*asterisco
Vol. 3, 2025, 55-78
ISSN electrónico: 2952-3567
DOI: https://doi.org/10.14201/ast.202535578

Estratos de distanciamento léxico en español. Análisis dialectométrico

Layers of Lexical Distancing in Spanish. Dialectometric Analysis

Francisco MORENO FERNÁNDEZ
Heidelberg Center for Ibero-American Studies. Universität Heidelberg
francisco.moreno@uni-heidelberg.de
https://orcid.org/0000-0002-3136-4443
Jana WECKESSER
Heidelberg Center for Ibero-American Studies. Universität Heidelberg
jana.weckesser@uni-heidelberg.de
https://orcid.org/0009-0003-2770-5094

Recibido: 16/05/2024. Aceptado: 14/11/2024.

Resumen: Este artículo analiza la variación léxica en los países hispanohablantes mediante técnicas dialectométricas, a partir de los datos del proyecto Varilex-R. Se identifican patrones de distanciamiento léxico entre regiones y países, empleando enfoques cuantitativos para determinar similitudes y diferencias lingüísticas. A través de métodos como el análisis de *clusters* y escalamiento multidimensional, se detectan seis grandes áreas léxicas: España-Guinea Ecuatorial, México-Caribe, América Central, Andes, Río de la Plata y Chile. Posteriormente, un análisis más detallado de 12 *clusters* revela mayor complejidad en Centroamérica y diferencias específicas en Venezuela y Ecuador. Los estudios referenciales muestran que países como España, Chile, Argentina, Costa Rica y México presentan distancias léxicas variables con el resto del mundo hispano. Destaca la correlación entre distancia geográfica y distancia léxica, aunque con excepciones, como la afinidad léxica entre España y Guinea Ecuatorial, y la cercanía de México con el Caribe. Los análisis evidencian que la diversidad léxica del español no solo responde a la geografía, sino también a factores históricos y socioculturales. La metodología dialectométrica permite una visión objetiva de estas variaciones, abriendo

el camino para futuros estudios con mayor profundidad en niveles regionales y en otros aspectos lingüísticos.

Palabras clave: dialectometría, variación léxica, zonificación lingüística, distanciamiento léxico.

Abstract: *This article analyzes lexical variation in Spanish-speaking countries using dialectometric techniques, based on data from the Varilex-R project. Patterns of lexical distancing between regions and countries are identified, employing quantitative approaches to determine linguistic similarities and differences. Through methods such as cluster analysis and multidimensional scaling, six major lexical areas are detected: Spain-Equatorial Guinea, Mexico-Caribbean, Central America, Andes, Río de la Plata, and Chile. A more detailed analysis of 12 clusters later reveals greater complexity in Central America and specific differences in Venezuela and Ecuador. Reference studies show that countries such as Spain, Chile, Argentina, Costa Rica, and Mexico exhibit varying lexical distances from the rest of the Hispanic world. The relationship between geographical and lexical distance stands out, although with exceptions such as the lexical affinity between Spain and Equatorial Guinea and Mexico's proximity to the Caribbean. The findings demonstrate that Spanish lexical diversity is influenced not only by geography but also by historical and sociocultural factors. The dialectometric methodology provides an objective perspective on these variations, paving the way for future studies with greater depth at regional levels and in other linguistic aspects.*

Keywords: *dialectometry, lexical variation, linguistic zoning, lexical distancing.*

1. INTRODUCCIÓN

El análisis de la diversidad dialectal de una comunidad tan grande y extensa como la hispánica admite la aplicación de técnicas múltiples planteadas desde estrategias o enfoques diferentes. En líneas generales, se han seguido tres estrategias para reunir, ordenar y presentar los datos lingüísticos de un territorio, de modo que quedara reflejada la particularidad de cada área frente a las demás.

a) La primera estrategia ha consistido en reunir información de un área y contrastarla con la información más completa y sistemática disponible de otra área o de un corpus de referencia, para poder efectuar un contraste entre lo que es particular y lo que es compartido (*enfoque diferencial*).
b) La segunda estrategia consiste en presentar los rasgos lingüísticos de cada territorio (p. e. léxicos) como un todo, prescindiendo de la distinción entre lo que es común y lo que es compartido con otras áreas (*enfoque integral*).
c) La tercera estrategia consiste en recurrir a la información aportada por expertos procedentes de distintas áreas de interés, para cruzar sus informaciones y datos e identificar lo que es compartido y lo que no lo es (*enfoque cualitativo*).

A estas estrategias, practicadas principalmente en el campo de la lexicografía (Zimmermann, 2018), podría añadírseles una más, que consiste en reunir datos de todas las áreas analizadas y realizar una comparación sistemática de todos ellas (*enfoque multilateral*). Esta es la estrategia seguida en el análisis estadístico realizado por Moreno Fernández y Ueda (2018) sobre los materiales lingüísticos procedentes de todo el espacio hispánico reunidos en el proyecto *Varilex-R* (Ueda y Moreno Fernández, 2016).

El objetivo de este artículo es presentar un análisis léxico-estadístico desde un enfoque multilateral a partir de los materiales aportados por el proyecto *Varilex-R* y como desarrollo del análisis realizado por Moreno Fernández y Ueda en 2018. Para ello, se procederá a aplicar técnicas dialectométricas y a discutir los resultados más relevantes que el análisis arroja.

2. ANTECEDENTES: EL PROYECTO *VARILEX*

El proyecto *Varilex* (Variación léxica del español en el mundo) es una iniciativa de Hiroto Ueda, de la Universidad de Tokio, desarrollada desde los años noventa, cuyos primeros frutos ya permitieron una zonificación dialectal del español (Ueda y Ruiz Tinoco, 2007; Ueda, 2015) y que en 2016 tomó como nombre *Varilex-R* (Ueda y Moreno Fernández, 2016). En esta última fase, todos los datos allegados con anterioridad fueron revisados por expertos y agregados por países.

En la actualidad, la base *Varilex-R* ofrece unas condiciones adecuadas para proceder al análisis cuantitativo de las diferencias y semejanzas entre los usos lingüísticos de todos los países hispánicos, así como de su cohesión interna. De hecho, Moreno Fernández y Ueda (2018) aplicaron una serie de técnicas estadísticas, a partir de los datos de *Varilex-R*, que dieron respuesta a preguntas de investigación de gran calado; fundamentalmente dos: cuál es el nivel de homogeneidad-heterogeneidad lingüística de las comunidades hispanohablantes y cuáles son las áreas más particulares en cuanto a las características del español en ellas utilizado. Para responder a esas preguntas de investigación, se realizaron análisis de correlación, de *clúster*, de componentes principales y de asociación, con cálculos de índices de generalidad y particularidad.

La batería de análisis estadísticos realizados en 2018 manejó datos de distintos niveles lingüísticos (léxico, gramática, fraseología) sin proceder a un análisis segregado o parcial de cada uno de ellos. En aquel momento se primó una visión holística de la lengua sobre una disección por niveles, con el fin de apreciar la realidad dialectal en su conjunto. Por otro lado, algunas de las pruebas aplicadas manejaron todas las variables de cada variante en un plano de igualdad, sin considerar qué usos eran primarios o secundarios, cuáles pertenecían a la nómina activa y cuáles a la pasiva de cada hablante de cada región y de cada país. Además, el manejo de unidades territoriales nacionales impidió la obtención de informaciones y conclusiones relativas tanto a los

espacios internos que las conforman, como a las regiones transnacionales que sin duda existen dentro del espacio hispanohablante.

Como resultado del mencionado análisis, Moreno Fernández y Ueda presentaron un gráfico elaborado desde un análisis de componentes principales, en el que se apreciaba las distancias dialectales entre territorios (Gráfico 1).

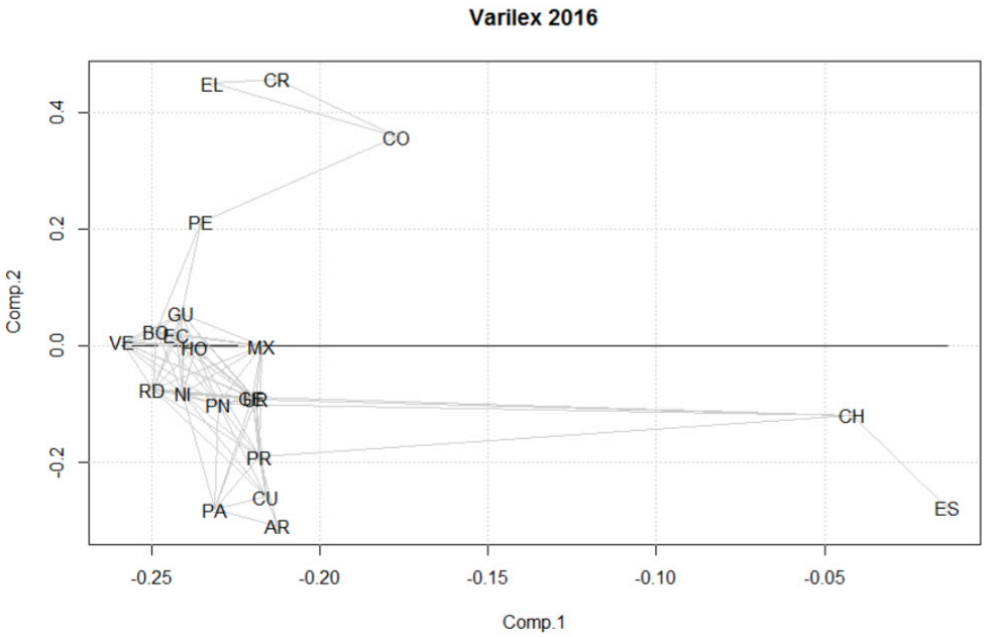


GRÁFICO 1. Representación de los dos primeros componentes principales en la base de datos *Varilex-R* (Moreno Fernández y Ueda, 2018).
Códigos de países: AR: Argentina, BO: Bolivia, CH: Chile, CO: Colombia, CR: Costa Rica, CU: Cuba, EC: Ecuador, EL: El Salvador, ES: España, GE: Guinea Ecuatorial, GU: Guatemala, HO: Honduras, MX: México, NI: Nicaragua, PA: Paraguay, PE: Perú, PN: Panamá, PR: Puerto Rico, RD: República Dominicana, UR: Uruguay, VE: Venezuela. Fuente: Moreno Fernández y Ueda, 2018.

El estudio de 2018 mostraba que la mayoría de los países hispanohablantes confluyen en un concentrado espacio de diversidad y con un equilibrio entre formas particulares y compartidas, lo que justificaría tanto el sentimiento de comunidad existente entre los hispanohablantes, como la conciencia de una identidad compartida. En ese concierto dialectal, las notas más discordantes eran los espacios de España y Chile, por un lado, y Argentina y Costa Rica, por otro. En el caso de España, su distanciamiento se derivaba del importante número de usos particulares asociados a su geografía periférica, mientras que la generalidad se explicaba por el hecho de que los usos lingüísticos de España también aparecen en otros muchos países, como primeras, segundas o terceras opciones. Por su parte, la personalidad de Chile se justificaba en su alto índice de

particularidad. Tanto Argentina, como Costa Rica y El Salvador ofrecieron una particularidad léxica mayor que otros territorios. Con todo, los aspectos comentados merecen una mayor profundización y un análisis pormenorizado por niveles lingüísticos y por país.

3. MARCO TEÓRICO Y PREGUNTAS DE INVESTIGACIÓN

Este estudio se inscribe dentro de los análisis dialectométricos, aquellos que miden las similitudes y disimilitudes entre puntos o áreas geolingüísticas diferentes. En este sentido, como los antecedentes inmediatos, nuestro análisis se adscribe a la tradición dialectométrica creada por Jean Séguy (1971) y desarrollada por investigadores como Henri Guiter (1973) o Hans Goebl (1982, 2010). Asimismo, este estudio se plantea desde un enfoque multilateral, pues aborda un análisis cruzado de datos procedentes de una multiplicidad de territorios.

Como aportación teórica, presentamos el concepto de «estrato de distanciamiento». Este concepto se refiere a los diferentes niveles de diferenciación o especificidad que pueden establecerse entre distintas áreas lingüísticas; en nuestro caso, áreas léxicas. El nivel de diferenciación propuesto tiene una base cuantitativa, condicionada por el detalle que aportan en cada momento los análisis estadísticos por medio de distintas pruebas. No se trata, pues, ni de identificar estratos de superposición léxica, como los que se han distinguido en lingüística histórica y románica (sustrato, superestrato) (Alonso, 1941), ni de catalogar elementos constitutivos léxicos (Alvar *et al.,* 1967). En nuestro caso, se trata de estratos que representan diversos niveles de distanciamiento entre áreas de acuerdo con el nivel de precisión estadística que se maneje.

A partir de los análisis comentados, hemos planteado para este trabajo unos objetivos generales y unas preguntas de investigación concretas. Como primer objetivo general, destacamos el análisis de la variación léxica entre los países hispanohablantes, que había quedado como una de las tareas pendientes en el análisis anterior (Moreno Fernández y Ueda, 2018). Como segundo objetivo general, nos proponemos analizar las distancias lingüísticas de base léxica que existen entre varios países. En cuanto a las preguntas de investigación, proponemos las siguientes:

1. ¿Cuáles son las principales zonas hispánicas por sus usos léxicos?
2. ¿Cómo se manifiesta la distancia léxica cuando se establecen niveles o estratos de diferenciación o particularidad?
3. ¿Cuál es la relación existente entre las distancias lingüísticas y las distancias geográficas según los usos léxicos de los países hispanohablantes?
4. Dada la personalidad de los países que destacaban en los análisis previos, ¿cuál es la distancia léxica de España, Chile, Argentina, Costa Rica y México respecto al resto de los países hispanohablantes?

Las respuestas a estas preguntas se enfrentan a algunas de las limitaciones ya encontradas en análisis estadísticos anteriores. Por un lado, el análisis se realiza sobre la base de países, no de regiones ni de espacios transnacionales, lo que ofrece una imagen condicionada de los espacios geolingüísticos. Por otro lado, nuestro análisis se centra exclusivamente en el nivel léxico, no atiende a todas las esferas semánticas posibles ni distingue entre respuestas primeras o segundas respuestas aportadas por los informantes.

4. METODOLOGÍA

La base de este análisis está formada por una selección de conceptos (ítems) incluidos en el corpus *Varilex-R*. Los datos se recopilaron en los 21 países hispanohablantes entre los años 1993 y 2005 a través de encuestas. En 2015, se procedió a la revisión de los materiales lingüísticos obtenidos por parte de investigadores de todos los países hispanohablantes estudiados (Ueda y Moreno Fernández, 2016). Los datos encuestados y los revisados se pueden consultar en la siguiente página web: https://h-ueda.sakura.ne.jp/varilex-r/.

La selección de los conceptos de *Varilex-R* que finalmente han sido objeto de análisis dialectométrico incluyó de forma exclusiva a ítems o preguntas que implicaban variación léxica, por lo que se dejaron a un lado conceptos de naturaleza gramatical, fonética o estilística. A modo de ejemplo, este último criterio supone dejar fuera cuestiones como la referida al concepto de «mujer atractiva» (D129): en este caso concreto, las respuestas obtenidas no reflejaban realmente diferencias geolingüísticas en su mayor parte, sino diferencias de registro o estilo (más formal: «atractiva») o más informal o coloquial («bombón», «pastelito»). Asimismo, nuestro análisis ha neutralizado las diferencias fonéticas y morfológicas de algunas respuestas, por no implicar variación léxica propiamente dicha. Por ejemplo, no se consideraron variantes distintas «jersey» y «jersey [yérsi]» en «jersey» ((R) A004 [SWEATER]), «platito» y «platillo» en «platito» ((R) A058 [COASTER]) o «mahón» y «mahones» en «mahón» ((R) A008 [JEANS]). Como segundo criterio de selección de los conceptos que implicaban variación léxica, se decidió que los conceptos analizados mostraran datos o respuestas en los 21 países analizados, para evitar una distorsión de las distancias generales basadas en distancias regionales.

La selección final arrojó una cifra de 196 conceptos o ítems relativos principalmente al hogar, la vestimenta, la automoción, las acciones y las emociones humanas. El número total de variantes léxicas únicas ha sido de 1725, que se refleja en un conjunto de 9.144 datos léxicos.

CUADRO 1. Ámbitos semánticos analizados y número de conceptos por ámbito

Hogar: 57
Vestimenta: 31
Vehículos – tipos: 13
Vehículos – partes: 15
Acciones: 33
Emociones: 3
Instalaciones / edificios: 4
Comida: 5
Negocio: 4
Calles / vías: 5
Televisión / Radio: 5
Tiempo libre: 9
Cuerpo humano: 3
Características / profesiones de una persona: 6
Tiempo/clima: 3

El número total de variantes léxicas obtenidas en cada uno de los países analizados se muestra en el cuadro 2. En él se observa que el número de respuestas reunidas desde Cuba es de 702, frente a las 294 que se reunieron en Colombia. La razón de estas diferencias no debe atribuirse sin más a una mayor riqueza léxica de unos países sobre otros, dado que, en la recogida de materiales, pudieron influir factores que quedan fuera de nuestra actual capacidad de análisis.

CUADRO 2. Número de respuestas obtenidas por países en los 196 ítems analizados

CU	PR	MX	NI	AR	CH	VE	PN	ES	RD	PA	PE	EL	HO	EC	UR	CR	BO	GU	GE	CO
702	649	625	606	558	458	458	455	447	446	432	389	373	356	339	334	335	323	314	305	294

Una vez realizado un recuento de los conceptos, datos y variantes considerados para el análisis, procedimos a crear los mapas de distanciamiento léxico. Para ello, recurrimos al programa *Gabmap* (Leinonen, Çöltekin, y Nerbonne, 2016; Nerbonne, Colen, Gooskens, Kleiweg, y Leinonen, 2011), aplicación web de acceso libre y código abierto. Aparte del material lingüístico obtenido de los 196 conceptos de *Varilex-R* seleccionados, el programa requirió de un archivo con un mapa base para poder ejecutar los cálculos de las distancias lingüísticas entre los puntos de encuesta, a partir

de las cuales se crearon los gráficos y posteriormente los mapas dialectométricos. El archivo del mapa base fue elaborado a través del programa *Google Earth Pro* (Versión: 7.3.6.9345 del 29 de diciembre de 2022).

Gabmap es una aplicación destinada especialmente al analisis dialectométrico de diversos niveles lingüísticos, como la fonética, la sintaxis o el léxico (Leinonen, Çöltekin, y Nerbonne, 2016, p. 71). La aplicación mide los siguientes tipos de diferencias:

- Diferencias categóricas: variantes léxicas de un solo concepto o afijo.
- Diferencias numéricas: frecuencias de formantes de vocales.
- Diferencias basadas en cadenas: transcripciones fonéticas.

Gabmap permite realizar exploraciones y cálculos para estudiar el material lingüístico disponible desde diferentes puntos de vista: por ejemplo, diferentes medidas de distancia. En este estudio, hemos optado por analizar las diferencias categóricas a través del valor de identidad ponderado (Goebl, 2010), que se consigue mediante una fórmula que discrimina las características lingüísticas raras o esporádicas de las que son más importantes, concediéndoles un mayor peso.

Por otro lado, *Gabmap* ofrece la posibilidad de contar con varias pruebas estadísticas con el fin de evaluar la validez del estudio. Entre ellas, merecen mencionarse la incoherencia local y el alfa de Cronbach. La incoherencia local es una medida que representa un valor numérico de estrés local, asignada a un conjunto de diferencias entre elementos, relacionadas con las distancias geográficas. En general, los valores más cercanos a 0 sugieren una mayor coherencia, mientras que los valores positivos son menos coherentes. En nuestro estudio, la incoherencia local tiene un valor de 0,66, lo que indica que las relaciones entre algunos puntos de encuesta presentan proximidad geográfica, pero distanciamiento léxico. El alfa de Cronbach es un coeficiente de fiabilidad que en *Gabmap* se usa para las mediciones de diferencias entre datos, dependiendo de la calidad del material lingüístico analizado y del método utilizado. En general, los valores superiores a 0,7 se pueden considerar aceptables. Nuestro análisis presenta un valor 0,94, que puede considerarse alto.

Asimismo, la aplicación *Gabmap* ofrece la posibilidad de crear mapas tomando uno de los puntos (países) como referencia, así como el cartografiado a partir de los resultados aportados por una prueba de escalamiento multidimensional y por el análisis de *clusters* jerárquicos. El escalamiento multidimensional es una herramienta estadística que representa las similitudes entre objetos dentro de un espacio geográfico (Schiffman, Reynolds y Young, 1981). En la aplicación *Gabmap*, el escalamiento multidimensional usa una matriz de distancia (*full sites x sites distance matrix*) como *input*, a partir de la cual se genera una representación de un espacio n-dimensional donde las distancias se aproximan a las distancias lingüísticas originales. *Gabmap* traza los resultados de estos cálculos en un sistema de coordenadas cartesianas de dos

dimensiones, donde los puntos de encuesta de datos similares o con rasgos comunes se representan más cercanos entre sí. En cuanto a los tipos de *clusters* jerárquicos disponibles en *Gabmap*, estos se basan en distintos métodos y algoritmos con el fin de agrupar áreas de dialectos diferentes y delimitar fronteras entre ellos. El análisis de *clusters* practicado en este estudio sigue el criterio de la media o promedio ponderado. Este método calcula la distancia entre dos grupos a partir de la media de las distancias entre todos pares de elementos, donde cada par está formado por un elemento de cada grupo, y subraya el peso de las características que aparecen esporádicamente (Everitt, Landau, Leese y Stahl, 2011). Su algoritmo es el siguiente (*Gabmap,* s. f.):

$$d_{k[ij]} = \left(\frac{1}{2} \times d_{ki}\right) + \left(\frac{1}{2} \times d_{kj}\right)$$

Uno de los fines de la dialectometría es trazar y delimitar áreas de variedades dialectales con el fin de examinarlas y compararlas entre sí. En este sentido, el escalamiento multidimensional es un método muy útil para detectar *clusters* y descubrir continuos dialectales. Sin embargo, esta técnica no agrupa los puntos geográficos analizados en grupos de dialectos, como suelen hacer los métodos de *clusters* jerárquicos, sino que indica similitudes entre los puntos estudiados (Nerbonne, Kleinweg, Manni, y Heeringa, 2008).

Los resultados dialectométricos de los diferentes métodos aplicados los presentamos en mapas y gráficos. El mapa 1, que incluye los 21 países estudiados, ha servido como base para representar las distancias analizadas.

MAPA 1. Mapa de los países estudiados elaborado con *Gabmap*.

5. ANÁLISIS DIALECTOMÉTRICO

Los análisis estadísticos cuya metodología acaba de describirse han buscado respuestas a nuestras preguntas de investigación iniciales, con los resultados que se exponen a continuación. En lo que se refiere a la zonificación léxica, el análisis de escalamiento multidimensional nos ha proporcionado un gráfico en el que se aprecia con relativa claridad la formación o agrupamiento de varios conjuntos de países, indicador de su menor distanciamiento léxico.

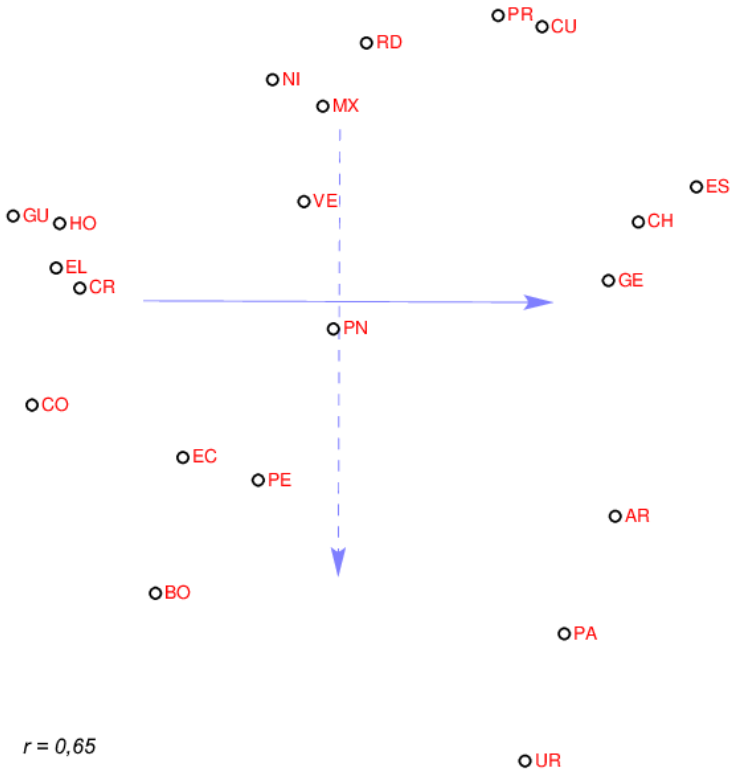
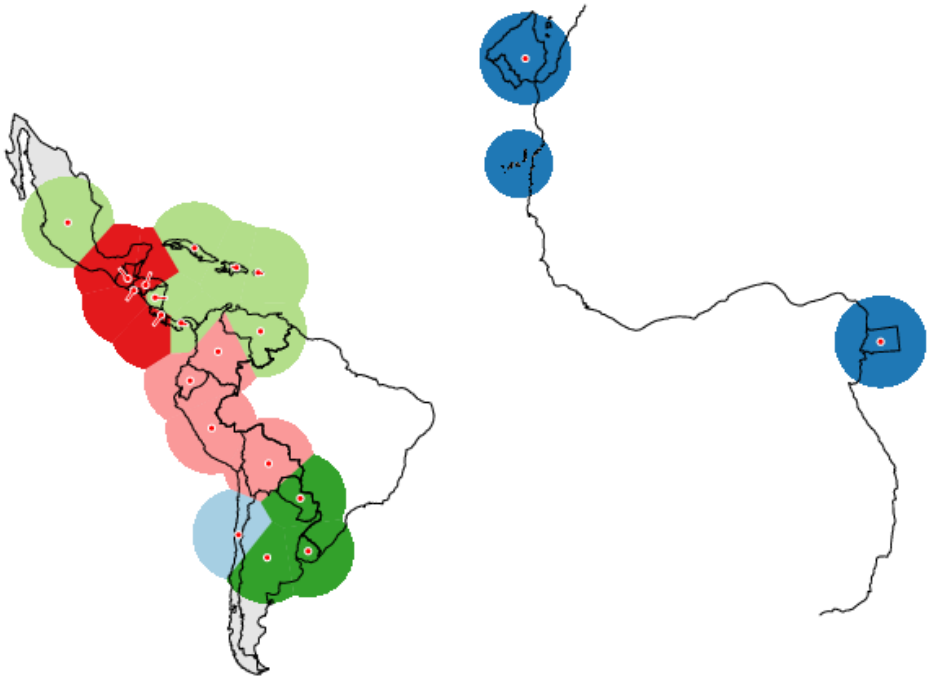


GRÁFICO 2. Resultados del escalamiento multidimensional de dos dimensiones a partir de datos léxicos de *Varilex-R.*

El gráfico muestra agrupamientos como el de los países del Río de la Plata (AR, PA, UR), el de América Central (GU, EL, HO, CR), con Nicaragua y Panamá algo más distanciados, por aproximarse en el eje de abscisas hacia los países caribeños (CU, RD, PR), que a su vez constituyen otro agrupamiento. Se agrupan igualmente, por su menor distancia léxica, los países andinos (EC, PE, BO, CO). México y Venezuela, por otra parte, muestran una cercanía al espacio de los países caribeños. Por último, el gráfico muestra en un extremo a España, Chile y Guinea Ecuatorial, tanto por su

particularidad, como por la distancia léxica que mantienen respecto de los demás países. La cercanía de España y Guinea Ecuatorial responde a su grado de coincidencia en los usos léxicos; la de Chile marca una distancia respecto de los demás países americanos, que se interpreta como equivalente a la distancia que mantiene España y no tanto como coincidencia de los usos léxicos entre la propia España y Chile. En este aspecto, la posición de estos dos países aparece muy bien marcada en coincidencia con los resultados del análisis practicado por Moreno Fernández y Ueda en 2018 sobre materiales de varios niveles lingüísticos, no solo léxicos (v. Cuadro 1).

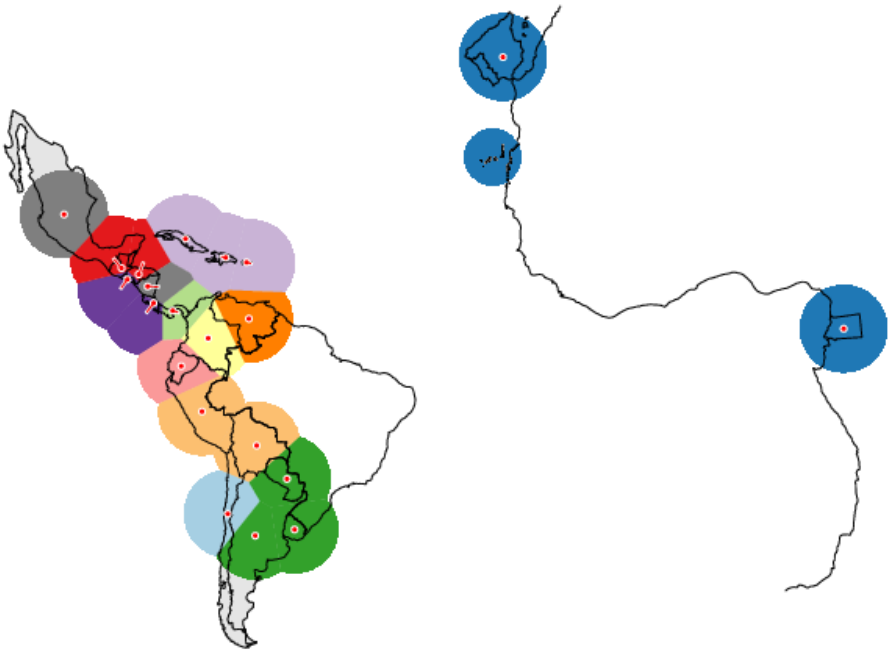
Seguidamente, a partir de los datos de *Varilex-R*, practicamos un análisis de *clusters,* basado en correlaciones, cuyos resultados, obtenidos mediante el método de la media ponderada y la formación de 6 *clusters,* queda representado en un mapa que dibuja las seis áreas léxicas más destacadas del espacio hispanohablante.



MAPA 2. Mapa de áreas léxicas a partir de análisis de *clusters* (media ponderada; 6 *clusters*).

Las áreas léxicas representadas en el mapa son las siguientes:
España y Guinea Ecuatorial
México y Caribe, incluido Venezuela, Nicaragua y Panamá
América Central, excluido Nicaragua y Panamá
Área andina
Área rioplatense
Chile

Ahora bien, cuando abordamos un estrato de distanciamiento más profundo, mediante un análisis estadístico del doble de *clusters* (12), los resultados muestran un mapa diferente.



MAPA 3. Mapa de áreas léxicas a partir de análisis de *clusters* (media ponderada; 12 *clusters*).

El estrato de distanciamiento de 12 *clusters* vuelve a mostrar la cohesión de algunas de las áreas identificadas en el estrato de 6 *clusters*: España y Guinea Ecuatorial; Chile; los países del área rioplatense; el Caribe insular; buena parte del área andina. Sin embargo, este estrato, más profundo, nos revela unas distancias léxicas no detectadas en el estrato de 6 *clusters*.

a) la especificidad de México (esto es, su distanciamiento relativo del Caribe);
b) la complejidad del territorio centroamericano, dentro del cual se distinguen hasta cuatro áreas léxicas;
c) la relativa distancia de Venezuela respecto de las Antillas hispanas;
d) la distancia de Colombia, por un lado, y de Ecuador, por otro, respecto de Perú y Bolivia

Sin poner en duda la significación de estos resultados, no perdemos de vista que un análisis más profundo, con un mayor número de *clusters*, aportaría nuevos matices y distancias entre territorios, llevando a nuevos estratos de distanciamiento.

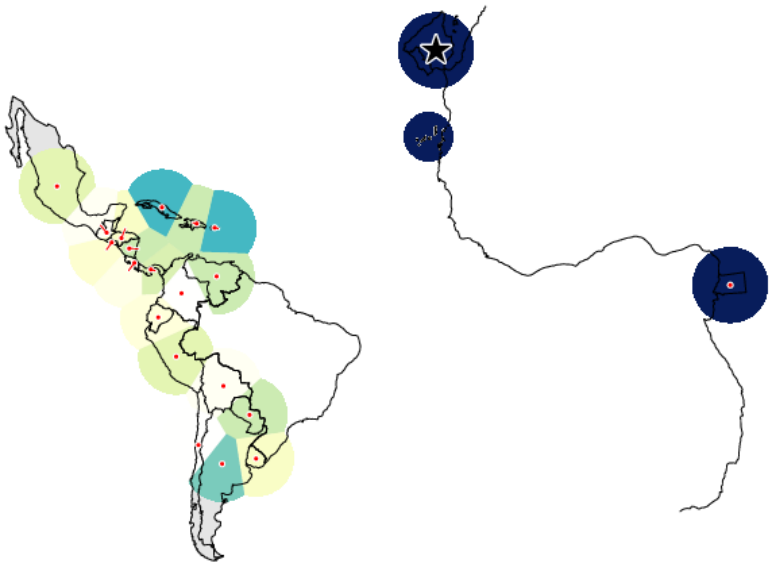
Por otra parte, dado que el estudio de Moreno Fernández y Ueda (2018) dibujaba perfiles diferenciados, en cuanto al grado de cohesión y de particularidad, de países como España, Chile, Argentina o Costa Rica, hemos abundado en el análisis de las distancias léxicas que cada uno de estos países revela respecto de todos los demás. Estos análisis, que denominamos «referenciales», toman cada uno de los países mencionados como referencia para proceder a la medición de las distancias y para observar la relación existente entre las distancias léxicas calculadas y las distancias geográficas. A los países mencionados, hemos añadido México para nuestro análisis, dada su entidad demográfica, su posición estratégica en el proceso de difusión del español por toda la geografía americana y su particularidad de acuerdo con los resultados del análisis en 12 *clusters*.

Los análisis referenciales que hemos practicado aportan dos tipos de información. En primer lugar, la aplicación *Gabmap* ofrece un mapa en el que se marcan las distancias relativas entre cada país de referencia y el resto de los países analizados. En segundo lugar, la aplicación provee un gráfico en el que se correlaciona la distancia geográfica entre el país de referencia y los demás (eje de abscisas) y la distancia lingüística entre los usos léxicos del país de referencia y los correspondientes a cada uno de los demás países. En los mapas, los tonos más oscuros indican mayor afinidad léxica. La interpretación de resultados nos indicará cuáles son las áreas más afines por sus usos léxicos al punto de referencia, así como si la correlación entre distancia geográfica y distancia lingüística es lineal o existe alguna quiebra en ella; es decir, si hay países con los que cada punto de referencia establece una relación particular.

5.1. *España*

Moreno Fernández y Ueda (2018) señalaban en sus análisis que España era un país con una diferenciación marcada respecto de los demás del territorio hispánico, si bien, al mismo tiempo, mostraba un alto índice de comunalidad con la mayoría de los países, al tener en cuenta todos los datos, obtenidos como primeras respuestas o como alternativas.

El mapa de nuestro análisis referencial muestra mediante colores y tonos diferentes cómo España presenta unas distancias marcadas, aunque similares (tonos claros), respecto de la mayoría de los países, si bien algo menor en cuanto a Argentina, Cuba y Puerto Rico. La correlación de distancias sitúa a Guinea Ecuatorial como un país cercano a España en el léxico y en la geografía, mientras que Chile aparece en el extremo opuesto por su lejanía geográfica y lingüística. Resulta interesante comprobar cómo son los países andinos y parte de América Central los que quiebran la correlación, al mostrarse más lejanos en lo léxico que en lo geográfico. Asimismo, se aprecia que la mayor parte de los países muestran unos índices de distanciamiento lingüístico que se mueven entre 0,75 y 0,85, frente al 0,60 de Guinea Ecuatorial; esto es, unos índices relevantes en cuanto a la distancia.



MAPA 4. Mapa con España como punto de referencia.

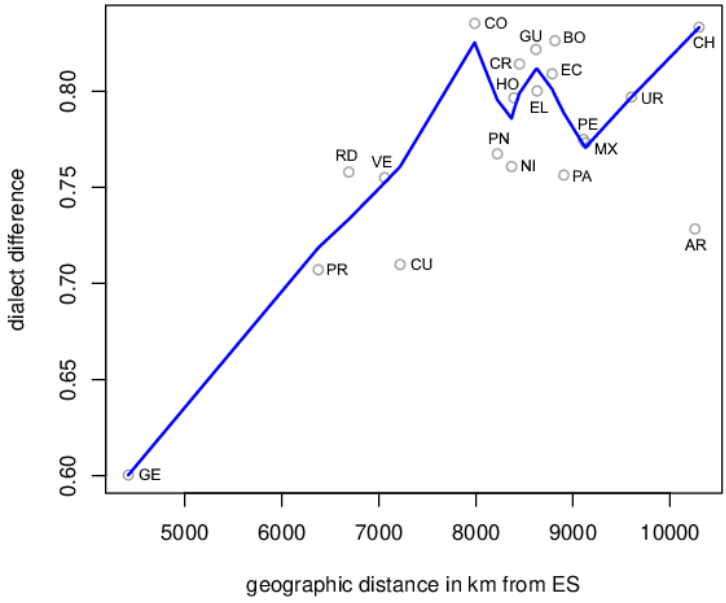
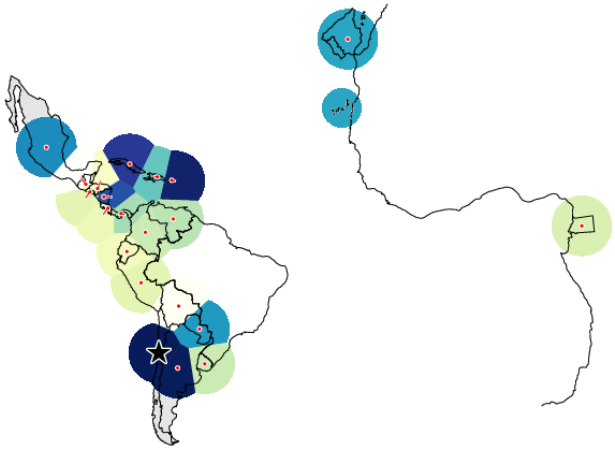


GRÁFICO 3. Gráfico de distancia geográfica y distancia lingüística con España como punto de referencia.

5.2. *Chile*

Los análisis de 2018 mostraban que el español chileno se caracterizaba por su alto índice de particularidad. Esto haría suponer que los usos analizados distan de los habituales en otros países. Y, efectivamente, el mapa de nuestro análisis referencial nos destaca las distancias respecto de los demás países, que aparecen menos intensas en relación con El Caribe y parte de Centroamérica, España, México y Paraguay.



MAPA 5. Mapa con Chile como punto de referencia.

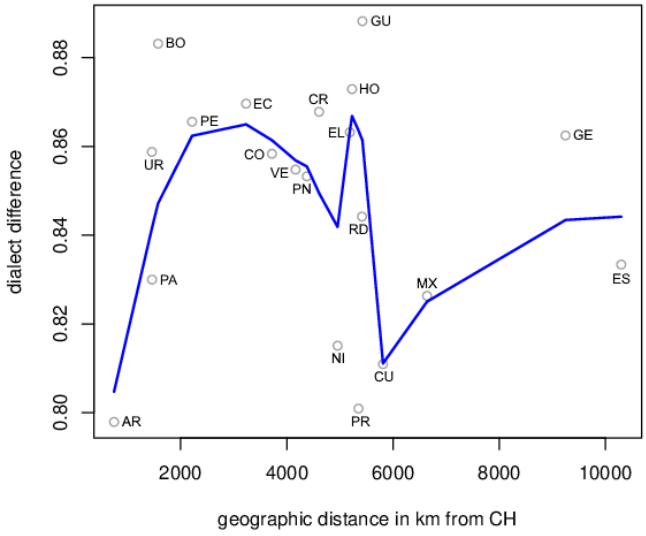
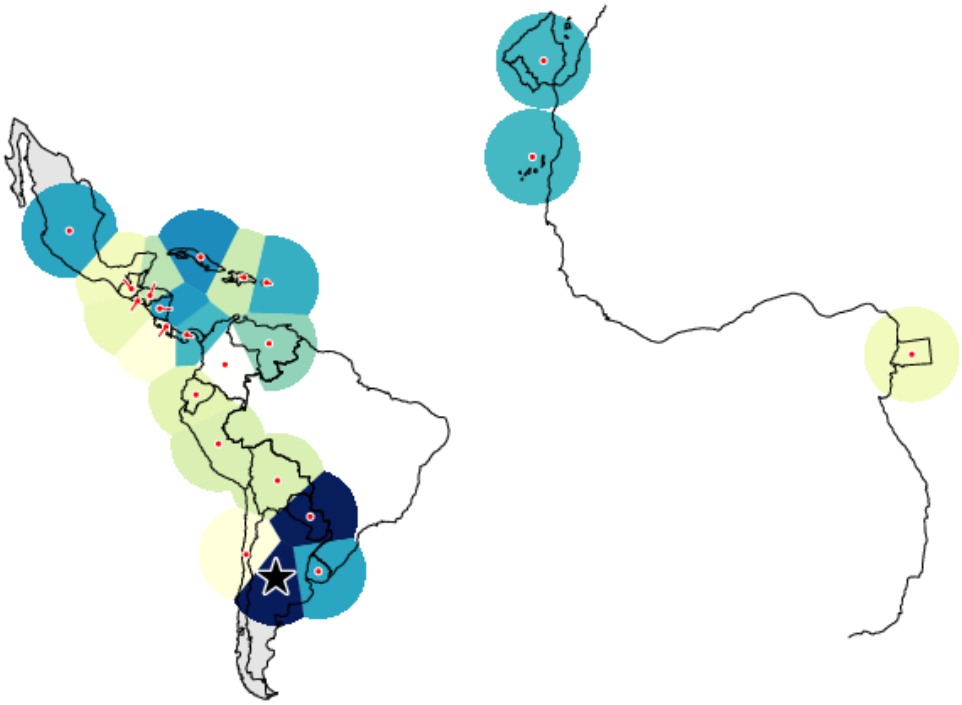


GRÁFICO 4. Gráfico de distancia geográfica y distancia lingüística con Chile como punto de referencia.

Las particularidades léxicas de Chile muestran correlaciones muy interesantes en la curva de distancias. Efectivamente, la distancia respecto de España es clara en la geografía; el léxico, en cambio, revela que más de una docena de países muestran mayor distanciamiento que España. Entre los más distanciados se encuentran varios países centroamericanos (Honduras, Guatemala, Costa Rica, aunque no así Nicaragua). Junto a este hecho, llaman también la atención dos datos: por un lado, la distancia léxica respecto de los países andinos, a pesar de la cercanía geográfica; por otro, la cercanía lingüística de Cuba y Puerto Rico, a pesar de la distancia geográfica que los separa. Con todo, los índices de distancia lingüística oscilan entre 0,80 (Argentina) y 0,90 (Guatemala), que son relativamente altos para el conjunto de los países.

5.3. *Argentina*

La particularidad del español rioplatense es bien conocida y está bien analizada en numerosos estudios de lingüística histórica, de dialectología y de lingüística perceptiva (Moreno Fernández, 2000). Entre las hablas rioplatenses, las argentinas ocupan un lugar relevante tanto por razones demográficas, como por motivos sociopolíticos.



MAPA 6. Mapa con Argentina como punto de referencia.

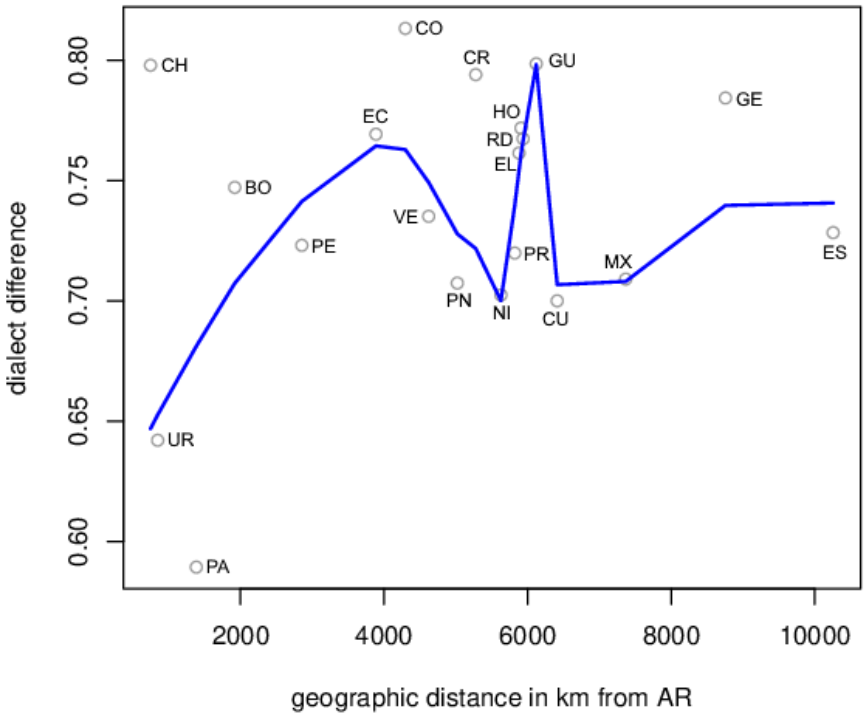
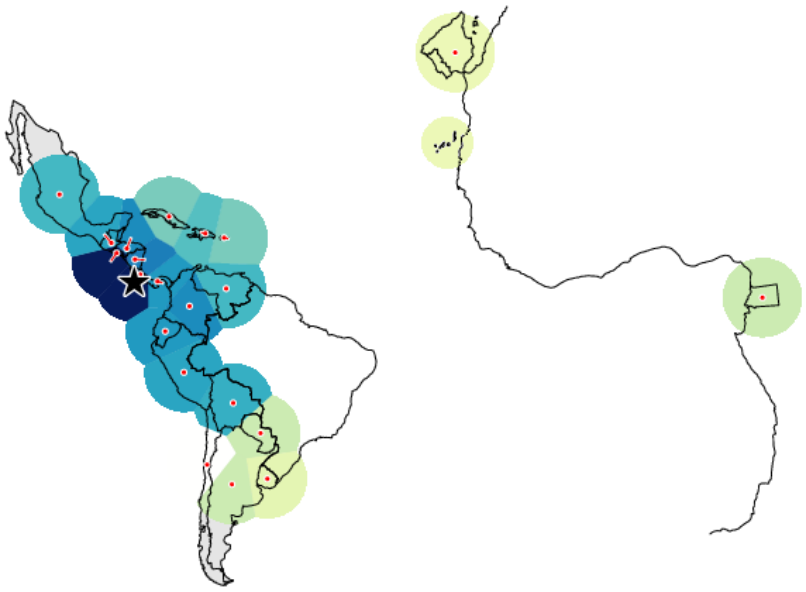


GRÁFICO 5. Gráfico de distancia geográfica y distancia lingüística con Argentina como punto de referencia.

El mapa que nos proporciona el análisis referencial de Argentina muestra a este país con mayores afinidades con Paraguay y Uruguay, así como con Cuba y Nicaragua. El gráfico de la correlación de distancias, por su lado, nos ayuda a apreciar la mayor distancia lingüística respecto a Guatemala, que no se correspondería con la distancia que los separa en la geografía, pero también la distancia respecto de los países andinos, que podría ser menor, dada la relativa cercanía geográfica. Llama la atención el caso de Chile, muy cercano en la geografía, pero bien distante en el léxico. El rango de los índices de distancias oscila entre 0,55 y 0,85, lo que significa que las diferencias en el distanciamiento de Argentina respecto a los demás países son amplias, si bien la mayor parte de los países se alinea entre 0,70 y 0,80.

5.4. *Costa Rica*

El mapa de distanciamiento de Costa Rica muestra la proximidad del léxico costarricense, no solo respecto a otros países de la América Central, sino también en relación con México y el área andina.



MAPA 7. Mapa con Costa Rica como punto de referencia.

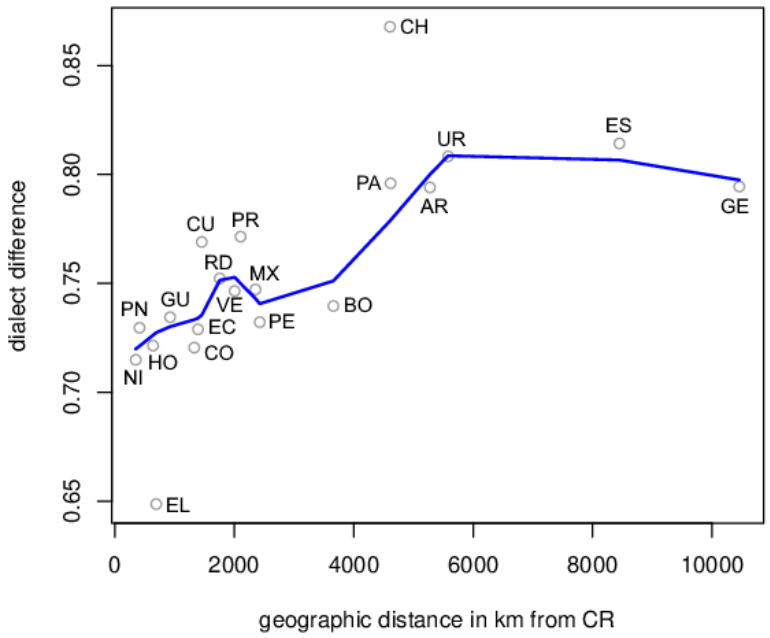
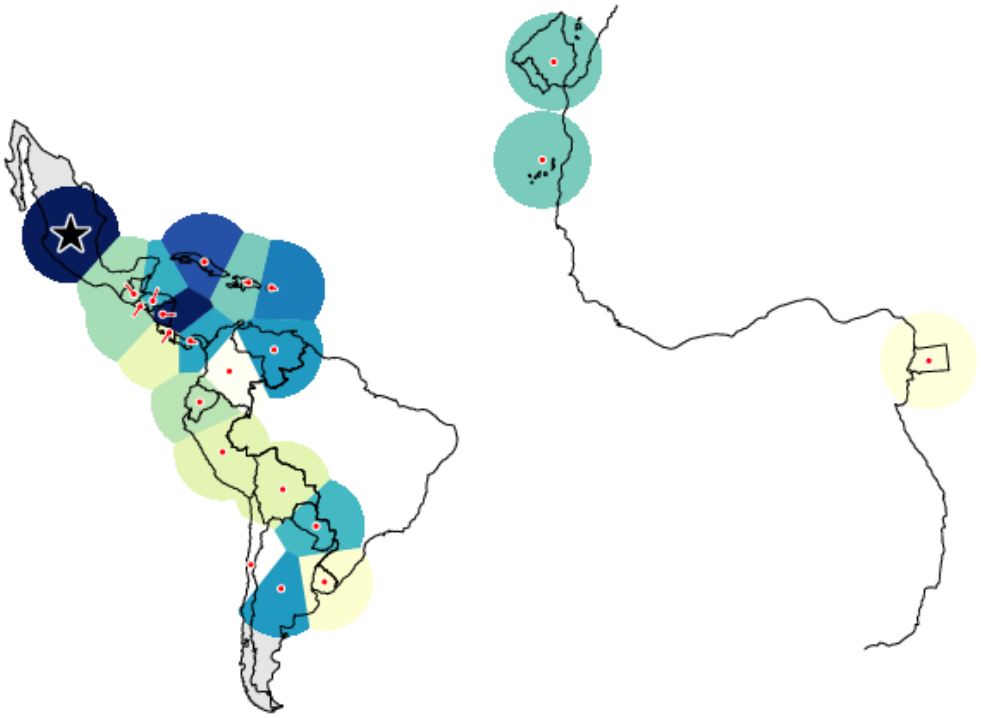


GRÁFICO 6. Gráfico de distancia geográfica y distancia lingüística con Costa Rica como punto de referencia.

A una mayor distancia de Costa Rica se hallan, en menor grado, las Antillas y, de forma más marcada, España y Guinea Ecuatorial, junto a los países del Cono Sur: Uruguay, Paraguay, Argentina y, particularmente, Chile. No obstante, con excepción de Chile, la mayor parte de los países se agrupan en torno en unos índices entre 0,70 y 075, con los casos del área rioplatense, España y Guinea Ecuatorial que apenas superan el índice 0,80. Costa Rica es, pues, un país de distanciamiento relativamente menor respecto a un gran número de países hispanohablantes.

5.5. *México*

El análisis del léxico mexicano resulta interesante por su relativa cercanía al del caribeño y su relativa lejanía respecto al centroamericano, con la excepción de Costa Rica. De hecho, el mapa muestra una mayor afinidad de México con Argentina que con Guatemala.



MAPA 8. Mapa con México como punto de referencia.

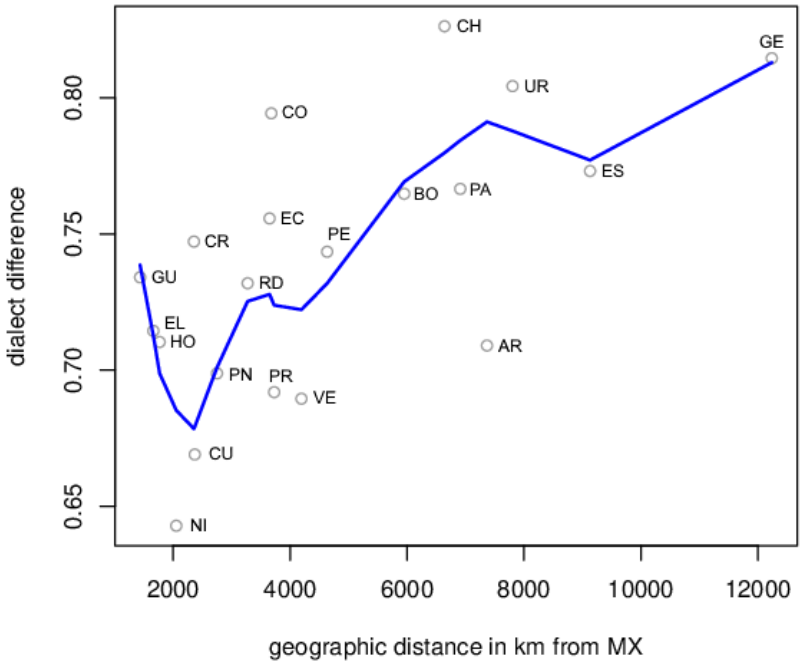


GRÁFICO 7. Gráfico de distancia geográfica y distancia lingüística con México como punto de referencia.

La correlación de distancias revela una curva en la que las distancias geográficas y léxicas se correlacionan con relativa claridad, con la salvedad de la mayor distancia lingüística que geográfica respecto a Guatemala, El Salvador, Honduras, Costa Rica y Colombia. Puede observarse, además, que el rango de distancias es bastante amplio, ya que oscila entre los índices 0,60 y 0,85. Existe, pues, una mayor dispersión de distanciamientos que en el caso de Costa Rica, por ejemplo.

6. DISCUSIÓN DE LOS RESULTADOS

Los intentos de zonificación del español cuentan con una larga tradición en los estudios dialectales (Henríquez Ureña, 1921; Rona, 1964; Zamora Munné, 1979-1980; Cahuzac, 1980; Ueda, 1985, 2007, 2023; Alba, 1992; Moreno Fernández, 1993). Las áreas léxicas derivadas del análisis de seis *clusters* que aquí se ha expuesto vienen a coincidir con la propuesta presentada por Moreno Fernández (2000), para la que se tuvieron en cuenta, entre otras causas, las percepciones predominantes en el conjunto de la comunidad hispánica y otros factores lingüísticos como la pronunciación y la sintaxis, además del léxico (Moreno Fernández, 2015). Esto significa que los territorios de

España, el Caribe, México, los Andes, el Río de la Plata y los Andes, revelan en el léxico una personalidad que también se ha identificado desde otras perspectivas.

Los análisis de *clusters* y los análisis referenciales que se han realizado revelan, sin embargo, otros hechos interesantes. Uno de ellos es la complejidad del territorio de la América Central que, si bien ha mantenido una relación histórica con México, especialmente el norte centroamericano, se aparta del léxico mexicano de forma ostensible, probablemente por la influencia de las hablas amerindias y por una fragmentada geografía e historia del istmo americano (Quesada Pacheco, 2010, 2013). Asimismo, resulta interesante comprobar la intensa afinidad del léxico español y el ecuatoguineano, marcado por un pasado colonial reciente y por la presencia continua de españoles en sectores estratégicos del país africano, como el religioso, el educativo y, en general, el cultural. Otro aspecto significativo que revela el análisis del distanciamiento léxico es la proporcional cercanía del léxico mexicano respecto al Caribe, debida probablemente a la posición de México en el proceso de expansión del español en América entre los siglos XVI y XVII principalmente, donde fue determinante la estrecha conexión entre La Habana con Veracruz y Ciudad de México (Lara, 2008).

Esta propuesta de zonificación léxica del español mediante un análisis dialectométrico nos ha permitido observar la existencia de estratos que revelan un nivel progresivo de distanciamiento en la particularidad de los territorios. Tal hecho nos indica que los niveles de estratificación léxica son múltiples y que pueden llevarse no solamente al nivel de los países, sino también al de las regiones dentro de cada uno de ellos, como fruto de una diversidad y un nivel de variación que entrecruza lo geolingüístico, con lo sociolingüístico y lo estilístico.

El hecho de que los análisis practicados no tengan en consideración factores sociales ni situacionales supone una de sus grandes limitaciones para un conocimiento holístico de la realidad léxica. Además, a tales limitaciones habría que sumar las derivadas de prescindir de la atención a espacios diferentes de los países y de atender exclusivamente a un número limitado de esferas léxico-semánticas. Con todo, estas técnicas simplemente buscan poner de relieve una visión de la diversidad desde un enfoque de lejanía, de acuerdo con la teoría de los focos (Moreno Fernández, 2023).

La historia del léxico hispánico está condicionada por una multiplicidad de factores que resultan complicados de abordar de un modo conjunto. El estudio de los estratos y de las distancias léxicas, no obstante, es capaz de revelarnos, no solamente las afinidades entre distintas áreas, sino cómo el léxico de cada país presenta elementos de una historia léxica compartida, junto a elementos compartidos con otras áreas parciales y junto a elementos propios de cada país derivados de su propia historia. Esta es la realidad que subyace a las repetidas ideas de la unidad y la diversidad del español (Alvar, 1969; Moreno de Alba, 1978; Lapesa, 1980; Rabanales, 1998).

7. CONCLUSIONES

Este estudio dialectométrico ha tenido como objetivo principal el análisis de la variación léxica en 21 países hispanohablantes, identificando las principales zonas léxicas, analizando la distancia léxica en niveles o estratos de diferenciación e identificando la relación entre las distancias léxicas y las distancias geográficas del espacio analizado. Además, nos hemos adentrado en el análisis de las distancias léxicas de varios países que, por diferentes factores, resultan particularmente interesantes: España, México, Chile, Argentina y Costa Rica. Las técnicas cuantitativas de la dialectometría han demostrado ser una herramienta eficaz para la medición de diferencias léxicas entre distintos lugares, a partir del manejo de una gran cantidad de datos para medir las distancias existentes entre los puntos geográficos considerados; en nuestro caso, los países hispanohablantes.

La base de datos *Varilex-R* proporciona un total de 981 conceptos en 21 países hispanohablantes. El análisis dialectométrico que aquí se ofrece ha trabajado con unidades de variación léxica a partir de 196 de los conceptos que integran *Varilex*, para lo cual se han desestimado conceptos que implicaban diferencias fonéticas, gramaticales y estilísticas y aquellos que no ofrecían respuestas en todos los países estudiados. Sobre este corpus de datos, hemos recurrido a la aplicación *Gabmap* para los análisis cuantitativos dialectométricos. Con los materiales obtenidos de esos 196 conceptos de *Varilex-R* y un archivo que incluía un mapa base, hemos podido calcular y representar las distancias lingüísticas entre países. El repertorio de posibilidades técnicas que el programa ofrece ha hecho posible la aplicación de diversas técnicas estadísticas y la elaboración de mapas y gráficos, siguiendo diferentes métodos y algoritmos con el objetivo de agrupar áreas dialectales y delimitar fronteras lingüísticas. Los métodos priorizados han sido el escalamiento multidimensional y el análisis de *clusters* jerárquicos.

Los resultados de las pruebas estadísticas son interesantes desde diferentes perspectivas. Así, los análisis revelan de forma clara la complejidad léxica de América Central, la afinidad de España y Guinea Ecuatorial o la cercanía léxica de México respecto al Caribe. Asimismo, hemos podido constatar la existencia de estratos de distanciamiento con diferentes niveles de diferenciación léxica, niveles cuya existencia también podría constatarse posteriormente en nivel transnacional y en un nivel regional dentro de cada país.

Nuestro estudio dialectométrico no es más que una pieza de un complejo analítico más amplio que en parte se ha trabajado en el pasado, pero que se habrá de trabajar en el futuro implicando espacios de diversa dimensión, así como niveles lingüísticos que aquí no han podido tenerse en cuenta.

REFERENCIAS BIBLIOGRÁFICAS

Alba, Orlando (1992). Zonificación dialectal del español de América. En C. Hernández (Coord.), *Historia y presente del español de América* (pp. 63-84). Junta de Castilla y León.

Alonso, Amado (1941). Substratum y superstratum. *Revista de Filología Hispánica, III*, 209-217.

Alvar, Manuel (1969). *Variedad y unidad del español*. Prensa Española.

Alvar, Manuel, Badía, Antoni, Balbín, Rafael de y Lindley Cintra, Luis F. (Eds.). (1967). *Enciclopedia lingüística hispánica* (Tomo 2). CSIC.

Cahuzac, Philippe (1980). La división del español de América en zonas dialectales. Solución etnolingüística y semántico-dialectal. *Lingüística Española Actual, II*, 385-461.

Everitt, Brian S., Landau, Sabine, Leese, Morven, y Stahl, Daniel (2011). *Cluster analysis*. Wiley.

Gabmap. (s.f.). *Manual to GABMAP – dialect analysis – Cluster analysis*. https://gabmap.nl/doc/manual/clustering.html [12/1/24].

Goebl, Hans (1982). Ansätze zu einer computativen Dialektometrie. En W. Besch *et al*. (Eds.), *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung, 1* (pp. 778-792). Walter de Gruyter.

Goebl, Hans (2010). Introducción a los problemas y métodos según los principios de la Escuela Dialectométrica de Salzburgo (con ejemplos basados del *Atlante Italo Svizzero*, AIS). En G. Aurrekoetxea Olabarri y J. L. Ormaetxea Lasaga (Eds.), *Tools for linguistic variation* (pp. 3-39). Universidad del País Vasco.

Guiter, Henri (1973). Atlas et frontières linguistiques. En G. Straka y P. Gardette (Eds.), *Les dialectes de France à la lumière des atlas régionaux* (Colloque de Strasbourg, 1971) (pp. 61-109). CNRS.

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

Lapesa, Rafael (1980). América y la unidad de la lengua española. *Documentos Lingüísticos y Literarios, 5*, 74-89.

Lara, Luis Fernando (2008). Para la historia de la expansión del español por México. *Nueva Revista de Filología Hispánica, 56*(2), 297-362.

Leinonen, Therese, Çöltekin, Çöltekin y Nerbonne, John (2016). Using Gabmap. *Lingua, 178*, 71-83.

Moreno de Alba, José G. (1978). *Unidad y variedad del español en América*. UNAM.

Moreno Fernández, Francisco (1993). *La división dialectal del español de América*. Universidad de Alcalá.

Moreno Fernández, Francisco (2000). *Qué español enseñar*. Arco/Libros.

Moreno Fernández, Francisco (2015). La percepción global de la similitud entre variedades de la lengua española. En K. Jeppesen Kragh y J. Lindschouw (Eds.), *Les variations diasystématiques et leurs interdépendances dans les langues romanes* (pp. 217-238). Éditions de linguistique et de philologie.

Moreno Fernández, Francisco (2023). Distancias reales y ficticias en los espacios lingüísticos. *Energeia, VIII*, 82-103.

Moreno Fernández, Francisco y Ueda, Hiroto (2018). Cohesion and particularity in the Spanish dialect continuum. *Open Linguistics, 4*(1), 722-742.

Nerbonne, John, Colen, Rinke, Gooskens, Charlotte, Kleiweg, Peter y Leinonen, Therese (2011). Gabmap – a web application for dialectology. *Dialectología, II*, 65-89.

Nerbonne, John, Kleiweg, Peter, Manni, Franz y Heeringa, Wilbert (2008). Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. En C. Preisach, H.

Burkhardt, L. Schmidt-Thieme y R. Decker (Eds.), *Data analysis, machine learning and applications. Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e. V.* (pp. 647-654). Springer.
Quesada Pacheco, Miguel Ángel (Ed.). (2010). *El español de América Central: nivel fonético*. Iberoamericana / Vervuert.
Quesada Pacheco, Miguel Ángel (Ed.). (2013). *El español de América Central: nivel morfosintáctico*. Iberoamericana / Vervuert.
Rabanales, Ambrosio (1998). Unidad y diversificación de la lengua española. *Onomazein, 3*, 133-142.
Rona, José P. (1964). El problema de la división del español americano en zonas dialectales. En *Presente y futuro de la lengua española, I* (pp. 215-226). OFINES.
Schiffman, Susan S., Reynolds, M. Lance y Young, Forrest W. (1981). *Introduction to multidimensional scaling: Theory, methods and applications*. Academic Press.
Séguy, Jean (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane, 35*, 335-357.
Ueda, Hiroto (1995). Zonificación del español del mundo. Palabras y cosas de la vida urbana. *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. En A. Ruiz Tinoco (Ed.), *Jornadas sobre métodos informáticos en el tratamiento de las lenguas ibéricas* (pp. 121-140). Centro de Estudios Hispánicos – Universidad Sofía.
Ueda, Hiroto (2023). Dialectología del español y dialectometría. En F. Moreno Fernández y R. Caravedo (Eds.), *Dialectología hispánica. The Routledge Handbook of Spanish Dialectology* (pp. 87-104). Routledge.
Ueda, Hiroto y Moreno Fernández, Francisco (2016). VARILEX-R: variación léxica del español en el mundo. http://goo.gl/BENLPL
Ueda, Hiroto y Ruiz Tinoco, Antonio (2007). Investigaciones sobre la variación léxica del español: Proyectos y resultados de 1992 a 2007. *VARILEX, 15*, 1-19.
Ueda, Hiroto (2015). http://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/. [26/12/2023].
Zamora Munné, Juan C. (1979-1980). Las zonas dialectales del español americano. *Boletín de la Academia Norteamericana de la Lengua Española, 4-5*, 57-67.
Zimmermann, Klaus (2018). Lexicografía diferencial y lexicografía integral. En M. Álvarez de la Granja y E. González Seoane (Eds.), *Léxico dialectal y lexicografía en la Iberorromania* (pp. 121-144). Iberoamericana-Vervuert.