UNHA FERRAMENTA INFORMÁTICA PARA A ANÁLISE DIALECTOMÉTRICA DA PROSODIA

A COMPUTING TOOL FOR THE DIALECTOMETRIC ANALYSIS OF PROSODY

ADELA MARTÍNEZ CALVO
Universidade de Santiago de Compostela
adela.martinez@usc.es

ELISA FERNÁNDEZ REI
Universidade de Santiago de Compostela
elisa.fernandez@usc.es

Artículo recibido el día: 07/04/2015
Artículo aceptado definitivamente el día: 26/05/2015
Estudios de Fonética Experimental, ISSN 1575-5533, XXIV, 2015, pp. 289-303

RESUMEN

Neste traballo preséntase unha ferramenta informática, desenvolvida co software estatístico R, que lle permite ao usuario realizar unha análise dialectométrica dos datos prosódicos do corpus fixo recollido no proxecto AMPER. Non se precisa coñecementos previos no uso do software R, polo que a ferramenta pode ser empregada por calquera usuario interesado neste tipo estudos. A análise dialectométrica realizada pola ferramenta inclúe o cálculo de correlacións entre curvas de F0 e a obtención das distancias prosódicas entre as distintas localizacións (falantes ou outra variable de interese) existentes no corpus. Unha vez construída a táboa de distancias prosódicas, o usuario pode aplicar mediante a ferramenta técnicas da estatística multivariante, como o escalado multidimensional (MDS) e a análise de conglomerados. Con esta última metodoloxía, o usuario pode detectar agrupamentos nas localizacións (falantes ou outra variable de interese) segundo a súa proximidade en termos de distancia prosódica.

Palabras clave: *análise de conglomerados, correlación, dialectometría, distancia prosódica, MDS, proxecto AMPER.*

ABSTRACT

This paper presents a computing tool, developed with the statistical software R, which allows a user to perform a dialectometric analysis of the prosodic data in the fixed corpus collected in the AMPER project. Since no previous knowledge about using R is assumed, this tool can be used by any user interested in such studies. The dialectometric analysis that is performed includes calculation of correlations between F0 curves and of prosodic distances between different locations covered by the corpus (for speakers or any other variable). Once the table of prosodic distances has been generated, the user can then apply multivariate statistical techniques to it, such as multidimensional scaling (MDS) and cluster analysis. The latter allows the user to detect the existence of clusters of locations (by speaker or another variables) according to their closeness in terms of prosodic distance.

Keywords: *cluster analysis, correlation, dialectometry, prosodic distance, MDS, AMPER project.*

1. INTRODUCIÓN

Nos últimos anos realizáronse diversas investigacións dirixidas a analizar e comparar os trazos prosódicos en distintas variedades lingüísticas. Un exemplo da importancia desta liña de traballo é o proxecto internacional AMPER que ten como obxectivo principal o estudo e a descrición da prosodia dentro do ámbito da Romania (AMPER s.d.). Como parte deste proxecto elaborouse un corpus fixo para cada unha das linguas románicas consideradas cunha metodoloxía común. Deste corpus, constituído por unha serie de estruturas sintácticas en que se controlan variables como o número de sílabas ou a posición acentual, extráense os valores de frecuencia fundamental (F0), duración e enerxía para permitir a comparación da entoación nas diferentes variedades lingüísticas.

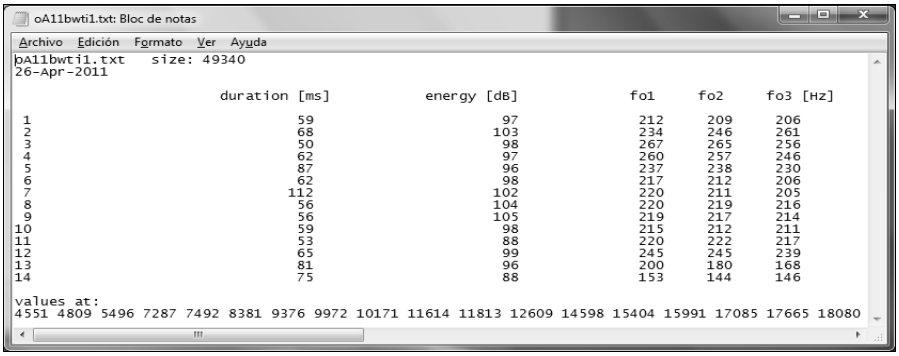
Neste traballo preséntase unha ferramenta informática para a análise dialectométrica deste tipo de datos prosódicos desenvolvida co software libre R (R Core Team, 2014). A ferramenta, actualmente na súa última fase de desenvolvemento, estará dispoñible na web do Instituto da Lingua Galega (http://ilg.usc.es/) a finais do ano 2015. Para a súa utilización a nivel usuario, non é necesario ter coñecementos previos no manexo do software R, xa que a ferramenta vai guiando ao usuario en cada paso da execución, e só se precisa instalar R no equipo (con certos paquetes adicionais) e cargar os arquivos *.R da ferramenta. A ferramenta informática permite:

1. Cargar os datos prosódicos do corpus fixo do proxecto AMPER e realizar unha análise descritiva das variables F0, duración e enerxía.
2. Obter e representar graficamente as correlacións (ponderadas pola enerxía) entre as curvas da F0 e as distancias prosódicas asociadas.
3. Calcular e representar graficamente o escalado multidimensional (MDS) derivado das distancias prosódicas.
4. Realizar unha análise de conglomerados baseada nas distancias prosódicas e representar nun dendrograma os agrupamentos detectados.

A continuación descríbense con máis detalle cada unha das funcionalidades da ferramenta. Os exemplos de representacións gráficas das seccións seguintes corresponden á análise dialectométrica do corpus fixo recollido no proxecto AMPER-Galicia (AMPER-Galicia s.d.).

2. CARGA DOS DATOS PROSÓDICOS

O usuario pode cargar na ferramenta os datos prosódicos de forma rápida e sinxela a partir dos arquivos *.txt formatados segundo a estandarización do proxecto AMPER. Cada un destes arquivos contén a F0 (medida en Hz), a duración (medida en ms) e a enerxía (medida en dB) para cada vogal da estrutura sintáctica á que corresponden os datos prosódicos. Na figura 1 pode verse un exemplo de arquivo *.txt con este formato.


oA11bwti1.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

oA11bwti1.txt size: 49340
26-Apr-2011

	duration [ms]	energy [dB]	fo1	fo2	fo3 [Hz]
1	59	97	212	209	206
2	68	103	234	246	261
3	50	98	267	265	256
4	62	97	260	257	246
5	87	96	237	238	230
6	62	98	217	212	206
7	112	102	220	211	205
8	56	104	220	219	216
9	56	105	219	217	214
10	59	98	215	212	211
11	53	88	220	222	217
12	65	99	245	245	239
13	81	96	200	180	168
14	75	88	153	144	146

values at:
4551 4809 5496 7287 7492 8381 9376 9972 10171 11614 11813 12609 14598 15404 15991 17085 17665 18080

Figura 1. *Exemplo de arquivo *.txt formatado segundo a estandarización do proxecto AMPER (datos do AMPER-Galicia).*

Cómpre sinalar que, ademais de obter os datos prosódicos do contido dos arquivos, a ferramenta tamén é quen de extraer a información codificada no propio nome do arquivo, formado por nove caracteres onde:

1. Os caracteres 1 a 3 indican a localización do falante.
2. O carácter 4 é un número enteiro que indica o sexo do falante (par: home; impar: muller).
3. Os caracteres 5 a 7 indican a estrutura sintáctica e acentual da frase.
4. O carácter 8 indica a modalidade (a: afirmativa; i: interrogativa; n: negativa; m: negativa interrogativa).
5. O carácter 9 é un número enteiro que indica o número de repetición.

Esta información adicional permite, por exemplo, que o usuario poida realizar a análise dos datos segundo a localización xeográfica dos informantes, o sexo do falante, a modalidade oracional, a estrutura acentual ou calquera outra variable de interese que poida ser identificada a partir da codificación dos nomes dos arquivos.

Para poder traballar adecuadamente coas estruturas sintácticas presentes nos datos prosódicos a analizar, é preciso que a ferramenta teña acceso tamén a dous arquivos *.csv auxiliares relativos ás estruturas sintácticas e ás localizacións, respectivamente. O primeiro deles é un arquivo *.csv con tres columnas de datos que debe conter os códigos das estruturas sintácticas presentes nos datos (columna *cod*), a lingua á que corresponden os datos (columna *ling*) e o código ANEPVANE da estrutura sintáctica (columna *anepvane*). O código ANEPVANE dunha estrutura sintáctica constrúese indicando o número de sílabas que ten para cada elemento da estrutura básica das frases consideradas no corpus fixo do proxecto AMPER: (Determinante + Nome + Extensión) + (Partícula + Verbo) + (Determinante + Nome + Extensión). Por exemplo, no caso do galego, a estrutura sintáctica *fwt* ("o médico pequeno falaba co cabalo") codifícase como *A1N3E3 P0V3 A1N3E0*. O segundo arquivo *.csv requirido debe conter, polo menos, dúas columnas de datos correspondentes ao código da localización (columna *cod*) e á lingua dos datos prosódicos rexistrados nese punto de enquisa (columna *ling*).

Por outra banda, se o usuario posúe un mapa da zona na que se sitúan as localizacións dos datos prosódicos en formato multiarquivo shapefile, poderá cargalo a través da ferramenta e obter algunhas representacións xeográficas dos resultados obtidos durante a análise dialectométrica. O multiarquivo shapefile é un formato de arquivo desenvolvido pola compañía ESRI (ESRI, 1998) e de uso común para o almacenamento de datos espaciais e xeográficos. Este formato pode constar de varios arquivos informáticos, dos cales son imprescindibles tres con extensións *.shp, *.shx e *.dbf, respectivamente. Poden obterse mapas neste tipo de formato para uso non comercial en Internet. Por exemplo, a base de datos GADM (GADM, s.d.) recolle información de bases de datos espaciais de varios organismos gobernamentais e doutras organizacións, todos elas dispoñibles en Internet.

Para as representacións xeográficas, ademais do mapa, é preciso que o arquivo auxiliar *.csv correspondente ás localizacións conteña dúas columnas de datos, a maiores das que se indicaron anteriormente, onde o usuario debe ter introducidas as coordenadas de lonxitude e latitude de cada localización (columnas *lonx* e *lat*) no mesmo sistema de referencia empregado no multiarquivo shapefile. Con esta información, a ferramenta é quen de obter os polígonos de Thiessen (tamén

chamados diagramas de Voronoi ou teselación de Dirichlet). Estes polígonos veñen determinados polas intersecciónss das mediatrices dos segmentos que unen cada par de localizacións consideradas e constitúen un método de interpolación simple e empregado habitualmente na representación espacial de datos non cuantitativos. A combinación destes polígonos co mapa orixinal permite representar algúns dos resultados dialectométricos xerados mediante a coloración de cada polígono segundo a escala de cores de interese en cada momento. A figura 2 amosa un exemplo deste tipo de representacións gráficas.

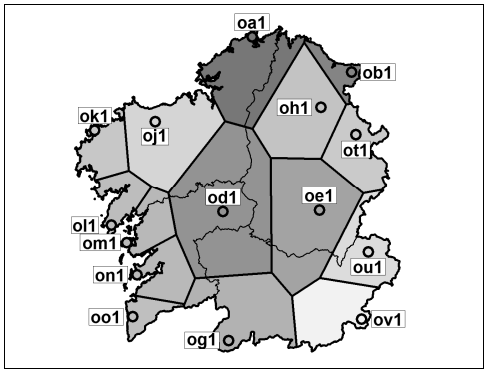


Figura 2. *Exemplo de representación gráfica mediante a combinación dun mapa e os polígonos de Thiessen (puntos de enquisa do AMPER-Galicia).*

Unha vez que todos os datos foron cargados na ferramenta, o usuario pode realizar unha análise descritiva das variables de interese e xerar un resumo estatístico para cada unha delas (mínimo, máximo, media…). Tamén ten a posibilidade de representar graficamente os datos, incluíndo nunha mesma figura, por exemplo, as distintas repeticións dunha estrutura para unha localización (falante ou outra variable de interese) concreta. Na figura 3 vemos todas as realizacións dos enunciados interrogativos seleccionados para a análise en Cariño (punto de enquisa *oa1*), tanto do informante masculino (*oa14*) como da informante feminina (*oa11*). Preséntanse as diferentes estruturas aliñadas pola posición do acento no núcleo do suxeito e na última palabra do complemento: núcleo do suxeito oxítono e última palabra do complemento paroxítona (*ox-pa*), núcleo do suxeito paroxítono e última palabra do complemento proparoxítono (*pa-pr*) etc. Así mesmo, resáltase co borde en negro a mediana de todas as realizacións para cada informante e cada estrutura.

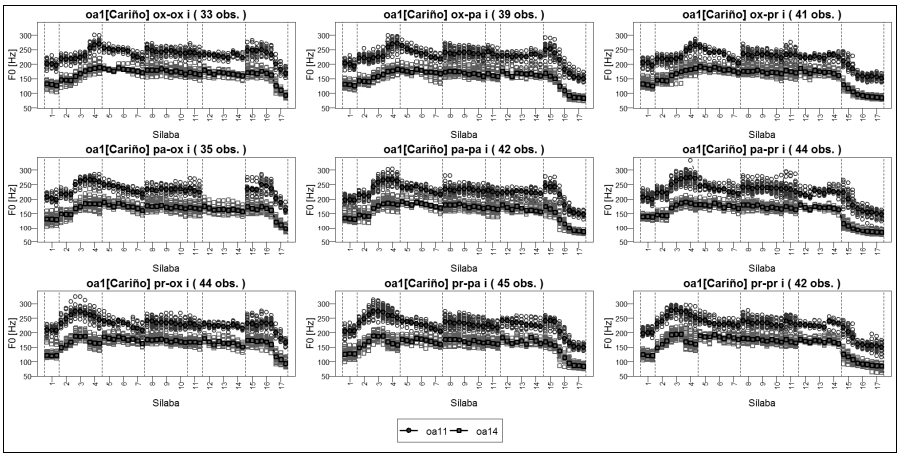


Figura 3. *Curvas de F0 das distintas estruturas acentuais dos informantes (masculino e feminino) de Cariño, A Coruña (datos do AMPER-Galicia).*

3. CORRELACIÓNS PONDERADAS E DISTANCIAS PROSÓDICAS

Para cada par de curvas entoativas coa mesma estrutura e modalidade, a ferramenta informática calcula a medida de correlación entre curvas de F0 ponderada pola enerxía do sinal, definida por:

$$corr(f_1, f_2) = \frac{\sum_i w(i)(f_1(i) - m_1)(f_2(i) - m_2)}{\sqrt{\sum_i w(i)(f_1(i) - m_1)^2 \ \sum_i w(i)(f_2(i) - m_2)^2}}$$

onde f_1 e f_2 son as curvas de F0 para cada unha das frases, m_1 e m_2 os valores medios de F0 para f_1 e f_2, e w a media dos valores de enerxía das dúas curvas (Hermes, 1998; d'Alessandro *et al*, 2011, Moutinho *et al*, 2011). Esta medida de correlación avalía de forma cuantitativa a similitude perceptiva entre dúas curvas de F0.

Aplicando esta medida de correlación ponderada a pares de curvas orixinadas na mesma localización (falantes ou outra variable de interese) pode obterse unha medida da variabilidade asociada a cada localización (falantes ou outra variable de interese). Na figura 4 amósanse, en forma de boxplots, os resultados obtidos para esta medida de variabilidade intra-localización no caso das frases interrogativas. Neste caso, os rexistros asociados a Cangas (*on1*) semellan ser os que presentan unha maior variabilidade.

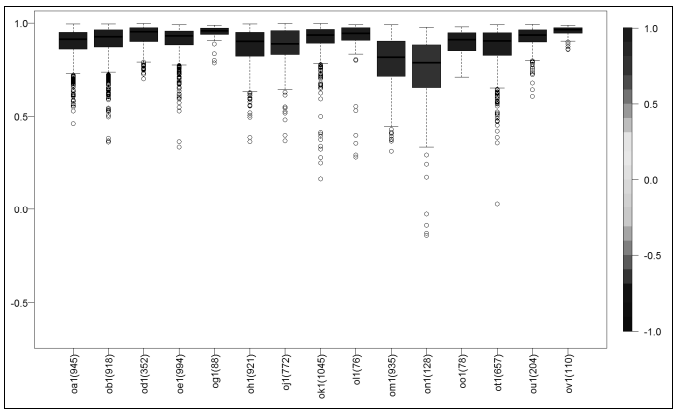


Figura 4. *Exemplo de representación gráfica da variabilidade intra-localización (datos do AMPER-Galicia).*

Por outra banda, a aplicación da medida de correlación a pares de curvas orixinadas en localizacións (falantes ou outra variable de interese) distintas permite definir unha distancia prosódica entre cada par de localizacións (falantes ou outra variable de interese). A ferramenta informática obtén a mediana das correlacións ponderadas entre cada par de curvas coa mesma estrutura e modalidade de dúas localizacións (falantes ou outra variable de interese) diferentes e permite ao usuario xerar a táboa de correlacións correspondente á proximidade prosódica entre as diferentes localizacións e representala graficamente. Na figura 5 vemos un exemplo das correlacións que presentan as interrogativas absolutas en todos os puntos de enquisa do AMPER-Galicia. Como se indica na lenda, a escala de cores indica a correlación (canto máis cálida sexa a cor, máis proximidade; canto máis fría, menor proximidade). Deste xeito, podemos comprobar que hai un importante grupo de puntos de enquisa, os que se corresponden co denominado *galego común*,

que presentan entre eles unha alta correlación (desde *oa1* ata *ov1*); un segundo grupo, os das Rías Baixas, que presenta unha correlación tamén bastante alta entre eles (*om1*, *on1* e *oo1*), pero máis baixa con respecto ao *galego común* (especialmente Oia, *oo1*); e, finalmente, un punto de enquisa, Camelle (*ok1*), con menor correlación cos puntos de enquisa do *galego común*, pero tamén con baixa correlación coas variedades das Rías Baixas.

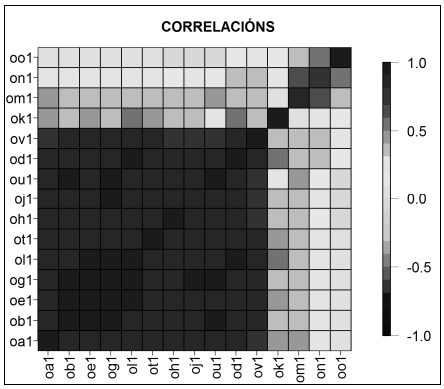


Figura 5. *Exemplo de representación gráfica da táboa de correlacións ponderadas (datos do AMPER-Galicia).*

A partir da táboa de correlacións ponderadas, a ferramenta permite construír a táboa de distancias prosódicas asociada empregando algunha das medidas dispoñibles: euclidiana, supremo, Manhattan, Canberra ou Minkowski. Entre elas, a máis empregada é a distancia euclidiana, que é a distancia cadrática usual entre dous vectores de datos numéricos. Nese caso, para cada par de localizacións (falantes ou outra variable de interese), a distancia prosódica entre elas calcúlase como:

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

onde os vectores $x = (x_1,...,x_n)$ e $y = (y_1,...,y_n)$ conteñen, respectivamente, as correlacións ponderadas de dúas seleccionadas con todas as n existentes no conxunto de datos prosódicos. Unha vez xerada a táboa de distancias prosódicas,

coma no caso da táboa de correlacións ponderadas, a ferramenta ofrece a posibilidade de obter unha representación gráfica da mesma. Na figura 6 vemos agora os mesmos datos ca na figura 5, as interrogativas recollidas nos distintos puntos de enquisa, pero neste caso represéntanse as distancias entre elas. Os resultados indícannos, con máis claridade aínda, que as variedades do *galego común* presentan pouca distancia prosódica entre elas e tamén que se distancian de todas as outras variedades, tanto de Camelle (*ok1*) coma das Rías Baixas (*om1*, *on1* e *oo1*).

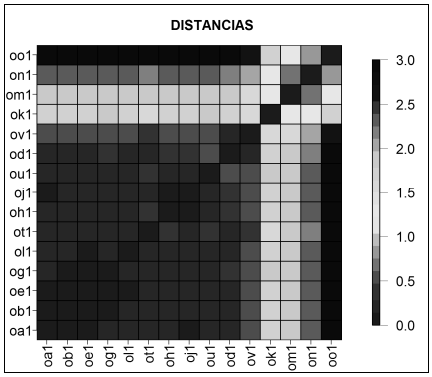


Figura 6. *Exemplo de representación gráfica da táboa de distancias prosódicas (datos do AMPER-Galicia).*

4. ESCALADO MULTIDIMENSIONAL

O escalado multidimensional (MDS pola súa denominación inglesa: MultiDimensional Scaling) é o nome que reciben unha serie de técnicas usadas para visualizar o nivel de similitude entre os elementos dun conxunto de datos. A clave destes métodos é considerar unha táboa de disimilitudes (por exemplo, unha táboa de distancias) entre os datos e obter un conxunto de puntos tales que as distancias entre eles aproximan as disimilitudes da táboa inicial. Se as dúas ou tres primeiras dimensións MDS aproximan ben as disimilitudes iniciais, as coordenadas MDS poden empregarse para facer gráficos de dispersión ou outro tipo de figuras que faciliten a detección de patróns na táboa de disimilitudes orixinal. Isto permite ás veces atopar factores asociados aos datos que inflúen (ou incluso determinan) nas disimilitudes que se observan entre eles.

Partindo da táboa de distancias prosódicas, o usuario pode calcular o MDS multidimensional clásico, tamén chamado análise de coordenadas principais (Gower, 1966; Mardia, 1978), e representar graficamente as localizacións (falantes ou outra variable de interese) empregando as súas coordenadas en termos das primeiras dimensións MDS obtidas. As coordenadas nas dúas primeiras dimensións MDS das distancias xa presentadas na figura 6 represéntanse agora na figura 7. Neste gráfico, o grupo de puntos situados á esquerda son os que se corresponden co *galego común*; os tres puntos que se atopan no cuadrante superior dereito son os das Rías Baixas; e, finalmente, o punto que se atopa separado tanto dese primeiro como desoutro segundo grupo é Camelle. A representación a través do MDS permítenos situar de maneira máis clara a este punto illado, Camelle, posto que ilustra moi ben como se atopa fóra dos outros dous grupos antes mencionados.

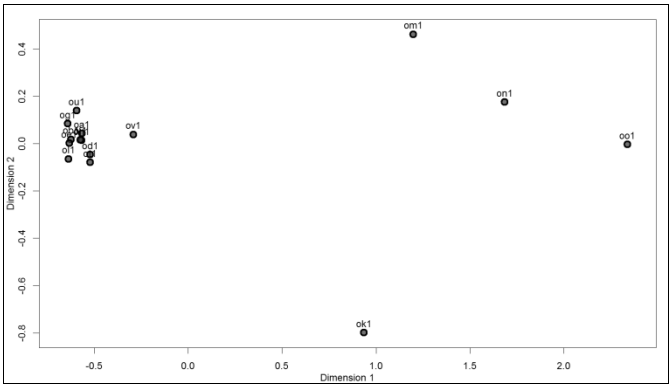


Figura 7. *Exemplo de representación gráfica das dúas primeiras dimensións MDS baseadas na táboa de distancias prosódicas (datos do AMPER-Galicia).*

5. ANÁLISE DE CONGLOMERADOS

A análise de conglomerados (tamén coñecida por análise clúster) é un procedemento que agrupa os datos dun conxunto de tal forma que os datos que caen nun grupo son máis similares aos datos de dito grupo (segundo certo criterio fixado de antemán) que aos datos que pertencen aos outros grupos. En xeral, os

criterios que se empregan para a definición dos grupos están baseados nalgunha medida de similitude (por exemplo, a correlación) ou disimilitude (por exemplo, unha distancia). Ademais existen varios algoritmos para a definición dos grupos que difiren substancialmente tanto na noción de que é un grupo como na forma de construílos eficientemente. Lamentablemente, dado un conxunto de datos, non hai *regras* que suxiran que medida de similitude/disimilitude ou que algoritmo considerar, polo que en cada situación haberá que probar varias combinacións e elixir a que máis se adecúe aos obxectivos do estudo.

Concretamente, a ferramenta informática permítelle ao usuario aplicar un algoritmo de agrupamento xerarquizado divisivo (Kaufman e Rousseeuw, 1990) e varios aglomerativos (Murtagh, 1985) á táboa de distancias prosódicas. Posteriormente, os grupos de localizacións (falantes ou outra variable de interese) detectados mediante esta metodoloxía poden ser representados nun dendrograma. Na figura 8 recóllense os datos xa presentados anteriormente das interrogativas galegas e nela observamos como se realizan os agrupamentos antes xa descritos: o *galego común* (de *oa1* a *ov1*), por un lado, e as Rías Baixas e Camelle por outro lado (se ben se fan dous subgrupos: un para Camelle e outro para o resto dos puntos). Os resultados desta análise permiten definir os agrupamentos de variedades presentes nos datos do corpus fixo segundo a distancia prosódica que presentan as interrogativas e cartografalos (figura 9).

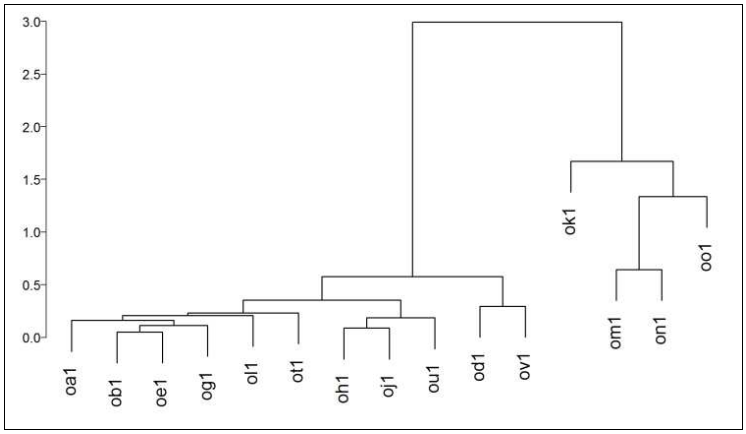


Figura 8. *Exemplo de dendrograma baseado na táboa de distancias prosódicas (datos do AMPER-Galicia).*

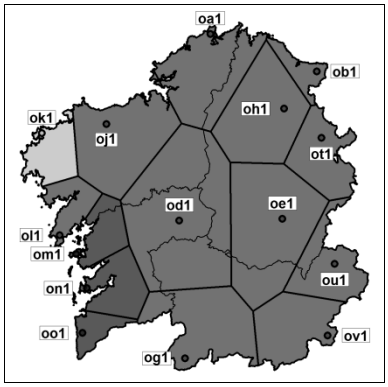


Figura 9. *Exemplo de representación gráfica dos agrupamentos detectados na análise de conglomerados (datos do AMPER-Galicia).*

6. CONCLUSIÓNS

A ferramenta informática que se presenta neste traballo permite ao usuario realizar unha análise dialectométrica dos datos prosódicos recollidos no corpus fixo do proxecto AMPER ou en calquera outro que teña o mesmo formato. A análise inclúe o cálculo das correlacións entre curvas da F0 e das distancias prosódicas entre localizacións (falantes ou outra variable de interese), a obtención do escalado multidimensional (MDS) derivado das distancias prosódicas e a elaboración de dendrogramas a partir da análise de conglomerados.

Aínda que a ferramenta foi desenvolvida co software estatístico R, non se precisan coñecementos previos no uso deste software polo que pode ser empregada por calquera usuario interesado. Na actualidade, estase traballando na elaboración dun manual de axuda da ferramenta que conterá información sobre a instalación e execución da mesma, e una descrición máis ampla de todas as súas funcionalidades e das metodoloxías estatísticas subxacentes.

AGRADECEMENTOS: As autoras queren dar as grazas aos revisores polos seus acertados comentarios e suxestións que contribuíron á mellora substancial do artigo. Este traballo foi financiado pola Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia a través da rede de investigación Tecnoloxías e Análise dos Datos Lingüísticos (rede TecAnDaLi, R2014/007).

7. REFERENCIAS BIBLIOGRÁFICAS

AMPER (s.d.): *Atlas Multimédia Prosodique de l'Espace Roman*. http://dialecto.u-grenoble3.fr/AMPER/amper.htm [31/03/2015].

AMPER-GALICIA (s.d.): *Atlas Multimedia Prosódico del Espacio Románico (AMPER-Gal)*. http://ilg.usc.es/amper/ [31/03/2015].

D'ALESSANDRO, C.; A. RILLIARD e S. LE BEUX (2011): «Chironomic stylization of intonation», *Journal of the Acoustical Society of America*, 129, pp. 1594-1604.

ESRI (1998): *ESRI Shapefile Technical Description. An ESRI White Paper*, Redlands, Environmental Systems Research Institute. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf [24/09/2015].

GADM (s.d.): *GADM database of Global Administrative Areas*. http://www.gadm.org/ [05/05/2015].

GOWER, J. C. (1966): «Some distance properties of latent root and vector methods used in multivariate analysis», *Biometrika*, 53, pp. 325-328.

HERMES, D. J. (1998): «Measuring the perceptual similarity of pitch contours», *Journal of Speech, Language and Hearing Research*, 41, pp. 73-82.

KAUFMAN, L. e P. J. ROUSSEEUW (1990): *Finding Groups in Data*, Nueva York, JohnWiley & Sons.

MARDIA, K. V. (1978): «Some properties of classical multidimensional scaling», *Communications on Statistics – Theory and Methods*, A7, pp. 1233-1241.

MOUTINHO, L. C.; R. L. COIMBRA, A. RILLIARD e A. ROMANO (2011): «Mesure de la variation prosodique diatopique en portugais européen», *Estudios de Fonética Experimental*, XX, pp. 33-55.

MURTAGH, F. (1985): *Multidimensional Clustering Algorithms*, Heidelberg e Viena, Physica-Verlag.

R CORE TEAM (2014): *R: A language and environment for statistical computing*, Viena, R Foundation for Statistical Computing http://www.R-project.org/ [31/03/2015].