UNIVERSIDAD DE MÁLAGA

Subsidia: Tools and Resources for Speech Sciences

José María Lahoz-Bengoechea
and Rubén Pérez Ramón (Eds.)

SUBSIDIA:
Tools and resources for speech sciences

José María Lahoz-Bengoechea
and Rubén Pérez Ramón (Eds.)

Universidad de Málaga

ISBN: 978-84-09-12881-5

Subsidia: Tools and Resources for Speech Sciences
Subsidia: Herramientas y Recursos para las Ciencias del Habla

Libro publicado por la Universidad de Málaga (UMA)

Subsidia: Tools and Resources for Speech Sciences is a scientific publication arising from the organisation of a congress with the same name, carried out in the city of Málaga (Spain) in June, 2017. Its main goal is to give voice to tools and resources developed with the aim of facilitating research in the field of speech sciences. This framework embraces subjects such as phonetics, experimental phonetics, phonology, discourse analysis or dialectology, among others. This book, outcome of the collaboration of expert researchers on their respective areas, aims to be an aid to the scientific community in the sense that it compiles and depicts a series of materials that, we hope, may result beneficial to keep moving forward in the research.

The papers collected in this volume are a selection of those submitted to the above-mentioned congress, and have undergone peer review.

Subsidia: Herramientas y Recursos para las Ciencias del Habla es una publicación científica resultante de la organización del congreso del mismo nombre desarrollado en la ciudad de Málaga (España) en junio del año 2017. Su objetivo principal es dar a conocer herramientas y recursos desarrollados con el objetivo de facilitar la investigación en el campo de las ciencias del habla. Dentro de este marco se engloban disciplinas tan variadas como la fonética experimental, la fonología, el análisis del discurso o la dialectología, entre otras. Este libro, resultado de la colaboración de investigadores expertos en sus respectivas áreas, pretende ser una ayuda a la comunidad científica en tanto en cuanto recopila y describe una serie de materiales que esperamos resulte provechoso para continuar avanzando en la investigación.

Los artículos recogidos en este volumen son una selección de los que se presentaron en dicho congreso y han pasado una evaluación por pares.

TECHNICAL EDITION / EDICIÓN TÉCNICA

Cover design / Diseño de la cubierta: Rubén Pérez Ramón & José María Lahoz-Bengoechea.

Subsidia: Tools and Resources for Speech Sciences

CREDITS

ARTICLES

Subsidia: Tools and Resources for Speech Sciences

Subsidia: Tools and Resources for Speech Sciences

Génesis y aspectos fundamentales de ProDis

Ana Maria Fernández Planas[1], Paolo Roseano[1], Wendy Elvira-García[1] y Simone Balocco[1]

[1] Universitat de Barcelona
e-mail: anamariafernandezp@ub.edu

Citation / Cómo citar esta publicación: Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea & R. Pérez Ramón (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

RESUMEN: El objetivo de este trabajo consiste en presentar ProDis, una herramienta informática para el análisis dialectométrico de la entonación creada por el equipo del *Laboratori de Fonètica* de la *Universitat de Barcelona*. El trabajo presenta un breve estado de la cuestión sobre los sistemas de dialectometría disponibles en la actualidad y se hace especial hincapié en los sistemas capaces de tratar datos prosódicos numéricos y en la necesidad del Atlas Multimedia de la Entonación del Espacio Románico (AMPER) de crear una herramienta propia que sirva para gestionar los datos recogidos mediante el corpus fijo del proyecto en cuestión. Asimismo, se detallan los sistemas anteriores que han perimido el nacimiento de la herramienta. Posteriormente, se realiza una presentación de ProDis en la que se abordan el método que usa para calcular las distancias prosódicas, las salidas del programa y la información que se puede obtener de dicho análisis.

Palabras clave: dialectometría; entonación; lenguas romances.

ABSTRACT: The aim of this paper is to present ProDis, a software for the dialectometric analysis of intonation. In the first part of the article we present a state of the art of currently available dialectometrical tools. Since the prosodic data that are collected within the *Atlas Multimedia de la Entonación del Espacio Románico* (AMPER) are numeric, we pay special attention to the tools that can dialectometrize numeric prosodic data. After resuming the features of the existing tools that, to a certain extent, can carry out the dialectometric analyses of numeric data, we present more in detail the characteristics of ProDis, the prosodic dialectometrical tool created at the Phonetics Laboratory of the University of Barcelona. The aspects of ProDis that are described are the method used to calculate prosodic distances, the outputs of the program, and the information that can be obtained thanks to the prosodic dialectometry.

Keywords: dialectometry; intonation; Romance languages.

1. INTRODUCCIÓN

Los atlas lingüísticos clásicos constituyen una fuente maravillosa de datos fonéticos, morfológicos, sintácticos y, sobre todo, léxicos (recuérdese por ejemplo el ALiR ―Atlas Linguistique Roman―, el ALPI ―Atlas Lingüístico de la Península Ibérica―, el ALE ―Atlas Linguistique de l'Europe―, o el ALAC ―Atlas Lingüístico de América Central―, por ejemplo, entre otros muchos). El estudio de sus datos fue el punto de partida para el establecimiento de isoglosas y fronteras dialectales que clasificaban los puntos de encuesta en grupos a partir de rasgos considerados muy relevantes de forma cualitativa por los investigadores.

Desde los años 70 y 80 del siglo XX se ha producido un paso natural en el desarrollo dialectal de la mano de la llamada dialectometría, que pretende establecer agrupaciones entre la masa de datos empíricos obtenidos en grandes bases de datos a partir de criterios cuantitativos y de procedimientos estadísticos objetivos. El término dialectometría se debe a Séguy (1973), uno de los padres de dicha disciplina junto con Guiter, aunque fue Goebl quien le dio un impulso definitivo. Goebl (1981) define la dialectometría como una alianza metodológica entre la geolingüística y la taxonomía numérica como disciplina matemática. Exactamente, el autor lo expone de forma sintética de la siguiente manera: dialectometría = geografía lingüística + taxonomía numérica (Goebl, 1981, p. 349).

Ciertamente, lo que los estudios dialectométricos pretenden es utilizar una enorme cantidad de datos que se han generado a través de los estudios dialectológicos y los atlas lingüísticos para establecer agrupaciones entre la masa de datos empíricos disponibles y obtener una distribución en el espacio virtual de los datos (Fernández Planas, Roseano, Martínez Celdrán y Romera Barrios, 2011, p. 145), o también en forma de agrupaciones reflejadas en dendrogramas. Sus resultados permiten una rápida asociación entre los

elementos considerados a partir de su cercanía o su lejanía —es decir, de sus semejanzas o de sus diferencias— y posibilitan condensar una gran cantidad de información cuantitativa en un espacio relativamente reducido.

La dialectometría no pretende eliminar el estudio dialectológico tradicional, sino que busca completarlo y erigirse como una herramienta esencialmente útil cuando se manejan cantidades enormes de datos a partir de grandes bases. Sin embargo, ofrece ciertas ventajas respecto a la dialectología tradicional: (1) permite gestionar sin gran esfuerzo por parte del investigador grandes cantidades de datos, de donde se puede inferir que permite llegar a conclusiones estadísticamente fiables, en el sentido de que van más allá de las meras intuiciones de los investigadores; (2) no hay apriorismos en el tratamiento de los datos, ya que el estudio se centra en valoraciones cuantitativas y no cualitativas. Este hecho supone un cambio radical respecto a la dialectología tradicional, más bien cualitativa, y constituye el fundamento de la reticencia de algunos autores; y (3) la forma de presentación de los datos (análisis de clúster en dendrogramas y escalamiento multidimensional, básicamente) es totalmente visual y favorece una comprensión bastante rápida de los hechos.

Por lo que respecta a las lenguas romances, el método se ha aplicado principalmente a las áreas lingüísticas del ladino (Goebl, 1993; Bauer, 2005), el italiano (Bauer, 2003), el francés (Séguy, 1973; Verlinde, 1988; Goebl, 1987; Goeble, 2003), el gallego (Álvarez Blanco, Dubert, y Sousa, 2006; Sousa, 2006; Saramago, 2002), el bable (D'Andrés Díaz, Álvarez-Balbuena García, y Suárez Fernández, 2003) o el catalán (Clua, 2004, Polanco, 1992). Fuera de la Romania se utiliza también en estudios dialectológicos de lenguas como el holandés (Heeringa y Nerbonne, 2001), el inglés (Goebl y Schiltz, 1997), o el euskara (Aurrekoetxea, 1992). Normalmente se ha trabajado con datos fonético-segmentales, morfológicos o léxicos, como hemos dicho. El estudio de los aspectos prosódicos de diferentes variedades se ha trabajado muchísimo menos.

En el seno de macroproyecto AMPER, Atlas Multimedia de Prosodia del Espacio Románico (Contini, 1992; Contini *et al.*, 2002; Romano y Contini, 2001; Contini, Lai y Romano, 2002; Romano, 2003; Fernández Planas, 2005), tras tener muy avanzada una enorme base de datos prosódicos acústicos, se impone trabajar en la comparación y clasificación de las variedades románicas. De hecho, que en el marco de AMPER se llegara a utilizar la dialectometría era también un desarrollo natural y un paso esperable. Así pues, el llamado "corpus fijo" en el proyecto constituye un terreno óptimo para el estudio dialectométrico. AMPER trabaja con habla cercana a habla de laboratorio en el llamado corpus fijo que cuenta con frases enunciativas e interrogativas absolutas que presentan la misma estructura SVO y número de sílabas en todas las lenguas (más o menos), con todas las combinaciones acentuales posibles en todas las posiciones de la frase salvo en el verbo, con dos hablantes por punto de encuesta (como mínimo) que repiten cada uno de ellos tres veces cada frase.

2. LA DIALECTOMETRÍA Y SUS DATOS

En realidad, cuando hablamos de dialectometría no estamos refiriéndonos a una única técnica, sino a un paraguas metodológico que incluye técnicas distintas que trabajan con el mismo objetivo (la aplicación de técnicas estadísticas a grandes bases de datos para averiguar cómo se agrupan a partir de las distancias que mantienen los elementos entre sí a partir de sus características) pero con diferentes algoritmos y con diferentes tipos de datos.

Desde sus inicios, igual que la dialectología clásica, la dialectometría ha usado datos fonético-fonológicos segmentales, morfológicos, léxicos o sintácticos que se almacenan en bases de datos alfabéticos. Desde el punto de vista estadístico, eso implica que se necesitan algoritmos capaces de establecer distancias cuantitativas a partir de variables nominales, principalmente Levenshtein (Kessler, 1995). El uso de datos prosódicos provenientes de análisis acústicos, más recientemente, ha planteado una cuestión metodológica crucial: conviene operar con variables numéricas, lo cual implica el hecho de necesitar otro tipo de métricas para trabajar. Si los datos prosódicos se transcriben con símbolos alfabéticos, tanto en un nivel más superficial y cercano a las melodías acústicas, como en un nivel mucho más profundo o fonológico, de acuerdo con los postulados para el sistema, conseguimos una cadena alfabética que vuelve a necesitar algoritmos que trabajen con datos alfabéticos o nominales.

Existen herramientas disponibles para trabajar en dialectometría tanto con datos numéricos como con datos alfabéticos. Gabmap (Nerbonne, Colen, Gooskens, Kleiweg, y Leinonen, 2011) puede trabajar con ambas, pero con muchas limitaciones porque el programa no permite vectores y reduce cada variable a un único valor. VisualDialectometry (Goebl y Haimerl, 2004) o DiaTech (Aurrekoetxea, Fernández-Aguirre, Rubio, Ruiz, y Sánchez, 2013) operan con datos nominales pero también ofrecen restricciones severas en el tratamiento de datos prosódicos porque, por ejemplo, no aceptan caracteres que se utilizan en el etiquetaje con los sistemas ToBI ("%" o "*", por ejemplo). Además, los algoritmos no son suficientemente robustos como para hacer frente a diferencias en variabilidad en la longitud de las etiquetas prosódicas.

3. ¿POR QUÉ CREAR UNA HERRAMIENTA NUEVA PARA EL TRATAMIENTO PROSÓDICO DE LAS DISTANCIAS ENTRE LOS DATOS?

Necesitamos una herramienta que trabaje con los datos prosódicos obtenidos en el marco AMPER y que refleje las especificidades de dicho tipo de datos.

La prosodia, por una parte, se manifiesta, fundamentalmente, en tres parámetros: f_0, duración e intensidad, de forma numérica; respectivamente, en Hz o semitonos, en segundos (o milésimas de segundo) y en

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

decibelios. Por otra parte, la prosodia vehicula información sobre la modalidad oracional, el acento léxico, la estructura sintáctica o la manifestación del foco. Finalmente, la prosodia expresa diferencias diafásicas, diastráticas o diatópicas (en terminología coseriana). En este último terreno, AMPER (Martínez-Celdrán y Fernández Planas, 2003–2016), junto con el IARI (Prieto, Borràs-Comes y Roseano, 2010–2014), ha ido conformando una enorme base de datos prosódicos acústicos que es susceptible de ser sometida a estudios dialectométricos a partir de algoritmos que trabajen con datos numéricos.

En el seno de este proyecto hubo una propuesta de herramienta dialectométrica, Stat-Distances (Rilliard y Lai, 2008), que no se acabó de desarrollar del todo y, a pesar de proporcionar resultados interesantes, no ofrecía datos imprescindibles como la matriz de distancias o la explicitación de los algoritmos que usaba. Nuestro primer contacto con la metodología dialectométrica fue de la mano de este programa y nos sirvió para empezar a constatar que nuestros resultados no siempre eran totalmente coincidentes con los obtenidos por la dialectología tradicional, lo cual es plausible ya que la dialectología clásica nunca había tratado este tipo de datos. En seguida Stat-Distances dejó de estar disponible para los investigadores que trabajamos en el proyecto. Entre los diferentes grupos implicados se han sucedido otras propuestas. A saber: un script en R (Martínez Calvo y Fernández Rei, 2015) y el programa que se presenta en este trabajo, que es la versión mejorada de unas rutinas previas que llamamos Calcu-Dista (Roseano, Elvira-García, y Fernández Planas, en revisión; Elvira-García, 2014). La especificidad de los datos prosódicos y la falta de una herramienta útil para operar con ellos nos llevó a proponer nuestra herramienta, a la que llamamos ProDis, procedente de la expresión *Prosodic Distances* (Elvira-García, Balocco, Fernández Planas, Roseano, y Martínez Celdrán, 2015; Fernández Planas, 2016a, 2016b), que es la versión mejorada de Calcu-Dista (Roseano, Fernández Planas, Elvira-García, Cerdà Massó, y Martínez Celdrán, 2015) y también se inspira en Stat-Distances. ProDis funciona, como es habitual en los trabajos en el seno de AMPER, en el entorno MatLab.

Así pues, nuestra herramienta constituye nuestra contribución para construir un programa potente capaz de trabajar en dialectometría a partir de los datos prosódicos numéricos que obtenemos en nuestros análisis en el seno de AMPER. En concreto, responde, en esta primera fase, a las exigencias técnicas que habían surgido dentro del marco del proyecto. En concreto, se necesitaba que nuestra herramienta fuera "amigable" y flexible, que fuera capaz de trabajar con datos numéricos, de solucionar los problemas debidos a las diferencias en número de sílabas entre frases en distintas lenguas en el marco AMPER (por ejemplo, *O pássaro gosta de Renato*, 10 sílabas, *vs. La guitarra se toca con paciencia*, 11 sílabas), de considerar de cada frase 3 repeticiones, de ponderar f_0 por duración, por intensidad o por ambos parámetros a la vez y, finalmente, de trabajar con nuevas lenguas adicionalmente.

4. LA GÉNESIS DE PRODIS: DE CALCU-DISTA A PRODIS

ProDis, como antes Calcu-Dista y antes todavía Stat-Distances, utiliza la fórmula (1), que se inspira en la que propuso Hermes (1998) y que resulta ser una fórmula sencilla para calcular distancias entre datos acústicos numéricos.

$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(f_0x_i - f_0y_i)^2}{N}} \quad (1)$$

El antepasado más cercano de ProDis, conocido como Calcu-Dista, era, más que una herramienta, una rutina para el cálculo de distancias prosódicas a partir de los datos numéricos de las melodías en semitonos. Esa rutina se fundamentaba en tres programas bien conocidos: Praat v. 5.4.01 (Boersma y Weenink, 2014), Excel (Microsoft Office 2007) y SPSS Statistics 20 (IBM). Los programas en cuestión tomaban con punto de partida los datos acústicos previamente procesados por tres programas creados en el seno del Laboratori de Fonètica de la UB y circunscritos al ámbito AMPER: AMPER-Reno, AMPER-Extra y AMPER-Eti (Roseano, 2012).

En primer lugar, un script de Praat creado *ad hoc* extraía, a partir de los archivos txt de cada repetición de las frases proporcionados por AMPER-2006 (López Bobo, Muñiz Cachón, y Díaz Gómez, 2007), los valores de f_0 en semitonos y los colocaba en una matriz de datos comparando cada repetición de una frase en un mismo hablante y entre hablantes distintos considerando tres valores por vocal.

En segundo lugar, un análisis en Excel sobre la salida de Praat aplicaba la fórmula de las distancias escogida. Se escogió como índice de la distancia entonativa entre dos frases, que podemos llamar x e y, la media cuadrática de la diferencia entre los valores de f_0 de la frase x y de la frase y en cada uno de los puntos de medición. Para los dos conjuntos x e y de valores de f_0 $\{f_0x_1, f_0x_2, ..., f_0x_N\}$ y $\{f_0y_1, f_0y_2, ..., f_0y_N\}$, donde N es el número de puntos de medición de f_0 en cada una de las dos frases, mientras que f_0x_i y f_0y_i son los valores de f_0 en semitonos en cada uno esos puntos.

Esta fórmula proporcionaba la distancia entre dos frases con la misma estructura (por ejemplo, entre dos declarativas SVO con sujeto llano, verbo llano y objeto esdrújulo) de dos puntos de encuesta. Para determinar la distancia general entre todas las frases de dos puntos de encuesta, puesto que la distribución de las distancias no es normal, de acuerdo con De Castro Moutinho, Coimbra, Rilliard, y Romano (2011, p. 44) se escogió la mediana de las RMS calculadas por cada pareja de frases x e y. A partir de las medianas de las distancias entre cada par de puntos de encuesta se pudo construir la matriz de distancias correspondiente.

En tercer lugar, la matriz de distancias constituía, a su vez, la base para la fase final del proceso de análisis,

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga

que se efectuaba con SPSS y consistía en lo siguiente. En primer lugar, en un análisis de conglomerados clúster, técnica multivariante cuya finalidad es clasificar los puntos de encuesta en grupos a partir de la semejanza entre sus características entonativas donde como método de comparación se utiliza la media de las distancias entre los grupos tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. El resultado se expresa en forma de dendrograma que permite ver, en forma de árbol invertido, cómo se agrupan los datos hasta el nivel que se considera oportuno. Y en segundo lugar, un análisis de tipo escalamiento multimdimensional (EMD o MDS) que representa bidimensional o tridimensionalmente de forma gráfica las distancias entre los sujetos o puntos de encuesta de la manera más objetiva posible en un espacio virtual. Este método estadístico pretende construir un espacio métrico con el menor número de dimensiones posibles, de tal manera que permite representar las proximidades o preferencias entre los objetos con el mayor grado de fidelidad. Desde un conjunto de objetos se establecen sus propiedades numéricas a partir de las cuales se elaboran las tablas de proximidad (o de similitud) y, finalmente, se trasladan estas proximidades a un espacio, un mapa de objetos (Matas Crespo, 2006). En realidad, ambos tipos de gráficos —dendrogramas y espacios MDS— proporcionan la misma información, y así se puede comprobar en el apartado de resultados. La ventaja de ambas formas de representación es la de permitir captar la distribución y la agrupación de los datos sin necesidad de tener que recurrir a una matriz de distancias numérica de proporciones enormes.

Como medida utilizamos el intervalo de distancia euclidiana (2).

$$(2) \qquad d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

El funcionamiento de Calcu-Dista, como era esperable, ha sido validado estadísticamente mediante la comparación con resultados obtenidos con métodos comparables (Fernández Planas *et al.*, 2015).

5. CARACTERÍSTICAS FUNDAMENTALES DE PRODIS

La idea de partida era: (1) cómo poder establecer una matriz de distancias por informantes y por puntos de encuesta (reuniendo en un bloque los distintos informantes del mismo punto de encuesta) a partir de una matriz inmensa de datos obtenidos en AMPER-2006 en txt de datos en semitonos; y (2) cómo ver reflejadas las matrices de distancias (o de proximidades) de forma gráfica en forma de dendrogramas y de distribución en espacios virtuales.

ProDis realiza la media y la mediana de correlación por informantes y por punto de encuesta. A continuación, a partir de estos datos realiza un análisis de clúster que permite clasificar los informantes o los puntos de encuesta en diferentes grupos, tanto en forma de dendrograma como de distribución en un espacio virtual, según su semejanza o su proximidad.

Concretamente, a partir de los datos en semitonos de la f_0 de las frases, computa las correlaciones entre ellos, analiza la media y la mediana de las correlaciones para cada hablante, construye la matriz de correlaciones entre hablantes y entre puntos de encuesta, lleva a cabo el análisis de clúster entre hablantes y entre puntos de encuesta y prepara las diferentes salidas gráficas (los dendrogramas y EMD), pero también los gráficos que permiten establecer la validación estadística de los datos como los mapas de correlación, de desviación estándar, los gráficos de silueta o los gráficos de Shepard.

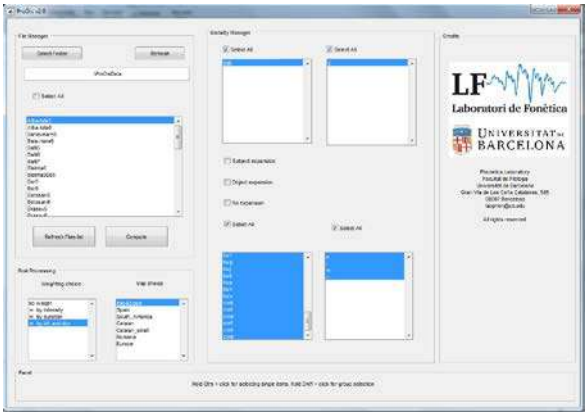
En la Figura 1 aparece una imagen de la interfaz de la herramienta.

Para el cálculo de las correlaciones se utilizan las de Pearson con cuatro métricas diferentes para cada análisis: (1) sin ponderar; (2) ponderadas por la intensidad (a la manera de Hermes, 1998), ponderadas por la duración, y ponderadas por la intensidad y la duración. Existen diferentes métodos para mediar las distancias en dialectometría y la correlación de Pearson (la que hemos estimado mejor) es una de ellas. Se presentó una revisión crítica de ellos en Elvira-García, Balocco, Roseano, y Fernández Planas (2016) y también en Hermes (1998). Serían, entre otros, el algoritmo de Levenstein (Kessler, 1995), la distancia euclidiana (Nerbonne *et al.*, 2011; Roseano *et al.*, 2015); Mahalanobis (Wouters y Macon, 1998), correlación de Spearman (Hermes, 1998), correlación de Pearson (Heeringa y Gooskens, 2003), tau de Kendall (Hermes, 1998).

Para las correlaciones, el programa compara cada repetición de una frase de un hablante de un punto de encuesta con las otras dos repeticiones del mismo informante de la misma frase y con las tres repeticiones de la frase con la misma estructura sintáctica de otro informante. Ello permite obtener una matriz de correlaciones como la que aparece en la Figura 2. Los valores, lógicamente, van de −1 a 1.

Los mismos datos los podemos ver de forma completa y más fácilmente aprehensible en un mapa de correlaciones como el que aparece en la Figura 3. Cabe destacar que en el gráfico en cuestión no aparecen valores numéricos, sino gradaciones de colores que

Figura 1: Interfaz de ProDis.



Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.

Figura 2: Ejemplo de matriz de correlaciones.

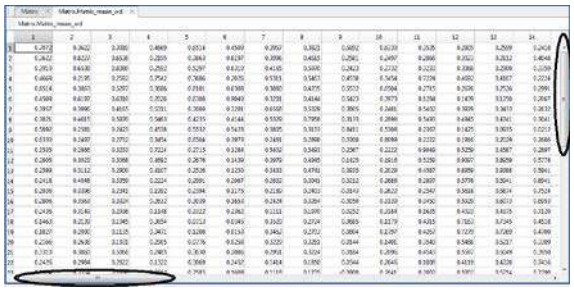


Figura 3: Ejemplo de mapa de correlaciones.

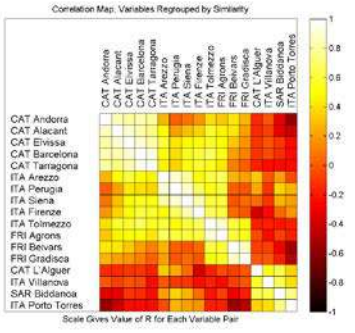
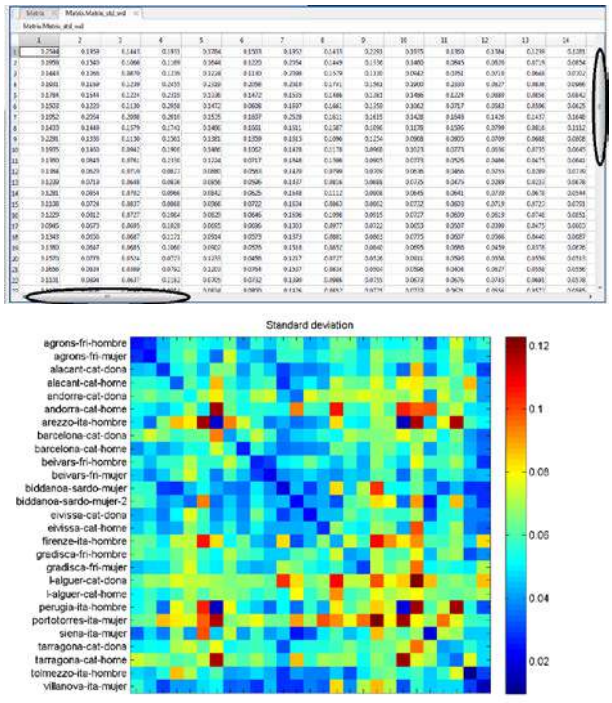


Figura 4: Matriz y mapa de desviación estándar.



indican grados de correlación distintos. De esa manera, la matriz resulta de comprensión más inmediata.

La herramienta ofrece también la matriz y el mapa de desviación estándar por informante y por punto de encuesta (Figura 4), lo cual es muy interesante para comprobar la coherencia intrasujetos y entre sujetos, por un lado, y también intrapuntos de encuesta y entre puntos de encuesta, por otro lado.

Se aprecia fácilmente cómo los puntos más tendentes a rojo (por ejemplo los valores de la voz femenina de L'Alguer en catalán, en el mapa de ejemplo) permiten adivinar que el informante (o el punto) en cuestión es menos homogéneo en los patrones melódicos que utiliza.

ProDis proporciona también gráficos de silueta que, de alguna manera, indican los grupos que matemáticamente sería óptimo establecer en los dendrogramas que agrupan progresivamente los elementos hasta unirlos todos, aunque es la decisión del investigador la que establece el límite de agrupaciones. Para las agrupaciones se toma como referencia el elemento más lejano. Podemos ver un ejemplo de dendrograma en la Figura 5.

El dendrograma funciona con una técnica multivariante que crea clústering aglomerativo jerárquico usando un método de agrupación completo, grupos basados en el elemento más lejano.

El gráfico de Shepard, que se refiere a los gráficos en EMD, nos demuestra cuando tiende a una línea, que los gráficos son fiables y válidos.

En la Figura 6 vemos dos formas de visualizar los gráficos EMC en ProDis: a) en dos dimensiones; b) en tres dimensiones. Aunque el gráfico en tres dimensiones es más recomendable porque el valor de Stress tiende a ser más bajo que en el gráfico en dos dimensiones, en ocasiones puede ser más difícil de interpretar.

Finalmente, mediante ProDis podemos obtener mapas geográficos donde se representen en colores coincidentes con los dendrogramas y con los EMD las localizaciones de los informantes o de los puntos de encuesta. Véase un ejemplo en la Figura 7.

6. CONCLUSIÓN. MEJORAS DE PRODIS RESPECTO A STAT-DISTANCES Y LÍNEAS DE FUTURO

ProDis satisface con creces la idea de partida que señalábamos en el inicio del aparatado anterior, ya que por un lado permite establecer una matriz de distancias por informantes y por puntos de encuesta a partir de los datos numéricos obtenidos en AMPER-2006 y, por otra parte, transforma las matrices de distancias en gráficos de más fácil interpretación. Además, mejora y supera en prestaciones las que ofrecía Stat-Distances porque: (1) considera repertorios de datos no coincidentes en el

Figura 5: Ejemplo de dendrograma con una línea que establece el límite de agrupaciones que se sería óptimo tener en cuenta.

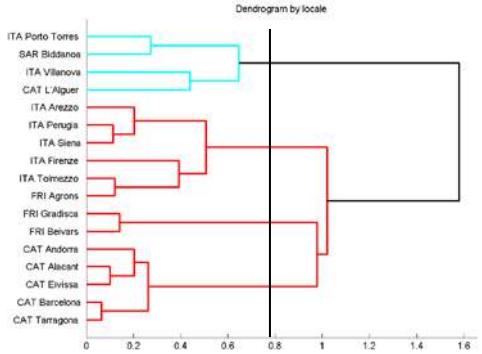


Figura 6: Gráficos EMC en dos dimensiones (arriba) y en tres dimensiones (abajo).

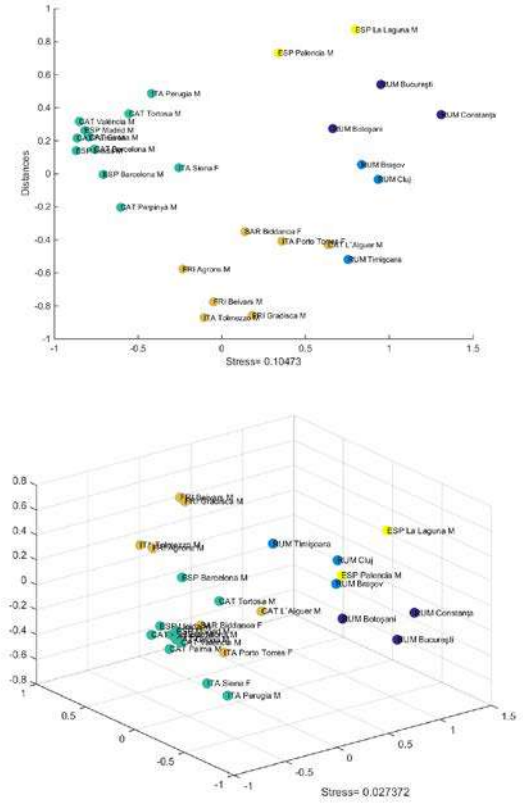
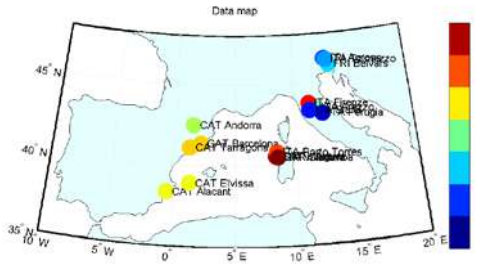


Figura 7: Ejemplo de mapa geográfico con la localización de los puntos de encuesta considerados.



número de sílabas; (2) puede ponderar por duración o por intensidad; (3) permite obtener matrices de confusión y de desviación estándar; (4) es capaz de comparar repertorios con diferente números de frases y diferente composición estructural; (5) puede exportar sus resultados a Excel, SPSS o R; (6) ofrece, entre sus resultados, gráficos de correlación, dendrogramas y MDS; (7) trabaja con semitonos, es decir, con datos normalizados; (8) es relativamente fácil de usar; (9) permite al usuario cierta flexibilidad.

Entre las líneas de desarrollo futuro inmediato, nos proponemos acelerar el software e incorporar EtiToBI (Elvira-García *et al.*, 2016) para llevar a cabo análisis dialectométricos con datos nominales (es decir, a partir de los etiquetajes de los datos en el sistema ToBI).

7. REFERENCIAS

Álvarez Blanco, R., Dubert, F., y Sousa, X. (2006). Aplicación da análise dialectométrica aos datos do Atlas Lingüístico Galego. En R. Álvarez Blanco, F. Dubert, y X. Sousa (Eds.), *Lingua e territorio* (pp. 461–493). Santiago de Compostela: Instituto da Lingua Galega – Consello da Cultura Galega.

Aurrekoetxea, G. (1992). Naffaroako Euskara: azterketa dialektometrikoa. *Utzaro*, *5*, 59–109.

Aurrekoetxea, G., Fernández-Aguirre, K., Rubio, J., Ruiz, B., y Sánchez, J. (2013). DiaTech: A new tool for dialectology. *Literary and Linguistic Computing*, *28*(1), 23–30.

Bauer, R. (2003). Sguardo dialettometrico apoyo alcune zone di transizione dell'Italia norte-orientale (lombardo vs. Trentino vs. Veneto). Parallel X. Sguardi reciprocidad. Vicende linguistiche e cultural dell'area italofona e germanófona. En R. Bombi y F. Fusco (Eds.), *Atti del Decimo Incontro italo-austriaco di linguisti* (pp. 93–119). Udine, Italia: Forum Editrice.

Bauer, R. (2005). La classificazione dialettometrica dei basiletti altoitaliani e Ladino rappresentati nell'Atlante linguistico del ladino dolomitico e dei dialetti limitrofi (ALD - I). En C. Guardiano *et al.* (Eds.), *Lingue, istituzioni, territorio. Riflessioni teoriche, proposte metodologiche ed esperienze di politica linguistica* (347–365). Roma, Italia: Bulzoni.

Boersma, P. y Weenink, D. (2014). Praat: doing phonetics by computer (Versión 5.4.01). Disponible en http://www.praat.org/.

Clua, E. (2004). El método dialectométrico: aplicación del análisis multivariante a la clasificación de las variedades del catalán. En M. P. Perea (Ed.), *Dialectología y recursos informáticos* (pp. 59–88). Barcelona: Universitat de Barcelona.

Contini, M. (1992). Vers une géoprosodie romane. *Actas del Nazioarteko Dialektologia Biltzarra Agiriak, Bilbao 1991* (pp. 83–109). Bilbao: Publicacions de la Real Academia de la Lengua Vasca.

Contini, M., Lai, J. P., y Romano, A. (2002). La géolinguistique à Grenoble: de l'AliR à AMPER. *Revue Belge de Philologie et d'Histoire*, *80*(3), 931–941.

Contini, M., Lai, J. P., Romano, A., y Roullet, S. (1998). Vers un atlas prosodique parlant des variétés romanes. En J.-C. Bouvier *et al.* (Eds.), *Mélanges offerts à X. Ravier* (pp. 73–84). Toulouse: Université de Toulouse – Le Mirail.

Contini, M., Lai, J. P., Romano, A., Roullet, S., De Castro Moutinho, L., Coimbra, R. L., Pereira Bendiha, U., y Secca Ruivo, S. (2002). Un projet d'atlas multimédia prosodique de l'espace roman. *Proceedings of the International Conference Speech Prosody 2002* (pp. 227–230). Aix-en-Provence: Laboratoire Parole et Langage.

D'Andrés Díaz, R., Álvarez-Balbuena García, F., y Suárez Fernández, X. M. (2007). Proxecto ETLEN para o estudio dialectográfico e dialectométrico da zona Eo-Navia, Asturias: fundamentos teóricos. Actas VII Congreso Internacional de Estudos Galegos: mulleres en Galicia: Galicia e os outros pobas da península (pp. 749–759). A Coruña: Ediciós do Castro.

De Castro Moutinho, L., Coimbra, R. L., Rilliard, A., y Romano, A. (2011). Mesure de la variation prosodique diatopique en portugais européen. *Estudios de Fonética Experimental*, *20*, 33–55.

Elvira-García, W. (2014). Calcu-Dista scripts package. Praat script. Disponible en http://stel.ub.edu/labfon/en/praat-scripts

Elvira-García, W., Balocco, S., Fernández Planas, A. M., Roseano, P.. Martínez Celdran, E. (2015). Presentació d'una aplicació informàtica per a l'anàlisi dialectomètrica de dades prosòdiques en el marc de l'Atles Multimèdia de la Prosòdia de l'Espai Romànic. Presentado en el VII Workshop sobre prosodia del catalán, Universitat de Barcelona, 22/06/2015.

Elvira-García, W., Balocco, S., Roseano, P., y Fernández Planas, A. M. (2016). Comparació de mesures de distància prosòdica entre varietats dialectals. Presentado en el VIII Workshop sobre prosodia del catalán, Universitat Pompeu Fabra, 04/07/2016.

Elvira-García, W., Roseano, P., Fernández Planas A. M. y Martínez Celdrán E. (2016). A tool for automatic transcription of intonation: Eti-ToBI a ToBI transcriber for Spanish and Catalan. *Language Resources and Evaluation*, *50*(4), 767–792.

Fernández Planas, A. M. (2005). Datos generales del proyecto AMPER en España. *Estudios de Fonética Experimental*, *14*, 13–27.

Fernández Planas, A. M. (2016a). Aspectos de ProDis, una nueva herramienta para el análisis dialectométrico prosódico. Presentado en el Workshop «Approaches to Sociolinguistic Aspects of Romanian and Spanish Intonation», Alexandru Ioan Cuza University of Iasi (Rumanía), 21/10/2016.

Fernández Planas, A. M. (2016b). Características generales de ProDis (herramienta para analizar distancias prosódicas). Presentado en el Servei de Tractament de la Parla i el So (STPS) de la Universitat Autònoma de Barcelona, 18/11/2016.

Fernández Planas, A. M., Dorta, J., Roseano, P., Díaz, X., Elvira-García, W., Martín Gómez, J. A., Martínez Celdrán, E. (2015). Distancia y proximidad prosódica entre algunas variedades del español: un estudio dialectométrico a partir de datos acústicos. *Revista de Lingüística Teórica y Aplicada*, *53*(2), 13–45.

Fernández Planas, A. M., Roseano, P., Martínez Celdrán, E., y Romera Barrios, L. (2011). Aproximación al análisis dialectométrico de la entonación en algunos puntos del dominio lingüístico catalán. *Estudios de Fonética Experimental*, *20*, 141–178.

Goebl, H. (1981). Eléments d'analyse dialectométrique (avec application à l'AIS). *Revue de Linguistique Romane*, *45*, 349–420.

Goebl, H. (1987). Encore un coup d'oeil dialectométrique sur las Tableaux phonétiques de patois suisses romands (TPPSR). *Vox Romanica*, *46*, 91–125.

Goebl, H. (1993). Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data. En R. Köhler y B. B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 277–315). Dordrecht: Springer.

Goebl, H. (2003). Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur. *Estudis Romànics*, *25*, 60–117.

Goebl, H. y Haimerl, E. (2004). Visual Dialectometry. http://www.dialectometry.com/dmdocs/index.html [28/11/2016].

Goebl, H. y Schiltz, G. (1997). Dialectometrical compilation of CLAE 1 and CLAE 2. Isoglosses and dialect integration. En W. Viereck, H. Ramisch, H. Händler, y C. Marx (Eds.), *The computer developed linguistic Atlas of England*, *Vol. 2* (pp. 13–21). Tübingen: Niemeyer.

Heeringa, W. y Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, *37*(3), 293–315.

Heeringa, W. y Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, *13*, 375–400.

Hermes, D. J. (1998). Measuring the perceptual similarity of pitch contours. *Journal of Speech Language and Hearing Research*, *41*(1), 73–82.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. En *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 60–66).

López Bobo, M. J., Muñiz Cachón, C., Díaz Gómez, L., Corral Blanco, N., Brezmes Alonso, D., y Alvarellos Pedrero, M. (2007). Análisis y representación de la entonación. Replanteamiento metodológico en el marco del proyecto AMPER. En J. Dorta (Ed.), *La prosodia en el ámbito lingüístico románico* (pp. 17–34). Santa Cruz de Tenerife: La Página Ediciones.

Martínez Calvo, A., y Fernández Rei, E. (2015). Unha ferramenta informática para a análise dialectométrica da prosodia. *Estudios de Fonética Experimental*, *24*, 289–303.

Martínez Celdrán, E. y Fernández Planas, A. M. (Coords.) (2003–2016). *Atlas Multimèdia de la Prosòdia de l'Espai Romànic*. http://stel.ub.edu/labfon/amper/cast/index_ampercat.html

Matas Crespo, J. (2006). La técnica del Escalamiento Multidimensional en el vocalismo: un análisis comparativo (Tesis Doctoral). Universitat de Barcelona.

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., y Leinonen, T. (2011). Gabmap: A web application for dialectology. *Dialectologia*. Special issue II, 65–89.

Polanco, L. (1992). Lengua y dialecto: una aplicación dialectométrica a la lengua catalana. *Miscelánea, 3*, 5–28.

Prieto, P., Borràs-Comes, J., y Roseano, P. (Coords.) (2010–2014). Interactive Atlas of Romance Intonation. http://prosodia.upf.edu/iari/.

Rilliard, A. y Lai, J. P. (2008). Outils pour le calcul et la comparaison prosodique dans le cadre du projet AMPER: L'exemple des variétés Occitane et Sarde. En A. Turculeţ (Ed.), *La variation diatopique de l'intonation dans le domaine roumain et roman* (pp. 217–229). Iaşi, Rumanía: Editura Universităţii Al. I. Cuza.

Romano, A. (2003). Un projet d'Atlas multimédia prosodique de l'espace roman (AMPER). En F. Sánchez Miret (Ed.), *Atti del XXIII CILFR, Vol. 1* (pp. 279–294). Tübingen: Niemeyer.

Romano, A. y Contini, M. (2001). Un progetto di Atlante geoprosodico multimediale delle varietà linguistiche romanze. En E. Magno Caldognettoy P. Cosi (Eds.), *Multimodalità e Multimedialità nella Comunicazione. Atti delle XI Giornate di Studio del "Gruppo di Fonetica Sperimentale" dell'Associazione Italiana di Acustica* (pp. 121–126). Padova: Unipress.

Roseano, P. (2012). La prosòdia del friulà en el marc de l'Atles Multimèdia de Prosòdia de l'Espai Romànic (Tesis Doctoral). Universitat de Barcelona.

Roseano, P., Elvira-García, W., y Fernández Planas, A. M. (en revisión). Calcu-Dista: A Tool for Dialectometric Analysis of Intonational Variation. En I. Feldhausen, M. M. Vanrell y U. Reich (Eds.), *Empirical Methods in Romance Prosody Research.* Language Science Press.

Roseano, P., Fernández Planas, A. M., Elvira-García, W., Cerdà Massó, R., y Martínez Celdrán, E. (2015). Contacto lingüístico y transferencia prosódica: El caso del alguerés. *Dialectologia et Geolinguistica*, *23*(1), 95–123.

Saramago, J. (2002). Diferenciação lexical interpontual nos territórios galego e português (Estudo dialectométrico aplicado a materiais galegos do ALGa). En R. Álvarez, F. Dubert García, y X. Sousa Fernández (Eds.), *Dialectoloxía e léxico* (pp. 41–68). Santiago de Compostela: Instituto da Lingua Galega – Consello da Cultura Galega.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane*, *37*, 1–24.

Sousa, X. (2006). Análise dialectométrica das variedades xeolingüística galegas. En M. C. Rolão Bernardo y H. Mateus Montenegro (Eds.), *Actas do I Encontro de Estudos Dialectológicos* (pp. 345–362). Ponta Delgada, Portugal: Instituto Cultural de Ponta Delgada.

Verlinde, S. (1988). La dialectométrie et la détection des zones dialectales: L'architecture dialectale de l'Est de la Belgique romane. *Revue de Linguistique Romane*, *51*, 151–172.

Wouters, J., y Macon, M. W. (1998). A perceptual evaluation of distance measures for concatenative speech synthesis. En *Proceedings of the Fifth International Conference on Spoken Language Processing*.

Fernández Planas, A. M., Roseano, P., Elvira-García, W., & Balocco, S. (2019). Génesis y aspectos fundamentales de ProDis. In J. M. Lahoz-Bengoechea, R. Pérez Ramón, & J. Villa Villa (Eds.), *Subsidia. Tools and resources for speech sciences* (pp. 125–132). Málaga: Universidad de Málaga.