La dialectometría y su aplicación en el estudio de las variedades dialectales del catalán

por Maria Pilar Perea
Universitat de Barcelona

Introducción

La metodología de análisis dialectométrico se ha aplicado recientemente en diversos proyectos centrados en el estudio de las variedades dialectales de las lenguas peninsulares.

En el País Vasco, Gotzon Aurrekoetxea (2004) y Charles Videgain han elaborado el «Atlas vasco de la Colección Bourciez», que se basa en el *Recueil des Idiomes de la Région Gasconne*, recopilado por el lingüista Edouard Bourciez. A partir de la transcripción de los textos correspondientes a diferentes versiones de la *Parábola del Hijo Pródigo*, se han analizado diversas características lingüísticas relacionadas con 151 conceptos y 112 rasgos gramaticales. La base de datos incorpora 52.904 palabras a partir de las cuales se han definido 12.499 variantes derivadas de los 151 conceptos aludidos. El fondo del mapa se construyó con el programa VDM (Visual DialectoMetry) y posteriormente el método dialectométrico de Hans Goebl, desarrollado en la Universidad de Salzburgo, se ha aplicado a los citados datos léxicos.

RFA 9-10 (2009/10) páx. 109-130 [ISSN: 1578-9853]

En Galicia, Rosario Álvarez, Francisco Dubert y Xulio Sousa (2006) han utilizado también el método de Goebl en el tratamiento de los datos del *Atlas linguistico galego*. Se seleccionaron 321 preguntas de 167 localidades, extraídas tanto del material editado como del inédito, las cuales representaban del modo más adecuado las particularidades fonéticas, morfológicas y léxicas del territorio.

Aplicando parcialmente la metodología y los instrumentos de análisis de Goebl, pero adoptando también algunos aspectos metodológicos de Francisco Moreno Fernández y de Pilar García Mouton, José Luis Aliaga Jiménez (2003) ha analizado el léxico de las hablas de Teruel. Los datos provienen del *Atlas lingüístico etnográfico de Aragón, Navarra y Rioja* (1979-1983).

Por otro lado, Ramón d'Andrés dirige el proyecto ETLEN (Estudiu de la Transición Llingüística na zona Eo-Navia) con el objetivo de dialectometrizar los datos dialectales encuestados en la frontera entre las lenguas gallega y asturiana. La novedad respecto a los proyectos anteriores es que los dados se obtienen a partir de una encuesta creada específicamente para esta finalidad.

Puesto que existen diversos proyectos de análisis dialectométrico que se han desarrollado fuera del ámbito peninsular, es interesante citar brevemente los trabajos que se llevan a cabo en Europa adoptando una aproximación metodológica distinta del método de Goebl. Es el caso de la escuela holandesa (en la Universidad de Gröningen), que se basa en el cálculo de la distancia de Levenstein para determinar las diferencias entre dos dialectos. Esta metodología se ha aplicado a los dialectos del holandés a través de los estudios de John Nerbonne y de Wilbert Heeringa (2001). Heeringa, junto con Charlotte Gooskens (2006), lo ha aplicado igualmente a los dialectos del noruego.

En Japón también se han desarrollado estudios relacionados con la dialectometría. Fumio Inoue y Chitsuko Fukushima (1997), por ejemplo, han

aplicado el análisis multivariante a datos dialectales del japonés y también al *Survey of English Dialects* (SED) para determinar distancias dialectales.

En este trabajo me centraré en el catalán, y concretamente en el análisis dialectométrico aplicado a los datos dialectales que se han extraído de *La flexió verbal en els dialectes catalans*, de Antoni M. Alcover. El corpus de referencia, obtenido mediante encuestas que se efectuaron entre los años 1906 y 1928, contiene casi medio millón de formas verbales.

La dialectometría y el estudio de las variedades dialectales del catalán

Desde los años setenta se han aplicado en los territorios de habla catalana diversos métodos de clasificación dialectal basados en criterios cuantitativos. Los resultados de diferentes propuestas han servido para cuestionar y replantear la división dialectal, centrada en la fragmentación tradicional –catalán oriental *versus* catalán occidental–, y para formular una nueva propuesta. Los primeros estudios fueron llevados a cabo por Sardà y Guiter (1975) y Guiter (1978) en áreas dialectales distintas.

La propuesta de clasificación de Sardà-Guiter (1975), basada en criterios científicos derivados de la aplicación de métodos dialectométricos, contiene, sin embargo, diversos aspectos discutibles. Por un lado, el uso de una fuente de información hasta cierto punto cuestionable, el *Atlas lingüístic de Catalunya* de Antoni Griera; y, por otro, la inserción del aranés, que no pertenece a los dialectos del catalán, y que muestra una distancia que lo separa de las hablas catalanas vecinas parecida a la que separa las variedades de Tarragona y Falset, dos localidades geográficamente muy próximas. Estas imprecisiones han hecho desconfiar de la bondad del método a los dialectólogos tradicionales.

Polanco (1992) ha aplicado el método dialectométrico a los datos catalanes del *Atlas lingüístico de la Península Ibérica*. El autor combina dos

metodologías: utiliza el método de intervalización de Guiter y el Índice General de Permeabilidad de Goebl. Los resultados coinciden a grandes rasgos con la clasificación dialectal tradicional (1992: 207); es decir, parecen confirmar la partición entre catalán oriental y catalán occidental.

Otras aplicaciones de técnicas dialectométricas, con la finalidad de determinar fronteras y de cuantificar las diferencias que presentan las variedades dialectales, han sido desarrolladas por Clua (1999) y Viaplana (1999). Los datos de partida han sido los del Corpus Oral Dialectal (COD)[1], desarrollado en el Departament de Filologia Catalana de la Universidad de Barcelona. Este corpus no se concibió inicialmente para su aplicación en estudios dialectométricos. Sin embargo, con ciertas adaptaciones, se han analizado según este método la conjugación regular del catalán occidental (Viaplana) y la morfología verbal y pronominal del valenciano (Clua). El modelo de cuantificación se basó en la elección de un índice de similitud utilizado en biología y en matemáticas. Actualmente se aplica este método, el de Goebl, y el de la Universidad de Gröningen, de manera paralela, a los datos restantes del COD.

Por otro lado, Xavier Casassas, colaborador de Hans Goebl en la Universidad de Salzburgo, está aplicando el tratamiento dialectométrico al *Atlas lingüístic del domini català,* aún en proceso de publicación.

El tratamiento dialectométrico de *La flexió verbal en els dialectes catalans*

Después de esta breve presentación, me centraré en la descripción del método dialectométrico de Hans Goebl aplicado a *La flexió verbal en els dialectes catalans*.

[1] Para una descripción del COD, cf. Lloret y Perea (2002).

En primer lugar, describiré los datos que he utilizado para la dialectometrización, indicaré a continuación el proceso seguido, utilizando el programa VDM, y finalmente señalaré ciertas dificultades que han surgido, así como, en algún caso, la posible manera de superarlas.

La descripción de los materiales

Los datos de *La flexió verbal en els dialectes catalans* constituyen, en conjunto, un compendio de morfología dialectal del catalán. Se trata de un proyecto iniciado en los primeros años del siglo XX por el fundador de la dialectología en Cataluña, Antoni M. Alcover (1862-1932), con la finalidad de recoger la conjugación completa de 75 verbos regulares e irregulares en 149 localidades del dominio lingüístico catalán.

Alcover inició las encuestas en 1906 y se tiene constancia de que aún en 1928 encuestaba (o completaba) la flexión verbal de la ciudad de Palma. Es cierto que, vinculado a otro proyecto que desarrollaba –el *Diccionari català-valencià-balear*–, Alcover visitó más de una vez algunas de las localidades seleccionadas, con lo cual a veces existe, aunque no de manera sistemática y general, información relativa al cambio lingüístico y también a ciertos aspectos sociolingüísticos: esto lo dicen los jóvenes o los viejos; es propio de clases cultas, son realizaciones populares, etc.

Es muy posible que en los 22 años de duración del proceso de encuestación (en realidad 16, porque la mayoría de las encuestas se acabaron en 1922) exista en algunos casos un posible desfase entre los datos encuestados. Cuando se editó la obra, entre 1929 y 1933, ya se hace constar en algunos casos la posible variación morfológica que puede surgir con el paso del tiempo.

Si estos materiales, por su tipología, difieren de los datos provenientes de los atlas, en los cuales la dialectometría se ha aplicado con más frecuencia, también discrepan en otros aspectos, como la tipología de

informantes que fueron encuestados: niños y jóvenes entre 10 y 14 años, y también en el método de encuestación. Pero la diferencia sustancial estriba en el hecho de que son muy frecuentes y abundantes las respuestas múltiples. Los atlas, sin embargo, contienen, mayoritariamente, respuestas únicas. Esto significa que los programas estadísticos que trabajan con este tipo de respuestas, como en el caso del VDM, no se adaptan a los datos múltiples, y, por lo tanto, son los datos los que deben adaptarse al programa; y eso es lo que, con ciertos costes, se ha intentado llevar a cabo.

Algunos de los aspectos metodológicos utilizados en las encuestas de la flexión verbal se describen en diversos documentos: en un buen número de diarios, redactados por Alcover, y en los cuadernos de campo. Allí constan las características de las localidades, los nombres y las edades de los informantes, diversos aspectos relacionados con la encuesta, siempre colectiva (en grupos de cinco o seis), las dificultades surgidas, la indistinción en la elección entre informantes de ambos sexos, etc.

En cuanto a la distribución de las 149 localidades, 25 pertenecen al dialecto rosellonés; 38, al catalán oriental; 34, al catalán occidental; 24, al valenciano; 30, a las Islas Baleares; y una localidad corresponde a la ciudad de Alguero en Cerdeña. De estas localidades, 51 eran rurales (de menos de 1.000 habitantes), 62 tenían carácter intermedio (entre 2.000 y 10.000 habitantes) y 26 eran urbanas (de más de 10.000 habitantes).

El uso de la dialectometría con los datos de Alcover no fue una de las primeras ideas que surgieron al trabajar con estos datos. De hecho, el primer proyecto fue la informatización de los materiales, es decir, la creación de una base de datos con las formas verbales.

Francesc de B. Moll, discípulo y continuador de diversos proyectos de Alcover, inició, entre los años 1929 y 1933, la edición del proyecto, titulado inicialmente «Estudi de la conjugació catalana», y, en el marco

de las publicaciones de la Oficina Romànica de Lingüística i Literatura, en aquel momento dirigidas por el gramático Josep Calveras, vieron la luz cuatro fascículos con los datos de la flexión verbal, en un total de 368 páginas. Faltaban, sin embargo los datos correspondientes a los verbos irregulares, que quedaron inéditos.

El objetivo de Moll, al editar *La flexió verbal*, fue sintetizar el conjunto de los datos, adaptándolos a su publicación y economizando su presentación. Cada localidad se asoció con un número, se estableció una transcripción ortográfica de referencia y se la vinculó con la transcripción fonética de cada persona verbal. A esta forma de referencia ortográfica, que no coincidía con el catalán estándar, se la ha llamado «forma arbitraria». Su función es hacer abstracción de la la realización fonética y, en realidad, se convierte en una especie de representación subyacente de la gramática generativa. La transcripción fonética, en la mayoría de casos, sólo recogía las desinencias verbales. La forma referencial era el infinitivo o, en el caso de los verbos de la segunda conjugación, el radical verbal. Con todo, y a pesar de su valor, el resultado editado en papel es de difícil consulta, y ello explica que haya sido tenido muy poco en cuenta en los estudios posteriores sobre la morfología verbal. La excesiva simplificación de la presentación ha impedido explotar todas sus posibilidades. En contrapartida, la versión informática consigue superar estas dificultades prácticas.

El tratamiento informático de los materiales

Cuando se inició la entrada de los materiales publicados en la base de datos (Microsoft Access 98), se pudo acceder también a los cuadernos de campo de Alcover, depositados en aquel momento, junto con todo su legado, en la Editorial Moll, en Palma de Mallorca. Estos cuadernos contenían todos los resultados de la encuesta sobre la conjugación, así

como una interesante introducción manuscrita, que habría tenido que figurar en el último fascículo, el cual no se llegó a publicar.

El acceso a los cuadernos de campo (existen doce, de un total de setenta y siete, que incluyen este tipo de información) ofreció la posibilidad de reconstruir los paradigmas, la flexión de los verbos irregulares y en algunos casos de completar la conjugación de algunas localidades que no se habían incluido en la selección efectuada en la obra editada. Se corrigieron también algunas erratas. La adición de los datos de los cuadernos incrementó el número de localidades a 170 (de 149) y el número de verbos (de 69) a 117 (aunque en algunos casos sólo se había documentado la conjugación incompleta de un verbo).

El conjunto de los datos llega casi a los 500.000 registros (470.255), y con ellos se puede trazar una panorámica del estado de la morfología del catalán de principios del siglo xx. La información permite desarrollar estudios sobre morfología verbal (y también de fonética), tanto desde un punto de vista diacrónico como sincrónico, a través de la comparación con datos actuales.

En 1999, los datos introducidos en la base de datos, que permitía un acceso rápido a los materiales, fueron publicados en dos volúmenes y en un CD-ROM por el Institut d'Estudis Catalans (cf. *Compleció i ordenació de la flexió verbal en els dialectes catalans d'A. M. Alcover i F. de B. Moll*).

El segundo proyecto relacionado con los datos de Alcover tenía como objetivo efectuar la cartografía automatizada de las formas verbales. Aunque *La flexió verbal* no se concibió como un atlas lingüístico, los materiales son tan sistemáticos y completos que son idóneos para ser presentados en forma de atlas, en este caso de tipo morfológico. El 2001 se publicó en CD-ROM, *La flexió verbal en els dialectes catalans d'A. M. Alcover i F. de B. Moll. Les dades i els mapes*, que contiene tanto la base de datos previa

como el programa de cartografía. La última actualización en CD-ROM se publicó en 2005, con el título de *Antoni M. Alcover. Dades dialectals*, y a estos datos se ha añadido un programa informático que gestiona los materiales dialectales incluidos en 65 cuadernos de campo[2].

Desde el punto de vista de *La flexió verbal*, el atlas se concibe como una compilación de mapas individuales, cada uno de los cuales muestra la distribución geográfica de una determinada forma morfológica correspondiente a la conjugación de uno de los 117 verbos de que consta el programa. El proceso de cartografía de los datos ha resultado de transferir la estructura de la base de datos a la representación topográfica de los registros.

El programa permite la elaboración de mapas y listas (Figura 1). Esta presentación simplifica los resultados y facilita su comparación. En realidad, esta es la única manera de mostrar cartográficamente los datos, puesto que en papel, en forma de atlas convencional, el número total de mapas superaría los 6.000.

Aunque se trata de un atlas descriptivo, las representaciones simbólicas y fonéticas muestran la formación de áreas y subáreas dialectales, informan sobre la acción del proceso de cambio lingüístico, a través del polimorfismo de resultados, y ofrecen datos representativos para su posterior estudio e interpretación.

Precisamente la posibilidad de interpretación, además de los estudios sincrónicos o diacrónicos que se llevan a cabo, basados en la variación de la morfología verbal, es lo que impulsó a llevar a cabo el tercer proyecto. Se trataba de aplicar metodologías cuantitativas y estadísticas a este gran corpus de datos.

---

[2] Ahora los materiales pueden consultarse en internet en el portal Antoni M. Alcover de l'Institut d'Estudis Catalans: http://alcover.iec.cat/.

La dialectometrización de los materiales

Metodológicamente, para la dialectometrización de los datos, se ha utilizado el programa VDM, desarrollado por Hans Goebl en la Universidad de Salzburgo. Este tratamiento permite reexaminar los materiales desde un punto de vista distinto al meramente descriptivo. Puesto que se contaba con un número de datos muy elevado, en un primer momento el estudio se limitó a las formas verbales correspondientes a la primera conjugación. Posteriormente se procedió a aplicar el método al conjunto total de los datos.

Como es sabido, el objetivo primordial de la geografía lingüística ha sido la creación de atlas dialectales. La unidad de clasificación dialectal en un mapa es la isoglosa. Así, en cada uno de los mapas de *La flexió verbal* es posible trazar fronteras dialectales que muestran el inicio y el final del uso de una determinada forma verbal u observar áreas donde resultados idénticos se solapan. El problema es que esta es una manera, de las distintas posibles, de representar la realidad. En geografía lingüística tradicional estudios simultáneos del conjunto total de los datos son irrealizables.

Por el contrario, la dialectometría analiza la realidad desde una perspectiva global y generalizadora y evita así los problemas debidos a las idiosincrasias de datos particulares (Goebl, 2003: 61). Esta metodología remarca agrupaciones internas y estructuras de los datos lingüísticos que la simple observación o que un enfoque meramente descriptivo no pueden indicar. Goebl (2003: 61) señala además que la dialectometría permite mostrar las estructuras profundas derivadas de las superficiales.

Cuando se tiene acceso a un número elevado de datos, los procedimientos estadísticos, especialmente las metodologías que permiten las clasificaciones y visualizaciones de estructuras superiores en redes

lingüísticas, son absolutamente necesarios. Y eso es lo que se planteó cuando se emprendió la dialectometrización de los datos de Alcover.

La dialectometría, a pesar del número elevado de datos con que trabaja, permite, a través de la clasificación numérica, la simplificación de estos datos y a la vez da respuesta a cuestiones que, sin este tratamiento, sólo se pueden examinar parcialmente. En catalán, en particular, y con estos materiales, nos cuestionábamos si se podía mantener la división dialectal en los seis dialectos actualmente en uso, o si la tan discutida y primigenia división dialectal entre catalán oriental y catalán occidental estaba aún vigente, al menos desde el punto de vista de la morfología verbal.

La clasificación de los datos que realiza el análisis dialectométrico se basa en la sustitución del concepto de isoglosa como unidad básica de la clasificación dialectal por el concepto de distancia lingüística. Así, mientras la isoglosa se concibe como una línea ideal que señala, en un mapa, el límite entre la presencia o la ausencia de un rasgo determinado, la distancia lingüística se relaciona con la cuantificación de las similitudes entre las realizaciones lingüísticas de dos variedades dialectales que pertenecen a una misma lengua o bien que se hallan en áreas de transición.

Ampliando los trabajos realizados por Jean Séguy y Henri Guiter, Goebl ha construido un base teórico-instrumental para la taxonomía numérica; es decir en la representación esencial de la variación de los datos originales por medio de un conjunto de unidades taxatorias (taxados). Este procedimiento se ha mejorado progresivamente a través de la experiencia obtenida aplicando su método a diversos atlas lingüísticos: el *Atlas linguistique de la France* (ALF), el *Atlante italo-svizzero* (AIS), el *Survey of english dialects* (SED) o el *Atlante linguistico del ladino dolomitico e dei dialetti limitrofi*. Posteriormente se ha cedido el programa a otros proyectos, como los citados relativos al vasco y al gallego, al *Atlante lessicale toscano* (ALT) o a los datos del *Reeks nederlandse dialectatlassen*,

y se ha desarrollado un conjunto de métodos de visualización distintos (los llamados mapas de rayos, mapas de isoglosas, mapas de densidad o mapas de síntesis), o en forma de árboles multicromáticos (dendrogramas). Los resultados obtenidos son de fácil y atractiva visualización.

El proceso que se ha seguido para aplicar el método dialectométrico de Hans Goebl a los datos dialectales de *La flexió verbal* ha constado de cuatro fases:

1. La preparación cartográfica de los materiales, que se llevó a cabo a través de la construcción de una red poligonal apropiada, es decir, de los polígonos de Thiessen, aplicando los principios de la geografía de Delaunay-Voronoi.

La flexió verbal tiene 342 lados de polígono. Una vez dibujados, estos polígonos constituyen el soporte adecuado para las representaciones dialectométricas. Las isoglosas se trazan con la ayuda de un número determinado de polígonos.

2. La aplicación del método dialectométrico a los datos de *La flexió verbal* requirió un tratamiento informatizado. Para su dialectometrización, se desestimaron los datos correspondientes a conjugaciones incompletas o a localidades con un número de respuestas muy reducido. Se utilizaron las 149 localidades iniciales publicadas en la obra editada, que requerían idealmente, para su tratamiento, 149 respuestas. En el caso en que el número de datos fuera menor, estos no podrían ser inferiores a 30. En conjunto, se aplicó el método a una red de 149 localidades y a 75 verbos en total. Y el resultado fueron 6.000 «mapas de trabajo» (Figura 2).

En la versión actual, la visualización de las formas fonéticas que aparecen como taxados no es correcta. Es necesario normalizarlas, puesto que, en la informatización de *La flexió verbal*, los símbolos fonéticos se representaron mediante la fuente SILDOULOS IPA 93, pero estos caracteres no son reconocibles por el programa VDM.

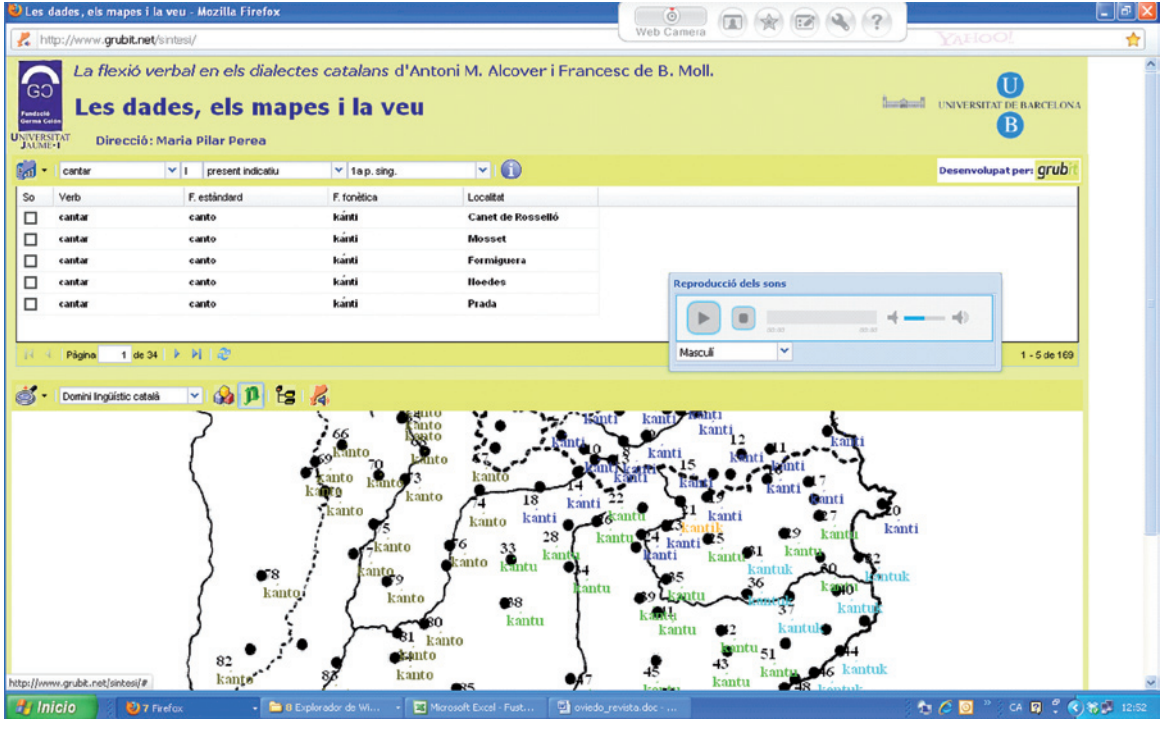

Figura 1: Mapa y lista generados por el programa de cartografía automática en internet (http://alcover.iec.cat/).

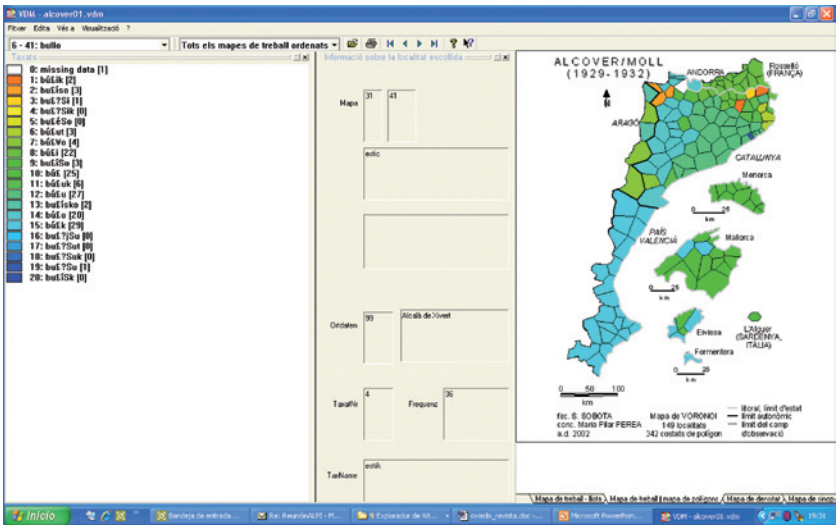

Figura 2: Uno de los mapas de trabajo.

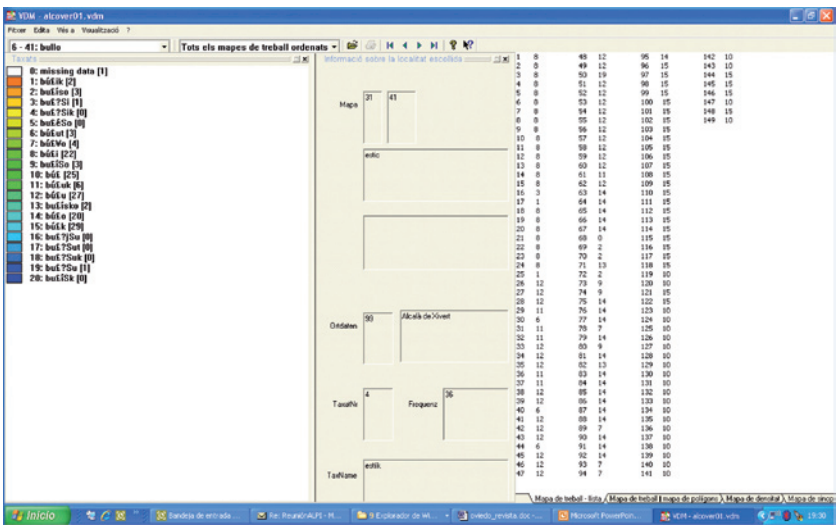

Figura 3: Matriz de datos.

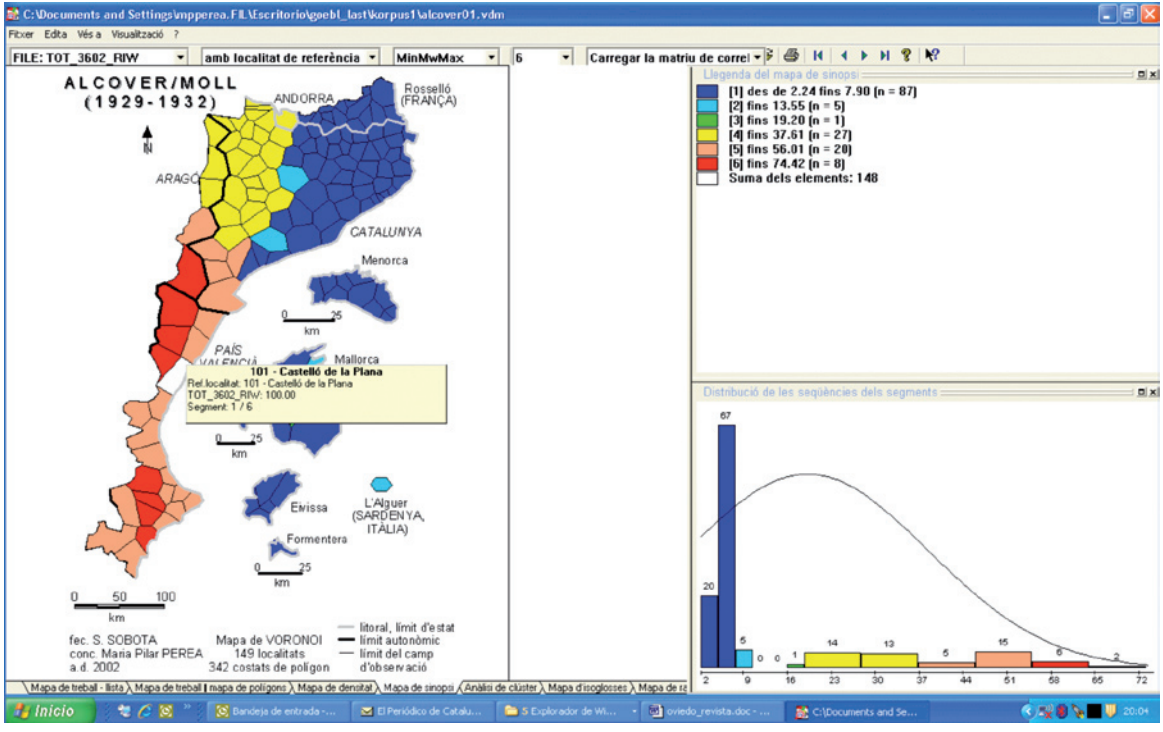

FIGURA 4: Cálculo de la distancia RIW en Castelló de la Plana.

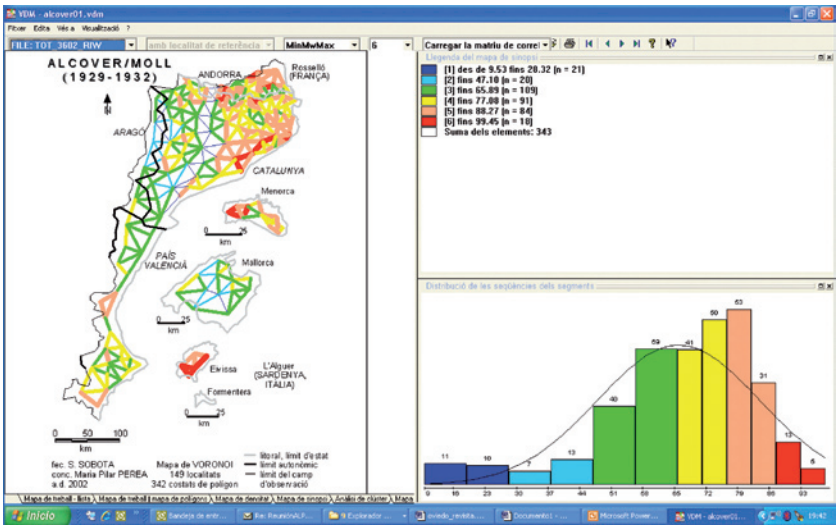

Figura 5: Mapa de rayos.

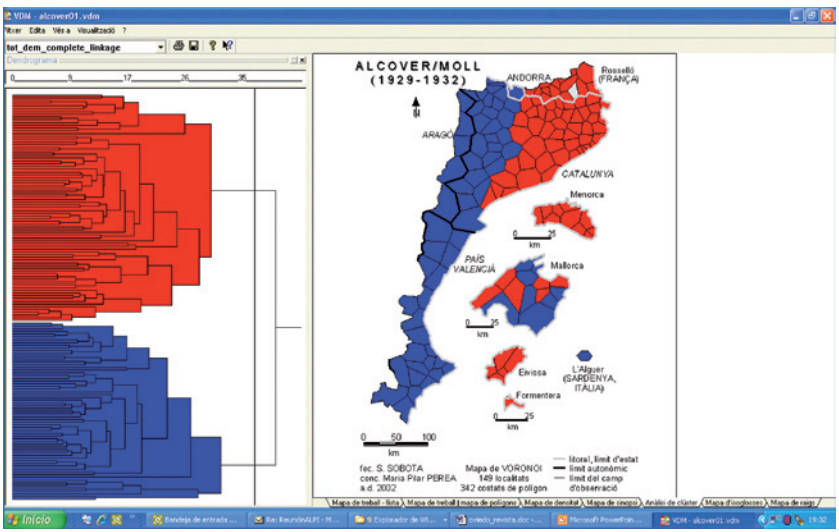

Figura 6: Dendrograma.

Antes de explicar cómo se han aplicado los procedimientos matemáticos a los datos de Alcover, cabe hacer tres consideraciones con relación a los materiales:

a) Las respuestas cero. Según Goebl (1997: 23), los datos ausentes son factores que distorsionan el tratamiento numérico y clasificatorio. En *La flexió verbal* hay algunas formas sin respuesta. A menudo la ausencia de datos se debe a la ignorancia de ciertos verbos que no son de uso general en todos los territorios de habla catalana (es el caso, por ejemplo, del verbo *caldre* 'ser necesario'). También pueden deberse a olvidos y errores.

b) Las respuestas múltiples, que constituyen realmente un problema. A diferencia de los atlas lingüísticos románicos, los datos de Alcover muestran respuestas múltiples en ciertas localidades. Esto es debido a la variación dialectal existente, que permite la convivencia entre formas tradicionales y formas innovadoras. El problema surge cuando estas respuestas múltiples no pueden incorporarse al programa VDM. En este caso es necesario elegir una forma que sea la más representativa para cada localidad. Para seleccionar respuestas únicas entre un conjunto de respuestas múltiples hay dos posibilidades:

1. dejar que el programa elija una respuesta al azar (esto constituye un error puesto que entonces no existe ningún criterio de selección y pueden mezclarse datos de carácter antiguo con otros más modernos);

2. hacer una selección de la forma que puede considerarse realmente representativa de la localidad, pero esta elección presenta dificultades, puesto que se desconoce lo que sucedía en la morfología verbal de hace cien años. Cabe indicar que cuando existen respuestas múltiples siempre puede observarse una sistematicidad entre formas antiguas y formas innovadoras. Así se sabe que los presentes de subjuntivo terminados en *-a* son más antiguos que los terminados en *-i* (*bega* versus *begui*, por ejemplo, 'beba'). Desde esta perspectiva y para solventar parcialmente

el problema, se llevó a cabo una selección manual de los datos, poniéndoles un índice (del 0 al 3; por lo que se elegían sólo cuatro respuestas, y se desestimaban, si existían, las soluciones restantes). Estos índices indicaban gradualmente el paso de las formas más antiguas a las más modernas, las cuales experimentaban a veces la acción de procesos analógicos. Así, en Organyà (una localidad del catalán occidental), la alternancia que muestra la 1.ª persona del plural del imperativo ('estemos') es la siguiente:

estem	estém	3
estiguem	astiɣém	4
estigam	astiɣám	2
estam	astám	1

Establecer esta gradación ha supuesto aplicar el programa cuatro veces a 4 subcorpus, teniendo en cuenta que los subcorpus 3 y 4 contienen un número importante de formas sin respuesta, porque no en todas las localidades se encontraban formas con cuatro variantes. La gradación se establecía tanto en casos de discrepancia morfológica como en casos de discrepancia fonética. Así, la 2.ª persona del condicional en Ribes de Freser (una localidad del catalán oriental) presenta una doble forma, que muestra una alternancia meramente fonética:

entendries	əntəndríës	2
entenries	əntənríës	1

Con la aplicación de este proceso de selección cuádruple, una parte de las variaciones dialectales se ha eliminado. Cabe considerar que en geografía lingüística, las respuestas múltiples son consideradas un problema cualitativo (Goebl, 1997: 23), pero es necesario que la dialectometría

sepa convertir este problema cualitativo en relaciones cuantitativas. Hay que tener en cuenta que la ausencia de respuestas múltiples distorsiona hasta cierto punto el valor de los resultados obtenidos.

c) La presión de la lengua estándar. Examinando otros trabajos de Goebl de tipo dialectométrico puede observarse la inserción de un punto de referencia artificial, que representa la forma estándar. En la dialectometrización de *La flexió verbal* no se ha añadido ningún punto de este tipo, ya que la forma estándar podia equipararse a los resultados obtenidos en Barcelona.

3. La tercera fase de la dialectometrización ha consistido en el proceso de adaptación de los materiales originales de Alcover (en Microsoft Access) para conseguir la taxación o clasificación de los datos. Aunque existe un programa de clasificación automático, en este caso el proceso se ejecutó manualmente. La taxación supone la agrupación de los resultados obtenidos en un mapa dialectal en tipologías. Cada mapa dialectal taxado crea un vector en una matriz de datos.

Puesto que la estructura de la base original de *La flexió verbal* estaba formada por 14 campos, fue necesario dar un número a todos los campos. Sin necesidad de eliminar ningún dato, el resultado fueron columnas con una secuencia de ocho dígitos, que representaban cada uno de los «mapas de trabajo». El número completo de dígitos representa las cifras asociadas con el número de verbo, el número de conjugación, el número de tiempo y el número de persona. Otras dos cifras adicionales están asociadas, respectivamente, con las cifras que enumeran las variantes morfológicas, que están vinculadas a la representación fonética; y el número de localidad, que coincide con la cifra indicada por Alcover.

4. Después de la codificación, la cuarta fase se desarrolló en la Universidad de Salzburgo. Goebl utiliza el llamado *distillation method*, que supone la construcción de una base de datos menor, la cual contiene las cifras indicadas anteriormente. Esta subbase fue la que se incorporó al

programa VDM. El programa, de tipo taxométrico y cartográfico, fue creado por Edgar Haimerl y ofrece una lista de algoritmos que facilitan el análisis de los datos (Figura 3)[3].

El primer resultado obtenido con el programa fue una matriz de datos. Esta matriz muestra una representación estructural de *n* elementos (puntos de la encuesta) y *p* (rasgos lingüísticos, comprendidos en cada uno de los mapas de un atlas). La matriz de datos puede ser computada en valores numéricos de similitud usando diversos índices. Los valores resultantes se guardan en una matriz de similitud (*n* x *n* = dialectos x dialectos), que es usada para posteriores análisis dialectométricos, por ejemplo, los dendrogramas y la visualización de similitudes.

El programa permite desarrollar un conjunto de cálculos de la matriz de datos a la matriz de similitud y a la matriz de distancias:

1. El cálculo de similitud entre variables. Hay distintos cálculos de similitud, cada uno de los cuales define la similitud entre localidades. Uno de ellos es el RIW (Relative Identity Value), parecido al coeficiente de combinaciones, pero modificado para incluir datos ausentes; otro es el GIW (Weighted Identity Value), en el cual se considera la frecuencia de la taxaciones individuales.

2. La clasificación numérica. Las matrices de similitud pueden visualizarse directamente a través del VDM. Una forma simple constituye un Perfil de similitud: una fila de la matriz de similitud contiene la similitud dialectal de la localidad de referencia (por ejemplo, Castelló de la Plana, *La flexió verbal* Nr. 101, cálculo de la distancia RIW) con respecto a las otras localidades (Figura 4). Localidades vecinas con una alta similitud comparten el mismo color rojo en el mapa; localidades más alejadas, en naranja (que indican áreas de transición), en amarillo, etc., a medida

[3] En http://ald.sbg.ac.at/dm/ se encuentran las características del programa y el software para su aplicación.

que se distancian, y las localidades que presentan la similitud más baja en un cromatismo que acaba en el azul oscuro. El punto de referencia es blanco. Los otros colores simbolizan diferencias cualitativas.

El programa VDM ofrece diferentes algoritmos de segmentación (MINMWMAX, MEDMW, MED, MMINMWMAXX), que son necesarios para definir la gradación cromática. El programa tiene una gradación de diez colores, los cuales comprenden entre 2 y 20 intervalos numéricos, que pueden variarse según la elección del usuario. El programa facilita también la visualización inmediata de los otros índices de similitud en la matriz respectiva. Así, cualquier localidad puede ser seleccionada y comparada con las otras. Desde esta perspectiva pueden definirse estructuras espaciales que se corresponden con áreas dialectales y pueden delimitarse áreas extensas a través de la asociación de localidades vecinas inmediatas.

Además de mapas de sinopsis, de densidad o de isoglosas, que visualizan de distintas maneras las representaciones gráficas de áreas dialectales, se muestran a continuación los mapas de rayos y los dendrogramas. Hay que tener en cuenta que las isoglosas dialectométricas no son equivalentes a las isoglosas tradicionales, sino que tienen carácter cuantitativo. La variación en el grosor en la medida de los polígonos indica la «disimilitud» o los valores de distancia.

El mapa de rayos (Figura 5) es el resultado del cálculo de la conexión interpuntual, que constituye la inversión lógica del mapa de isoglosas. La función del mapa de rayos no es delimitar sino más bien concentrar o contactar. La concentración de rayos gruesos de color rojo indica la aplicación de determinados algoritmos que simbolizan contactos intensos entre localidades. Las áreas que presentan contactos débiles se representan del modo inverso.

Los histogramas que aparecen al lado de la red poligonal de los diversos mapas son útiles para visualizar, en el mapa, el carácter estadístico de

la frecuencia de la distribución dialectométrica a través de dos columnas verticales. Se calcula a través de un algoritmo que siempre crea dos columnas, las cuales son equivalentes al mismo número de intervalos numéricos. La proyección de la curva de Gauss, que se obtiene a través de la media aritmética y la desviación estándar de la distribución de la frecuencia correspondiente en el histograma, remarca algunas particulares estadísticas.

Finalmente, aplicando métodos dendrográficos, el análisis de conglomerados usa jerarquías de clasificación ascendente, mediante la selección de diversos algoritmos. Unos procesos de clasificación adecuados generan árboles jerárquicos de clases disyuntivas. El resultado es un árbol genealógico (árbol de WARD), que es útil para describir la relación entre dialectos. Los resultados se pueden asociar con colores, habiendo seleccionado previamente un determinado número de conglomerados. En este caso (Figura 6), como indican los colores, y a partir de los cálculos aplicados, puede verse que no se puede establecer una división clara, desde un punto de vista morfológico, entre las variedades dialectales orientales y occidentales.

Según Goebl, aún son posibles dos interpretaciones más de las jerarquías: una sincrónica, que está relacionada con la relación espacial; y otra diacrónica, que incorpora aspectos temporales asociados con la relación espacial.

Conclusiones

Aunque todavía se está trabajando en la interpretación de los datos, este artículo ha mostrado que, pese a algunos problemas, la dialectometría no sólo trata los datos de manera global –una cosa imposible de efectuar con la representación individual de los mapas– sino que es útil para determinar y clasificar objetivamente los principales dialectos y subdialectos del catalán desde un punto de vista morfológico.

Además de cuestiones de presentación de los datos fonéticos, que no pueden ser adaptados en la visualización del programa, quedan aún diversas cuestiones a tener en cuenta, especialmente la problemática del tratamiento de datos múltiples. Para paliar estos efectos se ha iniciado una colaboración con Hiroto Ueda, profesor de la Universidad de Tokio, con el que se tratará numéricamente el polimorfismo de las respuestas[4].

Aunque algunos lingüistas contemplan con cierto escepticismo la dialectometría y sus resultados, es indudable que esta metodología constituye un procedimiento apto para tratar datos cuantitativamente numerosos y mostrar sus relaciones espaciales. Puesto que la realidad puede ser estudiada desde diversos puntos de vista, esta metodología, además de contrarrestar las críticas que ha sufrido la lingüística por no aplicar un método científico en sus estudios, ofrece efectivamente unas técnicas de clasificación numérica que son idóneas para progresar en el conocimiento de una realidad dialectal compleja. Así lo hemos experimentado con su aplicación a los datos de *La flexió verbal*.

BIBLIOGRAFÍA

Aliaga Giménez (2003) = José Luis Aliaga Jiménez, «Dialectometría y léxico en las hablas de Teruel», *ELUA, Estudios de Lingüística*, 17 (2003), págs. 25-55.

Aurrekoetxea (2004) = Gotzon Aurrekoetxea, «El atlas lingüístico vasco: 20 años de innovación tecnológica», in M. P. Perea, *Dialectologia i recursos informàtics*, Barcelona (PPU), 2004, págs. 15-41.

Álvarez, Dubert y Sousa (2006) = Rosario Álvarez, Francisco Dubert y Xulio sousa, «Aplicación da análise dialectométrica aos datos do *Atlas lingüístico galego*», *Lingua e territorio* (2006), págs. 461-493.

[4] Uno de los primeros resultados se encuentra en Perea & Ueda (2010).

Clua (1999) = Esteve Clua, *Variació i distància lingüística*, Tesis Doctoral presentada al Departament de Filologia Catalana de la Universitat de Barcelona, 1999.

Goebl (1997) = Hans Goebl, «Some Dendographic Classifications of the Data of CLAE 1 and CLAE 2», in W. Viereck and H. Ramisch (eds.), *The Computer Developed Linguistic Atlas of England 2*, Tübingen (Max Niemeyer Verlag), 1997, págs. 22-32.

Goebl (2003) = Hans Goebl, «Regards dialectométriques sur les donnés de l'Atlas Linguistique de la France (ALF): Relations quantitatives et structures de profondeur», *Estudis Romànics* XXV (2003), 61-117.

Gooskens & Heeringa (2006) = Charlotte Gooskens & Wilbert heeringa, «The Relative Contribution of Pronunciational, Lexical, and Prosodic Differences to the Perceived Distances between Norwegian Dialects», *Literary and Linguistic Computing*, 2006, págs. 1-16.

Guiter (1978) = Henri Guiter, «Panorama lingüístic des de Besalú», *Annals del Patronat d'Estudis Històrics d'Olot i Comarca*, 1978, págs. 35-48.

Heeringa & Nerbonne (2001) = Wilbert Heeringa, & John Nerbonne, «Dialect Areas and Dialect Continua», in David Sankoff, William Labov and Anthony Kroch (eds.), *Language Variation and Change,* New York (Cambridge University Press), 2001, vol. 13, págs. 375-400.

Inoue & Fukushima (1997) = Fumio Inoue & Chitsuko fukushima, «A quantitative approach to English dialect distribution: Analyses of CLAE morphological data», in Wolfgang Viereck and Heinrich Ramisch (eds.), *Computer developed linguistic atlas of England (CLAE)*, Tübingen (Max Niemeyer Verlag), 1997, vol. 2, págs. 57-65.

Lloret & Perea (2002) = M. Rosa Lloret, & M. Pilar Perea, «A report on 'The Corpus Oral Dialectal del Català Actual (COD)», *Dialectologia et Geolinguistica* 10 (2002), págs. 59-76.

Perea (2005) = M. Pilar Perea (ed.), *Dades dialectals. A. M. Alcover*, Palma (Conselleria d'Educació i Cultura, Govern de les Illes Balears), 2005.

Perea (2003) = M. Pilar Perea, «Dialectometry: A New Treatment of Dialectal Morphological Data», *Linguistica Atlantica* 27-28 (2003), págs. 86-91.

Perea (2008) = M. Pilar Perea, «Catalan verb morphology and dialectometric analysis», in G. Blaikner Hohenwart, E. Bortolotti, E. Lörincz, *Ladinometria*, *Miscellanea per Hans Goebl per il 65° compleanno*, 2008, vol. 2, págs. 61-77.

Perea & Ueda (2010) = M. Pilar Perea & Hiroto Ueda, «Applying Quantitative Analysis Techniques to "La flexió verbal en els dialectes catalans"», *Dialectologia et Geolinguistica*, 18 (2010), págs. 3-11.

Polanco (1992) = Lluís B. Polanco, «Llengua i dialecte: una aplicació dialectomètrica a la llengua catalana», *Miscel·lania Sanchis Guarner*, Barcelona (Publicacions de l'Abadia de Montserrat), 1992, v. III, págs. 5-28.

Sardà & Guiter (1975) = A. Sardà & H. Guiter, «L' *Atlas lingüístic de Catalunya* i la fragmentació dialectal del català», *Miscellania Barcinonensia* XL, (1975), págs. 93-112.

Viaplana (1999) = Joaquim Viaplana, *Entre la dialectologia y la lingüística. La distància lingüística entre les varietats del català nord-occidental*, Barcelona (Publicacions de l'Abadia de Montserrat), 1999.