Dialectologia 9 *(2012), 1-26.*
ISSN: 2013-2247

Received 29 November 2011.
Accepted 1 February 2012.

THE IMPORTANCE OF TEN PHONETIC CHARACTERISTICS TO DEFINE DIALECT AREAS IN SPANISH

Germán COLOMA
CEMA University, Buenos Aires, Argentina
gcoloma@cema.edu.ar

Abstract

This paper studies ten phonetic characteristics of the Spanish language (/s/-/θ/ merger, /ʝ/-/ʎ/ merger, /s/-aspiration, /x/-aspiration, /ʝ/-assibilation, /r/-assibilation, /n/-velarization, /tʃ/-deaffrication, /x/-uvularization and /tʃ/-voicing) and analyzes their ability to define dialect areas. We conclude that there are five of them (/s/-aspiration, /x/-aspiration, /n/-velarization, /x/-uvularization and /r/-assibilation) which are particularly useful for that task, since they define between six and fourteen compact dialect areas. Geographic coherence is the main element used to evaluate the usefulness of the studied characteristics, together with some statistic and dialectometric properties. An interesting corollary is that, although they are the most significant phonological variables in Spanish, neither the /s/-/θ/ merger (*seseo*) nor the /ʝ/-/ʎ/ merger (*yeismo*) are particularly relevant as geolinguistic markers.

Key words
phonetic characteristics, dialect zones, geographic coherence, compact areas

LA IMPORTANCIA DE DIEZ CARACTERÍSTICAS FONÉTICAS PARA DEFINIR ÁREAS DIALECTALES EN ESPAÑOL

Resumen

El presente trabajo estudia la capacidad de diez características fonéticas (seseo, yeísmo, aspiración de /s/, aspiración de /x/, asibilación de /ʝ/, asibilación de /r/, velarización de /n/, desafricación de /ʧ/, uvularización de /x/ y sonorización de /ʧ/) para delimitar zonas dialectales de la lengua española, y trata de aislar las que resultan más útiles para dicha tarea. Se concluye que hay cinco características (aspiración de /s/, aspiración de /x/, velarización de /n/, uvularización de /x/ y asibilación de /r/) que son particularmente significativas para delimitar zonas dialectales en el mundo hispanohablante, ya que permiten una zonificación que genera entre seis y catorce áreas dialectales compactas. La coherencia geográfica es el principal elemento utilizado para evaluar la utilidad de las características estudiadas, junto con ciertas propiedades estadísticas y dialectométricas de las variables. Un corolario de interés es que, a pesar de ser las variables fonológicamente más relevantes, ni el seseo ni el yeísmo son características demasiado importantes desde el punto de vista geolingüístico.

Palabras clave

características fonéticas, zonas dialectales, coherencia geográfica, áreas compactas

0. Introduction

This paper studies the geographic distribution of ten phonetic characteristics which are supposed to be useful to define dialect areas in the Spanish-speaking world. That geographic distribution produces isoglosses that define twenty-eight separate areas. The importance of those areas, and the importance of each of the ten phonetic variables that define them, are nevertheless not equivalent. That is why we evaluate the relevance of the analyzed variables, using a methodology that compares the geographic distribution of the characteristics and the clustering of areas induced by eliminating each of the variables. The main criterion for choosing a variable is the geographic coherence of the generated clusters, and the way in which our methodology will be applied is first simultaneous and then sequential.

The article is organized as follows. In the first section we explain the areas obtained by overlapping the isoglosses of the ten analyzed phonetic variables, while in the following section we study the relative importance of each of the variables (and

conclude that there are five of them which are more important than the others). Then we develop a sequential method that generates compact dialect regions characterized by a minimum number of phonetic variables, whereas in the last section we present the main conclusions of the whole paper, together with some final remarks.

1. Phonetic characteristics and dialect areas

The ten phonetic characteristics that we use to define dialect areas in the Spanish-speaking world are the following:

1) Seseo: It is the merger of the phonemes /s/ and /θ/ into a single one, typically pronounced using the alveolar fricative voiceless consonant [s].

2) Yeismo: It is the merger of the phonemes /ʝ/ and /ʎ/ into a single one, pronounced using one of the allophones of the first of those phonemes (which is generally the palatal approximant voiced consonant [j̝]).

3) Aspiration of /s/: It is the use of the glottal aspirated fricative consonant [h] as an allophone of the phoneme /s/, especially before another consonant.

4) Aspiration of /x/: It is the use of [h] as the main articulation of the otherwise velar fricative phoneme /x/.

5) Assibilation of /ʝ/: It is the articulation of the phoneme /ʝ/ through an assibilated postalveolar consonant, that may be a voiced affricate [dʒ], a voiced fricative [ʒ] or a voiceless fricative [ʃ].

6) Assibilation of /r/: It is the articulation of this phoneme through an assibilated alveolar or postalveolar fricative consonant [ɹ], instead of using the alveolar trill [r].

7) Velarization of /n/: It is the use of the velar nasal consonant [ŋ] as an allophone of /n/, not only when that phoneme appears before another velar consonant but also in a word-final position.

8) Deaffrication of /tʃ/: It is the use of the voiceless postalveolar fricative consonant [ʃ] to articulate the phoneme /tʃ/, either as the main pronunciation for that phoneme or as an alternative allophone.

9) Uvularization of /x/: It is the use of the voiceless uvular fricative consonant [χ] as an allophone of /x/, especially before /o/ and /u/.

10) Voicing of /ʧ/: It is the articulation of this phoneme through a partially voiced affricate consonant, whose sound can be represented as [tʒ] or [t̬].

No	Area / Characteristic	Se-seo	Ye-ismo	Aspir /s/	Aspir /x/	Assib /j/	Assib /r/	Velar /n/	Deaff /tʃ/	Uvul /x/	Voice /tʃ/
1	Traditional Castilian	No	No	No	No	No	No	No	No	Yes	No
2	Modern Castilian	No	Yes	No	No	No	No	No	No	Yes	No
3	Galician	No	Yes	No	No	No	No	Yes	No	Yes	No
4	Manchego-murcian	No	Yes	Yes	No	No	No	No	No	Yes	No
5	Extremaduran	No	Yes	Yes	No	No	No	Yes	No	Yes	No
6	Valencian	Yes	Yes	No	No	No	No	No	No	Yes	No
7	Eastern Andalusian	Yes	Yes	Yes	Yes	No	No	No	No	No	No
8	Western Andalusian	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No
9	Canarian	Yes	Yes	Yes	Yes	No	No	Yes	No	No	Yes
10	Northern Mexican	Yes	Yes	No	No	No	No	No	Yes	No	No
11	Central Mexican	Yes	Yes	No	No	No	No	No	No	No	No
12	Eastern Mexican	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No
13	Central American	Yes	Yes	No	Yes	No	No	Yes	No	No	No
14	Antillean Caribbean	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	No
15	Continental Caribbean	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No
16	Northern Andean	Yes	Yes	No	Yes	No	No	No	No	No	No
17	Equatorial Andean	Yes	No	No	Yes	No	Yes	No	No	No	No
18	Amazonic	Yes	Yes	No	Yes	No	Yes	No	No	No	No
19	Peruvian Coast	Yes	Yes	No	No	No	No	Yes	No	No	No
20	High Peruvian Andean	Yes	No	No	No	No	Yes	No	No	No	No
21	Eastern Bolivian	Yes	No	Yes	No	No	Yes	No	No	No	No
22	Paraguayan	Yes	No	Yes	No	Yes	Yes	No	No	No	No
23	Argentine-Bolivian	Yes	Yes	Yes	No	No	Yes	No	No	No	No
24	Tucuman-Saltean	Yes	Yes	Yes	No	Yes	Yes	No	No	No	No
25	Northern Chilean	Yes	Yes	Yes	No	No	Yes	No	Yes	No	No
26	Southern Chilean	Yes	Yes	Yes	No	No	No	No	Yes	No	No
27	Cuyan	Yes	Yes	Yes	No	No	No	No	No	No	No
28	River Plate	Yes	Yes	Yes	No	Yes	No	No	No	No	No

Table 1. Spanish dialect areas according to their phonetic characteristics

The geographic distribution of these ten phonetic characteristics is relatively well-studied in the Spanish dialectology literature. Based on the isoglosses proposed in that literature, we have identified twenty-eight dialect areas, which are the ones that appear on Table 1. Our main sources for the distribution of these phonetic variables in the Spanish-speaking world are Moreno-Fernández (2009) and Hualde (2005), and for some

particular regions and some phonetic characteristics we have used information from Borland (2004), Fontanella (2000), Lipski (2004), Martín-Butragueño (2010), Moreno de Alba (2001), Samper (2008), Utgard (2007) and Villena (2008). All the areas defined are different among each other, at least in one aspect of the distribution of the phonetic variables, and all of them are geographically compact.

Notice that the number of defined areas (28) is surprisingly small, if we take into account the quantity of binary variables used. In fact, as each phonetic variable can take two values, the number of possible permutations of those values in a group of ten elements is equal to "2^{10}". This implies that, in theory, there could be up to 1024 different dialect areas.

The demographic importance of the dialect areas described on Table 1 is very uneven, as can be seen on Table 2. That Table shows that some of the defined dialect areas have less than 0.3% of the total Spanish-speaking population (as is the case of the Extremaduran area, whose population share is 0.28%), while others have more than 15% of that population (as is the case of the Central Mexican area, whose share is 17%). These figures have been calculated using data from the World Bank (2011) and national complementary sources.[1] In order to calculate the population figures in a precise way, it was necessary to assume specific borders for each dialect area. Those assumptions are presented on Appendix 1.

Another figure reported on Table 2 is a "linguistic innovation index", that comes from converting the columns of Table 1 to numerical variables that assign a zero to the absence of the studied phonetic characteristics and a one to the presence of those characteristics. If, after doing that, we add up those figures horizontally, we obtain a number that in theory could lie between zero and ten, but that in practice goes from a minimum value of one (for the Traditional Castilian area) to a maximum value of seven (for the Western Andalusian area).

[1] For an explanation of the sources used, see Coloma (2011).

No	Code	Area / Characteristic	Population (2010)		Innovation Index	Differentiation Index
			Thousands	%		
1	CST	Traditional Castilian	2.429	0,58%	1	0,5567
2	CSM	Modern Castilian	22.114	5,29%	2	0,4778
3	GAL	Galician	2.933	0,70%	3	0,4974
4	MMU	Manchego-murcian	4.076	0,97%	3	0,4997
5	EXT	Extremaduran	1.156	0,28%	4	0,5184
6	VAL	Valencian	3.077	0,74%	3	0,3793
7	AOR	Eastern Andalusian	3.159	0,76%	4	0,3022
8	AOC	Western Andalusian	5.231	1,25%	7	0,4971
9	CAN	Canarian	2.043	0,49%	6	0,4577
10	MXN	Northern Mexican	16.214	3,88%	3	0,3611
11	MXC	Central Mexican	70.650	16,89%	2	0,2470
12	MOR	Eastern Mexican	17.147	4,10%	5	0,3953
13	CAM	Central American	43.161	10,32%	4	0,2984
14	ANT	Antillean Caribbean	25.409	6,07%	6	0,4241
15	CAC	Continental Caribbean	52.052	12,44%	5	0,3323
16	ANN	Northern Andean	30.837	7,37%	3	0,2645
17	ANE	Equatorial Andean	8.818	2,11%	3	0,4786
18	AMZ	Amazonic	2.656	0,63%	4	0,3840
19	RBP	Peruvian Coast	20.110	4,81%	3	0,2830
20	AAP	High Peruvian Andean	13.840	3,31%	2	0,4692
21	BOR	Eastern Bolivian	3.133	0,75%	3	0,4915
22	PAR	Paraguayan	10.229	2,44%	4	0,5557
23	ARB	Argentine-Bolivian	2.651	0,63%	4	0,4000
24	TCS	Tucuman-Saltean	3.648	0,87%	5	0,4767
25	CHN	Northern Chilean	2.018	0,48%	5	0,4789
26	CHA	Southern Chilean	15.117	3,61%	4	0,3897
27	CUY	Cuyan	2.217	0,53%	3	0,2871
28	RPT	River Plate	32.235	7,71%	4	0,3869
		Total	418.360	100,00%	3,60	0,3544

Table 2. Demographic and linguistic characteristics of the dialect areas

The interpretation of that number as a linguistic innovation index has to do with the idea that all the included variables represent some kind of change that occurred in a certain moment of the history of Spanish language, and therefore the areas that adopted that change can be considered “more innovative” than the areas that did not adopt the corresponding change. The obtained ranking can also be seen as compatible with the usual typology of the Hispanic dialectology literature, since in the group of areas with lower values for the innovation index we find the Modern Castilian, Central Mexican

and High Peruvian Andean areas (with a value of 2), while in the group of areas with higher values we find the Canarian and Antillean Caribbean areas (with a value of 6).

The last column of Table 2 shows the values of a "linguistic differentiation index", similar to the one employed in other works about Spanish language dialectometry.[2] That index has been calculated using the following formula:

$$\text{Differ}(n) = \sqrt{\frac{\sum_{i=1}^{10} (x_{in} - \mu_i)^2}{10}} \quad ;$$

where *Differ(n)* is the index corresponding to a particular dialect area, x_{in} is the value of a certain variable in that dialect area, and μ_i is the average value of that variable in the whole Spanish-speaking world. The idea behind this index is to measure how different a certain dialect area is from the rest of the areas. In order to do that, we have taken the average values of the variables as representative elements of the whole set. The farther an area is from those average values, the "more different" it is, and when we average those quadratic deviations (and we then apply the square root to that average), we obtain a number that is closer to zero if the area is similar to the general average and closer to one if the area is very different from that average.

The differentiation indices reported on Table 2 show that the less differentiated region is the Central Mexican area (*Differ = 0.2470*), followed by the Northern Andean area (*Differ = 0.2645*) and the Peruvian Coast area (*Differ = 0.2830*). On the other hand, the region with the largest differentiation index is the Traditional Castilian area (*Differ = 0.5567*), followed by the Paraguayan (*Differ = 0.5557*) and Extremaduran (*Differ = 0.5184*) areas. These results seem to coincide with the idea that the dialects spoken in the first three areas mentioned are closer to a sort of "neutral" or "standard" Latin American Spanish, while the last three areas would be representative of dialects with numerous idiosyncratic characteristics (either conservative or innovative).

Another way to evaluate the differences among the dialect areas is through a multimensional scale (MDS) plot, which translates the differences of values for the

[2] See, for example, García-Mouton (1999), who uses a different index based on the one proposed by Séguy (1973).

phonetic variables in each area into a measure of distance in a two-dimensional space.[3] That plot appears on Figure 1, in which each point represents one of the twenty-eight areas of Table 1, and the distances between it and the other points are based on the distances between those points in the ten-dimensional space of the analyzed phonetic variables.[4]

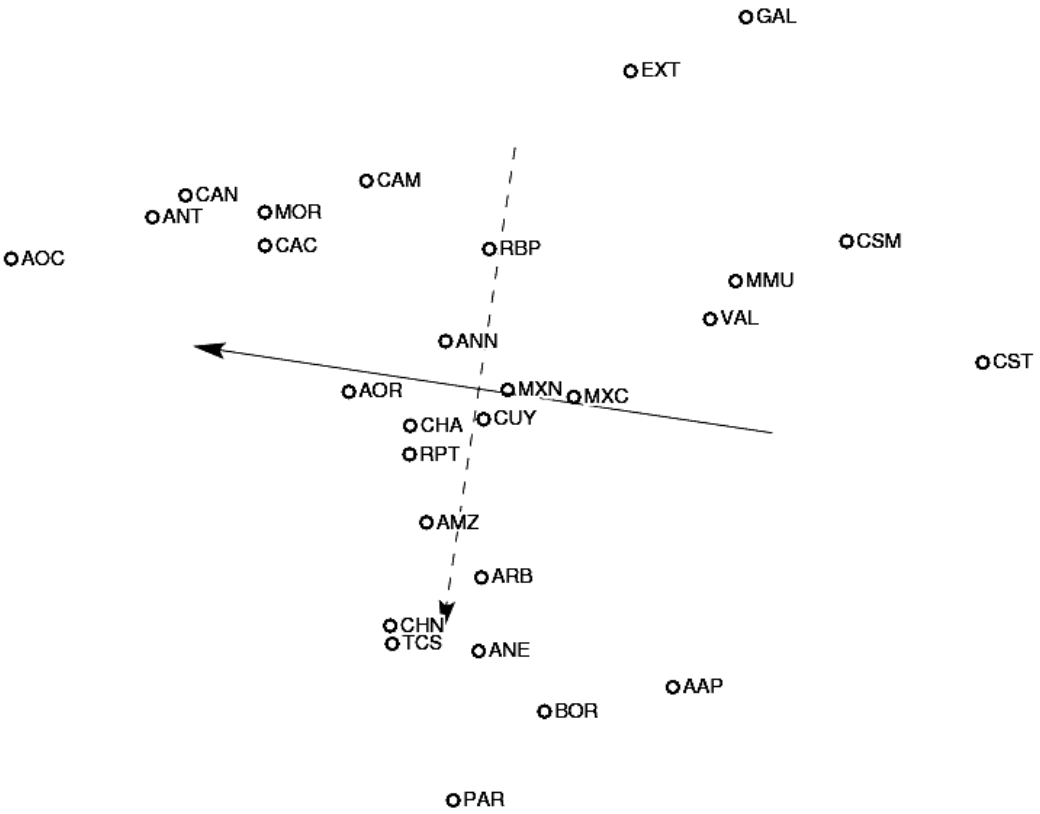

Figure 1. Multidimensional scale plot of the dialect areas

Figure 1 is useful to see that the areas whose phonetic characteristics are farther from the average of the Spanish-speaking world seem to be, due to different reasons, the Traditional Castilian (CST), Western Andalusian (AOC), Galician (GAL), Extremaduran (EXT) and Paraguayan (PAR) areas. In the plot we also see that some areas seem to constitute particularly homogeneous groups, such as the one integrated by the Antillean Caribbean, Canarian, Eastern Mexican and Continental Caribbean areas

[3] For an explanation of this concept and the logic behind the construction of an MDS plot, see Nerbonne (2010).

[4] This plot was generated using the Gabmap software, developed at the University of Groningen (Netherlands).

(ANT-CAN-MOR-CAC), and the one formed by the Northern Chilean, Tucuman-Saltean, Equatorial Andean and Argentine-Bolivian areas (CHN-TCS-ANE-ARB).

2. Importance of the phonetic variables

In order to study the relative importance of the phonetic variables described in the previous section, as possible criteria to define dialect areas in Spanish, in this section we will first calculate the correlation indices between the different variables. Due to the fact that the areas arising from overlapping the isoglosses defined by the ten variables differ substantially in size, our correlation indices will be weighted by the population of the corresponding areas. As was done with the linguistic innovation and differentiation indices calculated in the previous section, each concept is defined here through a binary variable that takes a value of zero when the corresponding phonetic characteristic is absent and a value of one when that characteristic is present. The formula for the correlation index for two variables "i" and "j" is therefore the following:

$$\mathrm{Corr}(i,j) = \frac{\mathrm{Cov}(i,j)}{\sigma_i \cdot \sigma_j} = \frac{\sum_{n=1}^{28}\left[\left(x_{in}-\mu_i\right)\cdot\left(x_{jn}-\mu_j\right)\cdot s_n\right]}{\sqrt{\sum_{n=1}^{28}\left[\left(x_{in}-\mu_i\right)^2\cdot s_n\right]}\cdot\sqrt{\sum_{n=1}^{28}\left[\left(x_{jn}-\mu_j\right)^2\cdot s_n\right]}} \quad ;$$

where *Cov(i,j)* is the covariance between the two variables, σ_i and σ_j are the corresponding standard deviations of those variables, μ_i and μ_j are the average values of those variables, and s_n is the population share of the nth area in the Spanish-speaking world.

The values of the calculated correlation indices appear on Table 3. Note that in all cases they take a value of one when "i" and "j" are the same variable and that, in general, those values can range from a minimum of -1 (perfect negative correlation) to a maximum of 1 (perfect positive correlation). As, additionally, it holds that "*Corr(i,j) = Corr(j,i)*", on Table 3 we only report the results that correspond to the upper-right half of the correlation matrix.

Variable	**Seseo**	**Yeismo**	**Aspir /s/**	**Aspir /x/**	**Assib /ʝ/**	**Assib /r/**	**Velar /n/**	**Deaff /tʃ/**	**Uvul /x/**	**Voice /tʃ/**
Seseo	1,000	-0,018	0,139	0,266	0,129	0,104	0,166	0,124	-0,952	0,020
Yeismo		1,000	0,030	0,144	-0,088	-0,831	0,262	0,135	0,025	0,022
Aspiration of /s/			1,000	0,128	0,323	0,050	0,193	0,308	-0,154	0,087
Aspiration of /x/				1,000	-0,114	-0,151	0,665	0,020	-0,280	0,077
Assibilation of /ʝ/					1,000	0,126	-0,070	-0,094	-0,135	-0,031
Assibilation of /r/						1,000	-0,293	-0,109	-0,109	-0,025
Velarization of /n/							1,000	0,064	-0,181	0,085
Deaffrication of /tʃ/								1,000	-0,130	-0,030
Uvularization of /x/									1,000	-0,021
Voicing of /tʃ/										1,000

Table 3. Correlation matrix between phonetic variables

The figures on Table 3 show two cases that exhibit a very high negative correlation, which are the /s/-/θ/ merger (*seseo*) with the uvularization of /x/ (*Corr = -0.952*), and the /ʝ/-/ʎ/ merger (*yeismo*) with the assibilation of /r/ (*Corr = -0.831*). This is due to the fact that, in general, in the Spanish-speaking world the absence of *seseo* is linked to the presence of /x/-uvularization, while the absence of *yeismo* is related to the presence of /r/-assibilation. To a lesser extent, the presence of /x/-aspiration seems to be linked to the velarization of /n/ (*Corr = 0.665*), while the other phonetic characteristics do not seem to be significantly correlated between each other.

Another element that could be useful to evaluate the relative importance of the analyzed variables is the average value of those variables in the total population, which is no other thing that the proportion in which each phonetic characteristics is present in that population. In this case the significant feature is that a variable has an average value close to 0.5, since a value which is very close to zero indicates that a characteristic is very unusual, while a value which is very close to one indicates that such a characteristic is so common that it is rare to find cases in which it does not appear. The values reported on the first column of Table 4 show that, according to this criterion, the most important variable is /x/-aspiration (μ *= 0.4554*), followed by the velarization of /n/ (μ *= 0,4045*) and the aspiration of /s/ (μ *= 0,3929*).

Variable	Average Value	R^2 when excluding		Regression w/ Difrer	
		w/ Innov	w/ Differ	Coefficient	t-stat
Seseo	0,9218	0,99992	0,99941	-0,0975	-79,54
Yeismo	0,9081	0,99968	0,99907	-0,0972	-6,43
Aspiration of /s/	0,3929	0,98259	0,99632	0,0344	16,10
Aspiration of /x/	0,4554	0,98850	0,99909	0,0157	8,09
Assibilation of /ʝ/	0,1637	0,99101	0,98572	0,0991	21,79
Assibilation of /r/	0,1123	0,99966	0,99892	0,1085	12,26
Velarization of /n/	0,4045	0,98944	0,99780	0,0324	20,12
Deaffrication of /tʃ/	0,1530	0,99140	0,98632	0,0991	14,88
Uvularization of /x/	0,0855	0,99991	0,99931	0,1304	56,99
Voicing of /tʃ/	0,0049	0,99996	0,99944	0,1262	80,50

Table 4. Statistical values associated with the phonetic variables

Besides the average value for each variable and its correlation indices with the other variables, the importance of the phonetic characteristics also has to do with their capability of explaining phenomena that the other variables do not explain. A way to evaluate that capability in this case is to perform a linear regression analysis of the linguistic innovation index on different sets of phonetic variables, and to evaluate the goodness of fit of the regressions through their coefficients of determination (R^2). As the innovation index is the sum of the values of the ten variables under analysis, a regression that had those ten variables as explanatory would have, by definition, an R^2 equal to one. If we alternatively calculate the R^2 of regressions that exclude one variable at a time (and that, therefore, use only nine of the ten phonetic variables), then we can obtain coefficients that show the reduction of the explanatory power of the regressions when we eliminate the variable under analysis. In order to perform those regressions, each of the 28 observations used (one for each dialect area) was weighted by the population represented by that observation.

The ranking of determination coefficients is therefore another clue for the relative importance of each phonetic variable (see Table 4, column 2). We can see that the largest reduction ($R^2 = 0.98259$) occurs when we exclude the aspiration of /s/, followed by the coefficients obtained when we exclude the aspiration of /x/ ($R^2 = 0.98850$) and the velarization of /n/ ($R^2 = 0.98944$). On the opposite extreme, the smallest reductions appear when we exclude the voicing of /tʃ/ ($R^2 = 0.99996$), the /s/-/θ/ merger ($R^2 = 0.99992$) and the uvularization of /x/ ($R^2 = 0.99991$). The origin of this ordering has to do with different factors, among which we can mention that /x/-aspiration, /n/-

velarization and /s/-aspiration are the variables whose average values are closer to 0.5, that /tʃ/-voicing is the variable with the smallest average value, and that *seseo* and /x/-uvularization are the variables with the largest correlation index in absolute value.

A similar procedure can be performed if we regress the linguistic differentiation index on the phonetic variables (see Table 4, column 3). As this index is not linear but quadratic, the regression analysis can also be carried out using the whole set of ten phonetic variables (i.e., without excluding any of them), in order to see which are the ones that generate more significant coefficients. The result of that analysis appears on the last two columns of Table 4, and shows that in this case, although all variables are statistically significant, the ones that exhibit higher t-statistics in absolute values are /tʃ/-voicing, *seseo* and /x/-uvularization (which seem to be the characteristics whose presence generate areas that are more differentiated in the Spanish-speaking world).

The individual exclusion of the different phonetic variables not only has quantitative effects on the coefficients of determination of an index's explanatory regression, but it also has qualitative effects on the dialect areas defined. By the way in which we have included the phonetic variables studied in this paper, dropping any of them from the matrix described on Table 1 has, as a direct consequence, the reduction in the number of dialect areas. Depending on which variable we exclude, the number of areas (which is equal to 28 when we use the ten variables under analysis) reduces to a value between 23 and 27, creating new areas that come from the union of the ones that disappear (see Table 5).

An important feature in this process of exclusion of variables is that, in some cases, the new regions are formed by the sum of areas that are not geographically contiguous. This occurs, for example, if we exclude the variable "aspiration of /s/", since when we do that we create a region which is the sum of the areas 7 (Eastern Andalusian) and 16 (Northern Andean), and another one which is the sum of the areas 11 (Central Mexican) and 27 (Cuyan). Similar problems of geographic incoherence arise when we try to exclude the variables "aspiration of /x/", "velarization of /n/" and "uvularization of /x/", since those exclusions create regions such as the ones that arise from joining the Central American area with the Peruvian Coast area (13+19), the Eastern Andalusian area with the Continental Caribbean area (7+15), and the Valencian area with the Central Mexican area (6+11).

Excluded variable	Number of areas				Non-compact areas
	Total	New	Compact	Non-comp	
Seseo	27	1	1	0	
Yeismo	25	3	3	0	
Aspiration of /s/	24	4	2	2	7+16, 11+27
Aspiration of /x/	24	4	1	3	7+27, 11+16, 13+19
Assibilation of /ʝ/	24	4	4	0	
Assibilation of /r/	25	3	3	0	
Velarization of /n/	23	5	2	3	7+15, 11+19, 13+16
Deaffrication of /tʃ/	24	4	4	0	
Uvularization of /x/	27	1	0	1	6+11
Voicing of /tʃ/	27	1	1	0	

Table 5. Result of the individual exclusion of phonetic variables

An alternative to find which are the most useful phonetic characteristics to define dialect areas in Spanish is therefore to choose the four variables whose exclusion creates non-compact areas (/s/-aspiration, /x/-aspiration, /n/-velarization and /x/-uvularization) and to discard the rest. The result of that alternative appears on Table 6, in which we find that, after following that procedure, we end up with eleven dialect areas (and ten of them are geographically compact). Five of the eleven areas belong to Spain, other five belong to Latin America, and the other one is the sum of two Spanish areas (Western Andalusian and Canarian) and two Latin American areas (Antillean Caribbean and Continental Caribbean). This last region is nevertheless compact, because the areas that belong to it are separated by the sea, but not by other intermediate areas in between. One of the Latin American regions that appear, however, does not satisfy this criterion of geographic coherence. That is the so-called "Mexican-High Peruvian" region (10+11+20), which arises from joining two contiguous North American areas with one South American area which is extremely far away from them.

Region / Variable	Aspir /s/	Aspir /x/	Velar /n/	Uvul /x/	Compact
Castilian (1-2/6)	No	No	No	Yes	Yes
Galician (3)	No	No	Yes	Yes	Yes
Manchego-murcian (4)	Yes	No	No	Yes	Yes
Extremaduran (5)	Yes	No	Yes	Yes	Yes
Eastern Andalusian (7)	Yes	Yes	No	No	Yes
Andalusian-Caribbean (8-9/14-15)	Yes	Yes	Yes	No	Yes
Mexican-High Peruvian (10-11/20)	No	No	No	No	No
Mexican-Central American (12-13)	No	Yes	Yes	No	Yes
Andean-Amazonic (16-18)	No	Yes	No	No	Yes
Peruvian Coast (19)	No	No	Yes	No	Yes
Southern Cone (21-28)	Yes	No	No	No	Yes

Table 6. Regions defined by variables whose exclusion generates non-compact areas

In order to divide this anomalous region that comes from the intersection of the isoglosses of the four isolated characteristics, it is necessary to include an additional variable, which can either be the /ʝ/-/ʎ/ merger (*yeismo*) or the assibilation of /r/. This is because, while the Northern and Central Mexican areas exhibit *yeismo* but no /r/-assibilation, the High Peruvian area exhibits no /ʝ/-/ʎ/ merger but its inhabitants typically assibilate the phoneme /r/. Using *yeismo* as an additional variable implies, consequently, to divide the region in two areas, which can be labeled "Western Mexican" (10+11) and "High Peruvian Andean" (20), but it also generates three new areas that are splits from the Castilian region, the Andean-Amazonic region and the Southern Cone region. Those are the Traditional Castilian area (1), the Equatorial Andean area (17) and a Bolivian-Paraguayan area (21+22), which are geographically compact but whose population does not exceed in any case the 3.2% of the Spanish-speaking world (see Table 7).

Region / Variable	**Innov Index**	**Differ Index**	**% Populat**
Variable: Yeismo			
Traditional Castilian (1)	1,00	0,5567	0,58%
Equatorial Andean (17)	3,00	0,4786	2,11%
Bolivian-Paraguayan (21-22)	3,77	0,5407	3,19%
Variable: Assibilation of /r/			
Amazonic-Equatorial (17-18)	3,23	0,4567	2,74%
Cordilleran-Chacoan (21-25)	4,12	0,5070	5,18%

Table 7. Comparison of splitted dialect areas

If, conversely, we use /r/-assibilation to divide the Mexican-High Peruvian region, we only obtain two additional areas (which are also compact), whose dimensions are a bit larger. Those areas are an Amazonic-Equatorial region (17+18) and a Cordilleran-Chacoan region (21+22+23+24+25).

Another way to compare the split of the Mexican-High Peruvian region that arises when we use the /ʝ/-/ʎ/ merger with the one that occurs when we use /r/-assibilation is to contrast the clustering of the different areas when we apply one criterion or the other. This can be represented through dendrograms such as the ones that appear on Figures 2 and 3. These dendrograms come from comparing the five chosen characteristics (the four main ones plus *yeismo*, on Figure 2, and the four main ones plus /r/-assibilation, on Figure 3), and the obtained clusters have therefore to do with the higher or lower dialect

closeness evaluated using those characteristics[5].

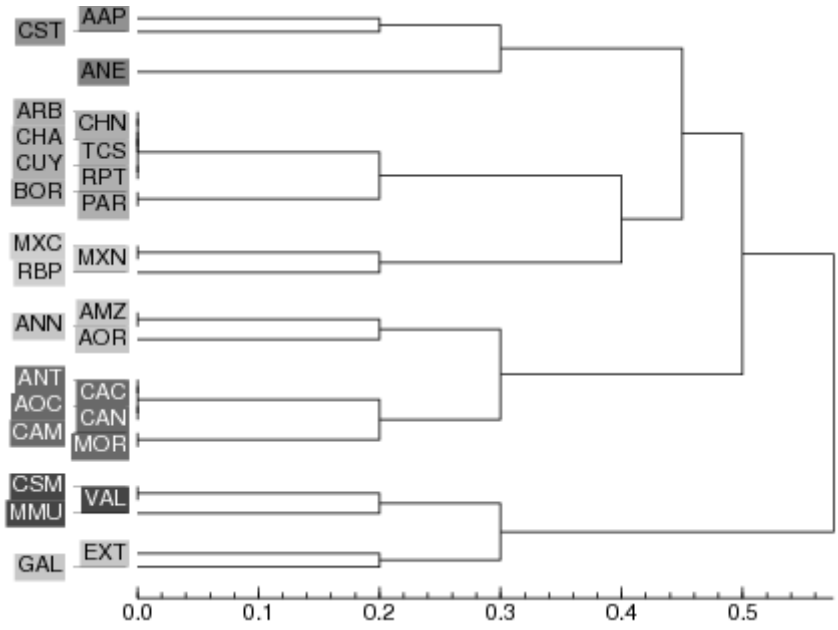

Figure 2. Dendrogram of eight clusters using *yeismo*

As Figures 2 and 3 show, the clusterings induced by the two schemes are somehow different. When we evaluate the closeness of Spanish dialect areas using *yeismo* as a relevant variable, the Traditional Castilian area (CST) is clustered with the High Peruvian Andean area (AAP) and the Equatorial Andean area (ANE). If we use /r/-assibilation, conversely, these two last areas group together with the Amazonic area (AMZ), and the Traditional Castilian area clusters with the Modern Castilian area (CSM), the Valencian area (VAL) and the Manchego-Murcian one (MMU). Therefore, if we evaluate these clusterings through a criterion of geographic coherence, there is a significant advantage for the use of /r/-assibilation as a relevant variable (Figure 3), in comparison with the use of the /ʝ/-/ʎ/ merger (Figure 2).

[5] For an explanation of this kind of analysis, see Nerbonne (2010). Figures 2 and 3 were generated using Gabmap.

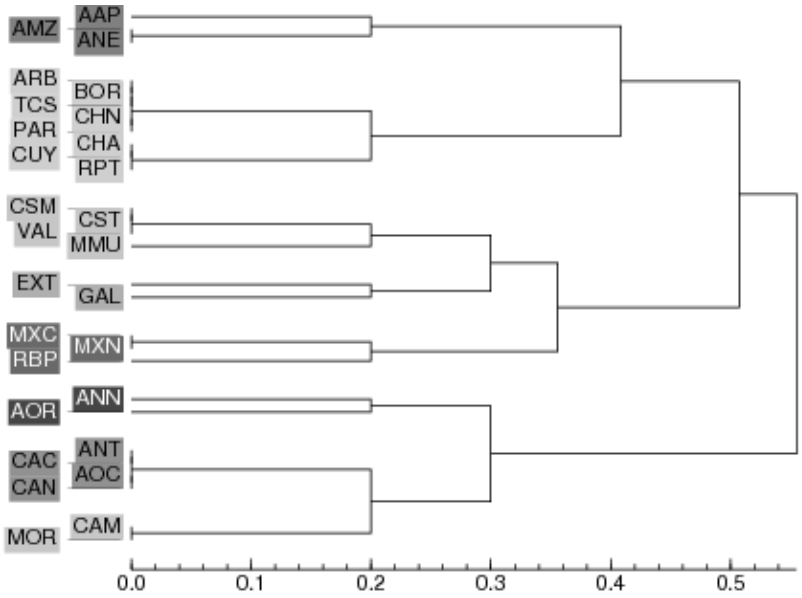

Figure 3. Dendrogram of eight clusters using /r/-assibilation

The results of our comparisons allow to state that the most important characteristics to define dialect areas in Spanish are /s/-aspiration, /x/-aspiration, /n/-velarization and /x/-uvularization, plus an additional characteristic that should be /r/-assibilation. Due to the inclusion of these last two variables, it is not necessary to include either *seseo* or *yeismo* (because these variables have a very large negative correlation with /x/-uvularization and /r/-assibilation). It is not relevant, either, to include the other analyzed phonetic variables (/ʝ/-assibilation, /tʃ/-deaffrication and /tʃ/-voicing), since they refer to relatively unimportant characteristics whose inclusion is not necessary for the defined regions to be compact.

3. A sequential method to define dialect areas

The method described in the previous section is based on the simultaneous definition of all the possible regions that arise from overlapping the isoglosses corresponding to the five most relevant phonetic variables. In some regions of the Spanish-speaking world, however, some of those variables could be relatively unimportant, and applying the same criteria to all areas can imply the identification of regions that are of little relevance as dialect units (see Table 6). That Galicia,

Extremadura, Eastern Andalusia and the Manchego-Murcian area are autonomous dialect regions, for example, is no doubt disproportionate if we observe that, on the other hand, the Western Andalusian area appears merged with the Canarian, Antillean Caribbean and Continental Caribbean areas, and that the Southern Cone region (Argentina, Chile, Paraguay and Uruguay) is also considered as a single dialect area (although the posterior introduction of /r/-assibilation as a relevant variable divides it in two regions).

A way to obtain compact areas of a larger dimension (and, presumably, of a higher significance as autonomous dialect regions) is to apply a sequential method that separates dialect areas through a minimum number of characteristics, and only includes new variables to divide regions that are considered too heterogeneous or non-compact. For the case of the phonetic variables analyzed in this paper, such a procedure can be applied using the five selected characteristics in a certain order. If, for example, we begin by including /x/-uvularization as a relevant variable, then we can isolate a compact region formed by the Traditional Castilian, Modern Castilian, Galician, Extremaduran, Manchego-Murcian and Valencian areas (areas 1 to 6), which can be jointly referred to as the "Northern Peninsular Region". After that, we could separate a second region characterized by the presence of /r/-assibilation (areas 17, 18, and 20 to 25), which we could name "Andean-Chacoan Region" (see Figure 4).

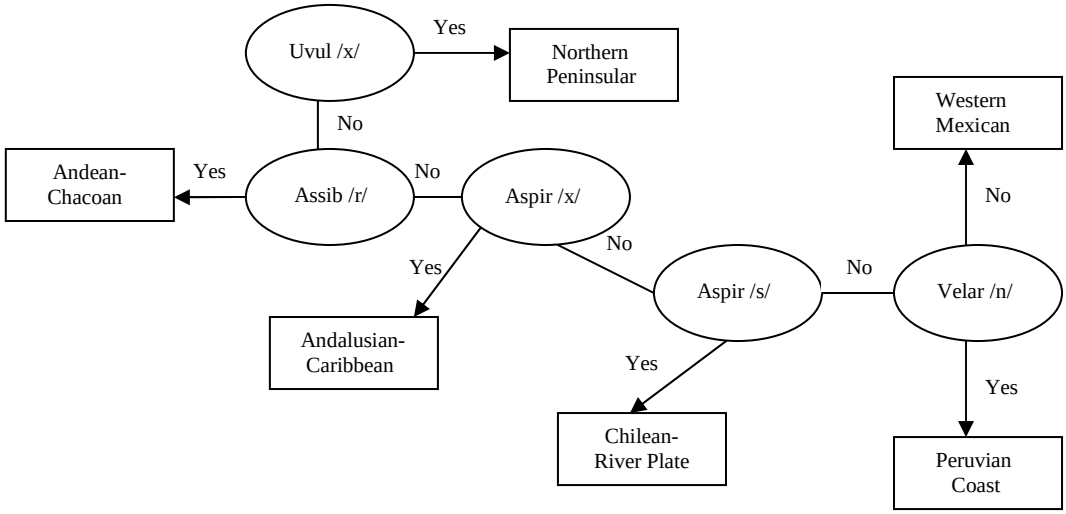

Figure 4. Sequential definition of dialect regions

Once we have defined these two first regions, the third compact region that can be splitted is the one that comprises the areas that aspirate the phoneme /x/, which in this case are areas 7 to 9 and 12 to 16 (and that can be generically referred to as "Andalusian-Caribbean region"). If we now subtract from the rest of the Spanish-speaking world the portion that aspirates the phoneme /s/, then we obtain a new compact region that includes areas 26, 27 and 28 (Chilean-River Plate region). The only additional task that we have is to separate the Peruvian Coast region (area 19) from the Western Mexican region (areas 10 and 11), for which we have to use the variable "velarization of /n/". With that the Spanish-speaking world ends up divided in only six regions, all of which are compact and dialectologically relevant. Although their dimensions are quite heterogeneous, all these regions have important cities inside their borders and none of them has less than 4.8% of the total population of the Hispanic world.

The division obtained through this sequential method can be compared with the one that arises from a clustering analysis using the variables /s/-aspiration, /x/-aspiration, /n/-velarization, /x/-uvularization and /r/-assibilation. On Figure 5 we see that such a procedure clusters the dialect areas in a very similar fashion than the sequential method, with the difference that the Central Mexican (MXC), Northern Mexican (MXN) and Peruvian Coast (RBP) areas appear together, and that the region formed by Paraguay, Eastern Bolivia, Northern Argentina and Northern Chile (ARB-BOR-TCS-CHN-PAR) appears together with the Chilean-River Plate region (CHA-CUY-RPT) instead of being clustered with the Andean region (AAP-AMZ-ANE).

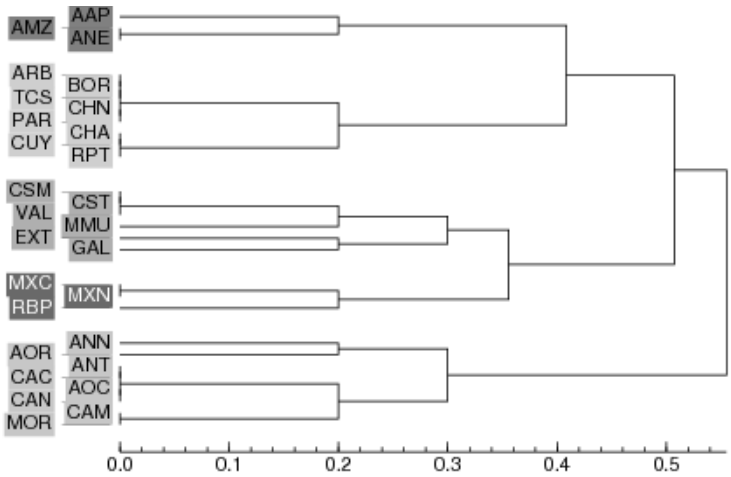

Figure 5. Dendrogram of five clusters and five variables

The sequential mechanism explained can also be generalized through the inclusion of new variables and the division of the obtained regions into smaller ones. On Figure 6, for example, we have used that mechanism to generate a division of the Spanish-speaking world in fourteen dialect regions, five of which belong to Spain and nine of which belong to Latin America. Those are a Northern Peninsular region (1+2+3+6), a Mixed Peninsular region (4+5), an Eastern Andalusian region (7), a Western Andalusian region (8), a Canarian region (9), a Western Mexican region (10+11), a Mexican-Central American region (12+13), a Caribbean region (14+15), a Northern Andean region (16), an Andean-Amazonic region (17+18+20), a Peruvian Coast region (19), a Cordilleran-Chacoan region (21+22+23+24+25), a Chilean-Cuyan region (26+27) and a River Plate region (28). In fact, the method is flexible enough to generate a larger or smaller number of regions according to the researcher's needs, up to a maximum of twenty-eight. That number could even be higher if we allow for other variables besides the ten phonetic characteristics analyzed in this paper.

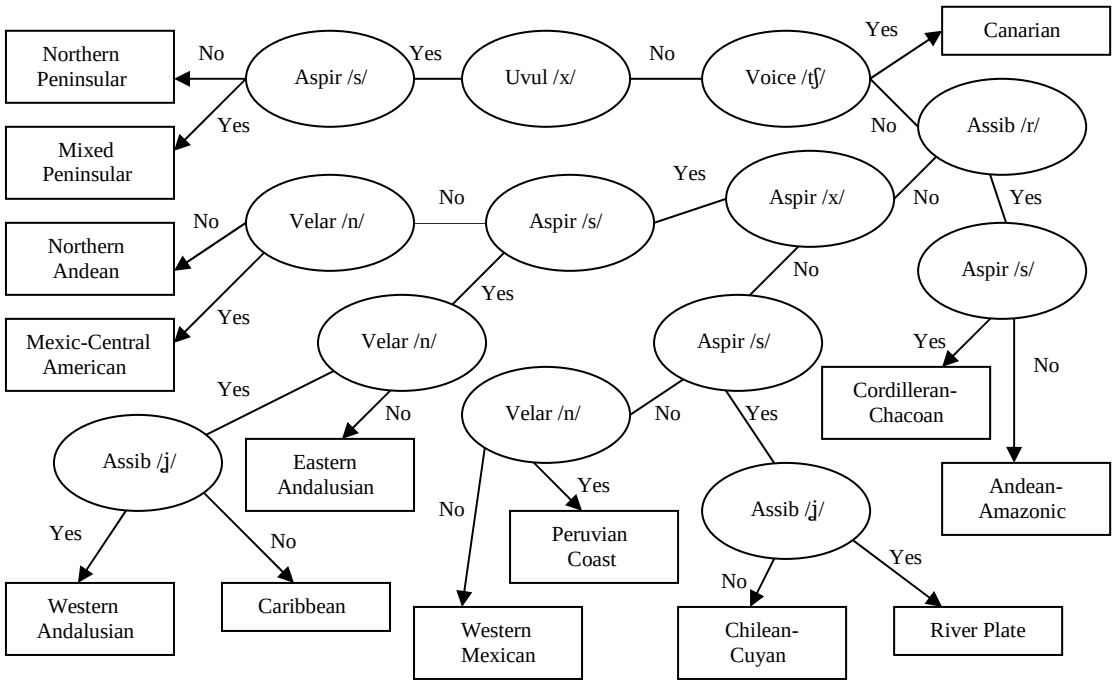

Figure 6. Generalized sequential definition of dialect regions

4. Final remarks

The main conclusion that we can obtain from the analysis performed in this paper is that the most important phonetic variables to define dialect areas in Spanish seem to be /s/-aspiration, /x/-aspiration, /n/-velarization, /x/-uvularization and /r/-assibilation. Each of them presents some advantage as a geolinguistic marker. While /s/-aspiration, /x/-aspiration and /n/-velarization have an average value which is close to 0.5 (that is, they divide the Spanish-speaking world into regions whose relative weight is relatively equivalent), /x/-uvularization and /r/-assibilation are by themselves able to generate compact regions with a certain linguistic homogeneity as dialect areas (the Northern Peninsular region, in the case of /x/-uvularization, and the Andean-Chacoan region, in the case of /r/-assibilation).

The abovementioned factors help to obtain a geographically coherent result when we use the five chosen variables for a clustering analysis. Besides, with the exception of /x/-aspiration with respect to /n/-velarization, these five phonetic variables have low correlation indices between themselves, so each of them is capable to explain different phenomena than the others. Finally, and as the main virtue of this combination of variables, we have found that they are the minimum possible set of characteristics whose isoglosses define compact dialect areas, and this is particularly true when we apply a sequential method like the one proposed in the previous section.

The reader may wonder why in this set of characteristics we have excluded both the /s/-/θ/ merger (*seseo*) and the /ʝ/-/ʎ/ merger (*yeismo*), which are supposed to be the most relevant phonological variables to describe the regional varieties of Spanish. This relevance is based on the fact that *seseo* and *yeismo* are the only characteristics that define "phonemic isoglosses", instead of purely phonetic or allophonic ones, and part of the dialectology literature considers that those isoglosses are generally more important.[6] This is not the case here, probably because *seseo* and *yeismo* are so widespread in the Spanish-speaking world that the population share of speakers that split the corresponding phonemes is relatively scarce. Besides, as the /s/-/θ/ split is so highly correlated with /x/-uvularization, and the /ʝ/-/ʎ/ split is so highly correlated with /r/-

[6] For an explanation of this structural theory of isogloss grading, see Chambers and Trudgill (1999), chapter 7.

assibilation, the additional inclusion of these characteristics does not help very much, provided that /x/-uvularization and /r/-assibilation are already included in the set of relevant variables that we have defined. Finally, we also have to point out that including *seseo* and *yeismo* as relevant geolinguistic variables can generate problems of geographic incoherence when we perform a clustering analysis. Indeed, those inclusions tend to induce that the Traditional Castilian area is grouped together with some South American areas, while the Valencian area tends to be clustered with some Latin American areas that do not aspirate the phonemes /s/ and /x/ (Peruvian Coast, Northern Mexican and Central Mexican).

Phonemic isoglosses are not always more important than purely phonetic ones in other languages besides Spanish. In Labov, Ash and Boberg (2007), for example, we find that the most relevant phonetic phenomena to define dialect areas in North American English are the "Northern Cities Chain Shift" (i.e., the shift in the articulation points of the phonemes /ɑ/, /æ/, /ɛ/, /ʌ/ and /ɔ/), the "Southern Vowel Shift" (i.e., the shift in the articulation points of the phonemes /i/, /ɪ/, /e/, /ɛ/, /o/ and /u/), the "Canadian raising" (i.e., the use of [ʌj] and [ʌw] as allophones for the diphthongs /aj/ and /aw/) and the "cot-caught merger" (i.e., the merger of the phonemes /ɑ/ and /ɔ/). Only the last of these four characteristics defines a phonemic isogloss, although in North American English we can find at least two additional phonetic variables (the so-called "father-bother" and "witch-which" mergers) that also generate phonemic isoglosses.

Summing up, our result about the importance of /s/-aspiration, /x/-aspiration, /n/-velarization, /x/-uvularization and /r/-assibilation as the main variables to define dialect areas in Spanish must be seen as relatively strong but provisional. This is so because it heavily depends on a given spatial distribution of the phonetic variables, and on a geographic coherence criterion. It nevertheless seems to us that this result can be relevant for future research that confirms or refute the existence of the postulated dialect borders, and that analyzes if those borders are actually perceived as important by the majority of speakers of the Spanish language.

Acknowledgements

I thank Florian Coulmas, Maria Pilar Perea, Daniel Perrin and Robert Shackleton for their useful comments. All remaining errors are mine.

References

BORLAND, Karen (2004) "La variación y distribución alofónica en el habla culta de Santiago de Chile", *Onomázein*, 10, 103-115.

CHAMBERS, J. K. & Peter TRUDGILL (1999) *Dialectology* (2nd edition), Cambridge: Cambridge University Press.

COLOMA, Germán (2011) "Valoración socioeconómica de los rasgos fonéticos dialectales de la lengua española", *Lexis*, 35, 91-118.

FONTANELLA, Beatriz (2000) *El español de la Argentina y sus variedades regionales*, Buenos Aires: Edicial.

GARCÍA-MOUTON, Pilar (1999) "Dialectometría", in José BLECUA (ed.), *Filología e informática*, Barcelona: Milenio, 335-357.

HUALDE, José (2005) *The Sounds of Spanish*, New York: Cambridge University Press.

LABOV, William, Sharon ASH & Charles BOBERG (2007) *Atlas of North American English: Phonetics, Phonology and Sound Change*, Berlin: Mouton.

LIPSKI, John (2004) *Lecture Notes on the Spanish of Spain*, University Park: Pennsylvania State University [Available at http://www. personal.psu.edu/jml34/readings.htm/].

MARTÍN-BUTRAGUEÑO, Pedro (2010) "La división dialectal del español mexicano", in Rebeca BARRIGA (ed.), *Historia sociolingüística de México*, Mexico City: El Colegio de México, chapter 24.

MORENO DE ALBA, José (2001)[3] *El español en América*, Mexico City: Fondo de Cultura Económica.

MORENO-FERNÁNDEZ, Francisco (2009) *La lengua española en su geografía*, Madrid: Arco Libros.

NERBONNE, John (2010) "Mapping Aggregate Variation", in Alfred LAMELI, Ronald KEHREIN & Stephan RABANUS (eds.), *Language and Space*, vol. 2, Berlin: Mouton, 476-495.

SAMPER, José (2008) "Sociolinguistic Aspects of Spanish in the Canary Islands", *International Journal of the Sociology of Language*, 193/194, 161-176.

SÉGUY, Jean (1973) "La dialectométrie dans l'Atlas Linguistique de la Gascogne", *Revue de Linguistique Romane*, 37, 1-24.

UTGARD, Katrine (2007) "Estudio geolingüístico de la fonética del español de Guatemala", *Voces*, 2, 137-190.

VILLENA, Juan (2008) "Sociolinguistic Patterns of Andalusian Spanish", *International Journal of the Sociology of Language*, 193/194, 139-160.

WORLD BANK (2011) *Population and Gross Domestic Product 2010*, Washington DC: International Reconstruction and Development Bank.

APPENDIX

REGIONS INCLUDED IN THE POPULATION FIGURES OF THE DEFINED DIALECT AREAS

1) Traditional Castilian Area (CST): Provinces of Alava, Burgos, Lérida, Palencia, Rioja, Segovia, Soria and Valladolid (Spain).

2) Modern Castilian Area (CSM): Provinces of Asturias, Avila, Baleares, Barcelona, Cantabria, Cuenca, Girona, Guadalajara, Guipuzcoa, Huesca, León, Madrid, Navarra, Rioja, Salamanca, Tarragona, Teruel, Toledo, Vizcaya, Zamora and Zaragoza (Spain).

3) Galician Area (GAL): Provinces of La Coruña, Lugo, Orense and Pontevedra (Spain).

4) Manchego-Murcian Area (MMU): Provinces of Albacete, Alicante, Ciudad Real and Murcia (Spain).

5) Extremaduran Area (EXT): Provinces of Badajoz and Cáceres (Spain).

6) Valencian Area (VAL): Provinces of Castellón and Valencia (Spain).

7) Eastern Andalusian Area (AOR): Provinces of Almería, Córdoba, Granada, Jaén and Melilla (Spain).

8) Western Andalusian Area (AOC): Provinces of Cádiz, Ceuta, Huelva, Málaga and Seville (Spain).

9) Canarian Area (CAN): Provinces of Las Palmas and Tenerife (Tenerife).

10) Northern Mexican Area (MXN): States of Baja California Norte, Baja California Sur, Chihuahua, Durango, Nayarit, Sinaloa, Sonora and Zacatecas (Mexico).

11) Central Mexican Area (MXC): States of Aguascalientes, Coahuila, Colima, Guanajuato, Guerrero, Hidalgo, Jalisco, Mexico, Michoacán, Morelos, Nuevo León, Puebla, Querétaro, San Luis Potosí, Tamaulipas, Tlaxcala and Distrito Federal (Mexico).

12) Eastern Mexican Area (MOR): States of Campeche, Oaxaca, Quintana Roo, Tabasco, Veracruz and Yucatán (Mexico).

13) Central American Area (CAM): Republics of Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica and state of Chiapas (Mexico).

14) Antillean Caribbean Area (ANT): Dominican Republic, Cuba and Puerto Rico.

15) Continental Caribbean Area (CAC): Republic of Panama, departaments of Atlántico, Bolívar, Cauca, César, Chocó, Córdoba, La Guajira, Magdalena, Nariño, San Andrés, Sucre and Valle (Colombia), provinces of El Oro, Esmeraldas, Guayas, Loja, Manabi and Galápagos (Ecuador) and states of Amazonas, Anzoátegui, Apure, Aragua, Balinas, Bolívar, Carabobo, Cojedes, Delta Amacuro, Falcón, Guárico, Miranda, Monagas, Nueva Esparta, Portuguesa, Sucre, Vargas, Yaracuy, Zulia and Distrito Capital (Venezuela).

16) Northern Andean Area (ANN): Departaments of Antioquia, Arauca, Bogotá, Caldas, Casanare, Cundinamarca, Huila, Meta, Norte de Santander, Quindío, Risaralda, Santander, Tolima and Vichada (Colombia) and states of Lara, Mérida, Táchira and Trujillo (Venezuela).

17) Equatorial Andean Area (ANE): Departaments of Boyacá and Putumayo (Colombia) and provinces of Azuay, Bolívar, Cañar, Carchi, Chimborazo, Cotopaxi, Imbabura, Los Ríos, Morona, Napo, Orellana, Pastaza, Pichincha, Sucumbios, Tungurahua and Zamora (Ecuador).

18) Amazonic Area (AMZ): Departaments of Amazonas, Caquetá, Guainía, Guaviare and Vaupés (Colombia) and departaments of Amazonas, Loreto, Madre de Dios y Ucayali (Peru).

19) Peruvian Coast Area (RBP): Departaments of Ancash, Arequipa, Cajamarca, Callao, Ica, La Libertad, Lambayeque, Lima, Moquegua, Piura, Tacna and Tumbes (Peru).

20) High Peruvian Andean Area (AAP): Departaments of Apurímac, Ayacucho, Cusco, Huancavelica, Huánuco, Junín, Pasco, Puno and San Martín (Peru) and departaments of Chuquisaca, Cochabamba, La Paz, Oruro and Potosí (Bolivia).

21) Eastern Bolivian Area (BOR): Departaments of Beni, Pando and Santa Cruz (Bolivia).

22) Paraguayan Area (PAR): Republic of Paraguay and provinces of Chaco, Corrientes, Formosa and Misiones (Argentina).

23) Argentine-Bolivian Area (ARB): Departament of Tarija (Bolivia) and provinces of Catamarca, Jujuy, La Rioja and San Juan (Argentina).

24) Tucuman-Saltean Area (TCS): Provinces of Salta, Santiago del Estero and Tucuman (Argentina).

25) Northern Chilean Area (CHN): Regions of Antofagasta, Arica, Atacama,

Coquimbo and Tarapacá (Chile).

26) Southern Chilean Area (CHA): Regions of Araucanía, Aysén, Biobío, Los Lagos, Los Ríos, Magallanes, Maule, O'Higgins, Santiago and Valparaíso (Chile).

27) Cuyan Area (CUY): Provinces of Mendoza and San Luis (Argentina).

28) River Plate Area (RPT): Republic of Uruguay and provinces of Buenos Aires, Chubut, Córdoba, Entre Ríos, La Pampa, Neuquén, Río Negro, Santa Cruz, Santa Fe, Tierra del Fuego and Capital Federal (Argentina).