Studia Romanica Posnaniensia 43/2 (2016): 5-27
Adam Mickiewicz University Press
Received: 31.01.2016 / Accepted: 1.05.2016

DOI: 10.14746/strop.2016.432.001
ISSN 0137-2475, eISSN 2084-4158

LINGÜÍSTICA

# La subagrupació romànica de la llengua catalana: una aproximació dialectomètrica de base fonètica a l'*Atlas Linguistique Roman*

## The Romance subgrouping of the Catalan language: A dialectometric approach to the *Atlas Linguistique Roman*

Esteve Valls

Centro Ramón Piñeiro para a Investigación en Humanidades & Departament de Filologia Catalana, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona
e-mail: e.valls@ub.edu

Manuel González

Centro Ramón Piñeiro para a Investigación en Humanidades & Departamento de Filoloxía Galega, Universidade de Santiago de Compostela, Av. Castelao, s/n, 15782 Santiago de Compostela
e-mail: manuel.gonzalez.gonzalez@usc.es

**Abstract**

The aim of this paper is to present a first dialectometric approach to the geolinguistic distribution of the Romance varieties using a new corpus based on the *Atlas Linguistique Roman* (*ALiR*) which is currently being developed at the Centro Ramón Piñeiro para a Investigación en Humanidades. However, the dialectometrization of *ALiR* will allow us not only to analyze the linguistic distance among the varieties of *Romania Antiqua*, but also to shed light from a quantitative perspective on several issues that have historically raised controversy. Specifically, this paper deals with the Romance subgrouping of the Catalan language, an issue that has been recently reopened by Radatz (2012).

**Keywords:** Romance studies, dialectology, dialectometry, Catalan subgrouping

## INTRODUCCIÓ

Tot i que l'anàlisi dialectomètrica de l'*Atlas Linguistique Roman* es trobava entre els objectius inicials d'aquest projecte –vg. Contini (1992, p. 345)–, aquest article constitueix –que en tinguem constància– la primera anàlisi quantitativa global que es du a terme a partir d'un corpus de dades que abraça la totalitat de la *Romania Antiqua*. Fins al moment, les anàlisis dialectomètriques de corpus de dades romàniques han provingut, sobretot, de l'Escola Dialectomètrica de Salzburg (EDS), encapçalada per Hans Goebl, que s'ha valgut del programari *Visual Dialectometry* (VDM) per analitzar la variació i la distància lingüística entre varietats a partir d'un nombre considerable d'atles lingüístics de l'àmbit romànic, com ara l'*Atlas Linguistique de la France* (Goebl, 2003) o l'*Atlante Italo-Svizzero* (Goebl, 2008)[1]. A diferència dels esmentats treballs, però, en aquest article s'aborda l'anàlisi de la variació i la distància lingüística entre les varietats romàniques a partir d'un subcorpus que conté dades de *tots* els atles lingüístics –publicats o inèdits– confegits fins al moment amb dades d'aquests parlars. Aquest subcorpus s'ha dissenyat i s'està desenvolupant al Centro Ramón Piñeiro para a Investigación en Humanidades (CRPIH).

El primer objectiu d'aquest treball és, doncs, presentar una aproximació dialectomètrica a la distribució geolectal de l'àmbit lingüístic romànic. A més, però, s'hi aborda un dels temes que han generat més controvèrsia en la història de la romanística: la qüestió de la subagrupació romànica de la llengua catalana, un debat que Radatz (2012) ha reobert recentment.

Finalment, i des d'un punt de vista estrictament metodològic, aquest article es proposa de contribuir a la valorització de la dialectometria com un conjunt de sistemes d'anàlisi necessaris per aproximar-se, des d'una òptica quantitativa global, a qüestions sovint abordades des d'un enfocament estrictament qualitatiu. Aquesta anàlisi serà possible gràcies a l'aplicació de la *distància de Levenshtein* mitjançant la interfície *Gabmap*, desenvolupada al *Center for Language and Cognition* de la Universitat de Groningen.

## 1. CORPUS

L'*ALiR* constitueix el primer intent de dur a terme una anàlisi global de la *Romania Antiqua* a partir de la síntesi de tots els atles lingüístics –publicats o encara inèdits– de l'àmbit romànic. Es tracta, doncs, d'un atles de segona ge-

[1] La tasca duta a terme per Goebl i els seus col·laboradors a l'Escola Dialectomètrica de Salzburg durant les últimes dècades ha estat ingent i queda fora del nostre abast repassar-la en aquest treball. Tanmateix, qui cerqui una introducció general als principis de l'EDS la podrà trobar, per exemple, a Goebl (1991, 2006) o a l'adreça <www.dialectometry.com> [consulta: 27.01.2016].

neració, interpretatiu: en comptes de partir de dades recollides *ad hoc* mitjançant la realització de noves enquestes dialectals, l'*ALiR* parteix de les dades obtingudes durant l'elaboració de la cinquantena llarga d'atles lingüístics que s'han confeccionat al llarg del segle XX en els diferents estats, regions o àmbits lingüístics de l'espai romànic[2].

Aquest treball es basa en un subcorpus de l'*ALiR* extret de la base de dades que s'està elaborant actualment al CRPIH per posar a l'abast del públic, d'una manera àgil i en accés obert, tota la informació –etimològica, dialectal, etnogràfica, etc.– que contenen les *Síntesis Romàniques* que els col·laboradors de l'*ALiR* han anat publicant fins al moment –vg. Tuaillon & Contini (1996) i Veny & Contini (2001, 2009). En aquestes síntesis, un o diversos autors analitzen les designacions romàniques d'un determinat concepte a partir de les informacions proporcionades pels comitès responsables de gestionar les dades dels diferents àmbits estatals o lingüístics. Concretament, el subcorpus a partir del qual s'ha realitzat l'anàlisi dialectomètrica consta de 36.808 ítems corresponents a les designacions romàniques de 48 conceptes en un total de 1.010 punts d'enquesta. Els conceptes escollits són els que ja s'han incorporat a la base de dades del CRPIH[3]: *abeille, alouette, araignée, aujourd'hui, belette, berceau, blatte et ténébrion, bousier, chenille, cigale, cloporte, courtilière, crapaud, demain, essaim, foie, forgeron, fourmi, fourmilière, grillon, guêpe, hêtre, hier, libellule, lundi, mante religieuse, mardi, merle, mille-pattes, miroir, moineau, moucheron, moustique, orvet, paille, papillon, perce-oreille, punaise des lits, rouge-gorge, salamandre, serpent, taureau, têtard, toile d'araignée, tortue, ver de terre, ver luisant*. En total, el subcorpus d'aquest treball consta de 273.499 segments fonètics, 136 dels quals són únics.

Com es desprèn de la figura 1, el nombre de respostes disponibles per localitat varia sensiblement en funció del concepte. Aquesta figura mostra la distribució dels 36.808 ítems en les 1.010 localitats enquestades. Les localitats amb un major nombre d'ítems disponibles per al càlcul d'interdistàncies s'han acolorit amb tonalitats de blau fosc; per contra, les localitats amb un nombre menor d'ítems disponibles presenten tonalitats de blau més clar. La quantitat d'ítems comparats oscil·la entre els 47 de les localitats acolorides de blau més fosc i els 4 de les localitats acolorides de blanc[4].

---

[2] Vg. Tuaillon & Contini (1996) per saber quins són els atles de primera generació en què es basa l'*ALiR*.

[3] Aquesta base de dades, així com tots els materials necessaris per a la seva dialectometrització, es poden descarregar a l'adreça <www.atlaslinguistiqueroman.wordpress.com> [consulta: 27.01.2016].

[4] Com s'observa, la minsa informació disponible en algunes localitats fa que els resultats s'hagin d'interpretar amb una certa cautela.

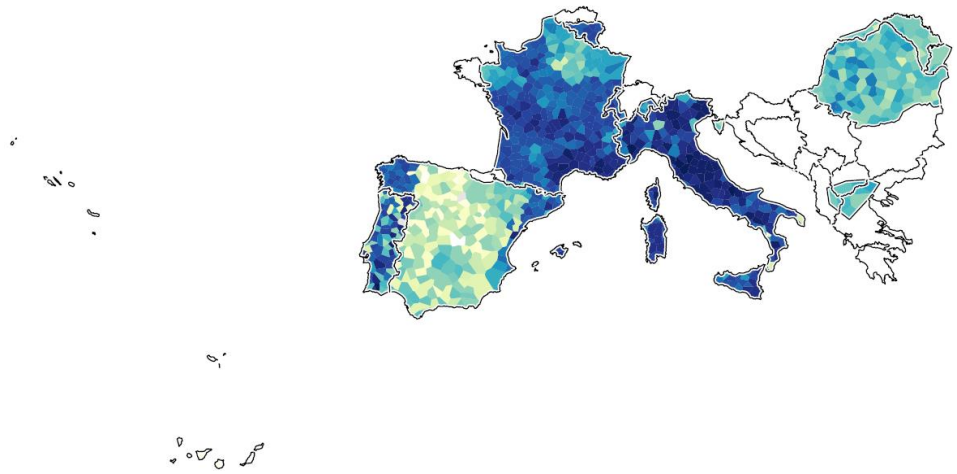


Figura 1. Distribució dels 36.808 ítems del corpus a les 1.010 localitats enquestades

Per comprendre aquesta figura cal tenir en compte tres fets: d'una banda, que s'han desestimat 27 punts d'enquesta dels 1.037 de què consta l'*ALiR* perquè pràcticament no es disposava de cap dada sobre les denominacions dels ítems seleccionats en aquelles poblacions. D'altra banda, que s'han mantingut diverses localitats dels àmbits lingüístics asturlleonès, castellà, aragonès i romanès que presenten una quantitat molt minsa de dades per evitar que la xarxa de localitats enquestades en aquests àmbits lingüístics fos massa desequilibrada respecte a les d'altres àmbits que compten amb xarxes de localitats més denses. La migrada informació disponible sobre els esmentats àmbits lingüístics limitarà, per força, la possibilitat d'analitzar-ne la variació interna. I, finalment, cal tenir en compte que, en els àmbits lingüístics del gallec i de les varietats gal·loromàniques i ita-loromàniques, els punts d'enquesta no representen localitats, sinó petits agrupaments de localitats lingüísticament molt properes anomenades *caselles* –o, en francès, *cases*.

## 2. METODOLOGIA

El càlcul de la distància fonètica entre varietats que és a la base de les anàlisis posteriors s'ha dut a terme mitjançant l'aplicació de l'anomenada *distància de Levenshtein* (LD). La distància de Levenshtein 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 de segments fonètics en una altra. Aquestes

operacions poden ser insercions, supressions o substitucions, i en la versió més simple de l'LD tenen totes tres un cost d'1. La distància final entre dos dialectes és, doncs, la que s'obté de transformar d'una varietat a l'altra les realitzacions fonètiques d'un nombre determinat de mots. Tot seguit s'exemplifica el funcionament bàsic de l'LD a partir de les realitzacions fonètiques del concepte *abeille* en dues varietats romàniques: una pertanyent al català oriental central barceloní i una altra pertanyent al romanx sutsilvà. En aquest cas, el cost final de la transformació és 5:

Taula 1. Exemple del funcionament bàsic de l'LD

| | | | |
|---|---|---|---|
| Català oriental central barceloní | ə ˈ β ɛ j ə | substitueix ə / ɐ | 1 |
| | ɐ ˈ β ɛ j ə | substitueix β / v | 1 |
| | ɐ ˈ v ɛ j ə | esborra ɛ | 1 |
| | ɐ ˈ v j ə | substitueix ə / ø: | 1 |
| | ɐ ˈ v j ø: | insereix l | 1 |
| Romanx sutsilvà | ɐ ˈ v j ø: l | | |
| Total | | | 5 |

Amb l'objectiu de reduir l'impacte que la diferent llargada dels mots comparats podria tenir en els resultats –dos ítems llargs podrien aportar més diferències que dos ítems curts, tot i que en termes relatius els dos ítems llargs podrien ser més semblants entre ells que no els ítems curts–, s'ha procedit a calcular la distància relativa entre cada parell de mots. Això s'ha fet dividint la distància total entre cada parell de pronúncies pel nombre de segments de cada alineament; en l'exemple que acabem de veure, la distància normalitzada seria de 0,8 –les 5 diferències dividides per 6, que és el total de segments alineats.

Una vegada obtinguda la matriu de distàncies fonètiques del conjunt de varietats, s'han aplicat diversos sistemes de classificació i visualització del paisatge dialectal: anàlisis jeràrquiques aglomeratives, mapes probabilístics a partir de dendrogrames de consens i anàlisis multidimensionals[5]. Tant el càlcul de l'LD com l'aplicació d'aquests sistemes taxonòmics i representacionals s'han dut a terme mitjançant la interfície *Gabmap*[6].

---

[5] Vg. Valls (2013, p. 55-58) per aprofundir en les característiques i el funcionament d'aquests sistemes de classificació i visualització de la distància lingüística.

[6] Aquesta interfície –vg. Nerbonne, Colen, Gooskens, Kleiweg & Leinonen (2011)– ha estat desenvolupada al *Center for Language and Cognition* de la Universitat de Groningen i és d'accés lliure a l'adreça <http://www.gabmap.nl/> [consulta: 30.01.2016].

## 3. LA SUBAGRUPACIÓ ROMÀNICA DEL CATALÀ

L'interès de la romanística per la llengua catalana s'ha focalitzat bàsicament en dues qüestions: d'una banda, el debat sobre la consideració del català com una llengua autònoma o com una varietat dialectal més del diasistema occità; i, de l'altra, el debat, reobert molt recentment –vg. Radatz (2012)–, sobre la seva subagrupació gal·loromànica o iberoromànica.

Com és sabut, el grau d'individuació de la llengua catalana respecte de l'occità va ser un dels punts en què els autors dels primers manuals de romanística van mostrar més vacil·lacions. Així, a la primera edició de la *Gramàtica de les llengües romàniques* (1836), Diez afirmava que el català era una varietat del provençal; en canvi, a la segona edició d'aquesta obra, publicada l'any 1856, el català ja va rebre la consideració de llengua autònoma. Aquest canvi de parer no es va reflectir en la primera edició del *Grundiss* de Gröber (1888), en la qual Morel-Fatio afirmava, com Diez cinquanta anys abans, que el català era una varietat dialectal de la llengua provençal. Novament, però, la segona edició d'aquest manual –de 1904-1906– va comportar el reconeixement, per part del mateix Morel-Fatio i de Saroïhandy, del català com a llengua romànica independent. El debat sobre la independència lingüística del català es va cloure amb la contribució de Meyer-Lübke *Das Katalanische* (1925), a partir de la qual "el català va ser acceptat unànimement dins el catàleg oficial de les llengües romàniques" (Radatz, 2012, p. 203).

Una vegada tancat el debat sobre la individuació de la llengua catalana, la romanística es va interessar per determinar si tipològicament aquesta llengua era més afí a les varietats gal·loromàniques o a les varietats iberoromàniques. Aquesta nova controvèrsia va girar a l'entorn de dos pols d'opinió: d'una banda, el pol format per autors com Meyer-Lübke o Griera, que defensaven el caràcter gal·loromànic del català, una llengua fortament emparentada amb l'occità; i, de l'altra, el pol format per autors com Menéndez Pidal, Meyer o Alonso, que consideraven que el català era la continuació orgànica de l'aragonès cap a l'est dins d'una cadena ininterrompuda d'idiomes iberoromànics (vg. Radatz, 2012, p. 204). Baldinger (1958, p. 52, cit. per Radatz, 2012, p. 204) esmenta un tercer grup de lingüistes que adopten una posició intermèdia, segons la qual el català formaria una unitat pirinenca amb l'aragonès, el gascó i l'occità.

Fins a la publicació, l'any 1976, d'*El léxico catalán en la Romania*, de Germà Colón, els arguments que s'havien esgrimit per defensar el caràcter gal·loromànic o iberoromànic del català havien estat bàsicament de tipus fonètic i fonològic. Aquest darrer autor, en canvi, va aportar nous arguments en favor de la hipòtesi gal·loromànica a partir, exclusivament, d'una anàlisi del lèxic català. Duarte & Alsina (1984) resumeixen així les principals contribucions de Colón:

> Germà Colón examina acuradament la història del lèxic català i en dedueix que: a) en el lèxic català dels segles XIII i XIV hi ha una afinitat evident amb l'occità, especialment amb el llenguadocià; b) a partir del segle XV, i sobretot del segle XVI, es produeix un canvi de rumb i un més gran acostament al castellà; [...] c) el lèxic català, nogensmenys, havia quedat molt intensament marcat per la tria portada a terme abans del segle XV. (vol. I, p. 20)

Tot i el convenciment expressat per Badia i Margarit que l'obra de Colón representava la fi de la polèmica sobre la subagrupació romànica del català, sembla que avui l'opinió més estesa entre la comunitat de lingüistes és que el català és una "llengua pont". Paradoxalment, aquest terme va ser encunyat pel mateix Badia i Margarit i dóna compte del fet que el català "comparteix més afinitats fonètico-fonològiques amb la gal·loromània però segueix majoritàriament la iberoromània en les seves solucions morfològiques. És una fórmula que d'aleshores ençà s'ha anat repetint *ad nauseam* en tots els llibres de text de la filologia romànica" (Radatz, 2012, p. 205).

Tal com es desprèn d'aquesta última afirmació, Radatz es mostra profundament crític amb el concepte de "llengua pont", que titlla d'estàtic, ahistòric i inoperant, perquè "al si de la Romània contínua de l'edat mitjana, totes les llengües eren 'llengües pont'" (p. 206). De fet, aquest autor considera que el debat sobre la subagrupació del català, pel fet de basar-se en la llengua medieval, obvia dues qüestions crucials: d'una banda, que la llengua és un fet dinàmic que ha evolucionat sensiblement al llarg dels darrers segles; i, de l'altra –i precisament per aquest motiu–, que la vigència actual del contínuum romànic medieval s'hauria de reconsiderar: en primer lloc, perquè la vitalitat etnolingüística de les varietats aragoneses, gascones i llenguadocianes que limiten amb el català al nord es troba en una situació d'extrema precarietat; en segon lloc, perquè el contacte amb el francès és mínim, atès que la frontera política ha esdevingut, amb els segles, també una frontera ideològica –o, més tècnicament, una isoglossa sociolingüística– que ha afavorit les dinàmiques endogrupals en el si dels estats-nació i ha afeblit les relacions transfrontereres; i, en tercer lloc, perquè l'augment de les interaccions entre els pobles de l'Estat espanyol ha afavorit que el català d'avui s'hagi subordinat a la llengua castellana fins al punt de mantenir-hi una relació d'advergència –és a dir, de convergència unilateral. Per a Radatz, doncs, la consideració del català com a "llengua pont" es buida de sentit des del moment que la perspectiva estrictament històrica se substitueix per una anàlisi de l'evolució diacrònica de la llengua fins als nostres dies:

> Més aviat, el català convergeix imparablement amb el castellà i cada vegada més elements gal·loromànics comencen a ser percebuts com a "arcaïtzants", són relegats als registres formals i acaben finalment als diccionaris històrics. Per totes les raons exposades, la resposta a la qüestió de la *subagrupació del català* només es pot donar de forma

> dinàmica: com un procés mil·lenari que ha anat convertint una llengua d'encuny clarament gal·loromànic en un idioma que convergeix cada vegada més amb la iberoromània. (p. 208)

A més, aquest autor també es plany que la sintaxi ha estat tradicionalment negligida en el debat clàssic sobre la subagrupació del català. Tanmateix, des d'un enfocament dinàmic com el que proposa, l'anàlisi de l'evolució sintàctica de la llengua és crucial perquè, com que es tracta del component més refractari al canvi lingüístic, qualsevol canvi en aquest nivell indicaria que el procés d'advergència amb el castellà ja estaria afectant els estrats més profunds de l'estructura lingüística. Amb l'objectiu de comprovar fins a quin punt les tendències evolutives observades en els àmbits de la fonologia, la morfologia o el lèxic es poden fer extensibles a l'àmbit de la sintaxi, Radatz (2012) analitza una dotzena de trets morfosintàctics tradicionalment compartits amb les varietats gal·loromàniques que semblen estar evolucionant cap a solucions de tipus iberoromànic. Els resultats a què arriba l'autor són concloents: el català, avui, també evoluciona tipològicament del tipus sintàctic gal·loromànic a l'iberoromànic.

Finalment, aquest autor també critica la concepció monolítica de la llengua que, al seu parer, ha primat a l'hora d'abordar la qüestió de la subagrupació romànica del català, perquè no s'ha tingut en compte cap tipus de variació interna, tot i que, de fet, per a ell és evident que la densitat d'elements gal·loromànics varia en funció de la varietat dialectal. Per això no s'està d'establir una jerarquia orientativa de les varietats diatòpiques del català –baleàric > central > barceloní > valencià– en virtut de la major o menor presència d'elements de tipus gal·loromànic[7].

En definitiva, doncs, sembla que per Radatz (2012) només es pot donar una solució adequada a la qüestió de la subagrupació romànica del català si es parteix d'una anàlisi complexa de la realitat que: 1) no se cenyeixi a un estadi sincrònic concret, sinó que tingui en compte l'evolució diacrònica de la llengua; 2) no se cenyeixi a una única varietat, sinó que tingui en compte la diversitat dialectal de la llengua; i 3) no se cenyeixi a un únic registre o grau de formalitat, sinó que tingui en compte tot el contínuum estílistic de la llengua. En conseqüència, l'autor creu que ja no té sentit referir-se al català com una "llengua pont" entre les varietats aragoneses i les occitanes –i encara menys entre el castellà i el francès–, sinó que cal tenir en compte els tres vessants que s'acaben d'esmentar:

---

[7] Com s'observa, Radatz no inclou el català nord-occidental –ni el septentrional, ni l'alguerès– en aquesta jerarquia. De fet, aquest "descuit" té una certa tradició en la història de la lingüística catalana i ha fet que es parli del català nord-occidental com de la varietat més menystinguda entre les grans varietats diatòpiques de la llengua –vg. Valls (2013, pp. 64-68).

1. [En relació amb el] català preliterari, la subagrupació s'ha de fer clarament amb la gal·loromània com a diasistema catalano-occità.
2. [En relació amb el] català modern, caldria diferenciar entre varietats diatòpiques (més gal·loromànic a les Balears, menys al País Valencià i amb Catalunya enmig) i diafàsiques (com menys formal, més iberoromànic); l'element gal·loromànic representa el passat i les solucions en retrocés mentre que l'iberoromànic representa el futur i l'element productiu del canvi lingüístic.
3. [En relació amb el] català en general, la resposta només es pot donar de forma dinàmica com un procés mil·lenari que ha allunyat el català de les seves arrels gal·loromàniques i l'ha acostat a la iberoromània, sense que s'hagi perdut el nucli profund d'elements gal·loromànics que al llarg dels segles s'han mostrat immutables davant les pressions del castellà: paraules com *parlar*, *sovint* o *voler* no corren perill de substitució en cap de les varietats del català. (Radatz, 2012, p. 216)

Malauradament, el corpus de què disposem no permetrà analitzar l'evolució de les varietats de la llengua catalana en els termes que Radatz reclama. Els resultats que es descriuen a continuació només permetran visualitzar la posició de les principals varietats de la llengua catalana respecte a la resta de varietats romàniques a principis del segle XX a partir de les transcripcions fonètiques d'una cinquantena de mots. Tanmateix, sí que permetran observar per primera vegada des d'una perspectiva quantitativa global la posició d'aquestes varietats en l'entorn romànic immediat. Aquesta serà, probablement, la principal contribució d'aquest treball a la controvèrsia sobre la subagrupació romànica de la llengua catalana.

## 4. RESULTATS

La figura 2 mostra un mapa de la Romània en el qual s'han projectat els dotze clústers principals de les varietats romàniques –que és el nombre màxim d'agrupaments que permet mostrar la interfície *Gabmap*. Com s'ha explicat a l'apartat 2, ens hem valgut d'aquesta interfície per calcular la distància de Levenshtein entre les varietats romàniques a partir dels 36.808 ítems del corpus. Una vegada obtingudes les interdistàncies entre les 1.010 varietats analitzades, s'ha aplicat un algorisme de clusterització –el mètode de Ward– a la matriu de distància fonètica. El resultat d'aquest procediment és el clúster de la figura 3, del qual s'han projectat els dotze agrupaments principals a la figura 2 per facilitar una primera aproximació als resultats.

En primer lloc, la figura 2 demostra que un subcorpus relativament petit –de només 48 conceptes, i sovint menys en la majoria de varietats– basta perquè emergeixin les principals unitats lingüístiques de què s'ha ocupat la romanística; en concret, s'identifiquen clarament els clústers –o grups més o menys homoge-

nis de varietats– corresponents als parlars portuguesos, gallecs, castellans, catalans, occitans, francoprovençals, d'oïl, italians septentrionals, italians centrals, italians meridionals i romanesos, i un darrer grup de varietats que integra, a grans trets, els parlars sards, corsos, friulans, ladins, rètics i els enclavaments lingüístics d'Itàlia. L'anàlisi d'aquest darrer grup de varietats, tan heterogeni, s'abordarà a partir del dendrograma probabilístic de la figura 11.

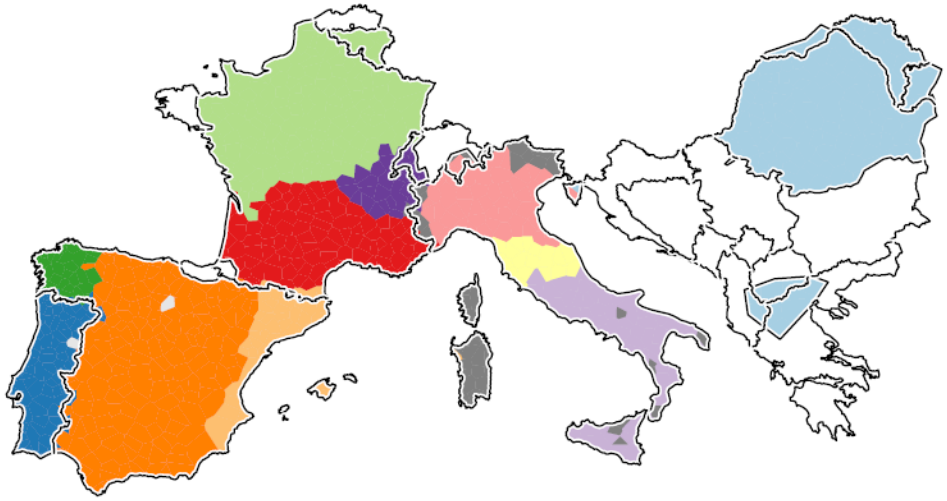


Figura 2. Els 12 clústers principals de les varietats romàniques obtinguts mitjançant el mètode de Ward. Aquesta figura es correspon amb el dendrograma de la figura 3

La projecció d'aquests dotze agrupaments en el mapa permet una primera aproximació a la qüestió de com s'organitza l'espai lingüístic romànic des d'un punt de vista geogràfic, però té l'inconvenient que no reflecteix les relacions internes de les varietats romàniques en termes de distància lingüística. Arran de l'observació d'aquest mapa, per exemple, hom podria pensar que entre les varietats d'oc i d'oïl, d'una banda, i les varietats italianes centrals i meridionals, de l'altra, hi deu haver una distància lingüística similar, atès que en tots dos casos es tracta de varietats contigües; o, per posar-ne un altre exemple, hom podria pensar, atès que la distància geogràfica que les separa és similar, que la distància lingüística entre el romanès i l'italià septentrional és semblant a la distància lingüística entre aquest darrer grup de varietats i el castellà. Com es desprèn de la figura 3, però, aquestes suposicions es troben molt allunyades de la realitat.

La figura 3 mostra el dendrograma jeràrquic aglomeratiu obtingut mitjançant l'aplicació del mètode de Ward a la matriu de distàncies. Aquest dendrograma permet visualitzar, d'una banda, com s'agrupen entre si les diferents varietats

analitzades; i, de l'altra, a quina distància ho fan, ja que la llargada de les línies horitzontals és directament proporcional a la distància cofenètica mitjana que separa entre si dos grups de varietats. Així doncs, com més a la dreta s'ajunten dos clústers, major és la distància que els separa. I, per contra, com més a l'esquerra s'ajunten dos clústers –o, a l'esquerra del tot, dues varietats– menor és la distància que els separa.

La figura 3 reflecteix clarament que les varietats occitanes i francoprovençals i, a una distància un xic major, les catalanes, formen un primer grup de varietats properes. Aquest primer grup s'uneix a un segon grup format per les varietats de l'italià –tant centrals com meridionals i septentrionals– i per les varietats sardes, corses, ladines, friulanes i rètiques. De la figura 3 es desprèn, doncs, que l'àrea mediterrània ha donat lloc al major contínuum de la Romània –si més no, en termes de distància fonètica global entre varietats–; aquest contínuum sembla estendre's del sud d'Alacant al sud de la península itàlica tot resseguint l'arc mediterrani i constitueix –incloses les Balears, les Pitiüses, Còrsega, Sardenya i Sicília– una gran àrea lingüística caracteritzada per la presència de varietats de transició que afavoreixen un trànsit suau entre els diferents grups lingüístics.

L'anàlisi de la resta d'agrupaments de la figura 3 permet constatar tres fets rellevants: en primer lloc, que el conjunt de dialectes que manté més similituds amb les varietats del contínuum mediterrani és el de les llengües d'oïl –tot i que s'hi agrupa a una gran distància, la qual cosa indica la marcada personalitat d'aquestes darreres varietats. En segon lloc, es constata que les relacions entre els parlars gallecs i els parlars castellans són semblants, en termes de distància lingüística, a les que mantenen entre si els parlars francoprovençals i els occitans o els italians septentrionals i la resta de parlars itàlics. I, en tercer lloc, s'observa que les àrees laterals de l'àmbit lingüístic romànic es troben a una gran distància de la resta de varietats estudiades: així doncs, tot i que el portuguès s'agrupa amb les varietats gallegues i castellanes en un conjunt de varietats iberoromàniques que no inclouria el català, ho fa a una distància notable. Per la seva banda, finalment, les varietats romaneses no apareixen emparentades amb cap altre grup de varietats, la qual cosa demostra un cop més l'especificitat d'aquest conjunt de parlars en el si de la Romània.

Cal tenir en compte, però, que l'anàlisi de conglomerats, tot i ser un sistema de classificació molt utilitzat en els estudis de tipus dialectomètric, és relativament inestable a causa del seu caràcter binari. Aquesta inestabilitat es concreta en el fet que, quan hi ha diversos parells d'elements que es troben a una distància similar a la matriu, petites diferències en els inputs poden donar lloc a classificacions dendrogràfiques diferents (Prokić & Nerbonne, 2010). Per superar aquesta inestabilitat, en aquest treball s'ha recorregut als sistemes d'anàlisi multidimensional, que presenten diversos avantatges respecte als sistemes de cluste-

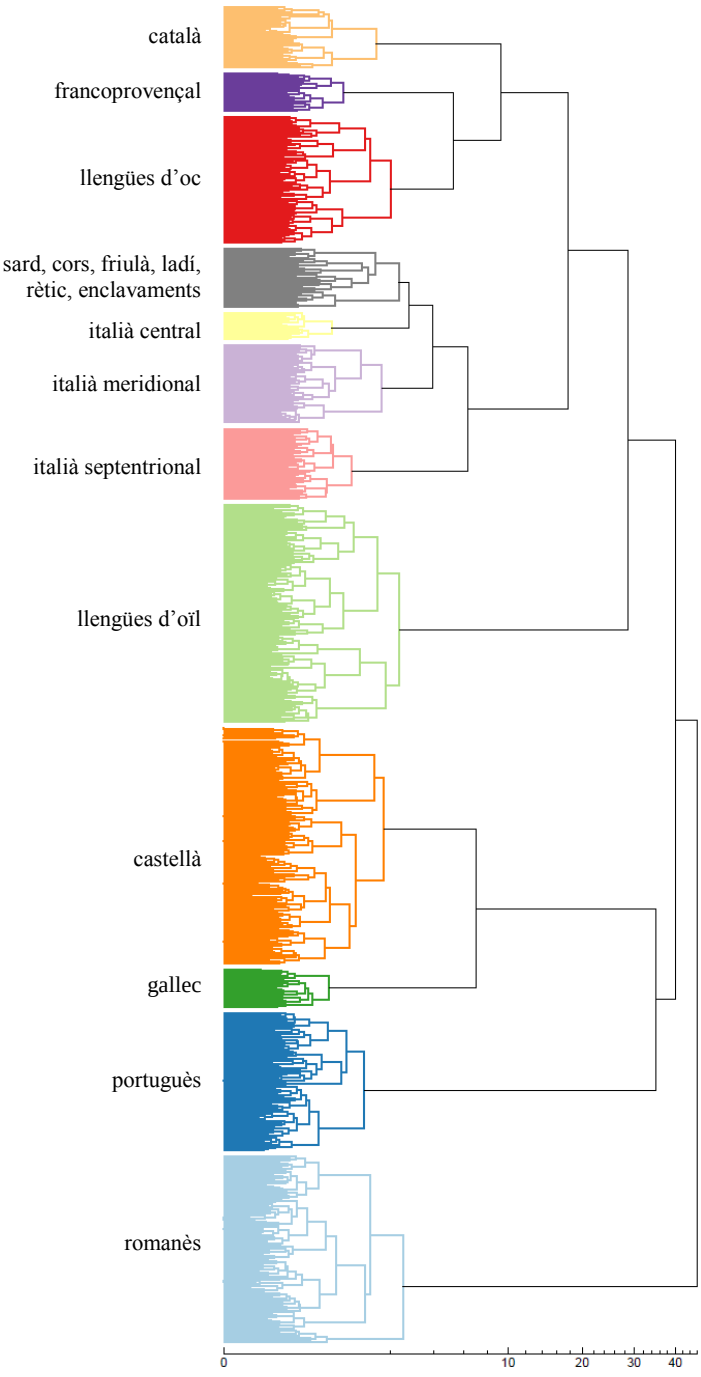



Figura 3. Dendrograma amb els 12 clústers principals de les varietats romàniques obtingut mitjançant el mètode de Ward. La figura 2 mostra la projecció d'aquests clústers en el mapa

rització tradicionals: en primer lloc, són estadísticament estables; i, en segon lloc, no agrupen les varietats entre si segons un procediment binari i jeràrquic, sinó que simplement reflecteixen en el pla les distàncies que mantenen entre elles totes les varietats examinades; això afavoreix una interpretació de la realitat lingüística en termes de contínuum, més que no pas en termes de grups excloents de varietats.

La figura 4 mostra els resultats de realitzar una anàlisi multidimensional a la matriu de distància obtinguda mitjançant l'aplicació de l'algorisme de Levenshtein. Els dotze clústers principals que emergien a la figura 3 s'han acolorit amb els mateixos colors en el gràfic per comprovar que realment es tracta de grups lingüístics independents i no de clústers emergits pel caràcter binari de l'anàlisi de conglomerats. Un primer cop d'ull a la figura 4 només permet arribar amb seguretat a una doble conclusió: s'observa, en primer lloc, que les varietats dacoromaneses, aromaneses, meglenoromaneses i istroromaneses formen un clúster que se situa a gran distància de la resta de varietats romàniques; en segon lloc, s'observa que la resta de varietats dibuixen un contínuum que s'estén de les varietats iberoromàniques, a la part inferior del pla, a les varietats gal·loromàniques, a la part superior del pla. El català se situa en una posició de transició entre aquests dos conjunts de varietats i les varietats italoromàniques. A partir de la informació que proporciona la figura 4 es fa impossible, en canvi, determinar els onze clústers de varietats romàniques –exclòs el romanès– que identificàvem a la figura 3.

Per poder analitzar la composició d'aquest gran contínuum romànic que s'estén per la meitat esquerra del pla de la figura 4 hem d'augmentar el zoom

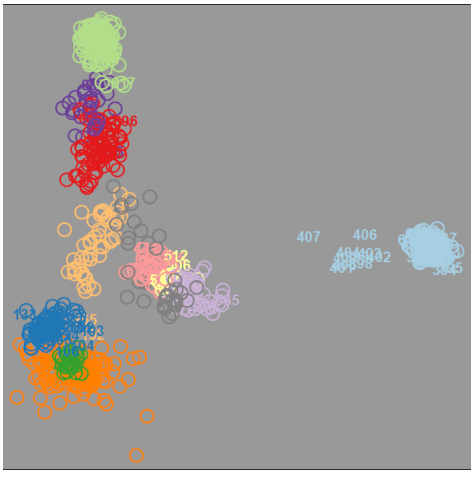



Figura 4. Resultats de l'anàlisi multidimensional amb els 12 clústers principals de les varietats romàniques. *R* = 0,70

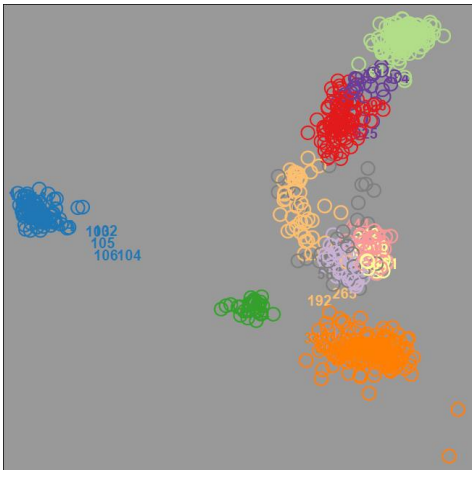



Figura 5. Resultats de l'anàlisi multidimensional de les varietats romàniques sense el romanès. *R* = 0,74

sobre la zona d'interès, i això és possible realitzant una nova anàlisi multidimensional que exclogui les varietats lingüísticament més distants, és a dir, les romaneses. Els resultats d'aquesta nova anàlisi –vg. la figura 5– demostren un fet que ja s'havia detectat en el dendrograma de la figura 3: que el portuguès, en tant que àrea lateral de l'àmbit lingüístic, se situa a una distància força important de la resta de varietats romàniques, incloses les que més se li acosten, les de la llengua gallega.

En aquest punt hom podria demanar-se quina és realment la posició del gallec respecte a les varietats portugueses i castellanes. Cal tenir en compte que la qüestió de la identitat del gallec ha estat una de les més debatudes pels romanistes, i que fins l'any 1988 els principals manuals sobre romanística solien considerar aquesta llengua com una varietat dialectal del portuguès o de l'espanyol; de fet, i per sorprenent que pugui semblar, no va ser fins llavors que alguns experts van optar per considerar-la una llengua independent. Per determinar el grau d'inviduació del gallec respecte al portuguès i al castellà s'ha elaborat, doncs, la figura 6, que mostra els resultats de l'anàlisi multidimensional en les varietats d'aquestes tres llengües. La conclusió és taxativa: les varietats del gallec formen un clúster tan homogeni i independent, com a mínim, com els clústers corresponents a les varietats portugueses i castellanes[8].

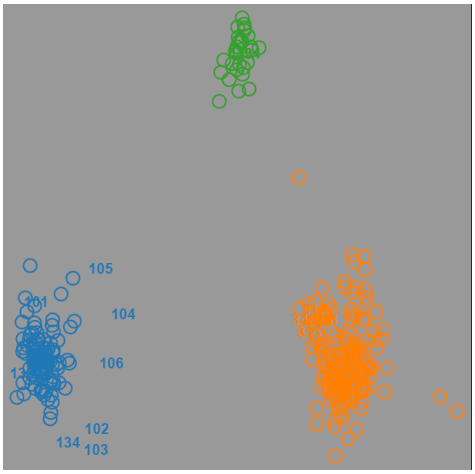



Figura 6. Resultats de l'anàlisi multidimensional de les varietats castellanes, gallegues i portugueses. $R = 0{,}95$

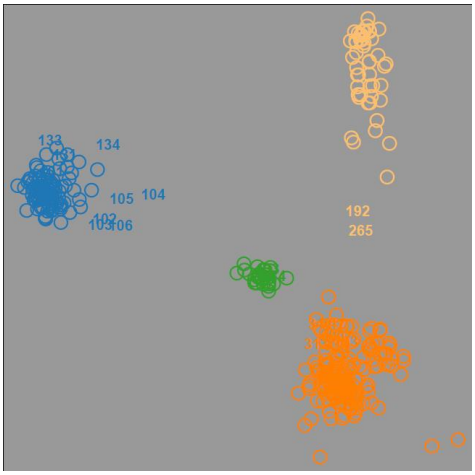



Figura 7. Resultats de l'anàlisi multidimensional de les varietats de la Península Ibèrica. $R = 0{,}94$

[8] A causa de la migradesa de dades disponibles sobre els parlars asturlleonesos i aragonesos en el corpus, aquestes varietats no emergeixen del clúster de parlars espanyols. La nostra hipòtesi és, però, que amb una quantitat major de dades aquests dos conjunts de varietats apareixerien com dos clústers identificables a una certa distància del de la llengua castellana.

La independència lingüística del gallec en termes de distància fonètica és, doncs, un fet empíricament contrastat. Ara bé, tot i que aquesta individualitat no es posa en qüestió, si s'incorpora una quarta varietat a l'anàlisi –la catalana, per ser també de l'àmbit ibèric–, es veu que realment la relació entre el gallec i el castellà i entre el gallec i el portuguès no és pas d'equidistància, com suggeria la figura 6. La figura 7, que plasma els resultats de l'anàlisi multidimensional en les varietats de la Península Ibèrica, mostra que la posició del gallec respecte al castellà és de major proximitat que no respecte al portuguès[9].

Les figures 8 i 9 mostren els resultats de sengles anàlisis multidimensionals dutes a terme per classificar les varietats gal·les i itàliques –incloses les de Còrsega, Sardenya i Sicília–, respectivament. De la figura 8 es desprèn clarament que les varietats d'oïl, d'oc i francoprovençals constitueixen tres grups lingüístics clarament independents; el mateix succeeix –vg. la figura 9– entre les varietats itàliques septentrionals, centrals i meridionals.

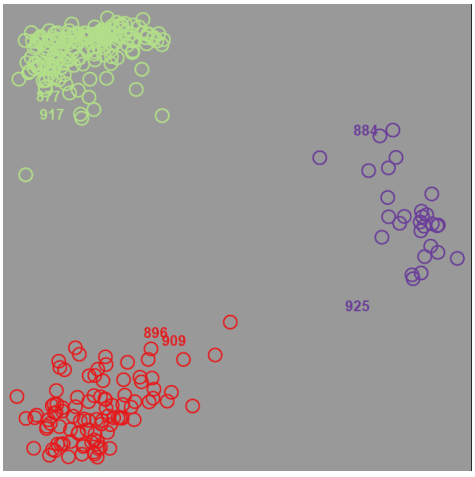



Figura 8. Resultats de l'anàlisi multidimensional de les varietats gal·les. $R$ = 0,71

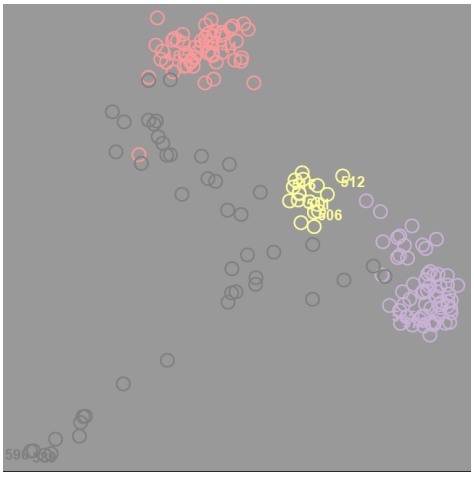



Figura 9. Resultats de l'anàlisi multidimensional de les varietats itàliques. $R$ = 0,71

La figura 10, finalment, permet abordar la qüestió que ha motivat aquest article: la subagrupació romànica del català. Per analitzar aquest punt s'ha realitzat una anàlisi multidimensional que inclogués únicament les varietats que la

[9] El fet que les varietats patrimonials de la llengua gallega sovint presentin un nombre elevat d'interferències lèxiques provinents del castellà s'ha de tenir en compte a l'hora d'interpretar aquests resultats. De fet, la depuració de les interferències que tot procés d'estandardització sol dur aparellat podria contrarestar o, si més no, mitigar, en els anys a venir, la tendència a l'advergència amb la llengua castellana que ha caracteritzat l'evolució del gallec en els darrers segles.

romanística ha considerat clarament iberoromàniques –les portugueses, gallegues i castellanes– i les que ha considerat clarament gal·loromàniques –les d'oïl, occitanes i francoprovençals. Els resultats mostren que, si més no durant les primeres dècades del segle XX, quan es van dur a terme les enquestes a partir de les quals s'ha confegit aquest subcorpus de dades fonètiques, l'agrupació de la llengua catalana amb algun d'aquests dos grups de varietats ja no era viable: la posició del català és clarament de transició entre les varietats gal·loromàniques meridionals –llenguadocianes i gascones– i les varietats iberoromàniques septentrionals.

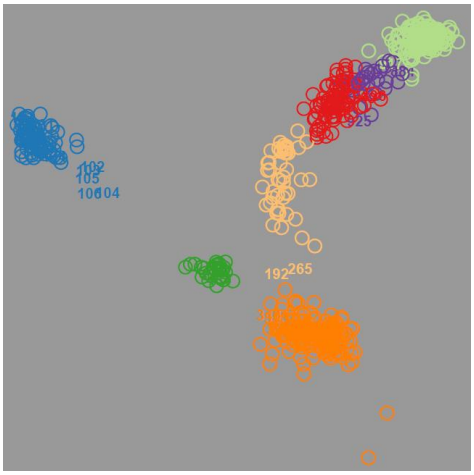



Figura 10. Resultats de l'anàlisi multidimensional de les varietats iberoromàniques i gal·loromàniques. *R* = 0,86

En aquest punt hem optat per utilitzar un segon sistema de clusterització que, com les anàlisis multidimensionals, permet aproximar-se a les relacions jeràrquiques entre grups de varietats sense les limitacions dels sistemes jeràrquics aglomeratius tradicionals. Es tracta del *noisy clustering* –vg. Nerbonne, Kleiweg, Manni & Heeringa (2008). Aquest sistema consisteix a afegir petites quantitats de soroll aleatori –*random noise*– a la matriu de distàncies en successives clusteritzacions i dóna com a resultat un *dendrograma probabilístic* o *de consens* –en anglès, *consensus* o *probabilistic dendrogram*. Aquest dendrograma proporciona dues informacions rellevants: en primer lloc, 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; en segon lloc, la llargada horitzontal de les línies reflecteix la distància cofenètica mitjana a què un conjunt de varietats s'han agrupat entre si en el total de clusteritzacions: com més llarga és una línia menys homogènies són les varietats que agrupa,

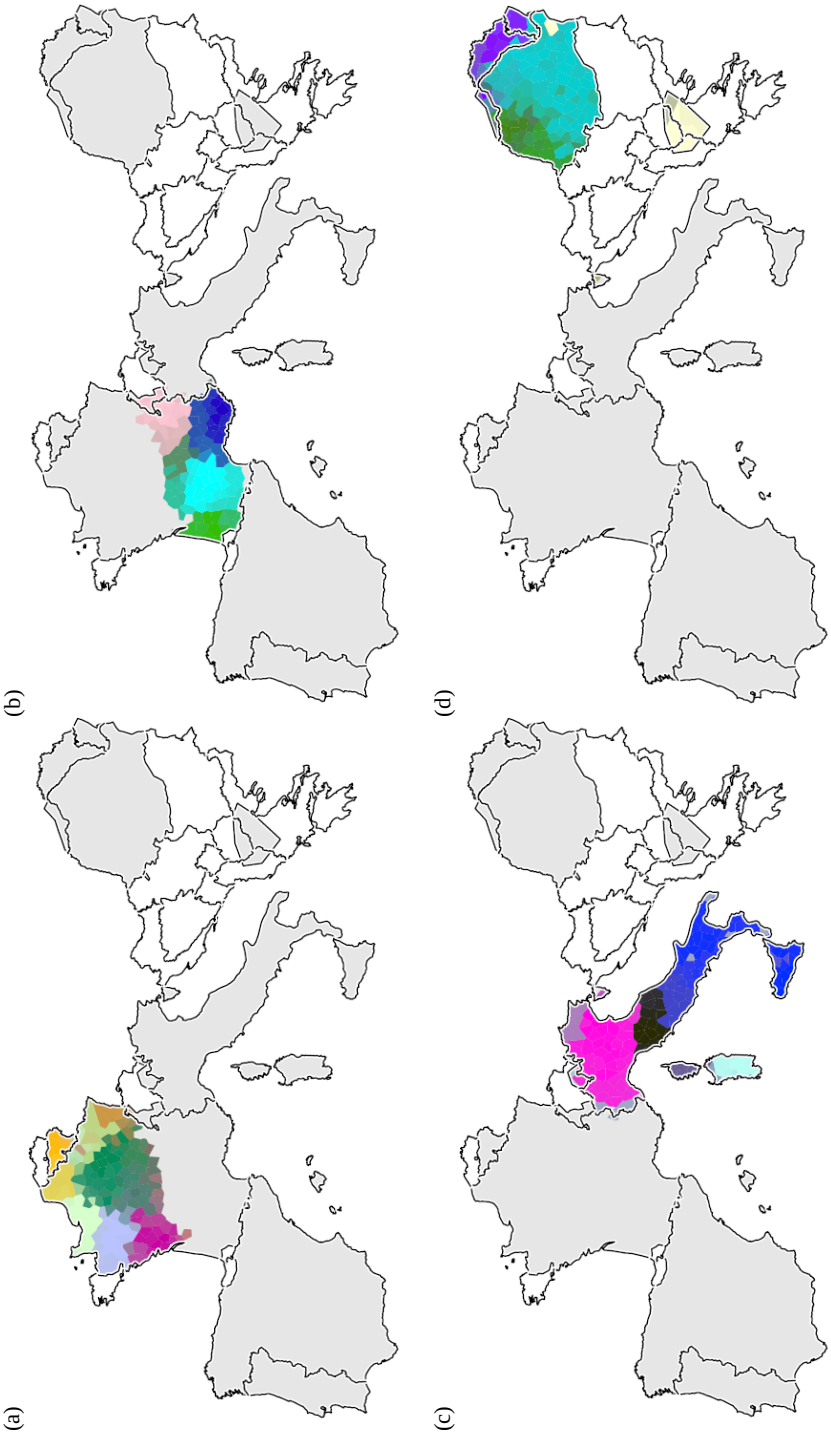
per molt que aquestes siguin més semblants entre elles que respecte a cap altra varietat del corpus.

Un primer cop d'ull al dendrograma probabilístic de la figura 11[10] ja permet arribar a una conclusió important: que les relacions entre varietats romàniques s'han d'entendre més en termes de contínuum que no pas en termes de subagrupacions; això és així perquè, tot i que aquest sistema de clusterització detecta molt bé les principals unitats lingüístiques romàniques, aquestes unitats no s'agrupen entre si de manera estable –això és, en més del 80% de clusteritzacions– en gairebé cap cas.

Novament, aquest sistema de classificació fa emergir els clústers de les varietats catalanes, castellanes, gallegues, d'oïl, francoprovençals, occitanes, italianes septentrionals, italianes centrals, italianes meridionals, romaneses i portugueses. A més, però, la major sofisticació del mètode permet que emergeixin altres grups homogenis de varietats que fins ara no s'havien detectat de manera nítida: en primer lloc, emergeix un grup de varietats corses, que formen un únic clúster amb una certa variació interna; en segon lloc, emergeix un grup de varietats sardes que s'agrupen entre si en el 100% de clusteritzacions, tot i que també presenten una certa variació interna –que s'identifica, com hem dit, per la llargada de les línies horitzontals del dendrograma. En tercer lloc, emergeix un grup de varietats ladines i friulanes –que al seu torn s'agrupen amb les varietats italianes septentrionals en un segon nivell. Arran d'aquest fet es pot concloure que aquestes varietats –antigues integrants d'un contínuum lingüístic avui fragmentat– continuen mantenint una relació de gran proximitat. Aquest no és el cas, en canvi, de les varietats retoromàniques, les terceres integrants d'aquest contínuum històric, que apareixen no només desvinculades dels parlars ladins i friulans sinó que –i això es podria deure a la influència de les varietats germàniques de Suïssa– no s'agrupen amb cap altre conjunt de varietats analitzades. La figura 11 permet identificar, finalment, un clúster prou homogeni format per les varietats aromaneses i, en un segon nivell, meglenoromaneses. Aquestes varietats, com l'istroromanès, a més distància, s'agrupen regularment amb les varietats dacoromaneses.

Un darrer aspecte interessant que es desprèn de l'observació de la figura 11 és la posició que ocupen les varietats gallegues: en el 96% dels casos apareixen agrupades entre si, com un grup de varietats molt homogènies en el context romànic, però en el marc del clúster de varietats de la llengua castellana. Això vol dir que, de totes les varietats romàniques que avui reben unànimement la consideració de llengües independents, les varietats de la llengua gallega són, pel

[10] Per manca d'espai no hem pogut incloure aquesta figura en l'article, però es pot visualitzar i descarregar a <https://ub.academia.edu/EsteveValls/Complementary-figures> [consulta: 27.01. 2016].




Figures 12a, b, c i d. Mapes probabilístics de les varietats d'oïl, d'oc i francoprovençals, italoromàniques i romaneses, respectivament

que sembla, les que presenten una distància lingüística més petita respecte a les varietats d'una altra llengua, en aquest cas la castellana. En aquest sentit, la relació entre el gallec i el castellà és, en termes de distància fonètica, similar a la que mantenen entre si varietats com el gascó i el llenguadocià, sobre les quals no hi ha un consens absolut a l'hora de considerar-les o no llengües autònomes. Cal tenir en compte, però, que aquests resultats depenen en bona mesura de les característiques *fonètiques* de les llengües analitzades, i que segurament variarien si es pogués dur a terme una anàlisi lingüística de tipus morfofonològic prèvia al càlcul de distàncies. Aquesta serà part de la tasca que es durà a terme al CRPIH en els propers mesos.

Els sistemes de clusterització estables, com el *noisy clustering*, també tenen l'avantatge que permeten projectar subsets de dades en els anomenats *mapes probabilístics* –o *fuzzy cluster maps*–, gràcies als quals és possible analitzar a partir d'un mapa la composició interna dels principals clústers detectats en un dendrograma de consens. Tal com s'explica al manual d'ús de *Gabmap*, un mapa probabilístic

> visualizes something between multidimensional scaling (MDS) and cluster analysis: main dialect groups are identified in the map, but continuous relationships are displayed for places which cannot be put in one group with high probability. The map is created by running MDS on the branch lengths of the dendrogram (so-called cophenetic distances) instead of on the original linguistic distances[11].

Les figures 12a, 12b, 12c i 12d són mapes probabilístics corresponents als següents subsets de dades: les varietats d'oïl, les d'oc i francoprovençals, les italoromàniques i les romaneses. Gràcies a aquests mapes es poden identificar les principals varietats que componen cadascun d'aquests clústers principals. A la figura 12a, per exemple, s'identifiquen les següents varietats d'oïl –de sud-oest a nord-est–: el poiteví-saintongès, el grup format pel gal·ló i l'angeví, el normand, el picard, el való, el lorenès, el franc-comtès i un gran *plateau* de parlars centrals que comprendria el borgonyó-morvandià, el xampanyès, el berrichon i el francès, tots ells pràcticament assimilats a aquesta darrera varietat. Pel que fa a la figura 12b, corresponent a les varietats d'oc i francoprovençals, s'hi identifiquen clarament els parlars gascons, llenguadocians, llemosins, alvernesos, vívaro-alpins i provençals, d'una banda, i un grup de parlars francoprovençals força homogenis, de l'altra.

La figura 12c, per la seva banda, identifica clarament les varietats italianes septentrionals, centrals i meridionals, com la majoria de tècniques que hem utilitzat fins ara. També identifica, però, molt clarament els parlars ladins i friulans,

---

[11] Aquest passatge s'ha extret del manual en línia de *Gabmap*, d'accés obert a l'adreça <http://www.gabmap.nl/~app/doc/manual/clustering-fuzzy.html> [consulta: 27.01.2016].

l'occità parlat a les *valadas* i el francoprovençal de la Vall d'Aosta, els parlars sards i, encara, els parlars corsos, que tenen continuïtat en el sassarès i el gal·lurès del nord de Sardenya. La figura 12d, finalment, assenyala quatre grans varietats dialectals de la llengua romanesa: d'una banda, els parlars aromanesos i meglenoromanesos, als Balcans; de l'altra, les tres varietats principals de dacoromanès: el valac, al centre i sud de l'actual Romania; el moldau, al nord-est; i els parlars de Banat, al nord-oest. Enmig d'aquestes varietats, els tons verdosos i blavosos, més clars, assenyalen l'àrea de transició formada pels dialectes transsilvànics.

## CONCLUSIONS

En aquest treball s'han presentat els resultats d'analitzar la variació i la distància lingüística entre les varietats de la *Romania Antiqua* des d'una òptica dialectomètrica. La mesura de distància fonètica utilitzada ha estat la distància de Levenshtein, que s'ha aplicat a un corpus de 36.808 ítems repartits en 1.010 localitats de l'àmbit romànic. Aquest corpus s'ha dissenyat al Centro Ramón Piñeiro para a Investigación en Humanidades i actualment es troba en procés d'ampliació. Tant el càlcul de la matriu d'interdistàncies com els diferents sistemes taxonòmics i de visualització emprats s'han aplicat mitjançant la interfície *Gabmap*.

La primera conclusió a què es pot arribar és que un corpus de dades relativament petit –de només 48 conceptes, i sovint menys en la majoria de varietats– basta perquè emergeixin les principals unitats lingüístiques de què s'ha ocupat la romanística. L'anàlisi quantitativa d'aquest corpus dibuixa un paisatge geolectal de l'àmbit romànic que es pot resumir en els següents punts:

a. En primer lloc, sembla –segons es desprèn de l'anàlisi jeràrquica aglomerativa– que l'àrea mediterrània ha donat lloc al major contínuum de la Romània; aquest contínuum s'estén del sud d'Alacant al sud de la península itàlica tot resseguint l'arc mediterrani i constitueix –incloses les Balears, les Pitiüses, Còrsega, Sardenya i Sicília– una gran àrea lingüística caracteritzada per la presència de varietats de transició que afavoreixen un trànsit suau entre els diferents grups lingüístics.
b. En segon lloc, s'observa –a partir de les anàlisis multidimensionals– que les varietats dacoromaneses, aromaneses, meglenoromaneses i istroromaneses formen un clúster que se situa a una gran distància de la resta de varietats romàniques. En contrast amb les varietats del romanès, les varietats iberoromàniques, italoromàniques i gal·loromàniques formen un contínuum de parlars separats per una distància fonètica molt més reduïda.

c. En tercer lloc, s'observa que el portuguès, en tant que àrea lateral de l'àmbit lingüístic, se situa a una distància notable respecte a la resta de varietats romàniques. Aquesta distància és important fins i tot respecte a aquelles varietats que més se li acosten, les de la llengua gallega.
d. En quart lloc, l'anàlisi del dendrograma probabilístic ha permès constatar que les relacions entre varietats romàniques s'han d'entendre més en termes de contínuum que no pas en termes de subagrupacions tancades. Aquest dendrograma també ha permès copsar que, a hores d'ara –s'entengui: en l'estadi sincrònic estudiat–, les varietats ladines i friülanes, antigues integrants d'un contínuum lingüístic avui fragmentat, continuen mantenint una relació de gran proximitat. Aquest fet contrasta amb les varietats retoromàniques –les terceres integrants d'aquest contínuum històric–, que apareixen no només desvinculades dels parlars ladins i friulans sinó que –i això es podria deure a la influència de les varietats germàniques de Suïssa– no s'agrupen amb cap altre conjunt de varietats analitzades.
e. En cinquè lloc, s'observa que les varietats del gallec formen un clúster tan homogeni i independent, com a mínim, com els clústers corresponents a les varietats portugueses i castellanes. Tot i amb això, la relació entre el gallec i el castellà i entre el gallec i el portuguès no és pas d'equidistància, ja que en les anàlisis multidimensionals s'observa una major proximitat d'aquesta llengua amb les varietats castellanes i en el dendrograma probabilístic fins i tot s'hi agrupa en el 96% de les clusteritzacions. Això sembla indicar que el procés de distanciament del gallec respecte al portuguès que ha tingut lloc al llarg dels darrers segles ha comportat, paral·lelament, un procés d'acostament d'aquesta llengua al castellà. Arran d'aquesta evolució, de totes les varietats romàniques que avui reben per consens la consideració de llengües independents, les de la llengua gallega són, pel que sembla, les que presenten una distància lingüística més petita respecte a les varietats d'una altra llengua –en aquest cas la castellana.
f. Pel que fa a la posició del català, les varietats d'aquesta llengua s'ubiquen clarament entre les varietats gal·loromàniques meridionals –llenguadocianes i gascones– i les varietats italoromàniques i iberoromàniques septentrionals. Així doncs, aquest treball permet constatar per primer cop, a partir d'una anàlisi quantitativa, que les varietats de la llengua catalana ocupen –o si més no ocupaven durant les pimeres dècades del segle XX– una posició intermèdia –de transició– entre els àmbits gal·loromànic, italoromànic i iberoromànic –amb cap dels quals, tanmateix, no és possible d'agrupar-les. En aquest sentit, doncs, sembla que els resultats de l'anàlisi dialectomètrica permetrien parlar, com han fet diversos autors a partir d'anàlisis estrictament qualitatives, del català com una "llengua pont" en-

tre els tres principals àmbits lingüístics peninsulars de la *Romania Antiqua*. Tot i amb això, considerem que estudis com el de Radatz (2012), que com hem vist qüestionen aquesta denominació, són avui dia més necessaris que mai per aprofundir en el coneixement de la posició *actual* de les varietats de la llengua catalana en el ja desdibuixat contínuum romànic. Si més no des d'una òptica sociolingüística, identificar els trets gal·loromànics de la llengua que han patit una erosió més forta a causa del contacte amb el castellà per revertir-ne el desgast –o com a mínim per conscienciar els parlants que *també* es poden utilitzar– podria ser una bona manera de contribuir a reforçar el capital lingüístic dels catalanoparlants: d'una banda, es rebaixaria la sensació que el català pràcticament ha esdevingut una llengua calcada del castellà; i, de l'altra, s'afavoriria un acostament més natural a l'occità –recordem-ho, una de les tres llengües cooficials a Catalunya– i al francès –la qual cosa donaria accés, de retruc, a la francofonia.

## AGRAÏMENTS

Aquest treball s'inscriu en el projecte "Descripción e interpretación de la variación dialectal: aspectos fonológicos y morfológicos del catalán" (FFI2010-22181-C03-02), finançat pel MICINN i el FEDER, i ha estat possible gràcies a un ajut del Centro Ramón Piñeiro para a Investigación en Humanidades de la Secretaría Xeral de Política Lingüística de la Xunta de Galícia.

## BIBLIOGRAFIA


Colón, G. (1976). *El léxico catalán en la Romania*. Madrid: Gredos.

Contini, M. (1992). L'Atlas Linguistique Roman. *IKER, 7*, 339-356.

Diez, F. (1836). *Grammatik der romanischen sprachen*. Bonn: E. Weber.

Duarte, C. & Alsina À. (1984). *Gramàtica històrica del català*. Barcelona: Curial.

Goebl, H. (1991). Dialectometry: A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data. Dins R. Köhler & B. Rieger (Eds.), *Contributions to quantitative linguistics. Proceedings of the First International Conference on Quantitative Linguistics* (pp. 277-315). Dordrecht: Kluwer.

— (2003). Regards dialectométriques sur les données de l'*Atlas Linguistique de la France* (ALF) : relations quantitatives et structures de profondeur. *Estudis Romànics, XXV*, 59-121.

— (2006). Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing, 21*, 411-435.

— (2008). La dialettometrizzazione integrale dell'AIS. Presentazione dei primi risultati. *Revue de linguistique romane, 72* (285-286), 25-73.

Gröber, G. (Ed.) (1888). *Grundiss der romanischen philologie*. Estrasburg: K. J. Trübner.

Meyer-Lübke, W. (1925). *Das Katalanische. Seine Stellung zum Spanischen und Provenzalischen, sprachwissenchftlich und historisch dargestellt*. Heidelberg: Winter.

Nerbonne, J., Kleiweg, P., Manni, F. & Heeringa, W. (2008). Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering. Dins Ch. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society* (pp. 647-654). Berlin: Springer.

Nerbonne, J., Colen, R., Gooskens, Ch., Kleiweg, P. & Leinonen, Th. (2011). Gabmap: A Web Application for Dialectology. *Dialectologia*, Special issue, II.

Prokić, J. & Nerbonne, J. (2010). Recognizing Groups among dialects. Dins J. Nerbonne, Ch. Gooskens, S. Kürschner & R. van Bezooijen (Eds.), *International Journal of Humanities and Arts Computing. Special issue on Language Variation* (pp. 153-172). Edinburgh: Edinburgh University Press.

Radatz, H. (2012). Per què els elements gal·loromànics fan 'heavy' en català. Arran del clàssic debat sobre la subagrupació del català. *eHumanista / IVITRA, 2*, 202-218.

Tuaillon, G. & Contini, M. (Eds.) (1996). *Atlas Linguistique Roman (ALiR). Volume I. Présentation; Volume I. Cartes; Volume I. Commentaires*. Roma: Istituto Poligrafico e Zecca Dello Stato.

Valls, E. (2013). *Direccionalitat, ritme, abast i naturalesa del canvi lingüístic en català nord-occidental. De l'anàlisi dialectomètrica a la reflexió sociolingüística* (tesi doctoral inèdita). Barcelona: Universitat de Barcelona.

Veny, J. & Contini, M. (Eds.) (2001). *Atlas Linguistique Roman (ALiR). Volume IIa. Cartes; Volume IIa. Commentaires*. Roma: Istituto Poligrafico e Zecca Dello Stato.

— (Eds.) (2009). *Atlas Linguistique Roman (ALiR). Volume IIb. Cartes; Volume IIb. Commentaires.* Roma: Istituto Poligrafico e Zecca Dello Stato.