Research Article

Francisco Moreno Fernández*, Hiroto Ueda

Cohesion and Particularity in the Spanish Dialect Continuum

https://doi.org/10.1515/opli-2018-0035
Received July 11, 2018; accepted December 7, 2018

Abstract: This paper studies the degree of cohesion among varieties of Spanish, proposing an analysis of Spanish dialectal variation and the internal cohesion of varieties using the Varilex-R database (2016). A battery of complementary statistical tests (correlation analysis, cluster analysis, association analysis) has been applied to these data in order to establish the distances between the principal modalities of the Spanish language. It also introduces the calculation of indices of generality and particularity, which, by establishing associations between linguistic uses within different countries, illustrate the extent to which each country's Spanish, by virtue of its linguistic uses, can be considered more general or more particular.

Keywords: dialectometry, linguistic cohesion, varieties of Spanish

Resumen: Este trabajo estudia el grado de cohesión entre las variedades del español, proponiendo un análisis de la variación dialectal hispánica y de su cohesión interna, a partir de la base de datos Varilex-R (2016). Sobre esos datos se ha procedido a aplicar una batería de pruebas estadísticas complementarias (análisis de correlación, análisis de clúster, análisis de asociación) con el fin de establecer las distancias existentes entre las principales modalidades del español. También se introduce el cálculo de un índice de generalidad y particularidad, que, estableciendo asociaciones entre los usos de unos países y otros, determina hasta qué punto el español de cada país, en virtud de sus usos lingüísticos, puede considerarse más general o más particular.

Palabras clave: dialectometría, cohesión lingüística, variedades del español

1 Introduction

The main goal of this study is to analyze the degree of cohesion among the varieties of the Spanish language that stretch across Europe, Africa, and the Americas. The analysis can be classified as being within the field of dialectometry, but it goes beyond the mere counting of coincidences and discrepancies to analyze the general degree of association among the language varieties, as well as their degree of particularity in comparison to the whole. This sort of analysis offers an interesting perspective for the study of major international languages, which until now has only been applied to languages other than Spanish, except in the cases of regional or qualitative approaches (Pennycook 1995, 2006; Lieberman 2007). These analyses have also been used to study affiliation and proximity between languages and dialects (Moore 1994; Borin and Saxena 2013), as this analysis does.

Spanish linguistics has generally maintained that the distance between varieties of Spanish is relatively

***Corresponding author: Francisco Moreno Fernández,** Cervantes Institute at Harvard University – Investigador del Instituto Franklin-UAH, E-mail: fmorenof@gmail.com
Hiroto Ueda, University of Tokyo

small, since interregional understanding poses few problems (Rosenblat 1962, Moreno de Alba 1978, Lapesa 1980, Thompson 1992, Lipski 1994, Rabanales 1998, Alvar 2002, Moreno Fernández and Otero 2016). However, the empirical evidence is not sufficiently broad or sound to demonstrate the degree of similarity among the different varieties of Spanish. The proposals for dialect zoning that have been made thus far have been based on partial and limited data or on inadequate methodology. Furthermore, these proposals have rarely assessed the linguistic relationships between geographically distant varieties.

This study proposes an analysis of the internal cohesion of the varieties of the Spanish language. To this end, our analysis will use the *Varilex-R* database: (Ueda and Moreno Fernández 2016) an updated, revised, and reordered version of the *Varilex* database (Ueda 1993). The *Varilex-R* database contains nearly 10,000 linguistic data points, related to 981 concepts, actions, expressions, or referents, which were gathered in 61 cities in 21 Spanish-speaking countries. A battery of complementary statistical tests (correlation analysis, cluster analysis, association analysis) has been applied to this vast collection of data in order to determine the distances between national modalities of Spanish. We also introduce indices of generality and particularity which, by establishing associations among the uses of language in different countries, determine the extent to which the Spanish of each country, by virtue of its linguistic uses, can be considered more general or more particular.

2 Spanish linguistic diversity

Over the last five centuries, the Spanish-speaking community has been shaped by geopolitical entities spanning several continents. These entities have maintained contact with one another in varying scopes and intensities, which, along with other linguistic and non-linguistic factors, has led to greater or lesser distances between their respective linguistic modalities (Moreno Fernández 2014a). Spanish-language scholars have generally considered the linguistic distance between varieties of Spanish to be relatively small due to the fact that interregional communication poses few problems of understanding, disregarding certain lexical and pragmatic discrepancies. Despite the near unanimity in this respect, there is, in fact, no empirical evidence that demonstrates exactly how close or how distant the different varieties of Spanish may be. They have, however, made different proposals for dialect zoning.

Indeed, the zoning of the Spanish language—that is, its division into dialect areas or zones—has been addressed with several approaches and from numerous angles over the last 150 years (Alba 1992, Moreno Fernández 1993a). Although the initial dialect division was made by Armas y Céspedes in 1882, the first proposal based on demographic, cultural, and geographical arguments was that of Pedro Henríquez Ureña (1921, 1930, 1931), who built upon earlier ideas to bring up another reasoning: regional Spanish was influenced by contact with indigenous languages. From a more narrow perspective, José Pedro Rona (1964) intended to differentiate dialect areas of American Spanish according to two phonetic factors (*yeísmo* and *rehilamiento*: *raya* ‘line’ and *ralla* ‘grate’, as [‘raj̯a] or [‘raʒa]) and two grammatical factors (*voseo* and verbal concordance: *vos tenés / vos tienes / tú tienes* ‘you have’). Using a similar method, Juan Zamora Munné (1975) proposed establishing the areas of American Spanish according to three features: *voseo*, the treatment of /s/, and the treatment of the velar /x/. Likewise, Raúl Ávila (2003) suggested the existence of three norms: Alpha Norm (*yeísmo*, no /θ/, no aspirated -/s/), Beta Norm (*yeísmo*, no /θ/, aspirated -/s/) and Gamma Norm (*yeísmo*, /θ/, no aspirated -/s/). Continuing in a phonetic vein, but without specifically aiming to discover dialect areas, Melvyn Resnick (1975) cataloged linguistic features according to geographical locations. Meanwhile, Philippe Cahuzac (1980) zoned American Spanish from the lexicon of agriculture, thus drawing attention to the field of dialectal lexicology.

In the 1990s, Hiroto Ueda used lexical data from 47 Spanish-speaking cities and referents for 206 concepts to calculate patterns of coincidences and correlations that enabled him to propose six large areas: Spain (and Africa); the Caribbean; Mexico; Chile; the Southern Cone; and, as a whole, Central America, Colombia, and Venezuela (Ueda 1995). For the first time, a proposal for dialect zoning was based on quantitative procedures applied to a comprehensive database of linguistic samples. The data handled by Ueda was more extensive than those previously used for zoning, though it was still limited (Ueda 2007, 2008).

Moreno Fernández’s subsequent proposal (2000, 2010) adopted a linguistic foundation and took a holistic perspective. He drew distinctions between the Castilian, Andalusian and Canary Island varieties of Spanish in Europe, while in the larger American territory he distinguished between the Mexican-Central American (including the southern United States), Caribbean, Andean, Chilean and Austral varieties. In addition to these, he categorized the Creole varieties of the Philippines and the Americas, as well as the Spanish from Equatorial Guinea. These are all worthy of consideration from the point of view of L1 and L2 language learning, as well as other areas of applied linguistics.

More recently, the Association of Academies of the Spanish Language has developed reference materials, including the *Corpus del español del siglo XXI* (CORPES XXI; Corpus of 21st-Century Spanish), with the collaboration and supervision of academics from various Spanish-speaking areas (RAE-ASALE 2005, 2010). The criteria used by the Academies for the representation of dialect areas are not strictly linguistic, nor are they presented as such. Instead, they combine the general geo-linguistic profiles of the main Spanish-speaking areas with the organizational function of the Academies.

Finally, it is worth mentioning the apparent contemporary tendency to distinguish two “superdialects” of Spanish (one urban and one rural) as a consequence of mass communication via the Internet and international media. According to this trend, the language varieties spoken in the world’s major Spanish-speaking cities will gradually coincide in their lexical uses (Gonçalves and Sánchez 2014, Moreno Fernández 2014b), while the particularities of different areas will be maintained in non-urban areas.

This presentation of proposals for geographic division of the Spanish-speaking space is neither detailed nor exhaustive, but are a minimum. Other approaches are possible, like studies of convergence and divergence processes (Moreno Fernández 1999-2000; Auer & Hiskens & Kerswill 2004; Soares da Silva 2006). All of them reflect the theoretical and practical difficulty of establishing a dialect inventory that is both principled and based on empirical evidence. See Table 1 for a summary of the proposals here presented.

Table 1. Proposals for dialect zoning of the Spanish language.

Authors	Areas	Criteria
Armas y Céspedes (1882)	Creole, Mexico and Central America, Pacific, Buenos Aires	Tendency to establish languages in the Americas
Henríquez Ureña (1921)	Mexico (+North America, Central America), Antilles, Andes, Chile, Argentina (+Uruguay, Paraguay, SE Bolivia)	American phonetics. Geography. Politics. Contact with indigenous languages
Rosenblat (1962)	Highlands / Lowlands	Phonetics. Influence of indigenous languages
Rona (1964)	16 areas defined by isoglosses	Isoglosses: *yeísmo, zeísmo, voseo*, verbal agreement
Resnick (1975)	256 areas. Feature index with geographical marking	Phonology. Identification of minimal dialect units
Zamora Munné (1980)	9 areas defined by the presence and absence of selected features	American phonetics: *voseo*, pronunciation of /x/, pronunciation of /s/
Cahuzac (1980)	4 large areas: I. Mexico, Central America, Antilles, Venezuela, Colombia. II. Andean Venezuela, Ecuador, Colombia, Peru, Bolivia, Northern Chile, Northwestern Argentina. III. Chile. IV. Argentina, Uruguay, Paraguay, Western Bolivia	Dialect lexicology. Agricultural lexicon
Ueda (1995)	Spain and Africa; Caribbean, Mexico; Central America, Colombia, and Venezuela; Andes; Southern Cone	Patterns of urban lexicon based on 206 concepts in 47 cities
Moreno Fernández (2000, 2010)	Mexican-Central American, Caribbean, Andean, Chilean, Austral, Castilian, Andalusian, Canary	Holistic linguistics
Ávila (2003)	Alpha, Beta & Gamma norms	Phonetics: use of aspirated /s/, use of /θ/ and *yeísmo*
RAE - ASALE (2005; 2010)	Chile, Rio de la Plata, Andes, continental Caribbean, Mexico and Central America, Antilles, United States and the Philippines, Spain	Academic committees
Gonçalvez and Sánchez 2014	Urban superdialect — rural superdialect	Lexical coincidences in Twitter messages

3 Preliminary methodological issues

One of the methods utilized in the general study of dialectal differences is the application of quantitative techniques from the field of study known as *dialectometry* (Séguy 1971; Guiter 1973; Goebl 1981, 1982, 2010). Quantitative methods are essential for avoiding impressionist approximations and subjective proposals, which often lead to interpretive biases (Moreno Fernández 1993b). Discussions of quantitative analysis have contributed to the understanding of variation and to the application of advanced methods of statistical inference and multivariate analysis whose use is regular in the dialectology of other languages, such as English, German, and Japanese (Houck 1967, Cichocki 1988, Viereck 1988, Thomas 1988, Ueda 2015), but is much less frequent in studies on Spanish linguistic geography (Aliaga Jiménez 2003, Ueda and Ruiz Tinoco 2003, García Mouton 1991, Moreno Fernández 1991). From this perspective, Hiroto Ueda's 1995 proposal was pioneering in its zoning of the Spanish-speaking space (Ueda 1995, Ruiz Tinoco 1999, 2002). In the last decade, diverse initiatives have attempted a quantitative analysis of linguistic distances between languages and varieties, and the use of information technology has been commonplace (Borin and Saxena 2016). However, statistical approaches have also been shown to have limitations; these limitations are due to diverse factors that have emerged not only in dialectology, but in various linguistic fields.

Overall, issues related to methodology and to the volume and quality of available data have most clearly influenced decisions about the study of the Spanish varieties, without prejudice toward traditions or ideologies that may have prevailed at prior times. This is because it is not possible to study the Spanish-speaking areas as a whole without having comparable data from each region. In the same way, a contrastive description of a specific dialect area cannot be carried out without sufficient data to make an adequate contrast between one area and another, or when there is only sufficient and reliable information for a single area. These hitches have largely delayed the dialectological task, along with intrinsic problems in the study of linguistic variation in general and lexical variation in particular. It could be said that, for decades now, linguistic knowledge has not been in a position to advance a new ideological paradigm for characterizing every country's specific features.

Analyzing the variation in a territory as large and extensive as the Spanish-speaking world requires a combination of appropriate methodology and adequate data. When these are unavailable, the task of gathering, organizing, and presenting linguistic information so that the particularity of each area is reflected in relation to the others can be organized in three veins. The first consists of gathering information from one area and contrasting it with the most complete and systematic information available from another area in order to compare and contrast both of them (*differential studies*). The second approach is to gather all of the linguistic features of a territory (e.g., lexical) and present them as a unified entity, omitting information about what is common or what is shared with other areas (*comprehensive studies*). The third is to draw on information provided collectively by experts from different areas of interest, in order to combine their data and identify what is shared and what is not (*complex studies*).

Finally, other seemingly minor methodological difficulties that may be decisive cannot be disregarded in the study of linguistic variation. Specifically, the use of nations or certain regions as units of reference for labeling linguistic features remains an artifice that overlooks geo-linguistic reality, which often contains international, transnational or local uses that are ignored, thus blurring the real landscape of the mosaic of language varieties. From this perspective, the linguistic geography of the Spanish language within Spain faces exactly the same problems as the language's other varieties. That is why the treatment of *españolismos* (Spain-isms), exclusive linguistic features from Spain, is as complex as that of the "-isms" in any other Spanish-speaking country (Moreno Fernández, in press).

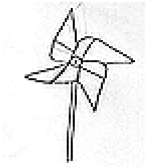
4 The *Varilex* project

The analysis of linguistic variation across the entire Spanish-speaking world requires managing large volumes of data using valid and reliable procedures. Only then will it be possible to understand the internal dynamics of this complex dialect continuum. The analysis proposed here is linked to the *Varilex* project

(V*ariación léxica del español en el mundo* —Lexical variation of worldwide Spanish), which was first developed by Hiroto Ueda (University of Tokyo) in the 1990s (Ueda 1995, Varilex 2015, Ueda and Ruiz Tinoco 2007). The most recent phase of the project, which began in 2016, is known as *Varilex-R* (Ueda and Moreno Fernández 2016).

Between 1993 and 2007, a massive amount of linguistic information was collected in 61 Spanish-speaking cities on several continents. The technique consisted of successively administered series of questionnaires that included questions regarding 981 lexical, phraseological, and syntactic aspects of contemporary Spanish spoken daily in urban environments. The lexical questions addressed nouns as well as verbs and adjectives. For example, see the question below regarding the concept B125: *windmill/pinwheel*.

Each item on the questionnaire consisted of words accompanied by a visual, and each offered options from which the informants could choose the most typical word or words (one or more) in their city; they could also provide alternatives that were not listed.



B125 [WINDMILL (US: PINWHEEL)] Juguete de papel recortado y doblado en forma de aspas que se fijan con un alfiler a un palito y que giran accionadas por el viento.[1]

1) abanico; 2) buscaviento; 3) estrella; 4) hélice; 5) molinete; 6) molinillo; 7) molinillo de viento; 8) molinito; 9) molino; 10) pajarita; 11) reguilete; 12) rehilete; 13) remolino; 14) remolino de papel; 15) ringlete; 16) veleta; 17) velete; 18) molinillo de papel; 19) molino de viento; 20) voladera. &) Otros: ______________________; #) No se me ocurre.

The questionnaires were completed by four individuals in each city: men and women, over and under 40 years of age. The responses from all 61 Spanish-speaking cities were combined in a large database. From that database, information could be presented in various formats and be subjected to different types of quantitative and qualitative analysis. For example, Table 2 shows the responses for the concept '*molinillo; hélice; rehilete*' ('windmill/pinwheel'), indicating the respondents' countries and cities (e.g., ES-MAD: Spain-Madrid), the responses obtained, and the number of informants that selected each response in each city.

In 2016, the *Varilex* database was subjected to a partial reconfiguration and a full revision. The reconfiguration consisted of combining the data across each country's cities so that the final classification would disregard information at the city level in favor of a classification by national territories. The country-level data was then subjected to a thorough review by experts from each of the countries. The purpose of this review was two-fold. First, it involved the correction of errors, which are inevitable when creating large databases. In addition, the experts went on to modify those linguistic elements that could not be considered to be general or majorities within each country, which led to the withdrawal of archaisms, dialectalisms, and especially jargon. The result of the reconfiguration was a new database, *Varilex-R* (2016), with information concerning the thousand concepts and referents considered during the first phase of the project and now reviewed and organized by country.

Thus, *Varilex-R* provides appropriate conditions for the quantitative analysis of the similarities and differences between linguistic uses across all Spanish-speaking countries, as well as of their internal cohesion. In fact, *Varilex*-R can be used to answer questions such as: What is the degree of similarity or dissimilarity of the variety of each of the Spanish-speaking nations with respect to all the others? What is the internal configuration of the Spanish dialect continuum, based on the lexical, phraseological, and grammatical uses analyzed? What is the ratio of every nation's exclusive linguistic features to the features in every other nation, according to the *Varilex-R* database? The collection of data at the national level necessarily means dispensing with the treatment and analysis of regional linguistic uses, although some of the data in the study may represent regional variations. At the same time, coincidence across different

[1] Engl$^{ish:}$ *Toy made from paper that has been cut out and folded into blade shapes, which are then affixed to a stick with a pin and which spins when blown by the wind.*

Table 2. Responses obtained for the lexical variants of the concept *'molinillo; hélice; rehilete*'' (windmill/pinwheel) by country and city.

Country codes: ES (Spain); CU (Cuba); RD (Dominican Republic); PR (Puerto Rico); EU (United States); MX (Mexico); EL (El Salvador); HO (Honduras); PN (Panama); CO (Colombia); VE (Venezuela); PE (Peru); PA (Paraguay); AR (Argentina); FIL: Philippines.
City codes: COR-A Coruña, SCO-Santiago de Compostela, STD-Santander, BAR-Barcelona VAL-Valencia SLM-Salamanca, GDL-Guadalajara, MAD-Madrid, VAL-Valencia, SEV-Seville, MLG-Málaga, IBI-Ibiza, TEN-Tenerife, PAL-Palma, HAB-Havana, SCU-Santiago de Cuba, STI-Santiago de los Caballeros, SJU-San Juan, DOR-Dorado, NOR-New Orleans, MON-Monterrey, MEX-Mexico City, SSA-San Salvador, TEG-Tegucigalpa, NAC-Nacaome, PAN-Panama City, MED-Medellín, BOG-Bogota, CBO-Maracaibo, TAC-Táchira, LIM-Lima, ASU-Asunción, SAL-Salta, BUE-Buenos Aires, NEU-Neuquén, MNL-Manila, ZBO-Zamboanga.

```
     E E E E E E E E E E E E E C C R P P E M M E H H P C C V V P P A A F F
     S S S S S S S S S S S S S U U D R R U X X L O O N O O E E E A R R I I
     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
     C S S B S G M V S M I T P H S S S D N M M S T N P M B C T L A S B M Z
     O C T A L D A A E L B E A A C T J O O O E S E A A E O B A I S A U N B
     R O D R M L D L V G I N L B U I U R R N X A G C N D G O C M U L E L O
                                                                          Total
  1) - - - - - - - - - - - - - - - 1 - - - - - - - - - 1 - 1 1 - - - - - -    4  1) abanico
  2) - - - - - - - - - - - - - - - - 2 2 1 - - - - - - - - - - - - - - 1 -    6  2) buscaviento
  3) - - - - - - - - - - - - - - - - 1 - - - - - - 1 - 1 - - - - - - - - -    3  3) estrella
  4) - - - - - - - - - - - - - - - 2 1 - 2 - - - 1 3 1 1 1 1 - - - - - - 1   14  4) hélice
  5) - - - - 1 - - 1 - - - - 1 - - - - - - 1 - - - - - - 1 - - - 2 3 2 - -   12  5) molinete
  6) 1 2 3 1 3 4 1 1 1 - 1 2 3 - - - - - - - - - - - - - - - - - - - - 2 -   25  6) molinillo
  7) 3 2 - 3 1 - 3 - - - 1 - 2 - - - 1 - - - - 1 - - 1 - - - 1 - 1 - - 1 -   21  7) molinillo de viento
  8) - - - - - - - - - - - - - - - - 1 1 1 - - - 2 - - 1 - - - - - - 2 - -    8  8) molinito
  9) - - - - - - - - - - - - - - - - 1 - - - - - - - - - 1 1 1 - - - - - -    4  9) molino
 10) - - 2 1 1 - - 1 - 1 - - 1 1 - - - - - - - - - - - - - - - - - - 1 1 -   10 10) pajarita
 11) - - - - - - - - - - - - - 3 2 - - - - - 3 - - - - - - - - - - - - - -    8 11) reguilete
 12) - - - - - - - - - - - - - - 2 - - - - 3 6 - - 1 - - - - - - - - - - -   12 12) rehilete
 13) - 1 - - - - - - - - - - - - - - - - - - - 1 - - - - - 1 1 - - - 1 - -    5 13) remolino
 14) - 2 - - - - - - - - - - - - - - - - - - - - - - - - - - 1 - - - - - -    3 14) remolino de papel
 15) - - - - - - - - - - - - - - - - - - - - - - - - - - 2 - - - - - - - -    2 15) ringlete
 16) - - - - - - - - - - - - - - - - - 1 - - - 1 - - 1 2 - - - - - - - - -    5 16) veleta
 18) - 2 1 1 - - - 2 2 - - - - - - - - - - - - - 2 - - - - - - - - - - 1 -   11 18) molinillo de papel
 19) 1 - - - 1 - - - - - - 2 - 1 - 1 - - - - - 1 - 1 1 - - - - - 1 - - 1 -   11 19) molino de viento
 20) - - - - - - - - 2 4 - - - - - - - - - - - - - - - - - - - - - - - - -    6 20) voladera
Sum. 5 9 6 6 7 4 4 5 5 5 2 4 7 5 4 4 7 4 4 4 9 4 5 6 4 6 5 4 5 0 4 3 6 7 1  170
```

national territories would allow for the discovery of broader linguistic areas, though not precisely enough to identify or define transnational areas by combining regions of several countries.

5 Quantitative analysis of Spanish dialectal variation

The large volume of data now available in *Varilex-R* (2016), the multiplicity of concepts and referents handled (nearly one thousand, all of which impact different classes of words, sentences, and phrases), as well as the number of geographical points included, make it possible to be optimistic about our goal. The intent is to know the level of cohesion of the dialect continuum of the Spanish-speaking community, including Spain, Equatorial Guinea, and all of the Spanish-speaking countries in the Americas.

To these data, sufficient in quantity and quality, a series of statistical techniques has been applied in order to answer several large-scale research questions, and two in particular: What is the level of linguistic homogeneity/heterogeneity across Spanish-speaking communities, and which areas are the most particular in terms of the characteristics of the Spanish language spoken there? To answer these research questions, we have carried out correlation, cluster, principal components, and association analyses, in addition to calculating indices of generality and particularity. Of all these analyses, the generality and particularity indices represent a contribution to the field of Spanish language dialectology.

6 Correlation analysis

Correlation calculations are widely practiced in dialectometry and are useful for determining the similarities between varieties of Spanish. The foundation of the dialectometric analysis offered here is a simple arrangement of the data, which are presented with their linguistic features (the responses obtained from the speakers) on one axis, and the country of origin of the linguistic uses on the other. This operation allows for the construction of co-occurrence tables, whose values represent the number of cases that appear at the intersection of each feature and each location. Therefore, data are displayed in a two-dimensional form, with linguistic forms on the vertical axis and countries on the horizontal axis.

Dialectometry often uses co-occurrence calculations; that is, counting the number of times that two specific locations agree on the choice of features and presenting that information in table format. From the table of co-occurrences, it is possible to calculate correlation coefficients. Specifically, in order to analyze language similarity among Spanish-speaking countries, we have chosen the correlation system known as the Jaccard index or Jaccard correlation coefficient (J.), which analyses the similarity between two or more sets of measurements, whatever type of data they contain. The formula is very simple:

$$J. = a \,/\, (a + b + c)$$

where a is the number of cases present in both sets of measurements; b, the number of cases that appear in the first set of measurements; and c, the number of cases that appear in the second set of measurements. Thus, instead of quantifying the absolute number of co-occurrences, a normalized figure on a scale between 0 and 1 is obtained. On that scale 0 means that the sets have no cases in common, and it tends to 1 as the two numbers b and c are reduced to zero. In this case, the groups are the 21 countries that have been taken into account for the analysis, and the correlations between pairs of countries are measured according to their similarity in the use of the 981 linguistic features analyzed.

Tabla 3. Jaccard correlation coefficients for linguistic traits and Spanish-speaking countries in the *Varilex-R* database.

Country codes: AR: Argentina, BO: Bolivia, CH: Chile, CO: Colombia, CR: Costa Rica, CU: Cuba, EC: Ecuador, EL: El Salvador, ES: Spain, GE: Equatorial Guinea, GU: Guatemala, HO: Honduras, MX: Mexico, NI: Nicaragua, PA: Paraguay, PE: Peru, PN: Panama, PR: Puerto Rico, RD: Dominican Republic, UR: Uruguay, VE: Venezuela.

J:(a)	ES	GE	CU	RD	PR	MX	GU	HO	EL	NI	CR	PN	CO	VE	EC	PE	BO	CH	PA	UR	AR
ES	1.000	.182	.317	.204	.284	.267	.172	.214	.115	.273	.089	.167	.067	.212	.194	.153	.159	.255	.241	.155	.291
GE	.182	1.000	.273	.331	.274	.260	.275	.273	.259	.259	.204	.266	.147	.312	.285	.279	.283	.153	.266	.279	.259
CU	.317	.273	1.000	.382	.478	.414	.290	.336	.205	.450	.158	.309	.129	.383	.324	.265	.272	.198	.377	.249	.404
RD	.204	.331	.382	1.000	.418	.321	.328	.329	.273	.343	.236	.314	.179	.387	.329	.321	.311	.159	.317	.270	.309
PR	.284	.274	.478	.418	1.000	.387	.318	.334	.207	.407	.180	.294	.141	.373	.324	.277	.283	.176	.336	.237	.376
MX	.267	.260	.414	.321	.387	1.000	.326	.382	.240	.468	.171	.318	.133	.387	.335	.275	.284	.186	.363	.261	.362
GU	.172	.275	.290	.328	.318	.326	1.000	.388	.290	.350	.233	.265	.184	.345	.332	.290	.325	.141	.273	.267	.277
HO	.214	.273	.336	.329	.334	.382	.388	1.000	.277	.430	.212	.296	.162	.360	.343	.284	.317	.173	.335	.274	.326
EL	.115	.259	.205	.273	.207	.240	.290	.277	1.000	.258	.485	.254	.251	.295	.260	.329	.271	.130	.221	.248	.202
NI	.273	.259	.450	.343	.407	.468	.350	.430	.258	1.000	.199	.338	.151	.389	.351	.268	.299	.196	.401	.264	.382
CR	.089	.204	.158	.236	.180	.171	.233	.212	.485	.199	1.000	.198	.241	.229	.216	.267	.233	.095	.170	.195	.162
PN	.167	.266	.309	.314	.294	.318	.265	.296	.254	.338	.198	1.000	.200	.345	.294	.276	.311	.144	.459	.312	.325
CO	.067	.147	.129	.179	.141	.133	.184	.162	.251	.151	.241	.200	1.000	.185	.225	.230	.201	.080	.144	.168	.132
VE	.212	.312	.383	.387	.373	.387	.345	.360	.295	.389	.229	.345	.185	1.000	.360	.349	.342	.180	.352	.313	.339
EC	.194	.285	.324	.329	.324	.335	.332	.343	.260	.351	.216	.294	.225	.360	1.000	.318	.356	.163	.336	.291	.321
PE	.153	.279	.265	.321	.277	.275	.290	.284	.329	.268	.267	.276	.230	.349	.318	1.000	.368	.143	.273	.290	.270
BO	.159	.283	.272	.311	.283	.284	.325	.317	.271	.299	.233	.311	.201	.342	.356	.368	1.000	.139	.324	.324	.305
CH	.255	.153	.198	.159	.176	.186	.141	.173	.130	.196	.095	.144	.080	.180	.163	.143	.139	1.000	.188	.152	.196
PA	.241	.266	.377	.317	.336	.363	.273	.335	.221	.401	.170	.459	.144	.352	.336	.273	.324	.188	1.000	.343	.475
UR	.155	.279	.249	.270	.237	.261	.267	.274	.248	.264	.195	.312	.168	.313	.291	.290	.324	.152	.343	1.000	.351
AR	.291	.259	.404	.309	.376	.362	.277	.326	.202	.382	.162	.325	.132	.339	.321	.270	.305	.196	.475	.351	1.000

A cursory review of the correlation coefficients between these Spanish-speaking countries reveals cases of marked similarity and cases of evident dissimilarity. Looking at the indices greater than .400 (greater similarity), those corresponding to the Caribbean island countries (Cuba, Puerto Rico, Dominican Republic) can be observed, as well as the Central American countries, with each other and with Mexico; Nicaragua in particular has high correlations with Central America and Mexico, as well as with the Greater Antilles. Looking at the indices below .150 (lower similarity), those corresponding to Spain with Central America, especially El Salvador, can observed; those of Chile with respect to Central America; and those of Colombia in relation to Chile and Argentina. The proximity between Panama or Cuba and Argentina, as well as the heterogeneity of Central America, would require a particular and more detailed analysis.

7 Cluster analysis

One of the most commonly used methods in dialectometric studies is cluster analysis, which is often used for taxonomic purposes (Clua 2010; Goebl 2010) and has been used by Hiroto Ueda on a subset of the *Varilex* data (Ueda 1995, 2008, 2015). Cluster analysis visualizes the way in which the data can be grouped by their similarities. To properly understand a cluster graph, such as the one shown in Figure 1, it is necessary to start on the right side, where the first branch is found at the point corresponding to EL (0.067): On one side, the countries from ES to EC and, on the other side, from EL to CO. Within the first group, at the GE point (0.139), there is a limit from ES to CH and from GE to UR. That being so, it is possible to reach successively the branches that represent each of the countries; the farther to the left in the graph, the lower the indices of union point. If the figure is seen as a tree (dendrogram), it can be said that the lower branches, located in the left side, correspond to greater distance between the countries, while the higher branches, to the right, correspond to countries more closely grouped together.

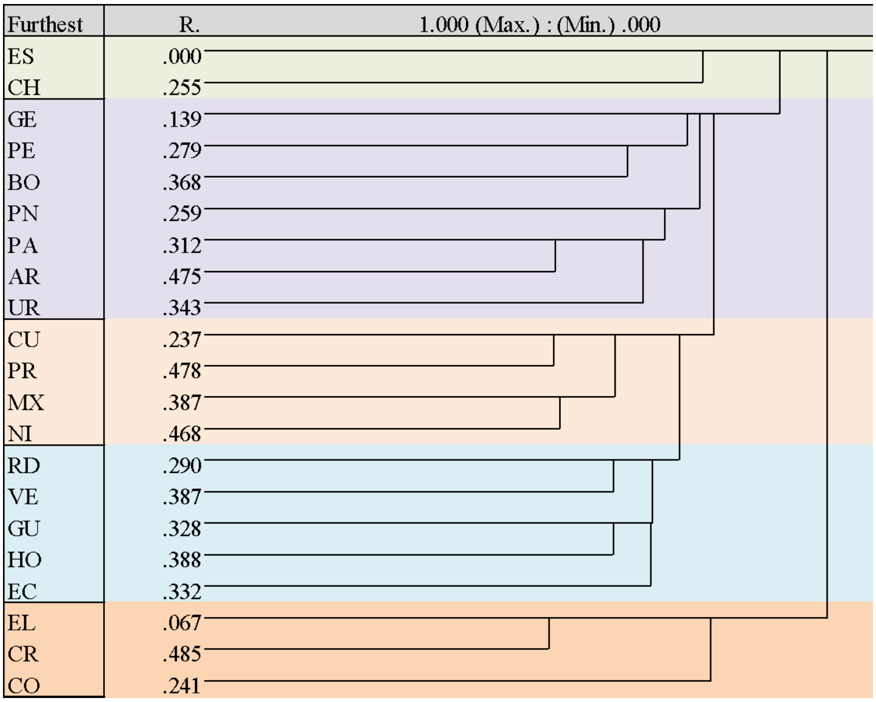


Figure 1. Cluster analysis (mean distance method) of Spanish-speaking countries in the *Varilex-R* database. Country codes: AR: Argentina, BO: Bolivia, CH: Chile, CO: Colombia, CR: Costa Rica, CU: Cuba, EC: Ecuador, EL: El Salvador, ES: Spain, GE: Equatorial Guinea, GU: Guatemala, HO: Honduras, MX: Mexico, NI: Nicaragua, PA: Paraguay, PE: Peru, PN: Panama, PR: Puerto Rico, RD: Dominican Republic, UR: Uruguay, VE: Venezuela.

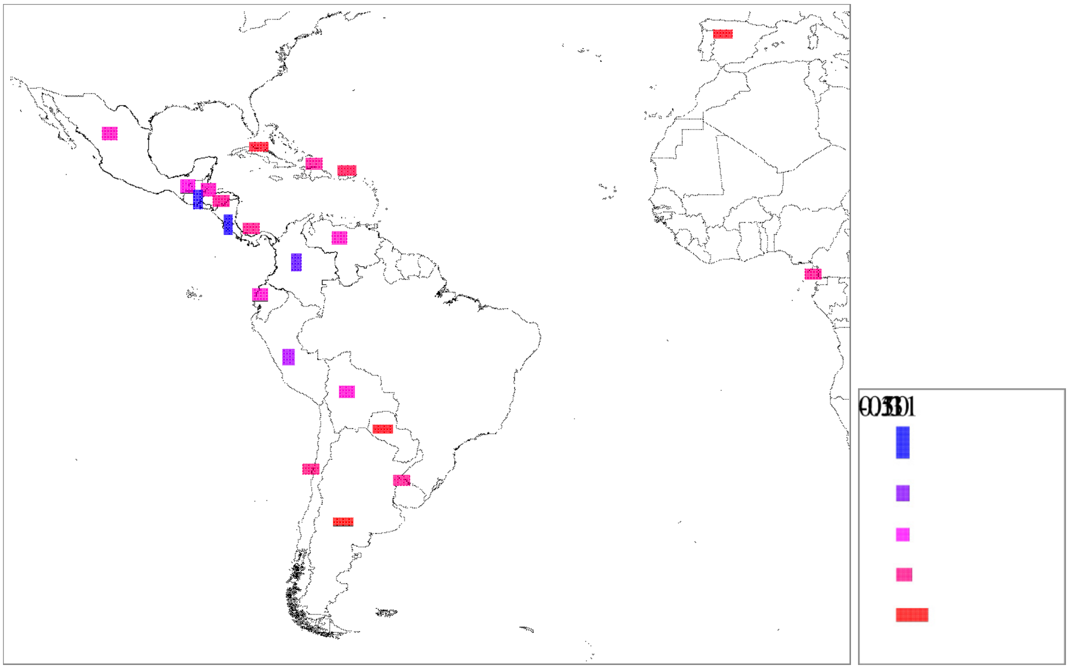
The linguistic features analyzed for the 21 Spanish-speaking countries allow us to build the dendrogram, whose most clear-cut grouping of countries is as follows: South America (Peru, Bolivia, Argentina, Uruguay, Paraguay) with Equatorial Guinea; the Caribbean, insular and continental (Cuba, Puerto Rico, the Dominican Republic, Mexico, Nicaragua); as well as, independently, Spain and Chile. El Salvador, Costa Rica and Colombia form their own group. It can be seen that Colombia and Ecuador are both separated from the Andean branch and closer to Central America and the continental Caribbean. The graph, therefore, breaks up some of the geo-linguistic units that have traditionally been conceived as blocks, in particular the Caribbean, Central America, and the Andes, while marking Spain, Chile, El Salvador, Costa Rica, and Colombia as clearly separate from the other varieties.

8 Principal component analysis

Principal component analysis is a complex technique that offers a multi-dimensional view of the distribution of both the linguistic features included in the database and of the countries. This is a somewhat more sophisticated method in which multivariate calculations on the coefficient matrix allow us to find several regression lines in the form of "components," which group together the greatest possible amount of information (Woods et al. 1986: 273-290). Applying principal components analysis to the geo-linguistic data of *Varilex-R* yields a table in which each country receives a different index according to the component in question (1, 2, 3, 4).

Table 4. Principal component analysis by country in the *Varilex-R* database. Eigenvalues: 6.645; 1.361; 1.182; 1.145; 0.945.

PCA	[1]	[2]	[3]	[4]	[5]
ES	-0.013	-0.279	0.649	-0.037	0.098
GE	-0.220	-0.090	0.192	0.019	0.144
CU	-0.216	-0.260	0.006	0.287	0.227
RD	-0.250	-0.075	0.106	0.228	0.277
PR	-0.218	-0.191	0.046	0.367	0.305
MX	-0.217	-0.001	-0.195	0.216	-0.359
GU	-0.241	0.055	0.030	0.219	-0.236
HO	-0.237	-0.004	-0.017	0.173	-0.401
EL	-0.232	0.450	0.144	-0.055	-0.028
NI	-0.241	-0.082	-0.093	0.215	-0.336
CR	-0.213	0.457	0.166	-0.030	0.069
PN	-0.230	-0.101	-0.229	-0.264	0.027
CO	-0.177	0.358	0.016	-0.152	0.169
VE	-0.259	0.006	0.011	0.072	0.023
EC	-0.242	0.018	0.002	-0.011	-0.025
PE	-0.235	0.212	0.067	-0.112	0.203
BO	-0.249	0.025	-0.027	-0.153	0.044
CH	-0.042	-0.120	0.566	-0.234	-0.45
PA	-0.231	-0.281	-0.202	-0.345	-0.045
UR	-0.219	-0.092	-0.075	-0.400	0.015
AR	-0.213	-0.310	-0.088	-0.295	0.076



Map 1.-Distribution of values of component [2] by country in the *Varilex-R* database.

These values can be shown graphically in a scatter plot of the first two components. The first component, plotted on the horizontal axis of Figure 2, shows values with a homogeneity that makes it difficult to individualize most of the countries. These values do, however, make it possible to distinguish clearly between Spain and Chile on one end of the figure and the rest of the countries on the other. As is well known, in a principal component analysis, the greatest variance by a projection of the data comes to lie on the first coordinate (Component 1), the second greatest variance on the second coordinate (Component 2), and so on.

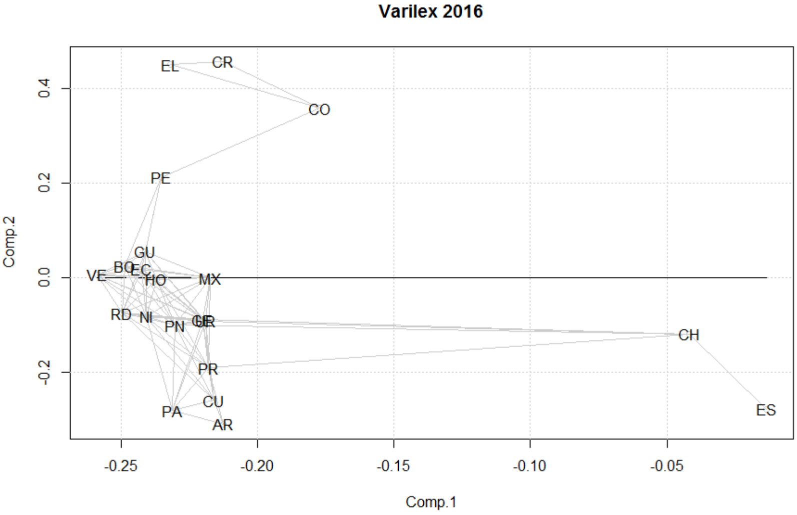


Figure 2. Representation of the first two principal components in the *Varilex-R* database. Country codes: AR: Argentina, BO: Bolivia, CH: Chile, CO: Colombia, CR: Costa Rica, CU: Cuba, EC: Ecuador, EL: El Salvador, ES: Spain, GE: Equatorial Guinea, GU:

Guatemala, HO: Honduras, MX: Mexico, NI: Nicaragua, PA: Paraguay, PE: Peru, PN: Panama, PR: Puerto Rico, RD: Dominican Republic, UR: Uruguay, VE: Venezuela.

The second principal component, shown on the vertical axis, takes both positive and negative values, which allows for other types of groupings. Thus, Spain, Argentina, Paraguay, and Cuba have the lowest values, while El Salvador and Costa Rica have the highest. Peru and Colombia also have values greater than 0.2. The remaining countries are grouped between -0.2 and 0.1.

The principal component analysis provides a graphical representation of the degree of cohesion of the set of Spanish modalities, a set in which Spain and Chile stand out for their dissimilarity with respect to the remaining countries. The close concentration of most of the countries, as well as the large distance separating Argentina and Paraguay from El Salvador and Costa Rica, is also remarkable. This graphical representation yields varied and useful conclusions regarding the conception and perception of the place that each nation occupies with respect to the others, with all of their associated historical, geopolitical, and ideological implications.

9 Association analysis

The analyses presented so far have used macroanalytic methods, which operate on complex, multidimensional data. However, it is also possible to approach dialectal reality from a microanalytic point of view. In order to achieve this, we propose a technique used in e-commerce systems to generate recommendations; for example, the purchase of a book A might trigger the recommendation of books B and C. What are these recommendations based on? Essentially, the system is based on data derived from previous sales. In this way, for a matrix consisting of five rows (1 - 5) and four columns (A, B, C, D), it is possible to record the number of matches or coincidences between two columns; that is, the number of rows in which both of the columns contain the value 1.

	A	B	C	D
1	1	1	0	0
2	0	0	1	0
3	0	1	0	0
4	0	0	1	1
5	1	1	1	0

In the sample matrix, there are two coincidences between A and B, i.e., two rows (1 and 5) where the value 1 appears in both A and B. Consequently, it is possible to define the Support between A and B as follows:

Support (A, B) = 2 / 5 = 0.4

The Support between A and B is equal to the number of matches divided by the sample size (in this case, 5). The greater the Support, the greater the degree of association between A and B.
Next, the degree of Confidence can be calculated, which is the conditional probability that when A is selected, B is also selected. This is equal to the number of coincidences between A and B, divided by the number of times A is selected:

Confidence (A, B) = 2 / 2 = 1.0

This represents the probability of buying B among those who have bought A. It is not necessarily equal to the probability of buying A among those who have bought B, since the impact of A on B may differ from the impact of B on A, as seen here:

Confidence(B, A) = 2 / 3 = 0.667

Association analysis involves another very important value, the Lift. The Lift compares the Confidence(A, B) with the probability of B.

> Lift(A, B) = Confidence(A, B) / Probability(B). En our example,
> Lift(A, B) = (2 / 2) / (3 / 5) = 1 / 0.6 = 1.667

Here, then, is the degree to which the purchase of A contributes to the purchase of B. The probability of B is obtained by dividing the frequency of B by the sample size (3 / 5 = 0.6). The higher the Lift, the greater the contribution of A to B. Like the Confidence, the Lift is unidirectional: A => B, i.e., Lift(A, B) is not in general equal to Lift(B, A).

This type of association analysis has been applied to the *Varilex-R* database, which contains 9,886 rows (variant linguistic forms of the 981 variables or concepts) and 21 columns (the countries). The results of the analysis are shown in Table 5, which shows the direction of influence (Left; Right) and the frequencies of each column (FLeft; FRight).

On the basis of this information, the Support (Sup), the Confidence (Conf), and the Lift (Lift) have been calculated. The table is sorted by descending order of Lift (i.e., from highest to lowest):

Table 5. Sample of association measures between countries based on the *Varilex-R* database.

Aso.	Left	=>	Right	FLeft	FRight	Sup.	Conf.	Lift
[1]	CR	=>	EL	986	1381	773	0.078	5.612
[2]	EL	=>	CR	1381	986	773	0.078	5.612
[3]	CO	=>	CR	835	986	354	0.036	4.251
[4]	CR	=>	CO	986	835	354	0.036	4.251
[5]	CO	=>	EL	835	1381	445	0.045	3.815
[6]	EL	=>	CO	1381	835	445	0.045	3.815
[7]	CR	=>	PE	986	1827	592	0.060	3.249
[8]	PE	=>	CR	1827	986	592	0.060	3.249
[9]	CO	=>	PE	835	1827	498	0.050	3.227
[10]	PE	=>	CO	1827	835	498	0.050	3.227

The Lift value shows the degree of association between the countries that appear on the left side of each row and in the indicated orientation. Clearly, given all possible combinations and orientations of countries, the amount of data processed is large and complex; hence the expression "data mining" (Hahsler, Grun, Hornik, Buchta 2016).

10 Generality and particularity indices

The arrangement of these 21 countries' linguistic features in a matrix format suggests an analysis of each country's degree of generality and particularity. Every cell in the matrix represents the appearance of a single feature in one country. From this matrix, it is possible to count the number of common features for each country; that is, the degree of "communality." Thus, the higher the communal value of a feature, the greater its degree of "generality," in the same way that someone who votes for the winning party represents the highest "generality" of the vote.

In the analysis of the *Varilex-R* database, when the common features for each country are added up, a sum is obtained that represents the degree of generality by the absolute frequency in each country (a.freq.G). The degree of generality may also be represented by the relative frequency (r.freq.G), obtained by scaling the absolute frequency so that its values are always between 0 and 1. Conversely, the degree of particularity can be measured (r.freq.P) from the degree of generality through simple subtraction.

r.freq.P = 1 - r.freq.G

In order to interpret these values, it should be taken into account the fact that the degree of generality is always considerably lower than the degree of particularity, not due to any characteristics of the data themselves, but simply because of the number of variables considered. To obtain a measure that is not affected by the number of variables, the "prominent relative frequency" (p.r.freq.G.) is calculated by increasing the value of the absolute frequency of the variable in question, multiplied by the number of variables being compared. From this prominent frequency of generality, it is also possible to calculate the prominent relative frequency of particularity (p.r.freq.P.). Table 6 shows all of these possible values calculated from the *Varilex-R* database.

Table 6. Values of generality and particularity of the Spanish-speaking countries in the *Varilex-R* database.

Generality	Sum	a.freq.G.	r.freq.G.	r.freq.P.	p.r.freq.G.	p.r.freq.P.
ES	4829	29547	.057	.943	.550	.450
GE	1645	18870	.037	.963	.433	.567
CU	4210	36195	.070	.930	.602	.398
RD	2203	24438	.048	.952	.500	.500
PR	3514	31971	.062	.938	.570	.430
MX	4253	36053	.070	.930	.601	.399
GU	1918	21711	.042	.958	.469	.531
HO	2575	26488	.052	.948	.521	.479
EL	1381	16703	.033	.967	.402	.598
NI	3715	34355	.067	.933	.589	.411
CR	986	12603	.025	.975	.335	.665
PN	2179	23170	.045	.955	.486	.514
CO	835	10143	.020	.980	.287	.713
VE	2635	27918	.054	.946	.535	.465
EC	2253	24348	.047	.953	.499	.501
PE	1827	20550	.040	.960	.454	.546
BO	1793	20987	.041	.959	.460	.540
CH	2336	16476	.032	.968	.398	.602
PA	3025	29472	.057	.943	.549	.451
UR	1803	19893	.039	.961	.446	.554
AR	3542	31978	.062	.938	.570	.430

Country codes: AR: Argentina, BO: Bolivia, CH: Chile, CO: Colombia, CR: Costa Rica, CU: Cuba, EC: Ecuador, EL: El Salvador, ES: Spain, GE: Equatorial Guinea, GU: Guatemala, HO: Honduras, MX: Mexico, NI: Nicaragua, PA: Paraguay, PE: Peru, PN: Panama, PR: Puerto Rico, RD: Dominican Republic, UR: Uruguay, VE: Venezuela.

Table 6 shows the ordering of the countries, where Spain has the largest sum of shared elements (column "Sum"), followed by Mexico, Cuba and Nicaragua. At the opposite extreme, the countries with the lowest values are El Salvador, Costa Rica, and Colombia, in that order. However, the simple sum does not represent the degrees of "generality" and "particularity" of each country. To that end, the weighted values in the column "a.freq.G.", generality by absolute frequency, should be considered. That being so, the order changes, with Cuba appearing first, followed by Mexico and Nicaragua, and with Chile, Costa Rica, and

Colombia at the end. The ordering is of course maintained in the three versions of "generality:" generality by absolute frequency (a.freq.G), by relative frequency (r.freq.G.) and by prominent relative frequency (p.r.freq.G). When considering each country's degree of particularity, it must be kept in mind that the degrees of "particularity" have the inverse order of the degrees of "generality." Therefore, the countries that present a higher degree of particularity, according to the *Varilex-R* database, are Colombia, Chile, and Costa Rica, while Spain's degree of particularity, contrary to what might be inferred from other tests, appears diluted among the majority of the Spanish-speaking countries.

11 Discussion

Quantitative analysis provides valuable information for interpreting the relationships between different varieties of the same language. In fact, the statistical evaluations applied here allow us to overcome some of the difficulties found in previous analytical approaches. Moreover, the compilation of linguistic uses linked to numerous concepts, referents, or expressions in all of the Spanish-speaking countries provides an ample base for new analyses and interpretations. All of this collectively has enabled us to propose hypotheses and interpretations based not on four phonetic features, as in Rona (1964), or even on 100 maps of linguistic features, as in the first dialectometry (Séguy 1971; Moreno Fernández 1991), but on nearly a thousand linguistic uses; based not on a single semantic sphere, as in Cahuzac (1980), but on numerous references; based not on hundreds of observations, as in the original *Varilex* of the 1990s (Ueda 1995), but on thousands; obtained not through a single analytical technique, as has been done in numerous studies, but through using a combination of statistical tests; using not disjoint data, but a cohesive database analyzed for its associations and internal correspondences.

Although this study has overcome some of the informational and methodological limitations that have traditionally affected dialectometric investigations, it of course has limitations of its own. This work has handled data from different linguistic fields (lexicology, grammar, phraseology), but has not conducted segregated or partial analyses of each of them individually, although such analyses may be undertaken in the future. We present a holistic view of the language, rather than a breakdown by levels, in order to appreciate the dialectic reality as a whole. Similarly, some of the tests that have been conducted treat all the versions of each variant equally, without considering which uses are primary or secondary or which belong to the active versus passive register of the speakers in each region of each country. The analysis of association and the indices of generality and particularity proposed here do account for the fact that the linguistic options for each variable are related to the speakers' choices, hence the interest and novelty in the panorama of contemporary dialectometry. Finally, the use of national territorial units prevents the discovery of information and conclusions relating to both the internal regions that comprise national territories and the transnational regions that undoubtedly exist within the Spanish-speaking world. In terms of territories by country, it is important to consider that *Varilex* deals with modern urban lexical variation, where a higher homogeneity is observed within a nation, because of the more intense communication that is presupposed within it. This situation would be opposed to traditional rustic lexical variation, where different degrees of heterogeneity within a territory would be expected.

The statistical operations conducted here from the *Varilex-R* database should be understood as complementary; the results are not absolute; they are relative, verifiable, and modifiable with respect to the results of the other tests. Together, these analyses reveal a great degree of cohesion among the varieties of the Spanish language, with a remarkable degree of coincidence of absolute values. This can be seen in the close concentration of countries revealed by the charts from both principal components analysis and association analysis. This study thus provides empirical support for the generally acknowledged homogeneity within the Spanish-speaking domain world, complementing the results of other studies that have reported a significant amount of shared uses. For example, several studies have reported the percentages of the lexicon shared in the press, radio, and television in the Spanish-speaking countries to be over 90% (López Morales 2006: 188-190).

Beyond the general cohesion of the Spanish dialect continuum, it is important to discover where the greatest discrepancies occur; that is, the national modalities that exhibit the greatest dissimilarity in their linguistic uses. The linguistic tests presented here show that the data coincide to a large degree, and also provide complementary results: The counts of the absolute frequencies of the national exclusive features, the cluster analysis, and the principal components and association graphs all reveal that Spain is the country with the most particular linguistic features, hence its distant location with respect to the rest of the countries in the graphical representations. Similarly, Chile also appears separated from the nucleus of countries with similar linguistic uses. The separation of these two countries may be interpreted with respect to their geographically peripheral positions, which would favor their archaistic nature. At the same time, this particular nature appears diminished when considered from the point of view of international communication, in which Spain shows a great ability to penetrate its modalities through the global media of social communication (López Morales 2006: 192-193).

This peculiar and geographically peripheral nature of Spain, contrasted quantitatively, supports the international community's perception of the Spanish spoken in Spain, which is identified with northern peninsular Spanish. Indeed, a study of perceptual dialectology conducted in 2015 showed that Spanish speakers perceive most of the Spanish-speaking domain as a homogeneous space, with little distance between most of its varieties, especially those located in the Americas. At the same time, among all the Spanish modalities, those that are perceived as more particular, distant, or different are undoubtedly the Castilian or peninsular Spanish, and the Argentinian and Uruguayan (from Rio de la Plata or *rioplatense*) (Moreno Fernández 2015). We are also disregarding any dialect fragmentation that exists within countries or national territories. According to our current data, the distinctive character perceived in Argentina has its empirical support in the extreme position occupied by Argentina in the vertical axis of the principal components graph.

On the other hand, Spain's linguistic separation must be qualified in light of the results provided by the indices of generality and particularity; that is, considering the degree of "communality" of the linguistic uses of each of the Hispanic countries. Our results show that Spain has one of the highest indices of generality. This means that a large part of the linguistic uses observed in Spain are also used in other areas of the Spanish-speaking community. To understand this apparent discrepancy, recall that the uses registered in *Varilex-R* do not include unique answers for the (nearly one thousand) concepts, expressions, and referents considered, but that many of these concepts often receive two, three, or more responses from the same speaker or from different speakers within the same country. This suggests that generality and particularity are not measured by the unique and active features within a community, but by recognizing a fluidity and multiplicity of expressive options.

Indeed, the previous discussion of linguistic distances and degrees of generality and particularity may seem paradoxical. How is it possible that the most distant country has a high index of "generality" or, equivalently, a low index of "particularity"? In the case of Spain, the answer must be methodological in nature. The index of "generality" is calculated from the coincidences with all of the other countries ("commonality"); that is, the information refers to each country. The principal components and association calculations, on the other hand, are made between all possible *pairs* of countries to find an overall configuration, thus the methodological solution to the apparent contradiction. The separation of Spain from the other countries is due to the significant number of particular uses associated with its distant geography; the generality comes from the fact that its linguistic uses also appear in many other countries as first, second, or third options. It is true that forms such as *chupa* (jacket), *jersey* (jumper or jersey), *papel de plata* (aluminum foil) or *vespino* (small motorcycle) can be treated as particularities of peninsular Spanish, i.e., as *españolismos*. But it is no less true that many uses common in Spain can serve as alternatives to other primary uses in many other countries, hence its generality. Such a coincidence could be attributable to historical contact as well as to the modern penetration of peninsular Spanish in the Spanish-speaking community through the media, among other possibilities.

Colombia's high particularity index, the highest among the Spanish-speaking countries, is an unexpected finding that merits more detailed study, although it could be indicative of the inversely proportional relationship between exclusive uses and non-shared uses. Meanwhile, Cuba's position as the country with the highest generality index could be due to its central location within the Spanish-speaking geography and to the frequent contacts that the Greater Antilles have maintained with the other Spanish-speaking territories throughout history—with the exception of last half century, which has also left its linguistic imprint.

Finally, it is worth briefly mentioning the implications and applications of this type of analysis for understanding the internal dynamics of the major international languages. From our point of view, the internal configuration of each linguistic space has proved essential for the future development of languages and their possibilities for growth as nodal languages (Moreno Fernández 2016).

12 Conclusions

The combination of statistical methods performed here has proved to be an effective tool for conducting an in-depth examination of a dialect continuum as dynamic and multidimensional as the Spanish-speaking community. These techniques will perform best when they are applied to a large, carefully collected volume of data that is adequately representative of Spanish language geography. In this sense, the *Varilex-R* database, completed and updated in 2016, provides an ample amount of information, representing 981 concepts, expressions, or references in 21 Spanish-speaking countries.

Correlation, cluster, principal components, and association analyses reveal a cohesive configuration within the Spanish-speaking world, with the majority of countries coming together in the same space of variation and with a balance between particular and shared forms that justifies both an existing sense of community and an awareness of a shared identity. In that dialectal concert, the most discordant notes are the spaces of Spain and Chile on one hand, and Argentina, Costa Rica, and Colombia on the other. In addition, the generality and particularity indices unveil relationships between varieties where the linguistic uses are not assessed independently, but rather interrelated, which leads us to appreciate that the particularity exhibited by Spain, for example, does not correspond to its notably high index of generality. Spain's high index of generality reveals agreement with the rest of the Spanish-speaking community that stems from the fact that the linguistic uses for the same concepts or referents are not typically unique or univocal, but diverse and multivocal, and it is in this diversity of solutions where confluences appear. Nevertheless, Chile's unique personality is reinforced by its high index of particularity.

Although the set of analyses performed here provides solid conclusions, it is far from exhaustive. Future work will involve the necessarily continuous task of refining and completing the database, and the detailed study of the linguistic features associated with each country in relation to the rest of the Spanish-speaking world. In addition, new studies will be needed for the analysis of other types of regional geographic entities, both intranational and transnational.

References

Alba, Orlando. 1992. Zonificación dialectal del español de América (Dialectal zoning of American Spanish. In C. Hernández (coord.), *Historia y presente del español de América (History and present of American Spanish).* Valladolid: Gobierno de Castilla y León, p. 63-84.

Aliaga Jiménez, José Luis. 2003. Dialectometría y léxico en las hablas de Teruel (Dialectometry and lexicon in Teruel speeches). *Estudios de Lingüística. Universidad de Alicante*, 17: 25-55.

Alvar, Manuel. 2002. ¿Fragmentación del español? (Fragmentation of Spanish language?). In M.T Echenique and J. Pedro Sánchez Méndez (coord.), *Actas del V Congreso Internacional de Historia de la Lengua Española, Valencia January 31 - February 4, 2000.* Madrid: Gredos, pp. 221-236.

Armas y Céspedes, Juan I. 1882. *Oríjenes del lenguaje criollo (Origins of Creole language)*. La Habana: Viuda de Soler.

Auer, Peter, Hiskens, Frans and Paul Kerswill (eds.). 2004. *Dialect change: Convergence and Divergence in European Languages*. Cambridge: Cambridge University Press.

Aurrekoetxea, Gotzon and Ormaetxea, Jose Luis (eds.) (2010): *Tools for linguistic variation.* Bilbao: Universidad del País Vasco.

Ávila, Raúl. 2003. La pronunciación del español: medios de difusión masiva y norma culta (Spanish pronunciation: social media and literate norm). *Nueva Revista de Filología Hispánica,* LI: 57-79.

Borin, Lars and Saxena, Anju (eds.) (2013): *Approaches to measuring linguistic differences*. Berlin: De Gruyter.

Cahuzac, Philippe. 1980. La división del español de América en zonas dialectales. Solución etnolingüística y semántico-dialectal (The division of Spanish in dialectal areas. An ethnolinguistic and semantic solution. *Lingüística Española Actual,* II: 385-461.

Cichocki, Wladyslaw. 1988. Uses of Dual Scaling in Social Dialectology: Multidimensional Analysis of Vowel Variation. In A. R. Thomas (ed.), p. 187-199.

Clua, Esteve. 2010. Relevancia del análisis lingüístico en el tratamiento cuantitativo de la variación dialectal (Relevance of linguistic analysis in the quantitative processing of dialect variation). In Aurrekoetxea / Ormaetxea, p. 151-166.

Clyne, Michael (ed.) (1992): Pluricentric Languages. *Differing norms in different nations*. Berlin: De Gruyter.

Enguita, José María (ed.) (1991): *I Curso de Geografía Lingüística de Aragón (I Course of Aragonese Linguistic Geography)*. Zaragoza: Institución Fernando el Católico.

García Mouton, Pilar. 1991. Dialectometría y léxico en Huesca (Dialectometry and lexicon in Huesca). In Enguita, p. 311-326.

Goebl, Hans. 1981. Eléments d'analyse dialectométrique (avec application à l'AIS) (Elements of dialectometric analysis (applied to AIS)). *Revue de Linguistique Romane*. 45: 349-420.

Goebl, Hans. 1982. Ansätze zu einer computativen Dialektometrie (Approaches to computative dialectometry). In W. Besch et al. (eds.), *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforshung, 1 (Dialectology. A handbook on German and general dialect research)*. Berlin: Walter de Gruyter, p. 778-792.

Goebl, Hans .2010. Introducción a los problemas y métodos según los principios de la Escuela Dialectométirca de Salzburgo (con ejemplos sacados del «Atlante Italo Svizzero», AIS) (Introduction to the problems and methods according to the principles of the Dialectometric School of Salzburg (with examples taken from Linguistic Atlas of Switzerland, AIS). In Aurrekoetxea / Ormaetxea, p. 3-39.

Gonçalves, Bruno and David Sánchez. 2014. Crowdsourcing Dialect Characterization through Twitter. ArXiv: 1407.7094. http://arxiv.org/abs/1407.7094 [Accessed: February 12, 2015]

Guiter, Henri. 1973. Atlas et frontières linguistiques (Atlas and linguistic borders). In G. Straka and P. Gardette (eds.), *Les dialectes de France à la lumière des atlas régionaux (Colloque de Strasbourg, 1971) (French dialects in the light of regional atlases (Strasbourg Colloquium)*. Paris: CNRS, p. 61-109.

Hahsler, Michael; Grun, Bettina; Hornik, Kurt; Buchta, Christian. 2016. Introduction to arules – A computational environment for mining association rules and frequent item sets. In: http://epub.wu.ac.at/132/ [Accessed: October 23, 2018]

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

Henríquez Ureña, Pedro. 1930. Observaciones sobre el español de América. II (Observations on American Spanish). *Revista de Filología Española*, XVII: 277-284.

Henríquez Ureña, Pedro. 1931. Observaciones sobre el español de América. III (Observations on American Spanish). *Revista de Filología Española*, XVIII: 120-148.

Houck, Charles. 1967. A Computerized Statistical Methodology for Linguistic Geography: A Pilot Study. *Folia Linguistica*, I, 1/2: 80-95.

Lapesa, Rafael. 1980. América y la unidad de la lengua española (America and the unity of the Spanish language). *Documentos Lingüísticos y Literarios* (*Linguistic and Literary Documents*), 5: 74-89

Lieberman, Erez; Michel, Jean-Baptiste; Jackson, Joe; Tang, T.; Nowak Martin A. 2007. Quantifying the evolutionary dynamics of language. *Nature,* 449: 713-716.

Lipski, John. 1994. *Latin American Spanish*. London: Longman.

López Morales, Humberto. 2006. *La globalización del léxico hispánico (The globalization of Hispanic lexicon)*. Madrid: Espasa.

Moore, Carmella C. 1994. Material culture, geographic propinquity, and linguistic affiliation on the North Coast of New Guinea: A reanalysis of Welsch, Terrell, and Nadolski (1992)." *American Anthropologist,* 96: 370-296.

Moreno de Alba, José G. 1978. *Unidad y variedad del español en América (Unity and variety of Spanish in the Americas)*. Mexico: UNAM.

Moreno Fernández, Francisco. 1991. Morfología en el ALEANR: Aproximación dialectométrica (Morphology in ALEANR. Dialectometric Approach). In Enguita 289-309.

Moreno Fernández, Francisco. 1993a. *La división dialectal del español de América (Dialect division of American Spanish)*. Alcalá de Henares: Universidad de Alcalá.

Moreno Fernández, Francisco. 1993b. Geolingüística y cuantificación (Geolinguistics and quantification). *Actas del III Congreso Internacional de Hispanistas de Asia (Proceedings of the III International Congress of Hispanists from Asia)*. Tokyo: University of Tokyo, p. 289-300.

Moreno Fernández, Francisco. 1999-2000. El estudio de la convergencia y la divergencia dialectal (The study of dialectal convergence and divergence). *Revista Portuguesa de Filologia*, XXIII: 1-27.

Moreno Fernández, Francisco. 2000. *Qué español enseñar (What Spanish to teach)*. Madrid: Arco/Libros.

Moreno Fernández, Francisco. 2010. *Las variedades del español y su enseñanza (Spanish varieties and their teaching)*. Madrid: Arco/Libros.

Moreno Fernández, Francisco. 2014a. *La lengua española en su geografía. Manual de dialectología hispánica (Spanish language in its geography. Handbook of Spanish dialectology)*. 2nd. ed. Madrid: Arco/Libros.

Moreno Fernández, Francisco. 2014b. Los súperdialectos del español global (The super-dialects of the global Spanish). *Medium*. https://medium.com/@FMORENOFDEZ/los-superdialectos-del-espanol-global-5aa83b2e9d85 [Accessed: October 23, 2018]

Moreno Fernández, Francisco. 2015. La percepción global de la similitud entre variedades de la lengua española (The global perception of similarity between Spanish varieties). In Kirsten Jeppesen Kragh and Jan Lindschouw (Eds.), *Les variations diasystématiques et leurs interdépendances dans les langues romanes (Diasistematic variations and their interdependencies in Romance languages)*. Strasbourg, Éditions de linguistique et de philologie, p. 217-238.

*Moreno Fernández, Francisco. 2016. La búsqueda de un español global (The search for a global Spanish). In VII Congreso Internacional de la Lengua Español*a (*VII International Congress of Spanish Language)*. San Juan, Puerto Rico: Cervantes Institute – RAE / ASALE. http://congresosdelalengua.es/puertorico/ponencias/seccion_5/ponencias_seccion5/moreno-francisco.htm [Accessed: October 23, 2018]

Moreno Fernández, Francisco. In press. Los "ismos" nacionales de la lengua española (Spanish language national «isms»). Boletín de la Real Academia Española.

Moreno Fernández, Francisco y Otero, Jaime. 2016 *Atlas de la lengua española en el mundo (Atlas of the Spanish Language in the world)*. 3rd ed. Barcelona: Ariel.

Moreno Sandoval, Antonio and Guirao Miras, José María. 2008. Frecuencia y distintividad en el uso lingüístico: casos tomados de la lematización verbal de corpus de distintos registros (Frequency and distinctiveness in linguistic use: cases taken from the verbal lemmatization of corpus from different registers). *Actas del I Congreso Internacional de Lingüística de Corpus (Proceedings of the I International Congress of Corpus Linguistics) (CILC-09)*, Murcia: Universidad de Murcia. 195-210.

Pennycook, Alastair (ed.). 1995. *English in the World / The world in English*. Cambridge: Cambridge University Press.

Pennycook, Alastair. 2006. *Global Englishes and Transcultural Flows.* London: Routledge.

Rabanales, Ambrosio. 1998. Unidad y diversificación de la lengua española (Unity and diversification of the Spanish language). *Onomazein,* 3: 133-142.

RAE-ASALE. 2005. *Diccionario panhispánico de dudas (Pan-Hispanic Dictionary of Doubts)*. Madrid: Santillana.

RAE-ASALE. 2010. *Nueva Gramática de la Lengua Española (New Grammar of Spanish Language)*. Madrid: Espasa.

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

Rona, José P. 1964. El problema de la división del español americano en zonas dialectales (On the division of the American Spanish in dialectal areas). *Presente y futuro de la lengua española (Present and future of the Spanish Language)*. I. Madrid: OFINES, pp. 215-226.

Rosenblat, Ángel. 1962. *El castellano de España y el castellano de América (Spanish from Spain and Spanish from America)*. Caracas: Universidad Central de Venezuela.

Ruiz Tinoco, Antonio. 1999. El Proyecto VARILEX en Internet. Base de datos compartida de variación léxica. *VARILEX* 7: 1-10.

Ruiz Tinoco, Antonio. 2002. Cartografía automática en Internet (Automated cartography on the Internet). VARILEX 10, p. 6-17.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale (The relationship between spatial distance and lexical distance). Revue de Linguistique Romane, 35: 335-357.

Soares da Silva, Augusto. 2006. Sociolinguística cognitiva e o estudo da convergência/divergência entre o português europeu e o português brasileiro (Cognitive Sociolinguistics and the study of convergence/divergence between European Portuguese and Brazilian Portuguese). *Revista Veredas*, 10, 1.2. In: https://veredas.ufjf.emnuvens.com.br/veredas/article/view/373/321 [Accessed: October 23, 2018]

Thomas, Alan R. (ed.) 1988. *Methods in Dialectology*. Clevedon: Multilingual Matters.

Thompson, R.W. 1992. Spanish as a pluricentric language. In M. Clyne (ed.), p. 45-70.

Ueda, Hiroto. 1995. Zonificación del español del mundo. Palabras y cosas de la vida urbana (Zoning of the worldwide Spanish. Words and things of urban life). *Lingüística*, 7: 43-86.

Ueda, Hiroto. 2007. Zonificación múltiple de las ciudades hispanohablantes según el léxico urbano moderno. Análisis clúster y análisis de componentes principales (Multiple zoning of Spanish-speaking cities according to modern urban lexicon. Cluster analysis and analysis of main components). In A. Ruiz Tinoco (ed.), *Jornadas sobre métodos informáticos en el tratamiento de las lenguas ibéricas (Conference on computational methods in the treatment of Iberian languages)*. Tokyo: Center of Hispanic Studies - Sophia University, p. 121-140.

Ueda, Hiroto. 2008. Análisis dialectométrico del léxico variable español: Interpretación taxonómica de resultados (Dialectometric analysis of the Spanish variable lexicon: Taxonomic interpretation of results). In *El español de América, Actas del VI Congreso Internacional de El español de América (American Spanish, Proceedings of the VI International Conference on American Spanish)*. Valladolid: Universidad de Valladolid, p. 813-822.

Ueda, Hiroto. 2015. *Métodos de análisis de datos cuantitativos en estudios lingüísticos (Methods for quantitative analyses in language studies)*. In: http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/4-numeros/numeros-es.pdf [Accessed: October 23, 2018]

Ueda, Hiroto and Moreno Fernández, Francisco. 2016. *VARILEX-R: Variación léxica del español en el mundo (VARILEX-R. Lexical variation of the Spanish language in the world).* In http://goo.gl/BENLPL.

Ueda, Hiroto and Ruiz Tinoco, Antonio. 2003. VARILEX. Variación léxica del español en el mundo. Proyecto internacional de investigación léxica (VARILEX. Lexical variation of the Spanish language in the world. International research project on lexicon). In Raúl Ávila *et al*. P*autas y pistas en el análisis del léxico hispano(americano) (Guidelines and clues in the analysis of the Hispanic (American) lexicon)*. Frankfurt: Iberoamericana, p. 141-278.

Ueda, Hiroto and Ruiz Tinoco, Antonio. 2007. Investigaciones sobre la variación léxica del español: Proyectos y resultados de 1992 a 2007 (Research on the lexical variation of Spanish: Projects and results from 1992 to 2007). VARILEX 15: 1-19.

Varilex (2015): http://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/ [Accessed: October 23, 2018].

Viereck, Wolfgang. 1988. The Computerisation and Quantification of Linguistic Data: Dialectometrical Methods. In A. R. Thomas (ed.), p. 524-550.

Woods, Anthony, Fletcher, Paul and Hugues, Arthur. 1986. *Statistics in language studies,* Cambridge, Cambridge University Press.

Zamora Munné, Juan C. 1979-1980. Las zonas dialectales del español americano (The dialectal zones of American Spanish). *Boletín de la Academia Norteamericana de la Lengua Española*, 4-5: 57-67.

APPENDIX

Analysis of generality and peculiarity

Based on the data organized in the matrix below, it is possible to calculate the degrees of "generality" and "particularity" to see which variables (v1-v5) possess them with respect to others.

G	v1	v2	v3	v4	v5		H	Suma
d1	1	1	1	0	0		d1	3
d2	1	1	0	1	1		d2	4
d3	0	1	0	1	0		d3	2
d4	1	0	1	1	1		d4	4

In order to measure the degree of "generality" of the variables, the vertical sum is not useful, although it is related to the strength of each variable:

V	v1	v2	v3	v4	v5
Suma	3	3	2	3	2

The vertical sum does not prove very helpful as it does not take into account the commonality that each positive value (1) has in the corresponding rows. Thus, an alternative must be sought. First, the values that each point has in relation to other points in the same row can be considered. For example, the value 1 of [v1: d1] is different from that of [v1: d2], since the first is 1 next to 2, while the second is 1 next to 3, so it is considered that the 1 of the second case has more generality than the first; that is, it follows the mode of the distribution. Therefore, the horizontal sum is used to represent the "generality" of the positive value (1) as follows:

H = sumH (G) - G

H	v1	v2	v3	v4	v5
d1	2	2	2	3	3
d2	3	3	4	3	3
d3	2	1	2	1	2
d4	3	4	3	3	3

The positive points (1) have now become values of the horizontal sum (3, 4, 2, 4) minus matrix G. Thus, the same values have incorporated values that represent the degree of commonality in each row. For example, the 1 of [v1, d1] has the communal value of 2, while the 1 of [v1, d2] has the communal value of 3. The greater the value of commonality, the greater the degree of "generality". According to common sense, the citizen who votes for the winning party of an election, along with the majority of the voters, represents the greatest "generality" in the voting.

G.a.val.	v1	v2	v3	v4	v5		H2	Sum
Sum	8	6	5	7	6		Sum	32

Next, it is possible to add the values vertically: (8, 6, 5, 7, 6). This horizontal vector represents, in some way, the degree of "generality:" "Generality by absolute value" (G.a.val.). It is important to calculate its relative value to evaluate the same degree within a scale from 0 to 1. Therefore, the horizontal vector (8, 6, 5, 7, 6) is divided by the sum of the same vector (32):

r.freq.G. = G.a.val. / Sum (G.a.val.)

r.freq.G.	v1	v2	v3	v4	v5
M*H/H2	.250	.188	.156	.219	.188

The degree of generality by relative value (G.r.val.) is obtained in this way. To measure the degree of particularity (P.r.val.), from generality (G.r.val.), a calculation through subtraction can be made:
P.r.val. = 1 - G.r.val.

r.freq.P.	v1	v2	v3	v4	v5
1-M*H/H2	.750	.813	.844	.781	.813

The degree of "generality" is always considerably less than that of "particularity," not because of the very nature of the data, but rather simply because of the number of variables. Among the five variables (v1-v5), the vertical sum corresponding to each column is divided. In this way, among twenty variables, it is further reduced without nearly reaching the first digit (0.1). And consequently, the value of the degree of particularity rises in a significant way.

To find the formula that is not influenced by the number of variables, an operation named "prominent relative frequency" (f.r.p.) can be proposed. It consists in increasing the value of the absolute frequency of the variable in question, multiplied by the number of variables in comparison.

a.freq.G.	v1	v2	v3	v4	v5	H2	Sum
Sum	8	6	5	7	6	Sum	32

For example, 8 divided by 32, is 8/32 = .250, which is the relative frequency. Now, it is considered that the value 8 is difficult to evaluate within the totality as great as five members; that is, it is not directly comparable with the remaining 4, since it is one value against 4 values. Hence the resulting value which is always considerably reduced in relative frequency: (.250, .188, .156, .219, .188). To correct this reduction, and to match the comparison condition, the value in question must be multiplied by the total number of remaining ones: 8 * (5 - 1) = 32, which is now comparable with the remaining 4 (6 + 5 + 7 + 6 = 24). To obtain the "prominent relative frequency" the value of 32 between the two figures is calculated (32 + 24 = 56): that is, 32 / (32 + 24) = 32/56 = .571. The formula for the prominent relative frequency (f.r.p.) is:

f.r.p = [f.a * (n-1)] / [f.a * (n-1) + (s.h. - f.a)]
= [8 * (5 - 1)] / [8 * (5 - 1) + (32 - 8)] = .571

where, f.a. is absolute frequency, n is number of variables, s.h. is horizontal sum.

And its corresponding degree of peculiarity by prominent relative frequency (p.f.r.p .: p.r.freq.P.) is:
P.p.r.val. = 1 - G.p.r.val.

p.r.freq.P.	v1	v2	v3	v4	v5
Sum	.429	.520	.574	.472	.520

All these calculations can be summarized in the following way:

Generality	v1	v2	v3	v4	v5
Sum	3	3	2	3	2
G.a.val.	8	6	5	7	6
G.r.val.	.250	.188	.156	.219	.188
P.r.val.	.750	.813	.844	.781	.813
G.p.r.val.	.571	.480	.426	.528	.480
P.p.r.val.	.429	.520	.574	.472	.520