CAP A UNA DIVERSIFICACIÓ METODOLÒGICA DE LA DIALECTOMETRIA CATALANA: PRIMERS RESULTATS D'APLICAR LA DISTÀNCIA DE LEVENSHTEIN AL *CORPUS ORAL DIALECTAL*[1]

ESTEVE VALLS
UNIVERSITAT DE BARCELONA
e.valls@ub.edu

1. Introducció

En paraules de Hans Goebl, membre fundacional de l'Escola Dialectomètrica de Salzburg, la dialectometria constitueix una simbiosi entre la geolingüística tradicional i la taxonomia numèrica (Goebl, 2003: 60). Pretén, d'una banda, analitzar la variació diatòpica d'una llengua des d'una perspectiva quantitativa –evitant, doncs, la càrrega subjectiva inherent en la noció tradicional d'isoglossa. A la base d'aquest plantejament hi ha el concepte de *distància lingüística* entre varietats. D'altra banda, la dialectometria posa també l'accent en l'optimització dels sistemes representacionals dels resultats, amb la intenció de proporcionar al geolingüista múltiples eines d'anàlisi i interpretació de la variació dialectal d'una determinada àrea lingüística.

Les aplicacions del mètode dialectomètric a corpus de la llengua catalana es remunten a l'article inaugural de la disciplina (Séguy, 1971). Tanmateix, no ha estat fins a finals de la dècada dels 90 que aquesta metodologia s'ha consolidat entre els lingüistes catalans, especialment arran de l'explotació del *Corpus Oral Dialectal* (COD) de la Universitat de Barcelona. La dialectometria desenvolupada a l'entorn del COD manté certs paral·lelismes amb el mètode de Salzburg –especialment pel que fa al còmput d'interdistàncies–, però se'n diferencia en considerar fonamental una anàlisi lingüística de les dades que permeti discriminar aquells elements impredictibles de la llengua d'aquells processos sistemàtics i, doncs, predictibles. Gràcies a aquest procés, previ al càlcul de distàncies, és possible de capturar diferències interdialectals que en les formes fonètiques restarien amagades (Clua & Lloret, 2006). Recentment (Clua et al., 2008) s'han publicat els primers resultats de dialectometritzar el COD des d'aquest enfoc teòric.

Ara bé, tot i que en els darrers anys la dialectometria ha anat guanyant adeptes entre els lingüistes catalans, hi ha un tercer mètode dialectomètric de validesa contrastada que a dia d'avui encara hi és pràcticament desconegut: el desenvolupat al *Center for Language and Cognition* de la Universitat de Groningen (Heeringa, 2004). Aquest sistema es basa en l'aplicació de l'anomenada *distància de Levenshtein* (LD), una mesura de la distància fonètica introduïda en dialectologia per Kessler (1995) en un estudi sobre les varietats del gaèlic irlandès.

L'objectiu d'aquest article és, doncs, doble: d'una banda, vol contribuir a difondre el mètode dialectomètric de Groningen en l'àmbit local; de l'altra, vol presentar els resultats d'una primera aplicació d'aquest mètode a un corpus en llengua catalana, el COD. A continuació (§2), es descriuen les principals característiques del *Corpus Oral Dialectal*; tot seguit (§3), es proporciona una explicació bàsica del funcionament de la distància de Levenshtein; a §4 s'introdueixen els sistemes taxonomètrics i de visualització escollits per a l'anàlisi dels resultats, que s'analitzen a §5; i, finalment, §6 recull alguns comentaris succints que pretenen estimular el debat sobre els avantatges i els inconvenients dels diferents mètodes dialectomètrics disponibles en l'actualitat.

[1]Aquest treball s'inscriu en el marc del projecte HUM2007-65531/FILO (*Explotación de un corpus oral dialectal: análisis de la variación lingüística y desarrollo de aplicaciones informáticas para la transcripción automatizada (2ª fase)*), finançat pel MICINN i el FEDER.

2. El *Corpus Oral Dialectal* (COD) de la Universitat de Barcelona

El *Corpus Oral Dialectal* és un corpus del català contemporani creat i sistematitzat des de l'any 1991 al Departament de Filologia Catalana de la Universitat de Barcelona. A partir d'un qüestionari de vora 600 ítems, es van recollir un total de 135.480 registres fonètics i 532.508 de morfològics. També es van compilar mostres d'uns deu minuts de conversa semidirigida amb cadascun dels informants. El treball de camp es va dur a terme als 86 caps de comarca –o localitats equivalents– del domini lingüístic, on es van entrevistar entre dos i tres informants, que havien de complir els següents criteris: ser oriünds de la localitat i descendents de pare i mare del mateix lloc, tenir una edat d'entre 30 i 45 anys, haver cursat només educació primària, ser de classe mitjana i haver viscut tota la vida en la localitat d'origen. L'objectiu era, doncs, confegir un corpus del català actual que reflectís la llengua de la majoria dels seus parlants, avui concentrats en zones urbanes. Es defugia, per tant, la tendència tradicionalment imperant en dialectologia, segons la qual l'informant ideal era aquell que reflectia la parla més conservadora d'un determinat indret. D'altra banda, la posterior dialectometrització de les dades va ser un criteri que es va tenir en compte durant el disseny del qüestionari.

Aquest treball presenta els resultats d'aplicar la distància de Levenshtein a un subcorpus del COD integrat per 29.192 ítems. Concretament, s'han seleccionat 356 registres per localitat, corresponents als següents àmbits gramaticals de la llengua: verbs (20.500 ítems), articles (1.312 ítems), possessius (1.968 ítems), clítics pronominals (4.264 ítems), pronoms personals (648 ítems), demostratius neutres (143 ítems) i adverbis locatius (143 ítems). En el cas de poblacions amb més d'un output per concepte, la forma escollida ha estat sempre la majoritària entre els informants.

3. La distància de Levenshtein (LD): una mesura de la distància fonètica

La distància de Levenshtein –en anglès, *Levenshtein Distance* o *edit distance*– és una mesura de càlcul de la distància fonètica entre dues línies de dades. Per determinar aquesta distància, l'algorisme de Levenshtein cerca quin és el menor conjunt d'operacions bàsiques necessari per transformar una línia en una altra. Aquestes operacions poden ser insercions, supressions o substitucions, i en la versió més simple de la LD tenen totes tres un cost d'1. La distància final entre dos dialectes és, doncs, la distància mitjana obtinguda de transformar d'una varietat a l'altra les realitzacions fonètiques d'un nombre determinat de conceptes. A (1) s'exemplifica el funcionament bàsic de la LD a partir de les realitzacions fonètiques de la primera persona del singular de l'imperfet de subjuntiu del verb *servir* a dues poblacions del català occidental. En aquest cas, el cost final de transformar la primera línia en la segona és 3:

(1)

Varietat 1	s e ɾ β i s ˈk e s	esborra s	1
	s e ɾ β i ˈk e s	substitueix k/ɣ	1
	s e ɾ β i ˈɣ e s	insereix a	1
Varietat 2	s e ɾ β i ˈɣ e s a		
Total			3

Des d'un altre punt de vista, aquest procediment també es pot entendre com el resultat d'alinear dos conjunts de segments fonètics. En aquests alineaments, la superposició fonètica és binària, de manera que només contribueixen a augmentar la distància fonètica entre línies els segments no coincidents (2):

(2) Varietat 1 s e ɾ β i s ˈk e s
Varietat 2 s e ɾ β i ˈɣ e s a

1 1 1 = 3

Amb l'objectiu d'augmentar la precisió dels resultats, en aquest treball s'ha adaptat l'algorisme de Levenshtein de tal manera que només permeti alineaments de vocals amb vocals i de consonants amb consonants. Tampoc no s'ha aplicat cap normalització dels resultats en funció de la longitud mitjana de les realitzacions fonètiques, tal com recomanen els principals estudis de validació de l'algorisme de Levenshtein (Heeringa et al., 2006).

D'ençà del treball inaugural de Kessler (1995) per al gaèlic irlandès, la distància de Levenshtein s'ha aplicat a corpus provinents de més d'una desena de llengües, com ara l'alemany (Nerbonne, 2008), el neerlandès (Heeringa, 2004), el frisó (Heeringa, 2005), el búlgar (Osenova et al., 2008) i les llengües de Gabon (Alewijnse et al., 2007). A més, s'ha consolidat també com una mesura versàtil, que s'ha emprat satisfactòriament en altres tipus de recerca: per exemple, Heeringa & Gooskens (2004) la incorporen en la seva anàlisi perceptual i acústica de 15 varietats del noruec; també Gooskens & Bezooijen (2006) en fan ús per comparar el grau de comprensibilitat mútua que comparteixen actualment els parlants de neerlandès i d'afrikaans.

Tant per a l'obtenció de la matriu de distàncies com dels agrupaments i representacions cartogràfiques posteriors, hem fet ús del paquet de programari integrat L04, dissenyat a Groningen per Peter Kleiweg[2]. Prèviament, va caldre convertir el corpus de dades a X-SAMPA, un sistema de codificació que permet que el programa informàtic processi les transcripcions fonètiques. Tot seguit presentem els cinc sistemes representacionals que hem emprat per visualitzar les relacions interdialectals recollides a la matriu de distàncies, obtinguda mitjançant l'aplicació de l'algorisme de Levenshtein.

4. Sistemes de visualització

En presentar públicament el seu nou mètode d'anàlisi de la variació diatòpica, Séguy (1971) va justificar aquesta innovació metodològica a partir d'un doble argument: calia, d'una banda, superar la subjectivitat del mètode tradicional, basat en els feixos d'isoglosses; i calia, en segon lloc, acabar amb la infraexplotació de les dades contingudes en els atles lingüístics publicats fins al moment. Séguy tenia *in mente*, tanmateix, un tercer objectiu crucial que va desenvolupar dos anys més tard: el d'optimitzar els sistemes de visualització dels resultats des d'una perspectiva *de conjunt*. Amb aquest objectiu va dissenyar els primers *network maps* utilitzats en dialectologia, que va incloure en el sisè i darrer volum de l'*Atlas Linguistique de la Gascogne* i que recollien, damunt del mapa, les distàncies de Hamming entre tots els punts d'enquesta de l'àrea lingüística estudiada (Séguy, 1973).

D'aleshores ençà, el desenvolupament d'eines que permetin una projecció cartogràfica global de la variació diatòpica ha estat un element central en la recerca dialectomètrica. A Salzburg, el programa VDM[3] ofereix múltiples sistemes de visualització dels resultats, d'entre els quals Goebl sol destacar els *mapes de similituds*. A Groningen, la diversificació ha estat encara major: Nerbonne (2008) introdueix detalladament les motivacions teòriques i les principals tècniques de mapatge de la variació diatòpica contingudes al paquet L04.

Tot seguit descrivim el funcionament bàsic dels cinc sistemes de visualització que hem escollit per analitzar la variació diatòpica del català. Es tracta, en tots els casos, de tècniques que s'apliquen per primera vegada a un corpus d'aquesta llengua; constitueixen, doncs, un pas més en la diversificació metodològica de la dialectometria catalana.

[2] Aquest paquet és d'accés lliure i es pot descarregar a l'adreça http://www.let.rug.nl/kleiweg/L04/ .
[3] Per a una introducció al programa VDM, dissenyat per Edgar Haimerl, vg. www.dialectometry.com .

4.1. Els *network maps*

Els *network maps* –també coneguts com a *link maps*– ofereixen una primera aproximació a la realitat dialectal d'una llengua traçant línies entre els diferents punts d'enquesta. La foscor d'aquestes línies és inversament proporcional a la distància lingüística entre poblacions: així, com més clares són, més distants lingüísticament són els punts que uneixen; i a la inversa, com més fosques són, més properes lingüísticament són les localitats que connecten.

Malgrat que es tracta de mapes amb un poder heurístic limitat, els *network maps* permeten detectar tant les àrees amb una major homogeneïtat lingüística com aquelles zones de transició entre blocs dialectals. Aquestes constitueixen franges de discontinuïtat, espais caracteritzats per la presència de línies clares entre punts contigus –tal com es desprèn de la figura 3. Aquest mapa forneix, alhora, una primera aproximació des d'una òptica global a la qüestió de com s'estructura geogràficament la variació dialectal de la llengua catalana.

4.2. A la recerca d'una clusterització estable: el *noisy clustering* i els *composite cluster maps*

L'anàlisi de conglomerats ha estat un dels mètodes de classificació de les varietats dialectals més utilitzats en dialectometria des del treball de Shaw (1974). En essència, es tracta d'un procediment iteratiu que selecciona i fusiona els dos punts més propers d'una matriu de distàncies. En formar un nou clúster, la distància entre aquest nou punt i la resta d'elements de la matriu s'ha de recalcular, i aquesta operació es repeteix a cada nova clusterització. Aquest procediment acaba produint un dendrograma d'estructura jeràrquica en el qual els conjunts de varietats lingüísticament més properes s'agrupen de forma binària. L'algorisme de clusterització utilitzat en aquest treball ha estat l'UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*).

Malgrat la gran quantitat d'estudis dialectomètrics que han fet ús de l'anàlisi de conglomerats, és un fet acceptat que es tracta d'un sistema taxonomètric inestable. Aquesta inestabilitat radica en el fet que, a voltes, pot haver-hi diversos parells d'elements amb una distància similar a la matriu, de manera que diferències mínimes en els inputs poden donar lloc a classificacions dendrogràfiques sensiblement diferents.

Precisament l'afany de superar aquesta mancança ha dut, en els darrers anys, a la presentació de dos nous sistemes de clusterització de major estabilitat: el *noisy clustering* i el *bootstrapping* (Nerbonne et al., 2008). Bàsicament, el *noisy clustering* consisteix a afegir petites quantitats de soroll aleatori –*random noise*– a la matriu de distàncies en successives clusteritzacions. Contràriament, en el *bootstrapping* no es parteix de la matriu de distàncies sinó dels propis inputs: el corpus d'entrada es modifica mínimament a cada clusterització, eliminant-ne alguns mots i permetent que n'hi apareguin de repetits. El resultat d'ambdues tècniques és un dendrograma de consens –o *composite dendrogram*.

Per interpretar correctament un dendrograma de consens –vg. la figura 4– cal tenir en compte dues dades: en primer lloc, que els números associats als clústers indiquen la quantitat de vegades que un conjunt de varietats han constituït un mateix grup en les diverses iteracions del procés; i, en segon lloc, que la llargada horitzontal de les línies reflecteix la distància cofenètica mitjana en què un conjunt de varietats s'han agrupat entre elles en el total de clusteritzacions. En el present treball s'ha optat per utilitzar el *noisy clustering*; la quantitat mitjana de soroll aleatori afegida ha estat de 0,33 i s'han realitzat 100 iteracions del procés.

El paquet de programari integrat L04 ofereix, a més, la possibilitat de projectar i visualitzar en un *composite cluster map* la informació continguda en un dendrograma de consens. Aquesta tècnica de mapatge es val de l'anomenada *poligonització de Voronoi* per traçar "fronteres" entre localitats adjacents i convertir l'àrea geogràfica estudiada en el que Goebl (1983) anomena un *parquet poligonal*. A continuació, aquestes línies s'ombregen de tal manera que el grau de foscor és inversament proporcional a la distància cofenètica entre punts contigus que registra el dendrograma de consens. D'aquesta manera –vg. la figura 5–, com més fosques són les línies divisòries, major és la distància lingüística entre el conjunt de varietats

situades a banda i banda de la ratlla; i a l'inrevés, com més clares són les línies, major és l'afinitat lingüística entre els grups de varietats entre les quals s'interposen.

4.3. L'anàlisi multidimensional (MDS) i els mapes RGB

La darrera tècnica estadística emprada per visualitzar les interdistàncies entre les varietats del COD ha estat l'anàlisi multidimensional (MDS), gràcies a la qual és possible reduir a un nombre limitat de dimensions informació continguda originalment en una gran quantitat de dimensions (en el cas del COD, 82, atès que es comparen entre elles 82 varietats de la llengua). Aquesta tècnica va ser utilitzada per primera vegada en dialectometria a Black (1976), un treball en què s'examinaven les relacions entre diversos dialectes de la llengua Bikol, parlada a les Filipines.

L'ús de l'anàlisi multidimensional presenta diversos avantatges respecte dels sistemes jeràrquics aglomeratius: en primer lloc, és un mètode de projecció de les distàncies lingüístiques estadísticament estable; i, en segon lloc, permet examinar les relacions entre varietats des d'una perspectiva més *global* que no els dendrogrames de consens. La figura 6 mostra els resultats d'aplicar l'anàlisi multidimensional a la matriu de distàncies provinent del COD.

Finalment, el paquet L04 permet també de visualitzar els resultats de l'MDS en els anomenats *mapes RGB* (o *RGB maps*). En aquests mapes, cada punt d'enquesta s'acoloreix de vermell, verd o blau en proporció al pes que hi tenen les respectives primera, segona i tercera coordenades de l'anàlisi multidimensional. El principal avantatge dels mapes RGB, tal com s'observa a la Figura 7, és que afavoreixen una interpretació dels resultats en termes de contínuums dialectals (Nerbonne, 2008).

5. Anàlisi dels resultats

A la llum dels resultats que recullen les figures 3, 4, 5, 6 i 7, sembla que la divisió tradicional de l'àrea lingüística catalana en dos grans blocs dialectals, l'oriental i l'occidental, es confirma en l'anàlisi dialectomètrica. Aquesta divisió es fa especialment palesa en el dendrograma de consens i el *composite cluster map* de les figures 4 i 5, que descrivim tot seguit.

Aquests gràfics reflecteixen diversos fets d'interès: d'una banda, confirmen la gran homogeneïtat lingüística de les varietats del català oriental central, que conformen un únic clúster sense pràcticament diferències internes. D'altra banda, el fet que s'incloguin en aquest grup les varietats tradicionalment adscrites al català septentrional de transició sembla assenyalar en la direcció que, a dia d'avui, ja no és possible de mantenir l'existència d'aquesta subàrea dialectal com una realitat diferenciada del català central. Més al nord, les quatre varietats sota domini administratiu francès es clusteritzen, també, amb les varietats centrals. Tanmateix, el fet que no constitueixin un grup sòlid entre elles i que es detectin diferències de pes entre Ceret, per un costat, i Sallagosa, Prada i Perpinyà, per l'altre, aconsella una revisió de les dades recollides en aquestes localitats. Altrament, és esperable una certa fragmentació interna d'aquesta àrea atesa la situació sociolingüística en què es troba immersa.

Pel que fa als parlars insulars, les figures 4 i 5 confirmen l'isolament de la varietat algueresa de la llengua, secularment aïllada de la resta del domini lingüístic. Mostren, també, que el clúster de les varietats tradicionalment anomenades baleàriques inclou, només, les parles de l'illa de Mallorca i de les Pitiüses; i, dins de les primeres, destaca encara el caràcter diferenciat de la parla de Sóller. Pel que fa a aquests parlars, el fet més remarcable és, potser, la posició indefinida en què apareixen les varietats de Maó i de Ciutadella, un grup homogeni a mig camí entre els parlars centrals i els insulars. Tornarem sobre aquest punt més endavant.

En l'àmbit ja del català occidental, el dendrograma de consens identifica clarament la frontera entre els dialectes nord-occidental i valencià. Cal, tanmateix, fer algunes puntualitzacions a aquesta classificació.

D'una banda, els resultats corresponents al valencià reflecteixen que es tracta, en conjunt, d'una àrea força homogènia. Tot i que el mapa de la figura 5 permet intuir les àrees del

valencià septentrional i de l'apitxat, el fet que constitueixin un clúster autònom un nombre poc significatiu de vegades –57 i 54 sobre 100, respectivament– desaconsella extreure'n conclusions categòriques. Pel que fa a les àrees del valencià meridional i central (vg. Veny, 1982 i Clua, 1998), no sembla que es puguin identificar a partir d'aquests resultats.

Sí que és clara, en canvi, l'adscripció de Vinaròs al català nord-occidental, així com l'isolament de la varietat de Benavarri, que en reflecteix el caràcter de parla de transició cap a l'aragonès. S'identifica, també amb claredat, el gran *plateau* del lleidatà, que comprèn les localitats de l'àrea central del nord-occidental. Aquest clúster coincideix amb els resultats de Viaplana (1999), el primer estudi dialectomètric que pren en consideració l'àmbit del català nord-occidental. Segons Viaplana, les varietats que conformen aquest clúster es caracteritzen per presentar un grau d'orientalització més elevat que no la resta de varietats nord-occidentals. Aquesta tesi sembla guanyar força en comprovar que l'altre gran conjunt de varietats nord-occidentals inclou, d'una banda, la resta de parlars de la Franja (Tamarit de Llitera i Fraga) i, de l'altra, certes varietats del pallarès, el ribagorçà i el tortosí (respectivament, Sort; el Pont de Suert; i Vall-de-roures, Tortosa, Amposta, Vinaròs i Gandesa). Es tracta, doncs, de varietats que semblen mantenir en major grau bé el seu caràcter de parles de transició (en el cas del tortosí), bé la seva idiosincràsia fruit d'un històric aïllament (en el cas del pallarès i del ribagorçà). La impermeabilitat de les varietats de la Franja a un hipotètic procés d'anivellament dialectal caldria cercar-la en factors de caire sociolingüístic i, en darrer terme, en la política lingüística de la comunitat autònoma de l'Aragó.

Per la seva part, els resultats de l'anàlisi multidimensional (vg. les figures 6 i 7) aporten un seguit d'informacions complementàries invisibles al dendrograma de consens de 4. Posen en dubte, per exemple, la filiació de l'alguerès al català oriental, una classificació no exempta de controvèrsia que tanmateix ha estat defensada, entre d'altres, per Veny (1982: 23). Des d'una òptica dialectomètrica, però, la ubicació de l'alguerès en relació amb la resta de varietats de la llengua és d'absoluta equidistància.

En segon lloc, les característiques diferencials del parlar de Solsona semblen tenir prou pes com per situar aquesta varietat a una petita distància del clúster format per la resta de varietats centrals. Es tracta, doncs, d'una ubicació que ve a refermar el caràcter de transició del parlar solsoní entre el català central i el nord-occidental.

D'altra banda, el gràfic de 6 permet d'identificar novament aquelles àrees que 3 ja assenyalava com a lingüísticament més homogènies: el central i, en menor grau, el lleidatà i el valencià. Tanmateix, el veritable interès de 6 quant a la homogeneïtat dels clústers radica precisament en el fet que situa determinades varietats entre clústers homogenis. Per exemple, de 6 es desprèn que les varietats septentrionals del País Valencià i les varietats meridionals del Principat i de l'Aragó constitueixen una zona de transició clara entre el nord-occidental i el valencià: el tortosí. D'altra banda, palesa també les diferències que mantenen entre elles les varietats baleàriques. Aquesta afirmació és vàlida especialment per al menorquí, que apareix a la figura 6 en una situació intermèdia entre el clúster format per les varietats centrals i el clúster integrat pels parlars de Mallorca, Eivissa i Formentera. La proximitat de la varietat de Maó al central és, encara, més acusada. En aquest sentit, cal tenir present que el subcorpus del qual s'han obtingut els resultats no recull dades relatives a la fonotàctica de la llengua, que probablement tendirien a augmentar el grau d'homogeneïtat lingüística d'aquests parlars.

6. Discussió i perspectives

Aquest article constitueix la primera aplicació de la distància de Levenshtein a un corpus dialectal del català. Alhora, presenta diverses tècniques de visualització de la distància lingüística entre varietats que mai no s'havien utilitzat en dialectologia catalana. D'una banda, doncs, contribueix a la diversificació metodològica de la dialectometria local i referma alguns posicionaments centrals de la dialectologia tradicional; d'altra banda, tanmateix, ofereix noves possibilitats d'anàlisi que poden aportar més dades a qüestions tradicionalment poc estudiades,

com ara la relació del menorquí amb la resta de parlars baleàrics, la filiació de l'alguerès o l'impacte de l'efecte-frontera entre varietats pertanyents a diferents territoris administratius.

Endemés, el paraigua dialectomètric aixopluga una multiplicitat de mètodes d'anàlisi en funció dels quals es poden obtenir resultats divergents. Avaluar quines d'aquestes diferències es deuen a qüestions estrictament metodològiques i determinar, doncs, la influència del mètode en els resultats, ha de ser un element central de la futura recerca dialectomètrica. Paral·lelament, s'està treballant per comprovar la validesa del mètode dialectomètric en l'estudi d'altres fenòmens lingüístics, com ara el procés de desdialectalització d'algunes varietats del català[4].

Finalment, diversos treballs han eixamplat el ventall d'aplicacions possibles de la LD a altres camps de recerca, com ara el dels estudis perceptius. Fóra interessant, doncs, aprofitar aquests avenços per estudiar les diferències entre la distància lingüística real i la distància percebuda pels parlants de les diferents varietats de la llengua catalana.

Agraïments

Vull agrair a diversos membres del Departament Alfa-Informatica de la Universitat de Groningen el seu ajut en l'anàlisi de les dades del COD: a Çagri Çoltekin, pel seu assessorament en la conversió a X-SAMPA de les transcripcions fonètiques; a Jelena Prokic i Martijn Wieling, pel seu suport en el procés d'aplicació de la LD i en la confecció dels mapes finals; i, molt especialment, a John Nerbonne, per la seva total disponibilitat a resoldre qualsevol dubte sorgit en el procés de dialectometrització. D'altra banda, aquest article tampoc no hauria estat possible sense el mestratge de Joaquim Viaplana i Esteve Clua en el camp de la dialectometria, així com de Maria-Rosa Lloret, investigadora principal del projecte en el qual s'inscriu aquest treball.

REFERÈNCIES BIBLIOGRÀFIQUES

ALEWIJNSE, BART; NERBONNE, JOHN; VAN DER VEEN, LOLKE; & FRANZ MANNI (2007): «A Computational Analysis of Gabon varieties». Dins de PETYA OSENOVA ET AL. (ed.) *Proceedings of the RANLP Workshop on Computational Phonology. Workshop at the conference* Recent Advances in Natural Language Processing. Borovetz, pàg. 3-12.

CLUA, ESTEVE (1998): *Variació i distància lingüística. Classificació dialectal del valencià a partir de la morfologia flexiva*. Universitat de Barcelona, tesi doctoral inèdita.

CLUA, ESTEVE; & MARIA-ROSA LLORET (2006) «New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)». Dins de MONTREUIL, JEAN-PIERRE Y. (ed.), *New Perspectives on Romance Linguistics*, vol. 2 (Phonetics, Phonology, and Dialectology). Amsterdam / Philadelphia: John Benjamins.

CLUA, ESTEVE; VALLS, ESTEVE; & JOAQUIM VIAPLANA (2008): «Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà». Dins de BLAIKNER HOHENWART, GABRIELE; BORTOLOTTI, EVELYN; & EMESE LÖRINCZ: *Ladinometria. Miscellanea per Hans Goebl per il 65º compleanno. Edizione multiligue*, vol. 2, pàg. 27-42. Vigo di Fassa: Istituto Culturale Ladino.

GOEBL, Hans (1983): «Parquet polygonal et treillis triangulaire: les deux versants de la dialectométrie interponctuelle». Dins de «Revue de Linguistique Romane», vol. 47, pàg. 353-412.

GOEBL, HANS (2003): «Regards dialectométriques sur les données de l'Atlas Linguistique de la France (ALF): relations quantitatives et structures de profondeur». Dins d'«Estudis Romànics», vol. XXV, pàg. 59-121.

HEERINGA, WILBERT (2004): *Measuring Dialect Pronuntiation Differences using Levenshtein Distance*. Groningen: Groningen Dissertations in Linguistics 46.

[4] Per a una anàlisi del procés d'anivellament dialectal del català nord-occidental, vg. VALLS (2008).

HEERINGA, WILBERT; KLEIWEH, PETER; GOOSKENS, CHARLOTTE & JOHN NERBONNE (2006): «Evaluation of String Distance Algorithms for Dialectology». Dins de NERBONNE, JOHN; & HINRICHS, ERHARD (ed.): *Linguistic Distances, ACL Workshop held at ACL/COLING*. Sydney Shroudsburg, PA: ACL, pàg. 51-62.

KESSLER, BRETT (1995): «Computational Dialectology in Irish Gaelic». In *Seventh Conference of the European Chapter of the Association for Computational Linguistics (Dublin, Ireland)*. San Francisco: Morgan Kaufmann Publishers, pàg. 60-66.

NERBONNE, JOHN; KLEIWEG, PETER; MANNI, FRANZ; I HEERINGA, WILBERT (2008): «Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering». Dins de PREISACH, CHRISTINE; SCHMIDT-THIEME, LARS; BURKHARDT, HANS; & REINHOLD DECKER (eds.): *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society.* Berlin: Springer, pàg. 647-654.

NERBONNE, JOHN (2008): «Mapping Aggregate Variation». Enviat a RABANUS, STEPHAN; KEHREIN, RONALD; & ALFRED LAMELI (eds.): *Mapping Language*, vol. de la sèrie *Language and Space.* Berlin: Mouton De Gruyter.

OSENOVA, PETYA; HEERINGA, WILBERT; & NERBONNE, JOHN (2007): «A Quantitative Analysis of Bulgarian Dialect Pronunciation». Enviat a *Zeitschrift für slavische Philologie.*

SÉGUY, JEAN (1971): «La relation entre la distance spatiale et la distance lexicale». Dins de «Revue de Linguistique Romane», vol. 35, pàg. 335-357.

SÉGUY, JEAN (1973): *Atlas Linguistique de la Gascogne.* París: Éditions du Centre National de la Recherche Scientifique, 6 vol.

SÉGUY, JEAN (1973): «La dialectométrie dans l'*Atlas linguistique de la Gascogne*». Dins de «Revue de Linguistique Romane», vol. 37, pàg. 1-24.

SHAW, DAVID (1974): «Statistical analysis of dialect boundaries». Dins de «Computers and the Humanities», 8, pàg. 173-177.

VALLS, ESTEVE (2008): «La desdialectalització del català nord-occidental: cap a una convergència total amb l'estàndard? Una anàlisi dialectomètrica sobre el procés d'anivellament dialectal en quatre generacions de parlants nord-occidentals». Universitat de Barcelona, treball d'investigació inèdit.

VALLS, ESTEVE; NERBONNE, JOHN; PROKIC, JELENA; WIELING, MARTIJN; CLUA, ESTEVE; & MARIA-ROSA LLORET (2009): «Applying the Levenshtein Distance to Catalan dialects: a brief comparison of two dialectometric approaches», en preparació.

VENY, JOAN (1982): *Els parlars catalans (síntesi de dialectologia)*. Palma: Editorial Moll.

VIAPLANA, JOAQUIM (1999): *Entre la dialectologia i la lingüística. La distància lingüística entre les varietats del català nord-occidental*. Barcelona: Publicacions de l'Abadia de Montserrat.

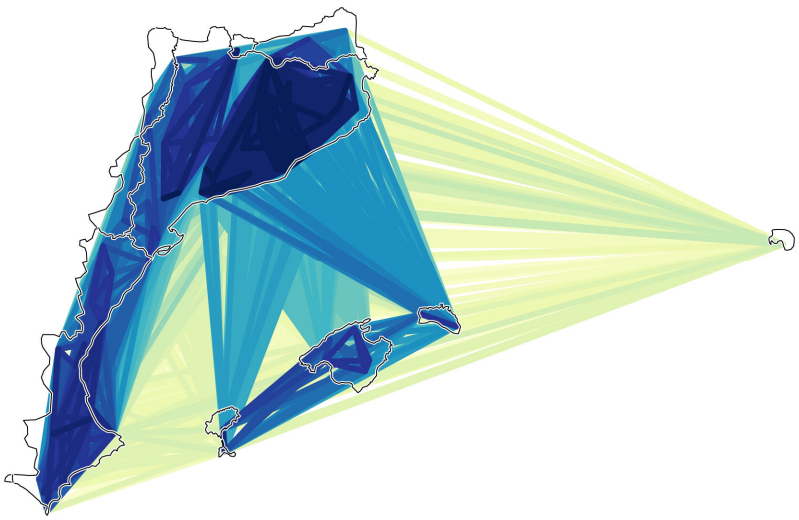

Figura 3. Network map *a partir de les dades del COD*

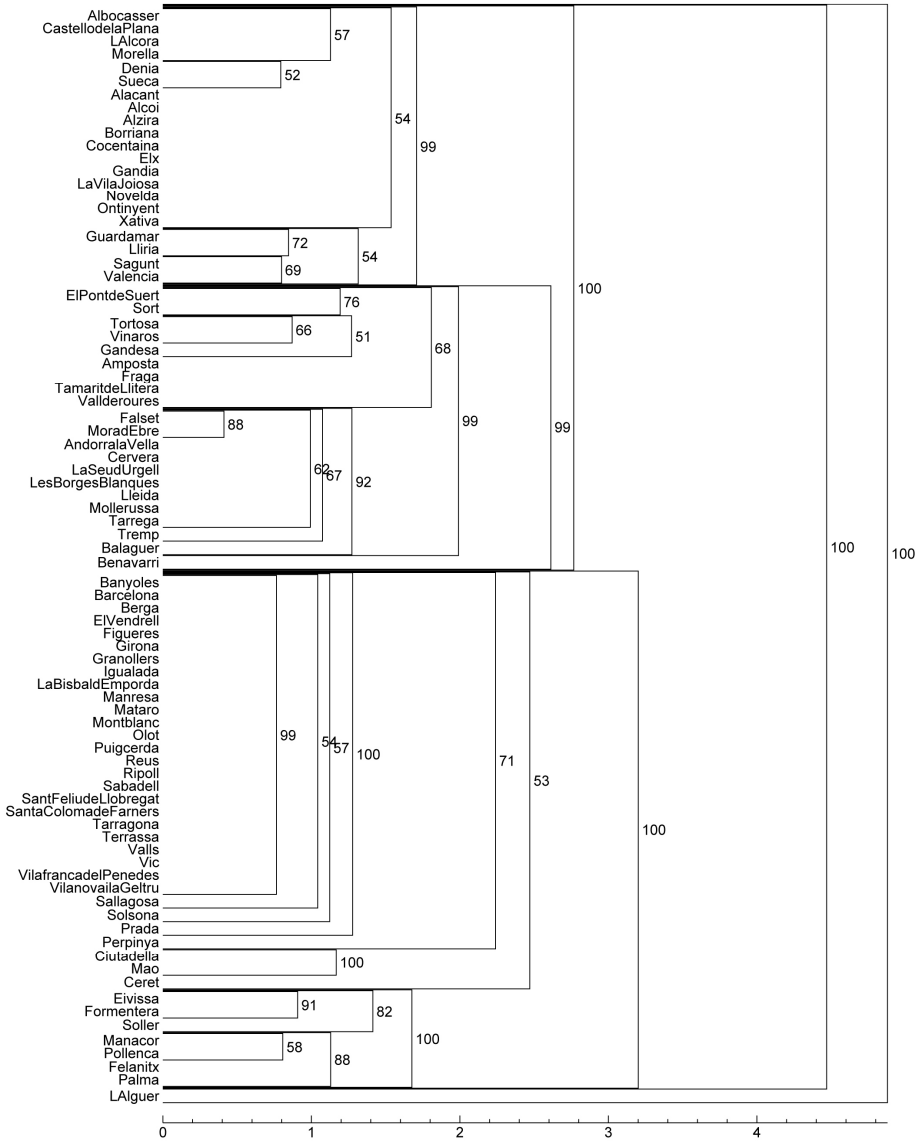

Figura 4. Dendrograma de consens obtingut mitjançat el noisy clustering*.*
Nombre d'iteracions: 100. Quantitat de soroll aleatori afegit: 0,33

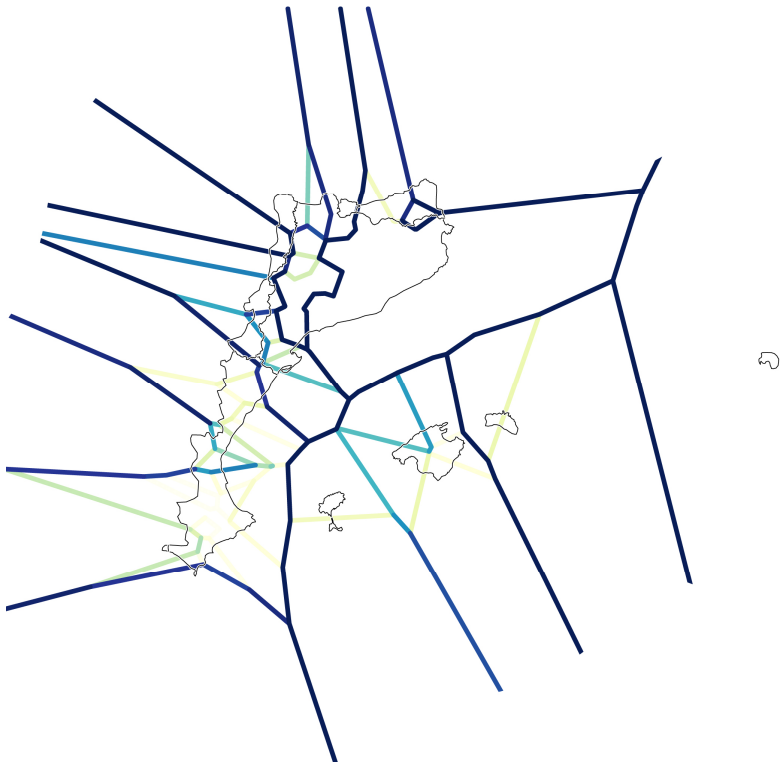

Figura 5. Composite cluster map*. Nombre d'iteracions: 100.*
Quantitat de soroll aleatori afegit: 1

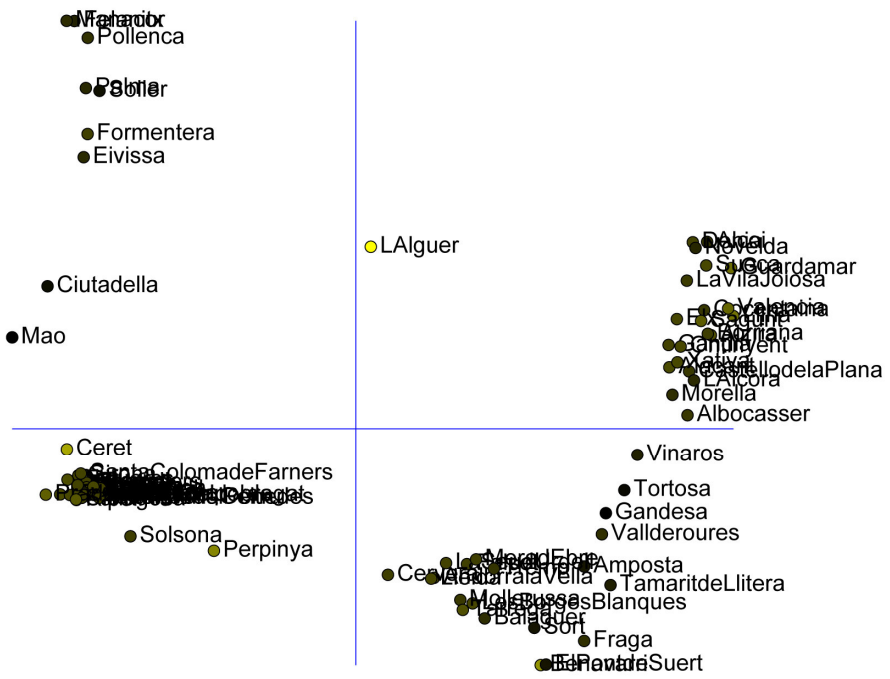

Figura 6. Resultats de l'anàlisi multidimensional.
R (correlació amb la matriu de distàncies) = 0,966

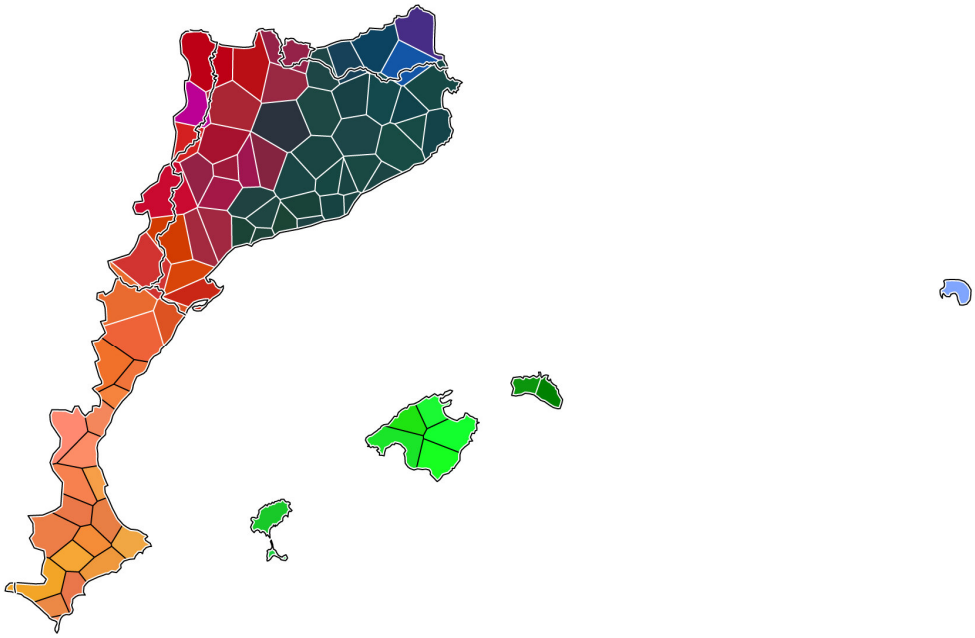

Figura 7. Mapa RGB.
R (correlació amb la matriu de distàncies) = 0,966