clac CÍRCULO
de
lingüística
aplicada a la
comunicación

24/2005

LOS CORPUS SONOROS EN LA INVESTIGACIÓN DE LA LENGUA HABLADA

José Antonio González Salgado
jantoniosalgado EN yahoo es

Resumen

Las modernas tecnologías están modificando el panorama de la investigación de la lengua hablada. En este artículo se ofrece un acercamiento al avance que suponen los corpus sonoros en la investigación dialectal. Para ello repasamos los principales proyectos que, utilizando una metodología similar, se están desarrollando actualmente en España.

Palabras clave: corpus orales, metodología dialectal, lingüística informática

Abstract

Modern technologies are changing the field of research in spoken language. This paper offers a first report of advances in dialect research through sound corpora. It surveys the main projects being carried out now in Spain with a common methodology.

Key words: oral corpora, dialectological methodology, computer linguistics

En los últimos años, el interés por recopilar muestras orales de hablas, dialectos y lenguas amenazadas de extinción es creciente[1]. A ello ha contribuido, sin duda, el avance tecnológico, con las posibilidades que ofrece de preservar, en formatos que no sucumben al paso del tiempo, las muestras reales que se toman. La geografía lingüística tradicional, en la que un explorador tenía que transcribir de oído las respuestas de los informantes, también ha visto en las nuevas tecnologías su tabla de salvación. Los atlas lingüísticos clásicos, en papel de gran formato, con la transcripción fonética de los resultados del interrogatorio, van dejando paso a proyectos en los que tenemos también la posibilidad de oír, en la viva voz de los informantes, las respuestas al cuestionario. El formato de publicación de estos atlas lingüísticos ya no es el incómodo papel, sino la edición electrónica, bien en soportes como el CD-Rom o el DVD, bien en la red de internet. El *Atlas Multimédia Prosodique de l'Espace Roman* (AMPER)[2], el *VIVaio Acustico delle Lingue e dei Dialetti d'Italia* (VIVALDI)[3], el *Atlas Linguistique Parlant d'une Région Alpine*[4] y el *Atlas Linguistique Audiovisuel du Valais Romand*[5] son los precursores de un modelo que ya empieza a tener resonancias en distintos espacios geográficos[6]. El formato electrónico, además, también comienza a ser utilizado para difundir atlas construidos con los parámetros clásicos, fundamentalmente porque los costes de la edición informática son mucho menores que los de la edición en papel; pero también porque los atlas electrónicos llegan donde los libros nunca podrían haber tenido acceso: internet permite la consulta de miles de mapas lingüísticos y etnográficos sin necesidad de recurrir a bibliotecas o de dedicar un gran espacio en las estanterías para guardar tan ingentes obras[7].

Además de los atlas lingüísticos, que al fin y al cabo lo que han hecho ha sido aclimatarse a los tiempos modernos, existen otros proyectos que sí son resultado directo de los avances tecnológicos: los archivos orales sonoros. Dentro de estos últimos

[1] Agradezco a la Dra. Inés Fernández-Ordóñez (Directora del *Corpus Oral y Sonoro del Español Rural*) la lectura del borrador de este artículo.
[2] http://www.ii.ua.pt/cidlc/gcl/fr/atlas/default.asp
[3] http://www2.hu-berlin.de/Vivaldi/
[4] http://www.u-grenoble3.fr/ellug/livres/atlas/
[5] http://www.unine.ch/dialectologie/AtlasPresent.html
[6] El 22 de mayo de 2005, en *El Faro de Vigo*, se publicó la noticia del proyecto de realización de un atlas lingüístico sonoro de las hablas gallego-lusas, dirigido por Xosé Henrique Costas González. El atlas, el primero de esas características que se planifica en la Península Ibérica, podrá ser consultado tanto en formato DVD como en una página web.
[7] La lista de los atlas lingüísticos accesibles actualmente a través de la red se puede consultar en http://www.geolectos.com/atlas.htm

conviene diferenciar los proyectos que suponen una recopilación de textos de referencia de distintas lenguas, más vinculados a la lingüística informática y, por tanto, con aplicaciones en el campo de la tecnología del habla, el reconocimiento de voz o la enseñanza de segundas lenguas, de los proyectos que se basan en la recogida de muestras orales de carácter dialectal, que persiguen la constitución de amplias bases de datos de etnotextos y archivos de sonidos con los que poder estudiar la variación diatópica (también la diastrática) en una región o país determinado. Mi interés se va a centrar en estos segundos modelos de investigación del habla (los corpus sonoros dialectales), aunque es necesario hacer una breve mención de los proyectos anteriores, porque a raíz de ellos pudieron surgir los archivos dialectales electrónicos y porque algunas cuestiones metodológicas aplicadas en los corpus dialectales tienen su origen en las características que presentan los proyectos de procesamiento del lenguaje natural.

LOS CORPUS AUTOMATIZADOS DE COMUNICACIÓN HABLADA

El campo que abarca el modelo de investigación de la comunicación hablada es amplísimo. Como Joaquim Llisterri[8] y otros autores han señalado en alguna ocasión, dentro de los corpus orales conviene diferenciar los que suponen "transcripciones de producciones lingüísticas más o menos espontáneas" de los que "ofrecen materiales mucho más controlados"[9]. Los segundos tienen especial aplicación en el desarrollo de sistemas de reconocimiento de voz (identificaciones de lengua y locutores), traducciones automáticas o sistemas de diálogo, y están especialmente diseñados para el mundo empresarial (telefonía e informática).

La lista de proyectos materializados o en curso de archivos orales es muy extensa[10]; ciñéndonos solo a los que recogen muestras del habla espontánea, cabe destacar los trabajos incluidos en el *C-Oral-Rom Project* (Integrated Reference Corpora

[8] Remito a los interesados en la materia a la extensa bibliografía recopilada por el Dr. Joaquim Llisterri, así como a los trabajos que ha desarrollado este investigador en el campo de las tecnologías del habla, la fonética experimental y la oralidad en general, todo ello disponible en http://liceu.uab.es/~joaquim/

[9] J. Llisterri, M.J. Machuca, C. de la Mota, M. Riera, A. Ríos, "Corpus orales para el desarrollo de las tecnologías del habla en español", Oralia. Análisis del discurso oral, 8, 2005 (en prensa). Disponible en internet en la dirección http://liceu.uab.es/~joaquim/curriculum/papers.html

[10] Véanse, a modo de ejemplo, la Bibliotèque de Corpus (http://elsap1.unicaen.fr/smonnier/corpus.html) del Laboratorio de Lingüística de la Universidad de Caen o los enlaces que propone Joaquim Llisterri sobre el particular (http://liceu.uab.es/~joaquim/).

for Spoken Romance Languages)[11], coordinado desde la Università degli Studi di Firenze y centrado en la recopilación de muestras orales de italiano, francés, portugués y español[12]; el proyecto *Português Falado: Variedades Geográficas e Sociais*, de gran interés para la dialectología, editado en cuatro CD-Rom por el Círculo de Lingüística de la Universidad de Lisboa y el Instituto Camões[13]; el proyecto *Français parlé: corpus et recherche dans les pays nordiques*, coordinado por lingüistas de Dinamarca, Noruega, Suecia y Finlandia[14]; o, en español, el *Corpus de Referencia de la Lengua Española Contemporánea* –dirigido por Francisco Marcos Marín– que es el título en que se inscriben tres proyectos distintos: el corpus argentino, el chileno y el corpus oral peninsular[15].

LOS CORPUS DE INTERÉS DIALECTAL

Hasta hace relativamente poco tiempo, el acceso a muestras reales de las hablas dialectales era una tarea casi imposible. Los dialectólogos teníamos que conformarnos con el estudio de los textos en transcripción fonética que nos proporcionaban las monografías y otros trabajos, perdiendo con ello todos los ricos matices que presentan los elementos suprasegmentales (acento y entonación) en determinadas regiones e ignorando, en la mayor parte de los casos, el contexto en que se originaron las grabaciones que dieron lugar a los textos transcritos. De igual manera que la sociolingüística revolucionó el panorama de la dialectología, hasta el punto de que hoy no tienen mucho sentido las monografías de las hablas populares en que no se atiende el componente social del lenguaje, las nuevas tecnologías modificarán (lo están haciendo ya) los trabajos dialectales, y llegará el día en que será impensable presentar un estudio sobre un habla determinada sin ofrecer las muestras orales que se hayan tomado.

---

[11] Véanse, a modo de ejemplo, la *Bibliotèque de Corpus* (http://elsap1.unicaen.fr/smonnier/corpus.html) del Laboratorio de Lingüística de la Universidad de Caen o los enlaces que propone Joaquim Llisterri sobre el particular (http://liceu.uab.es/~joaquim/).

[12] Evidentemente, me voy a referir solo a aquellos proyectos que presentan las grabaciones conseguidas –o, al menos, una parte de ellas– y no a los que suponen una transcripción de registros orales. Para estos últimos véase M. Chantal Pérez Hernández, "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento", en Estudios de Lingüística Española, n.º 18, 2002 (http://elies.rediris.es/elies18/).

[13] También son de destacar los registros sonoros de los dialectos portugueses que incluye el Instituto Camões en su página web (http://www.instituto-camoes.pt/cvc/hlp/geografia/mapa06.html).

[14] Las grabaciones orales sólo están disponibles en CD-Rom. La página web del proyecto es http://www.hum.au.dk/romansk/fransk/parle/

[15] http://www.lllf.uam.es/~fmarcos/informes/corpus/corpusix.html

Los corpus sobre los que voy a centrar mi atención en las siguientes líneas pertenecen a los ámbitos iberorrománico y vasco[16]. Son los siguientes: *Corpus Oral y Sonoro del Español Rural*, *Arquivo do Galego Oral*, *Corpus Oral Dialectal*, *Bizkaieraren Fonoteka*, *Archivu Oral de la Llingua Asturiana*, *Atlas Sonoru de la Llingua Asturiana*, *Eibartarren Ahotan* y *Corpus Dialectal de Extremadura*.

Corpus Oral y Sonoro del Español Rural (COSER)

El COSER[17], dirigido desde la Universidad Autónoma de Madrid por la Dra. Inés Fernández-Ordóñez, debe ser el proyecto de referencia de los modernos estudios dialectales españoles. El Corpus se empezó a compilar en 1990, y hasta el momento actual se han recogido grabaciones en más de 630 enclaves rurales (casi 800 horas de grabación) repartidos por las dos Castillas, Asturias, País Vasco, Navarra, La Rioja, Aragón, Madrid, Extremadura y Andalucía, datos que por sí solos ya dan idea de los interesantes resultados y de los éxitos investigadores que dicho proyecto está cosechando. La metodología empleada para desarrollar el *Corpus*, según Inés Fernández-Ordóñez, "ha sido la de la entrevista sociolingüística, dirigida por parte de los encuestadores hacia ciertos temas de la vida tradicional en el campo"[18]. Los informantes, por regla general, presentan las mismas características que los que se seleccionaban en los trabajos de geografía lingüística: naturales de la localidad, sin estudios, de edad avanzada…, y es de destacar el hecho de que no solamente no se discrimina a la mujer como sujeto de encuesta sino que, por los datos que han publicado los responsables del *Corpus*, las muestras de habla femenina son estadísticamente superiores a las muestras del habla de los varones.

El *Corpus Oral y Sonoro del Español Rural* es accesible a través de internet, en una página web en la que se informa de la metodología seguida para su confección, las campañas de encuesta que se han materializado desde 1988, las publicaciones

[16] Un par de ejemplos de lo que se está haciendo fuera del ámbito iberorrománico se pueden consultar en http://www.anglistik.uni-freiburg.de/institut/lskortmann/FRED/ (Freiburg English Dialect Corpus) y en http://www.ncl.ac.uk/necte/ (The Newcastle Electronic Corpus of Tyneside English).
[17] http://pidweb.ii.uam.es/coser/
[18] Inés Fernández-Ordóñez, "Nuevas perspectivas en el estudio de la variación dialectal del español: el *Corpus Oral y Sonoro del Español Rural*", en *Actes du XXIV Congrès de Linguistique et Philologie Romanes* (Aberystwyth, Wales, 2-5 August 2004) (en prensa). Disponible en internet en la dirección http://pidweb.ii.uam.es/coser/contenido.php?es&publicaciones

relacionadas con el proyecto, el historial del grupo investigador y una muestra de las grabaciones conseguidas, entre otras cuestiones.

Arquivo do Galego Oral

El *Arquivo do Galego Oral*[19], dirigido por Francisco Fernández Rei, es un proyecto del Instituto da Lingua Galega de la Universidad de Santiago de Compostela que se ha nutrido de las grabaciones efectuadas en el marco de un proyecto anterior denominado *Arquivo Dialectal da Lingua Galega* (1992-1995), de las realizadas para la confección del *Atlas Lingüístico Galego* (1974-1976) y de grabaciones recogidas por investigadores del Instituto da Lingua Galega (1979-1995)[20]. El número de archivos de sonido que contiene el *Arquivo* es enorme (sólo entre 1978 y 1985, los colaboradores del Instituto visitaron 723 núcleos del territorio lingüístico gallego), y esas grabaciones –unidas a los mapas del Atlas– son las que han permitido demostrar la existencia de tres bloques dialectales en Galicia:

> "O material recolleito para a elaboración do devandito Atlas e a análise de boa parte do corpus de gravacións existentes no ILG permitiron redactar un manual de dialectoloxía, no que se dividiu todo o dominio lingüístico actual (incluído o galego da franxa exterior) en tres grandes complexos ou bloques (occidental, central e oriental), con varias áreas e subáreas (e mesmo microsubáreas) de desigual extensión xeográfica en cada un deses bloques"[21].

Corpus Oral Dialectal

Desde la Universidad de Barcelona se está materializando el proyecto VALDIC[22] (Análisis e interpretación de la variación lingüística dialectal a partir de la explotación de un corpus oral). Como exponen los responsables del proyecto en su página web, sus propósitos son:

> "a) optimizar la accesibilidad y la presentación de dichos materiales mediante el desarrollo de aplicaciones de cartografía automatizada y de análisis dialectométrico;

[19] El *Arquivo* carece actualmente de página web; las grabaciones se han dado a conocer en varios CD-Rom y cassettes. Puede visitarse, no obstante, la dirección del *Arquivo Sonoro de Galicia* (http://www.consellodacultura.org/arquivos/asg/arq_son.html), que contiene alguna información sobre los materiales dialectales.

[20] Puede consultarse un resumen de la historia del proyecto, con ejemplos de etnotextos, en Francisco Fernández Rei, "Normas para a transcrición do Arquivo do Galego Oral", *Revista de Filología Románica*, 14, 1997, págs. 225-240.

[21] Francisco Fernández Rei y Carme Hermida Gulías, *A nosa fala: bloques e áreas lingüísticas do galego*, 2.ª ed., Santiago de Compostela, Consello da Cultura Galega e Arquivo Sonoro de Galicia, 2003, pág. 9.

[22] http://www.ub.es/lincat/

b) establecer los criterios y la metodología que rigen las clasificaciones dialectales y su aplicación a la lengua catalana;
c) determinar y valorar los efectos que el proceso de estandardización tiene sobre las variedades dialectales;
d) analizar gramaticalmente los datos dialectales con la finalidad de establecer una tipología de fenómenos de acuerdo con las tendencias universales que manifiestan las lenguas".

El investigador principal del proyecto es el Dr. Joaquim Viaplana Lleonart, quien ha dado a conocer, en formato CD-Rom, algunas de las grabaciones que se han realizado en el marco del VALDIC[23].

Bizkaieraren Fonoteka

El Departamento de Electrónica y Telecomunicaciones de la Universidad del País Vasco, en colaboración con la Diputación Foral de Vizcaya, está llevando a cabo un proyecto de fonoteca del vizcaíno[24] que consiste en la recopilación de grabaciones dialectales, y que comprende –en palabras de sus autores– "una base de datos oral y textual, que albergará los archivos sonoros junto con información adicional; un sistema de consulta que permitirá el acceso a los datos a través de internet; y el software de administración del sistema, que permitirá la actualización continua de la base de datos"[25]. En el año 2001 la fonoteca albergaba ya más de 100 horas de material sonoro procedente de 85 localidades (8 variedades dialectales vizcaínas). Los textos incluidos en la *Bizkaieraren Fonoteka* pertenecen a tres tipos: palabras aisladas, textos propiamente dichos (etnográficos, lingüísticos, narraciones y cuentos) y textos de literatura popular (adivinanzas, canciones, juegos, etc.)[26].

Archivu Oral de la Llingua Asturiana

Dentro de la sección *Espacio y tiempo de la lengua asturiana*, los responsables del diario electrónico *asturies.com* han incluido una selección de grabaciones dialectales recopiladas entre finales de los años 80 y 1997[27]. Las muestras sonoras proceden en su mayoría del archivo del Dr. García Arias y su duración oscila entre los 49 segundos (la

[23] Joaquim Viaplana y M. Pilar Perea, *Textos orals dialectals del català sincronitzats.Una selecció*, Barcelona, PPU, 2003
[24] http://bips.bi.ehu.es/ahoweb/bizkaifon/bizkaifon.html
[25] AA.VV., "Base de datos oral y textual para el euskera", en *XVI Simposium Nacional de la Unión Científica Internacional de Radio URSI 2001*, Villaviciosa de Odón, 2001.
[26] Algunos documentos relacionados con el proyecto se pueden consultar en la dirección http://bips.bi.ehu.es/ahoweb/bizkaifon/bizkaifon_documents.html
[27] http://www.asturies.com/asturianu/archoral/

grabación de Camonéu) y los 14 minutos (en La Canga). Desconozco si el proyecto va a tener continuidad, pero sus resultados podrían ser especialmente interesantes tratándose como se trata de una zona que carece de atlas lingüístico y que manifiesta una riqueza dialectal incomparable.

Atlas Sonoru de la Llingua Asturiana

Más continuidad y sistematicidad que el proyecto anterior presenta el *Atlas Sonoru de la Llingua Asturiana*,[28] dirigido por Jesús Suárez López, quien hasta la fecha ha editado dos CD-Rom que contienen diversas grabaciones efectuadas en distintas localidades asturianas. El primero, con la colaboración de Ignacio Martínez Alonso, incluye diecinueve grabaciones del suroccidente asturiano (Ibias, Degaña, Cangas del Narcea, Tineo, Miranda y Somiedo); el segundo, en colaboración con Xosé Antón Fernández Martínez, recoge fondos sonoros de Miranda, Grado, Tameza, Teverga y Quirós.

Eibartarren Ahotan

Un proyecto ciertamente curioso y muy bien planificado es el que están desarrollando en la ciudad de Eibar los investigadores Asier Sarasua e Itziar Alberdi desde 1999. Se trata del *Archivo Oral Eibartarren Ahotan* (*Eibar: la historia en boca de sus protagonistas*),[29] por medio del cual están reconstruyendo –a partir de los datos aportados por las personas mayores– la historia del siglo XX eibartarra:

> "El objetivo último es crear un archivo multimedia que ofrezca a las generaciones venideras información precisa sobre el modo de vida de la primera mitad del siglo XX, todo ello en formato multimedia y en boca de sus protagonistas"[30].

En el Archivo Municipal de Eibar se guardan las casi 400 grabaciones recopiladas hasta ahora (cintas en formato analógico, grabaciones digitales y vídeos). La duración media de los testimonio orales es de una hora.

[28] Las grabaciones no están disponibles en internet, aunque se puede consultar información sobre el *Atlas Sonoru* en la página de la Red de Museos Etnográficos de Asturias (http://www.redmeda.com/modules.php?op=modload&name=News&file=article&sid=11&mode=thread&order=0&thold=0)
[29] http://www.egoibarra.com/Egoibarra/Ikerketak/Eibartarren_Ahotan
[30] Resumen del proyecto: http://www.egoibarra.com/Egoibarra/Ikerketak/Eibartarren_Ahotan/Eibartarren%20Ahotan

Corpus Dialectal de Extremadura (CoDiEx)

El *Corpus Dialectal de Extremadura*[31] pretende albergar una colección de etnotextos (en formato oral y escrito) y archivos de sonido de reducida duración de las cuarenta localidades que se han seleccionado en la comunidad.[32] El objetivo fundamental del *Corpus* consiste en recopilar muestras de habla con las que poder analizar la situación actual de las modalidades lingüísticas más representativas de Extremadura. Como complemento a esa colección de textos sonoros, está prevista también la incorporación de reportajes etnofotográficos, grabaciones de vídeo e información diversa de cada localidad (apuntes geográficos, históricos, demográficos, etc.). En el aspecto metodológico, aunque en general se siguen los patrones marcados por otros proyectos en lo que respecta a la temática, selección de informantes o tratamiento de los datos obtenidos, es de destacar que –sistemáticamente– se recogen textos orales de sujetos masculinos y femeninos en cada una de las localidades de la red de puntos.

Actualmente el *Corpus* presenta los resultados obtenidos en las investigaciones llevadas a cabo en Madroñera, Guijo de Santa Bárbara y Baños de Montemayor, aunque ya dispone de materiales de otros núcleos de las dos provincias (Serradilla, Garrovillas, Malpartida de Plasencia, Ahigal, Guijo de Galisteo, Hervás, Alcuéscar, Cedillo, La Codosera, Cheles, Olivenza, Alange, Campanario, Fuente del Maestre y Peraleda del Zaucejo).[33]

FINAL

Los corpus que he reseñado en estas líneas son solo una muestra de lo que está suponiendo en el campo de la dialectología la aplicación de las modernas técnicas electrónicas. Falta mucho camino por recorrer, pero las bases están sentadas. Internet y los formatos digitales nos están permitiendo acceder a lo que hasta hace poco más de una década era impensable: la viva voz de los informantes, la materia prima de la que se nutre la dialectología.

[31] http://www.geolectos.com/codiex
[32] La región se ha dividido en varias zonas en función de las características lingüísticas que presentan esos cuarenta núcleos: *fala* de Xálima (3 pueblos), variante regional de influencia leonesa (10 pueblos), hablas populares de la frontera con Portugal (9 pueblos), hablas populares de la frontera castellana (8 pueblos) y hablas populares de mayor influencia meridional (11 pueblos).
[33] La mayoría de esas grabaciones proceden de la *Cartografía lingüística de Extremadura* (http://www.geolectos.com) y fueron realizadas entre 1995 y 1997.