Factores subyacentes de diferenciación en las lenguas románicas y grado de cercanía entre ellas, mediante el método ACP

© Davius Sanctex (2005)

Contenidos

1.1. Introducción: sobre el Análisis de Componentes Principales (ACP)

Cuando se examinan las calificaciones obtenidas por un cierto número de alumnos en diversas materias se observa por ejemplo que algunas de ellas muestran fuertes correlaciones. Así por ejemplo las calificaciones en matemáticas suelen estar fuertemente correlacionadas con las obtenidas en física y en menor medida en química. Mientras que diversas asignaturas dentro de las humanidades muestran cierta correlación entre sí. Además para la mayoría de alumnos el obtener buenas notas en humanidades no permite predecir si obtendrán también buenas notas en ciencias (algunos alumnos buenos en humanidades también lo son en ciencias pero no otros, y viceversa):
http://www.uoc.edu/in3/emath/docs/Componentes_principales.pdf

Esos hechos pueden ser investigados mediante una técnica estadística de Análisis de Componentes Principales de la varianza (ACP). La técnica permite identificar una serie de factores subyacentes que explican porqué las medidas obtenidas en un determinado item (e.g. matemáticas) aparecen correlacionadas con otro determinado item. Además el ACP permite construir índices que miden aproximadamente cada factor subyacente y cuanta de la información es predictible a partir de los mismos. Volviendo al ejemplo de las calificaciones se observa que existen dos factores o componentes principales F1 y F2. El primero de ellos F1 parece muy positivamente correlacionado con las asignaturas de ciencias, y los alumnos para los cuales el valor de F1 deducible de sus notas alcanza altos valores suelen ser buenos en matemáticas por lo cual F1 es de alguna manera una medida de las capacidades cognitivas implicadas en el razonamiento formal necesario para ciertas asignaturas. En cambio F2 parece más correlacionada con una capacidad cognitiva que permite asimilar correctamente las humanidades. El puro análisis estadístico de las calificaciones mediante ACP permite ver que existen estos dos factores, que cada uno de ellos parece involucrado en capacidades diferentes y que son independientes entre sí (a esto técnicamente se le conoce como condición de ortogonalidad de los componentes principales). Esta independencia significa que un alumno puede tener altos valores de F1 independientemente de F2 y de ahí que no tenga porqué existir en todos los casos una correlación clara entre las notas en ciencias y humanidades.

1.2. ACP: “Genes, pueblos y lenguas”

Otra aplicación de esta técnica la realizó Cavalli-Sforza que analizó la frecuencia de determinados genes en la población europea y las correlaciones que había entre las frecuencias de estos. El ACP proporcionó 5 componentes principales estadísticamente significativos cada uno de los cuales parecía positivamente correlacionado con diversas olas migratorias en Europa: Así F1 se parece a lo que se supone fue el doblamiento neolítico de Europa y el valor de F1 de las diversas poblaciones europeas parece relacionado con la distancia y el tiempo transcurrido desde el foco originario de difusión de al agricultura desde oriente medio. El componente F2 parece correlacionable con desplazamientos de población de norte a sur, que se ha querido identificar con la difusión originaria de pueblos uránicos en Europa. El componente F3 alcanza sus valores máximos en torno a las regiones desde las que se supone que partieron las migraciones de pueblos indoeuropeos y el valor parece variar a lo largo de las rutas seguidas hipotéticamente por estos pueblos indoeuropeos. Los componentes F4 y F5 parecen más difíciles de interpretar. Puede encontrarse más información sobre el ACP de la variación genética europea en:
http://www.angeltowns.com/members/racialreal/genetic_variation.html

2.1 Léxico románico diferencial y clasificación de las lenguas románicas

El problema de clasificación de las lenguas románicas y sobre qué tipo de teoría *Wellenstheorie* (teoría de ondas) o *Stammbaumtheorie* (árbol genealógico) puede resultar más útil para agrupar los subgrupos de lenguas románicas según su parentesco dentro del conjunto de lenguas románicas (de alguna manera se trata de identificar los parientes más próximos entre sí).

En este mensaje se usa la técnica del Análisis de Componentes Principales (ACP) [ver por ejemplo mensaje:] para estudiar cuales son los factores subyacentes que explican por ejemplo porqué dentro de la Romanía portugués y castellano se encuentran entre las lenguas más cercanas o también el occitano y el catalán, mientras que francés y portugués aparecen como relativamente lejanos (aunque ambos presentan desarrollos paralelos pero independientes como el tener vocales nasales en oposición fonológica). El análisis por ACP debe entenderse dentro del grupo de *Wellenstheorien* donde cada factor o componente principal parece coincidir más o menos con un área o centro difusor de innovaciones, la superposición de todas estas innovaciones es lo que daría la variedad interna dentro de las lenguas románicas ya que cada lengua habría sido alcanzada por un número diferente de "olas" o componentes principales.

El punto de partida de este estudio son 147 términos léxicos del vocabulario básico o derivado (excluyendo cultismos) que no son formas léxicas pan-románicas, es decir, son palabras que NO aparecen en todas las lenguas romances sino sólo en unas pocas de ellas. El ACP trata de ver cuales son los factores subyacentes que hacen que una determinada palabra (que puede ser una forma léxica innovadora difundida desde cierto foco difusor) esté en determinado grupo de lenguas. Si hubiéramos usado léxico pan-románico no habríamos observado bien la diferenciación[1], de ahí que necesitemos léxico en el que no todas las lenguas coinciden, el número de estos términos en las lenguas románicas está entorno al 20% del total de formas léxicas latinas heredadas. A continuación se presentan los datos del porcentaje de coincidencias en estas 147 formas léxicas de 7 lenguas románicas: aragonés (ar), castellano (cs), catalán (ct), francés (fr), italiano (sur-central) (it), occitano (oc) y portugués (pt) [hubiera sido deseable tomar también hablas italianas septentrionales o lenguas retorrománicas ya que constituyen excelentes ejemplos de transición entre las lenguas galo-romances y el italiano central estándar; y también habría sido deseable tomar datos del mozárabe y el asturiano-leonés para analizar posibles sub-agrupaciones dentro de las lenguas ibero-romances]. Los porcentajes tomados de una referencia estándar [1] son los siguientes:

Portugués (pt) con:	cs 85,7% / ar 53,7% / ct 21,0% / oc 15,6% / it 26,5% / fr 5,1%
Castellano (cs) con:	ar 59,1% / ct 27,2% / oc 14,2% / it 25,1% / fr <2%
Aragonés (ar) con:	ct 62,5% / oc 48,9% / it 31,9% / fr 29%
Catalán (ct) con:	oc 74,8% / it 33,3% / fr 41,4%
Occitano (oc) con:	it 50,3% / fr 73,4%
Italiano (it) con:	fr 53,7%

[nota: los porcentajes son simétricos y necesariamente la relación de X con Y es un porcentaje idéntico a la relación de Y con X]. Con estos datos expresados en tanto por 1 se puede construir una matriz de coeficientes de correlación como la siguiente:

[1] Mejor dicho, cuando eso se hace aparece un factor subyacente de peso cercano al 50% que participa por igual en todas las lenguas pero no permite identificar factores responsables de la diferenciación dentro de la familia románica. Por lo que para intentar clasificar las lenguas es mejor usar léxico en que las lenguas románicas presentarán diferencias entre sí porque así el método ACP obtiene una mayor resolución.

	pt	cs	ar	ct	oc	it	fr
pt	1,000	0,971	0,374	-0,579	-0,906	-0,545	-0,916
cs	0,971	1,000	0,460	-0,497	-0,896	-0,584	-0,947
ar	0,374	0,460	1,000	0,328	-0,197	-0,606	-0,530
ct	-0,579	-0,497	0,328	1,000	0,737	-0,138	0,302
oc	-0,906	-0,896	-0,197	0,737	1,000	0,306	0,826
it	-0,545	-0,584	-0,606	-0,138	0,306	1,000	0,526
fr	-0,916	-0,947	-0,530	0,302	0,826	0,526	1,000

Que rápidamente ya permite ver cuales son las lenguas más próximas entre sí (las que tienen un coeficiente de correlación más cercano a +1). También puede observarse que el catalán y el aragonés son las lenguas menos divergentes del conjunto, es decir, las que pueden considerarse más representativas del conjunto y que el francés es la más divergente de todas.

El ACP procede encontrando los autovalores y auto-vector de la anterior matriz de coeficientes de correlación (cada auto-vector sirve para identificar uno de los componentes principales que explicaría las correlaciones). El auto-valor correspondiente a auto-valor divido entre la suma total de los autovalores da el peso relativo del factor componente principal como factor explicativo. El análisis identifica 7 componentes explicativos, aunque en general sólo los primeros los de mayor peso pueden ser identificados y correlacionados positivamente con algo interpretable históricamente o lingüísticamente. A continuación se dan los pesos relativos de cada factor y entre paréntesis el peso relativo acumulado de todos los factores de los primeros n factores:

F1	F2	F3	F4	F5	F6	F7
64,5% (64,5%)	25,2% (89,8%)	5,9% (95,7%)	3,2% (98,9%)	0,8% (96,1%)	0,3% (100,0)	0,0% (100,0%)

Como puede verse los 3 primeros factores identificados F1, F2 y F3 tomados en conjunto explican el 95,7% de la variación observada en el léxico diferencial no pan-románico. Como veremos estos 3 factores son fácilmente correlacionables con regiones concretas y posibles centros difusores.

2.2.1. Factor F1 (ibero-romanicidad/galo-romanicidad) [peso relativo 64,5%]

Una vieja discusión dentro de la lingüística románica es la validez y extensión de los subgrupos galo-románico e ibero-románico. A priori uno podría pensar que pudieron existir dos focos innovadores: uno centrado en Hispania y otro en la Galia cada uno de los cuales produjo innovaciones que explican al existencia de estos dos grupos diferenciados. Si esto hubiera sido así habríamos encontrado dos componentes F_m y F_n diferentes asociados a cada uno de estos focos innovadores, pero esto no es lo que encontramos sino que solamente encontramos un factor. Es decir que sólo debió existir un foco innovador y la diferencia entre lenguas galo-románicas e ibero-románicas es una diferencia entre lenguas que fueron más o menos alcanzadas por esta innovación y lenguas que no.

Por lo que sabemos del latín clásico y las lenguas románicas en ciertas áreas léxicas el castellano y el portugués conservan términos más cercanos al latín clásico (antiguo) frente a otras lenguas que han substituido los términos clásicos por otros más recientes. Así podemos pensar que la única innovación que diferencia lenguas ibero-romances de galo-romances es una innovación cuyo foco difusor está dentro de la región galo-romance y las lenguas ibero-romances son lenguas que no fueron alcanzadas por dicha innovación. A continuación ofrecemos una medida de la galo-romanicidad. Las lenguas con un índice F1 muy cercano a +1 son muy galo-romances

en su léxico, mientras que las lenguas cercanas a -1 deben ser consideradas ibero-romances (se ha tomado la normalización de tal manera que la lengua más galo-románica, el francés, tenga el valor +1 y la menos el valor -1). los valores del índice de galo-romanicidad / ibero-romanicidad obtenidos de los datos iniciales son:

Lengua	índice [Gal / Iber]	Coeficiente Correlación
pt	-0,975	-0,985
cs	-1,000	-0,993
ar	-0,275	-0,461
ct	0,483	0,520
oc	0,900	0,919
it	0,522	0,580
fr	1,000	0,946

Puede verse que el índice de galo-romanicidad, debido a su alto peso, explica bastante bien las coincidencias encontradas entre unas lenguas y otras (como se aprecia en el coeficiente de correlación). En base a este índice podemos clasificar las lenguas en 4 grupos:

claramente ibero-romances:	portugués -0,975 / castellano -1,000
débilmente ibero-romances:	aragonés -0,275
débilmente galo-romances:	catalán +0,483 / italiano +0,522
claramente galo-romances:	occitano +0,900 / francés +1,000

2.2.2. Factor F2 (pirenaicidad) [peso relativo 25,2%]

El tercer factor subyacente parece deberse a algo que ha llamado poco la atención de los romanistas y es la similaridad de ciertas innovaciones léxicas del área pirenaica. De hecho este factor divide a las lenguas en lenguas positivamente correlacionadas con F3 y lenguas negativamente correlacionadas con F3. El valor en concreto del índice F3 para cada una de las lenguas no parece demasiado interesante (hasta donde yo alcanzo a observar sólo hay esto: aragonés, Catalán y occitano presentan valores del factor “pirenaicidad románica” o factor F3 positivos, mientras que el resto de lenguas presenta valores negativos, el valor de F3 parece ser tanto menor cuanto mayor es la distancia al área pirenaica). En lugar de reproducir el valor de F3 reproduzco el coeficiente de correlación entre la variación observada de las coincidencias de cada lengua con las otras y el valor predicho sólo a partir de F3:

Factor F3 largamente irrelevante:	pt -0,064 / cs 0,030 / fr -0,150
Factor F3 positivamente relevante:	ar 0,714 / ct 0,808 / oc 0,316
Factor F3 negativamente relevante:	it -0,670

Es decir, en el conjunto de lenguas pirenaicas el factor es importante y parece explicar gran número de coincidencias (tal vez podríamos asumir que durante el imperio esta era una región particularmente bien comunicad entre sí y las innovaciones léxicas se difundieron en la región). El que el italiano sea negativamente irrelevante significa que al haber incluido en nuestra muestra de 7 lenguas 3 lenguas pirenaicas, el italiano parece claramente diferente de estas por no compartir la pirenaicidad, tal vez si hubiéramos incluido un mayor número de lenguas románicas para el italiano F3 habría mostrado ser más bien irrelevante como sucede para el francés, el castellano o el portugués donde observamos coeficientes de correlación < 0,30 por lo que podemos considerar la correlación de estas lenguas con la

pirenaicidad estadísticamente no-significativos (es decir la pirenaicidad no ayuda predecir nada sobre las correspondencias léxicas de estas lenguas).

2.2.3. Factor F3 (hispano-italianidad) [peso relativo 5,9%]

El último factor identificado que parece tener un peso pequeño (y por tanto está en la frontera de ser considerado no significativo) parece deberse a los factores que acercaban el léxico peculiar de la península italiana con Hispania (en particular con el valle del Ebro y alrededores). El factor 3 tiene correlación negativa con las coincidencias léxicas de las lenguas claramente galo-románicas mientras que mantiene correlaciones positivas con las lenguas de Hispania (aunque sólo parece razonablemente alta cuando se compara con el aragonés).

Históricamente no parece fácil de interpretar este factor. Podría pensarse que la identificación como potencial factor explicativo por el método de ACP es sólo una anomalía estadística (típica de los factores con bajo peso, que generalmente son poco o nada significativos responden a peculiaridades de los datos de partida). Si lo incluyo como factor de cierta relevancia es porque parece que otras medidas estadísticas diferentes de la usada parecen detectar el mismo factor (aunque también con bajo peso). Sin embargo no hay que exagerar la importancia de este factor, dadas la bajas correlaciones que presenta con la variación de coincidencias de las lenguas:

Factor F3 claramente irrelevante:	pt -0,050 / cs 0,101 / ct 0,0209
Factor F3 posiblemente irrelevante:	oc -0,165 / fr -0,288
Factor F3 posiblemente relevante:	ar 0,314 / it 0,478

Estos bajos valores delatan la baja capacidad explicativa del factor, excepto para el italiano (lo que en realidad podría significar sólo que en la muestra elegida que consta casi exclusivamente de lenguas románicas occidentales el italiano es peculiar al ser la única lengua románica oriental).

2.2.4. Factores F4, F5, F6, F7 (estadísticamente no significativos)

En general los valores de estos factores son difíciles de interpretar y tienen pesos estadísticos en explicar la variación total pequeña (por lo que en gran parte podríamos considerar que se trata de factores aleatorios o fantasmas estadísticos sin significación para la clasificación de lenguas romances). Cuando se analizan datos procedentes de una muestra N elementos el método del ACP da también N factores explicativos que juntos explican el 100% de la variación. Sin embargo los factores con pesos más pequeños casi siempre son atribuibles a accidentes residuales inexplicables y peculiares de la muestra, por lo que sólo los primeros factores identificados (los de mayor peso relativo) representan factores realmente explicativos. Eso es lo que parece suceder con los factores analizados aquí donde F4, F5, F6 y F7 parecen ser “factores espurios” y no significativos ni estadística ni explicativamente.

Por ejemplo el **factor F4** (“hispanicidad general”) tiene muy bajo peso estadístico es difícil de interpretar, por lo que posiblemente no refleje ningún factor real de diferenciación, parece tener valores de correlación moderadamente positivos con portugués [r = 0,416] y castellano [r = 0,458] y claramente moderadamente negativos con francés [r = -0,654] y occitano [r = -0,302] . El patrón de reparto no es exactamente igual al de la diferenciación entre ibero-romance y galo-romance (factor 2) ya que aquí todas las lenguas hispánicas tienen correlación positiva con F4 y las exteriores a Hispania correlación negativa (aunque sólo en los casos mencionados parece la correlación significativa).

El **factor F5** es similar al anterior pero parece agrupar de un lado a portugués aragonés, castellano y portugués frente al resto de lenguas, pero el factor parece significativo nuevamente para solo las lenguas en extremos opuestos según el valor de F5 (portugués y castellano, muestran con F5 correlaciones positivas moderadas, y el francés correlación negativa moderada, el resto de lenguas presentan correlaciones positivas o negativas, pero cercanas a cero por lo que este factor no parece explicativo para ellas).

Los **factores F6 y F7** tienen pesos relativos tan ridículamente pequeños que con total seguridad reflejan peculiaridades estadísticas de la muestra.

A.1. Anexo 1 Datos numéricos

El método ACP identificar cuantos factores subyacentes significativos estadísticamente explican otro conjunto de *m* índices o factores aparentes de una población de *N* individuos. Los datos de entrada tienen comúnmente la forma:

	Individuo 1	Individuo 2	…	Individuo *N*
índice 1	f_1^1	f_1^2	…	f_1^N
índice 2	f_2^1	f_2^2	…	f_2^N
…	…	…	…	…
índice *m*	f_m^1	…	…	f_m^N

Donde fi^j es el valor del índice *i* medido para el individuo *j*. El objetivo es detectar los factores subyacentes F_k que expliquen como varían los valores de los *m* índices.

En nuestro caso haremos algo especial tomando $m = N = 7$; como población tomaremos 7 lenguas románicas (portugués, castellano, aragonés, catalán, occitano, italiano y francés), como índices a explicar tomaremos el porcentaje de coincidencias en una lista prefijada de vocabulario, así:

Índice 1: porcentaje de coincidencia de la lista con el portugués
Índice 2: porcentaje de coincidencia de la lista con el castellano
Índice 3: porcentaje de coincidencia de la lista con el aragonés
Índice 4: porcentaje de coincidencia de la lista con el catalán
Índice 5: porcentaje de coincidencia de la lista con el occitano
Índice 6: porcentaje de coincidencia de la lista con el italiano
Índice 7: porcentaje de coincidencia de la lista con el francés

La lista como se ha dicha está formada por 147 términos procedentes del latín, que no aparecen en toda lengua [es decir del léxico de origen tomamos sólo los términos que no aparecen en todas las lenguas, ya que los términos que aparecen en todas las lenguas no sirven para la comparación ya que no pueden diferenciar entre ellas. Un buen elejemplo lo constituyen los siguientes pares: *metus*/*pavor*, *fervere*/*bullire*, *frater*/**germanus* o *caput*/*testa* de los cuales las lenguas románicas han escogido bien, bien el otro). La tabla de entrada para los porcentajes de coincidencias es en nuestro caso:

	pt	cs	ar	ct	oc	it	fr
pt	100,0	85,7	53,7	21,0	15,6	26,5	5,1
cs	85,7	100,0	59,1	27,2	14,2	25,1	2,0
ar	53,7	59,1	100,0	62,5	48,9	31,9	29,0
ct	21,0	27,2	62,5	100,0	74,8	33,3	41,4
oc	15,6	14,2	48,9	74,8	100,0	50,3	73,4
it	26,5	25,1	31,9	33,3	50,3	100,0	53,7
fr	5,1	2,0	29,0	41,4	73,4	53,7	100,0

El objetivo del ACP es ver cosas como porqué por ejemplo la columna de Pt es notablemente parecida a la de cs, mientras que la de Pt es marcadamente diferente de la de fr. Se supone que cuando dos columnas son altamente parecidas (muestran fuerte correlación positiva) es porque hubo algún factor histórico común a la zona de desarrollo de las dos lenguas que mantuvo una estrecha vinculación en el desarrollo de ambas y de ahí sus similaridades y variación de coincidencia con respecto a las otras lenguas. Cada componentes principal (CP) representará en principio un factor subyacente importante a la hora de explicar la coincidencia. Se verá que si bien el ACP da como resultado, para una matriz de entrada con *m* índices, 7 posibles

componentes principales (CP) o factores explicativos sólo aquellos con algún alto peso estadístico suelen ser interpretables y representan factores reales. En el caso presente hemos encontrado que los dos principales factores juntos ya explican prácticamente el 89% de la variación observada, y los 3 primeros en conjunto pueden explicar prácticamente el 96% de los datos (los 2 primeros factores parecen interpretables).

El siguiente paso en el método de ACP es construir la matriz de coeficientes de correlación $R = [\, r_{jk} \,]_{j,k\,=\,1\ldots\,m}$ de acuerdo con la fórmula:

$$r_{ij} = \frac{cov(f_i, f_j)}{\sqrt{var(f_i)\,var(f_j)}} \qquad [1]$$

Es decir, se calculan los coeficientes de correlación entre las diversas columnas de la matriz de porcentajes y se construye la siguiente matiz simétrica [S_1]:

$[S_1] =$

1,000	0,971	0,374	-0,579	-0,906	-0,545	-0,916
0,971	1,000	0,460	-0,497	-0,896	-0,584	-0,947
0,374	0,460	1,000	0,328	-0,197	-0,606	-0,530
-0,579	-0,497	0,328	1,000	0,737	-0,138	0,302
-0,906	-0,896	-0,197	0,737	1,000	0,306	0,826
-0,545	-0,584	-0,606	-0,138	0,306	1,000	0,526
-0,916	-0,947	-0,530	0,302	0,826	0,526	1,000

Para esta matriz se buscan los valores propios o autovalores (eigenvalues) mediante el proceso llamado diagonalización (conviene usar el software adecuado). Puesto que la matriz es simétrica y definida positiva está garantizado que existen 7 valores propios reales y positivos $R_1 > R_2 > \ldots > R_N$. Cada uno de los valores propios dividido de la suma de todo es el peso w_k del factor subyacente k:

$$w_k = \frac{R_k}{\sum_{p=1}^{N} R_p} \qquad [2]$$

En nuestro caso los valores propios para la matriz de covarianzas son:

$[S_D] =$

4,517						
	1,766					
		0,414				
			0,223			
				0,059		
					0,021	
						0,003

[Los valores de la diagonal representan R_1 a R_7, de donde se deducen los pesos w_1 = 0,645, w_2 = 0,252, w_3 = 0,059, etc.] La interpretación tanto de la matriz original de covarianzas como de esta última matriz de covarianza diagonalizada es la siguiente: si imaginamos un espacio vectorial abstracto de 7 dimensiones donde cada lengua viene representada por un vector de la base canónica [el portugués sería (1,0,0,0, 0,0,0), el castellano (0,1,0,0, 0,0,0) y así hasta la última lengua el francés que representaría (0,0,0,0, 0,0,1)] la matriz de covarianzas serviría para medir distancias en esa base de tal manera que la “distancia” o “diferenciación” entre dos lenguas L1 y L2 con vectores de representación en este espacio v_{L1} y v_{L2} sería:

$$d(L1,L2) = (v_{L1} - v_{L2})^T \cdot [S](v_{L1} - v_{L2}) \qquad [3]$$

Siendo [S] la matriz métrica que si se usa la base original se tomaría [S_1] y si se trabaja en la base diagonalizada sería [S_D]. Es decir, [S_1] y [S_D] de hecho representan el tensor métrico expresado en dos bases diferentes.

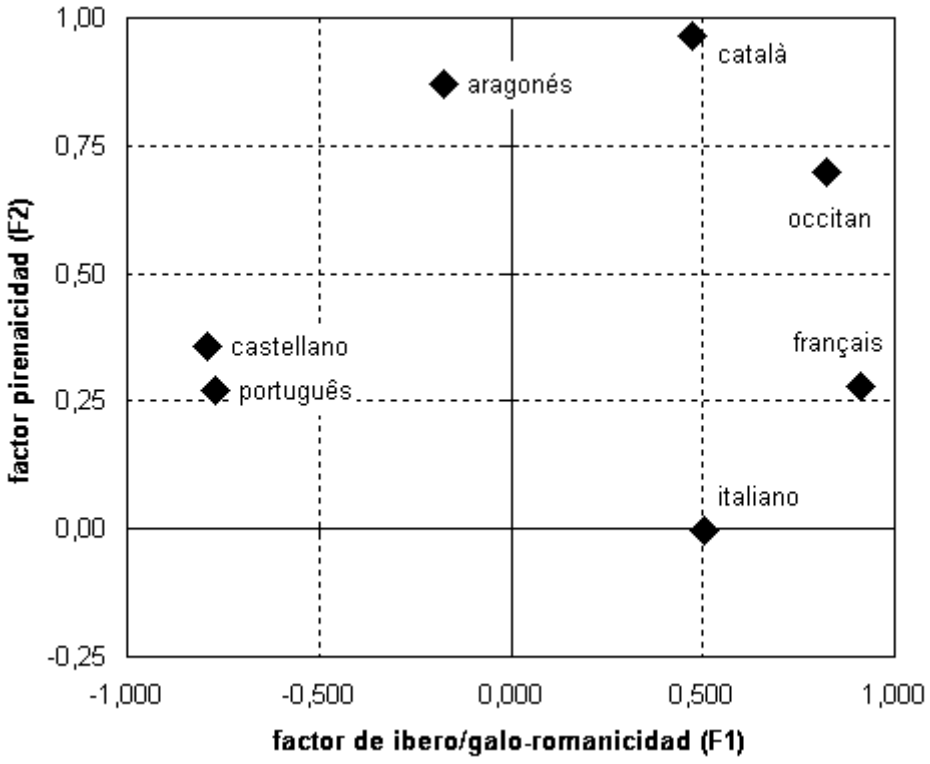
La segunda base es interesante porque los componentes principales vienen dados en ella por los siguientes vectores F1 = (1,0,0,0,0,0,0), F2 = (0,1,0,0,0,0,0), …, F7 = (0,0,0,0,0,0,1), mientras que en la primera base esto factores vienen dados por los vectores propios que también hemos calculado mediante programa. Como cualquier vector expresable en la primera base es interpretable en la segunda y viceversa, resulta que podemos interpretar cualquiera de las lenguas en consideradas en función de combinaciones lineales de los factores o componentes principales.

Como nuestro estudio reveló que sólo había como mucho 3 factores explicativos estadísticamente relevantes podemos imaginar un espacio tridimensional (F1 – F2 – F3) en el que podemos situar las diferentes lenguas tratadas:

	F1	F2	F3
Pt	-0,770	0,271	0,408
cs	-0,791	0,355	0,453
ar	-0,174	0,872	0,659
ct	0,470	0,965	0,494
oc	0,825	0,699	0,412
it	0,504	-0,005	0,774
fr	0,910	0,280	0,253

Las lenguas más cercanas entre sí quedan agrupadas más cerca entre sí en ese espacio abstracto. Por ejemplo si ignoramos la coordenada F3 la distribución de lenguas en este espacio abstracto es la siguiente:

FIG. 1.



El gráfico anterior expresa claramente hechos conocidos, que las dos lenguas románicas más cercanas de la muestra considerada son castellano y portugués. Que el siguiente grupo de lenguas cercanamente emparentadas es el de las lenguas románicas pirenaicas (aragonés, aragonés, catalán) estando el aragonés más o menos a medio camino, por lo que a variación léxica ser refiere, entre catalán y lenguas claramente ibero-romances (portugués y castellano). También se ve la cercanía del occitano de un lado con el catalán y de otro con el francés. En ese diagrama las lenguas galo-romances están situadas a la derecha del todo, mientras que las lenguas ibero-romances a la derecha del todo, siendo las lenguas más centrales sólo débilmente galo-romances o ibero-romances.

Por último para expresar las correlaciones entre las variaciones observadas entre como coinciden las lenguas entre sí y los factores explicativos encontrados se da la matriz de correlación entre factores explicativos y lenguas. Un valor superior en valor absoluto a 0,30 debe considerarse como orientativo de que el factor es explicativo para la lengua en cuestión (mientras que un valor del coeficiente de correlación comprendido entre -0,30 y +0,30 indica muy bajo poder explicativo). Los coeficientes de correlación encontrados que permitieron interpretar el significado tentativo de los factores F1 a F3 son:

	F1	F2	F3
Pt	-0,985	-0,180	0,050
cs	-0,993	-0,087	0,101
ar	-0,461	0,711	0,314
ct	0,520	0,860	0,029
oc	0,919	0,427	-0,165
it	0,580	-0,560	0,478
fr	0,946	-0,042	-0,288

Puede observarse que F1 es explicativo para casi todas las lenguas (en particular para las lenguas claramente ibero-romances o claramente galo-romances). El factor F2 es significativo para las lenguas pirenaicas (correlación positiva) y para el italiano (correlación negativa) siendo para el resto de lenguas poco explicativo. En cuanto al factor 3 puede verse que solo para italiano, aragonés y parcialmente francés parece ser significativo.

A.2. Anexo: Cercanía del parentesco entre las lenguas románicas

Con los datos anteriores puede construirse una matriz de "distancias" entre las lenguas románicas consideradas considerando como distancia abstracta la dada por la ecuación [3] del anexo anterior. Un valor alto de la distancia indica que una lengua está alejada de la otra mientras que un valor cercano a cero indica cercanía, la matriz de distancias es:

	Pt	cs	ar	ct	oc	it	fr
Pt	0,000	0,047	0,569	1,055	1,299	1,036	1,350
cs	0,047	0,000	0,562	1,059	1,310	1,059	1,368
ar	0,569	0,562	0,000	0,521	0,810	0,701	0,925
ct	1,055	1,059	0,521	0,000	0,315	0,493	0,496
oc	1,299	1,310	0,810	0,315	0,000	0,447	0,225
it	1,036	1,059	0,701	0,493	0,447	0,000	0,378
fr	1,350	1,368	0,925	0,496	0,225	0,378	0,000

Puede verse que en general lenguas centrales como el catalán o el aragonés son las que tienen distancias más cercanas con las demás mientras que las lenguas más divergentes dentro de la muestra son el francés, el castellano y el portugués como ya se observaba en la FIG. 1. Los valores promedio de distancias de una lengua con el resto son los siguientes:

Pt	cs	ar	ct	oc	it	fr
0,765	0,772	0,584	0,563	0,629	0,588	0,677

De hecho estas distancias se han calculado el peso de sólo los 3 primeros factores, ya que los otros 4 no contribuyen prácticamente en nada a la distancia. Como puede verse el catalán es las lenguas que representa una distancia promedio menor con el resto de lenguas. Una conclusión similar se obtiene si tomamos los datos iniciales de porcentaje compartido de cognados en el léxico no pan-románico donde se obtienen los siguientes promedios de concordancia:

Pt	cs	ar	ct	oc	it	fr
43,9%	44,8%	55,0%	51,5%	53,9%	45,8%	43,5%