Centro Internacional de Estudios Superiores del Español. Fundación Comillas. 2022.

Investigación de la variación léxica del español

Métodos de la zonificación multivariante y visualización cartográfica[1]

Hiroto Ueda (Universidad de Tokio)

§1. En 1993, en el X Congreso de ALFAL (Asociación de Lingüística y Filología de América Latina) en Veracruz de México, iniciamos un proyecto de investigación, denominado VARILEX («Variación léxica del español en el mundo»). Desde entonces hemos venido investigando la variación léxica del español moderno en España, Guinea Ecuatorial e Hispanoamérica *in situ* y por correspondencia. En la misma investigación, hemos recurrido a la ayuda de colaboración prestada por los investigadores locales, enviándoles cuatro cuestionarios con dibujos y explicación donde los 4 encuestados de dos sexos y dos generaciones seleccionarían las formas que utilizan en el sitio usualmente.

En la pregunta hemos pedido que subraye las palabras o expresiones que el encuestado mismo utilice, ofreciendo la explicación (acepción) del objeto de que se trata, seguida de una lista de palabras, de la siguiente manera:

Explicaciones

1. Subraye la(s) palabra(s) o expresión(es) si usted mismo la(s) utiliza.
Ejemplo:

[A001] JACKET: Prenda de vestir masculina, que forma con el chaleco y los pantalones el traje completo. No es de paño con botones dorados.

(1)americana, (2)capa, (3)chaleco, (4)chaqueta, (5)gabán, (6)leva, (7)paletón, (8)saco, (9)saco de terno, (10)saco de traje, (11)traje, (12)vestón.

&Otro(s)_____, #No se me ocurre.

$Comentario:

Fig. 1. Cuestionario, parte inicial

[1] Este documento es nuestra traducción al español de la comunicación que presentamos con el mismo título en japonés en I Congress of Geolinguistic Society of Japan, Aoyama Gakuin University (October 6, 2019).

Hemos subido el resultado de aproximadamente 2000 ítems en la página web de la Universidad de Tokio (Ueda 1993-):

Fig. 2. https://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/

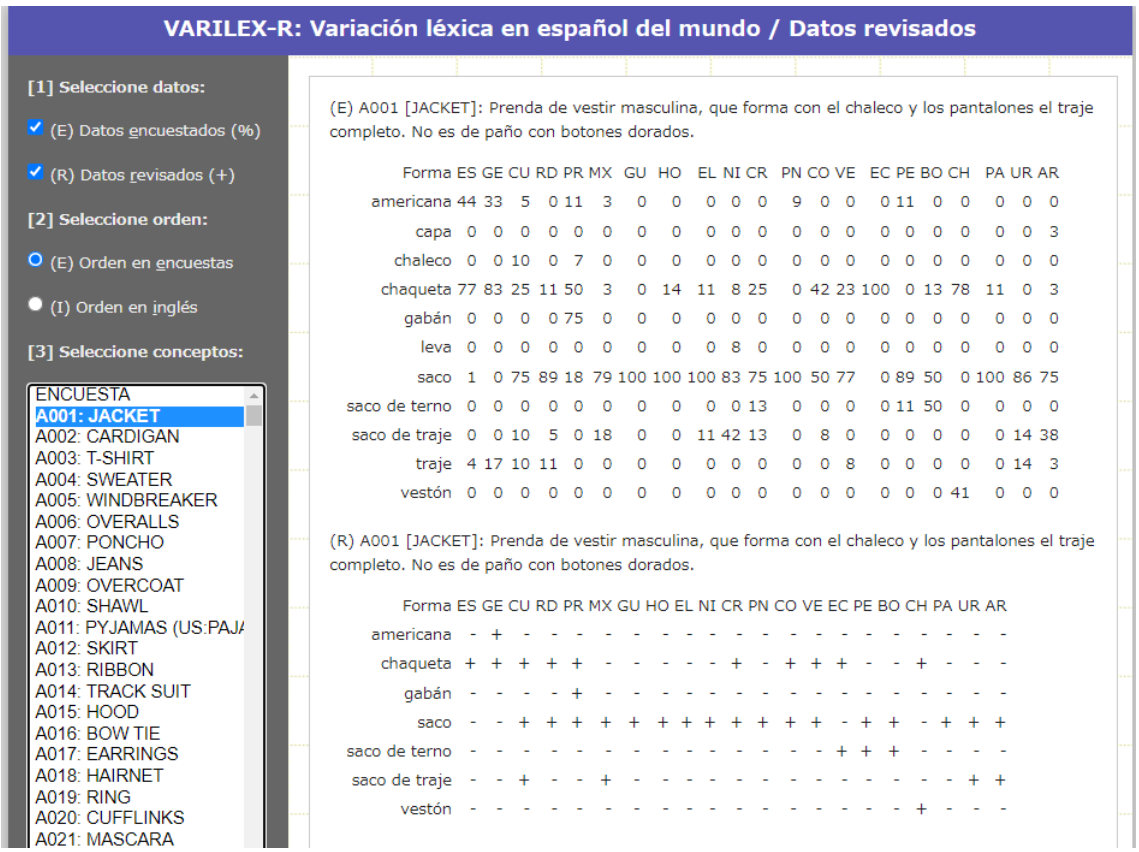


Fig. 3. https://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex-r/

En los métodos de la geografía lingüística tradicional de Europa y de la dialectología española, se practica la encuesta directa al informante anciano en pueblos pequeños sobre el léxico inmenso tradicional de la localidad y el investigador transcribe la respuesta en un cuaderno, que posteriormente se copia en los mapas de atlas lingüísticos. Nuestro método difiere del método tradicional en los siguientes puntos: grandes ciudades y pequeños pueblos, múltiples encuestados y único encuestado anciano, encuesta por correspondencia y encuesta directa, encuesta continua y encuesta única, procesamiento informático y trabajo manual. Los dos métodos diferentes en condiciones, objetivos y métodos no son excluyentes sino complementarios:

*	Dialectología tradicional	VARILEX
Área	País o región	Mundo hispanohablante
Sitio	Pueblos	Grandes ciudades
Encuestados	Ancianos	Jóvenes y mayores
Método	Entrevista	Correspondencia
Sucesividad	Única	Continua
Edición	Manual	Automática
Proceso	Atlas > Datos > Análisis	Datos > Atlas + análisis

Tabla 1.Comparación

En la misma investigación, hemos preparado un sistema de cuantificación y visualización manejable en la página web (Ueda y Moreno Sandoval 2017-). El programa produce la tabla bidimensional de forma y lugar junto con el mapa de distribución, donde se aplican métodos multivariantes para obtener la zonificación geográfica continua que combina las zonas en el esquema y puntos geográficos en el mapa. Para la visualización hemos adoptado el método de red (Fig. 9) y casco convexo (Fig. 10).

§2. En esta ocasión, tomando como ejemplo la variación del ítem número 1 (A001) 'jacket (for men)', observaremos su distribución geográfica en los países hispanohablantes y presentaremos nuestro método de zonificación y visualización. Utilizaremos los datos ofrecidos por investigadores locales (datos de tipo 1/0, representados por '+' en Tabla 2). En la siguiente tabla, se manifiesta la distribución de cada forma en los países correspondientes:

País	*americana*	*chaqueta*	*gabán*	*saco*	*saco de terno*	*saco de traje*	*vestón*
1.ES		+					
2.GE	+	+					
3.CU		+		+		+	
4.RD		+		+			
5.PR		+	+	+			
6.MX				+		+	
7.GU				+			
9.HO				+			
9.EL				+			
10.NI				+			
11.CR		+		+			
12.PN				+			
13.CO		+		+			
14.VE		+		+			
15.EC		+			+		
16.PE				+	+		
17.BO				+	+		
18.CH		+					+
19.PA				+			
20.UR				+		+	
21.AR				+		+	

Tabla 2. Distribución de formas en países

(1.ES: España, 2.GE: Guinea Ecuatorial, 3.CU: Cuba, 4.RD: República Dominicana, 5.PR: Puerto Rico, 6.MX: México, 7.GU: Guatemala, 8.HO: Honduras, 9.EL: El Salvador, 10.NI: Nicaragua, 11.CR: Costa Rica, 12.PN: Panamá, 13.CO: Colombia, 14.VE: Venezuela, 15.EC: Ecuador, 16.PE: Perú, 17.BO: Bolivia, 18.CH: Chile, 19.PA: Paraguay, 20.UR: Uruguay, 21.AR: Argentina)

La siguiente tabla muestra la distribución diagonalizada de las formas frecuentes (*chaqueta, saco, saco de traje, saco de terno*) y países correspondientes, donde hemos utilizado la distancia Minkowski de elevación 3 para obtener los coeficientes verticales (Xn) y los horizontales (Yp):

X1(EC) = [(1^3 + 2^3) / 2] ^(1/3) = 1.651

Dst.	1.s. de terno	2.chaqueta	3.saco	4.s. de traje	Xn
1.EC	+	+			1.651
2.ES.GE.CH		+			2.000
3.PE.BO	+		+		2.410
4.PR.RD.CR.CO.VE		+	+		2.596
5.GU.HO.EL.NI.PN.PA			+		3.000
6.CU		+	+	+	3.208
7.MX.UR.AR			+	+	3.570
Yp	2.410	4.165	5.372	6.538	

Tabla 3. Distribución diagonalizada

La distribución diagonalizada, obtenida de Análisis de correspondencia (Ueda y Moreno 2017a) indica con la flecha oblicua la correlación continua que hay entre las formas y países. Para obtener la zonificación separada, en primer lugar, preparamos la matriz simétrica de distancia unidimensional (UniDis) por medio del valor absoluto de la diferencia (Apéndice). Por ejemplo, la distancia entre 1. [EC] y 2. [ES.GE.CH] es:

UniDis(1. [EC] , 2. [ES.GE.CH]) = | 1.651 - 2.000 | = .349.

Distancia	1	2	3	4	5	6	7
1.EC	.000	.349	.759	.945	1.349	1.557	1.919
2.ES.GE.CH	.349	.000	.410	.596	1.000	1.208	1.570
3.PE.BO	.759	.410	.000	.186	.590	.797	1.160
4.PR.RD.CR.CO.VE	.945	.596	.186	.000	.404	.611	.974
5.GU.HO.EL.NI.PN.PA	1.349	1.000	.590	.404	.000	.208	.570
6.CU	1.557	1.208	.797	.611	.208	.000	.362
7.MX.UT.AR	1.919	1.570	1.160	.974	.570	.362	.000

Tabla 4. Matriz simétrica de distancia

Los dendrogramas que produce el Análisis de conglomeción (ing. 'Cluster analysis') no son convenientes para ver la sucesividad (o continuidad) de los datos, puesto que poseen una ambigüedad de posición que ocupa cada ítem. Veamos esta situación en forma de un objeto que se llama 'mobile' (en inglés), de estructura similar al dendrograma, donde observamos los cambios de sitio horizontal de cada ballena a pesar de tratarse del mismo objeto:

Fig. 4. Mobile (1)

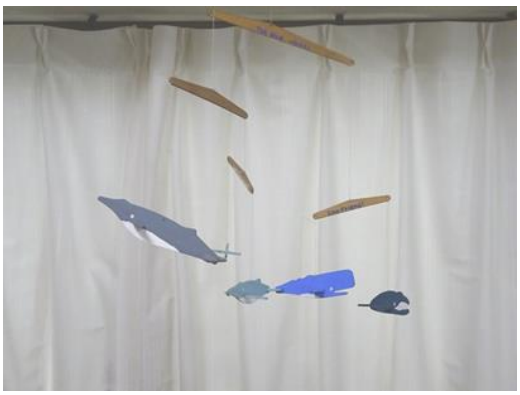


Fig. 5. Mobile (2)

Esto representa la ambigüedad de los sitios de objetos, es decir, no hay manera de determinar el orden de los objetos, por ejemplo:

$$\{ [(A\ B)\ ]\ [(C\ D)\ E]\} = \{[E\ (D\ C)]\ [B\ A)]\}\ (?)$$

No obstante, el orden de los países representados en la distribución oblicua (Tabla 3) está garantizado por el patrón diagonalizado y distancias por él ordenadas.

A partir de la misma matriz, por medio del Análisis de conglomeración (Ueda y Moreno 2017a), se obtiene el siguiente dendrograma:

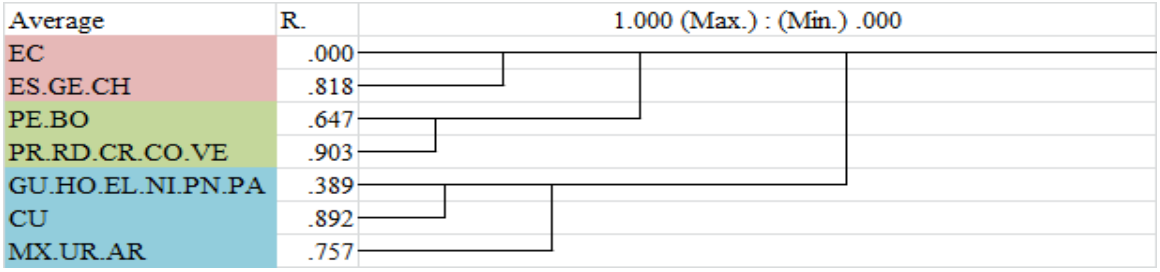


Fig. 6. Dendrograma de países

De la misma manera, realizamos el Análisis de conglomeración de formas léxicas (Yp):

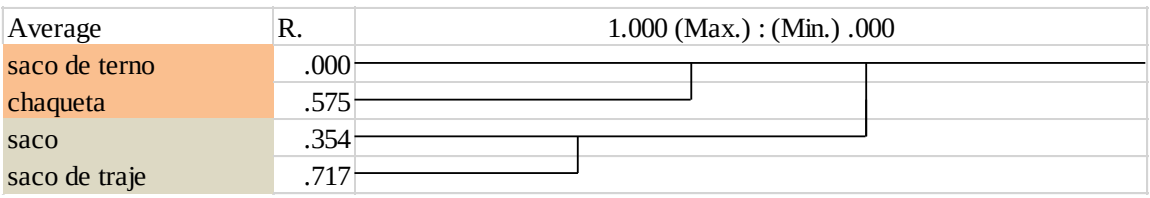


Fig. 7. Dendrograma de formas

Reflejamos el resultado del Análisis de conglomeración en la tabla diagonalizada anterior:

Dst.a	saco de terno	chaqueta	saco	saco de traje
EC	1	1		
ES.GE.CH		1		
PE.BO	1		1	
PR.RD.CR.CO.VE		1	1	
GU.HO.EL.NI.PN.PA			1	
CU		1	1	1
MX, UR.AR			1	1

Fig. 8. Distribución diagonalizada conglomerada

§3. A partir de esta tabla, vamos a considerar la distribución de dos variantes principales: *saco* y *chaqueta*. El mapa de red endocéntrica es idóneo para visualizar la distribución compleja de formas reducidas:

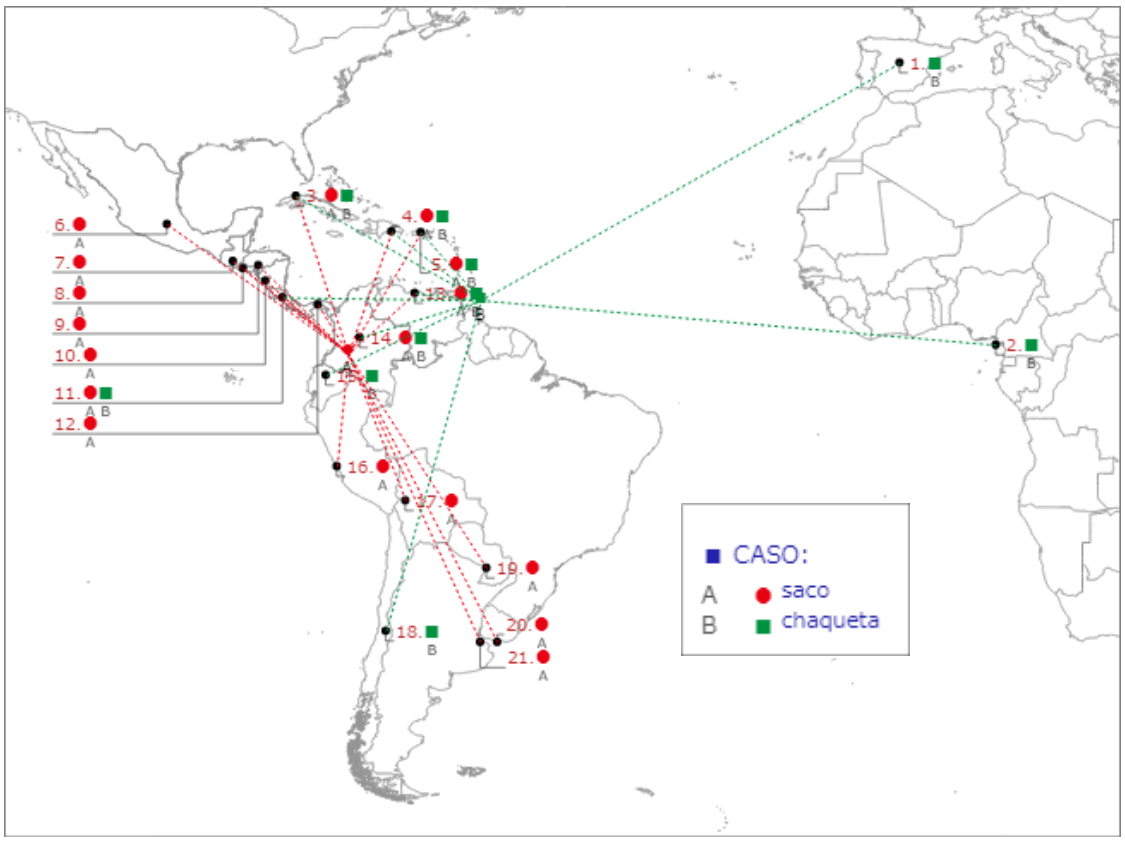


Fig. 9. Distribución geográfica de *saco* y *chaqueta*

Al observar este mapa, comprobamos la distribución de [A]: *saco* en

México y América Central, que corresponde a la zona occidental, la de [B]: *chaqueta* en España, Guinea Ecuatorial y Chile, que constituyen la zona oriental, y la de [A] + [B]: *saco* +*chaqueta* en los países de Caribe, es decir, zona central. De modo que observamos la continuidad geográfica en forma de:

[A] : [A+B] : [B].

En general, el cambio cronológico normal de la misma distribución es:

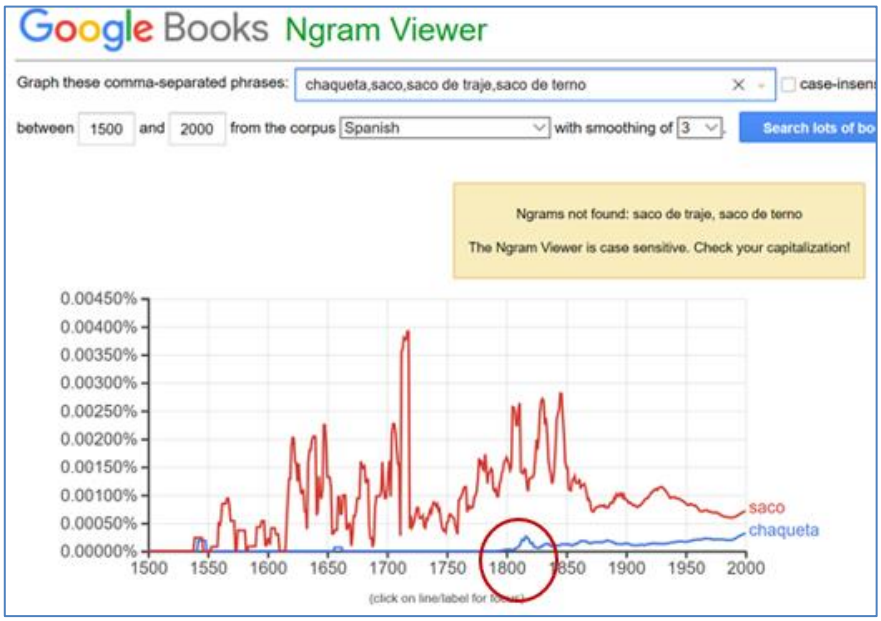
[A] → [A+B] > [B] o [B] → [A+B] → [A],

mientras que es difícil de pensar en:

*[A] → [B] → [A+B] (o *[B] → [A] → [A+B]), ni en:

*[A+B] → [A] → [B] (o *[A+B] → [B] → [A]).

Según Corominas y Pascual (1983, s.v. *saco, chaqueta*), *saco* de origen latino (< lat. SACCUS) se utilizaba como nombre de prenda de vestir en el siglo XIV, y a principios del siglo XIX se introdujo la forma francesa *jaquette* en forma de *chaqueta*. A continuación, exponemos el resultado de búsqueda de *saco* en *Google Books Ngram Viewer*, donde confirmamos la aparición de *chaqueta* en 1800:



Google Books Ngram Viewer [2019/10/03][2]

[2] Agradecemos al profesor Fumio Inoe, quien tuvo la gentileza de enviarnos esta imagen. 'saco' incluye la acepción de «Receptáculo de tela, cuero, papel, etc., por lo común de forma

Actualmente en España, Guinea Ecuatorial y Chile, la palabra *saco* con el significado de 'chaqueta' no se utiliza, de modo que se ha comprobado el cambio histórico de [A]: *saco* > [A+B]: *saco* + *chaqueta* > [B]: *chaqueta*, lo que es normal como hemos visto anteriormente.

§4. En la tabla diagonalizada anterior (Fig. 8), las distribuciones de *saco de terno* (Ecuador, Perú y Bolivia) y *saco de traje* (Cuba, México, Uruguay y Argentina) no están sobrepuestas, lo que se visualiza de la siguiente manera:

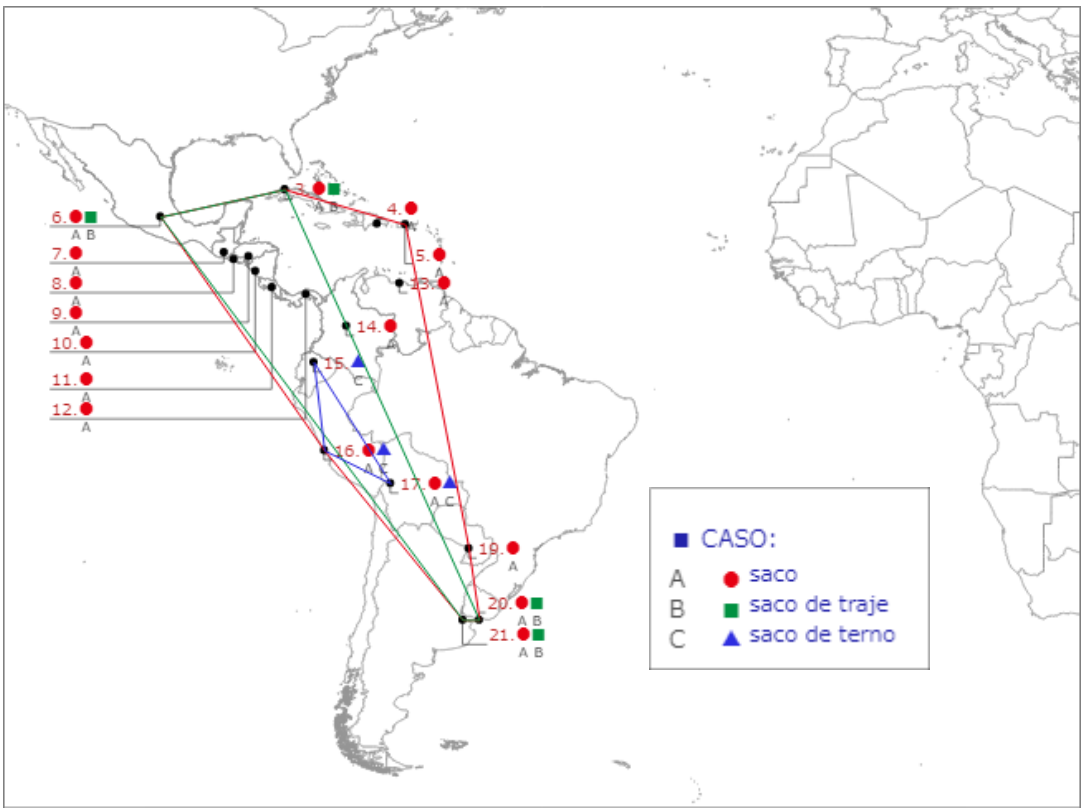


Fig. 10. Distribución geográfica de *saco, saco de traje, saco de terno*

Se trata del gráfico de casco convexo que encubre el conjunto de los puntos reactivos. La desventaja de este método está en que puede incluir los puntos negativos. No obstante, aquí es conveniente construir relaciones envolventes:

A: *saco* ⊃ B: *saco de traje* ⊃ C: *saco de terno*.

Generalmente cuando estamos ante la distribución envolvente:

rectangular o cilíndrica, abierto por uno de los lados» (DLE. s.v. *saco*).

$$X \supset Y : [X - Y - X],$$

suponemos dos posibilidades del cambio cronológico:

(1) [X] (forma antigua mayoritaria) → [Y] (forma nueva limitada),

y

(2) [Y] (forma antigua residual) → [X] (forma nueva mayoritaria).

Para apoyar la posibilidad (2), necesitamos satisfacer la condición de la continuidad geográfica de [X], puesto que es difícil suponer la aparición de una nueva forma accidentalmente coincidente en lugares separados y en especial distanciados. En nuestro caso concreto, los países de *saco de traje:* MX (México) - CU (Cuba), por una parte, y UR (Uruguay) - AR (Argentina), por otra, están separados y distanciados por los países de *saco de terno* son EC (Ecuador), PE (Perú) y BO (Bolivia). A pesar de que no encontramos información documental al respecto en los estudios anteriores ni en los corpus históricos (RAE-CORDE), podemos suponer el doble cambio de [A] → [B] y [B] → [C]:

[A]: *saco* → [B]: *saco de traje* → [C]: *saco de terno*,

basándonos en la relación envolvente arriba expuesta.

§5. Hasta hace poco, en los estudios dialectales, realizábamos todos los trabajos de investigación manualmente: recogida de datos, transcripción, procesamiento, edición de atlas, análisis de datos, visualización, publicación, etc. La situación ha cambiado en la actualidad de manera drástica. Los últimos desarrollos en la geografía lingüística, la informática, la estadística y la tecnología de comunicación son admirables. Nuestro proyecto VARILEX cuenta con ellos y no está solo, sino apoyado por los estudios de lingüística española general, que comparan sus datos con los de corpus grandes (Rojo 2021: 198), y por los profesores de escuelas y universidades que lo utilizan en las tareas de clase. Vamos a seguir estudiando la variación geográfica no solamente por encuestas personales sino también con corpus lingüísticos locales (Martínez y Ueda 2021). De esta manera, podremos aproximarnos al mundo inmenso y complejo del léxico español.

Referencia citada:

Corominas, J. y Pascual, J. A. (1983) *Diccionario crítico etimológico castellano e hispánico*. Madrid, Gredos.

Martínez, I. y Ueda H. (2021). *Inventario léxico de PRESEEA-Santander. Proyecto para el estudio sociolingüístico del español de España y América.* https://lecture.ecc.u-tokyo.ac.jp/~cuedákenkyúchiríinventario-santander.pdf [2022/12/14]

Real Academia Española. (2013) *Diccionario de la lengua española*. (23a ed.) [DLE] https://dle.rae.es/

RAE. CORDE. Real Academia Española. *Corpus Diacrónico del Español (CORDE)* http://corpus.rae.es/ [2022/12/14]

Rojo, G. (2021) *Introducción a la lingüística de corpus en español.* London: Routledge.

Ueda, H. (1993-) *VARILEX, Variación léxica del español en el Mundo* https://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/index.html

Ueda, H. y Moreno Sandoval, A. (2017a) *Análisis de datos cuantitativos para estudios lingüísticos.* [2022/12/14] https://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/4-numeros/doc/numeros-es.pdf [2022/12/14]

Ueda, H. y Moreno Sandoval, A. (2017b) *LYNEAL, Letras y números en análisis lingüísticos* [2022/12/14] https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/ [2022/12/14]

Ueda, H. y Moreno Ferna/ndez, Francisco. (2016) VARILEX-R. https://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex-r/ [2022/12/14]

Apéndice: Distancia unidimensional en R

```
UniDis=function(A=D){
   len=length(A); M=matrix(0, len, len)
   for(i in 1:len){
      for(j in 1:len){ M[i, j]=abs(A[i]-A[j])
   }}
   colnames(M)=rownames(as.matrix(A)); rownames(M)=colnames(M); M
} # Distancia unidimensional

> A=c(1.651, 2.000, 2.410, 2.596, 3.000, 3.208, 3.570)

> UniDis(A)
         [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,] 0.000 0.349 0.759 0.945 1.349 1.557 1.919
[2,] 0.349 0.000 0.410 0.596 1.000 1.208 1.570
[3,] 0.759 0.410 0.000 0.186 0.590 0.798 1.160
[4,] 0.945 0.596 0.186 0.000 0.404 0.612 0.974
[5,] 1.349 1.000 0.590 0.404 0.000 0.208 0.570
[6,] 1.557 1.208 0.798 0.612 0.208 0.000 0.362
[7,] 1.919 1.570 1.160 0.974 0.570 0.362 0.000
```