Métodos de la dialectología cuantitativa

Gotzon Aurrekoetxea
UPV-EHU
gotzonaurre@gmail.com

Resumen

La introducción de la cuantificación de la variación geolingüística ha traído consigo un espectacular auge de las publicaciones sobre la materia, que indican una renovada vitalidad de la disciplina. Uno de los mayores avances de la dialectología del siglo pasado, la dialectometría, se ha convertido en una realidad en prácticamente todas las lenguas cultivadas (Goebl 1992; Nerbonne 2013).

La variedad de técnicas cuantitativas utilizadas en la dialectología pone al alcance de los investigadores un amplio abanico de posibilidades de analizar los datos dialectales. Pero todo análisis cuantitativo necesita de una base de datos amplia que aleja al dialectólogo de las prácticas del denominado *(single) feature based* dialectología, ganando en la objetividad de la muestra del análisis.

En este trabajo se presentan los pasos que hay que seguir para desarrollar una investigación en dialectología cuantitativa. Además, se exponen algunas de las técnicas utilizadas, como las destinadas a la cuantificación de la distancia entre variedades, a la clasificación jerárquica, y/o al análisis del *continuum* dialectal. Así mismo, también se exponen métodos multivariantes para la identificación de patrones de variación, estudio de las variables que presentan similares patrones geográficos, analizar la probabilidad de pertenencia a determinados grupos dialectales, etc. La metodología de la dialectología cuantitativa se halla delimitada por los siguientes pasos: elección de un atlas lingüístico del que se proveerá su base de datos (que puede ser fonética, ortográfica o/y etiquetada), aplicación de una medida de distancia que proporciona una matriz de distancias y el uso de técnicas cuantitativas aplicadas a la matriz de distancias. La cuantificación se ha convertido en un paso obligatorio para expertos que se dedican al estudio de la variación lingüística.

Palabras clave: variación geolingüística, dialectometría, metodología.

Abstract

The introduction of the quantification of geolinguistic variation has brought a spectacular rise in publications on the subject, which indicate a renewed vitality of the discipline. One of the greatest advances in dialectology of the last century, dialectometry, has become a reality in practically all cultivated languages (Goebl 1992; Nerbonne 2013).

The variety of quantitative techniques used in dialectometry offers researchers a wide range of possibilities for analyzing dialectical data. But any quantitative analysis needs a broad database that distances the dialectologist from the practices of the so-called '(single) feature based' dialectology, gaining in the objectivity of the analysis sample.

ISSN: 1139-8736 https://infoling.org/elies/

The methodology of quantitative dialectology begins with the choice of a linguistic atlas from which its database will be provided (which can be phonetic, orthographic or/and labeled). The application of a distance measurement provides the distance matrix. The quantitative techniques applied to the distance matrix range from the quantification of the distance between dialectal varieties (interpunctual dialectometry), the hierarchical classification of dialectal varieties, the analysis of the dialectal *continuum* (with the technique of multidimensional scaling (MDS), the analysis of the correlation between geographical and linguistic distance, the detection of linguistic characteristics, etc. Quantification has become a mandatory step for experts who study linguistic variation.

Keywords: Regional variation, dialectometry, methodology.

1. Introducción

El camino recorrido en la cuantificación de la variación geolingüística ha sido muy fructífero desde el punto de vista de los avances, de la variedad de técnicas y del gran número de idiomas en los que se ha aplicado. La dialectometría (en adelante DM), que es como se denomina en este campo a la cuantificación de la variación geolingüística, se encuentra en un momento que era inimaginable cuando se elaboraron los primeros trabajos cuantitativos. Actualmente, se puede decir que la cuantificación es imprescindible en los estudios de la variación geolingüística cuando se trata de clasificar variedades dialectales. Ello no quiere decir que la cuantificación sea el objetivo de los trabajos geolingüísticos, sino la herramienta indispensable para obtener resultados fidedignos, que posteriormente deben de ser interpretados, como en todas las ciencias. Y decimos que es indispensable porque se necesita cuantificar esa variación para trazar y jerarquizar fronteras, delimitar áreas, medir la distancia lingüística entre localidades o variedades dialectales, etc. La DM no es el fin de nada, sino el medio que nos provee de herramientas adecuadas para que el análisis tenga mayor precisión y exactitud.

El gran acierto de Jean Séguy y de los creadores de esta disciplina en la década de los 70 del siglo pasado fue su visión de la realidad. La dialectología no podía evolucionar en aquella época en el plano teórico, porque las herramientas que tenía a mano no daban para más: no se podía avanzar en la delimitación y jerarquización de las fronteras, en la distinción entre zonas de transición y frontera dialectal, en la relación entre la distancia geográfica y la distancia lingüística, o en el análisis de los factores extralingüísticos en la diferenciación geolingüística. Hacían falta nuevas herramientas para analizar toda la información recogida en los atlas lingüísticos.

Hay que recordar, sin embargo, que antes de la aparición de la DM se veía ya una inexcusable necesidad de métodos estadísticos. Ya había propuestas cuantitativas en la literatura científica anterior, algunos loables intentos, también palos de ciego. Entre los precursores de la utilización de métodos cuantitativos en la dialectología se pueden citar entre otros a Davis y MacDavid (1950), Reed y Spicer (1952), Atwood (1955) y Houck (1967). En prácticamente todos ellos se conjuga lo cualitativo (elección de características a analizar por el investigador) y la cuantificación (se pueden ver buenos resúmenes de los trabajos de esa época en Shaw 1974, Schneider 1988 o Kretzschmar 1996).

En este sentido, el abanico del análisis cuantitativo se ha expandido enormemente con la

ISSN: 1139-8736 https://infoling.org/elies/

DM. Hay una gran variedad de técnicas que van introduciéndose y abriendo camino. Incluso ya hay investigadores que dividen la DM en una DM clásica o estándar y otra nueva DM (Pickl y Rumph 2012). Muchas de las técnicas se encuentran compiladas en los programas creados *ad hoc* y otras en paquetes del mercado (Grieve et al. 2011) o en R (www. r-project.org/) (Grieve y Jurgens 2019). A decir verdad, la DM ha dado un gran salto cualitativo y sirve actualmente para todo análisis cuantitativo de la variación geolingüística. Pero no solo es avance en el plano técnico o estadístico, sino que también se ha dado un avance en la concepción de qué es lo que hay que medir. La dimensión lingüística de los dialectómetras ha ido creciendo; tanto que, hoy en día, se plantea claramente que son inviables análisis basados en isoglosas (aunque estas sean cuantitativas) o en características elegidas por el investigador. Ante ello se plantean dos direcciones: una dialectología basada en grandes bases de datos o corpus, y una dialectología basada en sistemas o subsistemas lingüísticos. Esta última dirección, nacida a mediados del siglo XX, ha ido tomando más importancia en generaciones posteriores y ha sido tenida en cuenta por la escuela dialectométrica de Salzburgo y sus seguidores con la taxación o etiquetaje de los datos. La dialectología cuantitativa actual presenta medidas tanto para el análisis de la estructura superficial (nivel fonético), como de la estructura profunda (nivel fonológico) de la lengua. En la primera se encuentra la escuela de Groninga y en la segunda la escuela de Salzburgo.

El trabajo pretende dar una visión general de la DM (los diversos pasos en la cadena dialectométrica, diferentes escuelas, herramientas, etc.) y la aportación de la misma a los estudios de la variación geolingüística. Para ello, se ha estructurado en cinco secciones: en la segunda se explica para qué sirve la dialectometría, en la tercera se especifica cómo se trabaja en ella, exponiendo todos los pasos que hay que dar en un análisis dialectométrico; en la cuarta se analiza la relevancia de estos instrumentos metodológicos y en la quinta se exponen las conclusiones.

2. ¿Para qué sirve la dialectometría?

La DM surgió para superar la crisis epistemológica en la que estaba inmersa la dialectología. En efecto, esta disciplina lingüística se encontraba paralizada, sin mecanismos y técnicas adecuadas para avanzar en el conocimiento de la variación geolingüística, mientras que los atlas lingüísticos languidecían en los empolvados e inmanejables volúmenes. Se lamentaba Séguy de su nula utilización cuando afirmaba que "las ricas colecciones que constituyen los atlas lingüísticos permanecen infrautilizados" (1973a: 3) y Alvar, cuando afirmaba lo siguiente:

> Los atlas abruman por la inmensa cantidad de datos que encierran y si no facilitamos su consulta, quedarán como gigantescos archivos no utilizados. Desgraciadamente, tal es la situación en el mundo hispánico (1984: 112-113).

Se puede afirmar que la DM ha dado una segunda vida a los atlas lingüísticos que posibilita su explotación integral. La DM proporciona las herramientas adecuadas de las que carecía la dialectología tradicional para el análisis integral de la inmensa cantidad de datos que albergan los atlas lingüísticos, posibilitando la proyección cartográfica sintética de los mismos. La DM ha abierto una nueva vía a la dialectología, la vía de la cuantificación.

ISSN: 1139-8736 https://infoling.org/elies/

La DM sirve para cuantificar y objetivar los estudios sobre la variación geolingüística; cuantifica la variación, es decir, propone sistemas para contar las diferencias lingüísticas entre localidades y variedades dialectales. Goebl expresó concisamente el tema con la ecuación siguiente: "Dialectometría = geografía lingüística + taxonomia numérica" (1983: 113), que posteriormente desarrolló de la manera siguiente:

> Todas estas ciencias [botánica, biología...] se hallan ante el mismo dilema, es decir, la misma necesidad: controlar la desconcertante riqueza de los datos empíricos y clasificarlos en categorías para extraer una visión global, tipológica (1983: 113-114).

3. ¿Cómo se trabaja en DM para el análisis de datos lingüísticos?

Los métodos dialectométricos se han ido poco a poco diversificando y cada día hay más técnicas de análisis cuantitativos y más posibilidades de cartografiar dichos resultados. En sus primeros 50 años de existencia la DM ha hecho un recorrido ingente. Desde la distancia interpuntual de los primeros trabajos (Séguy 1973b; Guiter 1973; Goebl 1976) se pasa a la distribución de similitudes y se llega hasta el estudio de las zonas de transición con lógica *fuzzy,* pasando por la DM dendrográfica o clasificación jerárquica, la correlativa, etc. Repasaremos las principales técnicas utilizadas a lo largo de estos años.

Goebl ha explicado en numerosas ocasiones la mecánica de los trabajos dialectométricos o "cadena dialectométrica" (véase, por ejemplo, Goebl 2010). A continuación se explican en detalle los distintos pasos que se siguen en el desarrollo de esta técnica.

3.1. Tipificación de los datos

El primer paso en la cadena dialectométrica es la decisión de elegir los datos del atlas lingüístico en cuestión para el análisis dialectométrico. Se pueden tomar los datos en su integridad o solamente una muestra de los mismos, atendiendo a criterios objetivos (despojo de mapas que contienen lagunas en las respuestas, etc.). Posteriormente se decide la forma de los datos a analizar; se pueden tomar los datos directamente del atlas lingüístico, sin ningún tipo de tratamiento, o transformados de acuerdo a algún tipo de análisis lingüístico (por ejemplo, una lematización de los datos lexicales, o una tipificación o etiquetaje de los datos gramaticales). Esta fase es mucho más importante de lo que a primera vista pueda parecer, puesto que es el fundamento de la diferenciación lingüística. En el fondo lo que se está decidiendo es si la diferenciación lingüística se va a basar en la estructura superficial o en la estructura subyacente de la lengua. Si los datos del atlas lingüístico pasan directamente a la base de datos sin ningún análisis lingüístico (tanto en alfabeto fonético como en el ortográfico), se estará analizando la estructura superficial de la lengua, puesto que el análisis cuantitativo se llevará a cabo con datos de pronunciación, de la actuación. Sin embargo, si los datos son sometidos a un análisis lingüístico (se puede utilizar cualquier teoría lingüística para ello), la investigación será llevada a cabo de acuerdo a la estructura subyacente de la lengua en cuestión.

ISSN: 1139-8736 https://infoling.org/elies/

3.2. Creación de la base de datos

Una vez decidido el nivel lingüístico de los datos, se ha de crear la base de datos. Esta base de datos puede ser exhaustiva (que contenga todos los datos del atlas lingüístico elegido) o puede contener solo una muestra de los mismos. En este caso el requisito que debe cumplir es el de la representatividad, tanto desde el punto de vista de las características lingüísticas como de la población. Se trata del primer eslabón en la objetividad del análisis: cuanto más se aleje del *(single) feature based* dialectología más se acercará a la objetividad en la selección de los datos.

En geolingüística es una práctica habitual el uso de muestras, ante la imposibilidad de analizar el universo lingüístico que albergan las variedades dialectales; de hecho, un atlas lingüístico es una muestra que se supone es representativa del conjunto de rasgos de las variedades analizadas. La representatividad tiene que ser ratificada tanto en el plano lingüístico como en el de las poblaciones.

3.3. Elección de la unidad de distancia lingüística

Sin duda uno de los pasos más importantes dados por la DM ha sido el uso de una unidad de distancia para medir la similitud (o distancia) entre variedades. Es decir, ¿qué y cómo contar las diferencias lingüísticas? ¿En qué basarse para determinar que dos palabras o dos características lingüísticas son iguales o diferentes? Mayoritariamente han sido tres las unidades utilizadas a lo largo de estos años. Séguy (1971: 336; 1973b: 11) utilizó la distancia de Hamming, que consiste en el número de caracteres que difieren de una cadena a otra. Por su parte, Goebl usa una medida nominal o categórica (1983: 115), para ello se hace necesario tipificar las formas recogidas en los atlas lingüísticos. A este proceso de tipificación Goebl lo denomina "taxación" (etiquetaje). Como unidad de medida usa el índice relativo de identidad-IRI (Goebl 1987: 70), renombrado índice relativo de similitud-IRS (Goebl 2013: 145), un índice no ponderado en el que todos los *taxats* o etiquetas tienen idéntico valor (1).

Goebl utiliza también el índice de distancia-IRD. Como estas dos distancias se miden en porcentajes, la IRD = 100 – IRS (si la IRS es igual a 40, la IRD será 60). Tanto la IRS como la IRD son unidades de porcentajes de similitud o distancia que hay entre dos localidades o variedades teniendo en cuenta todas las características que han sido tenidas en cuenta para la investigación. La suma de las similitudes y las disimilitudes o distancias (= diferencias) es igual a 100.

(1) *apatx* *azazkal* *índice de similitud IRS = 0*
índice de distancia IRD = 100%

Goebl también propuso una unidad de distancia ponderada, denominada índice ponderado de identidad-IPI (Goebl 1987: 74), en el que toman más peso las formas menos usuales y menos expandidas.

La otra unidad de distancia que ha tenido gran recorrido es la distancia o algoritmo de *Levenshtein*, también llamado "distancia de edición" *(edit distance)* o "distancia de cadena" *(string distance)* y que se basa en la eliminación, sustitución o adición de sonidos para pasar de una palabra a otra o de una variante a otra (véase, entre otros, Kessler 1995; Nerbonne y Heeringa 2010; Valls et al. 2012). Esta unidad se usa para

ISSN: 1139-8736 https://infoling.org/elies/

medir la distancia fonética entre dos palabras basándose en la pronunciación, pero tiene el inconveniente de que no es una medida adecuada cuando se trata de contar las distancias entre dos palabras etimológicamente diferentes. No tiene sentido medir, por ejemplo, la distancia fonética entre las palabras vascas *apatx* y *azazkal* que significan 'pezuña', puesto que se trata de dos palabras completamente diferentes (2).

(2)	*apatx*	*azazkal*	*diferencia*
	a	*a*	*0*
	p	*z*	*1*
	a	*a*	*0*
	tx	*z*	*1*
		k	*1*
		a	*1*
		l	*1*

La diferencia entre estas dos palabras según la distancia *Levenshtein* es 5/7 (0,71). No es una medida adecuada, porque son dos vocablos diferentes y las coincidencias de sonidos son fortuitas.

Aunque no han logrado la popularidad de las anteriores, son conocidas también otras unidades de distancia como: distancia Euclídea, distancia Manhattan, distancia Canberra, distancia binaria o Minkowski, distancia cofenética, distancia distributiva o $\chi 2$, también llamada CHI-2 y KHI-2 (Philps 1984), entre otras.

Si el uso de una técnica estadística apropiada es importante, tiene mayor importancia si cabe la elección de una buena medida. Para ello se ha de saber qué es lo que queremos medir y cómo lo vamos a hacer. Podemos medir la distancia lingüística de los dialectos en la estructura fonética o superficial, o en la estructura subyacente o fonológica, con análisis lingüístico previo de los datos. Como ya ha quedado demostrado (Clua 2010) los dos tipos de análisis son interesantes, pero hay que tener en cuenta que los resultados son muy diferentes. Es, por tanto, labor del dialectólogo decidir el tipo de datos que debe utilizar en su investigación. De acuerdo con la decisión del plano lingüístico, el dialectólogo ha de decidirse por una u otra medida de distancia.

3.4. Matriz de distancias

La aplicación de la unidad de distancia a la base de datos proporciona la matriz de distancias. Esta matriz reúne las distancias lingüísticas entre todas las poblaciones analizadas. De hecho, se pueden crear matrices de distancias diferentes, según la medida de distancia que se aplique: si se aplica la media de similitud-IRS se obtendrá una matriz de similitudes, en la que el 100 % indica identidad absoluta; en cambio, si se aplican otras medidas de distancia, como la *Levenshtein*, se obtendrán matrices de distancia, en la que 100 % indicará máxima distancia (Figura 1).

ISSN: 1139-8736 https://infoling.org/elies/

	Ahetze	Aia	Aldude	Alkotz	Altzai	Altzürükü	Amezketa	Andoain	Aniz	Aramaio
Abaurregaina	40,61	43,99	35,64	33,05	48	47,8	46,59	45,74	35,07	53,69
Ahetze		41,85	26,25	36,92	40,65	41,56	43,3	43,68	35,81	52,92
Aia			46,54	37,74	53,22	54,53	22,51	21,97	41,89	44,14
Aldude				37,92	38,59	39,15	46,96	47,08	35,96	52,14
Alkotz					50,74	49,44	37,94	37,96	23,7	48,96
Altzai						14,17	56,12	54,77	49,98	60,03
Altzürükü							56,51	55,77	48,9	60,31
Amezketa								21,31	40,05	43,52
Andoain									41,65	44,76
Aniz										50,15

Figura 1. Matriz de distancias de localidades vascas con datos del EHHA (parte de la matriz)

Como se puede observar en la Figura 1, las distancias lingüísticas entre localidades se expresan utilizando el mismo sistema de las distancias geográficas expresadas en km. Así, por ejemplo, la distancia lingüística entre la localidad de Abaurregaina y Aia es un 43,99 % y de la misma localidad a Aramaio 53,69 % (es decir, entre las dos localidades más de la mitad de las características analizadas son diferentes).

La matriz de distancias lingüísticas puede ser un punto de partida interesante para tipos de análisis como el grado de dialecticidad que puede albergar una lengua, análisis de las localidades más aisladas, etc.

3.5. Análisis estadísticos de los datos

Una vez lograda la matriz de distancias lingüísticas el dialectólogo tiene ante sí dos puntos a decidir: la técnica estadística a llevar a cabo y la visualización de los resultados. El repertorio de técnicas estadísticas es grande y han sido muchas las utilizadas en la dialectometría. En esta contribución se citarán las más fructíferas en este campo del saber.

3.5.1. Primer paso: distancias entre localidades o variedades dialectales

La proyección de la cuantificación de las similitudes o diferencias lingüísticas entre localidades y variedades dialectales se llevó en un primer momento a mapas, y posteriormente también a diferentes recursos visuales.

La distancia interpuntual fue el primer análisis dialectométrico que se diseñó (Séguy 1973b; Guiter 1973; Goebl 1976). La distancia interpuntual proporciona las distancias lingüísticas entre dos localidades, tanto contiguas como alejadas. El procedimiento habitual es la visualización de la distancia entre localidades colindantes por medio de mapas isoglóticos: mapas en los que cada localidad analizada está representada por un polígono y los lados de los polígonos representan una “isoglosa cuantitativa” que permite visualizar el número de desigualdades entre dos localidades vecinas. Tanto Séguy (en lo que denominó DM lineal o unidireccional), como Guiter (en lo que denominó método global) y Goebl (en lo que denominó dialectometría interpuntual) utilizaron este procedimiento. Se trata de un paso importante en la cuantificación de las distancias lingüísticas entre localidades. De esta manera, Guiter, por ejemplo, pudo establecer una jerarquía de fronteras dialectales (1973: 63) con datos de diversos atlas lingüísticos del sur y sureste de Francia que abarcaban el vasco, el catalán y el occitano.

Otro de los resultados de la proyección de las distancias interpuntuales son los mapas de similitudes, en los que partiendo de una localidad se pueden visualizar las diferencias con las demás localidades. Utilizando esta técnica se pueden crear tantos mapas como

ISSN: 1139-8736 https://infoling.org/elies/

localidades se hayan analizado (Figura 2). Al proyectar los resultados en el mapa, los polígonos de las localidades son coloreados de acuerdo a la similitud que guardan entre ellas, lo que se representa por medio de colores (los colores rojos representan la máxima similitud, los colores azul oscuro, la mínima similitud). Se trata de una técnica cartográfica utilizada por geógrafos y usuarios de análisis de conglomerados jerárquicos, que fue incorporada a la DM por Goebl (1987).

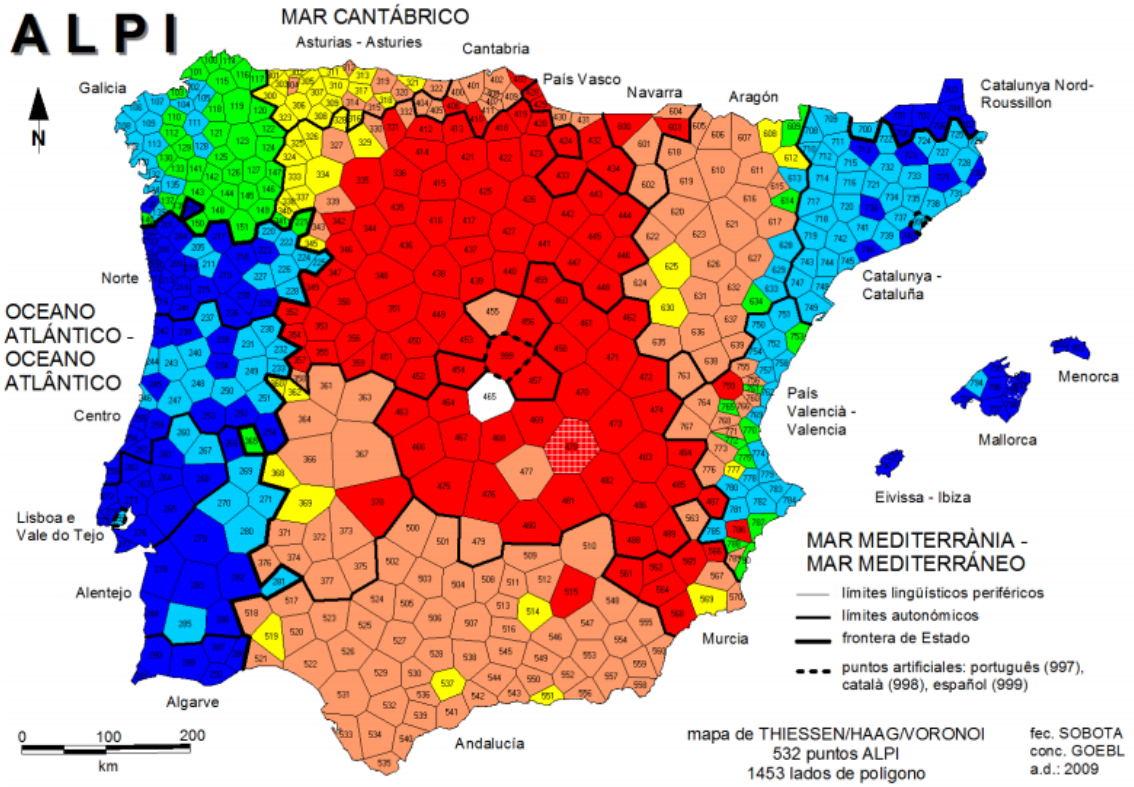

Figura 2. Distribución de similitud de la localidad Camarenilla-465 (Toledo) con relación a las hablas locales (ALPI) (mapa cedido por Goebl)

Las distancias lingüísticas entre localidades pueden ser agrupadas en distintos grupos. En el mapa de la Figura 2, se visualizan en 6 grupos (localidades en rojo, naranja, amarillo, verde, azul claro y azul oscuro). Como se puede comprobar en la Figura 2, la localidad de Camarenilla (polígono blanco en el centro del mapa) tiene altos niveles de similitud con localidades de su entorno y zona norte de la Península (localidades en rojo) y buenos niveles con localidades andaluzas y aragonesas (localidades en naranja). Sin embargo, la similitud decae según se acerca a zonas catalanas, gallegas (amarillos, verdes, azul claros y oscuros) y, sobre todo, portuguesas (con la mayoría de sus localidades en azul oscuro)[1].

3.5.2. Segundo paso: clasificación de los dialectos

Sin embargo, la gran explosión de la dialectología cuantitativa se dará con la aplicación de técnicas aglomerativas multivariantes *(multivariate analysis).* La DM dio un salto espectacular con el empleo de técnicas estadísticas más sofisticadas y de mayor alcance

ISSN: 1139-8736 https://infoling.org/elies/

en la delimitación de las distancias y fronteras dialectales. La posibilidad de la jerarquización de dichas fronteras (fronteras bruscas y fronteras tenues, por ejemplo) y la clasificación de los dialectos vino de la mano del análisis de conglomerados o clasificación aglomerativa jerárquica *(cluster analysis)*, llamada también DM dendrográfica y clasificación jerárquica. La dialectometría aglomerativa pone al alcance del investigador todo un arsenal de técnicas (métodos jerárquicos y no jerárquicos) y algoritmos (distancias mínimas, distancias máximas, distancias entre centroides, distancias ponderadas, Ward o varianza mínima, vecino más alejado, o más próximo), con la posibilidad de combinar todos ellos con distintas unidades de distancia.

A día de hoy, la dialectometría aglomerativa se ha convertido en una técnica estándar para la detección y clasificación de variedades y áreas dialectales, pero, al mismo tiempo, se trata de la parte más extraña para el lingüista no habituado en estadística. Por todo ello, se hace necesario aclarar algunos conceptos.

El primero de ellos es el concepto de "algoritmo", que se puede definir como un sistema determinado de agrupar las variedades basado tanto en la distancia intragrupal (distancia que hay dentro de un grupo de variedades dialectales) o intergrupal (distancia que hay entre dos grupos de variedades dialectales diferentes).

El segundo de los conceptos es el "análisis de conglomerados". Este análisis consiste en agrupar localidades dentro de una jerarquía, en la que paso a paso se van aglutinando empezando por las dos localidades que tengan más elementos en común; es decir, las dos localidades más próximas lingüísticamente. Una vez agrupadas las dos primeras localidades en adelante funcionan como si fueran una única localidad. Se recurre otra vez al mismo procedimiento para buscar las siguientes localidades más próximas lingüísticamente y se funden en un grupo. El mecanismo se va repitiendo hasta que todas las localidades se hallen agrupadas (Goebl 1991, 1992 y ss.; Nerbone et al. 2008: 2). El principio general del análisis de conglomerados es, por tanto, agrupar un gran conjunto de datos en subconjuntos más pequeños, en el que cada subconjunto sea lo más homogéneo posible en sí mismo y tan distinto como sea posible con los demás subconjuntos. Mediante esta técnica todas las variedades analizadas se integran en una jerarquía (Figura 3).

El gráfico resultante, llamado dendrograma, señala perfectamente la jerarquía en la que se engloba cada variedad dialectal. El dendrograma posibilita jerarquizar las fronteras que van surgiendo según se van agrupando las variedades. Cuanto más alto sea el nivel de configuración de un conglomerado la frontera que dibuja en el mapa tendrá mayor impacto y será lingüísticamente más relevante.

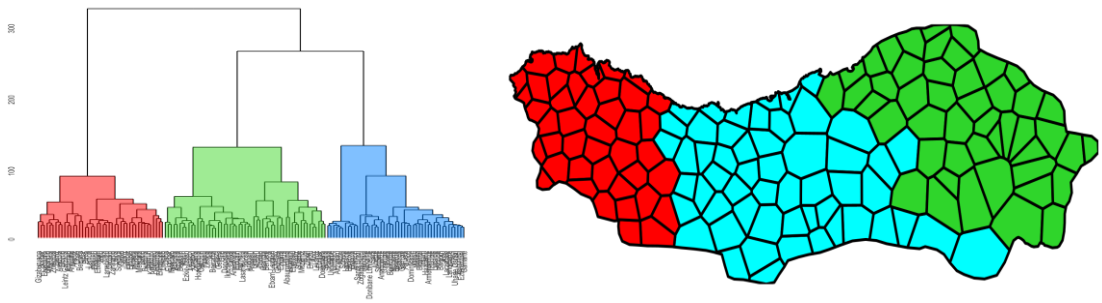

Figura 3. Dendrograma y mapa de los dialectos del euskera, respectivamente

ISSN: 1139-8736 https://infoling.org/elies/

El dendrograma (el gráfico de la izquierda de la Figura 3) recoge todas las variedades analizadas (en la parte baja del dendrograma) en el estudio de la clasificación de los dialectos vascos. El dendrograma, en su parte izquierda, se halla provisto de una escala que agrupa la distancia lingüística que contienen las variedades. En este caso la escala llega hasta el nivel 360, que es la diferenciación total entre todas las variedades; a mayor diferenciación lingüística entre localidades, más alto será el nivel de la escala. Para visualizar los dialectos en el mapa se ha dado un corte en el nivel 160 de la escala del dendrograma; el corte presenta las variedades agrupadas en tres grupos (rojo, verde y azul). A este nivel se halla representada el 44,44 % de la varianza (el nivel 360 representa el 100 % de la varianza). La proyección cartográfica del dendrograma refleja tres dialectos o grupos dialectales: occidental (en rojo), central (azul) y oriental (verde). Hay que señalar, sin embargo, que las fronteras entre la zona roja y la azul, y la frontera entre la zona azul y la verde no se encuentran al mismo nivel de la escala, puesto que el grupo rojo se une al grupo formado por el azul y el verde en el nivel más alto de la escala, mientras que estos dos grupos se unen más tempranamente. La frontera que separa el grupo rojo del azul es jerárquicamente superior a la frontera que separa los grupos azul y verde.

El mapa visualiza el espacio geográfico de los grupos dialectales definidos en el dendrograma. En este caso, los tres grupos dialectales han sido considerados como "dialectos". Se trata de tres dialectos (el occidental, el central y el oriental) con una distribución geográfica compacta. Es de resaltar que la frontera entre el dialecto occidental y el central es idéntica a la frontera trazada por otros procedimientos de la dialectología tradicional.

Los resultados obtenidos mediante esta técnica cuantitativa son mucho más fiables que las clasificaciones dialectales logradas por métodos de la dialectología tradicional. A pesar de ello, hay que decir que no se trata de una técnica perfecta. Estas clasificaciones pueden sufrir cierta inestabilidad ya que pequeños aportes de datos pueden derivar en diferencias significativas en los resultados. Por esta razón, es necesario que los resultados sean corroborados con otras técnicas. Nerbonne et al. (2008) proponen dos técnicas para superar esta inestabilidad: por un lado, el *bootstrapping* (denominado también *bootstrap clustering)* o método de remuestreo y, por otro, el *noisy clustering* o agrupamiento con ruido.

En primer lugar, el *bootstrapping* o método de remuestreo es una técnica de análisis que lleva a cabo diversos muestreos con reemplazamiento de datos en cada uno de ellos. El proceso de agrupación se repite un determinado número de veces con elección de diversos elementos de muestra en cada uno de ellos. El resultado final es un dendrograma compuesto *(composite dendrogram)* en el que consta el número de veces que aparece el grupo en las distintas iteraciones (véase, por ejemplo, Nerbonne et al. 2008: 651).

Por su parte, el *noisy clustering* o agrupamiento con ruido funciona a partir del cálculo de la varianza en la matriz de distancias, así mismo, se especifica un techo de ruido y se repite la operación 100 veces. El resto del procedimiento es idéntico al *bootstrapping.* Pero ¿qué es el ruido en una matriz de datos? El ruido en los datos está relacionado con la estabilidad y la variabilidad desde el punto de vista estadístico; cuando hay mucha inestabilidad y gran variabilidad se dice que hay mucho ruido en los datos y suele estar relacionado con errores tipográficos en la entrada de datos, formas raramente

ISSN: 1139-8736 https://infoling.org/elies/

documentadas, variables exógenas, etc. (véase Wieling y Montemagni 2016).

En todos los casos, la visualización de los resultados se logra por medio del dendrograma y de la cartografía. El dendrograma fruto de la clasificación ha de ser interpretado en primer lugar y, posteriormente, se ha de elegir una técnica de validación de la mejor partición para su proyección al mapa. Esta labor es facultad del investigador, pero la estadística proporciona una serie de criterios (coeficiente de correlación cofenético, coeficiente de pertenencia, replicación, simulaciones Monte Carlo, interpretabilidad teórico-práctica...) para certificar que dicha partición es correcta (véase más en Prokić y Nerbonne 2008; Mucha y Haimerl 2005).

3.5.3. Tercer paso: el análisis del *continuum* dialectal

La técnica aglomerativa divide las áreas dialectales de forma rígida, en grupos compartimentados en los que las fronteras dialectales se dibujan abruptas. Sin embargo, no siempre se da este tipo de fronteras entre áreas dialectales. Esta técnica no refleja siempre de modo adecuado la realidad dialectal, porque no siempre se da una frontera clara. Esta técnica anterior, por tanto, no visualiza los casos de *continuum* dialectal, en el que el paso de un área a otra se hace de forma gradual. La visualización de este *continuum* dialectal ha sido posible mediante la técnica *Multidimensional Scaling* (MDS) o escalamiento multidimensional, que fue aplicada por primera vez en dialectometría por Embleton (1987a, 1987b), siendo desarrollada más tarde por Embleton et al. (2013) y Nerbonne (2010), entre otros.

El escalamiento multidimensional reduce las dimensiones de la base de datos a una, dos o tres dimensiones. Las variedades analizadas se ubican en un espacio multidimensional abstracto. Esta técnica presenta las similitudes lingüísticas en un espacio bi o tridimensional, en el que la distancia lingüística entre variedades y áreas dialectales se expresa en el espacio con mayor o menor proximidad entre ellas, sin necesidad de ninguna frontera. Se han desarrollado diversas opciones para visualizar los resultados: 2D diagrama, 2D diagrama con localidades conectadas por líneas con el centro del área, 3D diagramas e incluso con posibilidad de interactuar para seleccionar el mejor plano (Embleton et al. 2013: 15), diagramas en diversos colores y mapas (véase Figura 4).

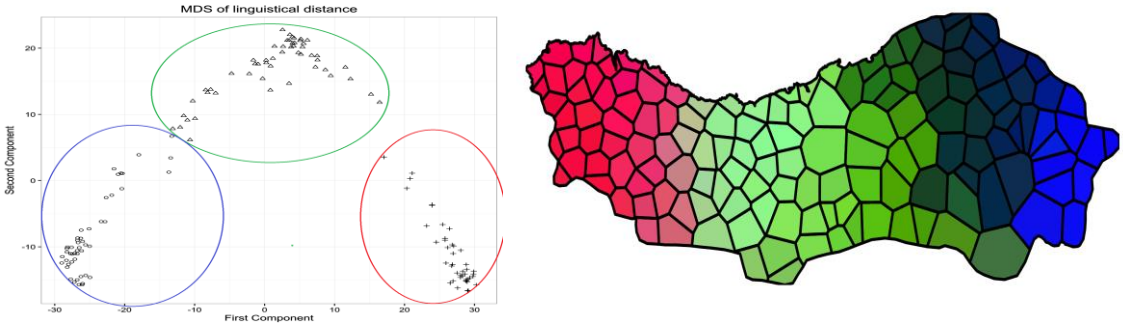

Figura 4. Diagrama y mapa correspondiente al MDS con datos del euskera

El gráfico de la Figura 4 presenta las variedades del euskera en una V invertida. Se han analizado los datos en tres dimensiones que en el gráfico se señalan con símbolos y en el mapa con colores. El gráfico señala claramente un alejamiento lingüístico entre las variedades que se hallan en la parte derecha (en círculo rojo, que corresponden a las

ISSN: 1139-8736 https://infoling.org/elies/

variedades en rojo en el mapa), mientras que no hay una clara separación entre las otras dos áreas dialectales (círculos en verde y azul) que en el mapa se señalan en color verde y azul. Esto indica que la frontera entre el área verde y azul no es tan nítida y abrupta como la frontera entre el área roja y verde.

3.5.4. Cuarto paso: el estudio de la relación entre la distancia geográfica y la distancia lingüística

La influencia de la geografía en la variación lingüística ha sido un tema recurrente en la historia de la geolingüística desde sus comienzos. La creencia general es que la distancia lingüística crece a medida que se va alejando desde un punto de partida. Esta relación no ha podido ser concretada hasta que los dialectólogos han podido disponer de herramientas cuantitativas. Fue Séguy el primero que formuló, en su curva de correlación (1971: 348), los principios de la DM correlativa, con datos de 16 atlas lingüísticos. Posteriormente, el tema ha sido una y otra vez analizado por distintos investigadores y distintos datos: Goebl con datos del *Atlas linguistique de la France*-ALF (2005) y *Atlas Lingüístico de la Península Ibérica*-ALPI (2013) (véase Figura 5), Nerbonne, en solitario y con colaboradores (2010, 2013; Nerbonne y Heeringa 2007; Spruit et al.2008; Nerbonne et al. 2010, entre otros), Aurrekoetxea con datos del corpus *Bourciez* (2010, 2016), etc. Esta relación ha hecho avanzar los estudios sobre los factores externos de la variación geolingüística hasta tal punto que Nerbonne ha propuesto un *Principio Fundamental de la Dialectología*, según el cual "las variedades lingüísticas próximas son generalmente –pero no siempre– más similares que las lejanas" (Nerbonne 2013: 222).

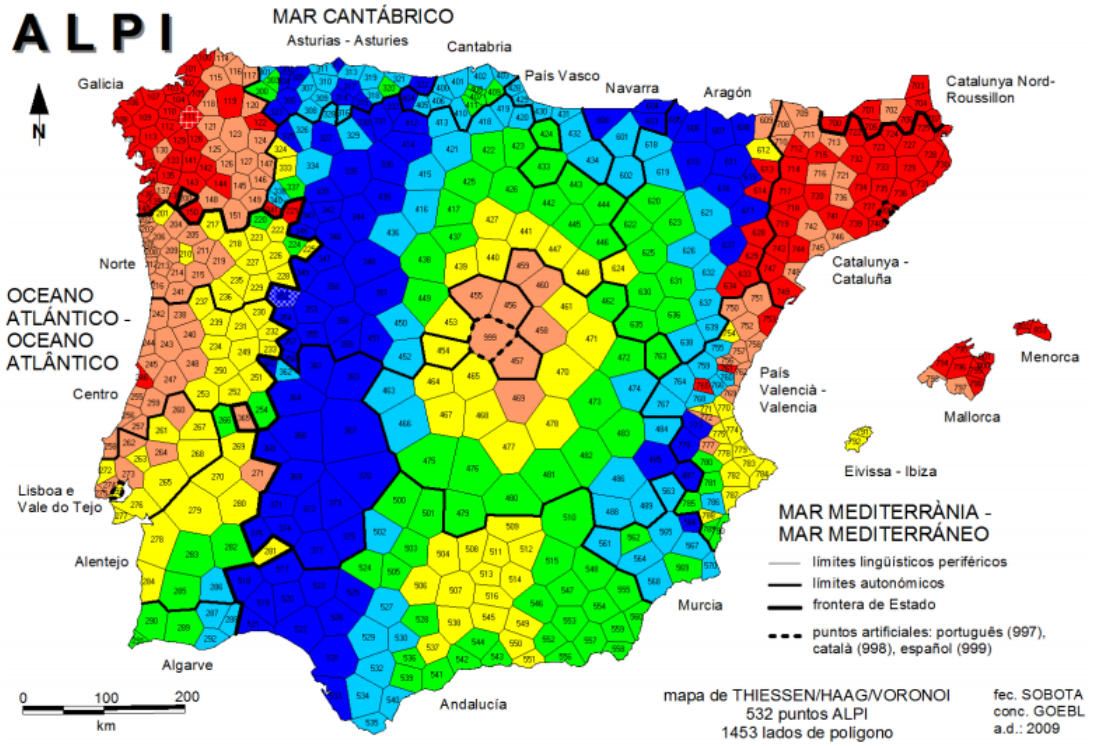

Figura 5. Correlación entre la distancia geográfica y la distancia lingüística en ALPI (mapa cedido por H. Goebl)

ISSN: 1139-8736 https://infoling.org/elies/

La Figura 5 presenta la correlación entre la distancia geográfica y la distancia lingüística en la Península Ibérica. Como se puede observar, hay una alta correlación en Galicia, Cataluña, Portugal y la zona central (zonas en rojo, naranja y amarillo); sin embargo, hay una baja correlación en regiones marcadas en azul, tanto oscuro como claro, y verde.

El factor de la distancia geográfica como el factor determinante en la diferenciación lingüística ha sido, sin embargo, puesto en duda: así, por ejemplo, Szmrecsanyi (2012) sostiene que la distancia geográfica es un mal predictor de la variabilidad morfosintáctica con datos del inglés. Otros investigadores propugnan el estudio de la correlación entre el tiempo necesario para trasladarse de una localidad a otra con la distancia lingüística. Por ejemplo, Gooskens (2005) analiza el tiempo de trayecto como indicador de la distancia lingüística: cuanto más tiempo de trayecto para trasladarse de una localidad a otra, mayor diferenciación lingüística hay entre dichas variedades. Posteriores investigaciones avalan esta vía.

Aparte de esta correlación la dialectología cuantitativa ha llegado a analizar la correlación entre distintas categorías gramaticales; como, por ejemplo, la correlación entre la distancia fonológica y morfológica, o entre esta última y la léxica, etc. De hecho, se pueden analizar las correlaciones entre todas estas categorías gramaticales.

3.6. Otros métodos dialectométricos

Si bien tanto el análisis de conglomerados como el escalamiento multidimensional han sido los métodos más utilizados en la dialectometría clásica, se han utilizado y siguen siendo utilizados cantidad de métodos y técnicas cuantitativas. En este apartado se expondrán sucintamente algunos de ellos: *Dual Scaling* o análisis de correspondencias, auto-correlación espacial, análisis de factores, análisis de componentes principales, agrupamiento difuso, glotograma, agrupamiento de bipartitos espectrales y dialectometría inversa.

3.6.1. *Dual Scaling* o análisis de correspondencias

Cichocki (1989) fue el primero en implementar esta técnica en geolingüística con el estudio sobre las frecuencias de tres variables en el francés canadiense. Los resultados de este análisis se representan en dos espacios multidimensionales, un espacio de localidades y un espacio lingüístico. En el diagrama la proximidad entre localidades indica afinidad lingüística, de tal modo que localidades lingüísticamente afines se encuentran más cercanas que las localidades lingüísticamente más alejadas. El espacio de las variantes lingüísticas se superpone a las localidades. La técnica del *dual scaling* extrae los patrones lingüísticos más importantes de las variedades y los representa en dos gráficos. Es una técnica que ayuda a conocer la estructura geolingüística de ciertas variables lingüísticas, pero no sirve para el estudio de muchas variables a la vez.

3.6.2. Auto-correlación espacial

Esta técnica fue propuesta por Kretzschmar (1992, 1996) y desarrollada por Grieve et al. (2011). En ella las variables lingüísticas se someten a un análisis de auto-correlación espacial para identificar patrones significativos de variación lingüística regional, que es

ISSN: 1139-8736 https://infoling.org/elies/

similar a proyección de isoglosas. Se trata de una medida de dependencia espacial que permite medir el grado de agrupamiento espacial en los valores de una variable.

3.6.3. Análisis de factores

El análisis de factores ha sido utilizado en más de una ocasión: Chauveau (1985) fue el primero en utilizarlo. Posteriormente Aurrekoetxea (1995), Inoue (2004), Shackleton (2005), Nerbonne (2006) y Leinonen (2008), entre otros, lo han utilizado en sus investigaciones. Se trata de una técnica de reducción de datos que permite al investigador descubrir las variables que muestran patrones similares de variación y si las variables se pueden dividir en subconjuntos independientes. Las variables que se correlacionan entre sí se combinan en factores, lo que significa que la cantidad total de datos se puede reducir.

3.6.4. *Principal Components Analysis*-PCA o análisis de componentes principales

El análisis de componentes principales es un método multivariante para capturar la variabilidad intrínseca en los datos: una combinación de características, las cuales explican la varianza en los datos de la mejor manera posible. Este es el primer componente principal. Posteriormente se procede buscando repetidamente una combinación de características que no esté correlacionada con las del componente principal o anteriores y que explica la mayor parte de la variación restante en los datos. Los resultados pueden ser cartografiados: las localidades encuadradas en cada componente principal son ploteadas con distintos colores (Shackleton 2005; Hyvönen et al. 2007; Ueda 2017).

3.6.5. *Fuzzy clustering* o agrupamiento difuso

El *fuzzy clustering* es un tipo de análisis de conglomerados que analiza la probabilidad de un agrupamiento de las variables analizadas. Hay algunas aplicaciones en geolingüística como Meschenmoser y Pröll (2012), Pröll (2013) o Aurrekoetxea, Iglesias et al. (2020) para el análisis de zonas de transición.

En la DM clásica la información detallada contenida en los mapas individuales se pierde en el proceso de construcción de áreas o grupos dialectales; es decir, se puede saber cómo se distribuyen los dialectos, pero no hay información sobre cuáles son las variables que actúan en cada caso. Se utiliza este procedimiento para construir grupos de mapas identificables que comparten patrones espaciales similares de variantes. Este método tiene como objetivo preservar la información contenida en mapas individuales y mantenerla accesible en cada paso en lugar de agregarlo, de esta forma no hay pérdida de información contenida en las variantes. Desde este punto de vista, el enfoque depende de la calidad de los métodos de búsqueda de corpus para recuperar características y compararlas con otras. De esta forma se crean grupos de mapas que exhiben similares propiedades estructurales. La similitud de dos mapas se mide mediante la función de covarianza.

La covarianza describe la relación de dos variables aleatorias y muestra valores positivos altos si las cantidades tienden a comportarse de manera similar, esto es, valores altos de A implican valores altos de B y valores bajos de A implican valores

ISSN: 1139-8736 https://infoling.org/elies/

bajos de B. La covarianza toma valores negativos en caso contrario, es decir, cuando estas implicaciones no se cumplen y valores altos de A no implican valores altos de B. La covarianza no depende de la posición de las cantidades de interés sino solo de su distancia. De esta manera, la función de covarianza describe la estructura de las propiedades de los mapas. Como la similitud se basa en la función de covarianza, dos mapas son similares si comparten las mismas propiedades estructurales. Estas varianzas, por tanto, son analizadas por *fuzzy cluster* análisis.

De este modo, esta técnica estadística proporciona la oportunidad de tratar los mapas de los atlas lingüísticos como un corpus, accesibles a través de procesos automatizados. Ello proporciona una nueva posibilidad de estudio de los corpus dialectales, que hasta ahora estaba restringida al uso de datos categóricos, así como a la falta de detección de patrones espaciales.

3.6.6. Glotograma

Esta técnica combina la edad de los locutores con el factor geográfico y parte de la premisa de que el lenguaje se difunde geográfica y socialmente. El glotograma fue concebido en la geolingüística japonesa en la década de 1980 y se utiliza para estudiar la dinámica de las lenguas o variedades en contacto. Los resultados se recogen en mapas y gráficos con signos en los que se señala el cambio lingüístico que se produce en el parámetro temporal y espacial, todo ello de acuerdo a la edad de los locutores y sus localizaciones (Sanada 2010).

3.6.7. *Bipartite-Spectral graph* o agrupación de gráficos bipartitos espectrales

La DM aglomerativa ha sido criticada porque ha priorizado el estudio de áreas dialectales en detrimento de las variables lingüísticas. En los últimos años se está revirtiendo la situación. Wieling y Nerbonne (2009, 2011) propusieron una nueva técnica que busca grupos de variedades y simultáneamente la asociación de sonidos relacionados con esos grupos. Por consiguiente, con esta técnica se estudia la relación dialectal de las localidades y su base lingüística conjuntamente. Para ello se miden las distancias de las localidades con el estándar usando la distancia *Levenshtein* (véase sección 3.3.). El objetivo no es lograr cuáles son los sonidos más característicos de cada grupo o los sonidos que más veces aparecen en dichas localidades. En realidad, el objetivo es, por una parte, lograr grupos de localidades con respecto a la distancia con la variedad estándar y, por otra, la correspondencia de sonidos entre la variedad estándar y los grupos de localidades.

Una variante es el *Hierarchical Spectral Partitioning of biparte graph* (Wieling y Nerbonne 2010) que conlleva una ventaja importante sobre otros métodos dialectométricos, puesto que la base lingüística se determina simultáneamente, cerrando la brecha entre la dialectología tradicional y la cuantitativa. Además de mostrar que los resultados del agrupamiento jerárquico mejoran con respecto al método de agrupamiento espectral plano utilizado en un estudio anterior (Wieling y Nerbonne 2009), esta técnica se utiliza para identificar las correspondencias de sonidos más importantes para cada grupo. Esta es una ventaja importante con respecto al método jerárquico, puesto que evita la necesidad de métodos externos para determinar las correspondencias de sonido más importantes para un grupo geográfico.

ISSN: 1139-8736 https://infoling.org/elies/

3.6.8. *Reverse dialectometry* o dialectometría inversa

La dialectometría inversa propone analizar la distribución espacial de cada una de las características lingüísticas. Se trata de detectar e identificar variables gramaticales midiendo la afinidad entre ellas, en función de las áreas que marca cada variable. Este punto de vista ha sido denominado como *reverse dialectometry* (Craenenbroeck 2014) o *variant-based dialectometry* (Pickl y Rumpf 2012). Aurrekoetxea et al. (2021) indagan en la misma vía, analizando las variables con un clúster jerárquico y un MDS de las características verbales de la lengua vasca, y llegan a la conclusión de que algunas características lingüísticas tienen los mismos patrones geolingüísticos.

3.7. Programas integrales creados para la DM

Los paquetes estadísticos del mercado presentan inconvenientes para su aplicación en la geolingüística. Entre estos se pueden citar, por ejemplo, que estos paquetes no presentan ningún tipo de cartografía, lo que hace prácticamente inviable su uso, o que no tienen en cuenta las múltiples respuestas. Por todo ello, los programas integrales de dialectometría son una buena opción, puesto que presentan únicamente las técnicas estadísticas requeridas en la DM y suelen incluir la opción de trasladar a un mapa los resultados de los procesos estadísticos. Actualmente los dialectómetras tienen distintas herramientas hechas exclusivamente para el estudio cuantitativo de la variación geolingüística. En las siguientes líneas se glosarán algunos de las más accesibles y eficaces. Es de destacar que todos ellos son programas de libre acceso.

3.7.1. *Visual DialectoMetry*-VDM (Universidad de Salzburgo)

La primera herramienta de DM, denominada *Visual DialectoMetry*-VDM (http://ald.sbg.ac.at/dm), fue diseñada por H. Goebl, creada por Haimerl y desarrollada en la Universidad de Salzburgo[2]. Goebl ha publicado en más de una ocasión las características de esta herramienta (Goebl 2006; Goebl y Smečka 2016).

Se trata de un programa que funciona en local, en la plataforma *Windows*, con dos unidades de distancia categóricas (índice relativo de similitud-IRS e índice ponderado de similitud-IPS) y una base de datos numeral (el dialectólogo debe adaptar sus datos lingüísticos a numéricos). Por todo ello, ha tenido un gran impacto en la geolingüística, sobre todo románica. El mayor hándicap de la herramienta es su programa cartográfico, que debe implementarse imperativamente en Salzburgo.

Mediante esta herramienta se pueden lograr mapas de similitud, mapas isoglóticos *(honeycomb map / isoglotic maps)*, mapas de rayos *(beam map)*, mapas de zonas de transición (desviación típica), mapas de zonas conservadoras *(Skewness o coefficient d'asymétrie de Fisher),* mapas de zonas de mayor cohesión lingüística (máxima, *synopsis of the maxima of N similarity distributions*), DM correlativa (correlación entre la distancia geográfica y la diferencia lingüística o entre dos categorías gramaticales), análisis de conglomerados (con diversos algoritmos: *Complete, Ward, Average...*), algoritmos de visualización / discretización de las distancias lingüísticas (MedMinMax, Med, MedMW), etc.

3.7.2. *Gabmap* (Escuela de Groninga)

ISSN: 1139-8736 https://infoling.org/elies/

Gabmap es una aplicación web destinada a facilitar las exploraciones en dialectología cuantitativa asistida por computadora (https://www.gabmap.nl/). *Gabmap* crea varias vistas de datos de dialectos, desde histogramas de caracteres utilizados para detectar errores de codificación, pasando por alineaciones de transcripciones fonéticas utilizadas para medir la distancia de pronunciación, hasta diagramas de escala de colores multidimensionales destinados a ilustrar resultados cuantitativos de forma intuitiva, etc. Muchos análisis van acompañados de instalaciones que permiten a los investigadores estudiar más, y por ejemplo, buscar las bases lingüísticas más importantes de una división de área, o examinar los resultados de la agrupación para la confiabilidad estadística.

El software, escrito de código fuente abierto (Nerbonne et al. 2011), funciona tanto con datos categóricos (lexicales o gramaticales), como con datos numéricos (vectores de frecuencias de formantes de las vocales...). Está dirigido especialmente al análisis de datos fonéticos y al estudio de las distancias de pronunciación entre localidades. Para ello usa la unidad de distancia *Levenshtein.* El usuario tiene, además, a su disposición distintas técnicas de análisis como mapas de rayos, análisis de conglomerados, MDS, agrupamiento difuso, agrupamiento con ruido, etc.

3.7.3. *Diatech* (UPV/EHU)

La herramienta *Diatech* es también una aplicación web (http://eudia.ehu.eus/diatech/), libre y multilingüe (Aurrekoetxea et al. 2013; Aurrekoetxea et al. 2016). Uno de los motivos principales para la creación de la herramienta fue posibilitar el tratamiento de las respuestas múltiples (MR) o polimorfismo (Aurrekoetxea, Nerbonne, Rubio 2020). En los paquetes estadísticos convencionales tan solo se podía introducir un dato por cada casilla de la base de datos. Por consiguiente, si se tienen datos con MR y se quieren analizar todos los datos a la vez, sin ningún tipo de selección y sin pérdida de información, inevitablemente se ha de elegir un programa *ad hoc,* como el *Diatech.* Por ejemplo, la herramienta permite la introducción de más de una respuesta a cada pregunta en la misma localidad. Un botón de referencia: El mapa correspondiente al caso “motivativo indeterminado” de la declinación de las palabras terminadas en la vocal “o” en vasco (*Euskararen Herri Hizkeren Atlasa-EHHA* V: mapa 1082); de 145 localidades, se han recogido más de una respuesta en 56 localidades.

La aplicación funciona con datos en alfabeto normativo, fonético o lematizado / etiquetado. Además, pone a disposición del usuario tres medidas de distancia: índice de similitud, índice ponderado de similitud y la distancia *Levenshtein;* es decir, unidades de distancia utilizadas tanto en la escuela de Salzburgo como Groninga. Los análisis estadísticos que se pueden hacer con esta herramienta son los siguientes: mapas sinópticos (mapas de similitudes, áreas de transición, correlación entre distancias geográficas y lingüísticas...), mapas de rayos, mapas isoglóticos, análisis de *cluster*, MDS y agrupamiento difuso.

3.7.4. *ProDis* (UB)

La DM o el análisis cuantitativo de los datos dialectales se ha expandido a todas las categorías lingüísticas. Quizás la que más ha tardado en ser analizada cuantitativamente

ISSN: 1139-8736 https://infoling.org/elies/

ha sido la prosodia. Con este fin, es decir, el de analizar los datos prosódicos con datos numéricos, el grupo del Laboratorio de Fonética de la UB (Elvira-García et al. 2018; Fernández Planas et al. 2019) implementaron un programa para el tratamiento cuantitativo de los datos acústicos llamado *ProDis*. El programa provee similitudes acústicas entre localidades, clasificaciones jerárquicas y agrupamientos multidimensionales.

3.8. Otras vías para hacer DM

El avance en dialectología cuantitativa ha sido inmenso, inimaginable al comienzo de su andadura (una buena visión de las técnicas utilizadas en *aggregate variationist analyses* se puede consultar en Nerbonne y Wieling 2018). En los últimos años ha ido creciendo la implementación de técnicas cuantitativas a la variación geolingüística. El *software R* ha propiciado una nueva vía para ello. La proliferación de técnicas no hace más que enriquecer el estudio en este campo. La DM se ha convertido, por tanto, en una tarea interdisciplinaria en el que la colaboración entre dialectólogos, geógrafos, matemáticos e informáticos es esencial. Una muestra de esta interdisciplinariedad se puede encontrar en Jeszenszky et al. (2018), en estudios basados en sistemas complejos con datos recogidos de Twitter (Gonçalves y Sánchez 2016; Donoso y Sánchez 2017; Ruiz-Tinoco 2018), en el uso de *USTM autoencoder* (Rama y Çöltekin 2016) o en técnicas para reducir los cuestionarios representativos mediante Coeficiente Simple de Emparejamiento y métodos *k-means* (Aurrekoetxea, Clua et al. 2020), entre otros.

4. ¿Por qué es relevante este instrumento metodológico?

La relevancia de la aportación de la DM a los estudios sobre la variación geolingüística se puede resumir en cuatro aspectos: (i) el uso de técnicas cuantitativas para la medición de la diversidad lingüística, (ii) la objetividad en la elección de los datos, (iii) la clasificación jerárquica de los dialectos y (iv) los avances recientes de la dialectología.

4.1. Técnicas cuantitativas para la medición de las distancias lingüísticas

El gran paso de Jean Séguy (1973a) fue su decisión de dar cuenta de las diferencias lingüísticas entre las localidades basándose en los datos del *Atlas linguistique de la Gascogne-ALG*. El resultado de esta decisión fue la plasmación de las diferencias lingüísticas entre las localidades en un mapa con la creación de lo que más tarde se denominó isoglosa cuantitativa; es decir, la distancia lingüística entre localidades se visualiza con el grosor de las isoglosas.

Su trabajo manual pasó rápidamente a ser elaborado por procedimientos automatizados y la cuantificación computarizada. Este planteamiento rompe radicalmente con el modo tradicional de investigación en geografía lingüística e inicia un modo de investigación nuevo, desconocido totalmente, y revolucionario con el uso de la estadística. A pesar de que algunos todavía tengan reparos para aceptarlo, lo cierto es que la dialectología se está poniendo en *modo ciencia*. Como ciencia empírica que es, la dialectología comienza a usar lo que se conoce como "método científico". En concreto se empiezan a utilizar dos procedimientos científicos: la medición y el razonamiento. Y dos principios: el de la reproductibilidad y el de la refutabilidad.

ISSN: 1139-8736 https://infoling.org/elies/

De hecho, esta tarea implica como primer paso la búsqueda de una unidad de medida de las diferencias lingüísticas (véase sección 3.3.). Por esta razón, Séguy, cuando propuso el uso de la *'distancia de Hamming'* como unidad de medida, dio un paso gigante: por primera vez en la historia de la geolingüística se proponía una unidad para medir las diferencias lingüísticas entre localidades. Este hecho tiene una relevancia incuestionable y marca un hito en el desarrollo de la cuantificación de la variación geolingüística. La relevancia no está en que haya acertado con una unidad de medida determinada, sino en el hecho mismo de proponer una unidad de medida.

En definitiva, la DM fundamentalmente proporciona una vasta colección de herramientas y técnicas para el análisis de la variación geolingüística, de tal forma que actualmente no hay dificultad alguna para medir cualquier distancia lingüística, ya sea fonética (segmental o suprasegmental), fonológica, sintáctica, morfológica o léxica. Por consiguiente, posibilita cuantificar las diferencias lingüísticas entre localidades, jerarquizar las fronteras dialectales, clasificar jerárquicamente áreas dialectales, detectar zonas conservadoras y áreas de transición, visualizar la correlación entre la diferenciación lingüística y la distancia espacial o geográfica, el influjo de la densidad lingüística en la variación, la visualización del *continuum* dialectal, etc.

4.2. Objetividad

La objetividad que proporciona la DM se basa en dos pilares: (i) el uso de grandes masas de datos y (ii) el uso de técnicas estadísticas de análisis.

Uno de los puntos más débiles de la dialectología tradicional era el uso de datos elegidos por el investigador, que, aunque no fuera de manera consciente, estaban sujetos a la subjetividad del autor. Por esta razón, el uso de amplias bases de datos, creadas con criterios objetivos, ha sido otra de las aportaciones de la DM a los estudios de la geografía lingüística. Gracias a este paso, se consiguió un cambio en el paradigma de estudio: de los *(single) feature-based Variation Studies*, estudios basados en unas pocas características (Nerbonne 2009) a la *aggregate dialectology* o dialectología aglutinativa. Además, esta decisión conlleva el uso exhaustivo de los atlas lingüísticos. Séguy expuso claramente esta determinación:

> Proponemos determinar la diferencia dialectal basándonos no en unos pocos criterios elegidos arbitrariamente, sino en la integración de todos los datos contenidos en los seis volúmenes del *Atlas lingüístique de la Gascogne* (1973a: 21).

Si bien los primeros proyectos de atlas lingüísticos no implicaban la creación de bases de datos, pronto comenzaron los proyectos atlantográficos: para su creación, se recurrió a la informática, y para su proyección, a la cartografía (se recomienda consultar Baiwir y Renders 2013 para la discusión sobre si los atlas lingüísticos pueden ser considerados como corpus). No hay que olvidar que los primeros dialectómetras tuvieron que crear sus bases de datos partiendo de datos de atlas lingüísticos ya publicados (por ejemplo, Séguy con su *Atlas linguistique de la Gascogne-ALG* y Goebl con su proyecto de dialectometrización del *Atlante Italo-Svizzero-AIS, Atlas linguistique de la France-ALF* o *Atlas lingüístico de la Península Ibérica-ALPI*, entre otros). Todo el tiempo, el esfuerzo humano y económico invertido en estas tareas ha merecido la pena, sin duda

ISSN: 1139-8736 https://infoling.org/elies/

alguna. Sin un atlas lingüístico o proyecto semejante de recogida y ordenación de los datos sería y es impensable hablar de DM. Los atlas lingüísticos son un requisito básico.

Cuando no es posible utilizar los atlas lingüísticos en su integridad, el dialectólogo debe atenerse a los requisitos estadísticos de las muestras, el segundo principio de objetividad anteriormente expuesto. La fiabilidad de la muestra es una condición básica en toda investigación puesto que la objetividad es uno de los factores de la fiabilidad. En lingüística, como en otras ciencias, es necesario utilizar muestras debido a la imposibilidad de abarcar la lengua o la variedad de una región en su conjunto y en su totalidad. De esta manera, el objetivo de los lingüistas es que su muestra de datos sea representativa del universo que pretende representar. En este sentido, la inmensa mayoría de los atlas lingüísticos son muestras representativas, sobre todo, en el aspecto léxico, morfológico y fonético. Quizás el aspecto sintáctico es el que queda infrarrepresentado, especialmente en los atlas más clásicos.

4.3. Clasificación dialectal jerárquica

La DM ha contribuido a solucionar el problema de la clasificación dialectal. No hay duda de que el análisis de conglomerados ha dado consistencia y seguridad a las clasificaciones dialectales. Las clasificaciones previas, llevadas a cabo por métodos tradicionales, han de ser revisadas y contrarrestadas con las realizadas con técnicas estadísticas. A pesar de que todo análisis estadístico reduce los factores de variabilidad, resaltando los factores que más incidencia tienen en ella, y que suponen una simplificación de la realidad lingüística, la clasificación obtenida por estos métodos cuantitativos es mucho más fiable que la llevada a cabo por métodos manuales.

El análisis de conglomerados, por otra parte, puede ser interpretado tanto desde el punto de vista sincrónico como del diacrónico: Goebl (2002) ha dado muestras de ello con datos del *Atlas linguistique de la France*-ALF. La interpretación sincrónica del dendrograma se apoya en una lectura ascendente en la creación del dendrograma y visualiza la diferenciación lingüística actual entre grupos dialectales, mientras que la interpretación diacrónica, en una lectura descendente, explica las fragmentaciones sucesivas a lo largo de la historia de dichas variedades lingüísticas.

4.4. Avance de la dialectología

La DM ha posibilitado un avance teórico de la dialectología. Algunas cuestiones que hasta ahora no podían ser resueltas o no podían avanzar han tenido un mayor recorrido gracias a la aportación de la DM. Por ejemplo, desde los primeros trabajos dialectológicos, nadie discutía que había una relación estrecha entre la distancia geográfica y la distancia lingüística. Sin embargo, esa suposición no ha podido ser formulada cuantitativamente hasta que los trabajos dialectométricos han determinado porcentajes de correlación que varían de una lengua a otra, de un contexto lingüístico a otro. A consecuencia de esto, se sabe que el factor de distancia geográfica no es un factor predominante en la variación geolingüística, o, al menos, no en todos los casos. Así mismo, gracias a los avances dialectométricos, hoy en día, se está en disposición de jerarquizar las clasificaciones de las variedades dialectales con la suficiente seguridad.

Bien es cierto que quedan por resolver muchos problemas de la geolingüística, sobre todo, en la interpretación de los resultados estadísticos. Por ejemplo, el dialectómetra ha

ISSN: 1139-8736 https://infoling.org/elies/

sido incapaz hasta el momento de denominar las áreas creadas por él mismo. Se da a este respecto una paradoja interesante: la dialectología tradicional, cuando llega a definir áreas dialectales, no duda ni un instante en nombrar dichas áreas y no tiene ningún reparo en clasificar dichas zonas como dialectos, subdialectos, variedades... Hay muy pocos casos en los que dudan o incluso prefieran no utilizar la denominación "dialecto", y uno de ellos es el gallego. García de Diego (1959: 130; 1984: 155) habla de la no existencia de dialectos en el gallego por ser poco profundas las diferencias entre las distintas modalidades, Zamora Vicente (1953: 80) habla de dos subdialectos. Fernández Rei corrobora la idea de García de Diego de que "non se pode falar de dialectos propiamente ditos no galego" (1990: 36), idea que ha sido corroborada por Dubert-García (2020) con procedimientos estadísticos.

5. Conclusiones

En este estudio, por consiguiente, se han expuesto detenidamente los pasos a seguir en la tarea dialectométrica, desde la elección del atlas lingüístico, la construcción y el etiquetaje de la base de datos, la elección de la unidad de distancia lingüística pasando por la elección de las distintas técnicas cuantitativas en consonancia a los objetivos de la investigación.

Como conclusión general, por una parte, hay que destacar que la DM ha delimitado la distancia lingüística en términos cuantitativos y ha posibilitado el avance teórico en temas propios de la dialectología, como la distancia entre variedades, la clasificación de los dialectos, la relación entre la distancia geográfica y la distancia lingüística, etc. Por otra parte, se han creado y desarrollado distintas técnicas utilizadas en DM para medir las distancias lingüísticas, delimitar grupos dialectales (MDS y el análisis de *cluster*), y delimitar y analizar las zonas de transición y zonas arcaicas, etc. Pero, ante todo, la aportación de la DM ha sido clave en el uso de grandes masas de datos para eliminar definitivamente la subjetividad del dialectólogo en la selección de las características. La superación de la fase *Single feature-based dialectology* para llegar al *aggregate Dialectology* ha sido un paso crucial.

En definitiva, la relevancia de la DM se mide en las técnicas cuantitativas para la medición de las distancias lingüísticas, en la objetividad de los resultados, en las técnicas jerárquicas de clasificación de variedades y en posibilitar el avance teórico de la dialectología en general.

6. Referencias

Alvar, Manuel. 1984. Automatización de los índices en los atlas lingüísticos. En M. Alvar, ed. *Informática y Lingüística.* Málaga: Ágora, pp. 107-119.

Atwood, E. Bagby. 1955. The phonological divisions of Belgo-Romance. *Orbis* IV: 367-389.

Aurrekoetxea, Gotzon. 1995. *Bizkaieraren egituraketa geolinguistikoa.* Bilbao: UPV/EHU.

Aurrekoetxea, Gotzon. 2010. The correlation between morphological, syntactic and phonological variation in the Basque language. En B. Heselwood y Cl. Upton,

ISSN: 1139-8736 https://infoling.org/elies/

eds. *Proceedings of Methods XIII. Papers from the Thirteenth International Conference on Methods in Dialectology, 2008.* Frankfurt: Peter Lang, pp. 207-218.

Aurrekoetxea, Gotzon. 2016. Distantzia geografikoaren eta hizkuntza distantziaren arteko korrelazioa. En G. Aurrekoetxea, J. M. Makazaga y P. Salaberri, eds. *Txipi Ormaetxea omenduz. Hire bordatxoan*, Bilbao: UPV/EHU, pp. 55-72.

Aurrekoetxea, Gotzon; Clua, Esteve; Iglesias, Aitor; Usobiaga, Iker; Salicrú, Miquel. 2020. Characterizing dialect groups: distance and informativeness associated with linguistic forms. *Zeitschrift für Dialektologie und Linguistik* 2020.2: 307-326. DOI: 10.25162/zdl-2020-0011

Aurrekoetxea, Gotzon; Fernández-Aguirre, Karmele; Rubio, Jesús; Ruiz, Borja; Sánchez, Jon. 2013. 'DiaTech': A New Tool for Dialectology. *Literary and Linguistic Computing* 28.1: 23-30. DOI: 10.1093/llc/fqs049

Aurrekoetxea, Gotzon; Iglesias, Aitor; Clua, Esteve; Usobiaga, Iker; Salicrú, Miquel. 2020. Analysis of transitional areas in Dialectology: Approach with Fuzzy Logic. *Journal of Quantitative Linguistics.* DOI: 10.1080/09296174.2020.1732765

Aurrekoetxea, Gotzon; Iglesias, Aitor; Santander, Gotzon; Usobiaga, Iker. 2016. Diatech: tool for making dialectometry easier. *Dialectologia* 17: 1-22.

Aurrekoetxea, Gotzon; Abasolo, Juan; Clua, Esteve; Usobiaga, Iker; Salicrú, Miquel. (En prensa). Hizkuntza-aldagaien aldakortasuna: lehen hurbilketa. *Uztaro 118*, 61-79.

Aurrekoetxea, Gotzon; Nerbonne, John; Rubio, Jesús. 2020. Unifying analyses of multiple responses. *Dialectologia* 25: 59-86.

Baiwir, Esther; Renders, Pascale. 2013. Les atlas linguistiques sont-ils des corpus? *Corpus* 12: 27-37.

Chauveau, Jean-Pierre. 1985. Etude de cartes linguistiques par l'analyse factorielle des correspondances. En A. Moll, eds. *Actes del XVI CILFR* (Ciutat de Mallorca, 7-12 d'abril de 1980), 2. Palma de Mallorca, pp. 379-397.

Cichocki, Wladyslaw. 1989. An application of dual scaling in Dialectometry. *Journal of English Linguistics* 22.1: 91-95.

Clua, Esteve. 2010. Relevancia del análisis lingüístico en el tratamiento cuantitativo de la variación dialectal. En G. Aurrekoetxea y J. L. Ormaetxea, eds. *Tools for Linguistic Variation.* Bilbao: UPV/EHU, pp. 151-166.

Craenenbroeck, Jeroen van. 2014. Quantity and quality in linguistics-reverse dialectometry. En *Methods in Dialectology XV. August 11–15 4, 2014*, Groningen, the Netherlands.
https://static1.squarespace.com/static/5217e223e4b090faa01f8f2d/t/55a4e0eee4b0d12431c376a2/1436868846371/project-phd.pdf

Davis, Alva L.; MacDavid, Raven I. 1950. A transition area. *Language* 26.2: 264-273.

Donoso, Gonzalo; Sánchez, David. 2017. Dialectometric analysis of language variation in Twitter. En P. Nakov, M. Zampieri, N. Ljubešić, J. Tiedemann, S. Malmasi y A. Ali, eds. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects.* Association for Computational Linguistics, pp. 16-25. DOI: 10.18653/v1/W17-12

Dubert-García, Francisco. 2020. O galego, unha lingua sen dialectos: olladas sociais e lingüísticas sobre a variación dialectal. *Estudios Románicos* 29: 147-163.

EHHA: Euskaltzaindia. 2013. *Euskararen Herri Hizkeren Atlasa.* Vol. V. Bilbao: Euskaltzaindia.

ISSN: 1139-8736 https://infoling.org/elies/

Elvira-García, Wendy; Balocco, Simone; Roseano, Paolo; Fernández-Planas, Ana María. 2018. ProDis: A dialectometric tool for acoustic prosodic data. *Speech Communication* 97: 9-18.

Embleton, Sheila. 1987a. A new technique for dialectometry. En V. B. Makkai, ed. *Twelfth LACUS Forum*. Jupiter Press: Lake Bluff, Illinois.

Embleton, Sheila. 1987b. Multidimensional scaling as a dialectometrical technique. En R. M. Babitch, ed. *Papers from the Eleventh Annual Meeting of the Atlantic Provinces Linguistic Association.* New Brunswick: Centre Universitaire de Shippagan in Shippagan, pp. 33-49.

Embleton, Sheila; Uritescu, Dorin; Wheeler, Eric S. 2013. Defining dialect regions with interpretations: Advancing the multidimensional scaling approach. *Literary and Linguistic Computing* 28.1: 13-22.

Fernández Planas, Ana María; Elvira-García, Wendy; Balocco, Simone; Roseano, Paolo. 2019. Análisis dialectométrico con *ProDis*: un paso más en los estudios prosódicos de AMPER. En J. Dorta, ed. *Investigación geoprosódica. Amper: análisis y retos.* Madrid/Frankfurt am Main: Iberoamericana Vervuert, pp. 119-135.

Fernández Rei, Francisco. 1990. *Dialectoloxía da lingua galega.* Vigo: Xerais.

García de Diego, Vicente. 1990. *Elementos de gramática histórica gallega (fonética-morfología).* Burgos: Hijos de Santiago Rodríguez.

García de Diego, Vicente. 1959. *Manual de dialectología española.* Madrid: Ediciones Cultura Hispánica.

García de Diego, Vicente. 1984. *Elementos de gramática histórica gallega (Fonética-Morfología).* Burgos, 1906.

Goebl, Hans. 1976. La dialectométrie appliquée à l'ALF (Normandie). En A. Várvaro, ed. *XIV Congresso internazionale di linguistica e filologia romanza. Atti.,* II. Napoles/Amsterdam: G. Macchiaroli / J. Benjamins, pp. 165-195.

Goebl, Hans. 1983. Matrices, distances et interpoints. Etude dialectométrique sur les isoglosses quantitatives. En Marie-Rose Simoni-Aurembou, ed. *Cahier des Annales de Normandie n°15. Dialectologie et littérature du domaine d'oïl occidental: actes du colloque tenu à l'Université de Caen en février 1981* pp. 113-137. DOI: 10.3406/annor.1983.3899

Goebl, Hans. 1987. Points chauds de l'analyse dialectométrique: pondération et visualisation. *Revue de linguistique romane* 51: 63-118.

Goebl, Hans. 1991. Una classificazione gerarchica di dati geolinguistici tratti dall'AIS. Saggio di dialettometria dendrografica. *Linguistica* 31:341-351.

Goebl, Hans, 1992. Problèmes et méthodes de la dialectométrie actuelle (avec application á l'ais). En G. Aurrekoetxea y X. Videgain, eds. *Actas del congreso internacional de dialectología*. Bilbao: Euskaltzaindia, pp. 429-475.

Goebl, Hans. 2002. Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66: 5-63.

Goebl, Hans. 2005. La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de linguistique romane* 69: 321-367.

Goebl, Hans. 2006. Recent advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21.4: 411-435.

Goebl, Hans. 2010. Introducción a los problemas y métodos según los principios de la Escuela Dialectométrica del Salzburgo (con ejemplos sacados del "Atlante italo-

ISSN: 1139-8736 https://infoling.org/elies/

Svizzero" AIS). En G. Aurrekoetxea y J. L. Ormaetxea, eds. *Tools for linguistic variation.* Bilbao: UPV/EHU. Anejos del Seminario de Filología Vasca Julio Urquijo LIII, pp. 3-39.

Goebl, Hans. 2013. La dialectometrización del ALPI: rápida presentación de los resultados. En E. Casanova Herrero y C. Calvo Rigual, eds. *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas (Valencia 2010).* Berlin/Boston: Walter de Gruyter, vol. VI, pp. 143-154.

Goebl, Hans; Smečka, Pavel. 2016. The quantitative nature of working maps (WM) and taxatorial areas (TA). A brief look at two basic units of Salzburg Dialectometry (S-DM). *Studies in Quantitative Linguistics 23. Issues in Quantitative Linguistics* 4: 11-127.

Gonçalves, Bruno; Sánchez, David. 2016. Learning about Spanish dialects through Twitter. *RILI* XVI-2: 65-75.

Gooskens, Charlotte. 2005. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica* 13: 38-62.

Grieve, Jack; Jurgens, David. 2019. Mapping word frequencies on Twitter using R and Python. En *Workshop presented at NWAV 48, University of Oregon, 2019-10-10.* http://jurgens.people.si.umich.edu/tutorials/Mapping_Word_Frequencies_on_Twitter_using_Python.html

Grieve, Jack; Speelman, Dirk; Geeraerts, Dirk. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23: 1-29.

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

Houck, Charles L. 1967. A computerized statistical methodology for linguistic geography: a pilot study. *Folia linguistica* 1.1-2: 80-95.

Hyvönen, Saara; Leino, Antti; Salmenkivi, Marko. 2007. Multivariate analysis of Finnish dialect data – An overview of lexical variation. *Literary and Linguistic Computing* 22.3: 271-290.

Inoue, Fumio. 2004. Multivariate analysis, geographical gravity centers and the history of the standard Japanese forms. *Area and Culture Studies* 68: 15-36.

Jeszenszky, Péter; Stoeckle, Philipp, Glaser, Elvira; Weibel, Robert. 2018. A gradient perspective on modeling interdialectal transitions. *Journal of linguistic geography* 6.2: 78-99.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. En *Proceeding of the European ACL*. Dublin: ACL, 60-67.

Kretzschmar, William. 1992. Isoglosses and predictive modeling. *American Speech* 67: 227-249.

Kretzschmar, William A. 1996. Quantitative areal analysis of dialect features. *Language Variation and Change* 8: 13-39.

Leinonen, Therese. 2008. Factor analysis of vowel pronunciation in Swedish dialects. *International Journal of Humanities and Arts Computing* 2.1-2: 189-204.

Meschenmoser, Daniel; Pröll, Simon. 2012. Using fuzzy clustering to reveal recurring spatial patterns in corpora of dialect maps. *International Journal of Corpus Linguistics* 17: 176-197.

Mucha, Hans Joachim; Haimerl, Edgar. 2005. Automatic validation of hierarchical cluster analysis with application in Dialectometry. En C. Weihs y W. Gaul, eds.

ISSN: 1139-8736 https://infoling.org/elies/

Classification—the ubiquitous challenge. Proc. of 28th Mtg Gesellschaft fur¨ Klassifikation, Dortmund, Mar. 9-11, 2004. Berlin: Springer, pp. 513-520.

Nerbonne, John. 2010. Mapping aggregate variation. En A. Lameli, R. Kehrein y S. Rabanus, eds. *Language and space. International handbook of linguistic variation. Vol. 2*. Berlin: De Gruyter Mouton, pp. 476-495.

Nerbonne, John; Colen, Rinke; Gooskens, Charlotte.; Kleiweg, Peter; Leinonen, Therese. 2011. Gabmap – a web application for Dialectology. *Dialectologia* Special issue II: 65-89.

Nerbonne, John; Heeringa, Wilbert. 2007. The geographic distribution of linguistic variation. *Studies in generative grammar*: 267-298. https://www.lingexp.uni-tuebingen.de/sfb441/LingEvid2006/abstracts/nerbonne.pdf

Nerbonne, John; Heeringa, Wilbert. 2010. Measuring dialect dierences. En P. Auer y J. E. Schmidt. eds. *Theories and Methods*, Vol.1. Berlin, New York: De Gruyter Mouton, pp. 550-567.

Nerbonne, John; Kleiweg, Peter; Manni, Franz; Heeringa, Wilbert. 2008. Projecting dialect distances to Geography: Bootstrap clustering vs. noisy clustering. En Ch. 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., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. Berlin / Heidelberg: Springer, pp. 647-654.

Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21.4: 463-475.

Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3.1: 175-198. DOI: 10.1111/j.1749-818X.2008.00114.x

Nerbonne, John. 2013. How much does Geography influence language variation? En P. Auer, M. Hilpert, A. Stukenbrock y B. Szmrecsanyi, eds. *Space in Language and Linguistics. Geographical, Interactional, and Cognitive Perspectives.* Berlin: De Gruyter, pp. 220-236.

Nerbonne, John; Prokić, Jelena; Wieling; Martijn; Gooskens, Charlotte. 2010. Some further dialectometrical steps. En G. Aurrekoetxea y J. L. Ormaetxea, eds. *Tools for Linguistic Variation.* Bilbo: UPV/EHU, pp. 41-56.

Nerbonne, John; Wieling, Martin. 2018. Statistics for aggregate variationist analyses. En Ch. Boberg, J. Nerbonne y D. Watt, eds. *The handbook of Dialectology.* Oxford: John Wiley & Sons, pp. 400-414.

Philps, Dennis. 1984. Dialectométrie automatique. En H. Goebl, ed. *Dialectology*, Série Quantitative Linguistics 21. Bochum: Dr. Brockmeyer, pp. 275-296.

Pickl, Simon; Rumpf, Jonas. 2012. Dialectometric concepts of space: Towards a variant-based dialectometry. En S. Hansen, Ch. Schwarz, Ph. Stoeckle y T. Streck, eds. *Dialectological and folk dialectological concepts of space. Current methods and perspectives in Sociolinguistic research on dialect change*, Series: linguae & litterae, 17, Berlin/Boston: De Gruyter, pp. 199-214. DOI: 10.1515/9783110229127

Prokić, Jelena; Nerbonne, John. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing* 2.1-2: 153-172.

Pröll, Simon. 2013. Detecting structures in linguistic maps−Fuzzy clustering for pattern recognition in geostatistical dialectometry. *Literary and Linguistic Computing* 28.1: 108-118.

Rama, Taraka y Çöltekin, Çağri. 2016. LSTM autoencoders for dialect analysis.

ISSN: 1139-8736 https://infoling.org/elies/

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, Osaka, Japan: The COLING 2016 Organizing Committee, pp. 25-32.

Reed, David W.; Spicer, John L. 1952. Correlation methods of comparing idiolects in a transition area. *Language* 28: 348-359.

Ruiz-Tinoco, Antonio. 2018. Geocorpus del español de las redes sociales y cartografía automática. *Monográficos SINOELE* 17: 598-608.

Sanada, Shinji. 2010. The "glottogram": A geolinguistic tool developed in Japan. *Dialectologia.* Special issue I: 185-196.

Schneider, Edgar W. 1988. Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from Georgia and Alabama. *Journal of English Linguistics* 21: 175-212.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335-357.

Séguy, Jean. 1973a. La dialectométrie dans l'atlas linguistique de la Gascogne. *Revue de Linguistique Romane* 37: 1-24.

Séguy, Jean. 1973b. *Atlas linguistique de la Gascogne. Complément du volume VI.* Paris: CNRS.

Shackleton, Robert G. Jr. 2005. English-American speech relationships: A quantitative approach. *Journal of English Linguistics* 33: 99-160.

Shaw, David, 1974. Statistical analysis of dialectal boundaries. *Computers and the Humanities* 8.3: 173-177.

Spruit, Marco René; Heeringa, Wilbert; Nerbonne, John. 2008. Associations among linguistic levels. *Lingua, special issue on Syntactic databases. Selected papers presented in the special session Comparing Aggregate Syntaxes, Digital Humanities conference, Paris, July 6, 2006*: 65-90.

Szmrecsanyi Benedikt. 2012. Geography is overrated. En S. Hansen, C. Schwarz, P. Stoeckle y T. Streck, eds. *Dialectological and folk dialectological concepts of space.* Berlin: de Gruyter, Walter GmbH & Co., pp. 215-231.

Ueda, Hiroto. 2017. Two statistical treatments of Spanish vocabulary: Composite indices of frequency and dispersion and principal component analysis applied to ordinal frequencies. *Dialectología*, Special Issue VII: 187-227.

Valls, Esteve; Nerbonne, John; Prokić, Jelena; Wieling, Martin; Clua, Esteve; Lloret, M-Rosa. 2012. Applying the Levenshtein distance to Catalan dialects: A brief comparison of two dialectometric approaches. *Verba* 39: 35-61.

Wieling, Martijn; Nerbonne, John. 2009. Bipartite spectral graph partitioning to co-cluster varieties andsound correspondences in dialectology. En M. Choudhury, S. Hassan, A. Mukherjee y S. Muresan, eds. *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing.* Suntec, Singapore: Association for Computational Linguistics, pp. 26-34.

Wieling, Martijn; Nerbonne, John. 2010. Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. En *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL, Uppsala, Sweden, July 16, 2010,* Stroudsbourg: Association for Computational Linguistics, pp. 33-41.

Wieling, Martijn; Nerbonne, John. 2011. Bipartite spectral partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25.3: 700-715.

Wieling, Martijn; Montemagni, Simonetta. 2016. Infrequent forms: Noise or not? En

ISSN: 1139-8736 https://infoling.org/elies/

M.-H. Côté, R. Knooihuizen, J. Nerbonne, eds. *The future of dialects: Selected papers from methods in Dialectology XV.* Berlin: Language Science press, pp. 215-223.

Zamora Vicente, Alonso. 1953. De geografía dialectal: -ao, -an en gallego. *Nueva revista de filología hispánica* VII. 1-2: 73-80.

---

[1] Los mapas se pueden consultar en la página web: http://dialektkarten.ch/dmviewer/alpi/index.es.html

[2] Para una revisión de los trabajos dialectométricos de la escuela de Salzburgo, se puede consultar la siguiente página web: http://dialektkarten.ch/dmviewer/index.es.html

ISSN: 1139-8736 https://infoling.org/elies/