RELEVANCIA DEL ANÁLISIS LINGÜÍSTICO EN EL TRATAMIENTO CUANTITATIVO DE LA VARIACIÓN DIALECTAL

Esteve Clua
IULA-Universitat Pompeu Fabra

Abstract

In previous studies we have argued in favour of the need for a phonological analysis of dialectal data prior to quantitatively determining linguistic distance. Our idea, based on generative linguistics, has to do with the fact that phonetic coincidences may frequently hide relevant phonological divergences. Thus, we consider that quantitative analyses carried out solely from phonetic data cannot reflect existing differences or similarities among dialectal varieties adequately. So far, our reasoning has always been backed by small-scale laboratory tests performed on small sets of data. In the present study, however, our reasoning in favour of linguistic analysis is corroborated by a tested and global treatment of data from the Contemporary Catalan Oral Dialect Corpus. To this end we have carried out two quantitative studies: one on the basis of phonetic data, and another one using a phonological analysis of the data.

Key words: *dialectometry, quantitative analysis, linguistic distance, phonological analysis, phonetic data.*

1. Introducción[1]

Una de las características distintivas de los trabajos dialectométricos[2] que se han desarrollado alrededor del *Corpus Oral Dialectal* (COD)[3] del catalán contemporáneo, es el hecho de basar los tratamientos cuantitativos para describir la variación lingüística del catalán en un análisis lingüístico previo. Como veremos, en estudios anteriores hemos justificado esta necesidad a partir de pequeños conjuntos de datos extraídos del corpus.

---

[1] Este trabajo forma parte del proyecto de investigación HUM2007-65531/FILO (ECOD "Explotación de un corpus oral dialectal: Análisis de la variación lingüística y desarrollo de aplicaciones informáticas para la trascripción automatizada, 2.ª fase"), financiado por el Ministerio de Ciencia e Innovación y el FEDER. Más información sobre el proyecto se encuentra disponible en http://www.ub.edu/lincat.

[2] *Vid.*, por ejemplo, Viaplana (1999), Clua (1999a y b, 2007), Clua *et alii* (2008 y 2009).

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

La finalidad de este trabajo es presentar por primera vez el contraste entre dos análisis cuantitativos globales del COD: uno realizado a partir de los datos fonéticos y otro a partir de los datos analizados fonológicamente. Nuestro objetivo es discernir si las posibles diferencias entre ambos tratamientos son pertinentes y justifican, por tanto, el trabajo de análisis lingüístico de los datos antes de llevar a cabo el tratamiento cuantitativo.

Para ello, en primer lugar (§2), argumentamos la necesidad de llevar a cabo un análisis lingüístico previo de los datos para poder captar todas las similitudes o diferencias que pueden existir entre determinadas variedades lingüísticas en relación a un determinado rasgo. En segundo lugar (§3), describimos el análisis fonológico que aplicamos a los datos del COD. A continuación (§4), a partir de una pequeña muestra de datos, como si se tratase de un pequeño ensayo de laboratorio, presentamos las implicaciones de tal planteamiento para determinar cuantitativamente la distancia lingüística entre variedades. En el apartado siguiente (§5), describimos y analizamos los resultados en contraste de los dos tratamientos cuantitativos aplicados a los datos del COD. Finalmente en (§6) presentamos las conclusiones de nuestro estudio.

2. El porqué del análisis lingüístico previo

Si alguna cosa caracteriza globalmente los diferentes métodos cuantitativos de descripción y clasificación dialectal que han proliferado durante las últimas cuatro décadas en Europa, es el hecho de compartir la adopción del concepto de distancia lingüística como pieza básica para la descripción de la variación lingüística. El concepto de distancia, adoptado del ámbito científico del análisis de datos, se asocia generalmente en nuestro campo a la cuantificación de las similitudes o diferencias que existen entre variedades lingüísticas en relación a un conjunto de datos. Se trata de un punto de vista sobre la variación lingüística que se aparta sustancialmente de la dialectología tradicional, pero que ya se podía vislumbrar en el siglo XIX en las palabras de Durand (1889):

> Et maintenant, qu'est-ce qui constitue le degré de ressemblance qui rapproche deux langues entre elles, et le degré de dissemblance qui les éloigne l'une de l'autre? La ressemblance se mesure à la proportion des caractères communs, la dissemblance à la proportion des caractères particuliers.

Pero ¿cómo determinamos las diferencias o similitudes entre variedades lingüísticas a partir de las cuales establecer el tratamiento cuantitativo? ¿Reflejan las representaciones fonéticas de los ítems de un determinado atlas o corpus todas las diferencias existentes entre las variedades lingüísticas? Vamos a intentar responder a estas preguntas a partir de los ejemplos siguientes en los que contrastamos las formas del numeral *dos* en dos variedades lingüísticas.

(1)	*Variedad 1*	*Variedad 2*
a.	dos ['dos]	dos ['dos]
b.	dos cafès ['doska'fɛs]	dos cafès ['doska'fɛs]
c.	dos bonsais ['dozβon'sajs]	dos bonsais ['dozβon'sajs]
d.	dos animals ['dozani'mals]	dos animals ['dosani'mals]

Si contrastamos la realizaciones fonéticas del numeral aislado (1a) en las dos variedades, vemos que no presentan ninguna diferencia. Tampoco hay diferencias cuando aparece delante de otra palabra empezada por consonante sorda (1b). Cuando precede una palabra empezada por consonante sonora (1c) tampoco encontramos diferencias entre ambas variedades, pero en este caso podemos observar que el sonido sibilante del final de la palabra presenta una realización sonora [z], que contrasta con las realizaciones sordas de este segmento en los contextos anteriores. Finalmente en (1d) si que podemos observar una realización diferente del numeral en las dos variedades contrastadas: mientras que la variedad 1 presenta una realización sonora de la sibilante final, en la variedad 2 encontramos una realización sorda.

Teniendo en cuenta esto podríamos concluir, en principio, que estas dos variedades no presentan ninguna diferencia en la representación fonética del numeral *dos* ['dos], porqué los tres segmentos que integran esta palabra coinciden totalmente; pero la coincidencia no puede ser total si tenemos en cuenta el comportamiento del último segmento, el sibilante [s], que se sonoriza siempre delante de un segmento sonoro [± vocálico] en la variedad 2, mientras que solo lo hace delante de un segmento no vocálico en la variedad 1. Es decir, las dos variedades coinciden totalmente en los segmentos fonéticos que constituyen el numeral *dos*, pero los procesos fonológicos que afectan dichos segmentos son diferentes. Si entendemos que estos procesos fonológicos son básicos para comprender la estructura sonora de las variedades lingüísticas, tendremos que colegir asimismo que no podemos pasarlos por alto al intentar captar las diferencias existentes entre ellas.

Para poder tener en cuenta todos estas diferencias, aplicamos a los datos fonéticos de nuestro corpus un análisis lingüístico basado en el modelo de la fonología generativa clásica, porque permite discriminar las diferencias superficiales o predecibles, que expresan las regularidades de las lenguas (y de las variedades que las componen), de las diferencias subyacentes o impredecibles, que afectan a la estructura léxica o gramatical de las palabras (*vid.* Lloret & Viaplana 1998). Desde nuestro punto de vista, la distinción entre estos dos niveles de análisis es fundamental para determinar la distancia lingüística entre variedades (*vid.* Clua 1999a, b y Viaplana 1999).

3. El análisis aplicado a los datos del COD

Para el análisis de fenómenos fonológicos específicos, hemos seguido mayoritariamente la orientación propia de la fonología generativa derivacional. 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 relaciones intraparadigmáticas e interparadigmáticas que se establecen entre palabras morfológicamente relacionadas.[4]

---

[4] Para un análisis derivacional y morfémico 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).

Podemos ver, a continuación, un caso de diferencias fonológicas (subyacentes) y otro de diferencias fonéticas (superficiales). El catalán presenta distintas terminaciones para la 1.ª 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 nordoccidental). 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 morfológica de las palabras. Podemos verlo en los ejemplos de (2), correspondientes al verbo *cantar*.

(2) Diferencias Fonológicas
1.ª persona del singular del Presente de Subjuntivo de *cantar*

a. Variedad 1: [ˈkante] /e/
b. Variedad 2: [ˈkanti] /i/
c. Variedad 3: [ˈkanta] /a/

Otro ejemplo de variación que afecta a las terminaciones verbales lo encontramos en las formas del gerundio. Muchas variedades, como es el caso del catalán de Barcelona, presentan una -[n] final en esta forma verbal; en cambio muchas de las variedades valencianas acaban el gerundio con -[nt]. Si realizamos una cuantificación de las diferencias a partir de los datos fonéticos, estas diferencias son iguales que las de (2). Desde la óptica generativa, sin embargo, la alternancia [n] ~ [nt] que se observa en la variedad 1 (3a) 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, ya que en estas variedades opera un proceso fonológico de simplificación del grupo consonántico [nt] en final de palabra, que no se produce cuando la forma verbal va seguida de un clítico pronominal; en cambio en las variedades de (3b) no se produce ningún tipo de alternancia, ya que estas variedades no conocen el proceso de simplificación del grupo [nt].

(3) Diferencias Fonéticas

a. Variedad 1:

cantant	‘cantando’	[kanˈtan]	/nt/
cantant-ho	‘cantándolo’	[kanˈtanto]	

b. Variedad 2:

cantant		[kanˈtant]	/nt/
cantant-ho		[kanˈtanto]	

Desde nuestra perspectiva las diferencias observadas en (2) son fonológicas, solo predecibles a partir de la estructura morfológica; mientras que en (3) las diferencias son meramente fonéticas y se pueden predecir por un proceso fonológico. Creemos que se trata de una distinción pertinente que debe tenerse en cuenta 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. Así pues,

el análisis lingüístico nos permite, por un lado, captar similitudes o diferencias entre variedades que a simple vista fonética serían imperceptibles, y por el otro, nos permite distinguir entre diferencias estructurales (las que aquí denominamos fonológicas) y diferencias predecibles (fonéticas) a partir de los procesos fonológicos sistemáticos que caracterizan las variedades lingüísticas.

4. Un ensayo de laboratorio

A partir del análisis de los clíticos pronominales, una de las áreas en que el catalán presenta más variación dialectal, hemos justificado en trabajos anteriores (*vid.* Clua y Lloret 2006 y 2007) la pertinencia del análisis lingüístico para realizar un tratamiento cuantitativo adecuado. Se trataba de lo que podríamos catalogar de pequeño ensayo de laboratorio a partir de la variación que presentan estos clíticos en tres variedades lingüísticas del catalán, entre las que medíamos y representábamos la distancia lingüística, primero, a partir de los datos fonéticos y, a continuación, a partir de los datos analizados fonológicamente.

Volveremos aquí a presentar nuestra argumentación, en este caso ciñéndonos al pronombre de 1.ª persona del singular *me* en tres variedades occidentales del catalán. El pronombre aparece en negrita en los ejemplos de (4).

(4) Clíticos pronominales de 1.ª persona del singular

	Variedad 1	*Variedad 2*	*Variedad 3*
a. *em pensaré* “me pensaré”	[**em**pensaˈɾe]	[**me**pensaˈɾe]	[**me**pensaˈɾe]
b. *m’esperaré* “me esperaré”	[**m**esperaˈɾe]	[**m**esperaˈɾe]	[**m**esperaˈɾe]
c. *vol esperart-me* “quiere esperar-me”	[espeˈɾaɾ**me**]	[espeˈɾaɾ**me**]	[espeˈɾaɾ**me**]
d. *esperà’m* “espérame!”	[esˈperə**m**]	[esˈperə**m**]	[esˈperə**me**]

Los ejemplos de (4) muestran que en estas variedades el pronombre *me* se presenta con una forma no silábica [m] y con dos formas silábicas distintas, [em] y [me]. En la variedad 1, [em] aparece delante de un verbo que empieza en consonante (4a), mientras que [me] aparece detrás de un verbo que acaba en consonante (4c). En la variedad 2, [me] aparece en los dos contextos anteriores (4a,c). En la variedad 3, [me] aparece en los contextos anteriores (4a,c) y también cuando el verbo acaba en vocal (4d). Desde el punto de vista tradicional la distancia lingüística entre estas variedades es similar: las tres coinciden en las formas de (4b,c), y mientras las variedades 1 y 3 difieren en dos casos: (4a) y (4d), la variedad 2 se distingue de la variedad 1 en una forma, (4a), y de la variedad 3 en otra forma, (4d).

Si basamos nuestro estudio de la variación dialectal únicamente en los datos fonéticos, la distancia lingüística entre las variedades 1 y 2 tiene el valor 1, ya que solo difieren en una forma: [**em**pensaˈɾe] vs. [**me**pensaˈɾe]. Las variedades 1 y 3, en cambio, presentan una distancia lingüística de valor 2 porque difieren en dos formas:

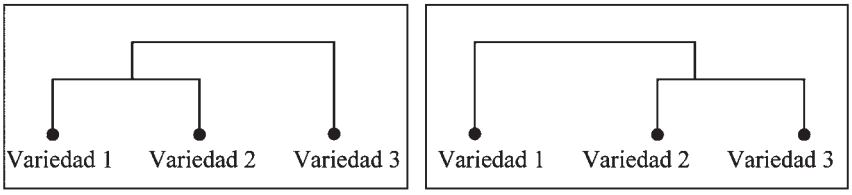
[**em**pensa'ɾe] vs. [**me**pensa'ɾe] i [es'peɾa**m**] vs. [es'peɾa**me**]. Por su parte, las variedades 2 y 3 tienen una distancia lingüística también de valor 1, ya que únicamente difieren en la forma: [es'peɾa**m**] vs. [es'peɾa**me**]. A continuación presentamos la matriz de distancias que corresponde a estas diferencias fonéticas.

(5) Matriz de distancias a partir de los datos fonéticos

	Variedad 1	Variedad 2	Variedad 3
Variedad 1	0		
Variedad 2	1	0	
Variedad 3	2	1	0

La representación dendrográfica de esta matriz de distancias presenta dos opciones posibles, dependiendo de si agrupamos primero las variedades 1 y 2, o 2 y 3, ya que ambas agrupaciones presentan una distancia mínima de valor 1. En todo caso, la distancia entre las tres variedades es mínima y podríamos decir que hay una distancia muy similar entre ellas.

(6) Representación dendrográfica de la distancia lingüística a partir de los datos fonéticos.



Estos casos de variación puede reanalizarse si se tienen en consideración aspectos relacionados con la estructura silábica y distinguimos entre formas subyacentes y formas predecibles. Así, en las variedades 1 y 2 las distintas formas del pronombre pueden explicarse a través de una única forma subyacente no silábica /m/. En este caso, la vocal [e] (que es la vocal no marcada del sistema vocálico átono de estas variedades: [a], [e], [i], [o], [u]) tiene carácter epentético; es decir, se añade para permitir la silabación adecuada de la secuencia formada por el pronombre y el verbo. La diferencia entre ambas variedades radica en la ubicación de la epéntesis. En la variedad 1, la vocal epentética aparece siempre en la periferia del grupo formado por el verbo y el pronombre: [**em**pensa'ɾe] (7a) vs. [espe'ɾaɾ**me**] (7c); en cambio, en la variedad 2 la epéntesis siempre aparece en un lugar fijo, a la derecha del pronombre: [**me**pensa'ɾe] (7a) y [espe'ɾaɾ**me**] (7c). Desde esta perspectiva, la variedad 3 es completamente distinta. El ejemplo determinante es (3d), [es'peɾa**me**], en donde la presencia de la vocal final del pronombre no puede justificarse por razones de silabación, pues esta variedad podría presentar perfectamente una forma [es'peɾa**m**] sin necesidad de recurrir a epéntesis alguna. En este caso, es más coherente postular una forma subyacente distinta, /me/, con una vocal final que se elide en determinados contactos vocálicos

[**m**espera'ɾe], en (7b)), al igual que ocurre en otros casos (cf. *entre amics: entr[a]mics*; *no és tan gran: n[o]s tan gran*).

(7)

	Variedad 1	Variedad 2	Variedad 3
a.	[**em**pensa'ɾe]	[**me**pensa'ɾe]	[**me**pensa'ɾe]
b.	[**m**espera'ɾe]	[**m**espera'ɾe]	[**m**espera'ɾe]
c.	[espe'ɾaɾ**me**]	[espe'ɾaɾ**me**]	[espe'ɾaɾ**me**]
d.	[es'pera**m**]	[es'pera**m**]	[es'pera**me**]
f.	/m/	/m/	/me/

Podríamos decir, en otras palabras, que la variedad 3 ha mantenido la forma originaria del clítico, *te* (coincidente con la forma del latín), aunque que elide la vocal por procesos fonológicos sistemáticos en contacto con determinadas vocales. En cambio, en la variedades 1 y 2 se ha producido una restructuración del sistema pronominal: la mayoría de clíticos pronominales de estas variedades suelen estar constituidos subyacentemente por una consonante (/m/, /t/, /s/…) a la cual añaden una vocal epentética cuando lo exigen las reglas de silabación. Teniendo en cuenta este análisis, los aspectos que presentan variación entre estas tres variedades son los de (8).

(8) Diferencias entre variedades

a. Variedad 1: /t/ y epéntesis en la periferia
b. Variedad 2: /t/ y epéntesis a la derecha del pronombre
c. Variedad 3: /te/ y elisión de vocales en contacto

Ahora la perspectiva de la distancia lingüística entre estas variedades es bastante diferente, mientras que las dos primeras coinciden plenamente en cuanto a la forma subyacente y solo difieren en el tipo de epéntesis utilizado, la tercera variedad se aparta de ellas considerablemente, tanto en el ámbito de las formas subyacentes como en el de los procesos fonológicos.

A continuación presentamos las matrices de distancias obtenidas a partir del análisis fonológico de los clíticos pronominales. En (9a), donde se presentan las diferencias morfológicas o subyacentes, podemos ver cómo mientras las variedades 1 y 2 presentan una distancia de valor cero, ya que ambas variedades coinciden en la forma /m/ del pronombre, la variedad 3 tiene un valor 4 respecto de las otras dos variedades.

(9) Matrices de distancias a partir de los datos analizados fonológicamente

a. Diferencias morfológicas (Variedades 1 y 2 /t/, Variedad 3 /te/)

	Variedad 1	Variedad 2	Variedad 3
Variedad 1	0		
Variedad 2	0	0	
Variedad 3	4	4	0

b. Diferencias en los procesos fonológicos

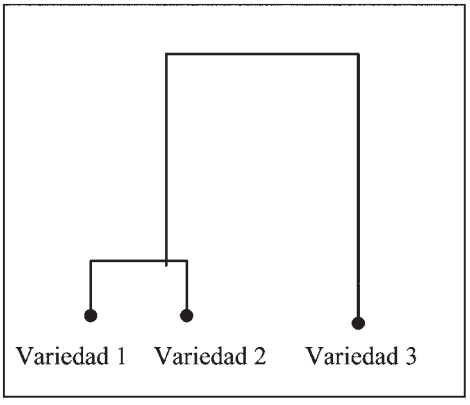
	Variedad 1	Variedad 2	Variedad 3
Variedad 1	0		
Variedad 2	1_a	0	
Variedad 3	2_b	2_c	0

a. Tipo de epéntesis: epéntesis en la periferia / epéntesis a la derecha del pronombre
b. Epéntesis y elisión de vocal
c. Epéntesis y elisión de vocal

En (9b) podemos ver las diferencias relacionadas con los procesos fonológicos. Entre las variedades 1 y 2 solo existe una diferencia derivada del tipo de epéntesis que experimentan: periférica en el caso de la variedad 1/ a la derecha del pronombre en la variedad 2. En cambio, entre estas dos variedades y la variedad 3 encontramos dos diferencias: la elisión de vocal, que se produce en 3 y no en las otras dos variedades; y la epéntesis, que no experimenta la variedad 3, a diferencia de lo que ocurre en las otras dos variedades.

De acuerdo con este análisis, el dendrograma siguiente muestra un grupo muy compacto entre las variedades 1 y 2, mientras que la variedad 3 presenta una distancia considerable en relación con este primer clúster.

(10) Representación dendrográfica de la distancia lingüística a partir de los datos analizados fonológicamente.



Con este ensayo de laboratorio que hemos realizado a partir de un ejemplo de variación extraído del clítico pronominal de 1.ª persona del singular, creemos demostrar claramente la pertinencia de basar el tratamiento cuantitativo de la distancia lingüística en datos analizados lingüísticamente. Sólo de esta manera podemos captar que las diferencias lingüísticas son menores entre las variedades 1 y 2, que la variedad es la que presenta una distancia mayor, a pesar de que fonéticamente parezca que la distancia entre las tres variedades sea similar.

5. Contraste de dos análisis dialectométricos de los datos del COD

A pesar de todo y en contra de la argumentación presentada en el apartado anterior, algunas veces se ha esgrimido que en un tratamiento cuantitativo global, con grandes cantidades de datos lingüísticos como las que pueden ofrecer un atlas o un corpus como el COD, estas diferencias pueden ser totalmente inapreciables e intrascendentes. Por eso, con la finalidad de intentar dilucidar este punto, hemos llevado a cabo un tratamiento cuantitativo contrastado de los datos del COD, que presentamos a continuación. Se trata de contrastar la representación de la distancia lingüística de los datos de dicho corpus previamente analizados fonológicamente (*vid*. Clua et *alii* 2009) con la que hemos obtenido a partir de los mismos datos, pero sin el análisis fonológico previo.

Los datos fonéticos que hemos tratado cuantitativamente coinciden con el subcorpus analizado fonológicamente de Clua et *alii* (2009). Son pues el resultado de aplicar una primera selección al conjunto total de los materiales del COD. Esta selección se basa en escoger la respuesta mayoritaria entre los informantes de un punto de encuesta, considerando que se trata de la opción más representativa del habla de una determinada localidad. A través de este proceso de filtrado, hemos elaborado 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 representativas 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: artículos, posesivos, clíticos pronominales, pronombres personales, demostrativos neutros y adverbios locativos.

Cabe decir que el cómputo de las diferencias entre variedades no se ha realizado a partir de la comparación de las diferentes formas verbales o nominales, sino a partir de la comparación de los segmentos morfológicos que las constituyen. Los segmentos morfológicos que se han tenido en cuenta en el caso de la flexión verbal han sido los de (11). ('Extensión' es un sufijo post-radical que en catalán determina la subclase verbal; 'TAM' representa las categorías de tiempo, aspecto y modo). En el caso de la flexión nominal hemos contrastado los segmentos: raíz, género ([± femenino]) y número.

(11)

	Raíz	Tema	Extensión	TAM	Número/persona
cantareu	kant	a		ɾé	w
serveixis	serb		éʃ	i	S

En la determinación de la distancia a partir del análisis lingüístico previo de los datos, además de la comparación de los segmentos morfológicos también se tuvieron en cuenta los diferentes procesos fonológicos que los afectan y que permiten explicar su realización fonética. Entre muchos otros, algunos de los procesos que hemos uti-

lizado como base de comparación en la flexión verbal son: desacentuación de la raíz (*canto* [kánto] pero *cantava* [kantáβa]), ensordecimiento de obstruyentes en situación final de palabra (*begui* [béɣi] pero *bec* [bék]), elisión de *-r* en situación final de palabra (*cantar-la* [kantárla] pero *cantar* [kantá]), etc.[5] En la flexión nominal hemos trabajado, entre otros, con los procesos de elisión de vocales y de inserción de epéntesis como los que hemos comentado en el apartado anterior.

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

$$(12) \quad dist\,(i,j) = \frac{\sum_{k=1}^{long} dif_{k}(i,j)}{long} \times 100$$

Es decir, la distancia lingüística entre dos variedades (*i, j*) es igual al sumatorio (∑) 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*) (*vid*. Sneath & Sokal 1973), 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.[6]

5.1. Representaciones dendrográficas de los dos análisis dialectométricos del COD

A continuación presentamos las dos representaciones dendrográficas de la distancia lingüística entre las variedades del COD. En la figura 1 tenemos la distancia lingüística resultante del análisis dialectométrico aplicado a los datos del corpus después de ser analizados fonológicamente; en la figura 2 podemos ver el resultado de aplicar el mismo análisis dialectométrico a los datos fonéticos del COD.

De una primera aproximación al dendrograma realizado a partir del análisis fonológico se desprende que son cuatro las áreas dialectales claramente diferenciadas en el marco de la lengua catalana, en este caso parce imposible vislumbrar agrupaciones de nivel superior. Observamos un primer grupo que está formado por las variedades baleares (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 conjunto que comprende el resto de

---

[5] Para una análisis dialectométrico de la morfología verbal del COD, *vid.* Clua (2007).

[6] 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).

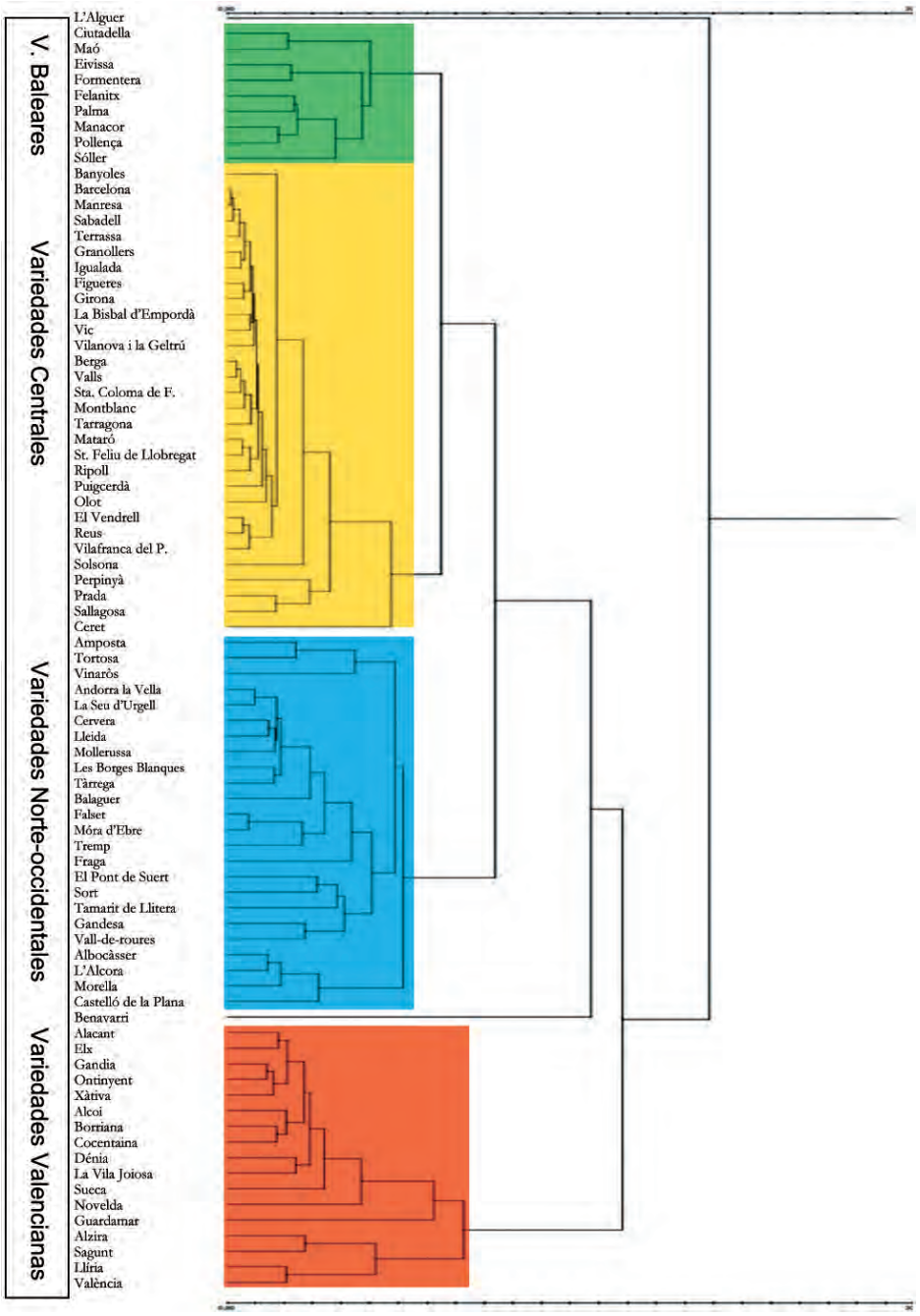


Figura 1

Distancia lingüística a partir de los datos del COD analizados fonológicamente

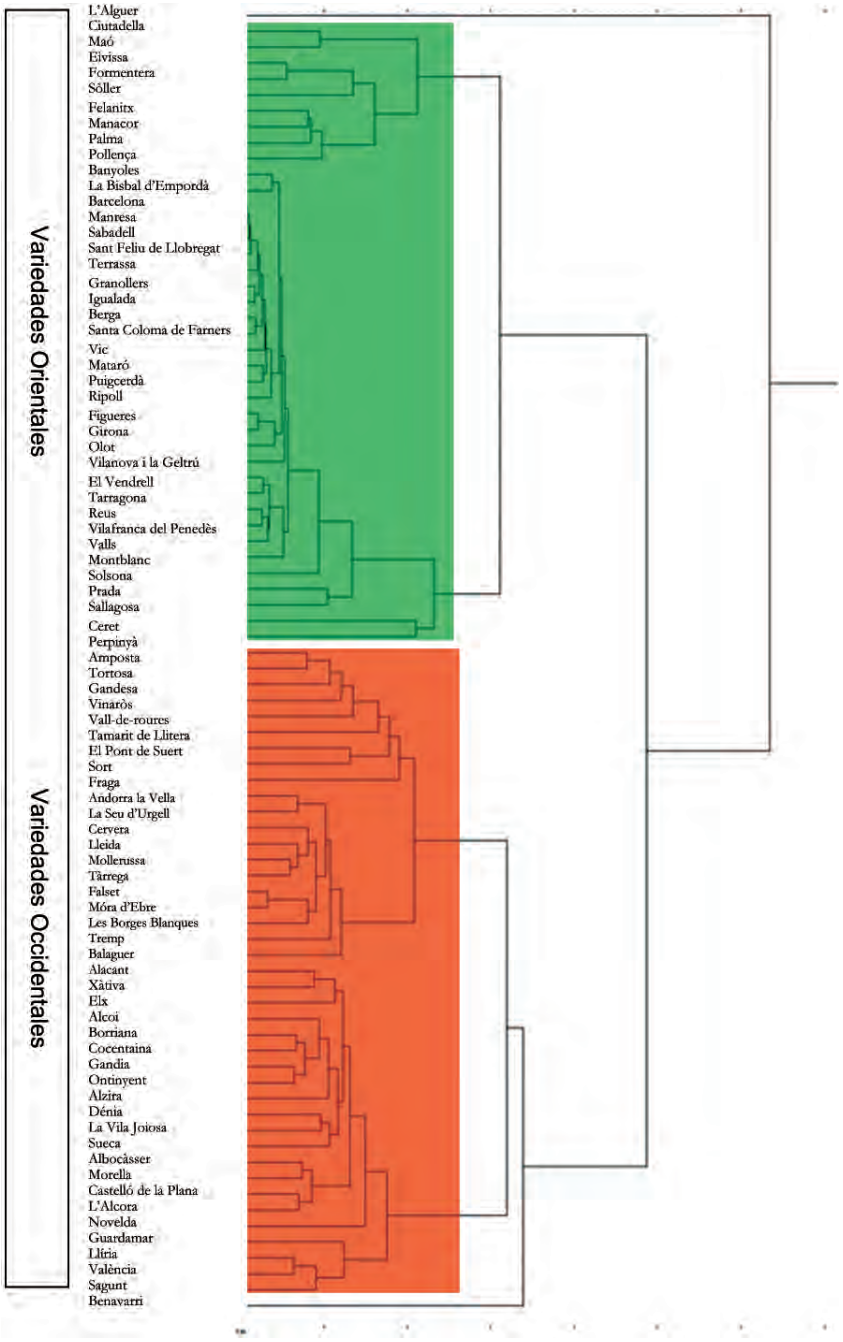


Figura 2

Distancia lingüística a partir de los datos del COD fonéticos

variedades valencianas. Por su parte, las variedades de Benavarri y, en especial, la 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.

En cambio, en la representación arbórea que resulta del tratamiento de los datos fonéticos podemos discernir claramente una agrupación de nivel superior a las anteriores que coincidiría, grosso modo, con la división que ha postulado tradicionalmente la dialectología clásica catalana. Una división dialectal del catalán en variedades orientales y occidentales. Esta clasificación estaba basada en un número limitado de isoglosas y principalmente, se ceñía a los diferentes procesos de reducción vocálica, por cuya acción los sistemas vocálicos tónico y átono de estas variedades difieren sustancialmente (*vid.* Veny 1982: 17-18). Si tenemos en cuenta que cuando se postuló esta clasificación tradicional, los estudios dialectales se basaban únicamente en los datos fonéticos, no es de extrañar que se produzca esta coincidencia. Por lo que respecta a las variedades del Alguer y de Benavarri, la coincidencia con el dendrograma fonológico es casi total; en este caso también aparecen como variedades aisladas del resto de los conglomerados.

En cuanto al resto de variedades agrupadas en la clasificación tradicional bajo el epígrafe de *catalán oriental,* hay que decir que las diferencias observadas en ambos tratamientos son poco remarcables. En todo caso, se puede afirmar que en el dendograma fonológico las agrupaciones parecen en general más compactas, con una distancia lingüística interna menor.

Así, en ambas representaciones las hablas insulares, las variedades de las Baleares *y* Pitiusas, constituyen 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.

También hay coincidencia en las variedades del catalán central, que en constituyen un grupo especialmente homogéneo en las dos estructuras arbóreas. 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, que es la única que se sitúa a una cierta distancia del resto, posiblemente a causa de su carácter de transición entre los subdialectos central y norte-occidental. A este conglomerado, que es el más compacto de todos, se le agrupan a una distancia considerable las variedades del llamado *catalán septentrional:* Perpinyà, Prada, Sallagosa y Ceret.

Donde sí que se aprecian claramente diferencias sustanciales entre ambos análisis dialectométricos es en las agrupaciones de las variedades que tradicionalmente se han reunido bajo el rótulo de *variedades occidentales.* Se trata de diferencias que, desde nuestro punto de vista, justifican por sí solas la necesidad de realizar un análisis fonológico previo para poder determinar adecuadamente la distancia lingüística entre variedades.

La primera diferencia importante tiene que ver con el hecho que el conjunto de variedades formado por el catalán norte-occidental y el tortosino en un sentido amplio (entendido como el conjunto de variedades que constituían la antigua diócesis de Tortosa) se agrupan en primera instancia con el conjunto de hablas orientales y, sólo a continuación, con el resto de hablas occidentales, es decir con el resto de hablas valencianas.

Veamos a continuación otras diferencias relevantes. En el dendrograma fonológico las variedades tradicionalmente adscritas al catalán occidental, presentan la siguiente estructura de grupos. En primer lugar, emerge un clúster que engloba tanto las variedades nord-occidentales como aquellas hablas de transición al valenciano. Concretamente, parece razonable una distinción en cuatro subgrupos principales: (i) un conjunto de variedades homogéneas que se corresponde, a grandes rasgos, con el denominado leridano: Cervera, Lleida, Mollerussa, les Borges Blanques y Tàrrega. 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; (ii) un segundo conjunto de variedades, probablemente más conservadoras,[7] 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; (iii) un tercer grupo, compuesto por las variedades de Amposta, Tortosa y Vinaròs, que constituyen el núcleo del dialecto tortosino, y, finalmente, (iv) un último cluster, que aunque aparezca físicamente alejado del anterior (entre ambos aparecen los grupos (i) y (ii)) en realidad está a muy poca distancia lingüística, 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 nord-occidentales, con lo cual creemos poder aclarar considerablemente la filiación de estas hablas, que a menudo ha sido objeto de discusión porque, dependiendo de la isoglosa utilizada (1.ª persona del presente de indicativo *cante*/*canto*; 3.ª persona del mismo tiempo *cante*/*canta*; 2.ª persona del imperfecto de subjuntivo *cantares*/*cantesses*…), se habían vinculado más a las variedades nord-occidentales o a las variedades valencianas.

En cambio, en el dendrograma obtenido a partir de los datos fonéticos del COD las variedades nord-occidentales presentan una estructura mucho menos coherente. En principio, las agrupaciones en torno al núcleo central, o leridano, son parecidas a las anteriores, pero a partir de aquí las divergencias son manifiestas. De entrada, al lado de este primer grupo, sólo existe un segundo clúster muy poco compacto donde se agrupan las variedades del tortosino estricto (Amposta, Tortosa, Gandesa y Vinaròs) con las hablas más occidentales (Fraga, Pont de Suert, Sort, Tamarit de Llitera y Vall-de-roures). Por su parte, el grupo de variedades del norte de Castelló (con la excepción de Vinaròs), que en la clasificación anterior se agrupaba con las variedades del tortosino, aquí se sitúa en el centro del clúster del resto de variedades valencianas, a una distancia demasiado importante, para ser coherente, del tortosino estricto.

---

[7] Sobre la base de las conclusiones de Viaplana (1999: 83-109).

Por lo que respecta al resto de variedades valencianas, en el diagrama fonológico 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;[8] y otro formado tan sólo por cuatro localidades, representativas del denominado valenciano *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, Llíria y València.

Por el contrario, en el dendrograma obtenido a partir de los datos fonéticos, estas variedades se agrupan en diferentes subgrupos con escasa coherencia. Aparte de la inclusión en el centro del clúster de las variedades del norte de Castelló, que ya hemos comentado, parece poco coherente que al grupo de variedades *apitxadas* nucleares (València, Llíria y Sagunt) se le añada antes Guardamar que Alzira, una variedad que tiene muchos rasgos del valenciano *apitxat*. Tampoco nos parece coherente la distribución de las variedades pertenecientes al valenciano meridional.

6. Conclusiones

Después de analizar las diferencias entre los dendrogramas de los análisis dialectométricos del COD, con y sin análisis fonológico previo, creemos que esta comparación realizada a partir de un corpus de datos considerable corrobora las hipótesis a las que habíamos llegado con los ensayos de laboratorio anteriores. En el sentido que los resultados de un análisis dialectométrico pueden ser considerablemente diferentes si partimos de datos fonéticos o si lo hacemos después de analizar fonológicamente estos mismos datos.

Es cierto que las diferencias no son tan grandes como las que se podían prever a partir del ensayo realizado con los clíticos pronominales, pero de todos modos consideramos que para describir adecuadamente la distancia lingüística entre un grupo de variedades cuanta más información pongamos en contraste más próxima a la realidad será la representación resultante.

Por otra parte, un análisis lingüístico de este tipo permite discriminar claramente el carácter lingüístico de los diferentes fenómenos que entran en juego en la variación. Nos permite introducir distinciones cualitativas en los resultados cuantitativos. Podemos discernir si la distancia obtenida está relacionada con los elementos subyacentes o con los procesos fonológicos, por ejemplo. Como se afirma en Viaplana (1994):

> La distinción entre elementos predecibles y elementos impredecibles en la estructura lingüística permite la discriminación de fenómenos diferenciales cruciales en la variación dialectal que, en ausencia de esta distinción, quedan amalgamados en la simple distinción de las formas.

En más de una ocasión se ha esgrimido como crítica a la descripción cuantitativa de la variación lingüística un cierto grado de menoscabo del análisis lingüístico. Así se ha señalado que una de las deficiencias que presentan algunos tratamientos cuantitativos de la variación dialectal tiene que ver con la falta de un análisis lingüístico

[8] En el sentido de Clua (1999a).

coherente previo a la transposición de los datos fonéticos 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 desde nuestro punto de vista el análisis lingüístico de los datos.

Referencias bibliográficas

Bonet, E. & M.-R. Lloret, 2005, «More on alignment as an alternative to domains: The syllabification of Catalan clitics», *Probus* 17, 1, 37-78.

Clua, E., 1999a, *Variació i distància lingüística. Classificació dialectal del valencià a partir de la morfologia flexiva*, tesis doctoral, Universitat de Barcelona.

—, 1999b, «Distància lingüística i classificació de varietats dialectals», *Caplletra* 26, 11-26.

—, 2007, «Distancia lingüística entre los dialectos del catalán a partir de los datos del COD», comunicación presentada en el *XXVième Congrès International de Linguistique et de Philologie Romanes***,** Innsbruck. (Aparecerá publicado en P. Danler *et alii* (eds.), *Actes du XXVième Congrès International de Linguistique et de Philologie Romanes (Innsbruck 2007)*, Niemeyer, Tübingen).

— & M.-R. Lloret, 2006, «New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)», en J-P. Montreuil (ed.), *New Perspectives on Romance Linguistics. Vol. 2: Phonetics, phonology, and dialectology*, Amsterdam/Philadelphia, John Benjamins, 31-47.

— & —, 2007, «Clasificación de variedades dialectales mediante técnicas de análisis multivariante, a partir de un corpus oral», en P. Cano López *et alii* (eds.), *Actas del VI Congreso de Lingüística General (Santiago de Compostela, 3-7 de mayo de 2004)*, vol. III: *Lingüística y variación de las lenguas*, Arco/Libros, Madrid, 3057-3068.

—, Valls, E. & J. Viaplana, 2008, «Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà», en G. Blaikner-Hohenwart *et alii* (eds.), *Ladinometria Festschrift für Hans Goebl zum 65. Geburtstag*, vol. 2, Universität Salzburg *et alii,* Salzburg, 27-42.

—; Lloret, M-R & E. Valls, 2009, «Análisis lingüístico y dialectométrico del *Corpus Oral Dialectal* (COD)», en P. Cantos Gómez & A. Sánchez Pérez (eds.) *A survey on corpus-based Research*, Asociación española de lingüística de corpus: 1033-1045.

Durand, J. P., 1889, «Notes de philologie rouergate», *Revue des langues romanes* 33, 47-84.

Lloret, M.-R., 2004, «The phonological role of paradigms: The case of insular Catalan», en J. Auger, C. Clements & B. Vance (eds.), *Contemporari Approaches to Romance Linguistics*, John Benjamins, Amsterdam/Philadelphia, 275-279

— & J. 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*, W. H. Freeman and Company, San Francisco.

Veny, J., 1982, *Els parlars catalans (síntesi de dialectologia)*, Moll, Palma.

Viaplana, J., 1999, *Entre la dialectologia i la lingüística*, Publicacions de l'Abadia de Montserrat, Barcelona.

—; Lloret, M.-R.; Perea, M.-P.; Clua, E., 2007, *COD. Corpus Oral Dialectal*. Barcelona, PPU. (Publicación en CD-rom).