El corpus ALDICAM-CM

Geografía lingüística diacrónica de la Comunidad de Madrid*

Pedro Sánchez-Prieto Borja
Universidad de Alcalá

The linguistic uses in Madrid have been object of references about the current popular use and the literary authors, but only in recent years there have been advances in the knowledge of the speech of Madrid, thanks to sociolinguistic studies. Rural Madrid has also been studied. There are, however, few studies on the historical variety of Madrid, and hardly ever from texts with an explicit place of origin. For this reason, we are elaborating the "Atlas Diacrónico e Interactivo de la Comunidad de Madrid" (ALDICAM), based so far on a corpus of 724 documents of 22 archives and 46 localities, and dated between 12th and 13th century. The results of any query on the corpus can be projected directly on a map of the Community of Madrid, according to the model already operational in CODEA + 2015.

Keywords: Corpus linguistics, Diachronic Dialectology, Linguistic geography, archival documents, Madrid region

1. Los documentos y los corpus lingüísticos

Los corpus lingüísticos han conocido un desarrollo impensable hace solo 30 años, hasta el punto de que hoy sería imposible cualquier investigación lingüística, al menos de metodología variacionista, tanto sincrónica como diacrónica que no estuviera basada en ellos. La posibilidad de disponer, de manera inmediata, de una ingente cantidad de datos permite al estudioso fundamentar mejor sus conclusiones.

Gracias al aumento en el número de datos disponibles, es posible establecer la frecuencia de las invariantes respecto de los distintos parámetros de la variación, lo que otorga fiabilidad estadística a las deducciones. Sin embargo, los textos ofrecidos no siempre son fiables, pues se utilizan ediciones anteriores, que no están expresamente elaboradas para ese corpus concreto. Además, los corpus no suelen hacer explícita la problemática de la edición de los textos incluidos, pues no se muestra la situación textual del testimonio seguido; este puede ser una copia

* Este estudio se ha llevado a cabo dentro del proyecto "Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid (ALDICAM-CM)", financiado por la Comunidad de Madrid (S2015 HUM 3443).

muy deturpada y distante en el tiempo del original (Rodríguez Molina y Octavio de Toledo 2017), y tampoco suelen indicarse los criterios editoriales.

Defendemos la elaboración de corpus abiertos a distintos fines y usuarios, pero con una orientación específica; en nuestro caso, al estudio lingüístico. Los materiales textuales que se ofrezcan han de ser preparados expresamente para ese corpus, y editados con criterios uniformes[1]. Frente a las fuentes literarias, hemos preferido acoger documentos de archivo, que tienen la ventaja de tener data *chronica* y *topica*[2].

2. El corpus CODEA

Dentro de estas premisas, nos planteamos en 1995 la elaboración del *Corpus de Documentos Españoles Anteriores a 1700* (CODEA), que contiene fuentes archivísticas de la Península Ibérica desde la época de orígenes hasta el s. XVII inclusive (en 2015 se amplió al s. XVIII, y en un proyecto en curso al XIX). El corpus contó en su primera versión con 1500 piezas, y CODEA+ 2015 con 2500 (http://www.corpuscodea.es). No se trata de un corpus monocorde registralmente, pues abarca toda la escala de formalidad, al partir de un concepto lato de documento, que incluye tanto el diploma regio en pergamino como la nota más humilde en un pedazo de papel mal cortado.

Resulta imposible con una sola versión abarcar todos los requerimientos de estudio, de la paleografía a la historia de las mentalidades; por ello se ha optado por una triple presentación: facsímil, transcripción paleográfica y presentación crítica, de acuerdo con los criterios CHARTA[3]. En cuanto a la elaboración informática, CODEA está concebido como una base de datos multirrelacional, en la que las búsquedas pueden filtrarse por fecha, lugar, tipología documental, diplomática, ámbito de emisión, etc. Pero la novedad más llamativa de la versión CODEA+ 2015 es la posibilidad de proyectar directamente a mapa los resultados de las consultas, lo que la convierten en un atlas lingüístico diacrónico y dinámico, pues, frente a otros atlas diacrónicos[4], no se ofrece un conjunto de mapas fijos, sino que el usuario puede elaborarlos de manera inmediata según sus necesidades

[1] Al respecto, hemos propuesto la distinción entre corpus primarios, o expresamente elaborados, y corpus secundarios (Miguel Franco y Sánchez-Prieto, B. 2015).

[2] El lugar de emisión, cuando falta, puede deducirse casi siempre. Los no datados deben descartarse, pero si interesaran por sus rasgos lingüísticos, pueden datarse conjeturalmente por comparación con otros, como se ha hecho para el corpus CODEA (Kawasaki 2014).

[3] "Criterios de edición de documentos hispánicos (orígenes-Siglo XIX) de la Red Internacional CHARTA" <http://www.redcharta.es>.

[4] Por ejemplo, *LAEME* (2008) para el inglés medieval o Dees (1980), para diplomas franceses.

y los parámetros que elija[5]. Así, p. ej., pueden buscarse las formas verbales en **edes*/**éis*/**és*, en documentos de entre 1450 y 11520, de ámbito privado y de Aragón.

3. El proyecto ALDICAM-CM

El modelo de corpus, y de atlas, peninsular puede tener gran interés para la historia de la lengua española, pues permite observar procesos evolutivos importantes con un grado notable de detalle. En particular, puede apreciarse la interacción lingüística entre las diferentes regiones, y así se identifican con facilidad distribuciones espaciales que tiene su foco en las áreas occidentales (*alguien*) u orientales (*asín*)[6]. Pero la necesidad de abarcar un territorio y un período muy dilatado obliga a ofrecer una red relativamente poco tupida de puntos. Por ello es necesario poner en marcha corpus y atlas de dominio menor, como el de Extremadura, basado en el léxico de inventarios de bienes[7]. En 2016 se inició la elaboración del "Atlas Lingüístico Diacrónico y Dinámico de la Comunidad de Madrid (ALDICAM-CM)", que acoge fuentes documentales de los ss. XIII al XIX, aunque estas solo abundan del XVI en adelante.

Hasta ahora, se ha trabajado en 23 archivos: Regional de la Comunidad de Madrid, de Villa, Histórico de Protocolos, Hospital de San José de Getafe y los municipales de Alcalá de Henares, Aranjuez, Arganda del Rey, Arroyomolinos, Buitrago del Lozoya, Coslada, Chinchón, Colmenar Viejo, Daganzo, El Escorial, Guadarrama, Hoyo de Manzanares, Moralzarzal, Navalcarnero, Parla, San Lorenzo de El Escorial, Torrelaguna, Torrejón de Ardoz y Valdemoro. Los lugares de emisión de documentos seleccionados hasta ahora son 47, y establecen una tupida red de puntos (Figura 1).

Los documentos transcritos son 724; las diferencias más significativas afectan a la tipología documental, y, a su vez, esta condiciona las posibilidades de estudio. Su extensión va de unas pocas palabras en las notas de abandono de niños conservadas en el fondo de la Inclusa a los varios folios de algunas piezas judiciales[8].

[5] La cartografía digital del corpus está siendo elaborada por Hiroto Ueda (Universidad de Tokio) mediante programación en JavaScript.

[6] Del Barrio (2016) ha identificado, basándose en CODEA, que la sustitución de de *haber* por *tener* tiene un foco aragonés.

[7] Es objeto de la tesis doctoral de Diego Sánchez Sierra, que será presentado en 2019.

[8] En atención a la uniformidad de emisor y, en su caso, copista, se ha considerado documento cada unidad funcional con una fecha y lugar de emisión propio y, casi siempre, explícito, y no al expediente completo de, por ejemplo, un auto judicial, pues este costa de piezas diferentes tanto desde el punto de vista diplomático (diferentes datas) y registral, según se observa al

Para la ciudad de Madrid, nuestras pesquisas se han orientado a aumentar la variación sociolingüística. En efecto, hemos podido recuperar un gran número de piezas de nivel sociolingüístico bajo y medio (Sánchez-Prieto y Vázquez Balonga 2017)[9].

Figura 1: Localidades de emisión de los documentos

comparar un auto con las declaraciones de testigos, pues esta muestran a veces segmentos que reflejan literalmente lo expuesto por el declarante: "a que volvió a dezir la testigo: –Pícaro, ¿a qué vienes aquí?" (El Escorial, 1708).

[9] P.ej., en una nota de abandono de una niña a la inclusa se escribe: "no esta ristiana ase de llamar franeçisca" (ALDICAM 27, 1593).

4. El corpus de ALDICAM

Como en CODEA, los documentos se presentarán en versión paleográfica y crítica, además del facsímil. La cabecera descriptiva incluirá los siguientes campos: [número de orden] 0476 / [corpus] ALDICAM / [fecha y lugar] s.f. [1585] (El Escorial, Madrid, España) / [archivo y signatura] Archivo Municipal de El Escorial, 3446-14 / [regesto] Denuncia de Juan Cabrera de Córdoba, criado del rey, a Francisco López, por sangrar mal a su caballo. / [ámbito] Judicial / [tipo] Denuncia / [temas] animales, justicia, rey.

Lo más significativo será la posibilidad de crear un número ilimitado de mapas, pues las consultas pueden filtrase por uno o varios campos de la base de datos. De este modo será posible examinar los parámetros de la variación (diacronía, diatopía, diastratía y diafasia); de estos, los dos primeros se presentan de manera inmediata; la diafasia puede objetivarse de acuerdo con la tipología documental, al distinguirse entre ámbito municipal y particular, más informal el segundo[10]. Un prototipo del corpus de ALDICAM estará accesible próximamente con una muestra de 100 documentos.

5. Geografía lingüística diacrónica en la Comunidad de Madrid

El examen de la documentación de las localidades madrileñas ha permitido apuntar diferencias diacrónicas entre las distintas localidades de la CM; los documentos nos ofrecen una suerte de mapa referencial en el que los temas tratados suelen ser diferentes; así, la ganadería aparecerá más en la Sierra N. y E. que en el S. de la Comunidad de Madrid.

5.1 Madrid comparado con otros territorios

Tradicionalmente, se ha señalado el papel de la corte madrileña (desde 1561) en la difusión de algunas innovaciones, como el leísmo. El fenómeno no tiene un foco único, y los territorios situados al norte son prioritarios en su consolidación (Fernández-Ordóñez 2001); la documentación examinada apunta a que el leísmo es un fenómeno tardomedieval en Madrid, y no parece ligado a la repoblación (Sánchez-Prieto y Vázquez Balonga 2018).

Encaja con lo que sabemos sobre los procesos de difusión del cambio el que la ciudad tuviera un peso considerable en la difusión de las innovaciones, es decir,

[10] Véase la sección "Materiales de trabajo" para los ámbitos deemisión, tipología documental y materias (http://aldicam.blogspot.com/p/material-de-trabajo.html).

que usos recibidos en el habla de Madrid cobraran impulso desde ahí y llegaran a estabilizarse y generalizarse en otros espacios. Ejemplo pueden ser los futuros en *-drá* (*tendrá*, *vendrá*, *pondrá*, etc.), documentados en el ámbito aragonés en la Edad Media, pero que desde la segunda mitad del XVI se difunden desde la capital[11]. Otros caso podrían ser los superlativos en *-ísimo*[12]. Para el léxico, resulta seguramente más difícil identificar los focos iniciales. Así el verbo *platicar*, tan difundido en el español de América, esp. México, se encuentra en la Península Ibérica entre los siglos XV y XVII sobre todo en Toledo y Madrid[13].

5.2 Diferencias geográficas internas en la Comunidad de Madrid

Si el leísmo es un fenómeno recibido tardíamente en la CM, suponemos que aquellas que están en contacto con las provincias castellano-viejas, en las que el fenómeno parece antiguo, lo reflejarán antes y con más intensidad. Así, El Escorial, antes integrado en la "tierra de Segovia", muestra *le* para referente contable inanimado ("el cual dicho *orno* içe y no conforme tiene obligación del acer*le*" (ALDI-CAM 472, de 1624). El proceso de expansión del laísmo, en cambio, parece seguir patrones diferentes. En el s. XVI parece del todo asentado en la ciudad de Madrid, pero no así en Arganda del Rey, donde, en el s. XVII, parece predominar todavía la variante etimológica *le* para O.I. femenino (Sánchez-Prieto y Vázquez Balonga 2018).

En el léxico material, los documentos reflejan las diferencias entre los distintos territorios en las actividades económicas, según se apuntó. Así, en Hoyo de Manzanares documentamos *fabriquero* (1682), para elaborador de carbón. Solo hemos encontrado en Hoyo de Manzanares *pasto siego*: "Más un pradito de *pasto siego* y monte de chaparro" (1704). Fuera de la provincia de Madrid no hemos documentado *cerviajo* 'borde alto de una finca o del camino': "cayó de un *cerviajo* y se halló sin lesión" (325, Daganzo de Arriba, 1782). En cuanto al léxico referido al ganado, citaremos *borro* ("tres borregas […] tres *borras*") y *ceajo* ("un *ceajo* de dos años") solo en Montejo de la Sierra (1894)[14].

[11] La difusión podrá deberse a los escribanos de origen navarro que trabajaron en la corte (Serrano Marín 2018).

[12] Se ha señalado la influencia del latín y del italiano. Una posible vía de entrada pudo ser la corona de Aragón, en particular el uso de la cancillería (Araque Comino y Sánchez-Prieto en este vol.)

[13] Puede establecerse de manera inmediata el mapa buscando "platic*" en CODEA.

[14] En el CORDE solo encontramos *borro* en dos manuscritos del Fuero de Navarra, de hacia 1300-1330. *Ceajo* se recoge en el *Diccionario* usual de 1956 como "chivo o cordero que no llega a primal", con la marca de argonesismo.

6. Conclusiones

Cabe señalar la prioridad que otorgamos a los documentos archivísticos para reconstruir etapas pasadas de la lengua, idea solo posible tras la ampliación del concepto de documento, no limitándolo a los escritos administrativos. Solo las piezas datadas y localizadas podían proporcionarnos información para poder contrastar la variedad madrileña con las de otras regiones, y, asimismo, examinar las diferencias geográficas internas de la Comunidad de Madrid.

Referencias

Del Barrio de la Rosa, F. 2016. De *haber* a *tener*. La difusión de tener como verbo de posesión en la historia del español: contextos y focos. En C. de Benito Moreno & Á. S. Octavio de Toledo (eds.), *En torno a* haber*: construcciones, usos y variación desde el latín hasta la actualidad*. Berna: PeterLang, 239-279.

Dees, A. 1980. *Atlas des formes et des constructions des chartes françaises du XIIIe siècle*. Tübingen: Beihefte zur Zeitschrift für romanische Philologie, vol. 178.

Fernández-Ordóñez, I. 2001. Hacia una dialectología histórica. Reflexiones sobre la historia del leísmo, el laísmo y el loísmo. *Boletín de la Real Academia Española*, LXXXI: 389-464.

Miguel Franco, R. & Sánchez-Prieto, B. 2015. CODEA: a 'primary' corpus of Spanish Documents. *Variants*, 12: 199-217. <http://variants.revues.org/364>.

LAEME: Lang, M. 2008. *Linguistic Athlas of Middle English.* <http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html> [18/05/2018].

Kawasaki, Y. 2014. Datación crono-geográfica de documentos medievales españoles, *Scriptum Digital*, 3: 29-63.

Rodríguez Molina, J. & Octavio de Toledo y Huerta, Á. 2017. La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística, *Scriptum Digital*, 6: 5-68.

Sánchez-Prieto, P. & Vázquez Balonga, D. 2017. Hacia un corpus de Beneficencia en Madrid (siglos XVI-XIX). *Scriptum Digital*, 6: 83-103.
http://scriptumdigital.org/numeros.php?opt=act&lang=es

Sánchez-Prieto , P. & Vázquez Balonga, D. 2018. Toledo frente a Madrid en la conformación del español moderno: el sistema pronominal átono. *Revista de Filología Española*, XCVIII, 1º, enero-junio: 157-187.

Serrano Marín, M. 2018. *Estudio de la morfología verbal del español en fuentes documentales de los siglos XIII-XVI*, Tesis doctoral, Universidad de Alcalá.