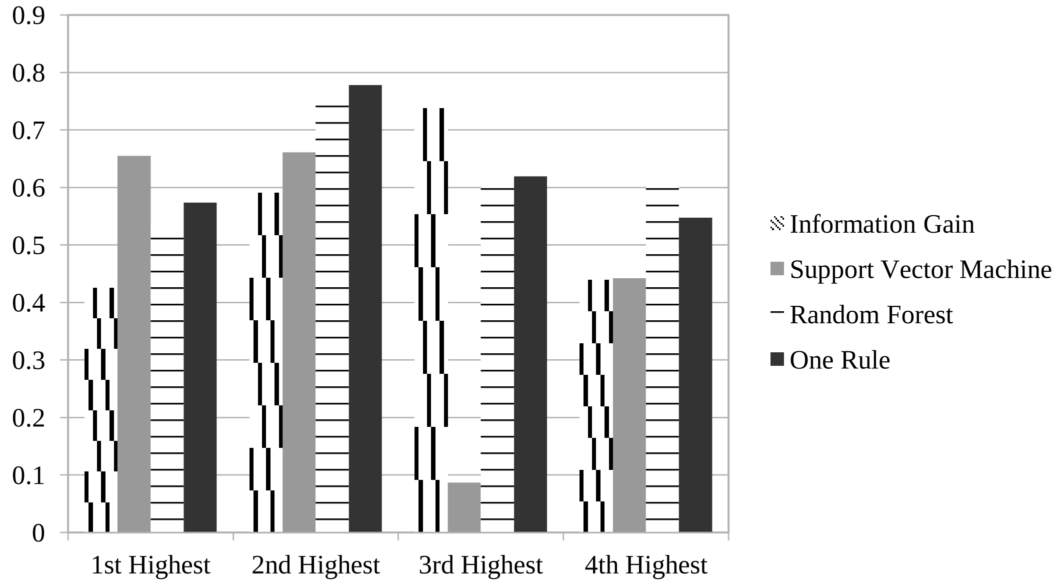


The question now is how to reduce the set of features used by finding the most relevant ones. Given the large numbers of algorithms that one may choose to eliminate variables, it could be tempting to sort through them until one is found that supports the researcher's hypothesis. In order to avoid this, four algorithms were applied to the data, all of which ranked the features in terms of how well they classify the countries. The first was a random forest analysis carried out in R (R Core Team 2020). The remaining three were carried out using WEKA (Frank et al. 2016). Holte's (1993) one rule attribute evaluator determines the worth of each feature in terms of how it much it contributes to classifying countries, and ranks the features accordingly. The information gain attribute evaluator calculates the value of each feature by measuring its information gain in respect to each country. The SVM classifier (Guyon et al. 2002) uses a support vector machine algorithm to determine the value of each feature in classifying countries. The resulting rankings from each algorithm were ordered from best to worst and divided into quartiles of 148 features. The rankings produced by each algorithm appear in the appendix.

As Figure 3 illustrates, the feature ranking algorithms produce different results. What is surprising is that the highest ranked 148 features chosen by three of the algorithms place the country's dialects farther from their capitals than do the second highest set of 148 features. However, this may merely suggest that the features that best group the countries do not correlate as highly with the geographical location of the country's capital cities, since the r^2 is not a measure of goodness of linguistic fit.

Nevertheless, we can now begin to use the data to determine dialect areas, as well as to hone in on the features that best define those dialect areas.

Figure 3. R^2 Between PCA and capital city distances for subsets of ranked features



An initial foray into dialect grouping was done with a hierarchical clustering dendogram (Seol 2020) using the Jamovi statistical package (Jamovi 2021) which is a graphical user interface for R (R Core Team 2020). Hierarchical clustering is a method commonly used in dialect studies (e.g. Leino, Antti, & Saara Hyvönen, Nagy et al. 2006, Moreno Fernández & Ueda 2018, Sato & Hefernan 2018). It clusters countries according to their similarities based on the features. Although we have seen that some features are more important than other in making dialectal groupings, as a point of comparison the dendogram in Figure 4 was built using all 592 features. The algorithm groups the southern cone countries of Uruguay, Paraguay, and Argentina together, and places Spain on its own branch, both of which seem intuitive. However, considering Cuba and Chile as isolated dialects, and combining countries in North, Central, and South America into a single dialect group runs counter to all previous classifications.

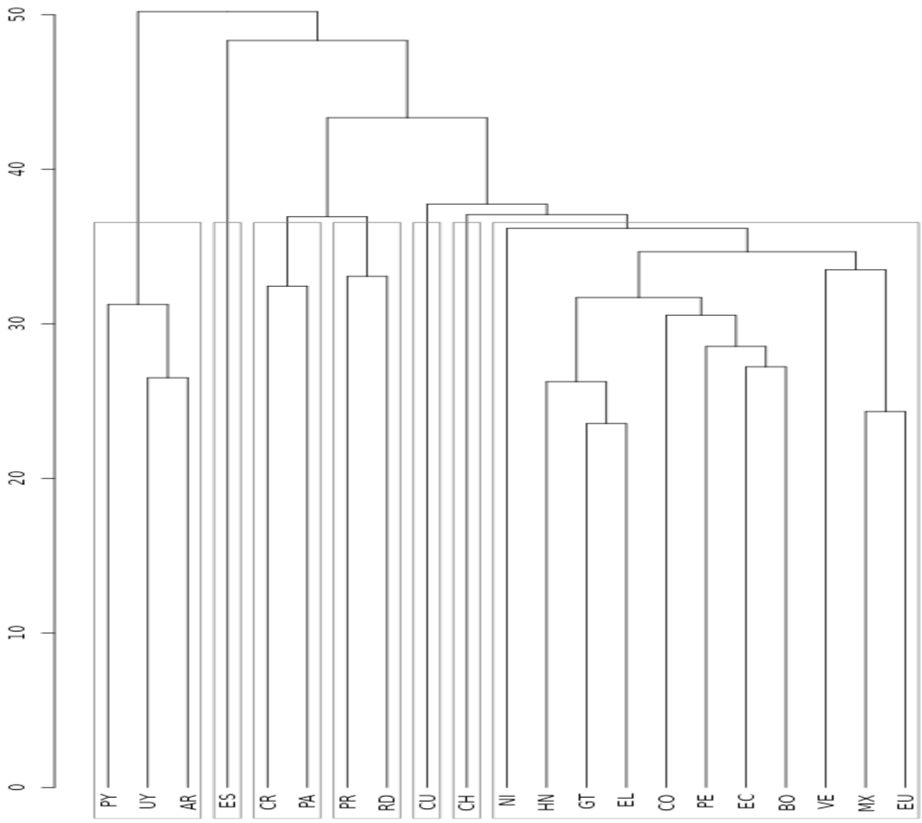
3.2. Evaluating the most highly ranked subset of features

Four algorithms were used to evaluate the worth of the features in terms of their ability to find similarities between countries. The 148 highest ranked features chosen by the support vector machine achieved the greatest overlap (66%) with the geographical location of the capital cities (Figure 3) and is used in this first analysis. When these features are considered a smaller, but tentative, grouping of countries into seven dialect areas emerges (Figure 5):

1. European (Spain)
2. Southern Cone (Uruguay, Argentina)
3. Southern Central America (Costa Rica, Panama)
4. Caribbean (Puerto Rico, Dominican Republic)
5. North America (Cuba, United States, Mexico)
6. Northern Central America (Nicaragua, El Salvador, Guatemala, Honduras)

7. Andean and Northern South America (Bolivia, Paraguay, Chile, Venezuela, Colombia, Ecuador, Peru)

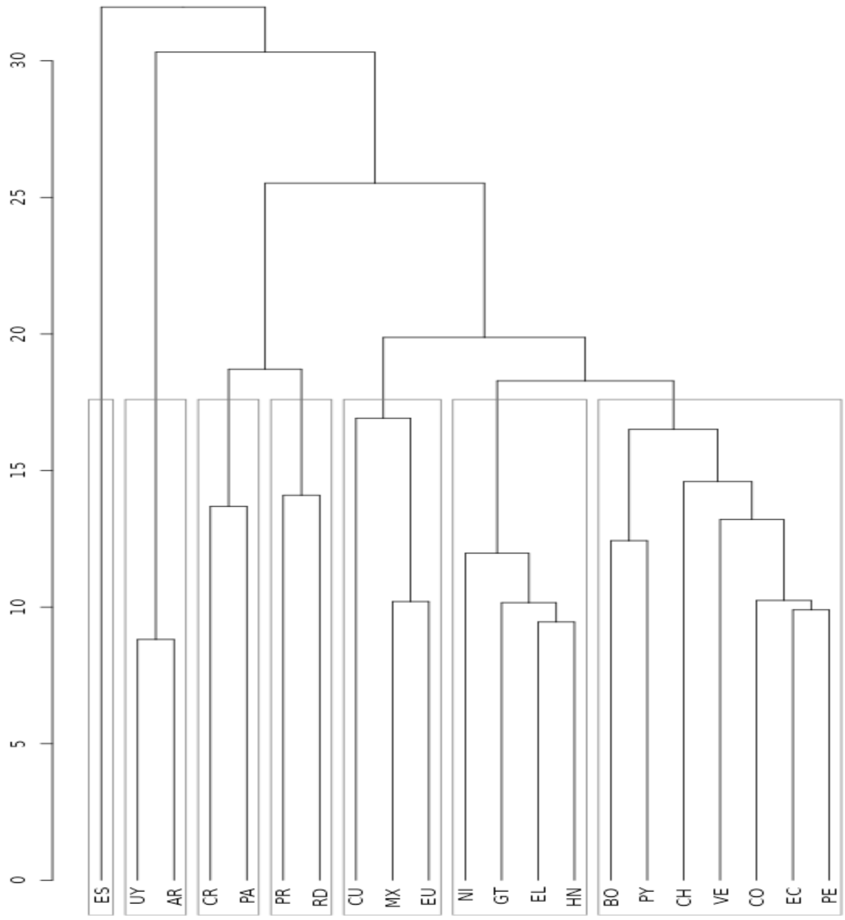
Figure 4. Hierarchical clustering dendogram using all 592 features



PY=Paraguay, UY=Uruguay, AR=Argentina, CR=Costa Rica, PA=Panama, PR=Puerto Rico, RD=Dominican Republic, CU=Cuba, NI=Nicaragua, GT=Guatemala, EL=El Salvador, CO=Colombia, PE=Peru, EC=Ecuador, BO=Bolivia, VE=Venezuela, MX=Mexico, EU=United States

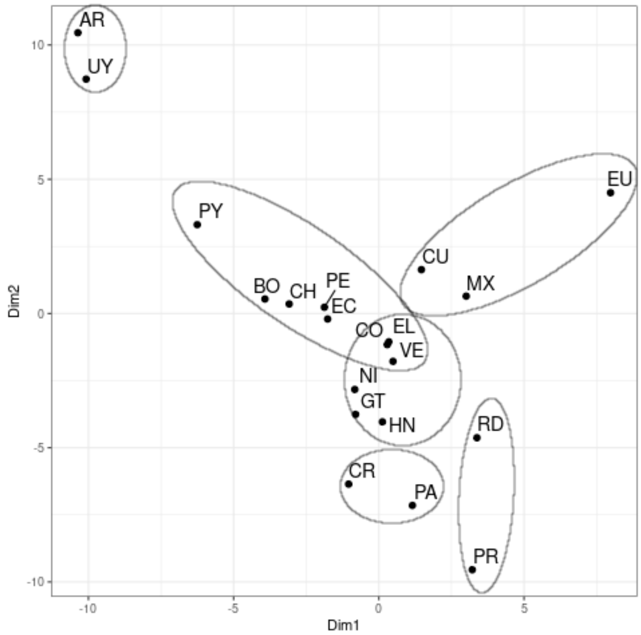
Another way of visualizing these dialect groups is by plotting the first two dimensions of the PCA of the same 148 highest ranked support vector machine features (Figure 6). There are three reasons for using both PCAs and cluster dendograms to evaluate groupings. The first is that when different methods produce similar outcomes, the results are considered more robust, and less algorithm specific. Secondly, while dendograms cluster countries, they do not give a sense of distance between countries and clusters than the two dimensional representation in a PCA provides. Finally, presenting the results of both analyses reduces the chances that a specific method is chosen simply because it yields the expected outcome. In any event, Spain is an outlier falling far from the Latin American countries. For this reason it was excluded from Figure 6 so that the remaining countries would be better separated.

Figure 5. Hierarchical Clustering Dendogram Using the 148 Highest Ranked SVM Features



The dendogram clusterings are indicated in the two dimensional space in Figure 6 with ellipses. It should be apparent that there are significant differences between the dendogram and the PCA plot. For example, the dendogram places Cuba and El Salvador in different groups, while in the PCA plot the two countries nearly overlap. The dendogram's placement of Cuba along with Mexico and the United States, rather than with other Caribbean countries, is unusual. While the three countries are grouped in the hierarchical dendogram, the PCA plot places Cuba closer to other Latin American countries than to the US. Most dialectologists include Cuba alongside the Dominican Republic and Puerto Rico (Canfield, 1962, Henríquez Ureña 1921, Rona 1964, Quesada Pacheo 2014, Zamora & Guitart 1988, Wagner 1920). On the other hand, the similarities between the countries capture the fact that the largest dialectal influences on the Spanish of the US are arguably Mexico and Cuba. The dialectal separation of Uruguay and Argentina from other South American nations has been noted since the 19th century (Armas Céspedes 1882). Paraguay falls closest to these countries, and some dialectologists include Paraguay, or parts of Paraguay, in southern cone varieties (Cahuzac 1980, Canfield, 1962, Henríquez Ureña 1921, Rona 1964, Quesada Pacheo 2014, Zamora & Guitart 1988).

Figure 6. PCA Plot Using the 148 Highest Ranked SVM Features



The dialectal placement of Central American countries is debated. Some cluster them with Mexico (Cahuzac 1980, Henríquez Ureña 1921), while others consider Central America a separate, but united dialect area (Rona 1964, Zamora & Guitart 1988). Still others separate Central American countries into different dialects (Canfield 1962, Quesada Pacheo 2014). The data from the present study support the existence of two dialects in Central America.

What countries Venezuela and Colombia should be aligned with is debated as well. This is principally because both countries are divided between Caribbean and non-Caribbean dialect zones. The similarities between Colombia and the Andean countries (Peru, Ecuador, Chile, Bolivia) that fall out in the present study may be explained by the fact that the majority of Colombians do not live on the Caribbean. As a result, this may skew the data compiled in the Corpus del Español toward non-Caribbean Colombian speech. The same argument, however, does not hold up for Venezuela, where the bulk of the population is concentrated on the Caribbean coast. In Figure 6, Venezuela and Colombia appear in overlapping area of the ellipses representing Northern Central American and Andean and Caribbean South American dialects, falling particularly close to El Salvador. In contrast, the dendogram does not capture the similarity between Venezuela and Colombia and Central American Countries.

The 148 highest ranked features chosen by the support vector machine provides an analyses that more closely corresponds to extant divisions than does the outcome using all 592 features. The question now is exactly how these features are related to the dialect areas, and if it is possible to further eliminate some. To do this, the seven dialect regions proposed were correlated with the 148 features (Table 1).

Table 1. Features Correlated With Dialect Regions

European	North Am.	Northern Central Am.	Southern Central Am.	Caribbean	Andean and Northern South Am.	Southern Cone
-ase past subj.	afiche (-)	auto (-)	acera	acera	arañazo (-)	alpargata
aero	ascensor (-)	autopista (-)	bar	anuncio	ascensor	atar
afiche (-)	bolígrafo	barca	bluyín	aro	bar (-)	auriculares
alubia	capaz	bocadillo	brassiere	autopista	barca (-)	auto
arañazo	colegio (-)	cantina	bus	brasiere	cartelón (-)	cabeza dura
armario	elevador	cartel (-)	cachetes	caldero	chancleta (-)	calesita
ático	habían with plural (-)	finca	cantina	cartelón	colegio	chance
auriculares	magnetófono	gasolinera	carro	cristal delantero	elevador (-)	colchón
balacera (-)	refrigerador	habían with plural	caterpillar	entretenimiento	escuela (-)	estación de servicio
barca	tú	magnetófono (-)	chancleta	escuela	falencia	frigorífico
bolígrafo		mico	escurridor	escurridor	gasolinera (-)	grabadora
cacahuete		quizás (-)	estola	estola	ocasión (-)	guardarropas
camarero		tablero	finca	goma de mascar	vitrina	habían with plural (-)
capaz			frigorífico	guineo		heladera
cazo			grasoso	ocasión		lavadora (-)
celular (-)			jeans	papel encerado		lavarropa
coger			lámina de queso	puerco		maleta (-)
constipado			lancha	salón		ocasión
escuela (-)			lavadora	sarten		parlante
estadía (-)			pasatiempo	ser consciente (-)		pasatiempo (-)
farola			poste de luz	sortija		perchero
gafas			queso americano	tiroteo		pizarrón
grasiento			rebanada de queso			present perfect (-)
guapa			sartén			reposera
lavadora			tiesto			ropero
lois			tirantes			ser consciente
mechero						sos
melocotón						sponsor
mofletes						valija
ordenador						veliz

palomitas						vidriera
pantalón vaquero						
papel de plata						
póster						
present perfect						
quizá						
sandalia						
se los/las decir (-)						
sois						
sujetador						
tiovivo						
tirantes						
vistazo						

(-) = negatively correlated

Of the 148 features, 25[1] were not significantly correlated with any dialect grouping which results in a smaller set of 123 whose features were correlated with at least one dialect zone. Features that are negatively correlated appear in the table with (-). A hierarchical dendogram based on these 123 features makes the same dialect groupings as it did with the 148 features (Figure 5). However, the PCA plot differs somewhat from that in Figure 6. The principal difference in Figure 7 is that Venezuela and Colombia are situated farther from the Central American countries. The analysis to this point has been based on the most highly ranked features determined by the support vector algorithm. The highest overlap between a group of features and geography was obtained by the features in the one rule algorithm's second quartile (Figure 3) so these 148 features were considered. As in the previous analysis, features were removed that did not correlate with any region, which reduced the number of features to 88. The hierarchical dendogram of these features (Figure 8) differs from the previous analysis in a number of different ways. First, the US forms a dialect group of its own rather than clustering with Cuba and Mexico. These two countries now fall into the Caribbean dialect area. In the previous analysis Colombia and Venezuela were placed with the Andean countries, while here they belong to the Caribbean dialect area. The inclusion of Ecuador with Caribbean countries is surprising in this clustering. The PCA plot of the same data appears in Figure 9 where the ellipses are placed according to the groupings made by the dendogram in Figure 8. The large amount of overlap between the clusters exemplifies how differently the PCA and hierarchical dendogram algorithms group the countries even when using exactly the same 88 features.

[1] Amarrar, audífonos, autopista, calzón, cassette, celofán, closet, computadora, distracciones, el sarten, encendedor, engrasado, ensalada de fruta, entrevistar, eres, extrañar, grabador, la calor, mantecoso, mejillas, miradita, nube de polvo, pómulo, seguidor, sostén

Figure 7. PCA Plot Using the 123 Highest Ranked SVM Features that Correlated with the 7 Proposed Dialect Areas

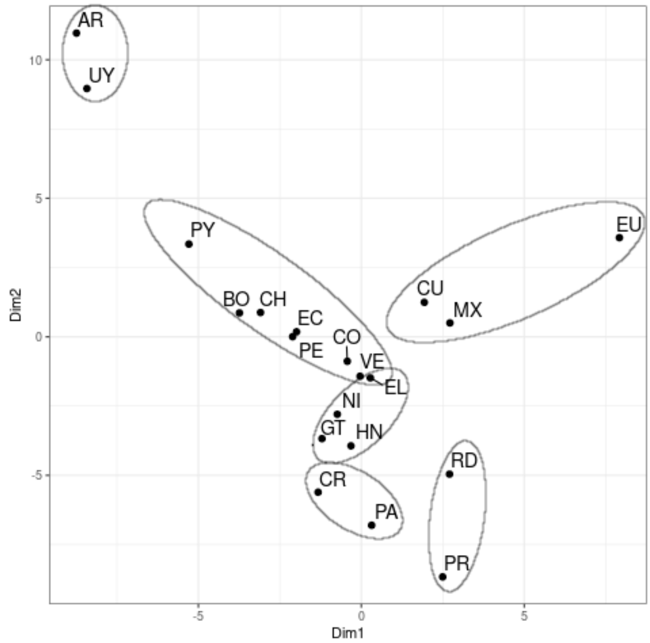


Figure 8. Hierarchical Clustering Dendogram Using the Features from the One Rule Algorithm

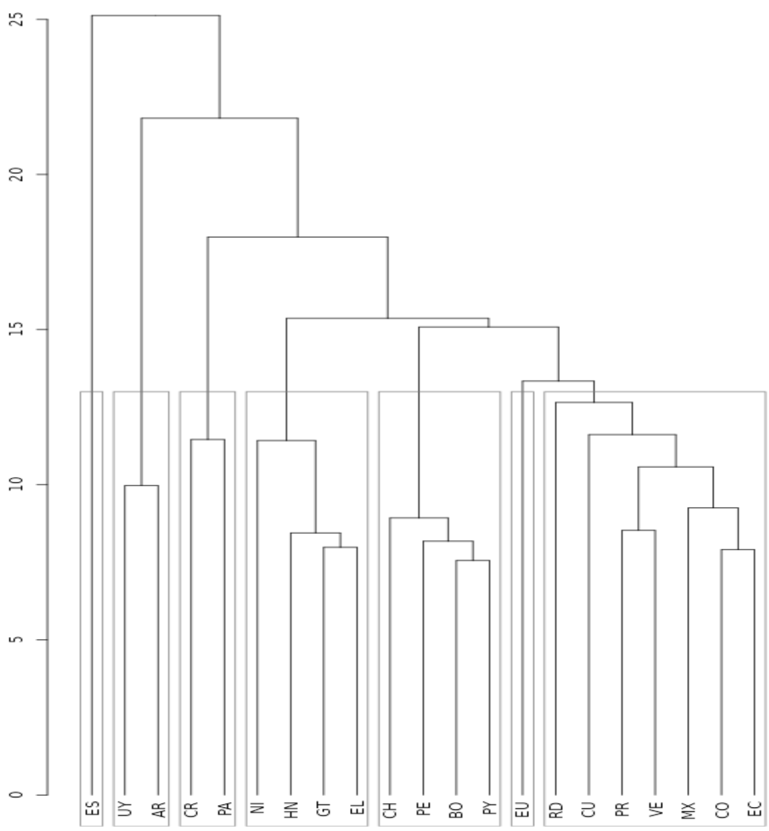
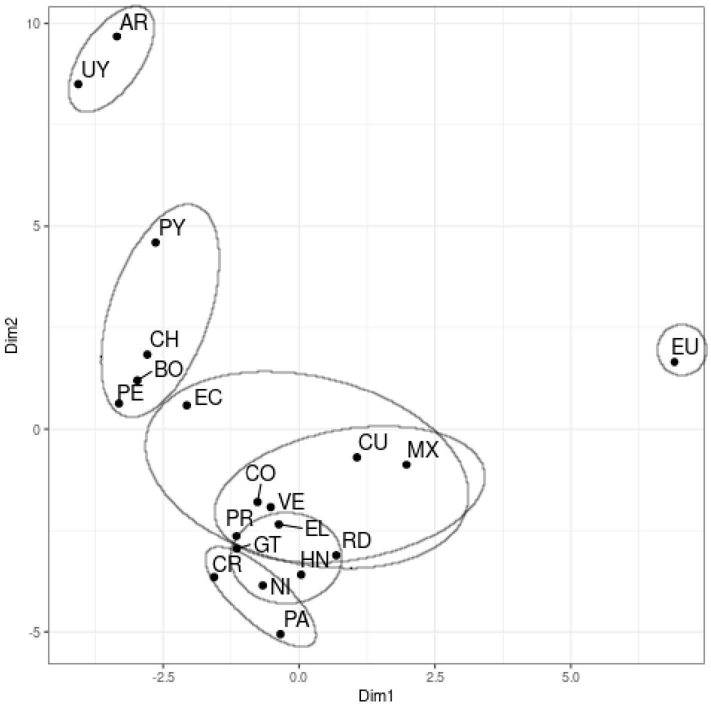


Figure 9. PCA Plot Using the Features from the One Rule Algorithm



4. Conclusions

The present study demonstrates how statistical and data mining methods can be applied to corpus data in order to group Spanish-speaking countries into dialect areas. In early attempts at delineating dialect areas researchers hand-picked features that formed the basis for their groupings. The variety of resulting clusterings may be due to the selection of features chosen. On the surface, using computational means to choose which features, and how many features to include in an analysis, seems to be a more objective method. In reality, it merely shifts the issue from cherry picking features to cherry picking the algorithm used to select the features.

As we saw in the final analysis, the hierarchical clustering dendogram groups countries differently than the PCA plot does even when they both use the same data. An additional problem with PCA plots is that they allow varying interpretations. While some countries form clear clusters situated at a distance from other countries, when the countries are plotted closer together, clustering them into groups with ellipses by hand can become an extremely subjective task.

In this analyses described above, only two clustering algorithms were used to determine dialect boundaries. There are, however, myriads of different methods for clustering features and many algorithms for measuring the distance between entities such as countries, the combination of which can result in widely varying outcomes. As a result, the temptation to search through all of the combinatorial possibilities until one stumbles upon an analytical method that supports one's hypothesis becomes real.

The solution to this dilemma is to not rely on a single analysis. When many feature sets are evaluated using a number of different computational means, a consensus should begin to emerge, and it is that consensus, not the results of a single study that should be the focus of attention. With the limited number of analyses presented above, 15 of the countries consistently cluster into six dialect areas:

1. European (Spain)
2. Southern Cone (Uruguay, Argentina)
3. Southern Central America (Costa Rica, Panama)
4. Caribbean (Puerto Rico, Dominican Republic)
5. Northern Central America (Nicaragua, El Salvador, Guatemala, Honduras)
6. Andean South America (Bolivia, Paraguay, Chile, Peru)

What the present analyses do not firmly establish is which dialect regions Cuba, Ecuador, Mexico,Venezuela, Colombia, and the US belong to, nor whether there may be additional dialect zones not considered. Another limitation of the present study is that it relies solely on political boundaries. Dialects do not necessarily align with such boundaries. Studies that make use of geotagged tweets are in a good position to find dialect zones that transcend the limits of individual countries. In sum, using computational methods and large numbers of features to delineate dialect boundaries is an improvement over earlier methods, the use of these methods opens up another set of issues that must be dealt with.

Acknowledgments

I appreciate the input by the reviewers as well as the suggestions provided by Earl Brown.

References

Alba, Orlando. 1992. Zonificación del español de America. In C. Hernández Alonso (ed.), *Historia y presente del español en America, 63-84.* Valladolid: Junta de Castilla y León.

Aliaga Jiménez, José Luis. Dialectometría y léxico en las hablas de Teruel. 2003. *ELUA. Estudios de Lingüística* 17: 5-55.

Brown, Earl. K. 2015. On the utility of combining production data and perceptual data to investigate regional linguistic variation: The case of Spanish experiential gustar ‘to like, to please’on Twitter and in an online survey.” *Journal of Linguistic Geography* 3(2): 47-59. https://doi.org/10.1017/jlg.2016.1

Burridge, J., Vaux, B., Gnacik, M., & Grudeva, Y. 2019. Statistical physics of language maps in the USA. *Physical Review E,* 99(3): 032305. https://doi.org/10.1103/PhysRevE.99.032305

Armas y Céspedes, Juan Ignacio. 1882. *Oríjenes del lenguaje criollo.* La Habana: Imprenta de la Viuda de Soler.

Cahuzac, Philippe. 1980. La división del español de América en zonas dialectales. Situación etnolingüística o semántico-dialectal. *Lingüística Española Actual* 2: 385-461.

Canfield, D. Lincoln. 1962. *La pronunciación del español en América*. Bogotá: Instituto Caro y Cuervo.

Davies, Mark. 2017. Corpus del Español, Web/Dialects. https://www.corpusdelespanol.org/web-dial/

Donoso, G., & Sánchez, D. 2017. Dialectometric analysis of language variation in Twitter. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects,* 16-25Valencia, Spain: Association for Computational Linguistics. 10.18653/v1/W17-1202

Eddington, David Ellingson. 2021. A corpus analysis of some usage differences among Spanish-speaking countries” *Dialectologia* 27: 71-95.

Frank, Eibe, Mark A. Hall, and Ian H. Witten. 2016. *Data Mining: Practical Machine Learning Tools and Techniques,* 4th Ed. San Francisco, CA: Morgan Kaufmann.

Embleton, Sheila, Dorin Uritescu, & Eric S. Wheeler. 2013. Defining dialect regions with interpretations: Advancing the multidimensional scaling approach. L*iterary and Linguistic Computing* 28: 13-22. https://doi.org/10.1093/llc/fqs048

Henríquez-Ureña, Pedro. 1921. Observaciones sobre el español en América. *Revista de Filología Española* 8: 357-390.

Holte, Robert C. 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning* 11.(1): 63-90.

García Mouton, Pilar. 1991. *Dialectometría y léxico en Huesca*. Mardid: Consejo Superior de Investigaciones Científicas.

Gonçalves, Bruno & David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS One, 9*(11): e112074. https://doi.org/10.1371/journal.pone.0112074; Data: http://www.bgoncalves.com/languages/spanish.html

Gonçalves, Bruno and David Sánchez. 2016. Learning about Spanish dialects through Twitter. *Revista Internacional de Lingüística Iberoamericana* 14: 65-75.

Grieve, Jack. 2011. A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16(4): 514-546. https://doi.org/10.1075/ijcl.16.4.04gri

Grieve, Jack. 2012. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory 8(*1): 39-72. https://doi.org/10.1515/cllt-2012-0003

Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In P. Auer, G. von Essen & W. Frick (eds.), *Aggregating dialectology, typology, and register analysis,* 53-88. Berlin: De Gruyter. https://doi.org/10.1515/9783110317558.53

Guyon, Isabel, Jason Weston, Stephen Barnhill, & Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning,*46(1): 389-422.

Henríquez-Ureña, P. H. 1921. Observaciones sobre el español en América. *Revista de Filología Española* 8: 357-390.

Holte, Robert C. 1993.Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning* 11(1): 63-90.

Huang, Yuan, Diansheng Guo, Alcie Kasakoff, & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment* *and Urban Systems* 59: 244-255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

The jamovi project. 2021. jamovi. (Version 1.6) [Computer Software.] Retrieved from https://www.jamovi.org)

Leino, Antti, & Saara Hyvönen. 2008. Comparison of component models in analysing the distribution of dialectal features. *International Journal of Humanities and Arts Computing* 2: 73-187. DOI: 10.3366/edinburgh/9780748640300.001.0001

Manni, Franz, Wilbert Heeringa, Bruno Toupance, & John Nerbonne. 2008. Do surname differences mirror dialect variation. *Human Biology* 80: 41-64.

Moreno Fernández, Francisco. 1991. Morfología en el ALEANR: aproximación dialectométrica. In *I curso de geografía lingüística de Aragón,* 289-309. Zaragoza: Institución Fernando el Católico.

Moreno Fernández, Francisco, and Hiroto Ueda. 2018. Cohesion and particularity in the Spanish dialect continuum. *Open Linguistics* 4: 722-742. https://doi.org/10.1515/opli-2018-0035

Nagy, Naomi, Xiaoli Zhang, George Nagy, and Edgar W. Schneider. 2006. Clustering dialects automatically: A mutual information approach. *University of Pennsylvania Working Papers in Linguistics* 12: 12.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1): 175-198. https://doi.org/10.1111/j.1749-818X.2008.00114.x

Quesada Pacheco, Miguel Ángel. 2014. División dialectal del español de América según sus hablantes Análisis dialectológico perceptual. *Boletín de Filología* 49(2): 257-309.

R Core Team 2020. R: A Language and environment for statistical computing. (Version 4.0) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2020-08-24).

Resnick, Melvyn C. 1975. Phonological Variants and Dialect Identification in Latin American Spanish. Mouton:The Hague.

Rodriguez-Diaz, Carlos A., Sergio Jimenez, George Dueñas, Johnatan Estivan Bonilla, & Alexander F. Gelbukh. 2018. Dialectones: Finding statistically significant dialectal boundaries using twitter data. *Computación y Sistemas* 22(4): 1213-1222.

Rodríguez Vázquez, Paloma. 2019. La zonificación dialectal del español de América: propuestas clásicas y propuestas actuales. Document, Universidade de Dantiago de Compostela. http://hdl.handle.net/10347/23567

Rona, José Pedro. 1964. El problema de la división del español americano en zonas dialectales. In F. Moreno Fernández (ed.), *Presente y futuro de la lengua española,* vol. I, 215-226. Madrid: Ediciones Cultura Hispánica

Sato, Yo, and Kevin Heffernan. 2018. Creating Dialect Sub-corpora by Clustering: a case in Japanese for an adaptive method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 3612-3616. Luxemburg: European Language Resources Association.

Sayce, David. n.d.. The Number of tweets per day in 2020. Accessed Feb. 2, 2022. https://www.dsayce.com/social-media/tweets-day/

Séguy Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335–57.

Seol, Hyunsoo. 2020. *SnowCluster: Cluster Analysis*. [jamovi module]. https://github.com.hyunsooseol/snowCluster

Shackleton Jr, R.obert G. 2005. English-American speech relationships: A quantitative approach. *Journal of English Linguistics* 33(2): 99-160. https://doi.org/10.1177/0075424205279017

Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1): 45-76.

Tellez, Eric. S., Daniela Moctezuma, Sabino Miranda, & Mario Graff. 2021. A large scale lexical and semantic analysis of Spanish language variations in Twitter. *arXiv preprint arXiv:2110.06128*.

Tinoco, Antonio. R., & Hiroto Ueda. 2007. The VARILEX Project-Spanish Lexical Variation. *Linguistica Atlantica* 27: 117-121.

Ueda, Hiroto. 2009. Resultados y proyectos en las investigaciones sobre variación léxica del español. *Dialectologia* 2: 51-80.

Wagner, Max Leopold. 1920. Amerikanisch-Spanisch und Vulgärlatein. *Zeitschrift für Romanische Philologie* 40: 286-312; 385-404.

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1(1): 243-264. https://doi.org/10.1146/annurev-linguist-030514-124930

Zamora, Juan y Jorge Guitart. 1988. *Dialectología hispanoamericana. Teoría, descripción, historia.* Salamanca: Almer.

Appendix

Rankings of features by importance according to four algorithms

Random Forest	Information Gain	One Attribute	Support Vector Machine	Feature
79	293	175	1	*bolígrafo*
40	355	346	2	present perfect
370	342	234	3	*capaz*
191	254	468	4	*mofletes*
139	251	207	5	*arañazo*
453	390	375	6	*tú*
194	341	165	7	*auto*
153	441	192	8	*-ase* subjunctive
2	298	115	9	*estadía*
94	465	210	10	*armario*
304	188	23	11	*entretenimiento*
311	173	155	12	*ático*
4	397	389	13	*valija*

335	461	462	14	*maleta*
190	400	405	15	*vistazo*
359	323	224	16	*chance*
344	445	490	17	*lavadora*
248	324	69	18	*coger*
418	264	321	19	*sois*
565	49	503	20	*lois*
201	306	251	21	*celular*
178	185	191	22	*alpargata*
218	360	206	23	*amarrar*
234	231	545	24	*pómulo*
372	127	564	25	*pantalón vaquero*
252	137	361	26	*reposera*
81	294	173	27	*bocata*
369	241	301	28	*sandalia*
289	31	83	29	*grasiento*
518	75	127	30	*escurridor*
73	402	199	31	*acera*
69	388	308	32	*salón*
297	165	403	33	*vidriera*
161	422	423	34	*sujetador*
309	297	122	35	*estación de servicio*
182	327	225	36	*cartel*
155	417	198	37	*afiche*
3	52	347	38	*poste de luz*
206	157	535	39	*pasatiempo*
399	118	280	40	*cabeza dura*
371	108	349	41	*póster*
402	460	147	42	*finca*
430	418	370	43	*sos*
267	329	18	44	*el sarten*
7	414	302	45	*sartén*
54	463	107	46	*guapa*
320	381	221	47	*anuncio*
129	73	242	48	*cassette*
390	94	96	49	*guineo*
483	361	352	50	*quizá*
179	412	316	51	*se los las decir*
89	440	212	52	*ascensor*
530	379	577	53	*papel de plata*
245	93	303	54	*ropero*
293	362	353	55	*quizás*
86	450	118	56	*gafas*
510	432	102	57	*habían* with singular
253	32	478	58	*melocotón*
432	291	162	59	*auriculares*
138	98	477	60	*mejillas*
227	257	284	61	*caldero*
200	300	171	62	*bocadillo*
285	38	105	63	*guardarropas*
277	20	455	64	*mantecoso*
8	284	26	65	*eres*

316	398	391	66	*veliz*
405	344	235	67	*cantina*
323	33	432	68	*tiovivo*
12	312	65	69	*computadora*
364	376	568	70	*ocasión*
72	289	159	71	*autopista*
76	263	408	72	*tiroteo*
25	76	114	73	*estola*
53	332	6	74	*elevador*
70	141	145	75	*farola*
535	99	460	76	*magnetófono*
172	295	113	77	*escuela*
440	195	550	78	*pizarrón*
492	216	61	79	*constipado*
305	431	101	80	*heladera*
187	61	252	81	*caterpillar*
168	81	366	82	*rebanada de queso*
450	196	335	83	*sostén*
419	117	579	84	*papel encerado*
14	317	46	85	*closet*
215	276	268	86	*bus*
195	189	295	87	*cacahuete*
273	346	10	88	*distracciones*
522	410	330	89	*ser consciente*
125	154	431	90	*tiesto*
473	459	186	91	*atar*
90	279	180	92	*barca*
11	153	533	93	*parlante*
34	521	29	94	*engrasado*
156	12	465	95	*mechero*
472	348	229	96	*camarero*
62	190	71	97	*colchón*
407	229	572	98	*nube de polvo*
193	444	211	99	*aro*
88	240	160	100	*automóvil*
415	452	93	101	*grasoso*
96	442	484	102	*lámina de queso*
424	128	256	103	*cazo*
127	210	131	104	*falencia*
365	406	201	105	*aero*
549	486	176	106	*bluyín*
247	411	328	107	*seguidor*
114	419	417	108	*tablero*
23	149	285	109	*calesita*
395	303	21	110	*entrevistar*
528	495	90	111	*gancho de ropa*
517	504	336	112	*queso americano*
223	187	496	113	*jeans*
213	133	288	114	*cachetes*
160	53	205	115	*alubia*
348	282	120	116	*extrañar*
181	255	185	117	*audífonos*

126	115	78	118	*goma de mascar*
425	484	13	119	*cristal delantero*
343	177	266	120	*calzón*
224	455	475	121	*mico*
9	84	228	122	*cartelón*
254	15	511	123	*lavarropa*
319	449	137	124	*frigorifico*
445	163	434	125	*tirantes*
481	162	51	126	*chancleta*
174	123	253	127	*celofán*
19	244	491	128	*la calor*
422	443	487	129	*lancha*
130	328	226	130	*carro*
374	236	28	131	*ensalada de fruta*
346	1	563	132	*palomitas*
159	260	86	133	*gasolinera*
120	239	157	134	*balacera*
396	68	521	135	*perchero*
246	277	178	136	*bar*
366	259	77	137	*grabadora*
386	19	269	138	*brassiere*
124	378	560	139	*ordenador*
157	60	409	140	*sponsor*
100	405	406	141	*vitrina*
77	308	73	142	*colegio*
117	171	79	143	*grabador*
51	176	337	144	*puerco*
52	416	323	145	*sortija*
165	29	472	146	*miradita*
431	368	368	147	*refrigerador*
326	337	32	148	*encendedor*
459	67	435	149	*tetera*
177	357	314	150	*sala de estar*
290	314	72	151	*colectivo*
340	319	58	152	*cochino*
387	464	457	153	*maní*
189	122	262	154	*buhardilla*
489	491	112	155	*gafotas*
93	436	515	156	*lentes*
464	131	357	157	*resfrío*
103	424	570	158	*mozo*
87	371	548	159	*pizarra*
91	167	567	160	*ojeada*
39	104	522	161	*penthouse*
578	503	376	162	*tutifruti*
140	234	152	163	*azafata o*
147	426	469	164	*mono*
417	34	7	165	*edredón*
146	92	260	166	*calzoncillo*
361	338	2	167	*echar de menos*
400	206	273	168	*bombacho*
375	343	237	169	*camioneta*

151	4	177	170	*bidón*
204	492	541	171	*polvero*
347	64	276	172	*bonita*
154	334	166	173	*autobús*
436	235	195	174	*agarrar*
119	70	74	175	*escaparate*
501	304	290	176	*cacerola*
29	485	427	177	*tartera*
18	82	272	178	*botella grande*
345	385	143	179	*fanáticos*
57	435	99	180	*hamaca*
169	156	558	181	*paila*
203	44	373	182	*tumbona*
286	140	259	183	*camastro*
258	219	351	184	*preciosa*
202	261	573	185	*obstinado*
580	569	261	186	*bomba de nafta*
104	96	540	187	*polvareda*
149	215	364	188	*remezón*
216	50	377	189	*tiza*
166	377	571	190	*nevera*
406	415	324	191	*simio*
60	309	257	192	*cazuela*
410	347	11	193	*descongelar*
291	14	204	194	*altoparlante*
373	287	277	195	*bote*
452	250	82	196	*grapadora*
82	27	557	197	*movimiento sísmico*
48	126	363	198	*rasguño*
487	479	154	199	*backpack*
570	582	80	200	*gramola*
456	437	495	201	*hostess*
394	222	111	202	*gallinita ciega*
13	364	334	203	*ropa interior*
205	227	243	204	*catarro*
272	25	507	205	*macaco*
334	373	559	206	*oportunidad*
141	321	53	207	*chicle*
339	386	220	208	*anteojos*
217	3	305	209	*scotch*
435	23	384	210	*tocadiscos*
242	174	442	211	*terco*
268	273	506	212	*luneta*
563	474	322	213	*soquetes*
180	213	543	214	*polvorín*
520	470	320	215	*sobrecama*
15	396	436	216	*terremoto*
404	194	289	217	*cabezón*
312	446	494	218	*jugo*
460	59	291	219	*cabezota*
105	350	1	220	*zumo*
512	175	585	221	*pantufla*

17	325	149	222	*chancho*
266	253	33	223	*encerado*
420	524	44	224	*cosedora*
465	89	348	225	*poste eléctrico*
488	146	100	226	*headphones*
341	129	223	227	*bomba de gasolina*
508	518	587	228	*papel albal*
383	74	325	229	*silla plegable*
555	526	542	230	*polvoreda*
112	63	158	231	*autovía*
67	322	132	232	*fallo*
116	290	174	233	*boliche*
148	313	59	234	*computador*
337	243	367	235	*recibidor*
107	83	148	236	*chancla*
92	182	36	237	*endulzante*
444	203	241	238	*casquitos*
101	363	354	239	*rancho*
136	36	338	240	*pulpería*
162	370	552	241	*plátano*
46	214	215	242	*amplificador*
495	510	380	243	*tranchete*
232	403	392	244	*vosotros*
447	228	443	245	*terral*
376	191	526	246	*pileta*
240	62	388	247	*tozudo*
238	8	387	248	*topadora*
551	467	317	249	*seboso*
429	262	569	250	*movimiento telúrico*
398	88	345	251	present for past subjunctive
131	286	119	252	*expendio*
128	394	299	253	*sandwich*
360	209	109	254	*guardafango*
476	274	586	255	*papel de estaño*
546	106	67	256	*cóctel de fruta*
439	409	319	257	*sismo*
208	301	279	258	*caballitos*
457	404	399	259	*vos*
427	221	246	260	*cerilla*
324	383	528	261	*pc*
513	489	190	262	*albal*
158	457	139	263	*fósforo*
325	448	485	264	*la margen*
28	152	456	265	*mansarda*
433	382	218	266	*anillo*
377	428	446	267	*hinchas*
467	476	20	268	*zippo*
106	265	428	269	*tasca*
188	292	116	270	*estancia*
329	380	538	271	*piscina*
382	100	590	272	*papel de aluminio*
579	481	108	273	*guardabarro*

330	204	271	274	*bóxers*
338	202	396	275	*zancudo*
494	299	123	276	*estacionar*
251	256	505	277	*luminaria*
133	408	415	278	*tal vez*
503	224	146	279	*farolillo*
545	580	126	280	*escurridero*
211	30	476	281	*megáfono*
118	72	532	282	*parabrisa*
184	367	369	283	*refrigeradora*
263	186	523	284	*percha*
308	179	502	285	*lavaplatos*
239	296	172	286	*bocadito*
26	145	265	287	*buldócer*
438	352	141	288	*fanaticada*
287	453	209	289	*armador*
455	237	414	290	*tirita*
442	482	151	291	*azulón*
540	525	574	292	*pantalones tejanos*
31	531	419	293	*tabanco*
75	494	8	294	*edulcorante*
207	369	553	295	*platicar*
448	155	95	296	*gripe*
560	515	371	297	*soutien*
235	57	68	298	*colgador*
219	393	383	299	*tragamonedas*
496	184	240	300	*canoa*
143	208	24	301	*equivocación*
264	584	556	302	*papel sanitario*
468	501	555	303	*paquete postal*
363	21	97	304	*habichuela*
262	5	451	305	*machina*
303	425	531	306	*patrocinador*
261	164	471	307	*mosco*
45	192	181	308	*barman*
350	468	315	309	*scuela*
66	58	294	310	*cacahuate*
37	87	429	311	*testarudo*
1	69	514	312	*lavavajillas*
47	22	64	313	*corpiño*
408	339	3	314	*diversión*
336	389	518	315	*mosquito*
480	462	106	316	*guagua*
562	497	424	317	*surtidor de gasolina*
294	519	398	318	*wurlitzer*
58	399	402	319	*vereda*
328	318	43	320	*cintillo*
362	349	12	321	*damasco*
257	336	34	322	*encomienda*
186	178	286	323	*calzada*
5	375	566	324	*olla*
237	211	150	325	*chancha*

22	310	66	326	*comedor*
471	351	516	327	*letrero*
167	430	512	328	*linda*
588	591	449	329	*máquina de música*
582	533	103	330	*guardilla*
135	218	463	331	*matera*
292	2	372	332	*porfiado*
282	101	48	333	*chongo*
497	530	501	334	*lavaloza*
446	469	410	335	*tajador*
212	283	130	336	*falda*
16	80	329	337	*seísmo*
434	77	75	338	*escarpa*
505	144	170	339	*blue jean*
416	205	133	340	*fosforera*
461	577	270	341	*brasiel*
516	79	440	342	*tejanos*
544	201	264	343	*brik*
271	438	467	344	*mochila*
111	421	197	345	*afición*
411	148	247	346	*cerillo*
576	550	341	347	*queso en lonchas*
493	527	459	348	*magnetofón*
349	107	549	349	*pochoclo*
183	487	393	350	*yesquero*
6	447	488	351	*maceta*
276	65	482	352	*macetero*
332	439	486	353	*la puente*
20	161	168	354	*banano*
21	55	365	355	*rayón*
367	28	275	356	*bombona*
10	395	433	357	*tiradores*
121	281	278	358	*butaca*
44	483	591	359	*papel de baño*
55	11	466	360	*microcomputador*
99	387	310	361	*rótulo*
536	553	546	362	*polvazal*
525	549	342	363	*queso en rebanadas*
280	311	231	364	*casete*
389	132	200	365	*acolchado*
491	168	54	366	*chimpancé*
502	520	91	367	*gandula*
554	575	40	368	*cinta scotch*
171	466	194	369	*altillo*
307	230	182	370	*barrero*
548	517	430	371	*tierral*
358	170	52	372	*chango*
318	134	519	373	*movi*
225	330	283	374	*caldera*
123	413	304	375	*saya*
256	103	576	376	*pantaloncillo*
479	247	454	377	*máquina de lavar*

249	238	255	378	*cayuco*
553	490	309	379	*salita de estar*
317	454	138	380	*frijol*
176	407	202	381	*aeromoza o*
43	139	313	382	*sacarina*
56	180	450	383	*marrano*
63	112	169	384	*banqueta*
265	366	356	385	*resfriado*
506	124	161	386	*autocar*
113	10	312	387	*sacapuntas*
403	42	401	388	*vitrola*
152	433	98	389	*hacienda*
385	169	483	390	*mostrador*
515	587	447	391	*máquina excavadora*
228	340	233	392	*carretera*
577	252	30	393	*engrapadora*
220	233	179	394	*bicoca*
97	150	470	395	*morral*
504	120	63	396	*coriza*
260	456	474	397	*medias*
302	151	390	398	*vaqueros*
198	401	397	399	*zapatilla*
352	114	588	400	*papel aluminio*
331	480	385	401	*tolvanera*
484	475	500	402	*lavadora de platos*
284	458	479	403	*mesero*
210	54	216	404	*aparador*
32	199	282	405	*calcetín*
401	105	14	406	*cuarto de estar*
583	564	422	407	*stewardess*
122	326	230	408	*cascos*
380	508	124	409	*espónsor*
498	560	426	410	*tapabarros*
355	111	374	411	*trusa*
298	429	296	412	*hermosa*
466	496	539	413	*pitusa*
279	271	50	414	*chispero*
412	267	481	415	*macuto*
50	280	153	416	*azotea*
547	477	497	417	*jokey*
490	26	292	418	*cabezudo*
24	270	318	419	*silla reclinable*
475	555	300	420	*salveque*
241	423	421	421	*-stes* for *-ste* preterite
327	198	81	422	*gallina ciega*
214	166	208	423	*argolla*
80	46	381	424	*trancazo*
175	102	87	425	*garrafón*
30	493	509	426	*loncha de queso*
392	391	379	427	*ustedes*
270	45	110	428	*guardapolvo*
65	85	464	429	*matero*

441	269	222	430	*bolso de viaje*
574	472	41	431	*cinta adhesiva*
281	511	213	432	*arañón*
71	353	520	433	*poncho*
173	56	128	434	*esparadrapo*
185	159	70	435	*colcha*
313	41	340	436	*rollo de papel*
278	499	62	437	*contén*
59	24	395	438	*yola*
275	434	492	439	*la azúcar*
98	331	19	440	*emparedado*
384	471	412	441	*tajada de queso*
150	356	217	442	*andén*
391	285	167	443	*banana*
315	223	15	444	*cubrecama*
33	512	498	445	*jorongo*
573	590	489	446	*lasca de queso*
463	372	214	447	*aparcar*
109	245	187	448	*bolsón*
110	288	183	449	*bella*
229	121	343	450	*propiciador*
27	90	250	451	*cercha*
378	138	238	452	*canchita*
95	13	452	453	*mascar*
36	568	583	454	*papel de váter*
299	51	94	455	*gripa*
393	158	530	456	*patrocinante*
300	160	219	457	*añorar*
342	500	56	458	*cocaleca*
295	335	163	459	*auspiciador*
145	358	547	460	*pollera*
83	384	529	461	*pava*
209	40	524	462	*pianola*
164	359	554	463	*pluma*
108	278	27	464	*error*
428	71	534	465	*parquear*
192	135	188	466	*alberca*
314	392	378	467	*usted*
283	320	57	468	*coche*
486	592	592	469	*abrochadora*
409	345	17	470	*curita*
353	507	444	471	*pororó*
379	116	355	472	*reportear*
507	589	493	473	*juke box*
142	181	254	474	*catre*
356	200	164	475	*auspiciante*
49	183	227	476	*carrusel*
38	585	581	477	*papel higiénico*
310	220	129	478	*espejuelos*
397	17	156	479	*badén*
134	136	332	480	*silla de playa*
557	509	382	481	*traganíquel*

351	193	144	482	*farol*
414	217	439	483	*tecle*
566	566	411	484	*tablón de anuncio*
243	47	510	485	*livin*
274	225	416	486	*tallador*
322	48	575	487	*pantalón de mezclilla*
575	523	136	488	*friegaplatos*
244	6	5	489	*echar en falta*
584	581	89	490	*gallo ciego*
469	572	31	491	*engrampadora*
333	498	407	492	*tocacintas*
199	37	22	493	*entretención*
132	78	350	494	*pota*
137	43	49	495	*chola*
231	266	135	496	*fréjol*
539	571	16	497	*cubrelecho*
581	570	525	498	*papel toalet*
85	110	362	499	*repo*
585	586	582	500	*papel para cocinar*
538	532	400	501	*vellonera*
144	316	248	502	*cervecería*
388	86	386	503	*tombo*
454	522	458	504	*magnavoz*
569	232	360	505	*rodaja de queso*
41	354	344	506	*propaganda*
354	212	287	507	*calcetas*
269	514	307	508	*salpicadera*
42	109	189	509	*ajustador*
381	35	536	510	*patera*
163	113	551	511	*placard*
226	374	565	512	*ómnibus*
541	567	580	513	*papel estañado*
534	506	425	514	*suspensores*
306	535	88	515	*gallito ciego*
221	226	453	516	*masticar*
470	557	331	517	*silla de extensión*
521	528	37	518	*crispeto*
64	7	480	519	*mesonero*
357	207	267	520	*buldózer*
250	39	441	521	*temblor de tierra*
421	91	461	522	*mahón*
368	95	117	523	*excavadora*
170	16	232	524	*carrillos*
564	545	578	525	*papel de water*
426	451	84	526	*giz*
301	197	121	527	*estada*
550	268	404	528	*vidrio delantero*
532	534	9	529	*diurex*
519	540	281	530	*cacillo*
196	365	193	531	*altavoz*
462	576	47	532	*cínife*
561	573	125	533	*escurreplatos*

482	559	394	534	*yins*
529	544	274	535	*bomba de bencina*
74	305	258	536	*celo*
571	505	561	537	*pala excavadora*
511	427	298	538	*hinchada*
542	562	413	539	*tajalápiz*
485	249	236	540	*cambur*
572	548	333	541	*rositas de maíz*
233	142	60	542	*concho*
443	246	517	543	*lighter*
115	18	104	544	*grúa*
514	513	504	545	*lonja de queso*
556	272	55	546	*chinela*
451	125	142	547	*foil*
78	307	244	548	*centro escolar*
500	119	249	549	*chalana*
533	558	438	550	*terregal*
236	172	76	551	*escondidas*
288	9	306	552	*salpicadero*
478	478	263	553	*breteles*
259	302	4	554	*durazno*
437	541	45	555	*clericó*
413	539	184	556	*bencinera*
524	583	562	557	*papel platina*
84	66	92	558	*garrafa*
35	546	358	559	*rocola*
458	143	311	560	*ruana*
230	473	39	561	*cotufa*
586	561	437	562	*taxibús*
523	537	42	563	*cinta pegante*
559	488	35	564	*endulzador*
102	315	245	565	*cerdo*
61	333	25	566	*enagua*
477	547	359	567	*roconola*
296	258	508	568	*macedonia*
68	420	418	569	*taberna*
568	529	513	570	*lavatrastos*
449	516	85	571	*gasolinería*
587	588	499	572	*judía verde*
197	147	140	573	*forofos*
591	556	326	574	*secaplatos*
527	565	420	575	*sutién*
592	551	339	576	*queso de sandwich*
589	242	589	577	*papel confor*
567	502	448	578	*máquina de lavar platos*
526	552	527	579	*pipocas*
222	579	239	580	*canguil*
531	574	38	581	*crispeta*
552	578	293	582	*cabrita de maíz*
543	554	544	583	*pomo plástico*
423	275	203	584	*aficcionados*
537	97	584	585	*papel de inodoro*

255	248	537	586	*piquera*
590	563	327	587	*silla de sol*
509	543	297	588	*hielera*
499	538	134	589	*freidero*
321	130	473	590	*microbús*
474	536	445	591	*poporopo*
558	542	196	592	*afilaminas*