APLICACIÓ DE L'ANÀLISI MULTIVARIANT A UN ESTUDI SOBRE LES LLENGÜES EUROPEES

F. OLIVA, C. BOLANCE i L. DIAZ*

Universitat de Barcelona

Utilitzant una informació complexa i qualitativa (l'escriptura dels deu primers nombres) es presenta un mètode que permet quantificar adequadament les diferències entre catorze llengües europees construint una matriu de distàncies. Per realitzar l'estudi comparatiu s'empren dues tècniques d'anàlisi multivariant: l'anàlisi de proximitats ("multidimensional scaling") i l'anàlisi de conglomerats jeràrquica ("hierarchical cluster analysis").

1. INTRODUCCIÓ

L'estudi que es presenta a continuació és una comparació de catorze llengües europees a partir d'una informació fàcilment accessible però complexa pel que fa a la seva quantificació: l'escriptura dels deu primers nombres naturals (vegeu la taula 1). Cal dir ja des d'ara mateix que els autors no pretenen extreure conclusions rigoroses des d'un punt de vista lingüístic, sinó mostrar com l'anàlisi multivariant pot proporcionar eines interessants si prèviament s'ha realitzat un esforç per quantificar la informació. La utilització dels deu primers nombres es justifica perquè són paraules força representatives i que no han experimentat canvis de significat en el curs de la història (recordem que solen formar part indefectiblement de la primera lliçó dels llibres de text per aprendre una llengua).

Un antecedent d'aquest estudi pot trobar-se a Johnson i Wichern (1988), on a partir de la informació esmentada anteriorment construeixen una matriu de distàncies entre onze llengües i realitzen posteriorment una anàlisi de conglo-

*Dept. d'Estadística. Facultat de Biologia. Universitat de Barcelona. Av. Diagonal, 645. 08028 Barcelona.

Taula 1.

Escriptura dels deu primers nombres de les catorze llengües considerades en l'estudi. S'ha prescindit dels accents i d'altres signes i s'han conservat sols els caràcters alfabètics. Entre parèntesi figuren les abreviatures que seran utilitzades en les representacions gràfiques.

Alemany	Anglès	Basc	Castellà	Català	Danès	Finlandès	Francès	Gallec	Holandès	Hongarès	Italià	Noruec	Polonès
(Al)	(An)	(Ba)	(Cs)	(Ca)	(Da)	(Fi)	(Fa)	(Ga)	(Ho)	(Hg)	(It)	(No)	(Po)
ein	one	bat	uno	u	en	yksi	un	unho	een	egy	uno	en	jeden
zwei	two	bi	dos	dos	to	kaksi	deux	dous	twee	ketto	due	to	dwa
drei	three	iru	tres	tres	tre	kolme	trois	tres	drie	harom	tre	tre	trzy
vier	four	lau	cuatro	quatre	fire	neua	quatre	catro	vier	negy	quattro	fire	cztery
funf	five	bost	cinco	cinc	fem	viisi	cinq	cinco	vijf	ot	cinque	fem	piec
sechs	six	tzei	seis	sis	seks	kuusi	six	seis	zes	hat	sei	seks	szesc
sieben	seven	tzaspi	siete	set	syv	seitseman	sept	sete	zeven	het	sette	sju	siedem
acht	eight	txortzi	ocho	vuit	otte	kahdeksan	huit	oito	acht	nyolc	otto	atte	osiem
neun	nine	beratzi	nueve	nou	ni	yhdeksan	neuf	nove	negen	kylenc	nove	ni	dziewiec
zehn	ten	amar	diez	deu	ti	kymmenen	dix	dez	tien	tiz	diez	ti	dziesiec

merats jeràrquica que permet observar algunes agrupacions. El mètode emprat per quantificar les diferències és ben senzill: consideren com a distància entre dues llengües el nombre de paraules que no comencen per la mateixa lletra. Per exemple, la distància entre el polonès i el castellà seria de tres, ja que en tres nombres ("1", "5" i "9") la lletra inicial és diferent; en canvi, la distància entre el castellà i l'italià és d'1, ja que sols el "4" no té la primera lletra igual. Aquest algorisme binari per comparar dues paraules desaprofita bona part de la informació disponible i comporta avaluacions poc afortunades: paraules força diferents aporten distància zero, mentre que paraules gairebé idèntiques són valorades com a diferència u. Un exemple ben esclaridor del que pot passar és el nombre "4" en català, castellà i polonès (*quatre, cuatro i cztery* respectivament). Segons la regla anteriorment descrita, per a aquesta paraula hem de considerar el català igual de diferent del castellà i del polonès (!?), mentre que el castellà i el polonès "coincideixen" (per un cop no valdrà l'acudit que els catalans sembla que parlem polonès!). És evident que el mètode resulta poc adequat perquè valora com a iguals paraules que només coincideixen en la primera lletra per casualitat i com a diferents paraules similars quant a l'escriptura i la pronunciació.

Per tal d'evitar aquest problema es proposa un mètode més laboriós però que permeterà considerar tota la informació per avaluar la diferència entre dues llengües. S'ha incorporat a l'estudi per raons d'interès evident el català, el basc i el gallec. Un cop obtinguda la matriu d'interdistàncies, l'anàlisi de proximitats o MDS (*multidimensional scaling*) serà utilitzada per aconseguir una representació adequada en dimensió reduïda que permeti relacionar les diferents llengües. Així mateix, s'empraran tres algorismes d'anàlisi de conglomerats jeràrquica i es construiran els dendrogrames corresponents. Es presentaran i es discutiran diverses mesures de l'ajust en ambdós casos per valorar la fiabilitat de les representacions obtingudes.

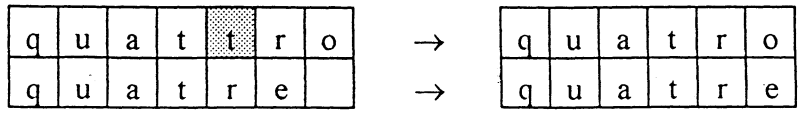
2. CONSTRUCCIÓ DE LA MATRIU D'INTERDISTÀNCIES

La construcció d'una dissimilitud entre dues llengües aprofitant al màxim possible la informació proporcionada per la taula 1 no és una tasca fàcil. Pot observar-se que s'ha realitzat una primera simplificació conservant sols les lletres i obviant en l'escriptura de les paraules els accents i la resta de signes. El criteri escollit per calcular la distància està basat en el nombre de no coincidències quan es consideren les paraules senceres, però amb unes normes addicionals que tot seguit exposem.

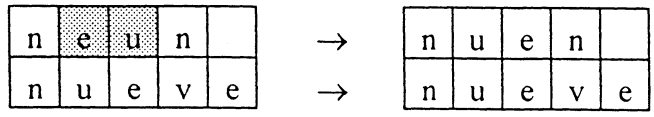
Suposem que volem calcular la dissimilitud entre dues llengües. Per a cada un dels deu nombres realitzarem el següent procés:

1) Utilitzarem si és convenient les regles:

   a) L'addició, deleció o duplicació d'una lletra és considerada com una diferència. Per exemple, considerem el "4" en italià i català. Suprimint una "t" de la paraula italiana (o bé addicionant una "t" al català) aconseguim amb una diferència



   b) La transposició entre dues lletres consecutives és considerada com una diferència. Per exemple, el nombre "9" en alemany i castellà. Transposant les lletres "e" i "u" en qualsevol de les dues paraules obtenim

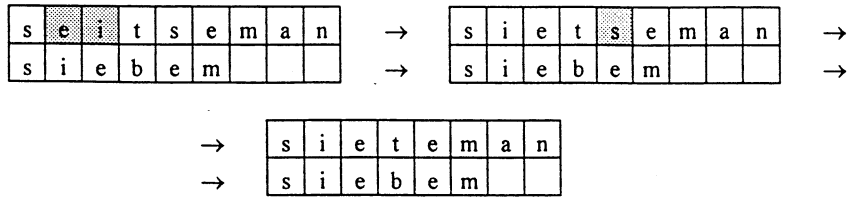


2) Superposarem ambdues paraules i cada lletra no coincident serà considerada una diferència (per exemple, una per al "4" en italià-català i dues per al "9" en alemany-castellà).

   Sumarem les diferències trobades en 1) i 2) (dues i quatre respectivament en els exemples anteriors). En alguns casos, segons l'ordre i el nombre de vegades que són utilitzades les regles, és possible obtenir diferents resultats; aleshores triarem el camí òptim, és a dir, el nombre mínim de diferències.

3) Repetirem el procés per a cada un dels nombres i considerarem com a mesura de dissimilitud entre dues llengües el nombre mitjà de diferències per paraula.

Pot comprovar-se com l'aplicació de les regles a) i b) permeten reduir el nombre de diferències d'una manera lògica. En efecte, sense la seva utilització els dos exemples abans presentats tindrien una diferència més en cada cas. Un bon exemple de l'aplicació completa del procés per a una paraula és el nombre "7" en finlandès (*seitseman*) i polonès (*siebem*):



El nombre de diferències és cinc, mentre que hauria estat vuit si la comparació fos considerant sols les lletres no coincidents.

Utilitzant aquest mètode s'ha construït la matriu de distàncies entre les catorze llengües. El resultat ha estat el següent (vegeu la taula 1 per al significat de les abreviatures)

	Al	An	Ba	Ca	Cs	Da	Fi	Fr	Ga	Ho	Hg	It	No	Po
Al	0.0													
An	2.9	0.0												
Ba	4.5	4.4	0.0											
Ca	3.4	2.8	4.5	0.0										
Cs	3.2	2.9	4.6	1.7	0.0									
Da	3.0	2.6	4.3	2.7	3.1	0.0								
Fi	5.8	5.5	5.9	5.7	5.5	5.9	0.0							
Fr	3.3	3.2	4.6	1.3	2.4	3.3	5.9	0.0						
Ga	3.2	2.7	4.4	1.3	0.7	2.6	5.5	2.3	0.0					
Ho	1.9	2.5	4.3	4.3	3.2	2.9	5.6	3.3	3.3	0.0				
Hg	4.2	3.8	4.5	4.0	4.2	3.6	5.6	3.8	4.0	3.7	0.0			
It	3.7	3.5	4.6	2.2	1.7	3.2	6.0	2.4	1.5	3.6	4.5	0.0		
No	2.9	2.7	4.3	2.9	3.2	0.3	5.8	3.3	2.7	2.8	3.6	3.3	0.0	
Po	4.5	4.4	5.3	4.4	3.6	4.4	5.6	4.5	3.8	4.2	5.2	4.2	4.4	0.0

3. L'ANÀLISI DE PROXIMITATS O MDS

Considerem un conjunt finit de n objectes (individus, poblacions, ...) $O_1, \ldots, O_n$.. La MDS (*multidimensional scaling*) és una tècnica multivariant d'anàlisi de dades que permet trobar una configuració de n punts en un espai euclidià utilitzant com a informació les proximitats (similituds o dissimilituds) entre els n objectes.

Direm que $\boldsymbol{D}_{n\times n} = (d_{ij})$ on $d_{ij} = d(O_i, O_j)$ és una matriu de distàncies si compleix les següents propietats

(a) simetria: $d_{ij} = d_{ji}$ (b) no negativitat: $d_{ij} \geq 0$, $i \neq j$ (c) $d_{ii} = 0$

Si a més compleix les propietats

(e) $d_{ij} = 0 \Leftrightarrow O_i \equiv O_j$ (f) desigualtat triangular: $d_{ij} \leq d_{ik} + d_{jk}$

aleshores $\boldsymbol{D}$ és una matriu de distàncies mètrica. Si una distància sols presenta les tres primeres propietats molts autors l'anomenen dissimilitud. Finalment una matriu de distàncies $\boldsymbol{D}$ és euclidiana si compleix les cinc propietats i a més és possible trobar una configuració de punts $P_1, \ldots, P_n$ en un espai euclidià amb coordenades $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ tals que

$$d_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

En algunes situacions, es disposa de les similituds entre els objectes i no pas de distàncies. Diem que $\boldsymbol{S}_{n\times n} = (s_{ij})$ on $s_{ij} = s(O_i, O_j)$ és una matriu de similituds si

(a) $s_{ij} = s_{ji}$ (b) $s_{ij} \leq s_{ii}$

És fàcil però comprovar que podem obtenir una matriu de dissimilituds a partir de les similituds mitjançant la transformació

$$d_{ij} = (s_{ii} + s_{jj} - 2s_{ij})^{1/2}$$

A partir de les dissimilituds o distàncies d_{ij}, l'objectiu de la MDS és trobar una configuració de n punts $P_1, \ldots, P_n$ en dimensió k tal que si denominem $d_{ij(k)}$ la distància euclidiana entre P_i i P_j, aleshores $\boldsymbol{D}_{(k)}$ sigui "semblant" a $\boldsymbol{D}$. La dimensió k és desconeguda, però a la pràctica es limita generalment a $k \leq 3$ per possibilitar la interpretació. És important assenyalar que la solució obtinguda és invariant quant a translacions, rotacions i reflexions. Hi ha nombrosos mètodes de MDS, però poden distingir-se dos grans grups:

- *mètodes mètrics:* intenten aconseguir la configuració de punts P_i utilitzant directament les distàncies entre els objectes.

- *mètodes no mètrics:* la informació utilitzada és sols el rang de les $n(n-1)/2$ distàncies entre tots els parells d'objectes

$$d_{i_1 j_1} < d_{i_2 j_2} < \cdots < d_{i_m j_m} \qquad m = n(n-1)/2$$

Atès que els rangs no varien per transformacions monòtones de les distàncies, la solució obtinguda serà també invariant respecte l'expansió o contracció uniforme.

En aquest estudi s'ha utilitzat el mètode mètric clàssic que serà breument descrit a continuació.

Solució mètrica clàssica

Sigui $\boldsymbol{D}$ la matriu d'interdistàncies entre els n objectes. Considerem les matrius $\boldsymbol{A}$ i $\boldsymbol{B}$ d'ordre n

$$\boldsymbol{A} = (a_{ij}), \quad a_{ij} = -\frac{1}{2} d_{ij}^2 \qquad\qquad \boldsymbol{B} = (b_{ij}), \quad b_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..}$$

on

$$a_{i.} = \frac{1}{n}\sum_{j=1}^{n} a_{ij} \qquad a_{.j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij} \qquad a_{..} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}$$

O expressat en forma matricial $\boldsymbol{B} = \boldsymbol{HAH}$ on $\boldsymbol{H} = \boldsymbol{I} - n^{-1}\mathbf{1}\mathbf{1}'$ és l'anomenada matriu centradora de dades d'ordre n. Es compleixen aleshores els següents resultats:

a) Si $\boldsymbol{D}$ és la matriu de distàncies euclidianas per a una configuració de punts $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ llavors $b_{ij} = (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_j - \overline{\boldsymbol{z}})'$, $i, j = 1, \ldots, n$. En forma matricial $\boldsymbol{B} = (\boldsymbol{HZ})(\boldsymbol{HZ})'$ i per tant $\boldsymbol{B} \geq 0$, és a dir, és semidefinida positiva (s.d.p.).

b) I a l'inrevés, si $\boldsymbol{B}$ és s.d.p. de rang $r \leq n-1$ pot aleshores construir-se una configuració de n punts $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, on $\boldsymbol{x}_i \in \mathbb{R}^r$ és la fila i–èsima de $\boldsymbol{X}$, tals que $d_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j)$. L'obtenció de $\boldsymbol{X}$ es immediata diagonalitzant $\boldsymbol{B}$. En efecte,

$$\boldsymbol{B} = \boldsymbol{T\Lambda T}' = (\boldsymbol{T\Lambda}^{1/2})(\boldsymbol{T\Lambda}^{1/2})' = \boldsymbol{XX}'$$

on $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_r)$ és la matriu diagonal de valors propis diferents de zero $\lambda_1 \geq \cdots \geq \lambda_r > 0$ i $\boldsymbol{T}$ és la matriu de vectors propis associats.

Algunes propietats importants són:

a) Les columnes de $\boldsymbol{X}$ són els vectors propis λ–normalitzats, és a dir, $\boldsymbol{x}'_{(i)}\boldsymbol{x}'_{(i)} = \lambda_i$, $\boldsymbol{x}'_{(i)}\boldsymbol{x}_{(j)} = 0$, $i \neq j$ $(i, j = 1, \ldots, r)$.

b) El centre de gravetat és l'origen de coordenades. Efectivament, $\boldsymbol{B1} = \boldsymbol{HAH1} = \boldsymbol{0}$ i per tant $\boldsymbol{1}$ és vector propi de valor propi 0. Llavors

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \boldsymbol{X}'\boldsymbol{1} = \boldsymbol{0}$$

c) Es compleix la relació $\sum_{i,j}^{n} d_{ij}^2 = \text{tr } \boldsymbol{D}^2 = 2n \text{ tr } \boldsymbol{B} = 2n\sum_{i=1}^{r}\lambda_i$ on $\boldsymbol{D}^2 = \boldsymbol{DD}$.

d) Si volem representar els objectes en un espai de dimensió reduïda k, la màxima resolució (separació entre els objectes) serà aconseguida utilitzant les k primeres coordenades (columnes de $\boldsymbol{X}$): $\boldsymbol{x}_{i(k)} = (x_{i1} \cdots \boldsymbol{x}_{ik})'$, $i = 1, \ldots, n$. La dispersió dels objectes en aquest espai euclidià serà

$$\sum_{i,j}^{n} d_{ij(k)}^2 = 2n \text{ tr } \boldsymbol{B}_{(k)} = 2n\sum_{i=1}^{k}\lambda_i \quad \text{on } \boldsymbol{B}_{(k)} = \boldsymbol{X}_{(k)}\boldsymbol{X}'_{(k)}$$

És a dir, de totes les possibles configuracions $\overline{\boldsymbol{x}}_{i(k)}$ dels n objectes en un espai euclidià de dimensió k, pot demostrar-se que la mesura de discrepància $\phi = \sum_{i,j}(d_{ij}^2 - \tilde{d}_{ij(k)}^2)$ és mínima si $\tilde{\boldsymbol{x}}_{i(k)} = \boldsymbol{x}_{i(k)}$. Aleshores $\min\phi = \text{ tr } (\boldsymbol{B} - \boldsymbol{B}_{(k)})$ i la mesura

$$c = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{r}\lambda_i} \times 100 = \frac{\text{tr } \boldsymbol{B}_{(k)}}{\text{tr } \boldsymbol{B}} \times 100 = 1 - \left[\frac{\text{tr } (\boldsymbol{B} - \boldsymbol{B}_{(k)})}{\text{tr } \boldsymbol{B}}\right] \times 100.$$

que ens indica el percentatge de dispersió (variabilitat), explicada per les k primeres coordenades, és màxima.

e) Si la informació inicial és una matriu de similituds $\boldsymbol{S}$ entre els n objectes, no és necessari transformar-les a dissimilituds, atès que és fàcil comprovar que $\boldsymbol{B} = \boldsymbol{HSH}$.

Aquesta solució, generalment coneguda com a mètode clàssic de la MDS, fou demostrada per Schoenberg (1935) i Richardson (1938) però ha estat popularitzada per Torgerson (1952, 1958), que introduí el terme *multidimensional scaling*. Més tard, Gower (1966) l'anomenà anàlisi de coordenades principals i va mostrar l'estreta connexió amb l'anàlisi de components principals.

Què passa però si la distància no és euclidiana? En aquest cas $\boldsymbol{B}$ tindrà valors propis negatius i no podem definir en els reals la potència $\boldsymbol{\Lambda}^{1/2}$, per tant no és possible λ–normalitzar els vectors propis. Suposem que $r(\boldsymbol{B}) = r$, $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_r)$ amb p valors propis positius i $q = r - p$ valors propis negatius

$$\lambda_1 \geq \cdots \geq \lambda_p > 0 > \lambda_{p+1} \geq \cdots \geq \lambda_r$$

Per aconseguir $\boldsymbol{x}'_{(i)}\boldsymbol{x}_{(i)} = \lambda_i\,(i = 1, \ldots, r)$ les últimes q coordenades han de ser imaginàries. En efecte,

$$\boldsymbol{x}_{(h)} = i\,|\lambda_h|^{1/2}\,\boldsymbol{t}_{(h)}, \quad h = p+1, \ldots, r \quad \text{on} \quad i = \sqrt{-1}$$

i les coordenades dels objectes serien $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp}, ix_{jp+1}, \ldots, ix_{jr})'$, $j = 1, \ldots, n$. Les interdistàncies entre els n objectes poden aleshores ser expressades com

$$d_{ij}^2 = \sum_{h=1}^{p}(x_{ih} - x_{jh})^2 - \sum_{h=p+1}^{r}(x_{ih} - x_{jh})^2$$

distàncies corresponents a una geometria ortogonal no representable en l'espai euclidià real (es pot observar que les dispersions dels objectes per a les q últimes coordenades representen de fet anti-distàncies).

La solució més senzilla al problema plantejat és considerar sols les p coordenades reals i obviar les q imaginàries, la qual cosa implica aproximar la matriu $\boldsymbol{B}$ per una altra semidefinida positiva de rang inferior $\boldsymbol{B}^* = \boldsymbol{B}_{(p)} = \boldsymbol{X}_{(p)}\boldsymbol{X}'_{(p)}$. Aquest problema fou estudiat per Eckart i Young (1936) en un context general i per Mardia (1978) en el context de la MDS, obtenint el següent resultat:

Si considerem

$$\psi = \sum_{i,j=1}^{n}(b_{ij} - \tilde{b}_{ij})^2 = \text{tr}\,(\boldsymbol{B} - \tilde{\boldsymbol{B}})^2$$

la mesura de la discrepància entre $\boldsymbol{B}$ i una altra matriu simètrica s.d.p. de rang inferior $\tilde{\boldsymbol{B}}$ es demostra aleshores que ψ es minimitza si $\tilde{\boldsymbol{B}} = \boldsymbol{B}^*$ i, per tant,

$$\min\psi = \text{tr}\,(\boldsymbol{B} - \boldsymbol{B}^*)^2 = \sum_{i=p+1}^{r}\lambda_i^2$$

Aquest resultat comporta la següent generalització: si volem aconseguir una configuració dels objectes en un espai euclidià de dimensió $k \leq p$ l'expressió ψ serà minimitzada quan $\tilde{\boldsymbol{B}} = \boldsymbol{B}_{(k)}$. Mardia (1978) proposa dues mesures de la dispersió explicada per la representació dels objectes en $\mathbb{R}^k$

$$c_1 = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{r}|\lambda_i|}\times 100 \qquad c_2 = \frac{\sum_{i=1}^{k}\lambda_i^2}{\sum_{i=1}^{r}|\lambda_i^2|}\times 100 = \left[1 - \frac{\text{tr}\,(\boldsymbol{B} - \boldsymbol{B}_{(k)})^2}{\text{tr}\,\boldsymbol{B}^2}\right]\times 100$$

La justificació de c_2 (mesura proposada també per Saito (1978)) és evident donada la seva relació amb ψ, mentre que c_1 és una generalització de la mesura c presentada anteriorment en el cas que $\boldsymbol{D}$ sigui euclidiana. No obstant això, en l'opinió dels autors d'aquest estudi, c_1 és menys "natural" atès que el denominador no correspon a cap dispersió global. En efecte

$$c_1 = \frac{\text{tr } \boldsymbol{B}_{(k)}}{\text{tr } \boldsymbol{B} + 2 \text{ tr } (\boldsymbol{B}^* - \boldsymbol{B})} \times 100 = \left[1 - \frac{\text{tr } (\boldsymbol{B} - \boldsymbol{B}_{(k)}) + 2 \text{ tr } (\boldsymbol{B}^* - \boldsymbol{B})}{\text{tr } \boldsymbol{B} + 2 \text{ tr } (\boldsymbol{B}^* - \boldsymbol{B})}\right] \times 100$$

Cal finalment assenyalar que el procediment exposat serà poc aconsellable si el "pes" dels valors propis negatius és alt o fins i tot impossible d'aplicar si $k > p$. En aquest cas poden emprar-se altres mètodes més adequats, com per exemple la solució de Mardia (1978), el mètode iteratiu lineal i els mètodes no mètrics (el lector interessat pot consultar els capítols sobre el tema de Cuadras (1991), Dillon i Goldstein (1984), Mardia *et al* (1979), Seber (1984) o l'obra específica sobre el tema de Davison (1983)).

4. L'ANÀLISI DE CONGLOMERATS JERÀRQUICA

Donat un conjunt de n objectes $=_1, \ldots, O_n$ dels quals es disposa una informació quantificable, l'anàlisi de conglomerats (*cluster analysis*) engloba una sèrie de tècniques que tenen com a finalitat l'agrupació dels objectes més semblants (mètodes aglomeratius) o bé la formació de subconjunts a partir del conjunt inicial (mètodes divisius). L'objectiu és classificar els n objectes de manera que, respecte a la informació coneguda, siguin el més homogenis possible dins dels grups i heterogenis entre els grups. Els mètodes d'anàlisi de conglomerats poden classificar-se també d'acord amb un altre criteri:

- *mètodes jeràrquics*: estableixen successives fusions o divisions dels objectes depenent de la seva homogeneïtat, formant una estructura de grups jerarquitzada. Presenten la particularitat que quan un objecte ha estat assignat a un grup no és ja possible reconsiderar la seva classificació. El resultat obtingut pot ser representat mitjançant un diagrama de dues dimensions en forma d'arbre anomenat dendrograma.

- *mètodes no jeràrquics*: pretenen aconseguir grups homogenis sense establir entre ells relacions de jerarquia. Són tècniques que realitzen una partició del conjunt dels n objectes optimitzant un determinat criteri formal predeterminat. En oposició als mètodes jeràrquics, un individu inicialment assignat a un grup pot ser posteriorment reclassificat.

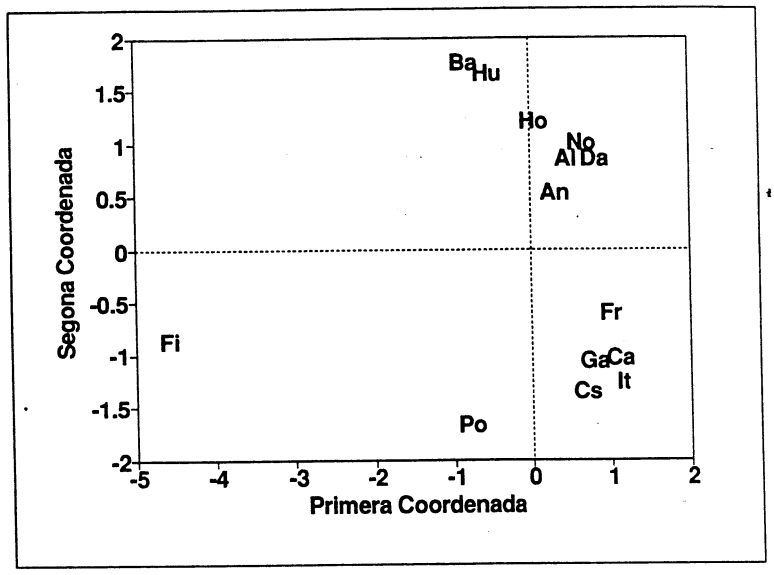


Figura 1. Representació de les llengües emprant les dues primeres coordenades obtingudes amb la MDS.

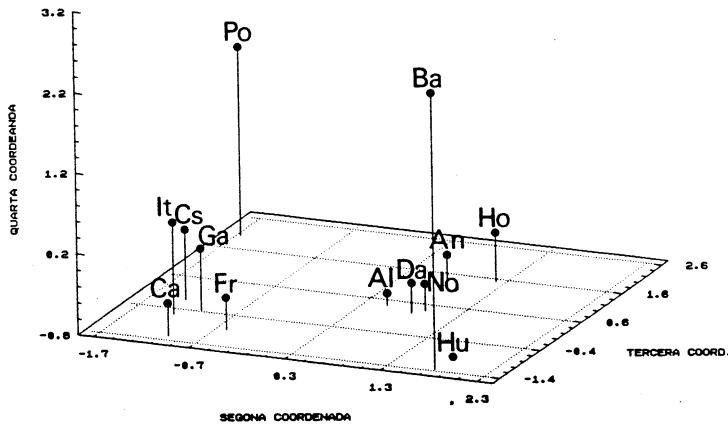


Figura 2. Representació de les llengües, exceptuant el finlandès, utilitzant la segona, tercera i quarta coordenades de la MDS.

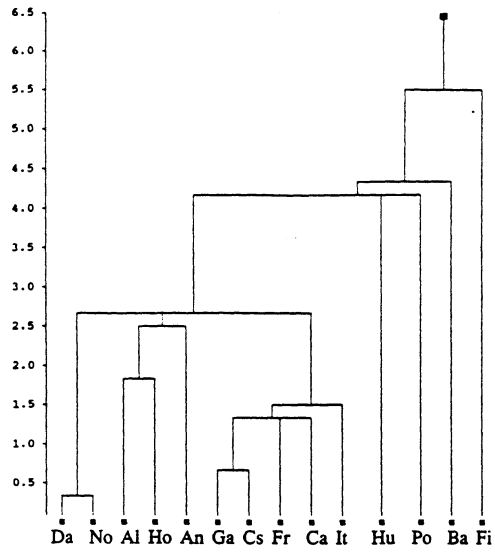


Figura 3a)

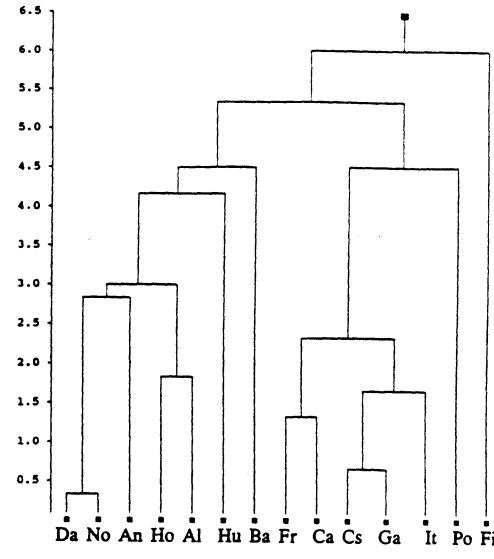


Figura 3b)

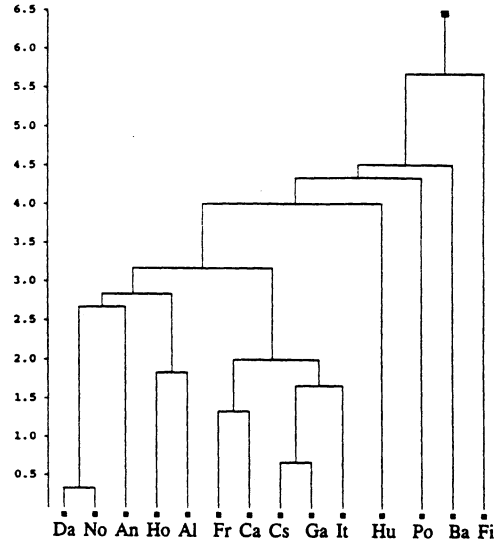


Figura 3c)

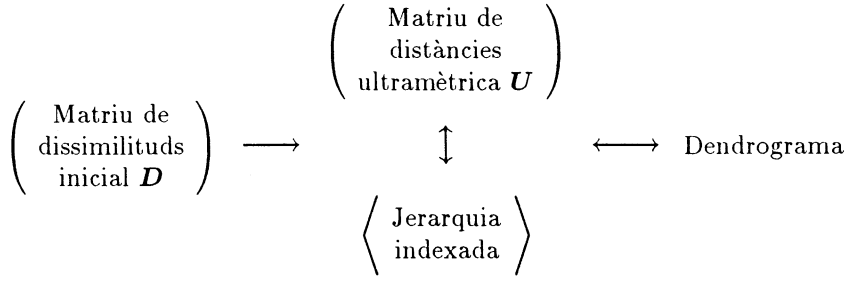
Figura 3. Dendrogrames obtinguts amb l'anàlisi de conglomerats jeràrquica: a) mètode del mínim, b) mètode del màxim, i c) mètode UPGMA.

En aquest estudi lingüístic s'han utilitzat mètodes jeràrquics aglomeratius, atès que la finalitat és aconseguir una agrupació successiva de les llengües europees per observar les "relacions de parentiu". Presentem a continuació una breu exposició dels mètodes utilitzats.

Sigui $\boldsymbol{D}$ una matriu de dissimilituds entre n objectes. Els mètodes d'anàlisi de conglomerats jeràrquica es basen en algorismes de construcció d'una matriu de distàncies ultramètrica $\boldsymbol{U} = (u_{ij})$, $u_{ij} = u(O_i, O_j)$, que sigui el més semblant possible a $\boldsymbol{D}$. Perquè una distància sigui ultramètrica ha de verificar la següent propietat:

$$u_{ij} \leq \max\{u_{ik}, u_{jk}\}$$

coneguda com a axioma ultramètric i que és més restrictiva que la desigualtat triangular. Com a conseqüència geomètrica d'aquesta propietat, tot triangle definit per les distàncies ultramètriques entre tres objectes és isòsceles, la qual cosa pot ser comprovada amb els dendrogrames representats en la figura 3. No és l'objectiu d'aquesta exposició aprofundir en aspectes formals i per tant només citarem que tota distància ultramètrica u definida sobre un conjunt finit de n objectes defineix una jerarquia indexada en $\{O_1, \ldots, O_n\}$; així mateix, tota jerarquia indexada implica una distància ultramètrica definida entre els n objectes. Finalment, un dendrograma no és més que la representació geomètrica d'una jerarquia indexada i, en conseqüència, d'una distància ultramètrica. Aquest procés pot ser esquematitzat de la següent manera



Hi ha nombrosos algorismes per construir una distància ultramètrica adequada (i, per tant, una jerarquia indexada) a partir d'una matriu de dissimilituds $\boldsymbol{D}$. S'han emprat en aquest estudi tres dels més freqüentment utilitzats:

a) **mètode del mínim**, *single linkage* o *nearest-neighbour* (Sneath (1957), Sokal i Sneath (1963), Johnson (1967)): si C_1 i C_2 són dos conglomerats o grups, la distància entre ells es defineix com la mínima distància observada entre un membre de C_1 i un altre de C_2, és a dir,

$$d_{(C_1)(C_2)} = \min\{d_{ij} : i \in C_1,\ j \in C_2\}$$

Mostrarem el procés de fusió amb un exemple senzill. Sigui

$$\boldsymbol{D} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} 0 & 7 & 1 & 9 \\ 7 & 0 & 6 & 3 \\ 1 & 6 & 0 & 8 \\ 9 & 3 & 8 & 0 \end{pmatrix} \end{array}$$

la matriu d'interdistàncies entre quatre objectes O_1, O_2, O_3 i O_4. La fusió ha de començar entre O_1 i O_3, ja que són els objectes més propers (homogenis). Els grups són ara $(O_1, O_3), O_2$ i O_4. Aleshores

$$d_{(2)(1,3)} = \min\{d_{21}, d_{23}\} = d_{23} = 6, \quad d_{(4)(1,3)} = \min\{d_{41}, d_{43}\} = d_{43} = 8$$

i la matriu resultant és

$$\boldsymbol{D}_1 = \begin{array}{c} \begin{matrix} (1,3) & 2 & 4 \end{matrix} \\ \begin{pmatrix} 0 & 6 & 8 \\ 6 & 0 & 3 \\ 8 & 3 & 0 \end{pmatrix} \end{array}$$

La unió serà ara entre O_2 i O_4 i es formaran els grups (O_1, O_3) i (O_2, O_4), amb distància $d_{(1,3)(2,4)} = \min\{d_{2(1,3)}, d_{4(1,3)}\} = d_{2(1,3)} = 6$. Finalment unirem els dos conglomerats en un sol conjunt global (O_1, O_2, O_3, O_4). El resultat ha estat una jerarquia i una distància ultramètrica definida entre el objectes

$$\boldsymbol{D} = \begin{pmatrix} 0 & 7 & 1 & 9 \\ 7 & 0 & 6 & 3 \\ 1 & 6 & 0 & 8 \\ 9 & 3 & 8 & 0 \end{pmatrix} \longrightarrow \boldsymbol{U} = \begin{pmatrix} 0 & 6 & 1 & 6 \\ 6 & 0 & 6 & 3 \\ 1 & 6 & 0 & 6 \\ 6 & 3 & 6 & 0 \end{pmatrix}$$

que permeterà la representació en un dendrograma com el de la figura 2. El mètode del mínim és un algorisme espai contractiu: la ultramètrica associada a la classificació jeràrquica tendeix a aproximar els objectes respecte les dissimilituds inicials.

b) **mètode del màxim**, *complete linkage* o *farthest-neighbour* (Sokal i Sneath (1963), McQuitty (1964)): oposat al mètode anterior, un cop units els dos grups més propers C_1 i C_2, la distància es defineix per

$$d_{(C_1)(C_2)} = \max\{d_{ij} : i \in C_1,\ j \in C_2\}$$

És un algorisme espai dilatant: tendeix a allunyar els objectes respecte la dissimilitud inicial.

c) **mètode UPGMA** (*Unweighted Pair Group Method Using Arithmetic Averages*) o *group average* (Sokal i Michener (1958), McQuitty (1964), Lance i Williams (1966)): la distància entre C_1 i C_2 és definida en aquest cas com la mitjana aritmètica de les $n_1 n_2$ distàncies entre totes les parelles d'objectes formades per un element de C_1 i un altre de C_2

$$d_{(C_1)(C_2)} = \frac{1}{n_1 n_2} \sum_{i \in C_1} \sum_{j \in C_2} d_{ij}$$

És un algorisme espai conservador i no modifica substancialment les dissimilituds inicials.

Algunes altres propietats importants d'aquests mètodes són:

- invariància monòtona: un algorisme verifica aquesta propietat si no varia l'estructura jeràrquica quan $\boldsymbol{D}$ és substituïda per una nova matriu de dissimilituds $\hat{\boldsymbol{D}}$ obtinguda mitjançant una transformació monòtona de $\boldsymbol{D}$. Els tres mètodes esmentats presenten invariància monòtona.

- no arbitrarietat: si dues o més dissimilituds són iguals, la classificació jeràrquica no depèn de l'ordre en què han estat agrupats els objectes. Dels tres mètodes, sols el del mínim garanteix aquesta propietat.

- continuïtat: petites pertorbacions en les dissimilituds inicials haurien de provocar sols petites modificacions en el dendrograma resultant. Novament sols el mètode del mínim assegura aquesta propietat.

- no encadenament: l'encadenament es produeix quan diversos conglomerats s'ajunten ràpidament a causa de l'existència de pocs individus intermedis. Aquest és un problema que presenta amb freqüència el mètode del mínim, perquè és espai contractiu.

Finalment, cal considerar alguna mesura de la qualitat de la classificació, ja que el procés $\boldsymbol{D} \longrightarrow \boldsymbol{U}$ provoca una distorsió (excepte si $\boldsymbol{D}$ ja és ultramètrica). Un procediment força utilitzat consisteix a calcular la correlació entre les $n(n-1)/2$ parelles d_{ij}, u_{ij}) i rep el nom de "correlació cofenètica" $0 \leq r_c \leq 1$). Introduïda per Sokal i Rohlf (1962) i analitzades amb profunditat les seves propietats per Farris (1969), proporciona una mesura de la distorsió: valors baixos adverteixen d'una forta discrepància entre les dissimilituds inicials i les distàncies ultramètriques, mentre que valors propers a 1 indiquen una evident estructura jeràrquica entre els n objectes. Una altra mesura, proposada per Jardine i Sibson (1968), és el coeficient

$$\lambda_\alpha = \frac{\left[\sum_{i,j} |d_{ij} - u_{ij}|^{1/\alpha}\right]^\alpha}{\left[\sum_{i,j} d_{ij}^{1/\alpha}\right]^\alpha} \quad 0 < \alpha < 1 \quad 0 \leq \lambda_\alpha \leq 1$$

que depèn del paràmetre α. Una bona classificació quedarà reflectida per un valor de λ_α proper a zero. Si escollim $\alpha = 1/2$ coincideix amb la popular mesura de *STRESS* utilitzada en diversos mètodes no mètrics de MDS.

5. MESURES DE L'AJUST

En l'exposició de la tècnica clàssica de la MDS s'han comentat dues mesures de la dispersió, c_1 i c_2, explicada en representar els n objectes en $\mathbb{R}^k$; en l'apartat anterior s'ha parlat de la correlació cofenètica r_c i del coeficient λ_α (infinites mesures depenent del valor de α) com a mesura de la distorsió d'un dendrograma. Es plantegen aleshores diverses qüestions: quan aquestes mesures indiquen un bon ajust? Hi ha altres mesures de l'ajust adequades o que s'hagin de considerar? Per què no utilitzar també altres mesures de la distorsió emprades en altres mètodes? Preguntes en efecte interessants, però que no tenen una fàcil resposta.

No ha estat pretensió dels autors solucionar aquestes qüestions, sinó evidenciar la prudència amb la qual han de ser considerades aquestes mesures. Amb aquesta finalitat s'han emprat, a part de les mesures ja esmentades, els següents coeficients:

- El *STRESS*, proposat per Kruskal (1964) i utilitzat com a coeficient a minimitzar per diversos programes de MDS no mètrica, com per exemple el KYST (Kruskal *et al.*, 1973)

$$S = \left[\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}\right]^{1/2} = \left[\frac{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}\right]^{1/2}$$

  on $\hat{d}_{ij}$ seran en el nostre cas les interdistàncies en dimensió k per a la MDS i les ultramètriques en les anàlisis de conglomerats. La utilització com a mesura d'ajust està ampliament justificada per la següent raó: si definim l'error quadràtic de la representació per una parella d'objectes (O_i, O_j) com $e_{ij}^2 = (d_{ij} - \hat{d}_{ij})^2$, el numerador de S^2 és aleshores l'error quadràtic total. El terme del denominador "normalitza" S i permet que sigui una mesura estandarditzada. Poden establir-se dues altres relacions interessants:

  a) S pot ser expressat com el quocient de les normes de dos vectors

$$S = \frac{\|\boldsymbol{d} - \hat{\boldsymbol{d}}\|}{\|\boldsymbol{d}\|}$$

on $\boldsymbol{d} = (d_{12}, \ldots, d_{1n}, d_{23}, \ldots, d_{2n}, d_{34}, \ldots, d_{nn})'$ i $\hat{\boldsymbol{d}}$ es defineix anàlogament.

b) $S^2 = \dfrac{\text{tr}\,(\boldsymbol{D} - \hat{\boldsymbol{D}})^2}{\text{tr}\,\boldsymbol{D}^2} = \dfrac{\text{tr}\,(\boldsymbol{B} - \hat{\boldsymbol{B}})}{\text{tr}\,\boldsymbol{B}} + \dfrac{2\,\text{tr}\,(\hat{\boldsymbol{D}}^2 - \boldsymbol{D}\hat{\boldsymbol{D}})}{\text{tr}\,\boldsymbol{D}^2}$

- El *S−STRESS*, plantejat per Takane *et al.* (1977) i Young i Lewyckyj (1979), utilitzat en l'algorisme ALSCAL de MDS no mètric

$$SS = \left[\frac{\sum_{i<j}(d_{ij}^2 - \hat{d}_{ij}^2)^2}{\sum_{i<j}(d_{ij}^2)^2}\right]^{1/2}$$

S'observa que la diferència consisteix en el fet que les distàncies són elevades al quadrat. Poden realitzar-se consideracions similars a la mesura anterior

a) $SS = \dfrac{\|\boldsymbol{d}^2 - \hat{\boldsymbol{d}}^2\|}{\|\boldsymbol{d}^2\|}$ b) $SS^2 = \dfrac{\text{tr}\,(\boldsymbol{A} - \hat{\boldsymbol{A}})^2}{\text{tr}\,\boldsymbol{A}^2}$

on $\boldsymbol{d}^2 = (d_{12}^2, \ldots, d_{1n}^2, d_{23}^2, \ldots, d_{2n}^2, d_{34}^2, \ldots, d_{nn}^2)'$ i $\hat{\boldsymbol{d}}^2$ es defineix anàlogament.

- coeficient d'alienació K, proposat per Guttman (1968)

$$K = \sqrt{1 - \mu^2}$$

on μ és l'anomenat coeficient de monotonicitat

$$\mu = \frac{\sum_{i<j} d_{ij}\hat{d}_{ij}}{\left[\sum_{i<j} d_{ij}^2 \sum_{i<j} \hat{d}_{ij}^2\right]^{1/2}}$$

Pot observar-se que $\mu = \cos\phi$ on ϕ és l'angle format pels vectors $\boldsymbol{d}$ i $\hat{\boldsymbol{d}}$ i per tant $K = \sin\phi$. L'única diferència entre μ i r_c és que en aquest darrer coeficient les distàncies són centrades. S'acompleix també la relació:

$$K^2 = \frac{\text{tr}\,(\boldsymbol{B} - \hat{\boldsymbol{B}})}{\text{tr}\,\boldsymbol{B}} - \frac{\text{tr}\,(\boldsymbol{D}\hat{\boldsymbol{D}} + \hat{\boldsymbol{D}}^2)\ \text{tr}\,(\boldsymbol{D}\hat{\boldsymbol{D}} - \hat{\boldsymbol{D}}^2)}{\text{tr}\,\boldsymbol{D}^2\ \text{tr}\,\hat{\boldsymbol{D}}^2}$$

Hem trobat per tant les relacions entre tres mesures considerades característiques de mètodes no mètrics de MDS i que poden ser també adequades per a mètodes mètrics i per a l'anàlisi de conglomerats. Així mateix, la correlació cofenètica pot ser una mesura més a considerar en el cas de la MDS.

6. PRESENTACIÓ I ANÀLISI DELS RESULTATS

En la taula 2 poden trobar-se els valors propis obtinguts en diagonalitzar $\boldsymbol{B} = \boldsymbol{HAH}$ i els coeficients c_1 i c_2. Pot observar-se que el rang de $\boldsymbol{B}$ és 13 i que $\boldsymbol{D}$ no és euclidiana, ja que han aparegut valors propis negatius. No obstant això, si considerem sols les 11 coordenades reals, la distorsió provocada és pràcticament negligible. Efectivament, $c_1 = 98,1\%$ i $c_2 = 99.8\%$, per tant podem obviar els valors propis negatius i aproximar $\boldsymbol{B}$ per $\boldsymbol{B}^*$.

Taula 2.
Valors propis obtinguts en diagonalitzar $\boldsymbol{B} = \boldsymbol{HAH}$ i coeficients c_1 i c_2 indicadors de la dispersió explicada (vegeu-ne la definició en el text).

λ	c_1	c_2	λ	c_1	c_2
28,24	27,3	47,0	3,83	95,0	99,4
19,44	46,0	69,3	2,38	97,3	99,7
13,80	59,4	80,5	0,76	98,0	99,8
12,35	71,3	89,5	0,06	98,1	99,8
8,80	79,8	94,0	0	98,1	99,8
7,64	87,2	97,5	−0,06	98,1	99,8
4,24	91,3	98,6	−1,94	100,0	100,0

La figura 1 mostra la representació MDS en $\mathbb{R}^2$ obtinguda al considerar les dues primeres coordenades. La primera coordenada presenta un "pes" considerable, però és fonamentalment degut a la clara separació entre el finlandès i la resta de llengües. Examinant les mesures d'ajust (vegeu les taules 2, 3 i 4), podem veure que considerar sols les dues primeres coordenades comporta una elevada pèrdua d'informació. Atès que la primera coordenada correspon pràcticament a la patent separació del finlandès, hem representat la resta de llengües en un diagrama tridimensional on els eixos són la segona, tercera i quarta coordenades. El significat i aportació de cada una de les coordenades pot sintetitzar-se de la manera següent:

- Primera coordenada: separació indubtable del finlandès com a llengua no comparable a cap de les altres.
- Segona coordenada: poden observar-se dos grups força evidents. Un grup està constituït pel català, castellà, gallec, francès, italià i polonès; l'altre és format per la resta de les llengües exceptuant el finlandès.

- Tercera coordenada: mostra una clara separació del polonès enfront del català, castellà, francès, gallec i italià. Per l'altre costat, basc i hongarès es mantenen properes, però ja es diferencien clarament de l'alemany, anglès, danès holandès i noruec.
- Quarta coordenada: basc i hongarès són una a cada extrem i per tant es fa evident que són dues llengües ben diferents.

Taula 3.
Diferents mesures d'ajust calculades per la representació MDS ($k = 2$ i $k = 4$) i els dendrogrames (vegeu la definició i el significat de cada un dels coeficients en el text).

	MDS k=2	MDS k=4	MÍNIM	MÀXIM	UPGMA
c_1	46,0	71,3			
c_2	69,3	89,5			
S	0,426	0,232	0,142	0,311	0,077
SS	0,516	0,266	0,221	0,579	0,128
K	0,348	0,200	0,108	0,198	0,076
r_c	0,851	0,943	0,956	0,788	0,971

Taula 4.
Mesures d'ajust per a la representació MDS i els dendrogrames adequadament transformades per permetre la seva comparació.

	MDS k=2	MDS k=4	MÍNIM	MÀXIM	UPGMA
$1-c_1/100$	0,540	0,287			
$1-c_2/100$	0,307	0,105			
S^2	0,182	0,054	0,020	0,097	0,006
SS^2	0,266	0,070	0,049	0,335	0,016
K^2	0,121	0,040	0,012	0,039	0,006
$1-r_c^2$	0,275	0,111	0,086	0,379	0,057

L'aplicació dels tres algorismes d'anàlisi de conglomerats jeràrquica (mètode del mínim, del màxim i UPGMA) ha permès construir els dendrogrames representats en la figura 3. La correlació cofenètica i les altres mesures discutides en l'apartat anterior es mostren en la taula 3 i 4. Els resultats obtinguts són semblants qualitativament i no difereixen substancialment de les conclusions que s'han obtingut amb la MDS: un grup format per les llengües romàniques (català, castellà, francès, gallec i italià), el grup de llengües germàniques (alemany, anglès, danès, holandès i noruec) i quatre llengües que no poden ser agrupades (basc, finlandès, hongarès i polonès). Pot observar-se clarament en els dendrogrames fins i tot la divisió entre llengües germàniques del grup septentrional (noruec i danès) i del grup occidental (alemany, anglès i holandès). Novament s'ha fet palès l'enigma basc, testimoni viu del passat lingüístic d'occident i que tradueix la tenaç adhesió d'una petita comunitat a la seva llengua contra les fortes pressions exteriors suportades durant més de dos mil·lenis.

Finalment caldria algun comentari respecte les mesures d'ajust. Les principals conclusions que es poden extreure són les següents:

- La primera impressió que produeixen les taules 3 i 4 és una certa sorpresa: les diferències entre les distintes mesures són molt notables! Pot observar-se la variació entre c_1 i c_2, K i r_c, S i SS. La comparació alhora de totes les mesures provoca una certa incomoditat. Quina mesura és més adequada? És l'ajust bo o dolent? Com hem mencionat anteriorment no pretenem respondre aquestes preguntes, sinó fer palès el problema amb unes dades que són prou evidents. En tot cas, es desprèn d'aquests resultats que és convenient tenir en compte diverses mesures i no pas una de sola.

- De les representacions obtingudes, el dendrograma aconseguit mitjançant el mètode UPGMA és el que més s'ajusta a les dissimilituds inicials. El mètode del màxim i la representació MDS per $k = 2$ presenten unes discrepàncies considerables (es poden observar però les diferències entre els valors de K i r_c). L'ajust de S millora substancialment augmentant la dimensió i considerant les quatre primeres coordenades.

7. CONCLUSIONS

Hem pogut comprovar en aquest estudi com l'anàlisi multivariant pot esdevenir una eina extraordinàriament eficaç en una àrea de recerca aparentment llunyana dels seus objectius com és el cas d'un estudi lingüístic. La quantificació acurada d'una informació qualitativa bastant reduïda (l'escriptura dels deu

primers nombres) ha permès agrupar adequadament catorze llengües europees, separant clarament les romàniques de les germàniques septentrionals i orientals. No ha estat l'objectiu obtenir conclusions rigoroses des d'un punt de vista lingüístic, sinó mostrar una metodologia que s'ha revelat apropiada. Un estudi més ampli i acurat, incorporant més llengües i un ventall més ampli de paraules, podria probablement aportar resultats encara més interessants. Tècniques semblants poden estendre's a d'altres estudis com ara l'evolució històrica d'una llengua o les variacions dialectals. La possibiliat de digitalitzar el so obre un altre possible camp d'aplicacions més sofisticades però essencialment fonamentades en mètodes similars. No obstant això, és el col·lectiu de lingüistes qui de ben segur sabrà trobar noves i més profitoses aplicacions.

8. AGRAÏMENTS

Els autors volen agrair molt especialment al Dr. C. M. Cuadras els suggeriments efectuats i el seu suport en el decurs de l'elaboració d'aquest treball. També agraïm la col·laboració de R. Serrano.

9. REFERÈNCIES BIBLIOGRÀFIQUES

[1] **Borg, I.**, i **Lingoes, J.** (1987). *Multidimensional Similarity Structure Analysis.* Springer-Verlag: New York.

[2] **Cuadras, C.M.** (1991). *Métodos de Análisis Multivariante.* PPU: Barcelona.

[3] **Davison, M.L.** (1983). *Multidimensional Scaling.* Wiley: New York.

[4] **Dillon, W.R.**, i **Goldstein, M.** (1984). *Multivariate Analysis. Methods and Applications.* Wiley: New York.

[5] **Eckart, C.**, i **Young, G.** (1936). "Approximation of one matrix by another of lower rank". *Psychometrika,* **1**, 211–218.

[6] **Farris, J.S.** (1969). "On the cophenetic correlation coefficient". *Syst. Zool.,* **18(3)**, 279–285.

[7] **Gower, J.C.** (1966). "Some distances properties of latent roots and vector methods used in multivariate analysis". *Biometrika,* **53**, 325–338.

[8] **Guttman, L.** (1968). "A general nonmetric technique for finding the smallest coordinate space for a configuration of points". *Psychometrika,* **33**, 469–504.

[9] **Jardine, N.**, i **Sibson, R.** (1968). "The construction of hierarchic and non-hierarchic classifications". *Comput. J.*, **11**, 177–184.

[10] **Johnson, S.C.** (1967). "Hierarchical clustering schemes". *Psychometrika,* **32**, 241–254.

[11] **Johnson, R.A.**, i **Wichern, D.W.** (1988). *Applied Multivariate Statistical Analysis.* Prentice-Hall International: New Jersey.

[12] **Kruskal, J.B.** (1964). "Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis". *Psychometrika,* **29**, 1–28, 115–129.

[13] **Kruskal, J.B., Young, F.W.**, i **Seery, S.B.** (1973). *How to use KYST, a very flexible program to do multidimensional scaling and unfolding.* Murray Hill: Unpublished manuscript, Bell Laboratories.

[14] **Lance, G.N.**, i **Williams, W.T.** (1966). "Computer programs for hierarchical polythetic classification ('similarity analysis')". *Comput. J.*, **9**, 60–64.

[15] **McQuitty, L.L.** (1964). "Capabilities and improvements of linkage analysis as a clustering method". *Educ. Psychol. Meas.*, **24**, 441–456.

[16] **Mardia, K.V.** (1978). "Some properties of classical multidimensional scaling". *Comm. Statist.-Theor. Meth.*, **A 7**, 1233–1241.

[17] **Mardia, K.V., Kent, J.T.**, i **Bibby, J.M.** (1979). *Multivariate Analysis.* Academic Press: London.

[18] **Richardson, M.W.** (1938). "Multidimensional psychophysics". *Psychol. Bull.,* **35**, 659–660.

[19] **Saito, T.** (1978). "The problem of the additive constant and eigenvalues in metric multidimensional scaling". *Psychometrika,* **43**, 193–201.

[20] **Schoenberg, I.J.** (1935). "Remarks to Maurice Fréchet's article 'Sur la definition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert'". *Ann. Math.*, **36**, 724–732.

[21] **Seber, G.A.F.** (1984). *Multivariate Observations.* Wiley: New York.

[22] **Sneath, P.H.A.** (1957). "The application of computers to taxonomy". *J. Gen. Microbiol.*, **17**, 201–226.

[23] **Sokal, R.R.**, i **Michener, C.D.** (1958). "A statistical method for evaluating systematic relationships". *Univ. Kansas Sci. Bull.,* **38**, 1409–1438.

[24] **Sokal, R.R.**, i **Sneath, P.H.A.** (1963). *Principles of Numerical Taxonomy.* Freeman: San Francisco.

[25] **Sokal, R.R.**, i **Rohlf, F.J.** (1962). "The comparison of dendrograms by objective methods". *Taxon,* **11**, 33–40.

[26] **Takane, Y., Young, F.W.**, i **De Leeuw, J.** (1977). "Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features". *Psychometrika,* **42**, 7–67.

[27] **Torgerson, W.S.** (1952). "Multidimensional scaling: I-theory and method". *Psychometrika,* **17**, 401–419.

[28] **Torgerson, W.S.** (1958). *Theory and Methods of Scaling.* Wiley: New York.

[29] **Young, F.W.**, i **Lewyckyj, R.** (1979). *ALSCAL 4 User's Guide.* Data Analysis and Theory Associates: Chapel Hill, NC.