**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 de los ítems en cada una de las poblaciones (Tabla 1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Ítem 1** | **Ítem 2** | **….** | **Ítem p** |
| **Población 1 (P1)** | α11 | α12 | …. | α1p |
| **Población 2 (P2)** | α21 | α22 | …. | α2p |
| **⁞** | ⁞ | ⁞ |  | ⁞ |
| **Población n (Pn)** | αn1 | αn2 | …. | αnp |

1. **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)(Pi, Pj)=d(k)(Población i, Población j)=d(k)(αik, αjk) k=1,2,…,p

y, si procede, las ponderan por un peso (λk). Formalmente, la matriz de interdistancias D=(dij) tiene la siguiente expresión:

1. **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.
2. **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:

G1={P11, P12,…., P1n1}; G2={P21, P22,…., P2n2}; …..; Gs={Ps1, Ps2,…., Psns}

1. **Representación gráfica**. En geolocalización (Figura 1a) o en un espacio de dimensión reducida (MDS) (Figura 1b)

|  |  |
| --- | --- |
| **a)** | **b)** |

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

--------------------------------------

**Problema.** Interpretar lingüísticamente los grupos: ¿qué caracteriza los grupos? y ¿en qué se diferencian los grupos Gi y Gj?.

**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 Gi (Gi={Pi1, Pi2,…., Pini}) y la variabilidad y diferenciación entre dos grupos Gi y Gj puede evaluarse de maneras distintas. En este documento proponemos tres alternativas.

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

siendo uik y uil las probabilidades de pertenencia de las poblaciones Pik y Pil al grupo Gi. La implementación se justifica para minimizar el efecto de las poblaciones atípicas, de transición o frontera.

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

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 Est1,h(Gi) se minimiza (cero) y la máxima variabilidad entre grupos se maximiza cuando Var1,h(Gi) es máxima, la capacidad discriminatoria se ha relacionado con el cociente entre la variabilidad y la estabilidad,

*Remark 1.* Cuando se considera la clasificación determinista asociada a lógicas fuzzy, la ponderación (probabilidad de pertenencia) que se aplica en las fórmulas Est1,h(Gi) y Var1,h(Gi) 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 IDi,h(Gi,Gj) se justica para evitar cocientes en los que el denominador sea cero y, el consecuencia, el cociente sea infinito.

*Remark 3*. Los denominadores ni(ni-1) y ninj que se utilizan en las fórmulas Est1,h(Gi) y Var1,h(Gi) deberán ajustarse si se considera la lógica fuzzy. **Ya se ha aplicado a las nuevas fórmulas. Si utilizamos en lugar de , la formulación será consistente con la variabilidad geométrica definida por generalización de la distancia euclídea.**

*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 ítem con valores bajos en Est1,h(Gi). 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 Est1,h(Gi) 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 IDi,h(Gi,Gj). 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 IDi,h(Gi,Gj) 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 Gi 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,

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

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

*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 Gi 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,

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

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

*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,

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **G1** | **····** | **Gs** | **G1G2** | **····** | **Gs-1Gs** | **G1G2** | **····** | **Gs-1Gs** |
| **Ítem 1** | Estr,1(G1) | ···· | Estr,1(Gs) | Varr,1(G1G2) | ···· | Varr,1(Gs-1Gs) | IDr,1(G1G2) | ···· | IDr,1(Gs-1Gs) |
| **Ítem 2** | Estr,2(G1) | ···· | Estr,2(Gs) | Varr,2(G1G2) | ···· | Varr,2(Gs-1Gs) | IDr,2(G1G2) | ···· | IDr,2(Gs-1Gs) |
| **⁞** | **⁞** |  | **⁞** | **⁞** |  | **⁞** | **⁞** |  | **⁞** |
| **Ítem p** | Estr,p(G1) | ···· | Estr,p(Gs) | Varr,p(G1G2) | ···· | Varr,p(Gs-1Gs) | IDr,p(G1G2) | ···· | IDr,p(Gs-1Gs) |

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.

**Entrada de datos**

**Conocer las funciones r complementarias**

**Calcular la distancia entre poblaciones por ítem**

**Obtener y ordenar estabilidad y capacidad diferenciadora**

**Clasificar fuzzy con distancia entre poblaciones por corpus**

**Obtener relaciones entre ítems (reglas subyacentes)**

**Información al usuario: salida**

1. Cálcular de ladistancia 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: dRIV(Pi, Pj)=0 si dLev(Pi, Pj)=0 y dRIV(Pi,Pj)=1 si dLev(Pi, Pj)>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.
2. **Clúster análisis.** Librerías de clasificación determinista: ‘mclust’, ‘ppclust’, . Librerías que trabajan con lógica fuzzy ‘fclust’, ‘ppclust’,….

\*Distancia Levenshein (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.