ESTEVE CLUA JULVE

DISTÀNCIA LINGÜÍSTICA I CLASSIFICACIÓ DE VARIETATS DIALECTALS

1. DELIMITACIÓ I CLASSIFICACIÓ DE VARIETATS DIALECTALS

Tradicionalment, els estudis dialectals, seguint criteris bàsicament qualitatius, han basat la delimitació de varietats geogràfiques en la noció d'isoglossa –línia imaginària que separa en un mapa dues zones divergents en relació amb un tret lingüístic determinat.[1] De vegades una única isoglossa, de vegades un conjunt d'isoglosses coincidents, formant un feix, són a la base de la majoria de partions dialectals amb què treballem. Posteriorment, hom va començar a tenir en compte els feixos d'isoglosses que dibuixaven conjuntament, d'una forma més o menys nítida, les fronteres dialectals; en aquesta segona etapa, tot i que continuava predominant el punt de vista qualitatiu, es començava a valorar l'aspecte quantitatiu, en el sentit que una frontera dialectal era més important segons el nombre d'isoglosses que constituïen el feix que la marcava. Finalment, amb la possibilitat de tractar informàticament les dades lingüístiques, han proliferat els estudis bàsicament quantitatius, que, en principi, poden facilitar la delimitació dels dialectes a partir del tractament estadístic de les similituds o les diferències lingüístiques detectades en un conjunt ampli de dades.

Sens dubte, la noció d'isoglossa ha tingut, i continua tenint, una importància cabdal en l'anàlisi i en la descripció de la variació lingüística diatòpica. Des d'una perspectiva que concep les varietats dialectals com a sistemes lingüístics, però, no és tan clar que

(1) Chambers & Trudgill (1980: 104-105) proposen també el terme *heteroglossa*, línia doble que marca dues zones divergents quant a un tret lingüístic, que permet deixar en el centre una zona neutra. La finalitat d'aquesta noció és ajustar-se més a la realitat geogràfica en el cas d'enquestes no exhaustives. Les poblacions no enquestades quedarien en aquesta zona neutra entre les dues línies de l'heteroglossa.

Caplletra 26 (Primavera 1999), pp. 11-26

aquesta noció constitueixi una eina adequada per a la delimitació i la classificació dialectals. Si més no, presenta alguns punts febles que cal tenir en compte.

El principal retret que s'ha fet a la noció d'isoglossa en tant que criteri bàsic per a la delimitació de varietats dialectals té a veure amb l'arbitrarietat que implica. Efectivament, les divisions d'àrees dialectals establertes mitjançant aquest mètode solen basar-se en un petit nombre de trets lingüístics, que són el resultat d'una tria mínima entre el gran conjunt de trets que constitueixen els sistemes lingüístics de les àrees analitzades. El problema rau en el fet que no disposem d'una jerarquització d'isoglosses prou fonamentada que justifiqui una determinada tria de característiques lingüístiques a l'hora d'establir les agrupacions dialectals; com a molt, es pot establir una ordenació de les isoglosses en funció de la seva rellevància estructural.[2] Per tant aquesta metodologia comporta ineludiblement una dosi important de subjectivitat per part de l'investigador.

La voluntat d'eludir aquesta arbitrarietat en la selecció dels elements lingüístics, sobre la base dels quals s'estableix la delimitació dialectal, és una de les raons principals que ha portat els investigadors a basar la determinació de varietats lingüístiques en el criteri quantitatiu.[3] Des d'aquesta perspectiva, la delimitació i la classificació dialectals es duen a terme a partir de l'aplicació de tractaments estadístics a tot el conjunt de dades que pot fornir una enquesta dialectal; a partir de la quantificació de les similituds o les diferències entre les varietats estudiades es pot comprovar la distància lingüística que les separa i, per tant, establir-ne la classificació, una classificació que serà el resultat del tractament sintètic i global del conjunt de dades lingüístiques.

2. EL CRITERI QUANTITATIU. LA DIALECTOMETRIA

Des del principi dels anys setanta s'ha desenvolupat i aplicat un nombre considerable de mètodes de classificació i delimitació dialectal basats en criteris quantitatius, que solen compartir el fet de tractar de forma sintètica i global el conjunt de dades d'un atles dialectal amb finalitats classificatòries. També els caracteritza el fet d'utilitzar anàlisis estadístiques multivariants per a la quantificació i el tractament de les dades; amb el terme anàlisi multivariant (*Multivariate Analysis*) hom fa referència a un nombre considerable de tècniques de descripció i d'anàlisi d'un conjunt d'elements a partir de diverses variables observades, sense que, *a priori*, s'atorgui un estatus qualitatiu especial a cap d'elles.

Tots aquests mètodes se solen agrupar sota el rètol de dialectometria.[4] Aquesta disciplina apareix al principi de la dècada dels setanta en l'àmbit de la lingüística romànica, lligada estretament a la geolingüística.[5] I es caracteritza per l'abandó de la noció d'isoglossa i l'adopció del concepte de distància lingüística com a eina bàsica de la classificació dialectal. El concepte de distància fou manllevat de l'àmbit científic de l'anàlisi de dades, en el qual s'associa –en general– a la quantificació de les similituds o les diferències entre individus, poblacions o grups de poblacions.

(2) Chambers & Trudgill (1980: 115).

(3) Vegeu Veny (1992: 205-206). Un altre dels avantatges del mètode quantitatiu, que sovint s'esmenta, té relació amb l'aprofitament *màxim* de les dades proporcionades per les enquestes dialectals, en contraposició al mètode qualitatiu, que en moltes ocasions suposava una subexplotació del gran cabal de dades dels atles lingüístics; vegeu Viereck (1988: 530; 1987: 11).

(4) De vegades, però, fora de l'àmbit de la lingüística romànica, s'han fet distincions entre els mètodes utilitzats originàriament pels «pares» de la disciplina (Séguy, Guiter i Goebl), que reben la qualificació de dialectomètrics, i altres mètodes, com ara l'anomenat anàlisi tipològica o anàlisi de conglomerats (Cluster Analysis) o el conegut com a *Multidimensional Scaling*; vegeu Viereck (1987:1; 1988:537).

(5) El primer a utilitzar el terme va ser Jean Séguy, que juntament amb Henri Guiter ha estat considerat el pare de la disciplina. Vegeu Séguy (1971 i 1973) i Guiter (1973). L'impuls més important d'aquesta disciplina es deu, però, a Hans Goebl, que amb una sòlida base de taxonomia numèrica ha anat polint els seus propis mètodes dialectomètrics a partir de treballs sobre materials de l'*Atlas Linguistique de la France* (*ALF*), de l'*Atlante Italo–svizzero* (*AIS*) i, fins i tot, del *Survey of English Dialects* (*SED*). Vegeu Goebl (1992).

L'aplicació de mètodes estadístics a ciències diverses és un fenomen molt corrent. La biologia, la medicina, l'economia, la psicologia –per esmentar-ne algunes– s'han beneficiat de l'aplicació de models propis de l'anàlisi de dades. En molts d'aquests casos es tractava d'establir classificacions de determinades entitats a partir de l'anàlisi multivariant; és a dir, de l'anàlisi de mesures associades a diferents factors o variables, que permeten establir una estructura d'interdistàncies entre les diferents entitats analitzades. La dialectometria, en part, és el resultat d'aquesta comunicació interdisciplinària entre la dialectologia geogràfica i l'anàlisi de dades.

Els estudis dialectomètrics s'han multiplicat darrerament arreu del món. La possibilitat generalitzada de disposar d'eines informàtiques cada cop més potents, que permeten el tractament de grans quantitats de dades i la realització de càlculs extensos, ha facilitat aquesta proliferació.[6] A la vegada, també s'ha ampliat molt l'espectre de tècniques que es fan servir per al càlcul de la distància lingüística. Així, al costat dels mètodes primerencs, com ara l'anomenat global de Guiter o l'interpuntual de Goebl, han aparegut nous mètodes basats en diferents tècniques de l'anàlisi estadística multivariant. Entre aquestes, les més utilitzades són les d'anàlisi de conglomerats (*Cluster Analysis*) –que a la vegada es pot subclassificar en tres versions bàsiques, segons el tipus d'algorisme matemàtic utilitzat per a la representació dendrogràfica de la matriu de dissimilitud:[7] *Single Linkage*,[8] *Complete Linkage*[9] i *Average Linkage* (UPGMA)[10]– i el *Multidimensional Scaling*.[11]

És evident que els mètodes dialectomètrics permetem eludir l'arbitrarietat que suposa la selecció d'un conjunt molt reduït de trets lingüístics com a base de la delimitació dialectal. Tanmateix, seria il·lusori pensar que el fet de defugir l'arbitrarietat en aquest aspecte concret implica la manca total de tries subjectives per part de l'investigador durant tot el procés. La necessitat d'elecció amb un cert grau de subjectivitat acompanya l'investigador en molts moments del procés classificatori: en la definició de l'enquesta dialectal, en el tipus d'anàlisi lingüística a partir de la qual establirà la comparació, en el tipus d'algorisme matemàtic utilitzat per establir els agrupaments de les diferents varietats lingüístiques... Intentar minimitzar les possibilitats d'arbitrarietat que encara implica actualment és una tasca fonamental per al desenvolupament d'aquesta metodologia, que a hores d'ara considerem la més adequada per determinar la classificació de les varietats dialectals.

3. MANCANCES DELS MÈTODES QUANTITATIUS

L'aplicació de criteris quantitatius a l'hora d'establir les agrupacions dialectals no ha estat mancada de crítiques. Algunes d'aquestes, provinents de la dialectologia tradicional, fan referència al menyspreu dels aspectes qualitatius. És evident que, malgrat la impossibilitat de disposar d'una jerarquització qualitativa dels diferents trets lingüístics –pertinent per a la classificació dialectal–, sembla clar que certs tipus de

(6) Vegeu a Goebl (1992: 433-434) un recull bibliogràfic de treballs dialectomètrics. Pel que fa a les aplicacions d'aquests mètodes a la llengua catalana, cal destacar el treball de Polanco (1992), que aplica els mètodes de Guiter i de Goebl a les dades de l'*ALPI*. Altres treballs dialectomètrics aplicats al català són:

Costa, J. (1977) «Aproximació lingüística al català de Cerdanya», dins *Actes del 8è centenari de la fundació de Puigcerdà*, Puigcerdà.

Guiter, H . (1978) «Panorama lingüístic des de Besalú», dins *Annals del Patronat d'Estudis Històrics d'Olot i Comarca*, Olot, pàgs. 35-48.

Sardà, A. & Guiter, H. (1975) «L'Atlas Lingüístic de la Catalunya i la fragmentació dialectal del català», dins *Miscellanea Barcinonensia* 40, pàgs. 93-112.

Tots ells basats en procediments dialectomètrics de les primeres èpoques de la disciplina. I, finalment, cal esmentar el treball d'Ortega (1998), que aplica el mètode del *Cluster Analysis* a la classificació dels parlars locals de la Marina.

(7) Per a una discussió sobre aquests aspectes, vegeu Sneath & Sokal (1973: 228-240), Woods *et alii* (1987: 260-261) i Manly (1986: 101-104).

(8) Vegeu Shaw (1974) i Morgan & Shaw (1982).

(9) Vegeu Goebl (1992) i (1997).

(10) Vegeu Inoue & Kasai (1989), Hoppenbrowers (1993), Sibata & Kumagai (1993) i Viaplana (1997).

(11) Vegeu Hoppenbrowers (1993) i Inoue & Fukushima (1997).

diferències lingüístiques són més rellevants que d'altres. Per tant, creiem que un tractament quantitatiu ha de tenir en compte, d'alguna manera, determinats paràmetres qualitatius.[12] L'únic camí possible que actualment ens pot permetre avançar en aquesta direcció passa forçosament per tenir en compte, d'una banda, la diferent rellevància estructural dels elements lingüístics que es tenen en compte per a la classificació dialectal i, de l'altra, la coherència estructural del conjunt de dades lingüístiques en relació amb els objectius classificatoris. Això vol dir que, si pretenem establir la classificació de determinades varietats lingüístiques a partir, per exemple, d'un conjunt de dades de caràcter bàsicament lèxic, els resultats no podran ser mai adients, encara que el tractament estadístic de les dades sigui totalment objectiu i adequat. El resultat d'aquesta anàlisi seria una classificació dialectal restringida a la variació lèxica de les zones estudiades, però no s'hauria de pretendre fer-la passar per una classificació del conjunt de les varietats lingüístiques respectives. Per fer una classificació global, les dades haurien de contenir, com a mínim, un conjunt prou representatiu dels aspectes més sistemàtics de la llengua: fonològics, morfològics, sintàctics... I, si és possible, un conjunt prou representatiu dels diferents estrats que constitueixen el sistema lingüístic. En una anàlisi quantitativa aquests aspectes s'han de tractar amb molta més cura que no pas en un tractament qualitatiu, sobretot en els tractaments quantitatius que, *a priori*, consideren totes les variables com si fossin del mateix rang.

Un altre problema que presenten sovint els estudis d'aquesta mena té a veure amb una certa utilització indiscriminada de les dades d'estudis geolingüístics, que no han estat concebuts per ser analitzats amb tècniques de caràcter quantitatiu. De fet, una de les finalitats dels mètodes dialectomètrics –inicialment– era poder tractar la gran quantitat de dades dels atles lingüístics, perquè aquests no es convertissin en 'cementiris de dades'.[13] Un tractament quantitatiu requereix que el conjunt de dades sobre el que s'aplica sigui representatiu de les varietats lingüístiques o dels estrats lingüístics que es vol analitzar. I aquesta no és una característica gaire comuna en alguns dels estudis geolingüístics que han estat tractats amb mètodes dialectomètrics, molt centrats en aspectes lèxics no sistemàtics.[14]

En alguns tractaments dialectomètrics hom hi troba a faltar una anàlisi lingüística coherent en la qual basar la transposició de les dades fonètiques, obtingudes a partir del qüestionari lingüístic, a les variables de comparació, que serveixen de base per al tractament quantitatiu. Atesa la gran variació que es pot donar en qualsevol estudi dialectal, els tractaments dialectomètrics han tendit a una certa 'tipificació' de les dades –és a dir, a una certa simplificació de la variació observada–; si aquesta 'tipificació' no recolza sobre criteris lingüístics coherents, pot ocórrer que es barregin criteris incompatibles (sincrònics i diacrònics, interdialectals i intradialectals) o que les variables en què recolza el procés classificatori no reflecteixin adequadament la variació real observada.

I, finalment, hi ha algunes qüestions relacionades directament amb el tractament estadístic de les dades que poden estar sotmeses excessivament a la subjectivitat de

(12) Estem, doncs, d'acord amb Veny (1992:214) en el sentit que els mètodes quantitatius s'han de complementar d'alguna manera amb aspectes qualitatius.

(13) Vegeu Viereck (1987: 11).

(14) Vegeu en aquest sentit Veny (1986: 26-27), que amb encert qüestiona l'aplicació del mètode Guiter al català –Sardà & Guiter (1975)–, tot remarcant que el problema d'aquesta aplicació sembla raure més *en la fragilitat dels materials que els van servir de base (ALC)*, que no pas en el mètode quantitatiu.

l'investigador i que han estat sovint el centre de moltes crítiques realitzades als mètodes dialectomètrics.[15] Concretament la tria del tipus de mesura de similitud, a partir de la qual s'estableix la distància lingüística, i de l'algorisme de classificació, que permet transformar la distància observada a partir de la mesura de similitud en una distància ultramètrica i obtenir la posterior representació dendrogràfica, implica un grau considerable d'arbitrarietat. L'investigador pot triar entre diferents possibilitats de mesura de similituds o d'algorisme classificatori, que donaran resultats amb diferències considerables. Per això és indispensable utilitzar mesures de control que permetin contrastar la fiabilitat de les classificacions obtingudes.

4. UNA PROPOSTA PER AL CÀLCUL I LA REPRESENTACIÓ DE LA DISTÀNCIA LINGÜÍSTICA

Tenint en compte aquests factors, creiem que la millora de l'adequació dels mètodes de classificació dialectal basats en criteris quantitatius passa per tenir en compte les característiques següents. En principi s'han de basar en el concepte de distància lingüística, entesa com a mesura de les similituds/divergències existents entre diferents varietats lingüístiques i s'han de servir de tècniques d'anàlisi multivariant per determinar la distància lingüística a partir de múltiples variables.

En el disseny dels diferents estadis del procés classificatori s'ha de tenir en compte el posterior tractament quantitatiu de les dades. Així, d'una banda, el punt de partida del procés ha de ser l'anàlisi de la variació d'un conjunt de dades representatives dels sistemes lingüístics o dels àmbits que es vulguin comparar, la qual cosa permet introduir consideracions qualitatives en el tractament quantitatiu. I de l'altra, cal minimitzar els aspectes distorsionadors dels corpus de dades analitzats (ens referim, per exemple, als casos de respostes nul·les, errònies o múltiples que poden aparèixer en les enquestes dialectals), que pertorben l'anàlisi estadística i poden incidir negativament en els resultats.

Les variables de comparació emprades per a la classificació de les varietats lingüístiques han de ser el resultat de l'aplicació d'una anàlisi lingüística –basada en els principis de la fonologia generativa– a les dades obtingudes a partir de les entrevistes. En aquest sentit, per exemple, una anàlisi lingüística de tipus generatiu permet simplificar i sistematitzar la variació observada, alhora que possibilita la distinció entre fenòmens variacionals de caràcter subjacent i fenòmens variacionals superficials.

Per últim, atesa la gran variabilitat que es pot produir en els resultats del tractament estadístic segons el tipus de mesura de similitud emprada o l'algorisme utilitzat per a la representació geomètrica de la distància observada, cal utilitzar mesures que permetin contrastar la fiabilitat dels resultats obtinguts.

Una metodologia d'aquestes característiques és la que s'aplica a Clua (1998),[16] treball que ens permetrà exemplificar tot el procés de classificació. Aquest estudi, en

(15) Vegeu Viereck (1988: 542-547), on es critiquen certes aplicacions dialectomètriques en l'àmbit dels dialectes anglesos; concretament les crítiques d'aquest autor s'adrecen a l'arbitrarietat que aquest mètode pot implicar quan es tracta d'establir una determinada interval·lització de les distàncies lingüístiques de cara a establir els grups dialectals.

(16) Aquest estudi té el seu punt de partida en Viaplana (1997), treball de caràcter interdisciplinari que combina diferents aspectes de la dialectologia geogràfica, de la dialectometria, de la lingüística generativa i de l'anàlisi de dades, amb l'objectiu de descriure els aspectes més rellevants de la morfologia regular de les varietats nord-occidentals i de proposar-ne una classificació a partir de la variació analitzada.

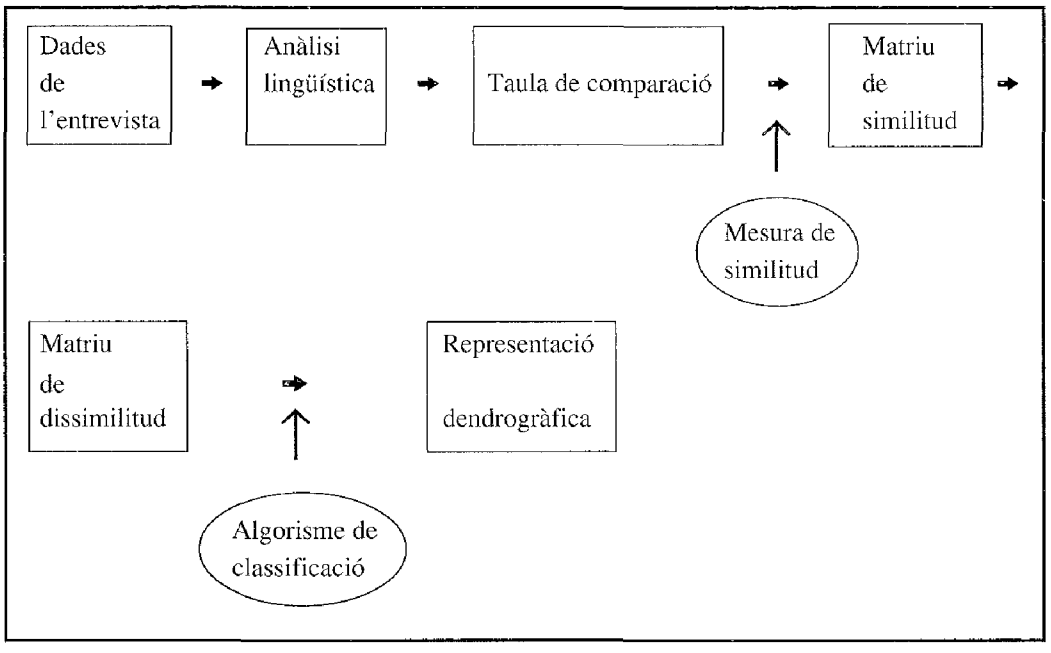
el qual es proposa una classificació de les varietats valencianes a partir de la morfologia flexiva, parteix de la idea que l'afinitat entre dues varietats és inversament proporcional a les divergències existents entre els sistemes lingüístics respectius. Aquestes divergències entre dos sistemes constitueixen el que anomenem distància lingüística, que com hem vist és una noció manllevada de l'anàlisi de dades. La major o menor distància lingüística és el factor que determina l'agrupament de les diferents varietats en classes (sub)dialectals. Atès que les característiques lingüístiques a partir de les quals s'estableix la distància (variables de comparació) són múltiples i que les dades resultants de l'anàlisi lingüística són considerables, ens servim de tècniques d'anàlisi estadística multivariant per poder-les tractar adequadament.

En principi, en aquest estudi no es tractava d'aplicar una determinada tècnica quantitativa a les dades d'una entrevista preexistent o als mapes d'un determinat atles dialectal, sinó que tant el disseny del qüestionari com la realització de l'entrevista, com els altres estadis de la investigació, s'han realitzat tenint en compte que un dels objectius finals era el tractament quantitatiu de la variació lingüística observada. Això vol dir que, per exemple, en el disseny del qüestionari –un cop descartat, per inviable, el tractament exhaustiu de la flexió– es va intentar que abastés una mostra prou representativa d'aquest àmbit lingüístic, que ens permetés arribar a conclusions generalitzables. El conjunt de dades analitzades comprèn exhaustivament les formes dels articles, les formes simples dels clítics pronominals i els diferents paradigmes dels verbs regulars; pel que fa als adjectius, ultra les formes regulars, s'analitzen els casos més generals entre els «irregulars». Creiem, doncs, que es tracta d'un conjunt representatiu de la flexió valenciana.

Quant a l'entrevista, després d'una primera anàlisi global de les dades obtingudes en l'aplicació principal, se'n va realitzar una de comprovació. Entre els objectius no secundaris d'aquesta segona recollida de dades hi havia el d'intentar cobrir les respostes deficients o nul·les de la primera entrevista, ja que aquest tipus de mancances pot distorsionar els resultats d'un tractament quantitatiu com l'utilitzat en aquest estudi.

Tot seguit presentem un esquema del procés seguit per a la classificació de les varietats valencianes. I, a continuació, n'expliquem els aspectes més rellevants des de l'òptica de l'anàlisi quantitativa.

(1)



El procés s'inicia, evidentment, amb la selecció del qüestionari, la determinació de l'àmbit geogràfic i del tipus i la quantitat dels informants, la realització de l'entrevista i la transcripció de les respostes.

L'anàlisi lingüística de les dades constitueix el següent pas del procés. Aquest és un factor força rellevant en l'anàlisi multivariant, ja que la matriu de comparació no està constituïda per les dades fonètiques resultants del procés de transcripció de les respostes de l'entrevista, sinó que els elements que la componen són el resultat de l'aplicació de l'anàlisi lingüística a aquestes dades. L'anàlisi lingüística emprada es basa en els principis de la fonologia generativa, que es caracteritza per analitzar la matèria sonora de les llengües a partir de dos nivells: el de les formes subjacents o bàsiques, constituïdes per segments fonològics, i el de les formes fonètiques o superficials. La relació entre ambdós nivells s'explica mitjançant els processos fonològics que permeten derivar les realitzacions fonètiques de les formes subjacents. D'aquesta manera s'aconsegueix simplificar la descripció lingüística i, a la vegada, es poden establir generalitzacions sobre el funcionament i l'estructura de les varietats lingüístiques.

Des d'aquest enfocament, la variació morfològica existent entre dues varietats pot ser bàsicament de dos tipus: de tipus morfològic, que implica l'existència de segments morfològics subjacents diferents; o de tipus fonològic, en què els segments morfològics subjacents són coincidents i la diferència es produeix per divergències en els processos fonològics que deriven les realitzacions fonètiques a partir de les representacions

fonològiques dels morfemes. D'acord amb aquests supòsits, l'anàlisi lingüística desenvolupada en aquest treball té com a finalitat la descripció explícita de les dades recollides en l'entrevista, mitjançant l'establiment de les formes subjacents i la determinació dels processos fonològics que relacionen aquestes amb les formes superficials.

A partir, doncs, de les formes subjacents i dels processos fonològics definits s'estableixen les taules de comparació, que són la base per poder definir posteriorment la similitud i la distància lingüístiques. Les taules de comparació posen en relació un conjunt de variables –en aquest cas, formes subjacents i processos fonològics– amb un conjunt d'individus – els informants– i presenten l'estructura següent:

(2)

Variables / Individus	X_1 X_2 X_3 ... ... X_p
Informant 1	X_{11} X_{12} X_{13} ... ... X_{1p}
Informant 2	X_{21} X_{22} X_{23} ... ... X_{2p}
Informant 3	X_{31} X_{32} X_{33} ... ... X_{3p}
...	... ...
Informant *n*	X_{n1} X_{n2} X_{n3} ... ... X_{np}

A partir de la taula de comparació, i mitjançant el paquet informàtic SAS,[17] es van elaborar les matrius de similituds, que constitueixen el següent pas del procés quantitatiu. Abans, però, com es pot veure a l'esquema (1), cal establir la mesura de similitud a partir de la qual el programa informàtic comptabilitzarà les coincidències i establirà les similituds entre els informants.

L'elecció de la mesura o índex de similitud té una gran rellevància per al resultat final del procés, ja que segons el tipus de mesura establert els resultats poden variar considerablement; sobretot quan les matrius de comparació presenten caselles nul·les, a causa de respostes errònies en l'entrevista o a causa de la manca de respostes, o quan les dades impliquen respostes múltiples. En aquests casos la mesura de similitud que se sol utilitzar és el percentatge de les coincidències, en relació amb el total d'elements comparats entre dues varietats.[18] En aquest estudi, el primer d'aquests factors, com ja hem assenyalat, era irrellevant, ja que no hi havia caselles buides;[19] pel que fa a les respostes múltiples, la constitució de la matriu de comparació a partir dels informants, i no de les poblacions, ens permetia tenir en compte la possibilitat de respostes diferents dins d'una mateixa varietat; d'aquesta manera se simplificava molt la possibilitat de comptar amb aquest tipus de resposta, que implica una gran complexitat en el moment d'establir la mesura de similitud.[20]

(17) Vegeu SAS Institute, Inc (1990). Per tenir una idea de la gran quantitat de dades que s'han hagut de comparar, només cal tenir present, per exemple, que el total de segments morfològics, potencials objectes de comparació, en la flexió verbal gira al voltant dels 70.000 elements –tenint en compte que hi ha en joc 17.500 formes verbals [5 verbs x 50 formes verbals x 70 informants] i que aquestes es poden fraccionar potencialment en quatre segments morfològics: extensió, tema, mode-temps i nombre-persona. Evidentment, una comparació d'aquesta mena no es pot dur a terme sense el suport electrònic adequat.

(18) Pel que fa a altres mesures de similitud utilitzades en aplicacions dialectomètriques, Vegeu Goebl (1981, 1984 I, 1993 i 1997).

(19) El fet de treballar bàsicament sobre dades de la morfologia regular, juntament amb la realització de l'entrevista de comprovació, expliquen aquesta característica.

(20) En els casos en què un mateix informant presentava respostes múltiples, es va intentar esbrinar *in situ* la variant preferent, o es va seleccionar contrastant les respostes del qüestionari amb el fragment de conversa lliure gravat.

Tenint en compte això, la mesura de similitud utilitzada és força simple i es basa en les coincidències, en relació amb les diferents variables, entre parells d'informants. Es pot definir de la manera següent:

(3) $$s(i,j) = \sum_k coin_k(i,j)$$

On $coin_k$ (i, j) pren el valor 1 quan, pel que fa a la variable lingüística k, coincideixen els informants i i j, i pren el valor 0 en cas contrari.

És a dir: la similitud entre dos informants (i, j) equival al sumatori (Σ) de coincidències en relació amb les diferents variables analitzades.

Un cop establertes les matrius de similituds i definides, per tant, les relacions d'interdistància associades a les similituds morfològiques entre les varietats analitzades, cal cercar una representació de l'estructura resultant que ens permeti visualitzar clarament les relacions de proximitat/llunyania entre aquestes varietats, ja que les matrius no permeten una bona interpretació dels resultats ni una visualització plausible de les distàncies entre varietats. Ens referim al darrer pas del procés esquematitzat a (1). Amb aquesta finalitat es va recórrer a les possibilitats que ofereix l'anàlisi de dades.

Un dels objectius de l'anàlisi de dades consisteix en la recerca, la representació i la interpretació de models que reflecteixin les analogies i diferències entre individus, poblacions o grups de poblacions. En aquest sentit, en general, es pretén aconseguir una representació en un espai de dimensió reduïda –per exemple, el pla–, de forma que produint una mínima distorsió en l'estructura original d'interdistàncies –que en el nostre cas és multidimensional– s'aconsegueixi una fàcil interpretació de les dades.

Hom sol agrupar les possibles representacions en dues classes segons el tipus d'espai geomètric utilitzat: representacions en models espacials o continus i representacions en models de xarxa o discrets.[21] Pel que fa als models continus o espacials, amb aquests models es pretén visualitzar l'estructura d'interdistàncies entre els individus o les poblacions, i per això es representen aquests com a punts d'un espai –el pla–, de forma que la distància entre els punts reflecteixi la distància entre els individus o les poblacions representades. Entre les tècniques que utilitzen aquest model de representació geomètrica destaca el *Multidimensional Scaling*.[22]

Quant als models discrets o de xarxa, la seva finalitat és posar de manifest les estructures de grups, i per això es representen els individus o les poblacions com a punts d'un espai en què tenim definits uns camins –xarxes– que permeten comunicar-los. Les longituds dels camins que uneixen les poblacions reflecteixen les distàncies entre les poblacions representades. Entre els models de xarxa, destaquen els arbres ultramètrics i els additius.[23]

Atès que un dels objectius bàsics d'aquest treball era determinar les possibles agrupacions entre varietats lingüístiques, a l'hora de representar les interdistàncies

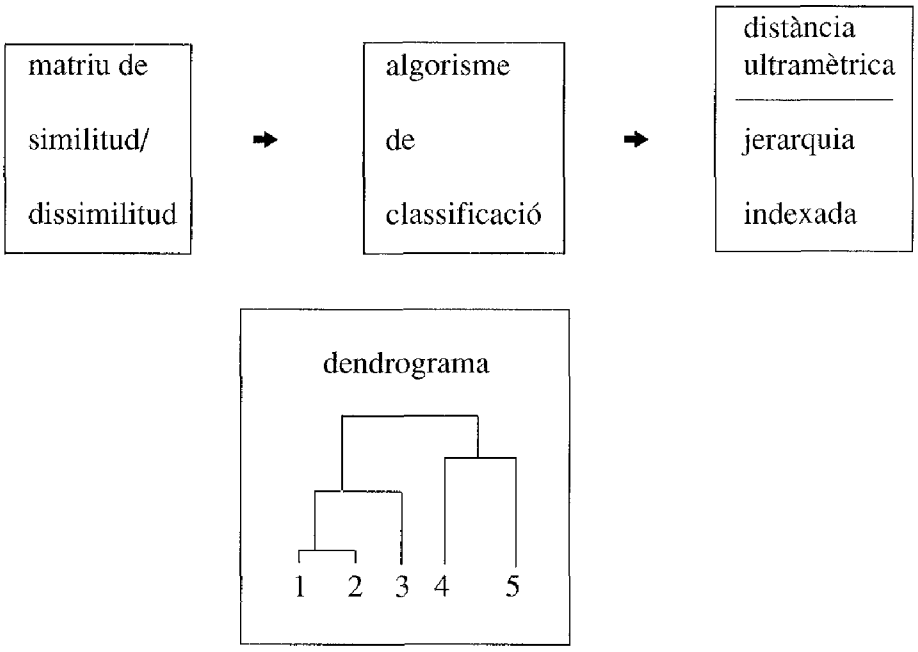
(21) Vegeu Arcas & Cuadras (1987).

(22) Vegeu Cuadras (1981: 371-418).

(23) Vegeu Arcas & Cuadras (1987).

entre les varietats estudiades ens vam decantar per la utilització de models discrets o de xarxa. Concretament, vam optar per una representació mitjançant arbres ultramètrics, que permeten una visualització òptima de l'estructura de grups. Aquest tipus de representació té com a finalitat la construcció de grups, basant-se en les relacions de proximitat/llunyania observades entre els diferents elements a partir d'una mesura de similitud adequada. L'esquema d'aquesta mena de classificacions és el següent:

(4)



És a dir, a partir d'una matriu de similitud/dissimilitud[24] i mitjançant l'algorisme de classificació, s'estableix una distància ultramètrica i una jerarquia indexada dels objectes que defineixen la matriu; finalment es representa en forma de gràfic –dendrograma– aquesta distància ultramètrica. L'element principal en aquest procés és l'algorisme de classificació, que permet dur a terme la necessària deformació de la dissimilitud inicial –la de la matriu– en una distància ultramètrica,[25] a partir de la qual obtenim la jerarquia indexada dels diferents elements de la matriu. La jerarquia indexada associada al dendrograma de (4) seria:

(5)

Jerarquia indexada de (4)

$J_1 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$

$J_2 = \{\{1,2\}, \{3\}, \{4\}, \{5\}\}$

(24) En general, la taxonomia numèrica fonamenta les classificacions jeràrquiques partint d'una matriu de similitud, però la descripció teòrica dels algorismes es basa en el concepte de dissimilitud.

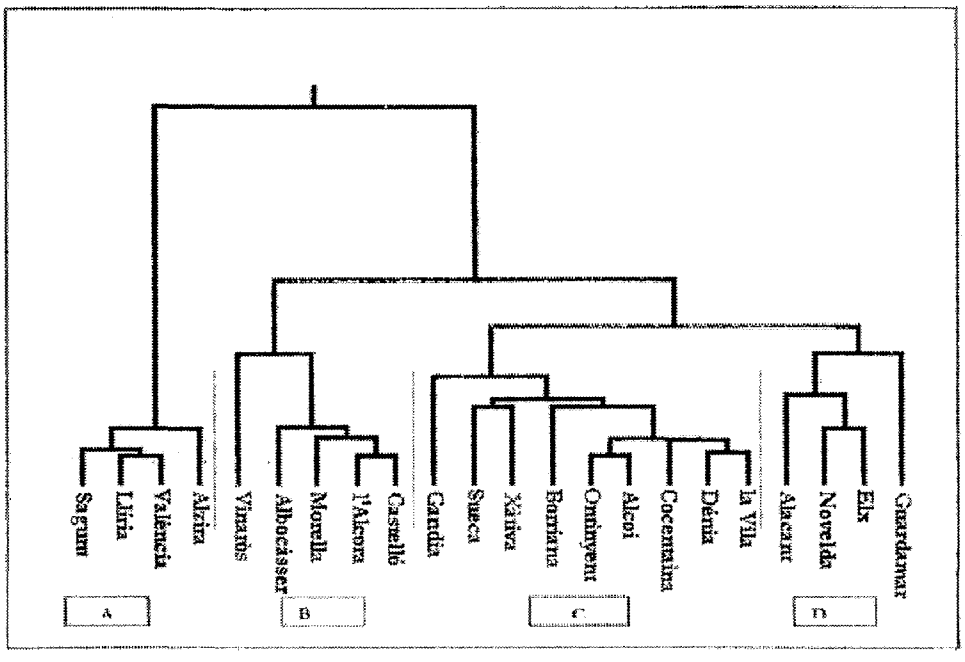
(25) Sobre les característiques de la distància ultramètrica vegeu Arcas & Cuadras (1987).

J_3= {{1,2,3}, {4}, {5}}
J_4= {{1,2,3},{4,5}}
J_5= {{1,2,3,4,5}}

La classificació realitzada es basa en el mètode UPGMA (*Unweighted Pair-Group Method using Arthmetic Averages*), definit per Sokal & Michener (1958). Es tracta d'un mètode utilitzat àmpliament en les aplicacions pràctiques de la taxonomia numèrica en diverses disciplines. Pel que fa a la qualitat de la representació obtinguda en relació amb el grau de fidelitat del model ultramètric a les dades originals, en general es considera el coeficient de correlació cofenètica, mesura d'ajust entre les dissimilituds de partida i les ultramètriques, que es caracteritza per: a) el coeficient de correlació cofenètica pren valors entre -1 i 1; b) el grau de distorsió del model considerat respecte de les dades originals es quantifica per la proximitat a 0 del coeficient de correlació. En aquest sentit, si el coeficient de correlació és molt pròxim a -1 o a 1, el model considerat reflecteix bé l'estructura d'interdistàncies original, i si el coeficient de correlació és pròxim a 0, llavors el model considerat distorsiona molt les dades originals. Els coeficients de correlació cofenètica obtinguts en els diferents gràfics han sortit suficientment pròxims a 1 per assegurar que els models jeràrquics obtinguts reflecteixen amb fidelitat l'estructura d'interdistàncies original.[26]

El resultat final del procés és un dendrograma que ens forneix una classificació de les varietats valencianes analitzades.

(6)



(26) A més, amb la finalitat de contrastar els resultats obtinguts, hem tingut en compte també representacions geomètriques mitjançant arbres additius, i n'hem comparat els resultats. Hem utilitzat, concretament, models additius basats en l'algorisme de Tversky (1977). Les representacions obtingudes mitjançant aquests models corroboren totalment les agrupacions dialectals obtingudes amb els arbres ultramètrics.

Aquesta representació posa de manifest l'existència de quatre grups dialectals –separats per les línies discontínues– amb diversos graus de cohesió interna. En primer lloc, a l'esquerra del dendrograma, s'hi troben agrupades les varietats apitxades –grup A–, amb València, Llíria i Sagunt fortament cohesionades –és a dir, separades per una distància lingüística molt petita– i Alzira, que s'hi aplega a continuació, per tant, a una major distància lingüística.

A continuació, el grup B està constituït per les varietats més septentrionals; es tracta de totes les varietats castellonenques, tret de Borriana. En aquest grup el primer conglomerat està integrat per les varietats de Castelló i l'Alcora, que presenten una distància lingüística mínima, semblant a la que es donava entre València i Llíria; tot seguit s'hi adjunten Morella i Albocàsser, a distàncies relativament petites. I, finalment, acaba incorporant-se en aquest grup la varietat de Vinaròs; en aquest cas, però, a un nivell de llunyania molt superior, la qual cosa trenca una mica la cohesió del grup.

El grup C, que és el més nombrós, està format per les varietats de València –tret de les apitxades–, és a dir: Sueca, Xàtiva, Gandia i Ontinyent; per les varietats del nord d'Alacant: la Vila Joiosa, Dénia, Alcoi i Cocentaina; i per Borriana.

Finalment, el grup D està format per les varietats alacantines situades al sud de la línia Biar-Busot:[27] Alacant, Elx, Novelda i Guardamar. Aquest grup és el que exhibeix una major distància lingüística entre els seus integrants. En aquest sentit destaca la varietat de Guardamar com la més allunyada de totes.

Aquesta agrupació dialectal coincideix amb la delimitació del valencià apitxat de Sanchis Guarner (1936). També està d'acord amb Colomina (1985) en la caracterització com a grup subdialectal de les varietats alacantines del sud de la línia Biar-Busot. En canvi, pel que fa a la resta de varietats, comporta diferències significatives amb les classificacions existents. D'una banda, el que aquí anomenem valencià septentrional no abasta totes les varietats castellonenques, ja que Borriana s'agrupa amb força claredat amb les varietats centrals. De l'altra, la nostra classificació implica definir un subdialecte, que anomenem central, integrat per les varietats compreses entre l'espai geogràfic delimitat per Borriana i la Vila Joiosa –amb l'excepció, de les varietats apitxades. Com que s'ha realitzat a partir de l'anàlisi de les varietats dels caps de comarca del País Valencià, tot i no ser geogràficament exhaustiva, es fonamenta en una perspectiva global de la variació geogràfica dels parlars valencians, la qual cosa li confereix rellevància en contraposició a altres delimitacions de varietats dialectals aïllades, realitzades bàsicament amb una perspectiva més local.

Evidentment, es tracta d'una classificació centrada en les característiques de la morfologia flexiva; cosa que permet introduir consideracions qualitatives en el tractament quantitatiu. Alhora, però, aquest fet comporta que l'assumpció d'aquests resultats en termes de la globalitat dels sistemes lingüístics valencians s'hagi de fer amb certes reserves, ja que la consideració d'altres estrats lingüístics –com el sintàctic, per exemple– podria comportar modificacions en la classificació establerta. En aquest sentit, és remarcable que la consideració de les varietats alacantines del sud de la línia

(27) Vegeu Colomina (1985:60).

Biar-Busot com a grup dialectal diferenciat de les varietats veïnes del nord –el grup C de (6)– només té sentit si tenim en compte el conjunt de la morfologia flexiva, ja que, si considerem aïlladament la flexió verbal, la distància lingüística existent entre ambdós grups no permet establir-hi cap mena de divisió.

5. CONCLUSIONS

Un dels principals reptes que es continua plantejant en l'anàlisi de la variació lingüística geogràfica és la delimitació i classificació de les varietats dialectals. Les classificacions tradicionals, que utilitzen com a eina bàsica les isoglosses i que es basen en criteris estrictament qualitatius, impliquen un grau considerable de subjectivitat, ja que es fonamenten en la selecció d'un petit nombre de trets lingüístics.

Amb l'objectiu de resoldre el problema de l'arbitrarietat que comporta el mètode qualitatiu, durant les darreres dues dècades han proliferat les classificacions basades en criteris quantitatius. La dialectometria, sorgida de la intercomunicació entre la geolingüística i l'anàlisi de dades, és la disciplina que aplega totes aquestes tècniques centrades en el tractament global de les dades d'un conjunt de varietats lingüístiques. Els tractaments dialectomètrics, però, no han estat exempts de mancances, que han hipotecat tant l'adequació dels seus resultats, com la seva pretesa objectivitat.

De la reflexió sobre les principals mancances observades en aquests mètodes, n'hem extret alguns trets que poden contribuir a potenciar-ne l'adequació. Estem convençuts que els tractaments quantitatius basats en el concepte de distància lingüística constitueixen una eina adequada per a la determinació d'agrupacions i classificacions de varietats lingüístiques. Permeten, si més no, que el recorregut entre la descripció de la variació i les conclusions classificatòries no es porti a terme únicament amb el suport del coneixement intuïtiu de l'investigador, sinó que es faci de forma fonamentada, amb el suport de l'anàlisi estadística multivariant, la qual permet un tractament global del conjunt d'elements lingüístics que determinen les semblances o les diferències entre varietats.

Això sí, sempre que aquest tipus de tractaments compleixin els requisits que hem comentat. En principi, cal que parteixin de l'anàlisi d'un conjunt de dades representatives dels sistemes o dels àmbits lingüístics que hom vulgui comparar. També és primordial que es fonamentin en una anàlisi lingüística coherent, que permeti sistematitzar i definir adequadament la variació lingüística observada. I, finalment, cal que es complementin amb mesures de control per tal de contrastar la fiabilitat de les classificacions obtingudes.

ESTEVE CLUA
Universitat Pompeu Fabra

REFERÈNCIES BIBLIOGRÀFIQUES

ARCAS, A. & C. M. CUADRAS (1987) «Métodos geométricos de representación mediante modelos en árbol», *Publicacions de Bioestadística i Biomatemàtica*, 20.

CHAMBERS, J. K. & P. TRUDGILL (1980) *Dialectology*. Cambridge: Cambridge University Press.

COLOMINA, J. (1985) *L'alacantí. Un estudi sobre la variació lingüística*. Alacant: Diputació Provincial d'Alacant.

CUADRAS, C. M. [1981] (1991) *Métodos de análisis multivariante*, Barcelona: PPU.

GOEBL, H. (1981) «Éléments d'analyse dialectométrique (avec application à l'AIS)». *Revue de linguistique romane*, 45, pàgs. 349-420.

– (1984) *Dialektometrishe Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. I, II i III, Tübingen: Niemeyer.

– (1987) «Points chauds de l'analyse dialectometrique: *pondération* et *visualisation*», *Revue de linguistique romane* 51, pàgs. 63-118.

– (1989) «Problèmes et méthodes de la dialectométrie», dins Schouten, M. E. H. & P. Th. van Reenen (eds.) *New Methods in Dialectology. Proceedings of a workshop held at the Free University, Amsterdam, December 7-10, 1987*, Dordrecht: Foris Pubications, pàgs. 165-184.

– (1991) «Una classificazione gerarchica di dati geolinguistici tratti dall'AIS. Sagio di diallettometria dendrografica», *Lingüística, 31*, pàgs. 341-351.

– (1992) «Problèmes et méthodes de la dialectométrie actuelle (avec application à l'AIS)», dins *Nazioarteko Dialektologia Biltzarra Euskaltzaindia. Bilbo 1991, X, 21/25*.

– (1993) «Dialectometry. A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data», dins Köhler, R. & B. B. Rieger (eds.) *Contributions to Quantitative Linguistics*, Dordrecht, Boston, London: Kluwer, pàgs. 277-315.

– (1997) «Some Dendrographic Classifications of the Data of CLAE 1 and CLAE 2», dins *The Computer Developed Linguistic Atlas of England 2* (1997), Tübingen : Max Niemeyer Verlag.

GUITER, H. (1973) *Atlas et frontières linguistiques*, dins Straka, G. & P. Gardette (eds.) *Les dialectes romans de France à la lumière des atlas régionaux. (Colloque de Strasbourg)*, París:Centre Nationale de la Recherche Scientifique, pàgs. 61-109.

HOPPENBROUWERS, C. & G. (1993) «Feature Frequencies and the Classification of Ductch Dialects», dins Viereck, Wolfgang (ed.) *Proceedings of the International Congress of Dialectologist. Bamberg 29.7-4.8.1990*, volume 1, Stuttgart: Steiner (*Zeitschrift für Dialektologie und Linguistik*, Beihefte,77), pàgs. 365-383.

INOUE, F. & FUKUSHIMA, C. (1997) «A Quantitative Approach to English Dialect Distribution: Analyses of CLAE Morphological Data» dins *The Computer Developed Linguistic Atlas of England 2* (1997), Tübingen: Max Niemeyer Verlag.

INOUE, F. & HISAKO K. (1989) «Dialect Classification by Standard Japanese Forms», dins Mizutani, S. (ed.) *Japanese Quantitative Linguistics* (*Quantitative Linguistics*, 39), Bochum, pàgs. 220-235.

MANLY, B. F. J. (1986) *Multivariate Statistical Methods*, London: Chapman & Hall [ed. 1992]

MORGAN, B. J. T. & D. J. SHAW (1982) «Graphical methods for illustrating data in the Survey of English Dialects», *Lore and Language*, 3, pàgs. 14-29.

ORTEGA , J. C. (1998) «Aplicació de tècniques de socioestadística avançada en l'anàlisi de dades dialectals: classificació dels parlars de la comarca de la Marina», *Sarrià*, núm. 0, Callosa d'en Sarrià, La Marina.

POLANCO, Ll. (1992) «Llengua i dialecte: una aplicació dialectomètrica a la llengua catalana», dins *Miscel·lània Sanchis Guarner, III*, Barcelona: Publicacions de l'Abadia de Montserrat, pàgs. 5-28.

SANCHIS GUARNER, M. (1936) *Extensión y vitalidad del dialecto valenciano «apitxat»*, dins *Revista de Filología Española*, 23, pàgs. 45-62.

SAS INSTITUTE, INC. (1990) *SAS User's Guide: Statistics. Edition 6.04*, Cary, North Carolina.

SÉGUY, J. (1971) «La relation entre la distance spatiale et la distance lexical», *Revue de Linguistique Romane*, 35, pàgs. 335-357.

– (1973) «La dialectometrie dans l'*Atlas linguistique de la Gascogne*», *Revue de Linguistique Romane*, 37, pàgs. 1-24.

SHAW, D. J. (1974) «Statistical analysis of dialectal boundaries», *Computer and the Humanities*, 8, pàgs. 173-177.

SIBATA, T. & Y. KUMAGAI (1993) «The S&K Network Method: Processing Procedures for Dividing Dialect Areas», dins Viereck, Wolfgang (ed.) *Proceedings of the International Congress of Dialectologist. Bamberg 29.7-4.8.1990*, volume 1, Stuttgart: Steiner (*Zeitschrift für Dialektologie und Linguistik*, Beihefte,77), pàgs. 459-495.

SNEATH, P. H. A. & R. R. SOKAL (1973) *Numerical Taxonomy. The Principles and Practice of Numerical Classification*, San Francisco: W. H. Freeman and Company.

SOKAL, R. R. i C. D. MICHENER (1958) «A statistical method for evaluating systematic relationships». *University of Kansas Sci. Bull.*, 38, pàgs. 1409-1438.

VENY, J. (1978) *Estudis de geolingüística catalana*, Barcelona: Edicions 62.

– (1982) *Els parlars catalans*, Palma de Mallorca: Moll.

– (1986) *Introducció a la dialectologia catalana*, Barcelona: Enciclopèdia Catalana.

– (1992) «Fronteras y áreas dialectales», dins *Nazioarteko Dialektologia Biltzarra Euskaltzaindia. Bilbo 1991. X. 21/25*, pàgs. 197-245.

VIAPLANA, J. (1997) *Entre la dialectologia i la lingüística: Una anàlisi dialectal aplicada al nord-occidental*, Barcelona. [Publicat a Publicacions de l'Abadia de Montserrat 1999.]

VIERECK, W. (1987) «Lowman's southern English dialectal data and dialectometry». *English World-Wide*, 8, pàgs. 11-23.

– (1988) «The Computerisation and Quantification of Linguistic Data: Dialectometrical Methods», dins Thomas, A. R. (ed.) *Methods in Dialectology. Proceedings on the 6th International Conference held at V. College of North Wales 3-7, 8, 87*, Clevedon Philadelphia: Multilingual Matters Ltd. Cop., pàgs. 524-550.

– (ed.) (1993) *Proceedings of the International Congress of Dialectologist. Bamberg 29.7-4.8.1990,* volume 1, Stuttgart: Steiner (*Zeitschrift für Dialektologie und Linguistik.* Beihefte, 77).

– (ed.) (1995) *Proceedings of the International Congress of Dialectologist. Bamberg 29.7-4.8.1990,* volume 4, Stuttgart: Steiner (*Zeitschrift für Dialektologie und Linguistik*, Beihefte, 77).

WOODS, A., P. FLECHTER & A. HUGES (1991) *Statistics in language studis*, Cambridge: Cambridge University Press.