

Dofe Man Ariola
S. T. Inguru
Baldoni Sodupe Oñatezulo (732283355)
ezagunen dialektoaren errepresentatuenak eta klasifikazioa eta multzoak (anafisika)
Documento de trabajo 2

Miquel Salicrú

Identificar características representativas y distintivas de los dialectos

Objetivo. Identificar de forma cuantitativa los ítems/formas y las reglas subyacentes que explican la clasificación dialectal.

Componentes del trabajo. a) Conceptualización cuantitativa; b) implementación en software; y c) aplicación en una base de datos lingüística.

Valor diferencial del trabajo. La caracterización dialectal se basa en el barrido manual y sistemático de los ítems que componen el Corpus Dialectal. La automatización del barrido en los ítems que intervienen en la clasificación dialectal se orienta a centrar al investigador en lo importante y, en consecuencia, a simplificar y reducir los tiempos de análisis.

Valorización esperada. Defensa de un TFG, implementación de una función en R o integración en DiaTech y publicación de trabajos en revistas académicas (herramienta y aplicación).

Antecedentes técnicos. En términos cuantitativos, la clasificación dialectal se compone de los siguientes elementos:

1. **Información base.** Tabla de doble entrada (Población/ítem) en la que se explicita los ítems en cada una de las poblaciones (Tabla 1).

	Ítem 1	Ítem 2	Ítem p
Población 1 (P_1)	α_{11}	α_{12}	α_{1p}
Población 2 (P_2)	α_{21}	α_{22}	α_{2p}
:	:	:		:
Población n (P_n)	α_{n1}	α_{n2}	α_{np}

2. **Distancia entre poblaciones.** Para caracterizar las analogías y diferencias entre las poblaciones se considera una medida de distancia. En lingüística, las distancias entre poblaciones (IRI, IPI, Lewenstein, COD,...) evalúan las diferencias en cada ítem

$$d^{(k)}(P_i, P_j) = d^{(k)}(\text{Población } i, \text{ Población } j) = d^{(k)}(\alpha_{ik}, \alpha_{jk}) \quad k=1,2,\dots,p$$

y, si procede, las ponderan por un peso (λ_k). Formalmente, la matriz de interdistancias $D=(d_{ij})$ tiene la siguiente expresión:

$$D = (d_{ij}) = (d(P_i, P_j)) = (\sum_{k=1}^p \lambda_k \cdot d^{(k)}(P_i, P_j))$$

3. **Clasificación dialectal.** En general, la clasificación dialectal utiliza técnicas deterministas jerárquicas (WARD o UPGMA) o no jerárquicas (K-mean) y técnicas no deterministas (fuzzy K-mean). La clasificación determinista proporciona clusters disjuntos y la clasificación no determinista permite poner de manifiesto las regiones de transición.
 4. **Resultado.** Una clasificación en s grupos, donde s es compatible con la maximización de los estadísticos ΔTESS , Silueta y Pseudo-F. En concreto, se dispone de un listado:
- $$G_1=\{P_{11}, P_{12}, \dots, P_{1n_1}\}; G_2=\{P_{21}, P_{22}, \dots, P_{2n_2}\}; \dots; G_s=\{P_{s1}, P_{s2}, \dots, P_{sn_s}\}$$
5. **Representación gráfica.** En geolocalización (Figura 1a) o en un espacio de dimensión reducida (MDS) (Figura 1b)

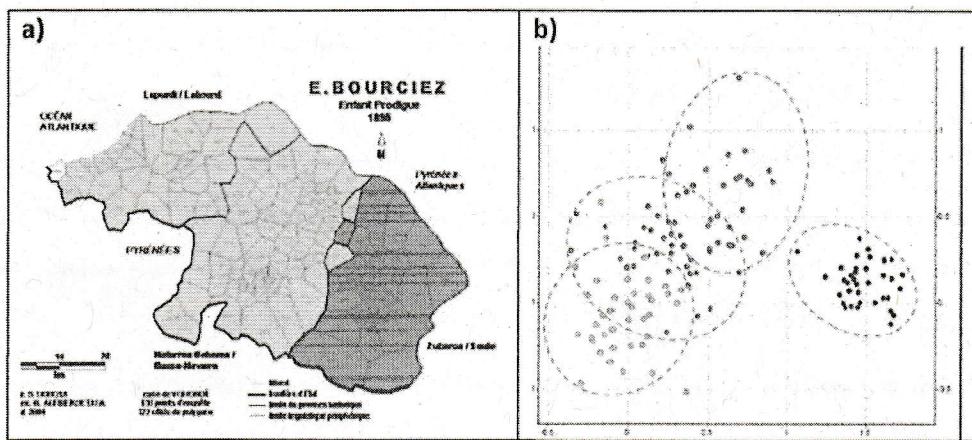


Figura 1. a) Representación geográfica de la clasificación, b) representación MDS.

Problema. Interpretar lingüisticamente los grupos: ¿qué caracteriza los grupos? y ¿en qué se diferencian los grupos G_i y G_j ?

Propuesta conceptual. Planteamos los conceptos de ítem estable dentro de grupo, ítem variable entre grupos y admitimos que los ítems que presentan un comportamiento diferencial entre dos grupos son los ítems estables dentro de sus grupos y variables entre los grupos.

Estabilidad dentro de grupo, variabilidad y diferenciación entre grupos. La estabilidad dentro de un grupo G_i ($G_i=\{P_{i1}, P_{i2}, \dots, P_{in_i}\}$) y la variabilidad y diferenciación entre dos grupos G_i y G_j puede evaluarse de maneras distintas. En este documento proponemos tres alternativas.

Alternativa 1. Para la matriz de interdistancias restringida al ítem h para cada grupo G_i , la media ponderada de las interdistancias entre los elementos que forman el grupo permite cuantificar la estabilidad del ítem en el grupo. Formalmente,

$$Est_{1,h}(G_i) = \frac{2}{n_i(n_i-1)} \sum_{k=1}^{n_i} \sum_{l=k+1}^{n_i} d^{(h)}(P_{ik}, P_{il}) u_{ik} u_{il}$$

siendo u_{ik} y u_{il} las probabilidades de pertenencia de las poblaciones P_{ik} y P_{il} al grupo G_i . La implementación se justifica para minimizar el efecto de las poblaciones atípicas, de transición o frontera.

La variabilidad del ítem h en los grupos G_i y G_j se cuantifica en la media ponderada de las interdistancias entre elementos de grupos distintos,

$$Var_{1,h}(G_i, G_j) = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{r=1}^{n_j} d^{(h)}(P_{ik}, P_{jr}) u_{ik} u_{jr}$$

y la capacidad diferenciadora de un ítem h se ha relacionado con la maximización de la estabilidad dentro de los grupos y de la variabilidad entre grupos. Atendiendo a que la estabilidad dentro de los grupos se maximiza cuando $Est_{1,h}(G_i)$ se minimiza (cero) y la máxima variabilidad entre grupos se maximiza cuando $Var_{1,h}(G_i)$ es máxima, la capacidad discriminatoria se ha relacionado con el cociente entre la variabilidad y la estabilidad,

$$ID_{1,h}(G_i, G_j) = \frac{Var_{1,h}(G_i, G_j)}{\max\{Est_{1,h}(G_i), Est_{1,h}(G_j), 0.000001\}}$$

Remark 1. Cuando se considera la clasificación determinista asociada a lógicas fuzzy, la ponderación (probabilidad de pertenencia) que se aplica en las fórmulas $Est_{1,h}(G_i)$ y $Var_{1,h}(G_i)$ puede obtenerse a partir de la verosimilitud C-mean. En otros casos, será necesario explorar alternativas.

Remark 2. El valor 0.000001 introducido en la fórmula $ID_{1,h}(G_i, G_j)$ se justifica para evitar cocientes en los que el denominador sea cero y, consecuencia, el cociente sea infinito.

Remark 3. Los denominadores $n_i(n_i-1)$ y $n_j(n_j-1)$ que se utilizan en las fórmulas $Est_{1,h}(G_i)$ y $Var_{1,h}(G_i)$ deberán ajustarse si se considera la lógica fuzzy.

Remark 4 (salida computacional 1). Interesados en buscar las características que difieren poco dentro de los grupos (características representativas del grupo), la atención se focalizará en los ítems con valores bajos en $Est_{1,h}(G_i)$. En consecuencia, los ítems más representativos de cada grupo se obtienen ordenándolos de menor a mayor por su valor de estabilidad y escogiendo los primeros de la lista. Salida computacional: listado de ítems por orden creciente en $Est_{1,h}(G_i)$ y valor de estabilidad.

Remark 5 (salida computacional 2). Interesados en buscar las características que difieren poco dentro de los grupos y difieren mucho entre grupos (características distintivas entre dos grupos), la atención se focaliza en los ítems con valores altos en capacidad diferenciadora $ID_{1,h}(G_i, G_j)$. Las diferencias altas entre grupos se obtienen cuando el numerador es un valor alto y las diferencias pequeñas entre grupos se obtienen cuando el denominador es bajo. En consecuencia, los ítems

más diferenciadores entre dos grupos se obtienen ordenándolos de mayor a menor por capacidad diferenciadora y escogiendo los primeros de la lista. Salida computacional 2: listado de ítems por orden decreciente en $ID_{i,h}(G_i, G_j)$ para cada pareja de grupos y valor en capacidad diferenciadora.

Alternativa 2. Para minimizar el efecto de las poblaciones de transición y poblaciones atípicas (sin necesidad de ponderaciones), la estabilidad del ítem h en el grupo G_i puede evaluarse a partir de la mediana de las interdistancias entre los elementos que forman el grupo. La robustez de las medianas hace innecesaria la ponderación por las probabilidades de pertenencia al grupo. Formalmente,

$$Est_{2,h}(G_i) = \text{med}_{k=1, \dots, n_i; j>k} \{d^{(h)}(P_{ik}, P_{il})\}$$

Acorde con este planteamiento robusto, la variabilidad del ítem h en los grupos G_i y G_j se cuantifica en la mediana de las interdistancias entre elementos de grupos distintos,

$$Var_{2,h}(G_i, G_j) = \text{med}_{\substack{k=1, \dots, n_i \\ r=1, \dots, n_j}} \{d^{(h)}(P_{ik}, P_{jr})\}$$

y, por paralelismo, la capacidad diferenciadora de un ítem h se ha relacionado con el cociente entre la variabilidad y la estabilidad,

$$ID_{2,h}(G_i, G_j) = \frac{Var_{2,h}(G_i, G_j)}{\max \{Est_{2,h}(G_i), Est_{2,h}(G_j), 0.000001\}}$$

Remark 6. Este planteamiento resulta especialmente interesante cuando no se dispone de la probabilidad de pertenencia a cada grupo y, particular, cuando la clasificación dialectal es de tipo determinista y esta clasificación no puede relacionarse con la lógica fuzzy. Este planteamiento no está exento de posibles contradicciones: a) la robustez de la mediana puede verse afectada cuando alguno de los grupos tenga un tamaño reducido (número de poblaciones reducido); y b) los indicadores pueden ser no representativos cuando más de la mitad de las interdistancias son iguales a cero (la mediana será cero independientemente del resto de valores).

Alternativa 3. La estabilidad del ítem h en el grupo G_i puede también cuantificarse a partir del porcentaje de distancias entre los elementos que forman el grupo i que son menores o iguales a un valor prefijado C. Formalmente,

$$Est_{3,h}(G_i) = \frac{1}{n_i(n_i-1)} \# \{d^{(h)}(P_{ik}, P_{il}); d^{(h)}(P_{ik}, P_{il}) \leq C, k = 1, \dots, n_i, l > k\}$$

La variabilidad del ítem h en los grupos G_i y G_j es el porcentaje de interdistancias entre elementos de grupos distintos que son mayores al valor prefijado C,

$$Var_{3,h}(G_i, G_j) = \frac{1}{n_i n_j} \# \{ d^{(h)}(P_{ik}, P_{jr}); d^{(h)}(P_{ik}, P_{jr}) > C, k = 1, \dots, n_i, r = 1, \dots, n_j \}$$

y la capacidad diferenciadora de un ítem h se ha relacionado con el cociente entre la variabilidad y la estabilidad,

$$ID_{3,h}(G_i, G_j) = \frac{Var_{3,h}(G_i, G_j)}{\max\{Est_{3,h}(G_i), Est_{3,h}(G_j), 0.000001\}}$$

Remark 7. En este contexto, falta ajustar la formulación con las ponderaciones por pertenencia a los grupos. Por otro lado, la elección de la constante C es subjetiva y poner C=0 puede ser muy exigente.

Remark 8. Para la/s alternativa/s escogida (r=1,2,3), el cálculo anterior proporciona una tabla de doble entrada ítem/grupo,

	G ₁	G _s	G ₁ G ₂	G _{s-1} G _s	G ₁ G ₂	G _{s-1} G _s
Ítem 1	Est _{r,1} (G ₁)	Est _{r,1} (G _s)	Var _{r,1} (G ₁ G ₂)	Var _{r,1} (G _{s-1} G _s)	ID _{r,1} (G ₁ G ₂)	ID _{r,1} (G _{s-1} G _s)
Ítem 2	Est _{r,2} (G ₁)	Est _{r,2} (G _s)	Var _{r,2} (G ₁ G ₂)	Var _{r,2} (G _{s-1} G _s)	ID _{r,2} (G ₁ G ₂)	ID _{r,2} (G _{s-1} G _s)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Ítem p	Est _{r,p} (G ₁)	Est _{r,p} (G _s)	Var _{r,p} (G ₁ G ₂)	Var _{r,p} (G _{s-1} G _s)	ID _{r,p} (G ₁ G ₂)	ID _{r,p} (G _{s-1} G _s)

Dependiendo del objetivo, las columnas se ordenan en orden creciente o decreciente y los listados se proporcionan por parejas de grupos.

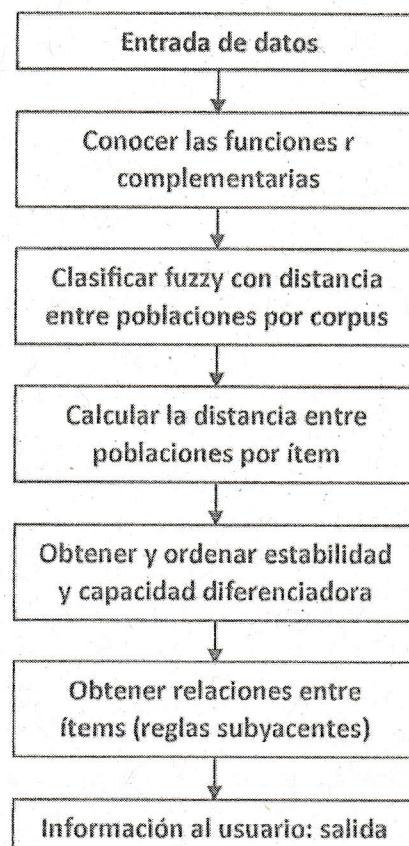
Ítems y reglas lingüísticas subyacentes. Las reglas lingüísticas subyacentes se han relacionado con esquemas de variación que se mantienen en las poblaciones.

verbo (raíz)	cantar	tancar	manar
kan'ta.	βas	βas	βas
	vas	vas	vas
	βes	βes	βes
	ves	ves	ves

EN CONSTRUCCIÓN

Estructura inicial del proyecto: construir una función en R (propuesta inicial)

1. Definir una estructura de fichero ('data frame') que sea compatible con las funciones r a emplear y con DiaTech.
2. Cálcular de la distancia entre dos vectores de caracteres (lingüística, genómica,...). Funciones en r: 'stringdist' y alternativas tienen implementadas distancias textuales. En lingüística, las distancias más empleadas son: Levenshtein y RIV (RIV se puede deducir de Levenshtein: $d_{RIV}(P_i, P_j)=0$ si $d_{Lev}(P_i, P_j)=0$ y $d_{RIV}(P_i, P_j)=1$ si $d_{Lev}(P_i, P_j)>0$). También, aunque menos utilizadas, las distancias WIV (RIV ponderada) y COD. En el ámbito de la genómica, las distancias implementadas en 'stringdist' pueden dar respuesta a las necesidades del investigador.
3. Clúster análisis. Librerías de clasificación determinista: 'mclust', 'ppclust', . Librerías que trabajan con lógica fuzzy 'fclust', 'ppclust',....



*Distancia Levenshtein (información) . https://ast.wikipedia.org/wiki/Distancia_de_Levenshtein ; https://en.wikipedia.org/wiki/Levenshtein_distance

Validación y ejemplo ilustrativo. La validación de resultados (conceptualización y implementación) se realizará con el Vasco Corpus Bourciez. **ESB.**

- meter en DiaTech
- función de R (y que reemplace DiaTech)
- Arriba