Dialectologia. Special issue, IV *(2013), 221-235.*
ISSN: 2013-2247

Received 1 November 2013.
Accepted 15 November 2013.

UNA NOTA SOBRE EL MÉTODO DE TAXONOMÍA CUANTITATIVA DE GRANDES DATOS. COEFICIENTES DE ASOCIACIÓN APLICADOS A LAS VARIANTES DEL *DICCIONARIO DE AMERICANISMOS*[1]

Hiroto UEDA
Universidad de Tokio
uedahiroto@jcom.home.ne.jp

Resumen

Tras clasificar los coeficientes de asociación utilizados en distintos estudios o nuevamente ideados, aplicamos el coeficiente Jaccard, que nos parece más sencillo y conveniente, a los materiales elaborados a partir del *Diccionario de americanismos* (Asociación de Academias de la Lengua Española: 2010). Al encontrar una dificultad, proponemos utilizar otro índice diferente para resolverla. El índice utilizado, denominado «Coeficiente de asociacion mayor», no entra en la clasificación de los coeficientes previamente realizada.

Palabras clave

coeficiente de asociación, taxonomía diatópica, variación léxica, español en Hispanoamérica, *Diccionario de americanismos*

A NOTE ON QUANTITATIVE METHODS OF TAXONOMY WITH LARGE DATA. COEFFICIENT OF ASSOCIATION APPLIED TO THE VARIANTS OF THE *DICCIONARIO DE AMERICANISMOS*

Abstract

After classifying the association coefficients used in different studies or newly designed, we apply the Jaccard coefficient, which seems to be simpler and most convenient, to the materials obtained from

[1] Agradezco de todo corazón a Humberto López Morales y Maria-Pilar Perea por las ayudas prestadas para realizar este trabajo. Hemos contado con la subvención ofrecida por Japan Society for Promotion of Science, 24520453. Este trabajo se adscribe en el proyecto de investigación "Portal de léxicos y gramáticas dialectales del catalán del siglo XIX" (FFI2010-18940 (subprograma FILO)), financiado por el Ministerio de Ciencia e Innovación.

the *Diccionario de Americanismos* (Asociación de Academias de la Lengua Española: 2010). After finding a problem, we propose to use a different index to solve it. The index used, called "greater association coefficient", is not classifiable between the coefficients previously defined.

Keywords
association coefficient, diatopic taxonomy, lexical variation, American Spanish, *Diccionario de americanismos*

1. Introducción

Para la taxonomía numérica de datos cualitativos, se han propuesto distintos coeficientes de asociación, de los más sencillos a los más sofisticados, seleccionados según el propósito de investigación y/o el carácter propio de los datos objeto de estudio. Algunos son tan utilizados que llevan nombres propios y otros no son tan conocidos como para ser utilizados con consenso general. En esta nota, en la sección 1, tratamos de enumerarlos clasificados según la constitución de las fórmulas anteriores, junto con el resultado de un pequeño experimento con parámetros controlados. Seguidamente, en la sección 2, los aplicamos a los datos del léxico variable registrado en el *Diccionario de americanismos* (Asociación de Academias de la Lengua Española: 2010) e informamos un problema general que se presenta en las fórmulas matemáticas a la hora de aplicarlas a datos de esta envergadura.

2. Coeficientes de asociación

Los datos cualitativos, a diferencia de los cuantitativos, suelen ser representados con un signo único, una letra o valor uno, en cada celda de la matriz, constituida de una columna de nombres individuales y una fila de parámetros variables. A partir de estos datos se calculan cuatro sumas (frecuencias) de cada valor y sus coocurrencias entre dos variables en cuestión: *a*, que son las veces de ocurrencias de reacciones positivas tanto del primer parámetro como del segundo; *b*, que son las veces de reacciones positiva en el primer parámetro y negativa en el segundo; *c*, que son, en

contraste, las veces de reacciones negativas en el primer parámetro y positivas en el segundo, y finalmente, *d*, que son las veces de reacciones negativas tanto en el primer parámetro como en el segundo.

A modo de ejemplo, el cuadro siguiente muestra las frecuencias de palabras registradas en dos corpus, seguidas de sus representaciones cualitativas y los valores de *a*(+/+), *b*(+/-), *c*(-/+) y *d*(-/-), con la frecuencia mayor que uno. Los valores de *a*, *b*, *c*, *d* son totales de cada columna (Hoz, 1953; Chang-Rodríguez, 1964).

Palabra	Hoz. Carta	Ch. &R. Drama		Hoz. Carta	Ch. &R. Drama		a++	b+-	c-+	d--
abajo	5	10		1	1		1	0	0	0
abandonar	9	6		1	1		1	0	0	0
abandono	0	0		0	0		0	0	0	1
abarcar	1	0		0	0		0	0	0	1
abastecimiento	2	0		1	0		0	1	0	0
abatir	0	1		0	0		0	0	0	1
abeja	2	3		1	1		1	0	0	0
abertura	0	0		0	0		0	0	0	1
abismo	0	0		0	0		0	0	0	1
abnegación	0	0		0	0		0	0	0	1
abogado	3	6		1	1		1	0	0	0
abonar	5	0		1	0		0	1	0	0
abono	0	0		0	0		0	0	0	1
abordar	0	0		0	0		0	0	0	1
aborrecer	0	6		0	1		0	0	1	0
					Total:		4	2	1	8

Tabla 1. Total de frecuencias de palabras registradas en dos corpus (cf. Hoz, 1953 y Chang-Rodríguez, 1964)

De estos cuatro valores se calcula, por ejemplo, el valor de correspondencia simple ('simple matching value'): $(a + d) / (a + b + c + d)$, que es 0,8. Este valor es sumamente alto, puesto que entre los parámetros (Hoz. Carta y Ch. & R. Drama) hay 4 coincidencias positiva (*a*) y 8 negativas (*d*), frente a las pocas veces de no correspondencia (b=2, c=1): $(4 + 8) / (4 + 2 + 1 + 8) = 0,8$. El máximo valor (1,0) se obtiene cuando $b = c = 0$; y el mínimo (0,0) cuando $a + d = 0$.

Desde esta fórmula tan sencilla hasta las correlaciones, se enumeran las fórmulas siguientes, clasificadas por su estructura matemática y sus componentes.[2]

[2] Anotamos los nombres detrás de las fórmulas correspondientes (Anderberg 1973: 89). Las que no llevan nombre son las que creemos posibles de utilizar con algún fundamento matemático.

[1] Coeficientes de asociación aditiva

[1a] $(a+d) / (a+b+c+d)$: Correspondencia simple

[1b] $a / (a+b+c)$: Jaccard

[1c] $2a / (2a+b+c)$: Dice-Sorensen

[1d] $a / (a+b+c+d)$: Russell y Rao

[1e] $3a / (3a+b+c+d)$

[1f] $(a+d-b-c) / (a+b+c+d)$: Hamann

[1g] $(a-b-c) / (a+b+c)$

[1h] $(2a-b-c) / (2a+b+c)$

[2] Coeficientes de asociación multiplicativa[3]

[2.1a] $ad / (ad + bc)$

[2.1b] $\sqrt{[ad / (ad + bc)]}$

[2.1c] $\sqrt{(ad)} / [\sqrt{(ad)} + \sqrt{(bc)}]$

[2.1d] $a^2 / (a^2 + bc)$

[2.1e] $a / \sqrt{(a^2 + bc)}$

[2.1f] $a / [a + \sqrt{(bc)}]$

[2.2a] $(ad - bc) / (ad + bc)$: Yule

[2.2b] $Sgn(ad - bc)\ \sqrt{[|ad - bc| / (ad + bc)]}$

[2.2c] $[\sqrt{(ad)} - \sqrt{(bc)}] / [\sqrt{(ad)} + \sqrt{(bc)}]$

[2.2d] $(a^2 - bc) / (a^2 + bc)$

[2.2e] $Sgn(a^2 - bc)\ \sqrt{[|a^2 - bc| / (a^2 + bc)]}$

[2.2f] $[a - \sqrt{(bc)}] / [a + \sqrt{(bc)}]$

[3] Coeficientes de asociación correlacional

[3a] $(ad - bc) / \sqrt{[(a+b)(c+d)(a+c)(b+d)]}$: Phi

[3b] $Sgn(ad - bc) * \sqrt{\{|ad - bc| / \sqrt{[(a+b)(c+d)(a+c)(b+d)]}\}}$

[3] La función «Sgn» devuelve más uno (+1) o menos uno (-1), según el valor que se encuentra entre paréntesis.

[3c] $[\sqrt{(ad)} - \sqrt{(bc)}] / \sqrt{\sqrt{[(a+b)(c+d)(a+c)(b+d)]}}$

[3d] $a / \sqrt{[(a+b)(a+c)]}$: Ochiai

Todas estas fórmulas son posibles y, naturalmente, devuelven resultados diferentes. Hemos preparado una hoja de Excel para hacer un pequeño experimento. Se trata de ver sus valores resultantes según el cambio del valor de *a* (de 0 a 10), con los valores fijos de *b* (=4), *c* (=6), *d* (=8); de modo que tenemos 11 columnas siguientes:

a	(+/+)	0	1	2	3	4	5	6	7	8	9	10
b	(+/−)	4	4	4	4	4	4	4	4	4	4	4
c	(−/+)	6	6	6	6	6	6	6	6	6	6	6
d	(−/−)	8	8	8	8	8	8	8	8	8	8	8

[1] Coeficientes de asociación aditiva

Los resultados del cálculo de [1] «Coeficientes de asociación aditiva» son siguientes:

Additive association	0	1	2	3	4	5	6	7	8	9	10
(a+d) / (a+b+c+d)	.444	.474	.500	.524	.545	.565	.583	.600	.615	.630	.643
a / (a+b+c)	.000	.091	.167	.231	.286	.333	.375	.412	.444	.474	.500
2a / (2a+b+c)	.000	.167	.286	.375	.444	.500	.545	.583	.615	.643	.667
a / (a+b+c+d)	.000	.053	.100	.143	.182	.217	.250	.280	.308	.333	.357
3a / (3a+b+c+d)	.000	.143	.250	.333	.400	.455	.500	.538	.571	.600	.625
(a+d-b-c) / (a+b+c+d)	-.111	-.053	.000	.048	.091	.130	.167	.200	.231	.259	.286
(a-b-c) / (a+b+c)	-1.000	-.818	-.667	-.538	-.429	-.333	-.250	-.176	-.111	-.053	.000
(2a-b-c) / (2a+b+c)	-1.000	-.667	-.429	-.250	-.111	.000	.091	.167	.231	.286	.333

Tabla 2. Resultado del cálculo del «Coeficiente de asociación aditiva»

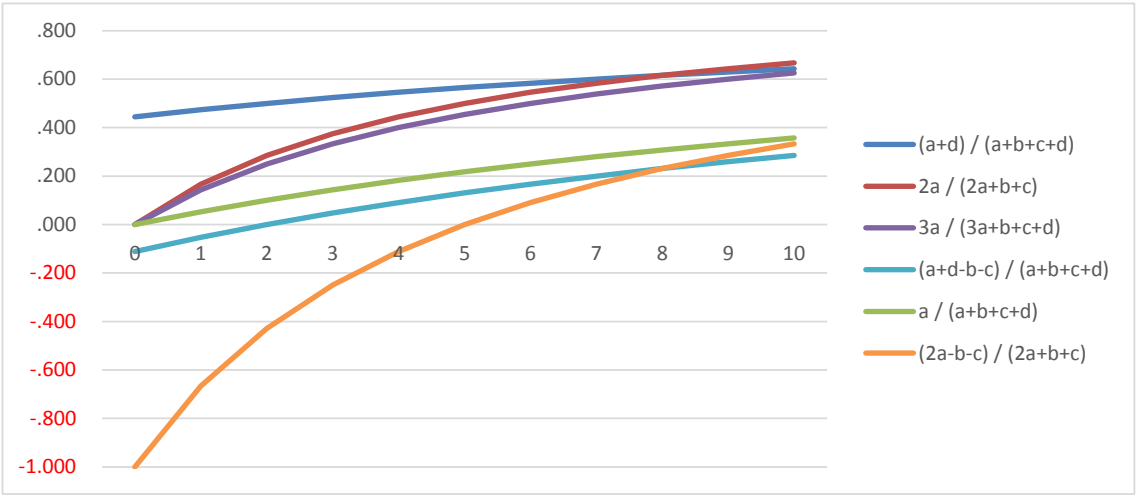

Gráfico 1. Resultado del cálculo del «Coeficiente de asociación aditiva»

En este grupo se encuentran coeficientes cuyos componentes (*a*, *b*, *c*, *d*) se combinan por adición y sustracción. Se observa que la «Correspondencia simple», (*a* + *d*) / (*a*+*b*+*c*+*d*), oscila con poca variación (0.444 - 0.643), por poseer un componente *d* tanto en el numerador como en el denominador, en contraste con los restantes [1b] - [1h]. Dice-Sorensen, 2*a* / (2*a*+*b*+*c*), y su correspondiente de tipo contrastivo, (2*a*-*b*-*c*) / (2*a*+*b*+*c*), poseen un rango de oscilación superior, por la intensificación que se hace con el valor 2*a*. Según nuestro sentido común, el valor 0.5 en el rango de [0.0 - 1.0] o 0.0 en el de [-1.0 - 1.0] son los más adecuados para el caso de [*a*=10, *b*=4, *c*=6], puesto que, en este caso, la frecuencia de *a* iguala a la suma de *b* (+/-) y *c* (-/+). Son los casos de los coeficientes [1b] *a* / (*a*+*b*+*c*): Jaccard y [1f] (*a*+*d*-*b*-*c*) / (*a*+*b*+*c*+*d*): Hamann.

Consideramos, no obstante, que Dice-Sorensen, 2*a* / (2*a*+*b*+*c*), lleva razón al duplicar el valor *a*, por ser un miembro de ecuación contra los dos, *b* (+/-) y *c* (-/+). En este sentido, el caso de [*a*=10, *b*=4, *c*=6] debería presentar un valor de 0.667 en la escala de [0.0 - 1.0]. La misma argumentación se aplica a nuestra fórmula propuesta, (2*a*-*b*-*c*) / (2*a*+*b*+*c*), que presenta el valor 0.333 en la escala de [-1.0 - 1.0].

[2.1] Coeficientes de asociación multiplicativa (1)

Multiplicative association (1)	0	1	2	3	4	5	6	7	8	9	10
ad / (ad + bc)	.000	.250	.400	.500	.571	.625	.667	.700	.727	.750	.769
√[ad / (ad + bc)]	.000	.500	.632	.707	.756	.791	.816	.837	.853	.866	.877
√(ad) / [√(ad) + √(bc)]	.000	.366	.449	.500	.536	.564	.586	.604	.620	.634	.646
a^2 / (a^2 + bc)	.000	.040	.143	.273	.400	.510	.600	.671	.727	.771	.806
a / √(a^2 + bc)	.000	.200	.378	.522	.632	.714	.775	.819	.853	.878	.898
a / [a + √(bc)]	.000	.170	.290	.380	.449	.505	.551	.588	.620	.648	.671

Tabla 3. Resultado del cálculo del «Coeficiente de asociación multiplicativa» (1)

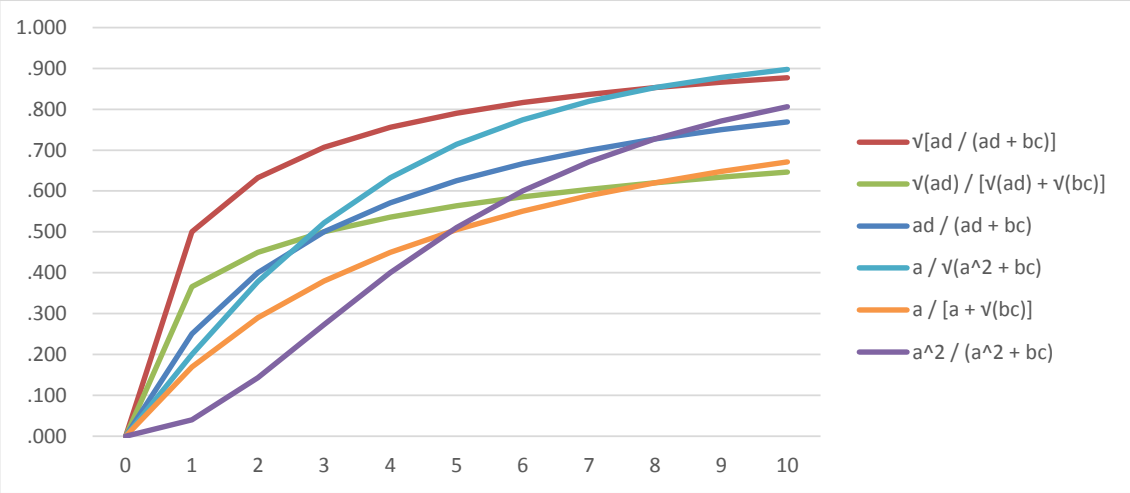

Gráfico 2. Resultado del cálculo del «Coeficiente de asociación multiplicativa» (1)

El grupo [2.1] se caracteriza por la multiplicación en los componentes *ad* y *bc*. Por no realizar la sustracción en el numerador, se oscila entre 0.0, cuando *a*=0, *ad*=0, y 1.0, cuando *bc*=0. La curva se presenta pronunciada por la multiplicación. Todos estos coeficientes presentan valores superiores a 0.5 en el caso de [*a*=10, *b*=4, *c*=6], donde supone el equilibrio entre el valor *a* y *b* + *c*.

[2.2] Coeficientes de asociación multiplicativa (2)

Multiplicative association (2)	0	1	2	3	4	5	6	7	8	9	10
(ad - bc) / (ad + bc)	-1.000	-.500	-.200	.000	.143	.250	.333	.400	.455	.500	.538
Sign(ad - bc)* √[|ad - bc| / (ad + bc)]	-1.000	-.707	-.447	.000	.378	.500	.577	.632	.674	.707	.734
[√(ad) - √(bc)] / [√(ad) + √(bc)]	-1.000	-.268	-.101	.000	.072	.127	.172	.209	.240	.268	.292
(a^2 - bc) / (a^2 + bc)	-1.000	-.920	-.714	-.455	-.200	.020	.200	.342	.455	.543	.613
Sign(a^2 - bc)*√[|a^2 - bc| / (a^2 + bc)]	-1.000	-.959	-.845	-.674	-.447	.143	.447	.585	.674	.737	.783
[a - √(bc)] / [a + √(bc)]	-1.000	-.661	-.420	-.240	-.101	.010	.101	.177	.240	.295	.342

Tabla 4. Resultado del cálculo del «Coeficiente de asociación multiplicativa» (2)

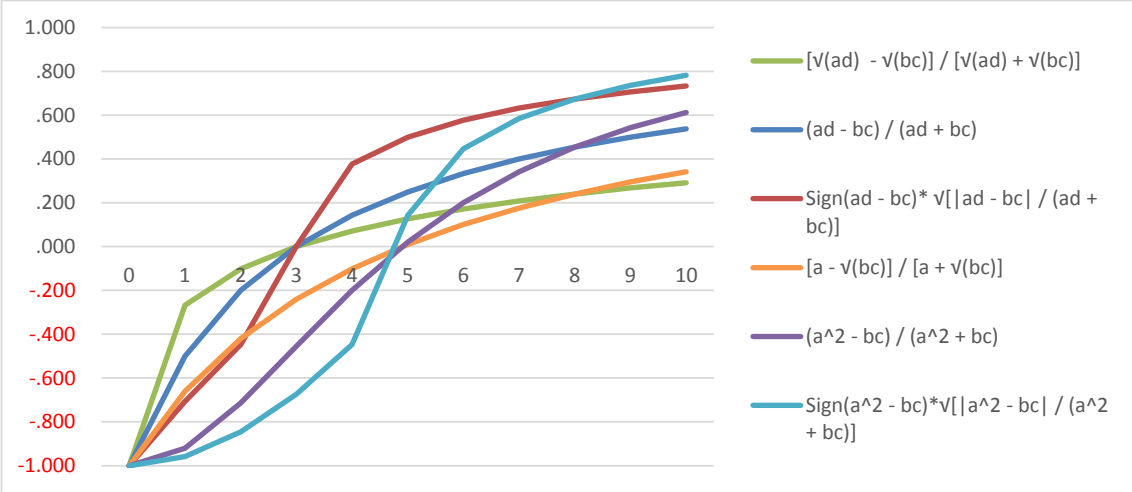

Gráfico 3. Resultado del cálculo del «Coeficiente de asociación multiplicativa» (2)

En este grupo se reúnen los coeficientes con la operación de sustracción en el numerador, cuyo efecto es la oscilación entre 0.0 y 1.0.

[3] Coeficientes de asociación correlacional

Correlational association	0	1	2	3	4	5	6	7	8	9	10
(ad - bc) / √[(a+b)(c+d)(a+c)(b+d)]	-.378	-.209	-.089	.000	.069	.124	.169	.206	.238	.265	.289
Sign(ad - bc)*√{|ad - bc| / √[(a+b)(c+d)(a+c)(b+d)]}	-.615	-.457	-.298	.000	.263	.352	.411	.454	.488	.515	.537
[√(ad) - √(bc)] / √√[(a+b)(c+d)(a+c)(b+d)]	-.615	-.236	-.095	.000	.070	.126	.170	.208	.239	.267	.290
a / √[(a+b)(a+c)]	.000	.169	.289	.378	.447	.503	.548	.585	.617	.645	.668

Tabla 5. Resultado del cálculo del «Coeficiente de asociación correlacional»

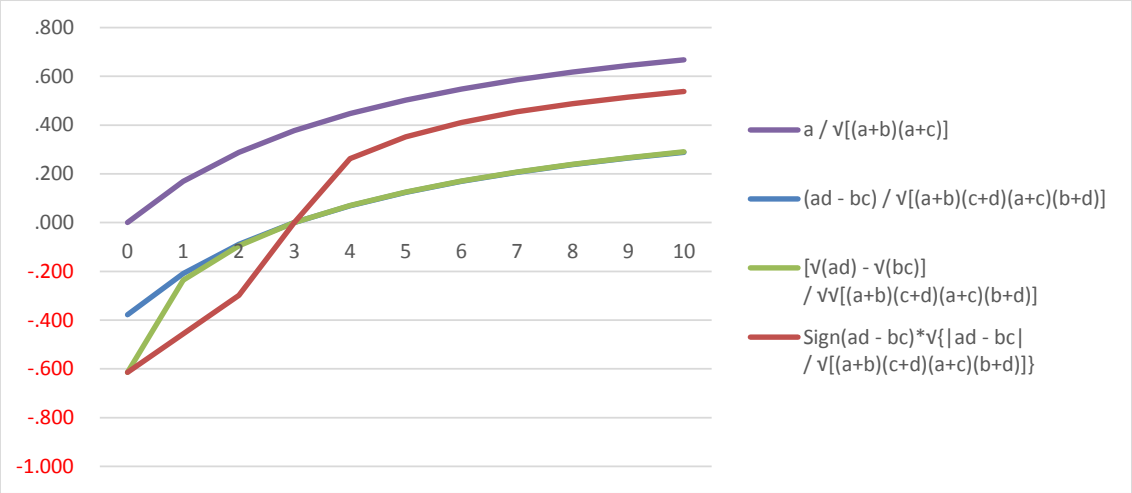

Gráfico 4. Resultado del cálculo del «Coeficiente de asociación correlacional»

Estos cuatro coeficientes sofisticados se derivan del coeficiente de correlación de Pearson. Entre el coeficiente Phi, (*ad* - *bc*) / √[(*a*+*b*)(*c*+*d*)(*a*+*c*)(*b*+*d*)], que incluye el valor de *d* y el de Ochiai, *a* /√[(*a*+*b*)(*a*+*c*)]; sin él, hubo una polémica en la historia de lingüística indoeuropea, en la que Kroeber (1937, 1969) utilizó Phi y Ellegard (1959) abogó por Ochiai. Según nuestra experiencia, debemos tener cuidado ante datos con alto valor de *d*, que influye más que el valor *a*, valor representativo de la asociación.[4]

3. El *Diccionario de americanismos*

Creemos haber enumerado la mayor parte de los posibles coeficientes indicadores de grado de asociación entre los dos parámetros variables, teniendo en cuenta la constitución y componentes de cada uno. Sin embargo, a pesar de efectuar diversos experimentos, no hemos agotado los casos posibles de distribución de los cuatro valores componentes. Se trata de los datos cuyos parámetros variables presentan las sumas, que varían considerablemente en el cuadro de matriz cualitativo. Por otra parte, el valor *d* (-/-) presenta una frecuencia sumamente alta cuando tratamos unos datos de envergadura grande.

[4] Véanse también Reed (1952) y Moore and Kimball. (1994).

Precisamente el *Diccionario de americanismos*, de la Asociación de Academias de la Lengua Española (2010), reúne más de 7.386 casos del léxico variable recogido en 19 países hispanoamericanos. Utilizamos el «Índice sinonímico» (p. 2.223-2.243) y sus correspondientes artículos dentro del cuerpo lexicográfico (p. 1-2.220). Reproducimos la parte inicial de nuestro material base del estudio, uno del dato original y otro del dato cualitativo:[5]

ARCHISEMA	FORMA	PAÍS	ACEPCIÓN
abandonar	amurar	Ar, Ur.	Abandonar una persona a alguien.
abandonar	botar(se)	Pe, Ch.	Abandonar algo o a alguien.
abandonar	botar(se)	Mx, Ho, ES, Ni, Pe	Abandonar o dejar sin cuidado a alguien o algo.
abandonar	chantar(se)	Co.	Abandonar, burlar el novio a la novia o viceversa. pop + cult→espon.
abandonar	cuitear	EU:SO.	Abandonar un juego o una tarea.

ARCHISEMA/FORMA	Ar	Bo	Ch	Co	CR	Cu	Ec	ES	Gu	Ho	Mx	Ni	Pa	Pe	PR	Py	RD	Ur	Ve
abandonar/amurar	v																	v	
abandonar/botar(se)			v											v					
abandonar/botar(se)	v		v					v		v	v	v		v					
abandonar/chantar(se)				v															
abandonar/cuitear																			

La tabla y la figura siguientes ofrecen el resultado del cálculo del Coeficiente Jaccard, a / (a+b+c), y su correspondiente dendrograma clúster (método de vecindad lejana) :

[5] Los países son: Ar (Argentina), Bo (Bolivia), Ch (Chile), Co (Colombia), CR (Costa Rica), Cu (Cuba), Ec (Ecuador), ES (El Salvador), Gu (Guatemala), Ho (Honduras), Mx (México), Ni (Nicaragua), Pa (Panamá), Pe (Perú), PR (Puerto Rico), Py (Paraguay), RD (República Dominicana), Ur (Uruguay), Ve (Venezuela). Hemos excluido abreviaturas indicadoras de la región, registros, frecuencias, etc.

Jaccard	Ar	Bo	Ch	Co	CR	Cu	Ec	ES	Gu	Ho	Mx	Ni	Pa	Pe	PR	Py	RD	Ur	Ve
Ar	1.000	0.305	0.266	0.170	0.160	0.183	0.202	0.222	0.200	0.219	0.210	0.212	0.162	0.238	0.184	0.193	0.188	0.500	0.184
Bo	0.305	1.000	0.282	0.213	0.180	0.197	0.241	0.300	0.247	0.302	0.266	0.301	0.197	0.294	0.218	0.153	0.212	0.248	0.217
Ch	0.266	0.282	1.000	0.171	0.153	0.161	0.170	0.205	0.177	0.208	0.178	0.204	0.165	0.237	0.201	0.144	0.180	0.239	0.180
Co	0.170	0.213	0.171	1.000	0.189	0.173	0.215	0.206	0.176	0.222	0.199	0.223	0.203	0.194	0.177	0.130	0.178	0.158	0.245
CR	0.160	0.180	0.153	0.189	1.000	0.150	0.196	0.192	0.228	0.250	0.196	0.260	0.232	0.168	0.186	0.128	0.202	0.152	0.203
Cu	0.183	0.197	0.161	0.173	0.150	1.000	0.147	0.223	0.156	0.223	0.201	0.225	0.163	0.173	0.213	0.115	0.233	0.162	0.211
Ec	0.202	0.241	0.170	0.215	0.196	0.147	1.000	0.188	0.193	0.197	0.205	0.200	0.195	0.230	0.163	0.166	0.180	0.185	0.185
ES	0.222	0.300	0.205	0.206	0.192	0.223	0.188	1.000	0.326	0.444	0.283	0.408	0.215	0.195	0.237	0.091	0.225	0.178	0.214
Gu	0.200	0.247	0.177	0.176	0.228	0.156	0.193	0.326	1.000	0.312	0.256	0.315	0.183	0.184	0.187	0.109	0.196	0.177	0.172
Ho	0.219	0.302	0.208	0.222	0.250	0.223	0.197	0.444	0.312	1.000	0.275	0.427	0.216	0.209	0.215	0.110	0.229	0.175	0.208
Mx	0.210	0.266	0.178	0.199	0.196	0.201	0.205	0.283	0.256	0.275	1.000	0.285	0.201	0.209	0.220	0.111	0.210	0.187	0.182
Ni	0.212	0.301	0.204	0.223	0.260	0.225	0.200	0.408	0.315	0.427	0.285	1.000	0.225	0.201	0.236	0.100	0.243	0.167	0.212
Pa	0.162	0.197	0.165	0.203	0.232	0.163	0.195	0.215	0.183	0.216	0.201	0.225	1.000	0.159	0.172	0.111	0.194	0.145	0.208
Pe	0.238	0.294	0.237	0.194	0.168	0.173	0.230	0.195	0.184	0.209	0.209	0.201	0.159	1.000	0.181	0.157	0.180	0.222	0.183
PR	0.184	0.218	0.201	0.177	0.186	0.213	0.163	0.237	0.187	0.215	0.220	0.236	0.172	0.181	1.000	0.099	0.248	0.164	0.208
Py	0.193	0.153	0.144	0.130	0.128	0.115	0.166	0.091	0.109	0.110	0.111	0.100	0.111	0.157	0.099	1.000	0.107	0.201	0.121
RD	0.188	0.212	0.180	0.178	0.202	0.233	0.180	0.225	0.196	0.229	0.210	0.243	0.194	0.180	0.248	0.107	1.000	0.172	0.232
Ur	0.500	0.248	0.239	0.158	0.152	0.162	0.185	0.178	0.177	0.175	0.187	0.167	0.145	0.222	0.164	0.201	0.172	1.000	0.161
Ve	0.184	0.217	0.180	0.245	0.203	0.211	0.185	0.214	0.172	0.208	0.182	0.212	0.208	0.183	0.208	0.121	0.232	0.161	1.000

Tabla 6. Resultado del cálculo del Coeficiente Jaccard

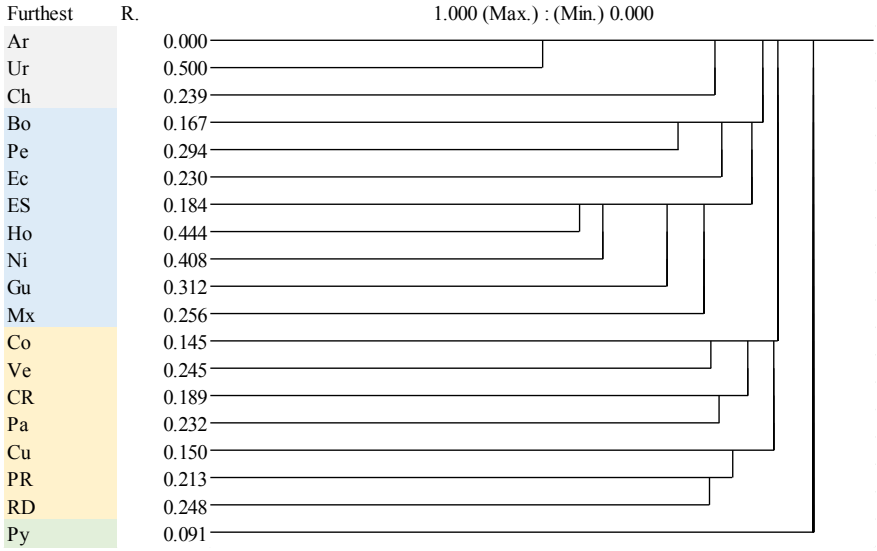

Gráfico 5. Dendrograma resultante del cálculo del Coeficiente Jaccard

Lo que nos sorprende es la situación de Py (Paraguay), que se encuentra al final, independiente de todos los restantes. Esto es debido al poco valor que posee con todos los otros países, tanto de la región de La Plata (Argentina, Uruguay) como del resto de ella. Sin embargo, en nuestros estudios sobre la variación léxica del español (Ávila *et al.* 2003, Ueda 2008), Paraguay suele agruparse con Argentina y Uruguay en la taxonomía diatópica. Al buscar las causas de esta independencia paraguaya,

encontramos el origen de la discrepancia en el cuadro siguiente, que expresa el valor *a* (+/+) entre 19 países tratados:

Valor a	Ar	Bo	Ch	Co	CR	Cu	Ec	ES	Gu	Ho	Mx	Ni	Pa	Pe	PR	Py	RD	Ur	Ve
Ar	1759	902	694	448	395	504	487	904	570	817	624	732	412	609	551	374	527	983	495
Bo	902	2100	803	602	488	592	627	1227	745	1134	829	1046	541	797	693	352	644	653	629
Ch	694	803	1547	419	352	424	389	809	482	746	513	674	387	567	557	265	477	529	453
Co	448	602	419	1328	386	417	436	777	447	748	526	683	426	444	468	216	438	344	542
CR	395	488	352	386	1099	340	366	696	512	777	481	726	432	360	451	187	449	302	427
Cu	504	592	424	417	340	1504	338	862	428	781	559	720	377	429	577	212	582	375	510
Ec	487	627	389	436	366	338	1133	687	452	646	505	592	379	475	409	240	414	363	399
ES	904	1227	809	777	696	862	687	3215	1200	1844	1114	1633	781	754	956	313	879	665	817
Gu	570	745	482	447	512	428	452	1200	1662	1058	714	978	442	478	541	217	530	430	454
Ho	817	1134	746	748	777	781	646	1844	1058	2785	998	1558	706	725	808	331	813	591	726
Mx	624	829	513	526	481	559	505	1114	714	998	1840	945	508	561	652	240	594	477	504
Ni	732	1046	674	683	726	720	592	1633	978	1558	945	2423	664	640	803	271	782	516	673
Pa	412	541	387	426	432	377	379	781	442	706	508	664	1192	357	435	174	449	301	451
Pe	609	797	567	444	360	429	475	754	478	725	561	640	357	1409	489	266	456	472	439
PR	551	693	557	468	451	577	409	956	541	808	652	803	435	489	1779	211	666	418	552
Py	374	352	265	216	187	212	240	313	217	331	240	271	174	266	211	553	206	292	214
RD	527	644	477	438	449	582	414	879	530	813	594	782	449	456	666	206	1577	407	566
Ur	983	653	529	344	302	375	363	665	430	591	477	516	301	472	418	292	407	1191	363
Ve	495	629	453	542	427	510	399	817	454	726	504	673	451	439	552	214	566	363	1428

Tabla 7. Valor a (+/+) entre los 19 países

En la Tabla 7 se observa que en la línea diagonal del rincón superior izquierdo al inferior derecho, se encuentran los valores de auto-correspondencia, que registra la frecuencia del léxico aparecido en la lista. En el cuadro, Argentina (Ar) posee 1.759 palabras registradas, mientras que Paraguay (Py), solo 553. Estas cifras corresponden a *a*+*b* y *a*+*c*, respectivamente. Notamos que el valor *a* (+/+), 374, es relativamente bajo para Ar (21,3%), mientras que para Py es considerablemente alto (67.6%). En todos los coeficientes tratados en la sección anterior no se hace distinción entre *b* y *c*, o lo que es lo mismo, entre *a*+*b* y *a*+*c*. Consideramos, por otra parte, que el grado de conexión entre Ar y Py es sumamente grande, si tenemos en cuenta el punto de vista de parte de Paraguay. Efectivamente, en la fila de Py, el valor máximo (374) de *a*(+/+) se encuentra con Ar, por encima de todos los otros países. En un caso como este, donde existe una diferencia grande entre las frecuencias del léxico tratado, proponemos utilizar otro índice de asociación, definido como sigue:

Asociación mayor = My [*a* / (*a*+*b*), *a* / (*a*+*c*)]

donde «My» representa una función que selecciona mayor cifra entre los dos valores dentro de los corchetes. Por ejemplo, entre Ar y Py, se hace el cálculo de 374/1759 y otro de 374/553 y se selecciona el valor del último: .676. El resultado es el siguiente cuadro y el dendrograma correspondiente:

Mayor	Ar	Bo	Ch	Co	CR	Cu	Ec	ES	Gu	Ho	Mx	Ni	Pa	Pe	PR	Py	RD	Ur	Ve
Ar	1.000	0.513	0.449	0.337	0.359	0.335	0.430	0.514	0.343	0.464	0.355	0.416	0.346	0.432	0.313	0.676	0.334	0.825	0.347
Bo	0.513	1.000	0.519	0.453	0.444	0.394	0.553	0.584	0.448	0.540	0.451	0.498	0.454	0.566	0.390	0.637	0.408	0.548	0.440
Ch	0.449	0.519	1.000	0.316	0.320	0.282	0.343	0.523	0.312	0.482	0.332	0.436	0.325	0.402	0.360	0.479	0.308	0.444	0.317
Co	0.337	0.453	0.316	1.000	0.351	0.314	0.385	0.585	0.337	0.563	0.396	0.514	0.357	0.334	0.352	0.391	0.330	0.289	0.408
CR	0.359	0.444	0.320	0.351	1.000	0.309	0.333	0.633	0.466	0.707	0.438	0.661	0.393	0.328	0.410	0.338	0.409	0.275	0.389
Cu	0.335	0.394	0.282	0.314	0.309	1.000	0.298	0.573	0.285	0.519	0.372	0.479	0.316	0.304	0.384	0.383	0.387	0.315	0.357
Ec	0.430	0.553	0.343	0.385	0.333	0.298	1.000	0.606	0.399	0.570	0.446	0.523	0.335	0.419	0.361	0.434	0.365	0.320	0.352
ES	0.514	0.584	0.523	0.585	0.633	0.573	0.606	1.000	0.722	0.662	0.605	0.674	0.655	0.535	0.537	0.566	0.557	0.558	0.572
Gu	0.343	0.448	0.312	0.337	0.466	0.285	0.399	0.722	1.000	0.637	0.430	0.588	0.371	0.339	0.326	0.392	0.336	0.361	0.318
Ho	0.464	0.540	0.482	0.563	0.707	0.519	0.570	0.662	0.637	1.000	0.542	0.643	0.592	0.515	0.454	0.599	0.516	0.496	0.508
Mx	0.355	0.451	0.332	0.396	0.438	0.372	0.446	0.605	0.430	0.542	1.000	0.514	0.426	0.398	0.366	0.434	0.377	0.401	0.353
Ni	0.416	0.498	0.436	0.514	0.661	0.479	0.523	0.674	0.588	0.643	0.514	1.000	0.557	0.454	0.451	0.490	0.496	0.433	0.471
Pa	0.346	0.454	0.325	0.357	0.393	0.316	0.335	0.655	0.371	0.592	0.426	0.557	1.000	0.299	0.365	0.315	0.377	0.253	0.378
Pe	0.432	0.566	0.402	0.334	0.328	0.304	0.419	0.535	0.339	0.515	0.398	0.454	0.299	1.000	0.347	0.481	0.324	0.396	0.312
PR	0.313	0.390	0.360	0.352	0.410	0.384	0.361	0.537	0.326	0.454	0.366	0.451	0.365	0.347	1.000	0.382	0.422	0.351	0.387
Py	0.676	0.637	0.479	0.391	0.338	0.383	0.434	0.566	0.392	0.599	0.434	0.490	0.315	0.481	0.382	1.000	0.373	0.528	0.387
RD	0.334	0.408	0.308	0.330	0.409	0.387	0.365	0.557	0.336	0.516	0.377	0.496	0.377	0.324	0.422	0.373	1.000	0.342	0.396
Ur	0.825	0.548	0.444	0.289	0.275	0.315	0.320	0.558	0.361	0.496	0.401	0.433	0.253	0.396	0.351	0.528	0.342	1.000	0.305
Ve	0.347	0.440	0.317	0.408	0.389	0.357	0.352	0.572	0.318	0.508	0.353	0.471	0.378	0.312	0.387	0.387	0.396	0.305	1.000

Tabla 8. Resultado del cálculo del «Coeficiente de asociación mayor»

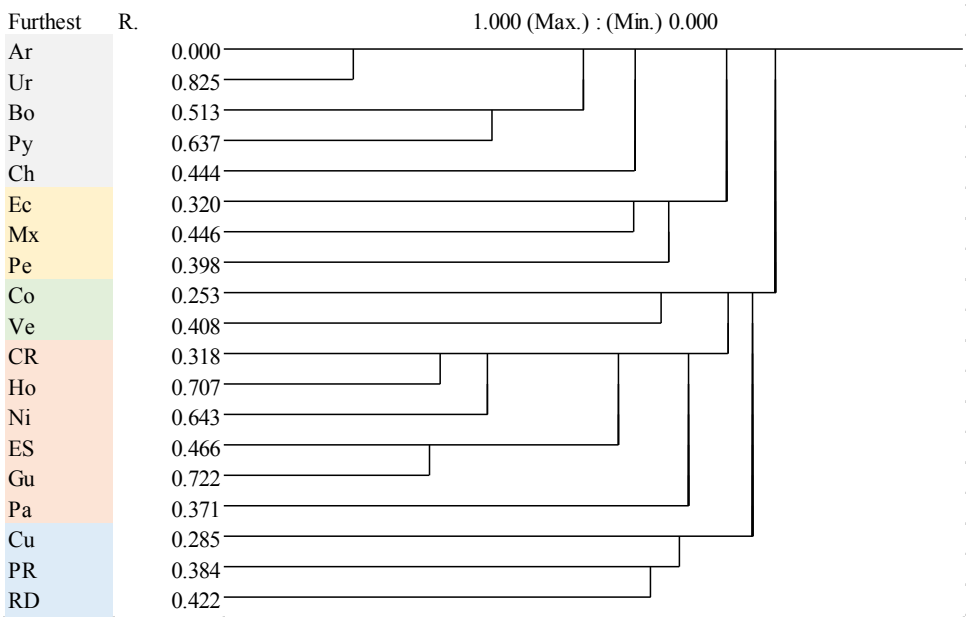

Gráfico 6. Dendrograma resultante del cálculo de «Coeficiente de asociación mayor»

4. Final

En este estudio hemos comparado distintas fórmulas para calcular el grado de asociación entre los parámetros variables, en nuestro caso, entre los países hispanoamericanos. En teoría, es posible formular múltiples ecuaciones para coeficientes de asociación y se pueden realizar unos experimentos de escala menor. En la práctica, por otra parte, podemos encontrarnos con unos datos interesantes y, al mismo tiempo, difíciles de tratar por la heterogeneidad numérica propia de materiales. Esto ocurre sobre todo cuando tratamos una base no preparada precisamente para el estudio taxonómico: diccionarios, mapas lingüísticos, documentos históricos, etc., a diferencia de los materiales recogidos con control de parámetros en las investigaciones sociolingüísticas y psicolingüísticas.

Para resolver estas dificultades, creemos que es conveniente poseer un amplio recurso de técnicas y métodos, observar las características propias de materiales objeto del estudio y aplicar a estos el método más apropiado posible. La función de «Asociación mayor» es simple y unilateral, en el sentido de que se toma en consideración solo uno de los dos posibles índices de asociación. Se selecciona el mayor de las ratios de coincidencia con el otro parámetro. A sabiendas de que este valor tiene el defecto de considerar solo una parte, lo hemos utilizado por su grado positivo mayor que posee con respecto al otro. Tras confirmar su utilidad, hemos incluido la función de «Asociación mayor» en nuestro paquete de programas para análisis de datos lingüísticos numéricos.[6]

Referencias

ANDERBERG, Michael R. (1973) *Cluster analysis for applications*, New York: Academic Press.

ASOCIACIÓN DE ACADEMIAS DE LA LENGUA ESPAÑOLA (2010) *Diccionario de americanismos,* Madrid: Santillana.

[6] Se encuentra el paquete NUMEROS en: http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/. Dentro de poco el paquete se convertirá en una versión multilingüe.

ÁVILA, Raúl, José Antonio SAMPER & H. UEDA (2003) *Pautas y pistas en el análisis del léxico* hispanoamericano, Madrid/Frankfurt: Iberoamericana Vervuert.

ELLEGÅRD, Alvar (1959) "Statistical measurement of linguistic relationship", *Language*, 35, 131-156.

GARCÍA HOZ, Víctor (1953) *Vocabulario usual, vocabulario común y vocabulario fundamental.* Madrid: Consejo Superior de Investigaciones Científicas.

JUILLAND, Alphonse & Eugenio CHANG-RODRIGUEZ (1964) *Frequency dictionary of Spanish words,* The Hague: Mouton.

KROEBER, Alfred L. & C. Douglas CHRÉTIEN (1937) "Quantitative classification of Indo-european languages", *Language,* 13, 83.

KROEBER, Alfred L. & C. D. CHRÉTIEN (1939) "The statistical technique and Hittite", *Language,* 15, 69-71.

KROEBER, Alfred L. & C. D. CHRÉTIEN (1960) "Statistics, Indo-European and taxonomy", *Language*, 36, 1-21.

MOORE, Carmella C. & A. Kimball ROMNEY (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.

REED, David W. & John L. SPICER (1952) "Correlation methods of comparing idiolects in *a* transition area", *Language,* 28, 348-359.

ROSEMBURG, Charles H. (1989) *Cluster analysis for researchers*, Florida: Robert E. Krieger Publishing Company, Inc. Malabar.

UEDA, Hiroto (2008) "Análisis dialectométrico del léxico variable español: Interpretación taxonómica de resultados", en *El español de América, Actas del VI Congreso Internacional de El español de América*, Valladolid: Instituto Interuniversitario de Estudios de Iberoamérica y Portugal, Universidad de Valladolid, 813-822.