

ANÁLISIS LINGÜÍSTICO Y DIALECTOMÉTRICO DEL CORPUS ORAL DIALECTAL (COD)¹

Esteve Clua
Universitat Pompeu Fabra
Maria-Rosa Lloret
Esteve Valls
Universitat de Barcelona

Resumen

En este trabajo presentamos los resultados del análisis dialectométrico que hemos aplicado a los datos del *Corpus Oral Dialectal* del catalán contemporáneo. Los datos morfológicos y fónicos de las diferentes variedades del catalán que contiene el corpus han sido analizados lingüísticamente en el marco de la fonología generativa clásica y posteriormente comparados entre sí hasta hallar la distancia lingüística entre dichas variedades y la correspondiente agrupación dialectal. En el análisis cuantitativo, además de un índice de distancia para comparar los datos, hemos utilizado el *Cluster Analysis* y para la representación gráfica hemos usado un algoritmo de clasificación basado en el método UPGMA. Los resultados obtenidos concuerdan en gran medida con las clasificaciones geolingüísticas tradicionales, pero permiten a su vez aclarar ciertas estructuras de interdistancias que hasta ahora eran discutibles, como ocurre por ejemplo con las variedades valencianas septentrionales.

1. Introducción

El *Corpus Oral Dialectal* (COD) del catalán contemporáneo, compilado a partir de 1991 por el Departamento de Filología Catalana de la Universitat de Barcelona, consta de materiales fonéticos, morfológicos y léxicos de las principales variedades del catalán. Los datos fueron obtenidos a través de un cuestionario de 600 ítems y de un conjunto de textos espontáneos, procedentes de las encuestas realizadas en cada una de las 82 capitales de comarca del ámbito lingüístico del catalán. En el momento en que se llevaron a cabo las encuestas, en 1993, los informantes tenían entre 30 y 45 años, que era la edad límite para que no hubieran sido escolarizados en catalán; además, habían de tener un nivel cultural medio. Los criterios utilizados para la selección de localidades e informantes se deben a la voluntad de compilar un corpus dialectal más urbano que los tradicionales, que mostrase de una manera más adecuada cuáles eran las características del catalán contemporáneo.²

El objetivo de esta comunicación es presentar los resultados obtenidos en la clasificación de las variedades dialectales del catalán mediante la aplicación de técnicas de análisis multivariante a los datos del COD, previamente analizados en el marco de la fonología generativa clásica. La descripción estadística realizada se basa en el análisis de conglomerados (*Cluster Análisis*), que nos ofrece una representación dendrográfica y jerárquica de las variedades dialectales. Anteriormente habíamos aplicado estas técnicas a algunas de las variedades catalanas (Clua & Lloret 2006), o a algún subconjunto de los datos del corpus (Clua 2007) pero esta es la primera vez que presentamos un estudio completo de la distancia lingüística de todas las variedades del catalán.

1 Este trabajo forma parte del proyecto ECOD (Explotación de un corpus oral dialectal: Análisis de la variación lingüística y desarrollo de aplicaciones informáticas para la transcripción automatizada, 2ª fase), financiado por el Ministerio de Ciencia e Innovación HUM2007-65531/FILO.

2 *Vid.* la versión en CD del COD: Viaplana *et alii* (2007).

Este análisis nos permitirá, además de comparar los resultados obtenidos con las clasificaciones tradicionales, aclarar algunos aspectos de la distribución dialectal que hasta ahora habían sido bastante controvertidos.

2. Análisis lingüístico de los datos

Una de las deficiencias que presentan algunos tratamientos cuantitativos de la variación dialectal tiene que ver con la falta de un análisis lingüístico coherente previo a la transposición de los datos fonéticos, obtenidos en las entrevistas, a las variables de comparación que sirven de base para el análisis cuantitativo. A causa de la gran variación observable en cualquier estudio dialectal, a menudo se ha tendido a una cierta tipificación de los resultados, es decir, a una simplificación de dicha variación; si este proceso no se sustenta en criterios lingüísticos coherentes puede pasar que se utilicen criterios muy dispares (sincrónicos y diacrónicos, interdialectales e intradialectales, etc.) o que las variables de las que parte el proceso clasificatorio no reflejen adecuadamente la variación real. De ahí la importancia que tiene en nuestra investigación el análisis lingüístico previo al tratamiento cuantitativo.³

El análisis lingüístico que aplicamos a los datos del COD se lleva a cabo en el marco del modelo generativo, porque permite discriminar las diferencias superficiales o reducibles, que expresan las regularidades de las lenguas (y de las variedades que las componen), de las diferencias profundas o irreducibles, que afectan la estructura léxica o gramatical de las palabras.⁴ Desde nuestro punto de vista, la distinción entre estos dos niveles de análisis es fundamental para determinar la distancia lingüística entre variedades.⁵

La metodología que utilizamos para el análisis es la propia de la gramática generativa: *a)* analizar un conjunto de datos para extraer generalizaciones, *b)* comprobar la viabilidad de las generalizaciones propuestas aplicándolas a otros conjuntos de datos y *c)* modificar, en el caso que sea preciso, las hipótesis de partida. En este sentido, disponer de un corpus informatizado de datos propio, homogéneo y representativo de la variación existente en una misma área lingüística supone tener acceso a un material empírico óptimo para constatar la validez y la adecuación de distintos marcos teóricos para explicar la variación lingüística y para aplicar los resultados obtenidos con estos análisis a la delimitación de la distancia lingüística.

Para el análisis de fenómenos fonológicos específicos, hemos seguido mayoritariamente la orientación propia de la fonología generativa derivacional, aunque en los últimos años hemos empezado a trabajar con la evaluación simultánea propuesta en el marco de la teoría de la optimalidad (*Optimality Theory*), en su versión de la teoría de la correspondencia (*Correspondence Theory*), para poder superar algunos de las deficiencias detectadas en el enfoque derivacional tradicional. En cuanto a los aspectos morfológicos, seguimos básicamente el enfoque morfémico clásico (*Item and Arrangement*), aunque también en esta área hemos empezado a avanzar en el campo de la morfología paradigmática (*Word and Paradigm*) para poder explicar los efectos analógicos y de contraste derivados de las

3 Véase en este sentido Clua & Lloret (2006 y 2007).

4 Vid. Lloret & Viaplana (1998).

5 Vid. Clua (1999a, b) y Viaplana (1999).

relaciones intraparamétricas e interparamétricas que se establecen entre palabras morfológicamente relacionadas.⁶

Podemos ver, a continuación, un caso de diferencias fonéticas (superficiales) y otro de diferencias fonológicas (subyacentes). El catalán presenta distintas terminaciones flexivas para la 1ª persona del singular del presente de subjuntivo: unas variedades presentan –[e] (en la mayor parte del área valenciana), otras –[i] (variedades del Catalán oriental) y otras –[a] (en algunas variedades del catalán norte-occidental). Ninguna de estas diferencias, sin embargo, puede ser atribuida a un fenómeno sistemático de la fonología del catalán; es decir, en estas variedades no existe ningún proceso regular por el cual –/e/ se convierta en –[i] o en –[a], o viceversa. Estas diferencias, por tanto, han de ser atribuidas directamente a la estructura profunda de las palabras. Podemos verlo en los ejemplos de (1), correspondientes al verbo *cantar*.

- (1)
- | | | | |
|----|-------------|-------------|-----|
| a. | Variedad 1: | [kánte]: | /e/ |
| b. | Variedad 2: | [kánti]:/i/ | |
| c. | Variedad 3: | [kánta]: | /a/ |

Otro ejemplo de variación que afecta a las terminaciones verbales lo encontramos en las formas del gerundio. Muchas variedades, como ocurre en el catalán de Barcelona, presentan una –[n] final en esta forma verbal; en cambio muchas de las variedades valencianas acaban el gerundio con –[nt]. Desde el punto de vista tradicional, estas diferencias son iguales que las de (1), ya que el enfoque tradicional sólo tiene en cuenta las realizaciones fonéticas. Desde la perspectiva generativa, sin embargo, la alternancia [n]~[nt] de (2a) puede ser atribuida a una única forma fonológica /nt/, que se realiza como [n] en posición final de palabra y como [nt] cuando le sigue un clítico en el caso de las primeras variedades (2a), ya que en estas variedades opera un proceso de simplificación del grupo consonántico [nt]; en cambio en las variedades de (2b) no se produce ningún tipo de alternancia.

- (2)
- | | | | | | |
|----|-------------|-----------|-----|------------------------------------|------|
| a. | Variedad 1: | [kantán] | vs. | [kantántu] (<i>cantant-ho</i>): | /nt/ |
| b. | Variedad 2: | [kantánt] | vs. | [kantánto]: (<i>cantant-ho</i>): | /nt/ |

En (1), la diferencia era subyacente; en (2), en cambio, la diferencia es meramente fonética. Bajo nuestro punto de vista, esta diferencia también se tiene que reflejar en la cuantificación de la distancia lingüística entre variedades, ya que de lo contrario el resultado del análisis cuantitativo puede apartarse considerablemente de la realidad.⁷

3. Análisis dialectométrico y representación de la distancia lingüística

El análisis cuantitativo que aplicamos al corpus pretende determinar la distancia lingüística existente entre las diferentes variedades dialectales mediante métodos multivariantes, con la

6 Para un análisis derivacional y morféxico de los datos del COD, *vid.*, entre otros, los trabajos citados anteriormente; para un enfoque simultáneo y paradigmático, *vid.*, por ejemplo, Bonet & Lloret (2005) y Lloret (2004)

7 Para una justificación empírica de la necesidad de analizar lingüísticamente los datos antes del tratamiento cuantitativo *vid.* Clua & Lloret (2006).

finalidad de poder establecer agrupaciones clasificatorias entre dichas variedades y descubrir estructuras de interdistancias que a simple vista son difíciles de captar.

El concepto de distancia, que procede del ámbito científico del análisis de datos, se asocia generalmente a la cuantificación de las similitudes o diferencias observadas entre individuos, poblaciones o grupos de poblaciones y permite establecer las clasificaciones sobre la base de un tratamiento sintético y global del conjunto de características (en nuestro caso lingüísticas) analizadas. Con el término análisis multivariante (*Multivariate Analysis*) se hace referencia a diversas técnicas de descripción y análisis de un conjunto de elementos a partir de diversas variables observadas, sin que *a priori* se otorgue a ninguna de ellas un estatus cualitativo de especial relevancia. Esto implica que las delimitaciones dialectales realizadas con esta metodología pueden eludir la siempre difícil jerarquización de características lingüísticas que subyace en toda clasificación dialectal basada en el concepto de isoglosa.

Una vez definida la distancia lingüística es necesario obtener su representación gráfica. Las matrices de similitudes/diferencias, resultantes de la comparación de las características lingüísticas de las diferentes variedades a clasificar, determinan las interdistancias lingüísticas entre estas variedades; pero, cuando se trabaja con cantidades importantes de datos, no permiten una buena interpretación ni una buena visualización de los resultados. Como se puede comprobar en (3), que es un pequeño fragmento de una de las matrices de similitudes con las que hemos trabajado, a partir de un conjunto de datos de este tipo podemos discernir cuáles son las variedades que presentan más coincidencias, pero resulta imposible establecer una clasificación global. Por eso es necesario buscar una representación de la estructura de interdistancias de la matriz que permita visualizar con claridad las relaciones de proximidad o lejanía entre las variedades lingüísticas analizadas. Se trata, pues, de lograr una representación en un espacio de dimensión reducida (como por ejemplo, el plano) que, con un mínimo de distorsión de la estructura de interdistancias original, permita una interpretación adecuada de los datos

	L'Alguer	Ciutadella	Eivissa	Felanitx	Formentera	Manacor	Maó	Palma	Pollença	Sóller	Banyoles	Barcelona	Berga
L'Alguer	0												
Ciutadella	2999	0											
Eivissa	2884	499	0										
Felanitx	3106	773	738	0									
Formentera	2688	815	449	971	0								
Manacor	3022	827	715	450	906	0							
Maó	3107	390	761	1031	1070	1067	0						
Palma	3034	535	471	362	730	415	845	0					
Pollença	3015	740	653	434	917	388	1051	416	0				
Sóller	2936	527	477	595	751	670	812	481	630	0			
Banyoles	3010	1175	1003	1482	1260	1511	1148	1265	1367	1368	0		
Barcelona	2969	1021	897	1407	1116	1419	971	1174	1255	1269	303	0	
Berga	2952	1000	865	1361	1084	1380	944	1129	1210	1230	309	82	0

Figura 1. Fragmento de una matriz de distancias

Con esta finalidad, el ámbito del análisis de datos nos ofrece diferentes modelos representacionales, según el espacio geométrico preferido. También existen diferentes algoritmos de clasificación para trasladar la distancia original a una representación visualmente óptima. El problema está en que la utilización de modelos o de algoritmos diferentes puede dar como resultado clasificaciones sensiblemente diferentes. Por eso, el investigador se ve obligado a escoger el modelo más adecuado a sus finalidades y a evaluar con objetividad los resultados de las representaciones obtenidas en relación con los datos originales.

En nuestro trabajo hemos optado por recurrir a modelos de representación en árboles ultramétricos, ya que generalmente permiten una buena visualización de las clasificaciones y de la estructura de grupos.

4. Distancia lingüística de los datos del COD

La distancia lingüística obtenida es el conjunto de similitudes/diferencias observado entre las variedades del catalán a partir de la comparación de los datos del corpus. Los datos que presentamos a continuación son el resultado de aplicar una primera criba a los materiales del COD. Concretamente, hemos elaborado un subconjunto a partir de los datos totales, de forma que la respuesta mayoritaria entre los informantes de un punto de encuesta se ha considerado la más representativa del habla de una determinada localidad. A través de este cuidadoso proceso de filtrado, hemos procedido a elaborar varias bases de datos –una para cada ámbito morfológico estudiado– que presentan dos ventajas respecto del corpus original: por una parte, reducen la cantidad de datos a comparar, de tal manera que se simplifica notablemente el tratamiento estadístico final; por la otra, recogen los rasgos más representativos de cada uno de los puntos de encuesta y reflejan, así, las características más comunes del habla de sus habitantes.

El cómputo final de registros que han sido objeto de comparación es de 29.364; de estos, la mayor parte (20.500) corresponden a la morfología verbal, mientras que el resto (8.864) pertenecen a las otras seis categorías lingüísticas analizadas, que desglosamos a continuación.

4.1. Flexión nominal

Con este epígrafe hacemos referencia a los siguientes ámbitos morfológicos: los artículos, los posesivos, los clíticos pronominales, los pronombres personales, los demostrativos neutros y los adverbios locativos. De éstos, sólo los cuatro primeros pertenecen estrictamente al ámbito de la flexión nominal, mientras que los dos últimos se han incluido porque presentan diferencias dialectales significativas: aportan, pues, una información sustancial para determinar la distancia entre las diferentes variedades de la lengua.

Cada uno de estos ámbitos morfológicos se ha tratado en una base de datos diferenciada, a saber:

- a) Artículos. Esta base de datos consta de 1.312 registros (16 para cada población) correspondientes a los artículos personales del catalán y a los artículos determinados. Con respecto a los contextos de aparición, tanto el artículo determinado como el artículo personal se han recogido seguidos de palabra empezada en vocal y en consonante. Además, en el caso del artículo determinado femenino singular también se han tenido en cuenta, específicamente, los contextos en qué precede a una palabra empezada por las vocales *i* o *u* átonas o tónicas.
- b) Posesivos. Esta base de datos consta de 1.968 registros (24 para cada población) correspondientes a las formas de los posesivos tónicos antepuestos al nombre que determinan.
- c) Clíticos pronominales. Esta base de datos consta de 4.264 registros (52 para cada población) correspondientes a las formas de los pronombres débiles *em*, *et*, *es*, *el*, *la*, *els*, *les*, *li*, *ens*, *us*, *en*, *hi*, *ho* en los siguientes contextos: en posición proclítica (tanto prevocálica como preconsonántica) y en posición enclítica (tanto postvocálica como postconsonántica).
- d) Pronombres personales. Esta base de datos consta de 648 registros (8 para cada población) correspondientes a la primera, segunda y tercera personas del singular y del plural, tanto masculinas como femeninas.
- e) y f) Demostrativos neutros y locativos. Cada una de estas bases de datos consta de 246 registros (3 para cada población) que se corresponden a los tres grados de proximidad que pueden presentar algunas de las variedades de la lengua catalana.

El análisis lingüístico de la flexión nominal se ha llevado a cabo, pues, a partir de los datos contenidos en este subcorpus. En primer lugar, hemos procedido a la segmentación morfológica de los elementos impredecibles, para los que hemos postulado las diversas formas subyacentes que han sido objeto de comparación en el ámbito de la morfología. En concreto, hemos optado por una segmentación en morfos correspondientes a los morfemas raíz, género (+femenino) y número. Ni que decir tiene que estos últimos morfemas son inexistentes en los datos de e) y f).

4.2. Procesos Fonológicos

Paralelamente, hemos definido un conjunto de hasta sesenta procesos fonológicos con el objetivo de sistematizar los fenómenos predecibles que actúan sobre los segmentos subyacentes y los convierten en realizaciones fonéticas palpables. La complejidad de trabajar con un número tan elevado de procesos fonológicos es notoria, tal y como se desprende, por ejemplo, del tratamiento de la epéntesis por sonicidad, que exponemos brevemente a continuación.

La inserción de una vocal de apoyo para reparar estructuras que violan la escala universal de sonicidad es un recurso muy generalizado en el marco de la lengua catalana. Pero desde un enfoque dialectométrico, postular una única regla general de epéntesis se revela insuficiente para dar cuenta de las diferencias entre variedades, dado que intervienen diversos factores de diferenciación que deben ser tenidos en cuenta. Los ejemplos de (3) extraídos de las bases de datos de artículos y de clíticos pronominales, así lo certifican:

(3)a.	gos	gossos	rento
	gos	gossos	rento
	gos	gossos	rento
b.		gossos	rento
			rento
			rento
c.	gos	gossos	posar- lo
	gos	gossos	posar-

A la luz de estos datos, en un primer momento contemplamos la posibilidad de tratar la inserción de cada vocal (^{u} ^{u} ^{u} ^{u}) como un proceso fonológico específico, diferente también en función de la posición (anterior o posterior al morfo respectivo) en qué esta se insertaba. En consecuencia, una forma como ^{u} estaría marcada con la regla r1, que insertaría una vocal [^{u} en el morfo de raíz /l/; la forma ^{u} estaría marcada con la regla r2; ^{u} derivaría de una regla r3; ^{u} , de una regla r4; ^{u} , de una regla r5; ^{u} sería el resultado de aplicar la regla r6, etc.

Pero este sistema de reglas impedía reflejar que entre las formas del tipo ^{u} y ^{u} hay una distancia menor que entre las formas ^{u} y ^{u} ; o que las formas del tipo ^{u} y ^{u} son más parecidos entre ellas que en comparación con la forma ^{u} . Hacía falta, pues, tener en cuenta dos factores relevantes para la distancia lingüística: el timbre de la vocal insertada y la posición en qué se insertaba dicha vocal.

A partir de este razonamiento determinamos tres reglas de inserción vocálica que nos permitieran caracterizar el timbre de las vocales de apoyo: R1 ($\emptyset \rightarrow \text{e}$), R2 ($\emptyset \rightarrow \text{ }^{\text{u}}$) y R3 ($\emptyset \rightarrow \text{a}$). Además, hemos postulado dos reglas más para especificar la posición de la epéntesis

respecto al morfo: R4 (epéntesis a la izquierda) y R5 (epéntesis a la derecha). En el caso de la epéntesis condicionada morfológicamente, los ejemplos de (3.c) evidencian que es innecesario determinar la posición de la inserción, dado que ésta se realiza siempre a la derecha del morfo de raíz; en consecuencia, hemos postulado únicamente una regla R6 de introducción de la epéntesis morfológicamente condicionada y dos reglas más, R7 y R8, que determinan el timbre (respectivamente, [] y []). De esta manera, y tal y como se puede ver reflejado en el análisis de (4), se ha conseguido discriminar los elementos de proximidad y de distancia (con respecto al fenómeno de la epéntesis) entre las diversas formas de los artículos y de los clíticos pronominales:

(4) a. R4	gos R1, R4	gossos	R1, R4	rento	R1,
	gos R2, R4	gossos	R2, R4	rento	R2,
	gos R3, R4	gossos	R3, R4	rento	R3,
R4					
b. R5		gossos	R1, R5	rento	R1,
				rento	R2, R5
				rento	R3, R5
c. R7	gos R6, R7	gossos	R6, R7	posar- lo	R6,
	gos R6, R8	gossos	R6, R8	posar-	R6,
R8					

4.3. Flexión verbal

La base de datos sobre morfología verbal de nuestro corpus, que como hemos apuntado es la que contiene más registros, recoge los paradigmas inflectivos de los cinco modelos de verbos regulares del catalán siguientes: *cantar*, *perdre*, *beure*, *sentir* y *servir*. Mayoritariamente, se han recogido las formas simples de estos verbos, aunque en algunos casos también se han tenido en cuenta las formas perifrásticas. Esto es así, por ejemplo, en el pasado perifrástico (*vaig cantar*), que es mayoritario entre las variedades del catalán, però que contrasta con una forma simple de pasado propia de algunas zonas, por ejemplo, valencianas (*cantí*). De esta manera la base de datos contiene unas cincuenta formas para cada uno de los verbos. Como ya hemos comentado, estas formas verbales han sido analizadas lingüísticamente para discriminar las diferencias fonológicas de las fonéticas. Además también se ha realizado una segmentación previa de las formas verbales para poder comparar los diferentes segmentos morfológicos entre sí.

Es de decir que el cómputo de las diferencias entre variedades no se ha realizado a partir de la comparación de las diferentes formas verbales, sino a partir de la comparación de los segmentos morfológicos que las constituyen. Los segmentos morfológicos que se han tenido en cuenta son los de (5).

(5)

	Raíz	Tema	Extensión	TAM	Número/persona
--	------	------	-----------	-----	----------------

cantareu	/kánt/	/á/		/ é/	/u/
serveixis	/sé b/		/é /	/i/	/s/

Como ya hemos comentado en el apartado anterior, para la determinación de la distancia lingüística también hemos tenido en cuenta, además de los segmentos morfológicos, los diferentes procesos fonológicos que los afectan y que permiten explicar su realización fonética. Entre muchos otros, algunos de los procesos que hemos utilizado como base de comparación son: desacentuación de la raíz (*canto* [kánto] → *cantava* [kantá a]), ensordecimiento de obstruyentes en situación final de palabra (*begui* [bé i] → *bec* [bék]), elisión de segmento rótico en situación final de palabra (*cantar-la* [kantárla] → *cantar* [kantá]), etc.⁸

4.4. Cuantificación y representación de la distancia lingüística

Para establecer la comparación entre estos elementos y poder definir la distancia lingüística hemos utilizado el siguiente índice de distancia:

(6)

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

Es decir, la distancia lingüística entre dos variedades (*i, j*) es igual al sumatorio (Σ) de las diferencias en cuanto a una variable *k* entre las variedades (*i, j*), dividido por *long*, que es la longitud (número de sonidos) de cada segmento morfológico comparado.

En cuanto al método de representación gráfica nos hemos servido del *Cluster Análisis* y hemos usado un algoritmo de clasificación basado en el método UPGMA (*Unweighted pair-group method using arithmetic averages*)⁹, que ha sido contrastado ampliamente en aplicaciones de la taxonomía numérica en múltiples disciplinas. Para la evaluación de la distorsión entre las representaciones y la distancia original, aplicamos el coeficiente de correlación cofenética, con unos resultados que corroboran la fidelidad de las representaciones jerárquicas en relación con la distancia lingüística de partida.¹⁰

5. Resultado del análisis dialectométrico de los datos del COD

De una primera aproximación al dendrograma obtenido de la matriz de distancias lingüísticas globales se desprende que son cuatro las áreas dialectales claramente diferenciadas en el marco de la lengua catalana: un primer bloque formado por las variedades baleáricas (mallorquín, menorquín e ibicenco); un segundo bloque que comprende el catalán central y el septentrional; un tercer grupo en el que se integran tanto el catalán norte-occidental como la mayoría de las variedades de la provincia de Castelló; y, finalmente, un cuarto y último

⁸ Para una análisis dialectométrico de la morfología verbal del COD *vid.* Clua (2007).

⁹ *Vid.* Sneath, P.H.A. & Sokal, R. R. (1973).

¹⁰ La definición de la distancia lingüística se ha realizado con el programa Microsoft® Excel. El análisis de conglomerados y la representación gráfica de la distancia lingüística se ha llevado a cabo con el sistema de análisis multivariante GINGKO (Departament de Biologia Vegetal, Universitat de Barcelona <http://biodiver.bio.ub.es/vegana/index.html>).

conjunto que comprende el resto de variedades valencianas. Por su parte, las variedades de Benavarri y, en especial, de la localidad sarda del Alguer, ocupan posiciones notablemente aisladas; el dendrograma refleja, por consiguiente, su carácter marcadamente idiosincrático en el conjunto de la lengua catalana.

Tradicionalmente, la dialectología clásica ha postulado una división dialectal del catalán basada en un número limitado de isoglosas; de éstas, la que ha tenido una mayor trascendencia a efectos clasificatorios es la que separa los dialectos orientales de los occidentales en virtud, principalmente, de los respectivos procesos de reducción vocálica, por cuya acción los sistemas vocálicos tónico y átono de estas variedades difieren sustancialmente¹¹. A continuación partimos de la distinción entre catalán oriental y occidental para describir de forma pormenorizada la información contenida en el dendrograma de la figura 2.

Las variedades tradicionalmente adscritas al catalán oriental aparecen agrupadas de la siguiente forma en el dendrograma:

- a) El *alguerés*, hablado en la ciudad de Alguer, en la isla de Cerdeña, es el dialecto catalán que presenta un mayor número de soluciones propias diferenciadas del resto de variedades de la lengua. Su ubicación en el dendrograma (aislado del resto de variedades y a una gran distancia de todas ellas) refleja claramente esta realidad.
- b) El resto de hablas insulares (el *catalán de las baleares y pitiusas*) forma un clúster bien definido y con una clara estructura interna: por un lado, las variedades de la isla de Mallorca (Felanitx, Palma, Manacor, Pollença y, a una mayor distancia, Sóller) se agrupan entre ellas; por el otro, lo mismo ocurre en los casos de Eivissa y Formentera y de Ciutadella y Maó. Emergen, por lo tanto, las tres principales variedades baleáricas: el mallorquín, el menorquín y el ibicenco.
- c) Las variedades del *catalán central* constituyen un clúster especialmente homogéneo. De hecho, tan sólo el habla de Solsona se sitúa a una cierta distancia del resto, posiblemente a causa de su carácter de variedad de transición entre los subdialectos central y norte-occidental. Específicamente, pertenecen a este grupo las variedades de Banyoles, Barcelona, Manresa, Sabadell, Terrassa, Granollers, Igualada, Figueres, Girona, la Bisbal d'Empordà, Vic, Vilanova i la Geltrú, Berga, Valls, Santa Coloma de Farners, Montblanc, Tarragona, Mataró, Sant Feliu de Llobregat, Ripoll, Puigcerdà, Olot, el Vendrell, Reus, Vilafranca del Penedès y Solsona.
- d) Las variedades tradicionalmente adscritas al llamado *catalán septentrional de transición* se diluyen, en el presente dendrograma, en el homogéneo clúster del catalán central. Parece dudoso, cuando menos, defender la vigencia de este subdialecto a partir de los datos del COD.
- e) Las variedades adscritas al *catalán septentrional* (Perpinyà, Prada, Sallagosa y Ceret) se agrupan en primera instancia con el *catalán central*, pero a una distancia considerable; y entre ellas tampoco gozan de la homogeneidad del clúster del catalán central.

A excepción del habla del Alguer, cuyas peculiaridades impiden que se integre en un clúster superior, el resto de grupos se asocian entre ellos en un nodo de segundo nivel; lo forman, concretamente, el clúster formado por las variedades baleáricas y el clúster formado por los dialectos *central* y *septentrional*.

Por lo que se refiere a las variedades tradicionalmente adscritas al *catalán occidental*, presentan la siguiente estructura de grupos:

11 Cf. Veny (1982: 17-18).

- a) En primer lugar, emerge un clúster que engloba tanto las variedades norte-occidentales como aquellas hablas de transición al valenciano. Concretamente, parece razonable una distinción en cuatro subgrupos principales: (1) un conjunto de variedades homogéneas que se corresponde, a grandes rasgos, con el denominado *leridano*: Cervera, Lleida, Mollerussa, les Borges Blanques y Tàrraga. A este grupo se añaden también dos variedades pirenaicas (Andorra y la Seu d’Urgell) y, cada vez a mayor distancia Balaguer, en primer lugar; Falset, Móra d’Ebre y Tremp, a continuación; y, finalmente, el habla de Fraga; (2) un segundo conjunto de variedades, probablemente más conservadoras¹², que se agrupan, en primer término, entre ellas, y a continuación, con el clúster anterior: se trata de las hablas del Pont de Suert, Sort, Tamarit de Llitera, Gandesa y Vall-de-roures; (3) un tercer grupo, compuesto por las variedades de Amposta, Tortosa y Vinaròs, que constituyen el núcleo del dialecto *tortosino*; y, finalmente, (4) una última unidad que incluye las variedades de Albocàsser, l’Alcora, Morella y Castelló de la Plana. Las variedades de este último grupo, que a menudo han sido denominadas variedades de transición entre el catalán y el valenciano, se unen claramente al clúster de variedades norte-occidentales, con lo cual creemos poder aclarar considerablemente la filiación de estas hablas que a menudo ha sido objeto de discusión, ya que dependiendo de la isoglosa utilizada (1ª persona del presente de indicativo *cante/canto*; 3ª persona del mismo tiempo *cante/canta*; 2ª persona del imperfecto de subjuntivo *cantares/cantesses...*) se habían vinculado más a las variedades norte-occidentales o a las variedades valencianas.

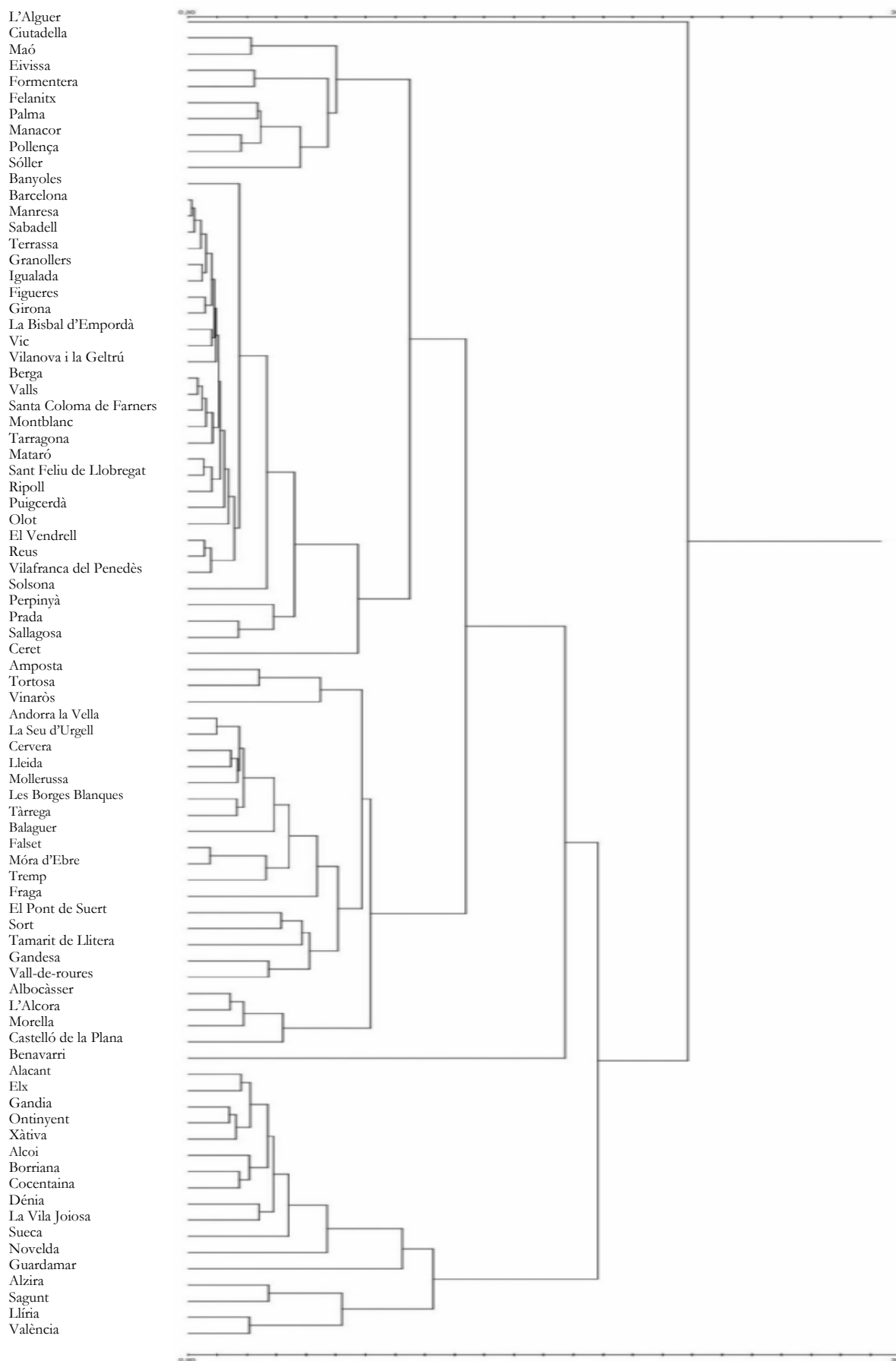
Contrariamente a los presupuestos de la dialectología clásica, el conjunto de variedades anteriores se clusterizan en primer lugar con el conjunto de hablas orientales y, sólo a continuación, con el resto de hablas valencianas. Lo cual tiene que ser objeto de un estudio más profundo por nuestra parte. En otro orden de cosas, el aislamiento de Benavarri evidencia que sigue vigente su carácter de variedad de transición al aragonés.

- b) Las hablas de las actuales provincias de València y Alacant (y también la variedad de Borriana) conforman, por último, un clúster que presenta, a su vez, una subdivisión en dos grupos: uno mayoritario, que engloba a las variedades del *valenciano central y meridional*,¹³ y otro formado tan sólo por cuatro localidades, representativas del denominado *valencià apitxat*. Se integran en el primer grupo las hablas de Alacant, Elx, Gandia, Ontinyent, Xàtiva, Alcoi, Borriana, Cocentaina, Dénia, la Vila Joiosa, Sueca y, a mayor distancia, Novelda y Guardamar. El segundo grupo incluye, en cambio, las variedades de Alzira, Sagunt, Lliria y València.

12 En base a las conclusiones de Viaplana (1999: 83-109).

13 En el sentido de Clua & Lloret (2007).

Figura 2. Rerepresentación de la distancia lingüística de las variedades del catalán a partir de los datos del COD



Referencias bibliográficas

- Bonet, Eulàlia & Lloret, Maria-Rosa (2005): "More on alignment as an alternative to domains: The syllabification of catalan clitics", *Probus* 17.1, 37-78.
- Clua, Esteve; Valls, Esteve; Viaplana, Joaquim (2008): "Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà". A: Blaikner-Hohenwart, Gabriele *et alii* (ed.), *Ladinometria Festschrift für Hans Goebler zum 65. Geburtstag*, Vol. 2. Salzburg, Universität Salzburg *et alii*, 27-42.
- Clua, Esteve (2007): "Distancia lingüística entre los dialectos del catalán a partir de los datos del COD", en Danler, Paul *et alii* (ed.), *Actes du XXVème Congrès International de Linguistique et de Philologie Romanes (Innsbruck 2007)*. Tübingen, Niemeyer. [en prensa].
- Clua, Esteve (1999a): *Variació i distància lingüística. Classificació dialectal del valencià a partir de la morfologia flexiva*. Tesis doctoral. Universitat de Barcelona.
- Clua, Esteve (1999b): "Distància lingüística i classificació de varietats dialectals", *Caplletra*, 26, 11-26.
- Clua, Esteve & Lloret, Maria-Rosa (2006): "New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)", en Montreuil, J-P. (ed.): *New Perspectives on Romance Linguistics. Vol. 2: Phonetics, phonology, and dialectology*. Amsterdam, John Benjamins, 31-47.
- Clua, Esteve & Lloret, Maria-Rosa (2007): "Clasificación de variedades dialectales mediante técnicas de análisis multivariante, a partir de un corpus oral", en: *Actas del VI Congreso de Lingüística General* (Santiago de Compostela, 3-7 de mayo de 2004). Madrid, Arco/Libros.
- Lloret, Maria-Rosa (2004): "The phonological role of paradigms: The case of insular Catalan", en Auger, Julie / Clements, Clancy / Vance, Barbara (ed.): *Contemporary Approaches to Romance Linguistics*. Amsterdam / Philadelphia, John Benjamins, 275-279.
- Lloret, Maria-Rosa & Joaquim Viaplana (1998): "Variació morfofonològica. Variants morfològiques", *Caplletra* 25, 43-62.
- Sneath, P.H.A. & R. R. Sokal (1973): *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. San Francisco, W. H. Freeman and Company.
- Veny, Joan (1982): *Els parlars catalans (síntesi de dialectologia)*. Palma, Moll.
- Viaplana, Joaquim; Lloret, Maria-Rosa; Perea, Maria-Pilar; Clua, Esteve (2007): *COD. Corpus Oral Dialectal*. Barcelona, PPU. (Publicació en CD-rom).
- Viaplana, Joaquim (1999): *Entre la dialectologia i la lingüística*. Barcelona, Publicacions de l'Abadia de Montserrat.