Esteve Clua
Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra

EL PROJECTE DiLET: UNA NOVA ANÀLISI DE LA DISTÀNCIA LINGÜÍSTICA*

RESUM

En aquest treball presentem els objectius, el marc teòric i la metodologia del Projecte DiLET, que pretén definir la distància lingüística entre les varietats del català, mitjançant una anàlisi dialectomètrica adequada; tant des del punt de vista espacial com des del punt de vista temporal, en aquest cas a partir del contrast amb el Corpus Oral Dialectal del català contemporani.

Paraules clau: dialectologia, dialectometria, distància lingüística, variació fonològica i morfològica, anàlisi de dades.

1. INTRODUCCIÓ

El projecte DiLET (*Distància lingüística entre les varietats catalanes en els eixos espacial i temporal: aspectes fonològics i morfològics*) del grup de dialectometria de l'Institut de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra, forma part dun projecte de recerca més ampli anomenat *Estudi de la fonologia i de la morfologia del català: descripció, teoria i variació*, un projecte compartit amb dos grups de recerca de la Universitat Autònoma de Barcelona (http://filcat.uab.cat/clt/index.html) i de la Universitat de Barcelona (http://www.ub.edu/GEVAD/), i finançat pel *Ministerio de Economía y Competitividad*.

Aquest projecte compartit es proposa, a partir de la coordinació entre els tres grups: a) estudiar un conjunt de fenòmens fònics segmentals, morfofonològics, i de morfologia flexiva centrat en les varietats dialectals del català, amb extensió comparativa a llengües tipològicament afins, amb l'objectiu de millorar el coneixement descriptiu d'aquests fenòmens i completar així punts febles en el nostre coneixement gramatical; b) avançar en l'estudi de la variació lingüística determinant factors d'expansió i anivellació en el canvi lingüístic i establint quantitativament distàncies *dialectométricas; c) establir la contribució dels fenòmens estudiats a la teoria fonològica i morfològica i proposar per tant modificacions de propostes teòriques vigents. En el plànol empíric específic se centrarà fonamentalment en el català, amb especial atenció a la variació dialectal, però recorrent així mateix a l'anàlisi comparativa amb altres llengües tipològicament afins en relació amb alguns dels fenòmens estudiats, especialment les romàniques. Per a això es constituirà un corpus específic de dades orals. Part de la base empíric-descriptiva s'analitzarà en el marc de la teoria de la Optimidad (TO) per determinar quines modificacions són necessàries en les versions clàssica i de serialisme harmònic d'aquest model; una altra part servirà per estudiar l'anivellació dialectal i el contacte lingüístic i per fer una anàlisi dialectomètrica de la distància lingüística interdialectal en el plànol sincrònic i en el diacrònic.

* Aquest treball ha rebut el suport del projecte d'investigació FFI2010-22181-C03-03, finançat pel MINECO.

Revue d'études catalanes, *nº 1* (2014), 135-149.
Date de réception : 15/07/2012 Date d'acceptation : 28/07/2013

2. EL PROJECTE DiLET

L'objectiu del projecte DiLET és contribuir a ampliar el coneixement sobre la variació lingüística en general i sobre la distància entre varietats lingüístiques, en particular, des d'una perspectiva doble: des de l'eix espacial i des del temporal. Tot això a partir de la delimitació de la distància lingüística entre les varietats catalanes. Amb aquesta finalitat ens proposem constituir un corpus oral actualitzat dels dialectes del català. Aquest corpus ens haurà de permetre, en primer lloc, definir la distància lingüística existent en l'actualitat entre aquestes varietats geogràfiques, mitjançant una anàlisi dialectomètrica adequada; i, posteriorment, a partir del contrast amb el Corpus Oral Dialectal de la Universitat de Barcelona (COD), determinar l'evolució temporal d'aquesta distància i, per tant, el canvi lingüístic experimentat.

Encara que només han transcorregut dues dècades des que es va iniciar la constitució del COD, considerem que durant aquest període l'escolarització en català i la generalització de la llengua estàndard en els mitjans de comunicació, si més no en alguns territoris, han pogut propiciar processos d'anivellament i convergència dialectal dignes de ser analitzats i mesurats. En aquest projecte ens proposem, doncs, en primer lloc, l'elaboració i la implementació d'una enquesta dialectal que ens permeti constituir un corpus oral representatiu de les varietats del català, que sigui comparable amb el COD. En segon lloc, ens centrarem en la creació d'una base de dades que faciliti l'accessibilitat i l'explotació del corpus per part de la comunitat científica. A continuació, treballarem en la transcripció i en l'adequació dels materials orals per al corpus. I, finalment, durem a terme la dialectometrització de les dades per determinar la variació i la distància lingüística de les varietats catalanes des de les dues perspectives esmentades.

En l'actualitat hem elaborat el qüestionari i, després de testar-lo en diferents àmbits dialectals, s'han començat a recollir les dades en nuclis de les varietats del català central, del valencià i de les illes. Atès que l'enquesta dialectal ens ha de permetre constituir un corpus oral representatiu de les varietats del català i alhora comparable amb el COD, el qüestionari elaborat, encara que més extens que el del COD, comprèn tots els àmbits lingüístics que apareixien en aquell per així poder comparar les mateixes dades lingüístiques en dos períodes de temps diferents. Posteriorment (§ 3.1) tornarem sobre aquest aspecte.

L'enquesta s'aplicarà a les mateixes poblacions que van fornir les dades del COD, és a dir a les 82 capitals de comarca (o ciutats equivalents) que apareixen a la figura 1. I en cada població s'entrevistarà com a mínim a tres informants de 30 a 45 anys i de nivell sociocultural mitjà.

Figura 1. Caps de comarca i altres poblacions enquestades al projecte DiLET.

3. MARC TEÒRIC

Tradicionalment els estudis dialectals, seguint criteris bàsicament qualitatius, han basat la descripció i 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

[1] Chambers i 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.

dialectals amb què treballem. Posteriorment, es 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 descripció i 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 aquesta noció constitueixi un 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 aquesta noció, 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 descripció, 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.

Els mètodes que segueixen aquestes pautes d'anàlisi de la variació dialectal se solen agrupar sota el rètol de dialectometria. Aquesta disciplina apareix al principi de la dècada dels setanta en l'àmbit de la lingüística romànica, lligada estretament a la geolíngüística. Com ja hem apuntat, la dialectometria es caracteritza per l'abandó de la noció d'isoglossa i l'adopció del concepte de distància lingüística com a eina bàsica de la descripció de la variació lingüística i de les classificacions dialectals. 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. En el nostre camp aquest concepte s'associa a la quantificació de les similituds o diferències que existeixen entre varietats lingüístiques en relació a un conjunt de dades.

---

[2] Chambers i 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).

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.

Però, com determinem les diferències o similituds entre varietats lingüístiques a partir de les quals establir el tractament quantitatiu i determinar la distància lingüistica? Realment, les representacions fonètiques dels ítems d'un determinat atles o corpus reflecteixen adequadament les diferències existents entre les varietats lingüístiques? Intentarem respondre aquestes preguntes a partir dels exemples següents en els quals contrastem les formes del numeral *dos* en dues varietats lingüístiques, la de Barcelona i la de València.

(1)	Varietat 1 (Barceloma)		Varietat 2 (València)	
	a. dos	[ˈdos]	dos	[ˈdos]
	b. dos pares	['dos'parəs]	dos pares	['dos'paɾes]
	c. dos llibres	['doz'ʎiβrəs]	dos llibres	['doz'ʎiβɾes]
	d. dos ous	['doz'ɔws]	dos ous	['dos'ɔws]

Si contrastem la realitzacions fonètiques del numeral aïllat (1a) en les dues varietats, veiem que no presenten cap diferència. Tampoc hi ha diferències quan apareix davant d'una altra paraula començada per consonant sorda (1b). Quan precedeix una paraula començada per consonant sonora (1c) tampoc trobem diferències entre ambdues varietats, però en aquest cas podem observar que el so sibilant del final de la paraula presenta una realització sonora [z] , que contrasta amb les realitzacions sordes d'aquest segment en els contextos anteriors. Finalment a (1d) sí que podem observar una realització diferent del numeral en les dues varietats contrastades: mentre que la varietat de Barcelona presenta una realització sonora de la sibilant final, en la varietat de València hi trobem una realització sorda.

Tenint en compte això podríem concloure, en principi, que aquestes dues varietats no presenten cap diferència en la representació fonètica del numeral dos ['dos], perquè els tres segments que integren aquesta paraula coincideixen totalment; però la coincidència no pot ser total si tenim en compte el comportament de l'últim segment, el sibilant, que apareix realitzat com a sonor davant d'un segment sonor [± vocàlic] en la varietat 1, mentre que solament ho fa davant d'un segment no vocàlic en la varietat 2. És a dir, les dues varietats coincideixen totalment en els segments fonètics que constitueixen el numeral dos, però els processos fonològics que afecten aquests segments són diferents. Si entenem que aquests processos fonològics són bàsics per comprendre l'estructura sonora de les varietats lingüístiques, haurem de deduir, en conseqüència, que no podem ignorar-los en intentar captar les diferències existents entre elles.

Per poder tenir en compte totes aquestes diferències, en les anàlisis quantitatives de les dades del COD, a diferència del que ocorre a altres centres on es treballa en dialectometria com Salzburg o Groningen, apliquem a les dades fonètiques del corpus una anàlisi lingüística basada en el model de la fonologia generativa clàssica, perquè

permet discriminar les diferències superficials o predictibles, que expressen les regularitats de les llengües (i de les varietats que les componen), de les diferències subjacents o impredictibles, que afecten l'estructura lèxica o gramatical de les paraules[4]. Des del nostre punt de vista, la distinció entre aquests dos nivells d'anàlisi és fonamental per determinar la distància lingüística entre varietats[5].

Per a l'anàlisi de fenòmens fonològics específics, hem seguit majoritàriament l'orientació pròpia de la fonologia generativa derivacional. Quant als aspectes morfològics, ens basem en l'enfocament morfèmic clàssic (Item and Arrangement).

Presentem a continuació un cas de diferències fonològiques (subjacents) i un altre de diferències fonètiques (superficials) per exemplificar la nostra anàlisi. El català presenta diferents terminacions per a la 1a persona del singular del present d'indicatiu: [ø], [u], [o], [e], [i] i [a]. Ho podem veure en els exemples de (2), corresponents al verb cantar.

(2) Diferències Fonològiques

1a persona del singular del present d'indicatiu del verb *cantar*

a. Varietat 1 (Palma)	['kant]	/ø/
b. Varietat 2 (Barcelona)	['kantu]	/u/
c. Varietat 3 (Lleida)	['kanto]	/o/
d. Varietat 4 (València)	['kante]	/e/
e. Varietat 5 (Ceret)	['kanti]	/i/
f. Varietat 6 (l'Alcora)	['kanta]	/a/

Aquestes diferències, però, no poden ser atribuïdes a cap alternança fònica sistemàtica de la fonologia del català; és a dir, en aquestes varietats no existeix cap procés regular pel qual –[u] es converteixi en –[i] o en –[a], o desaparegui en situació final de paraula. Per tant, han de ser atribuïdes directament a l'estructura morfològica de les paraules. Són, doncs, diferències subjacents, impredictibles.

Un altre exemple de variació que afecta les terminacions verbals, el trobem en les formes de gerundi. Les varietats del català central, per exemple, solen presentar una realització fonètica acabada en nasal: *emprant* [əm'pɾan]; mentre que la majoria de les varietats valencianes, balears i la varietat de l'Alguer marquen el gerundi amb l'aplec consonàntic [nt]. Si quantifiquem les diferències entre varietats a partir de les dades fonètiques, aquestes diferències són iguals que les observades a (2). Des de l'òptica generativa, però, l'alternança [n] / [nt] que s'observa en la varietat 2 (3b) pot ser atribuïda a una única forma fonològica /nt/, que se simplifica en posició final de paraula, però es manté quan la forma de gerundi va seguida d'un clític, ja que en aquestes varietats opera un procés fonològic de simplificació d'aplecs consonàntics [nt] en final de paraula, que no es produeix quan la forma verbal va seguida d'un clític pronominal començat per vocal; en canvi a la varietat de València (3a) no es produeix aquest procés de simplificació en posició final de paraula.

[4] Vegeu Lloret i Viaplana (1998).
[5] Vegeu Clua (1999a, b) i Viaplana (1999).

(3) Diferències Fonètiques

Gerundi del verb *emprar*

a. Varietat 1 (València):
   emprant [emˈɒrant] /nt/

b. Varietat 2 (Barcelona):
   emprant [əmˈɒran] /nt/
   emprant-ho [əmˈɒrantu]

Des de la nostra perspectiva, les diferències observades a (2) són fonològiques, solament predictibles a partir de l'estructura morfològica; mentre que a (3) les diferències són merament fonètiques i es poden predir per un procés fonològic. Creiem que es tracta d'una distinció pertinent que ha de tenir-se en compte en la quantificació de la distància lingüística entre varietats, ja que en cas contrari el resultat de l'anàlisi quantitativa pot apartar-se considerablement de la realitat. Així doncs, l'anàlisi lingüística ens permet, d'una banda, captar similituds o diferències entre varietats que a simple vista fonètica serien imperceptibles, i per l'altra, ens permet distingir entre diferències estructurals (les que aquí denominem fonològiques) i diferències predictibles (fonètiques) a partir dels processos fonològics sistemàtics que caracteritzen les varietats lingüístiques.

En treballs anteriors hem argumentat i justificat la pertinència de l'anàlisi lingüística prèvia per realitzar un tractament quantitatiu adequat de la distància lingüística. A Clua i Lloret (2006) ho vam fer a partir de l'anàlisi de la distància lingüística en els clítics pronominals, una de les àrees del català que presenta més variació fonètica. En aquest cas, es tractava del que podríem catalogar de petit assaig de laboratori a partir de la variació observada en aquests clítics en tres varietats lingüístiques del català, entre les quals mesuràvem i representàvem la distància lingüística, primer, a partir de les dades fonètiques i, a continuació, a partir de les dades analitzades fonològicament. Les diferències obtingudes en ambdós tractaments feien aconsellable partir d'una anàlisi lingüística de les dades fonètiques del corpus.

Atès que algunes vegades s'havia esgrimit que en un tractament quantitatiu global (amb grans quantitats de dades lingüístiques com les que poden oferir un atles o un corpus com el COD) aquest tipus de diferències (entre els resultats obtinguts a partir de les dades fonètiques i fonològiques) podia ser totalment inapreciable i intranscendent, a Clua (2010) i a Clua *et al.* (2010) vam realitzar un tractament quantitatiu contrastat de les dades del COD. En aquests treballs, vam contraposar els resultats de la representació de la distància lingüística de les dades d'aquest corpus prèviament analitzades fonològicament (vegeu Clua *et al.*, 2009) amb els obtinguts a partir de les mateixes dades, però sense l'anàlisi fonològica prèvia. El resultat de la comparació corrobora les hipòtesis a les quals havíem arribat amb els assajos de laboratori anteriors; en el sentit que els resultats de l'anàlisi dialectomètrica són considerablement diferents si partim de dades fonètiques o si ho fem després d'analitzar fonològicament aquestes mateixes dades.

3. ASPECTES METODOLÒGICS

Els antecedents metodològics del nostre projecte es troben en el marc dels estudis realitzats al voltant del COD. A continuació descriurem la metodologia emprada en l'anàlisi d'aquest corpus (l'anomenada MCOD) que és la mateixa que emprarem també en el projecte DiLET.

3.1. El Corpus Oral Dialectal de la Universitat de Barcelona

Des del 1991, el Departament de Filologia Catalana de la Universitat de Barcelona ha compilat i sistematitzat en bases de dades un Corpus Oral Dialectal (COD) del català contemporani. Aquest corpus, que conté informació de les sis principals varietats dialectals del català: alguerès, balear, central, nord-occidental, rossellonès i valencià, s'ha constituït a partir d'una sèrie d'entrevistes que es van dur a terme a cada cap de comarca (o ciutat equivalent) de l'àmbit lingüístic català.

El corpus consta de materials fonètics i morfològics obtinguts a través d'un qüestionari de 600 ítems i d'un conjunt de textos espontanis[6]. Les dades estan recollides en vuit bases de dades diferents: morfologia verbal regular, aspectes fonètics, clítics pronominals, articles, possessius, pronoms personals forts, demostratius i locatius[7]. I el nombre de registres que conté cada una d'aquestes bases de dades és el de (4).

(4)

Base de dades	Nombre de registres
Verbs	68262
Fonètica	55270
Clítics pronominals	24766
Articles	5049
Possessius	5223
Pronoms personals	1608
Demostratius	1512
Locatius	803

3.2. El procés d'anàlisi dialectomètrica de les dades del COD

En el disseny dels diferents estadis del procés de dialectometrització de les dades del COD s'han tingut en compte el posterior tractament quantitatiu de les dades. Així, d'una banda, el punt de partida del procés ha estat l'anàlisi de la variació d'un conjunt de dades representatives de les varietats lingüístiques catalanes, la qual cosa permet introduir consideracions qualitatives en el tractament quantitatiu. I de l'altra, s'ha tractat de 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.

Com hem justificat en l'apartat anterior, les variables de comparació emprades per a la classificació de les varietats lingüístiques s'han obtingut a partir d'una anàlisi

[6] Pel que fa a aquests textos, vegeu Viaplana i Perea (2003) i també l'apartat DADES de <http://www.ub.edu/lincat/>.

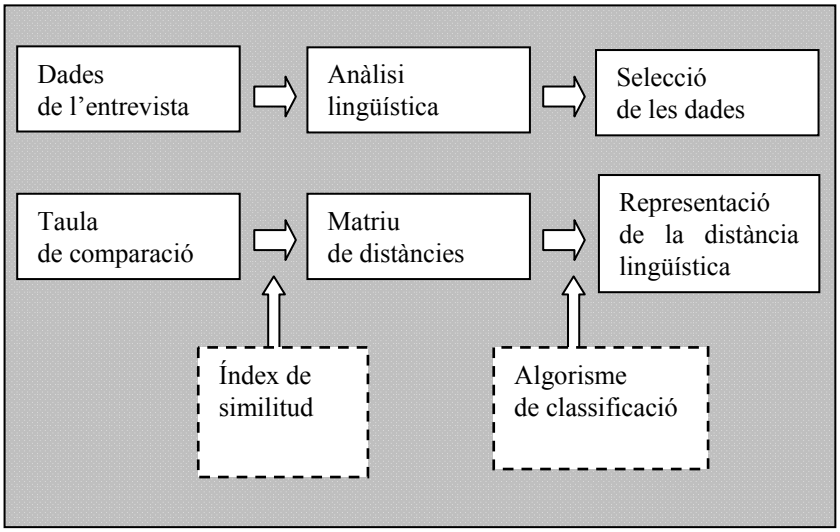
[7] Hi ha una edició en CD-Rom de les dades del COD a Viaplana *et al.* (2007); vegeu també <http://www.ub.edu/lincat/>.

lingüística —basada en els principis de la fonologia generativa— de les dades fonètiques obtingudes a les entrevistes.

Per últim, atès a la gran variabilitat que es pot produir en els resultats del tractament estadístic segons el tipus d'índex de similitud emprat o l'algorisme utilitzat per a la representació gràfica de la distància observada, s'han utilitzat mesures que ens han permès contrastar la fiabilitat dels resultats obtinguts.

Tot seguit presentem un esquema del procés seguit en l'anàlisi dialectomètrica de les dades del COD, i a continuació n'expliquem els aspectes més rellevants des de l'òptica de l'anàlisi quantitativa.

(5) Procés d'anàlisi dialectomètrica



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. Atès que ja n'hem comentat les característiques a l'apartat anterior, només esmentarem que l'aplicació d'aquesta anàlisi ens forneix les formes subjacents dels diferents morfemes i els processos fonològics a partir dels quals s'estableixen les taules de comparació, que són la base per poder definir posteriorment la similitud i la distància lingüístiques entre varietats.

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:

(6) Taula de comparació

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 les taules de comparació, es van elaborar les matrius de similituds, que constitueixen el següent pas del procés quantitatiu. Abans, però, com es pot veure a l'esquema (5), cal establir l'índex de similitud a partir del qual es comptabilitzaran les coincidències i s'establiran les similituds entre varietats[8]. En aquest cas l'índex de similitud utilitzat ha estat el següent:

(7) Index de similitud de l'MCOD

$$dist\ (i,j) = \frac{\sum_{k=1}^{long} dif_k(i,j)}{long} \times 100$$

És a dir, la distància lingüística entre dues varietats (i, j) és igual al sumatori (Σ) de les diferències quant a una variable k entre les varietats (i, j), dividit per long. que és la longitud (és a dir, el nombre de sons) de cada segment morfològic comparat.

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 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.

3.3. La representació de la distància lingüística (DL)

Un cop definida la DL, és necessari obtenir-ne representació gràfica. Les matrius de similituds / diferències, resultants de la comparació de les característiques lingüístiques de les diferents varietats a classificar, determinen les interdistàncies lingüístiques entre aquestes varietats, però, quan es treballa amb quantitats importants de dades, no permeten una bona interpretació ni una bona visualització dels resultats. Com es pot comprovar a (8), que és un petit fragment d'una de les matrius de similituds amb què hem treballat, a partir d'un conjunt de dades d'aquest tipus podem discernir quines són les varietats que presenten més divergències (en aquest cas serien les varietats de l'Alguer i Felanitx, que presenten un 33,67% de divergències en el conjunt de les dades

[8] La definició de la distància lingüística s'ha realitzat amb el programa Microsoft ® Excel. Concretament, hem utilitzat unes macros d'Excel dissenyades pel professor Sergi Civit del Departament d'Estadística de la Universitat de Barcelona.

del COD analitzades), però resulta impossible establir una classificació global i captar adequadament les relacions estructurals entre varietats. Per això és necessari buscar una representació de l'estructura d'interdistàncies de la matriu que permeti visualitzar amb claredat les relacions de proximitat o llunyania entre les varietats lingüístiques analitzades. Es tracta, doncs, d'aconseguir una representació en un espai de dimensió reduïda (com per exemple, el pla) que, amb un mínim de distorsió de l'estructura de interdistàncies original, permeti una interpretació adequada de les dades.

(8) Fragment de matriu de distàncies

	L'Alguer	Ciutadella	Eivissa	Felanitx	Formentera	Manacor	Maó	Palma
L'Alguer	0,00							
Ciutadella	33,01	0,00						
Eivissa	30,53	6,43	0,00					
Felanitx	33,67	9,62	7,96	0,00				
Formentera	29,59	8,00	3,43	8,75	0,00			
Manacor	32,91	8,29	6,20	4,88	7,11	0,00		
Maó	33,27	4,93	9,90	12,59	11,15	10,43	0,00	
Palma	32,21	6,44	4,77	4,38	5,84	3,17	10,03	0,00

Amb aquesta finalitat, l'àmbit de l'anàlisi de dades ens ofereix diferents models representacionals, segons l'espai geomètric preferit. També existeixen diferents algoritmes de classificació per traslladar la distància original a una representació visualment òptima. El problema és que la utilització de models o d'algoritmes diferents pot donar com a resultat classificacions sensiblement diferents. Per això, l'investigador es veu obligat a escollir el model més adequat a les seves finalitats i a avaluar amb objectivitat els resultats de les representacions obtingudes en relació amb les dades originals. En les nostres anàlisis, hem optat per recórrer a models de representació en arbres ultramètrics (o dendrogrames), ja que generalment permeten una bona visualització de les classificacions i 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 / dissimilitud adequada.

Concretament, el mètode de representació gràfica de què ens hem servit majoritàriament en les nostres anàlisis és el Clúster Anàlisi i hem usat un algorisme de classificació basat en el mètode UPGMA (Unweighted Pair-Group Method Using Arithmetic Average)[9], que ha estat contrastat àmpliament en aplicacions de la taxonomia numèrica en múltiples disciplines. Per a l'avaluació de la distorsió entre les representacions i la distància original, apliquem el coeficient de correlació cofonética, amb uns resultats que corroboren la fidelitat de les representacions jeràrquiques en relació amb la DL de partida. Vegeu-ne un exemple en la figura 2 que representa la DL de les varietats del català a partir de la metodologia MCOD, amb les dades de COD analitzades fonològicament.

[9] Vegeu Sneath i Sokal (1973).

D'altra banda, a partir de la col·laboració amb un dels dos centres de recerca europeus capdavanters en Anàlisi Estadística de la Variació Lingüística, l'Escola Dialectomètrica de la Universitat de Salzburg, dirigida pel professor Hans Goebl, hem pogut realitzar una anàlisi dialectomètrica completa de les varietats catalanes mitjançant el paquet VDM (Visual Dialectometry), desenvolupat en aquest centre. Es tractava de definir i representar cartogràficament i dendrogràficament la DL entre les varietats del català[10]. Vegeu-ne un exemple a la figura 3, on es mostra la representació cartogràfica de la DL de les varietats catalanes definida a partir del sistema VDM.

[10] Vegeu Goebl (1992 i 2010).

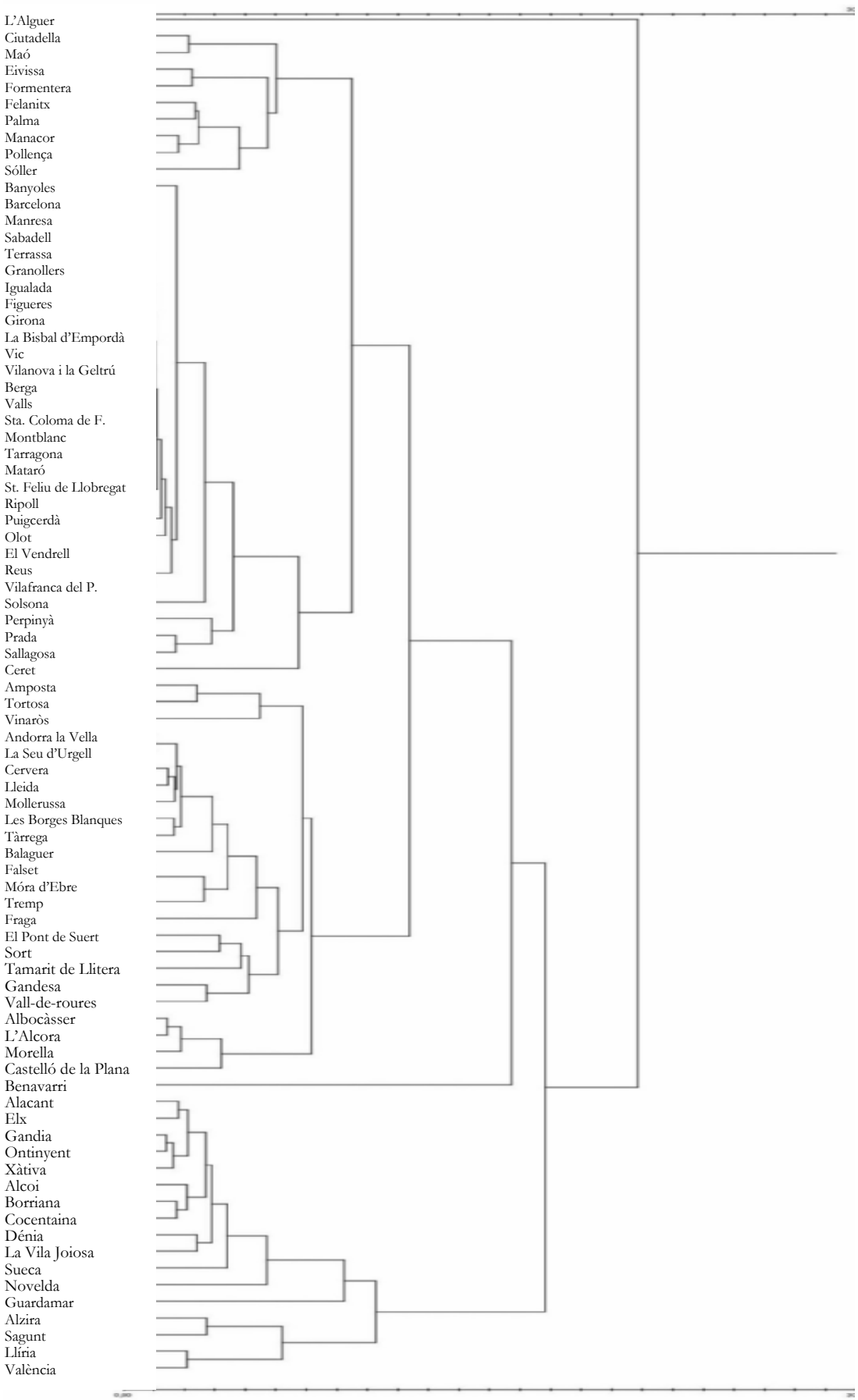


Figura 2. Representació de la DL entre les varietats catalanes amb l'MCOD.

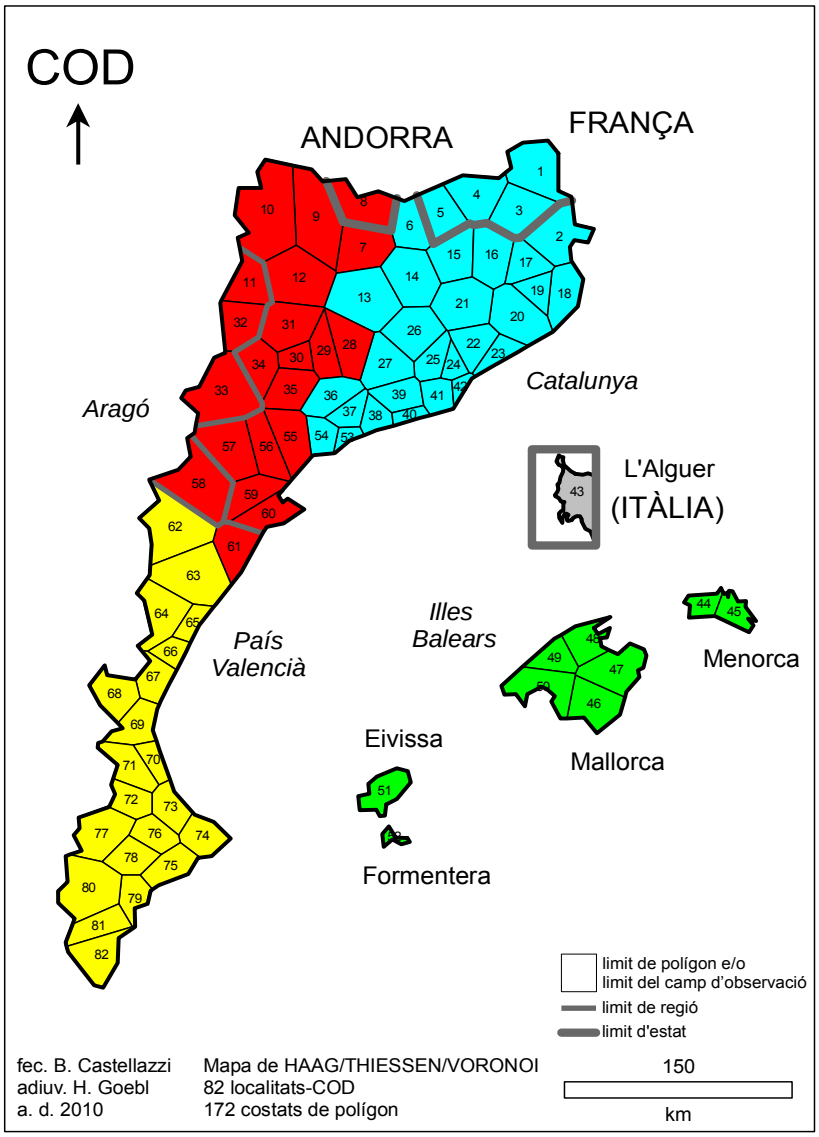


Figura 3. Representació de la DL entre les varietats catalanes amb el VDM.

BIBLIOGRAFIA

Chambers, Jack K. i Trudgill, Peter (1980). *Dialectology*. Cambridge: Cambridge University Press.

Clua, Esteve (1999a). *Variació i distància lingüística. Classificació dialectal del valencià a partir de la morfologia flexiva*. Tesi doctoral. Universitat de Barcelona.

Clua, Esteve (1999b). «Distància lingüística i classificació de varietats dialectals». *Caplletra, 26*, 11-26.

Clua, Esteve (2010). «Relevancia del análisis lingüístico en el tratamiento cuantitativo de la variación dialectal». Dins Gotzon Aurrekoetxea i José Luis Ormaetxea (eds.), *Tools for Linguistic Variation.* Bilbao: Universidad del País Vasco, pp.151-166.

Clua, Esteve i Lloret, Maria-Rosa (2006). «New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)». Dins Jean-Pierre Montreuil (ed.), *New Perspectives on Romance Linguistics. Vol. 2: Phonetics, phonology and dialectology*. Amsterdam/Philadelphia: John Benjamins, pp. 31-47.

Clua, Esteve; Lloret, Maria-Rosa i Valls, Esteve (2009). «Análisis lingüístico y dialectométrico del corpus oral dialectal (COD)». Dins Pascual Cantos Gómez i Aquilino Sánchez Pérez (eds.), *A survey on corpus-based research [Actas del I Congreso Internacional de Lingüística de Corpus (CICL-09), 7-9 Mayo 2009, Universidad de Murcia]*. Murcia: Asociación Española de Lingüística del Corpus, pp. 1033-1045.

Clua, Esteve; Valls, Esteve i Adrover, Margalida (2010). «Tractament quantitatiu de la variació dialectal i anàlisi lingüística: noves perspectives a partir de les dades del COD». Comunicació, *XXVI^e^ Congrès International de Linguistique et de Philologie Romanes* (València, 6-11 septembre 2010).

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

Goebl, Hans (2010). «Introducción a los problemas y métodos según los principios de la escuela dialectométrica de Salzburgo (con ejemplos sacados del "Atlante italo-svizzero")». Dins Gotzon Aurrekoetxea i José Luis Ormaetxea. (eds.), *Tools for Linguistic Variation*. Bilbao: Universidad del País Vasco, pp. 3-39.

Lloret, Maria-Rosa i Viaplana, Joaquim (1998). «Variació morfofonològica. Variants morfològiques». *Caplletra* 25, 43-62.

Sneath, Peter H. A. i Sokal, Robert R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. San Francisco: W. H. Freeman and Company.

Veny, Joan (1992). «Fronteras y areas dialectales». *Nazioarteko Dialektologia Biltzarra Euskaltzaindia. Bilbo 1991. X. 21/25*, pp. 197-245.

Viaplana, Joaquim (1984). «La flexió verbal regular del valencià». Dins Emili Casanova (ed.), *Estudis en memòria del professor Manuel Sanchis Guarner: Estudis de Llengua i Literatura Catalanes*, I. València: Universitat de València, pp. 391-407.

Viaplana, Joaquim (1986). «Morfologia flexiva i flexió verbal catalana». *Llengua i Literatura*, Barcelona 1, 385-403.

Viaplana, Joaquim (1999). *Entre la dialectologia i la lingüística*. Barcelona: Publicacions de l'Abadia de Montserrat.

Viaplana, Joaquim i Perea, Maria-Pilar (eds.) (2003). *Textos orals dialectals del català sincronitzats. Una selecció*. Barcelona: PPU. [Inclou CD-Rom].

Viaplana, Joaquim; Lloret, Maria-Rosa; Perea, Maria-Pilar i Clua, Esteve (2007). *COD. Corpus Oral Dialectal*. Barcelona: PPU. [Publicació en CD-Rom].