

DESARROLLO Y EXPLOTACIÓN DEL CORPUS DE DOCUMENTOS ESPAÑOLES ANTERIORES A 1700 (CODEA)*

PEDRO SÁNCHEZ-PRieto BORJA
Universidad de Alcalá

RESUMEN:

Este artículo tiene por objeto presentar el estado actual del «Corpus de Documentos Españoles Anteriores a 1700» (CODEA, <http://www.textoshispanicos.es>), elaborado por el «Grupo de Investigación de Textos para la Historia del Español» (GITHE) de la Universidad de Alcalá. El CODEA ofrece a día de hoy 1500 documentos de diferentes archivos, de todas las provincias peninsulares no bilingües de España, y de los ss. XII al XVII. La triple presentación (facsimilar, paleográfica y crítica) facilita su empleo en diversos ámbitos, de la paleografía a la sintaxis histórica. Se adelantan, además, los desarrollos previstos para un futuro inmediato, como la búsqueda por lemas, la presentación estadística directa de los resultados de las consultas y la visualización en forma de mapa de las respuestas a las búsquedas. Y se exponen las vías y posibilidades para su explotación lingüística en los diferentes niveles (grafía y fonética, sintaxis y léxico). El CODEA funda su utilidad en su característica de corpus «primario», en el que los elaboradores son responsables del proceso íntegro de edición de los textos que ofrecen.

Palabras clave: corpus lingüísticos, documentos archivísticos, historia de la lengua española

ABSTRACT:

This paper aims to present the current status of the «Corpus de Documentos Españoles Anteriores a 1700» (CODEA, <http://www.textoshispanicos.es>) prepared by the «Grupo de Investigación de Textos para la Historia del Español» (GITHE) of the University of Alcalá. CODEA offers 1500 documents from different archives produced in all of the non-bilingual peninsular provinces of Spain, spanning from the 12th to the 17th centuries. The triple presentation (facsimile, paleographic transcription and edition with normalized spelling) facilitates its use in various disciplines, from palaeography to historical syntax. In the paper we also present further developments that will be available in the near future, such as the possibility of conducting searches by lemmata, the presentation of results in statistical graphs, and the display of maps with the geographical distribution of the forms in the search results. Likewise the paper illustrates with various examples the potential of these developments for different types of linguistic analyses (spelling and phonetics, syntax and lexical studies). The most distinctive feature of CODEA is to be a «primary» corpus in which the authors provide the texts and are fully responsible for their editorial process.

Key Words: Language Corpora, Archival documents, History of the Spanish language

1. LOS CORPUS EN LA INVESTIGACIÓN DE HISTORIA LINGÜÍSTICA

1.1. Es ya un lugar común hablar de las posibilidades que brindan la lingüística, la filología y la informática cuando cooperan entre sí en la investigación sobre el cambio lingüístico y la historia de la lengua. Sin embargo, las expectativas creadas no se han visto del todo satisfechas, y cabe hablar de una cierta distancia entre lo que se promete y lo que, a día de hoy, está disponible como logro acabado¹. Un buen número de proyectos no pasan del grado de tentativa, y otros —sin excluir el nuestro— avanzan de manera mucho más lenta

* Este trabajo se ha elaborado gracias al proyecto «Edición y estudio lingüístico de documentos toledanos (ss. XVI y XVII)», del Ministerio de Ciencia e Innovación (FFI2009-10877).

¹ Entre las causas, cabe citar algunas deficiencias estructurales de la política científica: es fácil conseguir recursos para dotación en material informático, pero no para pagar personal investigador dedicado íntegramente a un proyecto.

de lo que sería deseable. Ello se debe a diferentes causas, entre las que cabe contar la incomunicación relativa entre filólogos, de una parte, e informáticos, de otra. Los primeros sabemos poco de los aspectos técnicos del desarrollo computacional, los segundos no comprenden cuales son los objetivos científicos que se pretenden cubrir, y desconocen los conceptos básicos de lingüística imprescindibles para la edición electrónica y la elaboración de recursos para la recuperación de la información. Con todo, no es este, a nuestro entender, el problema principal, pues de un modo u otro, cada vez más se tienden puentes que salvan esta distancia, y es ya más frecuente entre los jóvenes investigadores el conocimiento tecnológico necesario, de modo que hoy el término *tecnolingüística* puede resultar adecuado para expresar esta confluencia entre ramas que hasta hace poco iban casi siempre en paralelo. No se trata, desde luego, que el filólogo tenga que ser capaz del desarrollo íntegro de aplicaciones electrónicas, pero sí que cada vez es mayor la necesidad de preguntarse seriamente qué información se quiere poner al alcance de la comunidad científica y —volveremos sobre esto— de los demás usuarios y en qué modo. Simplificando algo, el problema no es solo saber qué se ofrece, sino cómo encontrarlo. Pensamos, como puede comprenderse, en la edición electrónica de textos. Porque nadie mejor que el editor —en cuanto estudioso de los textos en todos sus aspectos, propiamente textuales, lingüísticos, retóricos, literarios, conceptuales y culturales— sabe qué pueden darnos estos y qué querría él mismo encontrar en ellos. En definitiva, para saber qué buscar hay tener primero una noción de lo que hay. Ello no quiere decir, claro, que el aprovechamiento del texto se agote en lo previsto inicialmente; antes al contrario, ello es solo un punto de partida dinámico para que se pueda ir más allá, y aquí reside una ventaja conceptual de la edición electrónica; por ejemplo, es posible combinar parámetros de búsqueda que nos demuestren la relevancia de factores que no hubiéramos considerado al principio.

Es en el plano propiamente filológico donde encontramos las dificultades mayores, pues la exigencia de calidad en las ediciones electrónicas es, en general, menor que en las que se difunden en papel. Aunque solo fuera por esto, nos opondríamos ya de plano al rótulo de «edición tradicional» para la edición en papel, rótulo que rechazamos enérgicamente. La diferencia básica no es entre edición en papel y edición electrónica, sino edición rigurosa y edición no científica, del mismo modo que, como decía Orduna (1990), antes de enfrentar edición crítica a no crítica, hay que diferenciar entre ediciones buenas y malas. El mero soporte electrónico no es garantía de rigor; y, si acaso, hasta ahora ha sido más bien lo contrario. Pero si el carácter científico es el criterio clave, la edición electrónica tiene una ventaja enorme sobre la edición en papel, y es su naturaleza revisable y mejorable en actualizaciones sucesivas o nuevas ediciones².

1.2. Por otra parte, en lo que respecta a la recuperación de la información, las ventajas evidentes del uso de los corpus tienen también su cruz; la facilidad de empleo, el acceso inmediato a los datos lingüísticos conocen su contrapartida en la forma descontextualizada en que estos datos se ofrecen. La falta de conocimiento de la historia textual que suele acompañar al uso inmediato de los corpus dificulta la interpretación lingüística. En un corpus

² Caben dos modelos en la revisión de las ediciones electrónicas: (1) el que consiste en la corrección continua sin plazo, y (2) la revisión reglada en ediciones sucesivas, normalmente anuales. El primero conecta mejor con las posibilidades de las nuevas tecnologías; el segundo, facilita las referencias en estudios sobre el corpus.

de documentos medievales, como el que aquí vamos a presentar, no todas las piezas sirven para ilustrar la variación geográfica; en los diplomas de la cancillería es escasamente relevante el que un privilegio rodado, pongamos por caso, se emita en Burgos o en Sevilla, mientras que cartas de compraventa de dos monasterios próximos pueden mostrar significativas diferencias. En los impresos, el problema no es menor que en las fuentes manuscritas. Un libro impreso en Toledo, ¿debe llevar la marca geográfica «Toledo» aun cuando el autor sea de otro lugar?³ Es cosa sabida, pero mucho más olvidada, que el manejo indiscriminado de los datos que pueden obtenerse en los corpus de grandes dimensiones falsea en no pocos puntos la realidad lingüística. Toda prudencia es poca, pues la idea, fundada en la estadística, de que cuanto mayor es la muestra las desviaciones particulares quedan absorbidas, no nos parece de aplicación en parcelas humanísticas, donde la fiabilidad se encomienda al rigor⁴.

No conviene, pues, gastar ya palabras de asombro ante las no tan nuevas tecnologías⁵, sino que hemos de poner el acento en que las nuevas posibilidades nos obligan a ir más lejos de quienes nos precedieron, y podríamos preguntarnos si se ha superado la descripción de las áreas lingüísticas peninsulares que llevó a cabo Menéndez Pidal en *Orígenes del español*⁶. Si la respuesta no es unánimemente afirmativa, no es ciertamente por un peor planteamiento del problema, pues los progresos metodológicos son grandes en dos aspectos fundamentales, un mejor conocimiento de la relación entre grafía y fonética y la incorporación decidida de la perspectiva sociolingüística. Lo que nos lleva a plantear algunas reservas es que la fase empírica no ha avanzado de manera comparable, y no estamos seguros de manejar mejores datos de los que disponía Menéndez Pidal⁷ y que se concretaron luego en los *Documentos lingüísticos de España*, en los que la selección de piezas muestra no pocas intuiciones geniales, según ilustraremos en otro lugar.

Suele olvidarse, pues, que el uso de recursos electrónicos no es garantía de cientifismo, pues la ciencia requiere de (a) una fundamentación empírica sólida y (b) de un método riguroso de interpretación de los datos. El plan científico es previo a la investigación; sin embargo, esto no quiere decir que todo deba y pueda preverse. Antes al contrario, creemos que los corpus están abiertos a empleos tal vez no previstos en la concepción

³ Con esto no queremos invalidar los impresos como fuente de datos para el historiador de la lengua. Así, la *Historia o descripción de la imperial ciudad de Toledo*, de Pedro de Alcocer, toledano, fue impresa en Toledo en 1555. La autoría se ha discutido, pero la alternativa que se propone es el sacerdote toledano Juan de Vergara.

⁴ El problema de la representatividad cuenta con escollos difíciles de sortear. No es el menor el que la distribución de los fondos sea bastante caprichosa. De la primera mitad del s. XIV se conservan menos documentos que de la segunda del XIII.

⁵ Como gran mal de Internet se ha señalado su infinita capacidad distractora, que casa mal con la concentración y esfuerzo continuado que requiere el trabajo filológico.

⁶ El uso de corpus nos acerca a una concepción del saber ya conocida en la Edad Media, en la que la ciencia pertenecía a todos los que se esforzaban en ella, y primaba la citabilidad sobre la idea de autor. Para la Biblia, la circulación de concordancias antes de mediados del s. XIII permitía hacer alarde de un «conocimiento» — suponemos que también en la predicación — que para los profanos debía de causar asombro.

⁷ Sería ingenuo pensar que *Orígenes* se escribió solo con los documentos que anteceden a la obra. El tino de Menéndez Pidal y colaboradores del Centro de Estudios Históricos en la selección de las piezas no deja de sorprendernos casi un siglo después, lo que no implica que esté exenta de problemas, y menos aún que la valoración de los datos no sea errónea muchas veces. Ello no va en desdoro de la genialidad de las intuiciones, pues solo mediante hipótesis atrevidas es posible un progreso sustancial del conocimiento.

actual; de hecho, pueden tener, y, sin duda la tendrán, una mayor transcendencia en el futuro, tanto en el ámbito científico más estricto, por lo que nos compete vinculado a la historia de la lengua, como en el de la difusión del conocimiento y del acceso a la información. Constituyen el campo más innovador de las humanidades digitales, aunque no necesariamente el más llamativo —por ahora— ni el más conocido. La imbricación entre proceso editorial y recuperación de la información tiene perspectivas de desarrollo insospechadas, y por ello mismo la lingüística de corpus se configura no ya como una técnica auxiliar, sino como una verdadera corriente metodológica (Enrique Arias 2009), en la que los requisitos son la sólida fundamentación empírica y la comprobabilidad de las deducciones basadas en los textos. Y es que tanto como el conocimiento mismo importa conocer el proceso por el que este se adquiere. Por ello, no se trata solo de cómo encontrar la información, sino de cómo se construye ésta y cómo se somete a crítica⁸. Así, con sus recursos de navegación y búsqueda, el modelo del corpus podría contribuir a poner orden en la babel que es hoy Internet⁹.

1.3. La edición electrónica necesita, pues, unos fundamentos igual de sólidos que la que se ofrezca en cualquier otro soporte. Y, por esta vía insospechada, la filología (sin adjetivos, pero si se quiere con el que aparece en el sintagma «filología informática») pasa de ser un área para muchos tenida como marginal, instrumental de los estudios literarios, lingüísticos e históricos a ocupar un lugar destacado en las humanidades digitales, y aun más allá de ellas. La exigencia de rigor se hace más perentoria a medida que los textos se alejan del momento presente. No es que editar los textos del s. xx no requiera tomar continuas decisiones que precisen conocimientos filológicos, pero sí es cierto que ello afecta más profundamente a las obras antiguas, pues con las actuales el editor (y el lector) comparten código lingüístico, aparte de más experiencias humanas, lo que allana el camino hacia su comprensión. Por todo ello, la historia de la lengua y la lingüística histórica han tenido peso relevante en el desarrollo de los corpus. En conjunto, la edición de textos de épocas pasadas plantea problemas más complejos que la de los cercanos en el tiempo¹⁰, y la experiencia cumplida en ellos puede aprovecharse con ventaja en la de los contemporáneos¹¹.

2. PROBLEMAS DE MÉTODO

⁸ Lo contrario implica una trivialización del conocimiento; hoy muchos piensan que saber qué pasó en 1212, 1492 o 1815 no es relevante.

⁹ El exceso de información bloquea las funciones fundamentales de la memoria. Lo gracioso del caso es que esta idea era un lugar común entre los expertos en teoría del conocimiento de principios de los 80, antes de la difusión de internet.

¹⁰ Seguramente, los logros más llamativos se han obtenido con textos antiguos. Véase, por ejemplo, la edición interactiva de *Don Quijote de la Mancha* de la Biblioteca Nacional <http://quijote.bne.es/libro.html>. Es cierto que si ha habido más ensayos para textos del s. xix y anteriores, ha sido, en buena parte, a que no están sometidos a derechos de autor.

¹¹ Por ejemplo, las variante formales plantean un problema para la lematización y búsqueda por lemas: *como* ~ *cuemo*, *so* ~ *su*, *fazedera* ~ *fazedora*, etc. (3.9.2.). Salvados estos escollos para textos antiguos, resulta sencilla su aplicación a los modernos.

2.1. Ante la pregunta de cómo puede aprovechar el historiador de la lengua los corpus textuales, las respuestas pueden ser básicamente dos: (1) atestiguar, ilustrar, ejemplificar aquello que ya se conoce; (2) cuestionarse lo que se conoce, contrastándolo con los datos empíricos¹². Esta segunda orientación exige un tratamiento más riguroso de los datos textuales, que incluya el análisis cualitativo-cuantitativo, y que tenga en cuenta varios planos inherentes a la investigación histórica, al menos los siguientes: (a) el cronológico (¿de cuándo son los textos?), (b) el geográfico (¿de dónde proceden?), (c) el diafásico, registral y sociolingüístico, (d) el tipo (y subtipo) textual.

Pero una caracterización precisa de un texto en estos planos solo es posible con un conocimiento de su historia, que proporcione respuestas a las cuestiones acerca de la autenticidad del testimonio o su carácter de copia, de su distancia cronológica con el original y su distancia textual, y todo ello dentro de una exigencia de fiabilidad. ¿Nos dará el *Cantar de Mio Cid* claves para conocer los factores que intervienen en el desarrollo histórico de *a* ante objeto directo personal? Seguramente no, y no solo por la distancia de un siglo entre el texto original y la copia conservada. Aunque parezca lo contrario, es la sintaxis histórica la que más se ha visto perjudicada por esta falta de anclaje en la crítica del texto (Romero Cambrón 2005). ¿Por cuánto tiempo seguiremos ilustrando con el *Calila e Dimna* la sintaxis del s. XIII? Para que un corpus sea lingüísticamente aprovechable se exige: (1) control filológico del proceso de edición, lo que se relaciona con la fiabilidad, y (2) control filológico en el proceso de tratamiento de la información, es decir, de los datos obtenidos mediante la exploración y explotación del corpus. El editor ha de ofrecer instrumentos que hagan posible la verificación por el usuario. Estos instrumentos se identifican normalmente con el aparato crítico, pero el medio electrónico permite proporcionar las «fases» por las que editor llega a la propuesta editorial final (facsímil, transcripción, edición).

No me cabe duda alguna de que esto es solo posible cuando la elaboración de corpus incluye el proceso de edición, pues solo así es posible aplicar a todos los textos que lo integran unos criterios explícitos, coherentes y homogéneos. Llegamos así a la que me parece la principal característica, que permite distinguir entre:

- (a) Corpus primarios: los elaboradores utilizan materiales textuales propios y son responsables de todo el proceso de edición.
- (b) Corpus secundarios: se utilizan materiales textuales ajenos, y los elaboradores son responsables solo de la preparación electrónica¹³.

Naturalmente, ambas opciones están justificadas¹⁴, pero difieren radicalmente en el modo en el que nos sitúan ante el texto como fundamento genuino de las deducciones e hipótesis del historiador de la lengua. El corpus primario, incardinado en el proceso de

¹² En este sentido, es muy elocuente el título de las actas del primer congreso sobre corpus iberorrománicos celebrado en Palma de Mallorca en 2007 (Enrique-Arias 2009).

¹³ Naturalmente, hay vías intermedias o de compromiso, como la revisión de las ediciones previas, o la unificación de los criterios de presentación gráfica, lo que redundaría en una mayor coherencia de los textos finales (es el caso del *TMILG*, para el gallego medieval; véase también Pascual y Domínguez 2009 sobre el corpus del *Nuevo Diccionario Histórico del Español*).

¹⁴ Las diferencias de finalidad, concepción y tamaño condicionan el modo en que se afronta la edición de los textos que integran un corpus. Así, el *CICA* aprovecha las ediciones previas de textos en catalán de la Edad Media.

edición, nos pone ante los ojos lo que no sabíamos¹⁵. El conocimiento directo de los textos es la única vía, junto con la lectura completa de los mismos, para escapar a las limitaciones que nacen de un planteamiento bastante extendido: se sabe aquello que se estudia, se estudia aquello que se sabe. En el plano estricto de la crítica del texto, el antídoto es lo que podemos llamar el «elogio del hápax», o, como plantea Ferreiro (2010), «el hápax como problema y como solución»¹⁶.

Por otra parte, solo el control del proceso de edición garantiza la explicitación segura de las intervenciones en el proceso elaborativo (manos, correcciones, raspados, etc.) y, por descontado, del editor. Que de cada texto tenga, sin embargo, que mostrarse necesariamente todo el fundamento de las decisiones del editor resulta quimérico para corpus de cierta extensión. La edición crítica con aparato exhaustivo puede sustituirse por la que proponemos llamar «edición asertiva o documentada», en la que se hace explícito de dónde procede el texto ofrecido al lector o usuario: (a) del manuscrito base, si lo hay (o de varios alternativamente), (b) lecciones de otros manuscritos, (c) enmiendas del editor¹⁷.

Las ventajas del corpus primario se extienden también a la posibilidad efectiva de adoptar unos criterios unificados. Estos permiten a su vez recuperar la información contenida en el corpus de manera más eficaz. Así, unos criterios normalizados de presentación gráfica (en aquellos rasgos no relevantes fonéticamente) facilita el proceso de lematización (*vua* > *uva*; *en* 'en'/*en* 'ende' > *en/én*).

Por otra parte, no se nos escapa que los criterios de elaboración de un corpus están supeditados a sus objetivos. Dentro de la orientación hacia la historia lingüística en la que nos moveremos aquí casi siempre, cabe distinguir entre (a) corpus específicos y (b) corpus no específicos. Los primeros son los orientados hacia un objetivo concreto (elaborar un diccionario, una gramática histórica, estudiar la variación social en un ámbito determinado). La delimitación precisa de los objetivos es una clara ventaja metodológica (Arias Álvarez 2009), sobre todo en lo que concierne a la selección de los textos que integrarán el corpus. La orientación no específica presenta unos focos de interés más difusos. Por contra, implica un planteamiento más amplio y dinámico acerca del contexto mismo de texto y su utilización, lo que casa bien con la filología entendida como crítica del texto (y hoy no limitada a la edición) y que se orienta al estudio integral de éste. En el mismo sentido, ya hemos señalado que ni siquiera podemos prever qué utilización se hará de los corpus, aparte de que su uso no puede destinarse solo a los especialistas de una determinada disciplina.

¹⁵ Aunque resulte raro, la elaboración de corpus no se pone en relación con la metodología de edición. Desde nuestro punto de vista, el corpus, antes que una herramienta para ofrecer información, es un medio de edición innovador, en el que los textos aumentan su valor en la medida en que se articulan con otros para cubrir un arco cronológico amplio, un espacio geográfico y un espectro registral lo más completo posible, dentro de los objetivos que se hayan fijado previamente.

¹⁶ En el sentido de que la lectura directa de los textos, y más aun el trabajo de edición, nos pone ante los ojos estructuras sintácticas (y formas léxicas) sobre cuya carta de naturaleza en la lengua escrita y hablada de otras épocas es obligado dudar. No conviene dejar de registrar estos rasgos, por raros que parezcan, pues en otro momento de la investigación, o por parte de otros investigadores, tal vez se identifiquen como formas genuinas de la lengua antigua. Es el caso de *carrea* por *carrera* (en un documento de catedral de Toledo de 1235), a cuyo propósito me hace notar Fernando González Ollé en comunicación personal el interés de esta forma para trazar la historia de la «preposición» navarra y aragonesa *carria* 'hacia'.

¹⁷ En cambio, nos parece menos económico en un corpus recoger, en presencia de la lección correcta de un manuscrito, las variantes erróneas de otros códices.

2.2. En esta perspectiva metodológica se ha de plantear el problema de la cuantificación. Muchas de las objeciones nacen de la duda sobre la representatividad de los datos, pues el tratamiento estadístico acomuna datos que requieren valoración diferenciada. Muchas formas lingüísticas pueden ser convencionales, mientras que una que se presente aislada tal vez sea representativa de la lengua de uso. Ello puede ser así para el plano gráfico-fonético, por ejemplo en la atestiguación del seseo, aunque no es seguro que pueda atribuirse este valor a rasgos morfosintácticos. Si un texto se adhiere a las formas del sistema etimológico del pronombre átono objeto menos en una ocasión, ¿diremos que el texto, es, por ejemplo, leísta? ¿Y en las formas sometidas a variación en la lengua de uso, p. ej., el uso del artículo ante posesivo? (v. i. 5.3.). Y dentro de un mismo texto, ¿se han de computar más de una vez las formas? P. ej., ¿es significativo si un documento presenta la voz *campa* (Cantabria), el que aparezca una vez o cinco? Seguramente no, pero ¿negaremos que habrá que computar los varios empleos de *le* por *lo* en el mismo documento?

2.3. Hasta cierto punto, el problema de la representatividad y comparabilidad se supedita al de la tipología textual. Raramente los parámetros arriba señalados —cronológico, geográfico, social— se pueden establecer con seguridad en los textos literarios antiguos. ¿Son más representativos de una variedad lingüística determinada los documentos notariales? Sin lugar a dudas, una ventaja es la localización (como veremos, esta no siempre se muestra explícitamente). Sin embargo, ya hemos señalado en otros lugares (Sánchez-Prieto Borja en prensa a) la necesidad de ampliar el concepto de documento. Acogemos así tipos muy variados, que difícilmente pueden constituir una sola tradición discursiva en el sentido que se ha establecido (para una crítica de la aplicación de esta corriente a los textos archivísticos, *vid.* Díez del Corral Areta en prensa). Ello tiene una gran ventaja en cuanto que se amplía el ámbito sociolingüístico abarcado. En la cúspide se sitúan (1) los documentos de la cancillería; en el peldaño siguiente, (2) los jurídicos (pero registralmente la parte más baja la ocupan las «transcripciones» de la declaración de partes y testigos en juicio); a continuación, (3) los municipales; más abajo los (4) monásticos (con las cartas de compraventa y donación típicamente medievales), y por último, (5) los particulares¹⁸.

Mayores son las dificultades de validar cronológicamente los datos de estos documentos, no en el sentido de su emisión en un momento determinado (v. i. para la proporción de documentos datados), pero sí en cuanto a la interpretación respecto del uso efectivo en la lengua hablada. Porque los documentos no muestran todo el espectro comunicativo de la lengua, y de ahí que las variantes innovadoras se documenten mal, si nacen con una marca sociolingüística negativa, como *usted* y sus variantes *vusté*, *usté*, *osté*, etc. (García Godoy en prensa). ¿Qué significa que un rasgo o variante se documente hasta determinada fecha? ¿Qué entonces deja de tener vigencia, o qué dejó de tenerla tiempo

¹⁸ La cuestión no es tanto el que unos documentos reflejen usos más «cultos» que otros, sino cuál es el origen de las convenciones lingüísticas que muestran. En este sentido, se ha de señalar que el grado de localismo ha de leerse en sentido inverso, de más a menos, en esta escala. Ni mucho menos está representado en la escritura, máxime para la Edad Media, todo el espectro social: nobleza, clero, grupos urbanos (cargos en el *concejo*, *omnes buenos*). Por debajo, campesinos, obreros urbanos, sobre todo del s. xv en adelante, no están representados. Es seguro que si dispusiéramos del registro de estos últimos grupos la historia de la lengua sería otra. De ahí el valor de testimonios esporádicos, discontinuos, y aun breves que permitan vislumbrar como podía ser la lengua en el registro popular.

atrás?, y, en general, ¿cuánto tiempo antes? Lógicamente, para épocas pasadas, la datación en la lengua hablada solo es posible dentro de una hipótesis histórica. Pero para formular hipótesis al respecto es necesario saber siquiera hasta cuando se documenta un determinado fenómeno; alcanzaremos así parámetros de cronología absoluta, que pueden ser útiles para la datación de otros textos (Kawasaki en prensa).

El parámetro cronológico es casi siempre fácil de certificar en las fuentes documentales. No ocurre así con el geográfico. La tradicional postergación de las fuentes documentales en el estudio de los cambios lingüísticos, sobre todo sintácticos, desde la baja Edad Media, y más aun para los ss. XVI y XVII, y no digamos después, ha favorecido la idea de que no hay diferencias geográficas para los usos gramaticales del español moderno, cuando más bien la hipótesis de partida tendría que ser la contraria, la de que, en principio, todo cambio lingüístico sigue en su génesis y difusión un patrón geográfico. En realidad, como intentaremos mostrar con datos del corpus aquí examinado, habría dos modelos de difusión del cambio lingüístico: (a) areal, para los cambios de abajo a arriba en la escala sociolingüística y (b) registral, de arriba a abajo (lógicamente, el componente sociológico está por definición incluido también en a). La discriminación solo puede hacerse si se conoce el lugar en que se elaboró el texto, y para este viaje los documentos archivísticos son prioritarios respecto de las obras literarias.

3. EL *CODEA*: ESTADO ACTUAL Y DESARROLLOS PRÓXIMOS

3.1. Todas estas reflexiones metodológicas han estado presentes implícita o explícitamente en el origen y desarrollo del «Corpus de Documentos Españoles Anteriores a 1700» (*CODEA*). Ahora bien el *CODEA* no nació con la idea de divulgarlo como corpus independiente en internet, sino que surgió de la necesidad de integrar fuentes documentales en el «Corpus Diacrónico del Español» (*CORDE*). La primera fase de la edición se llevó a cabo entre 1996 y 1998. Finalmente, lo que por ahora llamaremos «ediciones críticas» de los documentos se incorporaron al corpus de la RAE entre 1999 y 2000. Pero la verdadera explicación de por qué se hizo el *CODEA* está en que conté en la Universidad de Alcalá con alumnos excepcionales: Cristina Castillo, Rocío Díaz, Carmen Fernández, José Manuel Lucía, Florentino Paredes, Carlos Sánchez y María Jesús Torrens, y lo que empezó siendo un trabajo de curso de cada estudiante sobre un documento archivístico tomó vuelos mayores en la serie *Textos para la historia del español*¹⁹. Fue fundamental la integración en el grupo de Bautista Horcajada. Nos atrevimos así a proponer a la RAE la elaboración de un gran corpus documental para el *CORDE*, propuesta que aceptó financiar su entonces director D. Fernando Lázaro Carreter. El equipo, con nuevas incorporaciones igualmente entusiastas (María Simón, Irene Vicente, Rocío Martínez, Ruth Miguel —latinista— y José Luis Ramírez) ha seguido durante años trabajando en las ya varias revisiones (lo que no quiere decir que esté exento de errores).

¹⁹ Hasta ahora, varios volúmenes publicados, sobre (i) fondos de diversos archivos, (ii) Archivo Municipal de Guadalajara, (iii) Municipal de Alcalá de Henares, (iv) Regional de la Comunidad de Madrid, (v) Municipal de Daganzo, (vi) Provincial de Guadalajara, (vii, en prensa) Municipal de Málaga, (viii, en prensa) Municipal de Arganda del Rey.

Con el tiempo, este entusiasmo juvenil por los documentos se renovó con nuevos argumentos en el marco, que se fue dibujando cada vez nítidamente, de una teoría de conjunto sobre la evolución de la lengua española y la constitución del español moderno. Los estudios acerca de la obra de Alfonso x, y en particular sobre la *General estoria*, pusieron de relieve la diferencia entre los escritos de la Cámara Regia y los documentos de la Cancillería. ¿Por qué esta diferencia? ¿Por qué muchos de los rasgos representados en los códices alfonsíes declinan, mientras que los de los documentos suelen tener más fortuna? Surgía así la necesidad de anclar en el espacio y en el tiempo los textos literarios, encontrar términos de comparación adecuados y buscar nuevos materiales en los que fundamentar el estudio del profundo proceso de renovación que lleva a la lengua moderna.

La proximidad del AHN fue decisiva en la orientación del corpus, pues facilitaba el objetivo de la comparación entre espacios distintos, y no limitarnos a un área. Pronto advertimos, sin embargo, la necesidad de hacerlo accesible en internet, y ello no solo por el valor que podía tener para la historia de la lengua. En este sentido, el interés lingüístico inicial se fue ampliando a medida que fuimos conscientes de que el estudio de esta no puede abordarse sin un anclaje sólido en la filología, entendida como crítica integral del texto. Así, el problema de la autenticidad «autorial» y la datación crónica y tópica ha de abordarse transversalmente, dentro de una indagación que considere los aspectos lingüísticos, desde luego, pero también los históricos y diplomáticos. El examen, pues, de los elementos de validación (diferentes para cada tiempo y espacio), tanto materiales (materia escriptoria) y figurativos (crismón, signo, rueda) como los textuales —dispositivos y validativos (fórmulas, nombres, firmas)— obligaba a ampliar el ámbito de estudio. A su vez, el examen de la lengua misma (grafía y fonética, morfología, sintaxis y léxico), no podía sino ampliarse hacia el nivel pragmático (o pragmlingüístico), lo que a su vez hacía necesario atender a las intenciones, valor, uso y contenido. Este último elemento se revelaba así como criterio de selección (ante la evidencia de que a mayor diversidad de asuntos, mayor variedad léxica y aun sintáctica). La variedad de contenidos se relacionaba asimismo con la amplitud registral. Finalmente, el corpus concebido de esta manera podía atraer a investigadores de diversas parcelas, pero antes que por los intereses sectoriales de paleógrafos, lingüistas, historiadores de la lengua, diplomatas, historiadores..., se justificaba el enfoque plural en el estudio íntegro del documento, campo natural del filólogo.

3.2. No es este, pues, un corpus exclusivamente pensado para el historiador de la lengua, lo que encaja con el sentido lato de documento que aquí se maneja, que identificamos con las fuentes archivísticas, y no solo con el diploma o documento regio, el documento administrativo y el legal. Se acoge pues el arco que va del privilegio rodado, por citar el instrumento público más solemne de la cancillería medieval, al billete particular o nota. Los contenidos reflejan lo divino y lo humano, pues antes o después se sintió necesidad de ponerlo todo por escrito. La distribución temática es transversal, en el sentido de que ni siquiera los documentos más solemnes se limitan a asuntos diplomáticos o de estado, sino que el legislador se sintió siempre tentado a intervenir en costumbres (pragmáticas de los ss. xv y xvi son CODEA 272, 284, 286, 305, 1193).

No falta en los antiguos tipos que nos pueden parecer modernos, como el «certificado médico»:

Certifico yo el doctor Luis Rodríguez, catredático de prima de medicina de esta universidad de Salamanca, que á muchos años que asisto a curar a mi señora doña Ana María de Espejo de muchas y graves enfermedades que á tenido, y de presente á más de tres meses que anda con calentura continua, con muchos accidentes muy a menudo, y que queda tres y cuatro horas sin sentido, y demás de esto le an sobrevenido una malencolía tan extraordinaria de que se puede temer pierda el juicio, y porque se an executado en esta señora varios y diferentes remedios sin aver sido de provecho, pienso ser necesario que salga de el convento así para hacer exercicio como para divertirse, que solo este remedio podrá restituirla la salud de que falta tiene. Y por ser así verdad lo juro a Dios y a esta (*cruz*). Salamanca, y 25 18 de abril de [...] (*CODEA* 1259, Salamanca, 1655).

La historia de las mentalidades hallará material abundante tanto en los documentos solemnes como en los particulares. En otro lugar hemos señalado algunos de los más llamativos o graciosos. Permítasenos referirnos aquí solo, para mostrar un tipo social que no falta en nuestros días, al hombre enriquecido pero cuyo nivel cultural no está a la misma altura, como se aprecia en esta carta sobre el recaudo de lienzos para la Armada, de indudable interés sociolingüístico²⁰:

Bien creo que *vuestraexcelencia* en este punto meacauan dedar lade*Vuestraexcelencia* y conella los dos pligos para cadiz y Gibraltar quese encaminaran luego con proprio ya *Vuestraexcelencia* sabe quela disculpa que ay denohauer pagado el lienço quede ay seasacado es nohauer prouehido su*Magestad. dineros. aestaproueeduria/* y assi meapesado enel alma deque agora seaya ydo a sacar sin *dinero*. pero ame forçado tanto aello el *señor* Adelantado quemeaobligado ahazer lo porser de tanta *Importancia*, el bestir la chusma, como *Vuestra excelencia* sabe/ y comosehallase lacantidad de brite queescrui a *Vuestraexcelencia* poca falta me haria Las lonas que *Vuestraexcelencia* me (*tachado*, escriue) dize (*sobre la línea*, d.) nosehallan Beso a*Vuestraexcelencia* lasmanos por la que meahecho conlanueua dela Armada que yua afrança/ aVnqueno es muy buena dios encamine el Resto della ylade buen Uiaje como sedessea/ y a *Vuestra excelencia* (*tachado*, con acrescentamiento) (*sobre la línea*) Guarde yacresçiente comolos Criados desucasa desseamos del puerto (*En vertical*) 1591 ***** en 27 de marco (*CODEA* 1148, El Puerto de Santa María, Cádiz, 1591).

3.3. En cuanto a la tipología documental, se acoge una amplia clase, que en el estado actual del *CODEA* se lista dentro del regesto («Privilegio rodado de Fernando II por el concede algunas heredades de Villa Mato y Villa Arias al Monasterio de Santa María de Vega», *CODEA* 370, de 1173), pero que en un desarrollo próximo constituirá un nuevo campo o se etiquetará para que pueda utilizarse como criterio/filtro para la navegación. No se pretende un listado exhaustivo, que dificulta la práctica de la asignación de tipo diplomático a cada documento, sino una nómina rentable, sobre todo, en términos lingüísticos, dentro de la hipótesis de que ciertos rasgos de lengua, vistos en clave registral, pudieran ir asociados al tipo de documento (la cuestión ha sido probada para los hechos gráfico-fonéticos en el privilegio rodado, notablemente continuista desde Fernando III a Fernando IV a través de tres cuartos de siglo, frente a las cartas plomadas, ya innovadoras bajo Sancho IV respecto de la

²⁰ Nótese 2 *pligos* ‘pliego’, 7 *poca falta me haría las lonas*, sin concordancia de número, y hasta la risible frase 8-9 «beso a vuestra excelencia las manos por la que me á hecho», con juego de palabras fallido entre el tratamiento de «vuestra merced» —sustituida por *vuestra excelencia*— y la omisión de *merced* ‘favor’ («la que me á hecho»).

cancillería de su padre Alfonso x. Provisionalmente, incluimos los siguientes (Puñal Fernández 2004): privilegio rodado, carta plomada, carta abierta, mandato, carta de compraventa, carta de donación, carta de merced, provisión real, apeo o deslinde, albalá, misiva, real cédula, pragmática, testamento, carta particular, papel. Esta lista tiene una operatividad limitada, pues los perfiles registrales (y, por ende, sociolingüísticos) no se corresponden con claridad con la tipología diplomática. Puede ser complementaria una distinción más elemental que tenga en cuenta el ámbito en el que se crea el documento, y así proponemos cinco clases básicas, ya operativas en el corpus digital *CODEA* 2011: documentos cancellorescos, jurídicos, municipales, eclesiásticos y particulares, que podrían corresponder, con matices, a una gradación de mayor a menor en la escala registral, pero quizá también a una escala de peculiaridad local de la lengua, con los mismos polos.

3.4. Arquitectura del *CODEA*. Archivos de procedencia

Hasta ahora se han incluido documentos de siete archivos, y está prevista la incorporación de otros fondos, como los de la Catedral de Toledo y el Archivo Municipal de Sevilla. La versión 2011 incluye las siguientes unidades:

- (1) Archivo Histórico Nacional 1060
- (2) Archivo General de Simancas 215
- (3) Archivo Municipal de Toledo 60
- (4) Archivo Histórico Provincial de Guadalajara 34
- (5) Archivo Municipal de Guadalajara 37
- (6) Archivo Municipal de Alcalá de Henares 29

Esta organización por archivos, con sus secciones y subsecciones, acierta a representar la arquitectura del corpus (puede verse en detalle en Sánchez-Prieto Borja en prensa a). Este acceso orgánico se complementa con otro alternativo en el que se listan todos los documentos por orden cronológico (con visualización directa de año y regesto). De este modo pretendemos superar el que consideramos uno de los defectos de la mayoría de las páginas electrónicas, que dan la opción de buscar sin que el usuario sepa qué puede encontrarse²¹.

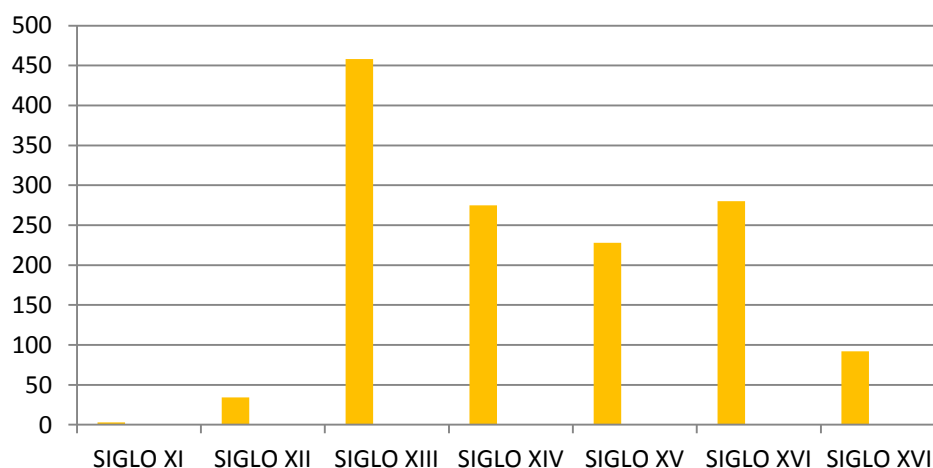
3.5. Criterios de selección: lengua, cronología y geografía

No es independiente de los parámetros cronológico y geográfico la cuestión de la lengua; en primer lugar, ¿solo debíamos acoger piezas en romance, de acuerdo con el nombre del corpus? Pensábamos así en un primer momento, pero pronto percibimos la dificultad de excluir lo latino si queríamos ofrecer la documentación de la época de los orígenes (lo anterior al s. XIII, y aun de la primera mitad del s. XIII mismo), pues incluso documentos puramente latinos tienen interés para la historia del léxico y la sintaxis del romance (por ejemplo, las piezas de la catedral de Toledo en latín contienen numerosos arabismos

²¹ No así en el «Portal de Archivos Españoles» (*PARES* <http://pares.mcu.es/>), donde, además de la búsqueda, es posible la navegación dentro de archivos, secciones y carpetas hasta llegar al documento mismo digitalizado). Es probable que la versión 2012 del *CODEA* incluya navegación estructural a través de su propia arquitectura.

incorporados al español). La versión 2011 incluye como criterio la separación entre documentos romances y latinos; también se han marcado en esta los segmentos identificables como lengua latina dentro de documentos romances, pero no cuando el sistema gráfico se mueve dentro de un código —o *scripta*— difícil de atribuir a un código u otro²².

En cuanto a la distribución por siglos, los datos son los que siguen: s. XI, 3; s. XII, 34; s. XIII, 458; s. XIV, 275; s. XV, 228; s. XVI, 280; s. XVII: 92; sin fecha: 82, lo que arroja un total de 1.452 (a 11 de enero de 2011). Los datos se resumen en el siguiente gráfico:



Progresivamente se irán completando los siglos con menos documentos, sobre todo el XII y el XVII. Con todo, la desproporción en número de piezas no perjudica en igual medida la comparabilidad, pues se ha de considerar también la extensión (muchas de la primera mitad del s. XIII tienen unas pocas líneas; algunas del XVII, un cuaderno).

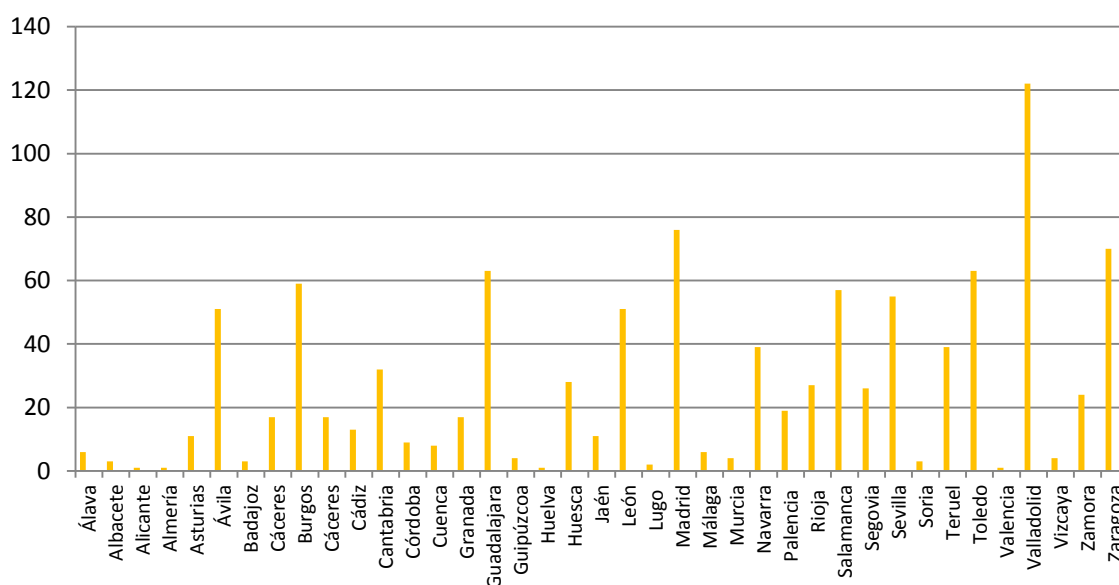
En orden a la fiabilidad cronológica del *CODEA*, se ha reflejado la fecha de los traslados, se ha marcado la duda acerca de la fecha de copias simples y se indica claramente cuándo la datación no es explícita sino deducida (*ca.* 1500). Los criterios de autenticidad pueden variar según emisores y época. En muchos diplomas de la catedral de Toledo hay firmas originales, mientras que no las habrá en privilegios o cartas plomadas.

El factor geográfico se concreta, lógicamente, no en el lugar de archivo actual, sino en el de emisión. Hemos optado por agrupar los documentos por las provincias actuales, pues aunque estas no reflejan sino la organización moderna del territorio, no deja de tener profundos anclajes históricos, además de ser el criterio más práctico y operativo²³. Un gran

²² Cabe separar la distinción neta de la difusa entre latín y romance. Ejemplo de la primera son las fórmulas latinas, muchas veces corrompidas, en contexto romance: «e quedar e aver por firme e cumplir e pagar todo lo qu'el dicho don Martín, abad del dicho monesterio de Matallana, procurador, sacare e fiziere e obligare e cumpliere e pagare todo lo que fuere judgado contra nós por esta personería *judicum sisti judicatum solvi*» (*CODEA* 403, Sevilla 1361); de la segunda, «Horta fuit contentio inter Gocher Stephanit e comitisse Agnetis de Fonte Euraldo super unum solarium deposado que est inter Villacent e Villaeta cum septem terras, e venerunt ad iudicium ante imperatorem e ante sua germana domna Sancia» (*CODEA* 363 [Monasterio de Santa María de Vega], 1097).

²³ Por ejemplo, el conde Pedro de León y la condesa Doña María dan unas tierras al monasterio vallisoletano de Santa María de Vega; la carta de donación fue emitida en el escritorio del conde, que firma de su puño y letra, así como su mujer, el día de Navidad de 1125 (*CODEA* 320). La pieza se conservó en el archivo del

número de piezas lleva en el *CODEA* la indicación «[s. l.]», puesto que el lugar de emisión no es explícito, pero en gran parte de ellos se sabe a ciencia cierta dónde se confeccionaron (caso de bastantes del Archivo Capítular de Toledo emitidas en esa ciudad, pero en las que solo se refleja la *data chronica*). La versión 2012 incluirá los lugares no acreditados por la data, pero de los que hay indicio sobrado sobre su origen (en todo caso, se ofrecerá dentro de la búsqueda y navegación por el parámetro geográfico la posibilidad de seleccionar «con lugar de emisión explícito»), de manera que el número de los no localizados es muy bajo²⁴. El cuadro, en la misma fecha de consulta que el cronológico, cuenta solo los documentos con lugar de emisión explícitos (para una exposición razonada de los criterios de la selección diatópica en un corpus documental, véase Sánchez-Prieto, Carrera, Isasi y Spence en prensa):



3.6. La edición de los documentos se ha llevado a cabo con criterios largamente ensayados, con los que básicamente coinciden los de la Red *CHARTA* (v. i. § 4). Su fundamento teórico y operativo está en la triple presentación de cada texto (facsimilar, paleográfica y crítica), acorde con la idea de que es imposible mediante una edición única proporcionar toda la información que los estudiosos esperan encontrar en un corpus, desde la forma de la letra al contenido, de la grafía a la sintaxis.

Es para nosotros un principio fundamental que los diferentes textos que integran un corpus no pueden estar editados con criterios diferentes y discrepantes, y es exigencia irrenunciable en un corpus primario.

No es corriente, ni siquiera en versiones electrónicas, que se ofrezca más de una edición del texto, que para fuentes documentales suele corresponder más o menos con la

monasterio y con la desamortización de Mendizábal pasó al Archivo Histórico Nacional, en cuya sección de Clero, provincia de Valladolid, mismo monasterio, se conserva hasta hoy.

²⁴ Ejemplo de la dificultad de localización son las «ordenanzas del gremio de pasteleros» (*CODEA* 1390), emitidas en Toledo el 8 de abril de 1625, pero que insertan la disposición del mismo gremio de Granada, y así reza el titulillo («Ordenanzas del gremio de pasteleros de la ciudad de Granada»), lo que puede dar lugar a confusión.

transcripción paleográfica (López Villalba 1998), casi siempre con desarrollo de los segmentos abreviados, que se marcan en cursiva.

3.7. A estos elementos básicos textuales múltiples se añade la «cabecera» (*head*), con unos datos mínimos, de acuerdo con el principio de que una plantilla exhaustiva es difícil de aplicar²⁵. De cada documento se indica (1) grupo investigador, (2) nombre del corpus, (3) archivo y signatura, (4) fecha y lugar de emisión, (5) regesto, (6) escribano y (7) transcriptor y revisores. Su ampliación es posible fácilmente; de hecho, ya la versión *CODEA* 2011 añade clases de documento, según se ha señalado. Elemento fundamental es el regesto, que sigue la estructura básica (1) tipo documental, (2) emisor, (3) asunto y (4) receptor (*CODEA* 6, de 1285, «Privilegio rodado de Sancho IV a los caballeros y *dueñas* de Guadalajara, que confirma otros dos de Alfonso X: uno sobre los excusados, los alardes y los pueblos que debían guardar, y otro que les eximía del pago de moneda»).

3.8. Desarrollo electrónico

El *CODEA* está disponible en la web desde el 2009 en versión de prueba y desde 2010 en versión estándar <http://demos.bitext.com/codea/> y <http://www.textohispanicos.es/>. La versión 2011 presenta un nuevo diseño y añade la búsqueda por las cinco clases documentales básicas antes establecidas. En cuanto a la preparación para la edición digital, se ha elaborado un XML no estándar. Hasta ahora no se ha introducido marcación TEI (se valorará su inclusión en el futuro, especialmente para indicar las manos que intervienen en la escritura del documento; una propuesta de etiquetado en Isasi 2010, así como en Spence, Isasi, Pierazzo y Vicente en prensa).

3.9. Recursos de navegación y búsqueda

La navegación es libre y gratuita en todo el corpus, de modo que, como se ha dicho, toda pieza es accesible en su integridad desde la caja «Documentos», que, sin otras restricciones de búsqueda o navegación ni contraseña alguna, contiene la lista ordenada cronológicamente de todas las piezas del *CODEA*:

²⁵ Un modelo muy completo es el del *TMILG* («Tesouro Medieval Informatizado da Lingua Galega»).



Es posible moverse en lista de documentos según los parámetros «Archivo» y «Provincia». Otras filtros combinables con los anteriores son la fecha (año concreto o fechas extremas) y el escribano, a los que la versión 2011 añade la tipología básica que separa documentos cancilleresco, jurídicos, municipales, eclesiásticos y particulares. El buscador es todavía básico en las versiones 2010 y 2011, y combina los mismos parámetros de búsqueda señalados para la navegación.

Como se dijo, aunque no se dispone de una clasificación acorde con la tipología diplomática, es posible seleccionar cualquier tipo documental mediante la búsqueda de la palabra o sintagma («privilegio rodado», «carta plomada», «pragmática», etc.). Desde la versión 2010 es posible la búsqueda de cualquier palabra en el texto, en el regesto o en ambos. Así, la búsqueda por «pragmática» en el regesto arroja una exigua pero sustanciosa lista; esta posibilidad tiene un interés indudable, pues, siguiendo con este ejemplo, las pragmáticas muestran por su temática una sintaxis menos formularia en sus partes argumentativas y dispositivas, y un léxico notablemente amplio, en tanto legislan sobre costumbres, como es el caso de la de la reina Doña Juana por la que prohíbe vestir telas de seda, de plata y oro a excepción de las personas de la realeza (*CODEA* 274, de 1515).

Todos los archivos
 Todas las provincias de emisión
 Documentos
 Búsqueda: En el registro
 Nº de Orden: Entre el año: y el año: Escribano:
 Transcripción paleográfica: ☒ Presentación crítica: ☒
 Resultados: 6. Mostrando del 1 al 6

Pragmática por la que Enrique IV baja la moneda a la mitad
 ID: 0272
 Signatura: AMTO, A.S. 586, cajón 8, legajo 1, nº 21
 Fecha: 1470 diciembre 24 (Segovia, España)

Pragmática de los Reyes Católicos en que se obliga a traer caballo y no cabalgar en mulas ni otras bestias
 ID: 0264
 Signatura: AMTO, A.S. 591, cajón 8, legajo 1, nº 26
 Fecha: 1499 septiembre 30 (Granada, España)

Pragmática de los Reyes Católicos por la que regulan el precio del trigo, la cebada y el centeno
 ID: 0266
 Signatura: AMTO, cajón 8, legajo 1, nº 32
 Fecha: 1503 enero 3 (Madrid, España)

Pragmática de la reina doña Juana en la que prohíbe vestir telas de seda, plata y oro a excepción de las personas de la realeza
 ID: 0274
 Signatura: AMTO A.S. 602, cajón 8, legajo 1, nº 37
 Fecha: 1515 julio 20 (Burgos, España)

Pragmática por la que el rey don Carlos prohíbe matar terneras para evitar la escasez de carne. Hecha pública por el pregonero del rey el 6 de septiembre de 1525.
 ID: 1193
 Signatura: AGS, Patronato Real, legajo 70, documento 21 (n. 6442)
 Fecha: 1525 agosto 27 (Toledo, España)

Pragmática de Carlos I y su madre Juana sobre los trajes
 ID: 0305
 Signatura: AMTO, A.S. 616, cajón 8, legajo 1, nº 51
 Fecha: 1537 junio 29 (Valladolid, España)

3.10. Desarrollos próximos

3.10.1. En cuanto a los parámetros para navegación y búsqueda, está previsto incluir en la versión *CODEA* 2012 la tipología diplomática amplia (según la lista antes presentada: privilegio rodado, carta plomada, provisión real, etc.).

3.10.2. Desarrollo del buscador

- (1) La búsqueda por lemas multiplica las posibilidades de explotación de un corpus (la lematización la está llevando a cabo Bautista Horcajada Diezma). Como es lógico, se hará sobre el texto crítico, pues las regularizaciones gráficas, así como la unión y separación de palabras y la acentuación, y la menor variabilidad gráfica respecto de la transcripción paleográfica, facilitan la tarea. No hay acuerdo en todos los casos sobre qué ha de entenderse, en la práctica, por lema en un texto antiguo. A primera vista parece facilitar la explotación de un corpus el que se aplique un concepto lato. No es este un problema irresoluble, pues siempre está la alternativa de buscar por formas exactas, pero sí es verdad que, en general, ambos niveles son complementarios: por formas, se pueden estudiar aspectos gráfico-fonéticos; el rastreo por lemas conecta mejor con el examen de la sintaxis y el léxico; así, por próxima que esté a la forma estándar *especificada*, seguramente no se nos ocurriría buscar la forma exacta *espacificada*; para evitar este inconveniente, encontraremos *nuef* con solo buscar *nueve*, de manera que no tendremos que pensar en todas las variantes formales no regularizadas en la presentación crítica (en cambio, es obvio que no aparecerá en esta *dyxo*, sino *dixo*), ni en las variantes con posible valor fonético (*gualardón* y *galardón*), o en las morfológicas (*coñocida* y *coñoçuda*, con el agravante de la variación entre nasal palatal y no palatal). Ello no quiere decir que la lematización no presente problemas ni exija tomar decisiones arriesgadas. Así, ¿listaremos *pora* como lema distinto de *para*? Aun a riesgo de dificultar a veces la consulta para el no especialista, hemos preferido separar como lemas distintos *pora* y *para*, pues sus valores semánticos —sin hablar

de los orígenes— no coinciden, ya que en muchos casos la forma más antigua corresponde a ‘por’. En cambio, nos parece conveniente agrupar bajo el mismo lema lo que llamamos variantes formales del lexema, por ejemplo, *aciprés* y *ciprés*, pero no las formas prefijales como *correr* y *acorrer*, por implicar probablemente una divergencia semántica que acabó cristalizando en el léxico.

- (2) El buscador de *CODEA* 2012 permitirá búsqueda por secuencias de palabras continuas y discontinuas.
- (3) También será posible localizar segmentos de la palabra, tanto iniciales como interiores o finales, sin que necesariamente correspondan al análisis morfológico (*renunciamos/renunciamos/renunciamos*).

3.10.3. Salida a mapa

Los resultados de las búsquedas se visualizarán en un mapa, donde se marcan con puntos los lugares de emisión de los documentos (la herramienta se está preparando con la colaboración de Bautista Horcajada).

3.10.4. Audio de una parte de los documentos

Se espera ofrecer una lectura por voces de diversa procedencia, que corresponderán de manera aproximada al lugar de emisión del documento.

3.10.5. Otro recurso que se incorporará será la salida estadística, de modo que los resultados de una búsqueda se muestren cuantificados. Ejemplo podría ser la oposición *el mi/Ø mi*, es decir, la segunda invariante excluye *el* ante *mi* (nótese que al buscar por lemas se incluyen todas las formas del artículo: *el, la, lo, las, los...*); se podrá obtener como resultado la proporción, expresada en % sobre el total de la variantes, es decir, (a) + (b); nótese, además, que la búsqueda podrá hacerse en el conjunto del corpus o en una parte, de acuerdo con los filtros posibles: por archivo, límite temporal, geográfico, por tipo documental, y aun escribano.

4. EL CODEA EN EL MARCO DE CHARTA

El *CODEA* adquirirá una nueva dimensión al integrarse, según lo previsto, en *CHARTA*, acrónimo de «Corpus Hispánico y Americano en la Red: Textos Antiguos», que propusimos en 2006 para dar nombre a la red internacional dedicada al documento archivístico, y que nació con los objetivos siguientes:

- (1) Establecer unos criterios de edición de fuentes documentales.
- (2) Elaborar un corpus de documentos españoles y americanos de los orígenes al s. xix inclusive.
- (3) Promover el intercambio entre grupos de investigación de diversos países sobre el documento hispánico, así como favorecer la transferencia metodológica con grupos y redes dedicadas al documento en otras lenguas.

La red *CHARTA* (<http://www.charta.es>) está integrada a día de hoy por los siguientes grupos (indicamos también el coordinador): Universidad de Alcalá (Coordinador de la Red: Sánchez-Prieto); Consejo Superior de Investigaciones Científicas (Torrens); Universidad de Deusto (Isasi); Universidad de Murcia (Díez de Revenga), Universidad de Salamanca (Sánchez González de Herrero), Universitat de les Illes Balears (Enrique-Arias), Universidad de Las Palmas de Gran Canaria (González Monllor); Universidad de Valladolid (Carrera); Göteborgs Universitet (Söhrman); King's College London (Spence); Université de Neuchâtel (Sánchez Méndez); Università di Padova (Rivarola); Universidad de los Andes (Obediente); Universidad de Tokio (Ueda); Universidad Nacional Autónoma de México (Arias); Universidad de Málaga (Carrasco), Universidad de Granada (Calderón).

Los criterios han sido ya aprobados, inscritos en el Registro de la Propiedad Intelectual de Madrid, publicados en la web citada, y acaban de aparecer en formato libro: «Criterios de edición de documentos hispánicos (orígenes - siglo XIX)», en *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*²⁶.

Algunos aspectos relevantes de estos criterios son la expresión «autorial», bajo la forma de (1) emisor que tiene la *sanctio*, y que se expresa de forma reglada, siempre que es posible, en la posición segunda del regesto: «Carta plomada del rey Sancho IV por la que...» y (2) notario o escribano responsable de la escrituración, pero que puede incluir también otro dato cual es (3) el amanuense. La expresión de esta «autoría compartida» puede llevarse a cabo mediante elementos añadidos en la cabecera descriptiva, y su formulación está prevista en TEI (Sánchez-Prieto, Carrera, Isasi y Spence en prensa). También cabe destacar el que se indiquen las correcciones de diferentes manos, así como la intervención editorial. Se pretende establecer un protocolo de edición electrónica que genere automáticamente la presentación crítica, mediante reglas de correspondencia entre los dos niveles del texto.

En cuanto a la integración del *CODEA* en *CHARTA*, los documentos se irán incorporando progresivamente, a medida que los vayamos adaptando a los criterios de la red, solo en parte diferentes de los utilizados para el *CODEA*. Este será, pues, un subcorpus dentro de *CHARTA*. Un objetivo inmediato es incluir nuevos documentos en *CODEA*, de modo que vayan completándose las lagunas geográficas y cronológicas que presenta la versión actual. Los desarrollos de los dos corpus serán, pues, paralelos, pero seguramente avanzaremos en el *CODEA* herramientas tecnolingüísticas a modo de ensayo.

5. EL *CODEA* COMO FUENTE DE DATOS

La investigación lingüística histórica encuentra su fundamento empírico en el uso de corpus, y ello se muestra con mayor evidencia en los niveles sintáctico y léxico. Para este último ámbito parece innecesario trabajar con corpus específicos, y, tampoco se exige que estos sean primarios ni orgánicos, sino que se considera preferible la amplitud en la que encontrar los datos que se buscan u otros no previstos inicialmente. La calidad y rigor dispares de las ediciones que se incorporen debería obligar a un control de las documentaciones, pero ello implica una conciencia filológica de la que muchos lingüistas e historiadores de la lengua

²⁶ Igualmente, *CHARTA* ha sido uno de los promotores, junto con la RAE y el Cilengua de la Fundación San Millán, de las normas para *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*, incluidas en el mismo volumen que las de fuentes documentales (Sánchez-Prieto Borja 2011b).

carecen. Este modo de ver las cosas parece verse corroborado por el tratamiento estadístico. Ante una inmensa masa de datos, ¿qué importancia tiene que unos pocos ejemplos sean anómalos, erróneos o directamente falsos? Para la documentación de una forma rara quisiéramos contar con corpus amplísimos. Pero es obvio que la investigación no acaba aquí. El léxico, y cualquier otro nivel de la lengua, pueden estudiarse en una perspectiva areal (v. i. § 5.4.). Ello precisa de textos con localización espacial precisa. Así, un macrocorpus como el *CORDE* delimita grandes dominios (los países hispanohablantes), pero no incluye divisiones menores, lo que dificulta enormemente la interpretación de los datos. Así, un estudio de los finales *-nt* arrojará datos sorprendentemente altos para el s. xiv si el corpus en cuestión incluye muchos textos aragoneses, sobre todo notariales, con lo que queda oculta la realidad de que en Castilla son raros en esa centuria. La estadística viene así a sepultar las diferencias espaciales.

Pero es en el plano de la sintaxis donde el rigor del corpus es una exigencia perentoria, porque la variabilidad es alta respecto de los parámetros cronológico, temporal, social y aun de idiolecto. La comparabilidad de los textos no viene definida de antemano. Es imposible documentar toda la variedad de la sintaxis histórica en los documentos, por lo que sería disparatado llegar al extremo opuesto de descartar los textos literarios, pero parece claro que los documentos archivísticos se prestan mejor a una cuantificación de sus rasgos, en especial sintácticos. Lo ideal sería, pues, contar con corpus rigurosos y amplios, de variada tipología textual, y en los que pudieran filtrarse las consultas por clase de texto. Pero pensamos que en la práctica de la investigación será frecuente utilizar varios corpus.

Aun en el estado actual, el *CODEA* puede ser una fuente de datos relevantes para el historiador de la lengua española y de sus variedades históricas. Las lagunas geográficas temporales son una merma significativa en el ámbito léxico y sintáctico, mientras que este déficit es mucho menor en el plano grafemático y fonético. De hecho, los documentos del *CODEA* se han utilizado ya en diferentes estudios. Entre otros muchos, sobre el establecimiento de parámetros para la datación de textos no datados (Kawasaki, en prensa), la grafemática (Sánchez-Prieto 2004), la apócope (Ueda, en prensa), los participios en *-udo* (Pato y Felíu Arquíola 2005), los pronombres átonos (Matute 2004), el indefinido *alguien* (Pato 2009), las formas *otro, otri, otre, otrie* (Paredes García en prensa), la preposición *a* con valor 'en donde' (Sánchez-Prieto Borja 2000), la geografía léxica de la Edad Media (Sánchez-Prieto Borja 2011a), el nombre de persona (Simón Parra 2009), o los pronombres de cortesía (Shishida 2010). Aparte de estos, son muchos los aspectos que pueden estudiarse. Presentamos a continuación una breve muestra de las posibilidades de estudio que ofrece el *CODEA* en los diferentes niveles.

5.1. Historia de la escritura, Escriplogía y Dialectometría

La inclusión de facsímiles en la versión 2011 (la de 2010 solo contaba con los del Archivo Municipal de Toledo) permite llevar a cabo estudios de paleografía; pero, lejos de la concepción tradicional, esta puede insertarse en la perspectiva más amplia de la historia de la escritura, basada en la correlación e interacción entre los niveles (a) paleográfico, (b) gráfico y (c) fonético. En otros lugares hemos ilustrado esta correlación con la existencia de dos líneas en la escritura alfonsí, la de la chancillería y la de la cámara regia; la primera adopta una minúscula diplomática de astiles prolongados; la segunda se configura en torno al

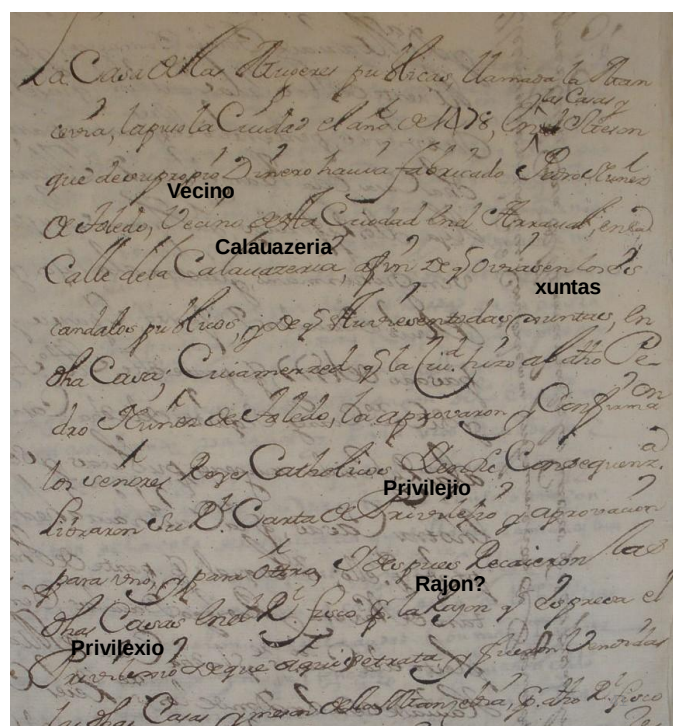
empleo de la gótica libraria o *textualis*, con escasa separación entre renglones y astiles cortos. Ambos tipos llevan aparejadas opciones gráficas diferentes, pues mientras la primera muestra formas angulosas (*visto*) y largas de las letras (*concejo*), la segunda favorece la homogeneidad del renglón, con preferencia por los trazos redondeados (*uisto*, *conceio*). Además, y ello otorga especial relieve a estas dos corrientes, se asocian a usos lingüísticos dispares (v. i. § 5.2.2. sobre la apócope).

En el plano general de los comienzos de la escritura romance, este corpus puede contribuir en su estado actual a identificar los escritorios donde se empieza a escribir de manera plena el romance, gracias a la rica documentación del s. XIII incluida, mientras que para la coexistencia entre latín y romance será innecesario incorporar más materiales del XII, hasta ahora débilmente representado.

En una perspectiva escriptológica (Goebel 1975), el CODEA puede contribuir a identificar las *scriptae* peninsulares; así, adelantamos que las tradicionales tres áreas leonesas, occidental, central y oriental, pueden corresponder con dos tradiciones de escritura, (1) centro-occidental y (2) oriental. El examen escriptológico parece confirmar también que Navarra y Aragón representan dos tradiciones de escritura diferentes en los usos monásticos, lo que no quiere decir necesariamente que se trate de dos variedades dialectales (Sánchez-Prieto Borja, en prensa b).

Como muestra de la interacción entre los niveles fonético, gráfico y paleográfico, rastreables en este corpus, véase en este fragmento de 1581 la confusión entre las grafías de sibilantes:

CODEA 270



AMTO
A.S. 343, c.4, l. 1, n.º 49
Toledo, 1581

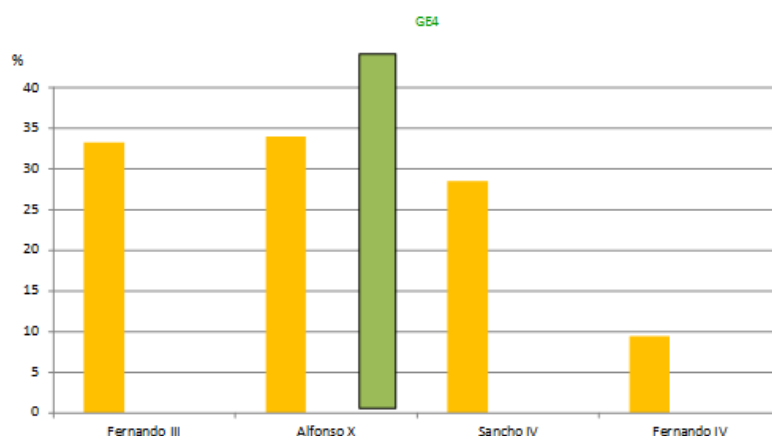
5.2. Fonética

5.2.1. El estudio del nivel fonético puede empezar por la cronología absoluta de las documentaciones. Un ejemplo puede ser el de los testimonios tardíos de *Castiella* en el área hispánica abarcada por el *CODEA*. La lista incluye los testimonios de 1400 y posteriores:

- (a) Pedro Pérez, clérigo de Potes, torna al Monasterio de Santo Toribio una viña y huerto que tenía arrendada al cuarto ID: 0735 Signatura: AHN, Clero, Santander, carpeta 1922, n.º 6 Fecha: 1400 marzo 5 (Santo Toribio de Liébana, monasterio de Cantabria).
- (b) Carta de confirmación de Enrique III de una carta otorgada por su padre Juan II el 6 de septiembre de 1380, en la que sostiene que las aljamas de Abanilla y Alguaza (Murcia) están libres de pagar maravedís ID: 1343 Signatura: AHN, Clero, Murcia, carpeta 1397, n.º 5 Fecha: 1406 agosto 23 (Valladolid).
- (c) Juan II confirma un privilegio otorgado por Fernando III al Monasterio de Valbuena, que es confirmado en sucesivas cartas por Alfonso X, Sancho IV y Fernando IV ID: 0345 Signatura: AHN, Clero, Valladolid, carpeta 3441, n.º 12 Fecha: 1407 agosto 26 (Segovia).
- (d) Carta de confirmación de los privilegios, fueros y demás prerrogativas del arzobispo, deán y cabildo de Toledo otorgada por Juan II. Contiene confirmaciones anteriores de Enrique II y Enrique III. ID: 0545 Signatura: AHN, Sellos, caja 3, n.º 22 Fecha: 1408 diciembre 31 (Guadalajara).
- (e) Testamento de Pedro Maitea ID: 0993 Signatura: AHN, Clero, Navarra, carpeta 1413, n.º 13 Fecha: 1417 enero 31 (Sangüesa, Navarra).
- (f) Carta de recepción de Suer Alfonso como nuevo vecino de Calatayud ID: 0919 Signatura: AHN, Clero, Zaragoza, Iglesia de Santa María de Calatayud, carpeta 3619, n.º 7 Fecha: 1417 agosto 6 (Calatayud, Zaragoza).
- (g) Carta de confirmación de una carta dada el 24 de agosto de 1328 en la que Juan II manda que no se embargue a los mensajeros del Monasterio de Santo Toribio de Liébana ID: 0739 Signatura: AHN, Clero, Santander, carpeta 1920, n.º 5 Fecha: 1417 septiembre 26 (Burgos).
- (h) El Monasterio de Santa María de Nájera entrega unas casas a Diego López de Ruego pagando como tributo y encenso dos florines de oro ID: 0612 Signatura: AHN, Clero, Logroño, carpeta 1036, n.º 21 Fecha: 1418 febrero 8 (Nájera, La Rioja).
- (i) Carta del rey Juan II mediante la cual confirma todo lo contenido en otra carta suya dada al Monasterio de Santa María de Aguilar de Campó en Alcalá de Henares, el 26 de junio de 1403, en la que les otorga y confirma todos los fueros y privilegios que ID: 1183 Signatura: AHN, Clero, Palencia, carpeta 1682, n.º 8 Fecha: 1420 abril 3 (Valladolid).
- (j) Lope García de Quintana, con poder del condestable de Castilla Bernardino de Velasco, vende al convento de Santa Olalla una campa y heredad con la casa y ermita de San Pelayo ID: 1405 Signatura: AHN, Clero, Santander, carpeta 1913, n.º 2 Fecha: 1495 julio 26 (Castro Urdiales, Cantabria).

Como se ve, no puede decirse que los testimonios se concentren en un área determinada, ni que una zona u otra sea particularmente conservadora; prima, pues, el factor cronológico, acerca del cual puede concluirse que de las 10 documentaciones más tardías 9 se concentran entre 1400 y 1420, mientras que la última es de 1490. De ello puede concluirse, de manera provisional y a falta de un examen más detallado que combine varios parámetros, que un texto no datado que presente *Castiella* tiene alta probabilidad de ser de hacia 1420 como tarde, al menos si es un documento administrativo.

5.2.2. Es cosa sabida que los códices alfonsíes tienen índices de apócope relativamente altos (datos en Harris-Northall 1991, Moreno Bernal 1993). Sin embargo, ¿los diplomas coetáneos muestran índices similares? Para dirimir esta cuestión hemos buscado las palabras que ofrecen porcentajes más altos de pérdida de la vocal final en los diplomas cancillerescos de Fernando III, Alfonso X, Sancho IV y Fernando IV. Como término de comparación hemos utilizado la Cuarta Parte de la *General estoria* (1280), de la que disponemos de transcripción paleográfica revisada con el manuscrito por nuestro grupo de investigación *GITHE* y por Gago Jover (<http://hispanicseminary.org/t&c/ac/index-es.htm>). Excluimos de la búsqueda en el texto historiográfico alfonsí *pendient*, un tecnicismo diplomático usual en los diplomas —*sello pendiente*— pero no en los códices, y *veint*, que en el código aparece en romanos (xx). Los resultados son los siguientes:



(a)delant, cort, end, fuent, -m(i)ent, mont, muert, part, pendiente, puent, regnant, suert, veint

General estoria, Cuarta Parte (1280): 46,74 % [excepto *pendient(e)* y *veint(e)*]

5.3. Morfología y sintaxis

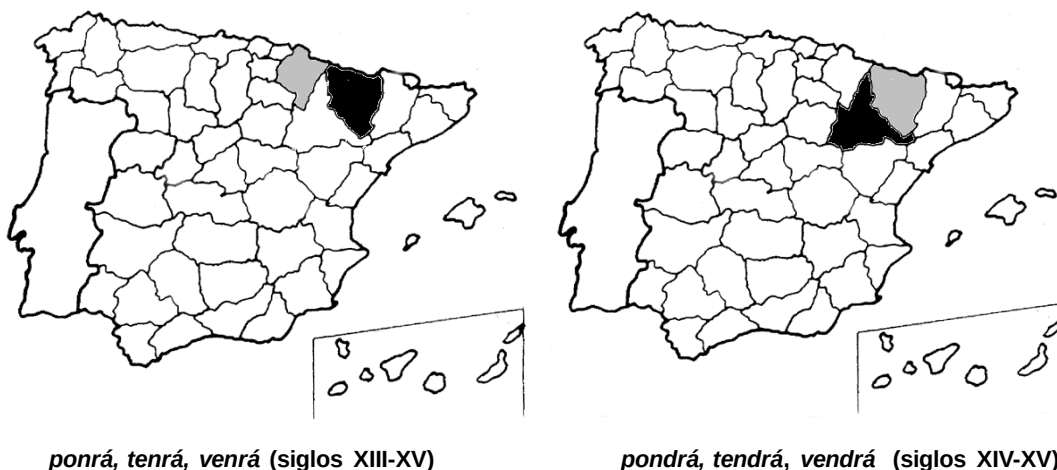
El objetivo general que podemos marcarnos en los dominios de la morfología y la sintaxis es cómo se manifiestan las tres coordenadas fundamentales de la variación en los textos, y en particular en las fuentes documentales incluidas en el *CODEA*: diacrónicas, diatópicas y diastráticas. El intento de establecer una cronología absoluta de los cambios morfológicos y sintácticos solo puede aspirar, de momento, a acotar aproximadamente el espacio y el tiempo de la variación. La aplicación de este parámetro temporal nos lleva a cuestionar la continuidad del castellano antiguo. Suponemos un modelo relativamente constante para la evolución de la lengua hablada, mientras que la escrita podría proceder por saltos, en tanto

sometida más directamente que la hablada a cambios culturales. Ello nos lleva a introducir el componente de la tipología textual. Si encontráramos indicios paleográficos y fonéticos de la falta de uniformidad del castellano alfonsí, ¿se observan también discrepancias sintácticas entre códices y diplomas?

Para los cambios morfológicos y sintácticos, la hipótesis de partida es que las variantes tienen una distribución geográfica, planteamiento no habitual entre los historiadores de la lengua sobre todo para aquellos procesos que se consideran relativamente tardíos. ¿Con todo, el modelo geográfico es general, o más bien es aplicable a unos cambios pero no a otros? Trataremos, si es posible, de determinar en cuáles se percibe un componente espacial y en cuáles este es irrelevante. Otra posibilidad explicativa podría estar en las limitaciones del método: determinada variación puede ser sensible al componente geográfico, pero nuestro método empírico (basado en el examen de las fuentes documentales) no permite, al menos por ahora, detectarlo. Subyace, pues, la pregunta de si la variación en los textos es la de la lengua de uso. Unas pocas calas en el *CODEA* acerca de variantes representativas pueden, tal vez, ayudarnos a responder a estas cuestiones, o al menos a plantearlas mejor.

5.3.1. Un cambio bien conocido en la morfología del verbo en español es el que afecta a los futuros analíticos (VENIRE HABEO), que, una vez alcanzada la forma sintética y consiguiente pérdida de la vocal intertónica (*venré*), conocen un resultado metatético (*verné*) o bien epentético (*vendré*). Por descontado, las formas se solapan en el tiempo, dentro del proceso de variación característico de constitución de los paradigmas verbales. Moreno Bernal (2003) examinó los orígenes de este desarrollo y adelantó la hipótesis de que la extensión de variante moderna (*vendrá*, *pondrá*, *tendrá*, etc.) se da antes en Aragón que en Castilla. El estudio de Moreno Bernal combinaba el examen de textos literarios y fuentes documentales, entre estas algunos de los documentos que forman el *CODEA*. Con los recursos que permite ahora la presentación electrónica, hemos examinado este corpus, y ello permite alcanzar la conclusión de que la hipótesis de la prioridad aragonesa de las formas epentéticas se corrobora plenamente. Si admitimos que el origen de estas formas está en las que tienen la secuencia *-nr*²⁷, interesará conocer si la forma en *-nrá* tiene una distribución coincidente con la epentética. Los datos de *CODEA*, presentados de manera muy simplificada tanto en lo temporal como en lo geográfico, muestran que *ponrá*, *tenrá*, *venrá* se localizan entre los ss. XIII y XV en la actual provincia de Navarra y más aun en la de Huesca, mientras que *pondrá*, *tendrá*, *vendrá*, en los siglos XIV y XV, están presentes en Huesca y alcanzan su mayor intensidad en Zaragoza:

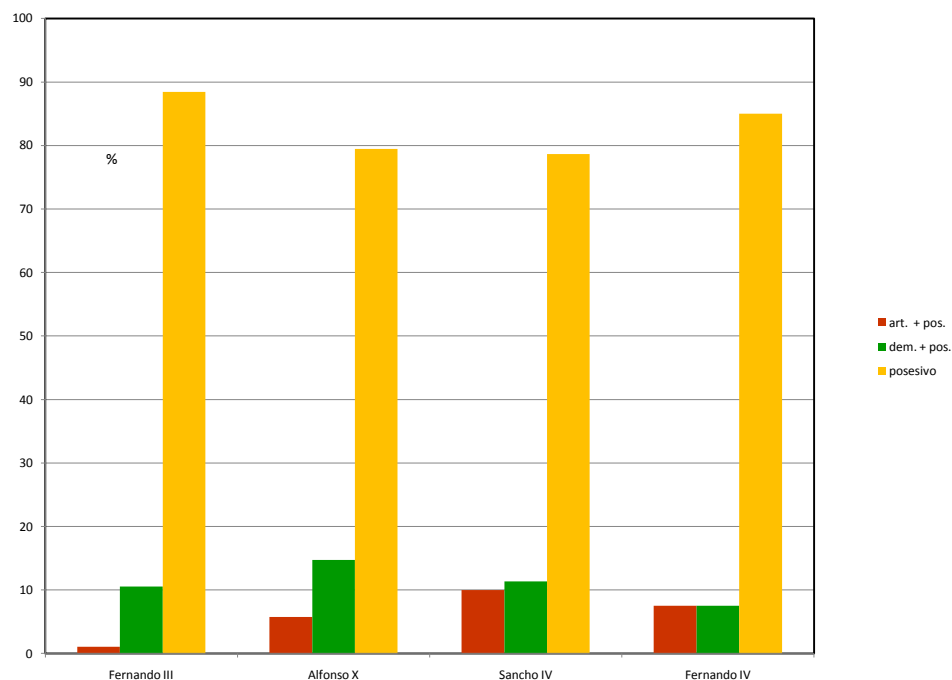
²⁷ Bautista Horcajada Diezma me señala que *venrá* fue una manera de escribir */bendrá/*. Es probable que esto sea cierto en parte, pero no creo que se pueda descartar una lectura de nasal + vibrante múltiple (cf. *onra*, *Enrique*, o el medieval *enrizar*; que la secuencia es espontánea del español parece mostrarse en el vulgarismo *enritar* por *irritar*).



Puede concluirse, por tanto, que el origen del futuro epentético hay que situarlo en el ámbito navarro-aragonés, pues son precisamente las formas en *-nrá* las que parecen estar en el origen de las en *-ndrá*, sin descartar, claro, que la forma sin *-d-* pudiera leerse como *-ndrá*. El cambio, pues, es claramente sensible al componente geográfico.

5.3.2. En una perspectiva diferente situamos el examen de la secuencia artículo + posesivo en los documentos de la cancillería castellana de Fernando III, Alfonso X, Sancho IV y Fernando IV. Ya Lapesa señaló que el uso estaba supeditado al tipo de texto. Enrique-Arias (2010: 78), que estudia esta variación en las versiones bíblicas medievales, señala la progresión desde los textos narrativos, sapienciales y proféticos hasta los poéticos, con índices superiores en la *General estoria* (ca. 1275-1280) respecto de Esc. I.I.6 (ca. 1250). Así, *Macabeos 1* presenta índices del 12% y 13,8% respectivamente, y *Cantar de los Cantares*, del 47,30% y 77,50% (con la precisión necesaria de que ninguno de estos libros de la *General estoria* se conserva en testimonio original procedente de la Cámara Regia).

En los que respecta a los documentos cancillerescos, se ha de notar que *el mi* tiene la competencia de *este mi*, con lo que la oposición netamente lingüística es entre sustantivo sin y con actualizador, y dentro de esta última, entre los distintos elementos con capacidad actualizadora, y que tiene uno de sus polos en las dos secuencias que aquí consideramos. En nuestros diplomas, la mayoría de los casos de demostrativo + posesivo son de *esta mi carta*, que se repite muchísimo. La secuencia de artículo + posesivo encuentra una restricción importante, y es la que establecen los nombres de parentesco (*mi(o) padre, mi(o) avuelo*), pues no se da en estos casos, de acuerdo con una comportamiento románico típico, como se ve en italiano. Cabe señalar un mayor empleo de artículo ante posesivo de primera persona (*el mi*) que de segunda (*el tu*). De manera simplificada, los datos son los que se observan en este gráfico:



Se aprecia que el posesivo con nombre alcanza en los tres cuartos de siglo examinados unos índices de frecuencia de entre el 79% y el 89%, mientras que la secuencia de demostrativo + posesivo se sitúa por encima del 10% en el reinado de Fernando II y alcanza el 15% con Alfonso X, para caer ligeramente bajo Sancho IV (11%) y algo más pronunciadamente con Fernando IV (8%). Por último, resulta curioso que artículo + posesivo apenas alcance el 1% en la cancillería sanfernandina, para elevarse hasta el 6% en la alfonsí, llegar al 10% en la de Sancho IV y bajar al 8% en la de su hijo. Las cartas reales presentan índices más bajos de la secuencia artículo + posesivo que los textos históricos de las biblias romanceadas, y quedan mucho más lejos de los poéticos (*Cantar de los cantares*). La tipología textual es un factor relevante en esta variación; dada la naturaleza de los textos examinados, podemos suponer una escala que fuera de denotativo a connotativo, de objetivo a subjetivo, de estilísticamente no marcado a marcado. No podemos concluir, sin embargo, que unos tipos textuales u otros reflejen mejor la lengua hablada, pues los polos señalados están todos ellos presentes en esta.

5.3.3. Volvemos a considerar la pertinencia o no del parámetro espacial con el estudio del desarrollo de los elementos de relación, conectores y marcadores discursivos en la lengua jurídica y administrativa. En concreto, examinaremos los marcadores que incluyen *tanto*. Este elemento forma en la Edad Media secuencias de valor cuantitativo: *por ál tanto* ‘por otra cantidad igual’; *por tanto*, *en tanto* ‘por tal cantidad’, *dos tanto*, *otro tanto*, etc. Desde la segunda mitad del s. XIV, *tanto* se empieza a usar en expresiones temporales y modales, indicadoras de la modalidad discursiva, que se acomoda muy bien a la configuración del texto administrativo y legal, caracterizado por un período frásico largo, sintácticamente complejo y que engarza varias oraciones. A veces la nueva secuencia es una alternativa a los usos medievales, que continúan empleándose, como *por ende*, que pierde vitalidad y queda

reservado para el discurso jurídico, mientras que *por tanto* conoce un desarrollo muy amplio que trasciende hoy el ámbito de la lengua escrita.

Presentamos la documentación en el CODEA de las secuencias *(h)asta (en) tanto (que)*, *en tanto (que)*, *por tanto* (usado en los documentos para «concluir y encerrar razones») ²⁸ y *con tanto que* ‘con tal que’:

(h)asta (en) tanto (que) 1395 Calatayud (Zaragoza) [“fasta tanto que”] 1404 Zaragoza 1405 Calatayud (Zaragoza) [“entro a tanto que”] 1485 Valladolid 1492 Toledo 1521 Vitoria (Álava) 1549 Badajoz 1576 Toro (Zamora) 1626 Andújar (Jaén) en tanto (que) 1416 Velilla de Jiloca (Zaragoza) 1458 Albarracín (Teruel) 1458 Toledo 1472 Jerez de la Frontera 1478 Granada 1479 Béjar (Salamanca) 1494 Madrid 1500 Huesca 1520 Hernani (Vizcaya) 1549 Badajoz entre tanto (que) 1469 Alcañiz (Teruel) [Enrique IV] 1470 Segovia [“entre tanto que se da orden”] 1492 Toledo 1508 Sevilla 1515 Burgos 1521 Vitoria (Álava) 1544 Valladolid 1552 Madrid 1544 Valladolid 1564 Madrid	por tanto 1364 Mayorga (Valladolid) [“por tanto yo torno al dicho prior... en la tenencia”] 1496 Calatayud 1497 Huesca 1499 Salamanca [ver] 1503 Santo Toribio de Liébana (Cantabria) 1506 Calatayud (Zaragoza) 1526 Zaragoza 1529 Zaragoza 1531 Santa María de Piedra (Zaragoza) 1532 Toledo 1539 Sevilla 1586 Salamanca con tanto que ‘con tal que’ 1500 Granada 1506 Bilbao 1505 Vega de Espinareda (León) 1512 Cáceres 1515 Burgos 1526 Granada 1537 Valladolid 1528 Trujillo (Cáceres) 1549 Guadalajara 1592 Burgos
--	--

De las documentaciones precedentes resulta difícil deducir una distribución espacial clara. El desarrollo de las secuencias estudiadas no responde a una geografía precisa, y ello se confirma por el examen de los contextos en los que se insertan, y que nos indican que el desarrollo de estos usos es propio de la lengua notarial y casi todas ellas han tenido notable fortuna en el español culto, e incluso han pasado al español hablado hoy. No cabe duda de que se trata de usos que empezaron en la lengua administrativa y jurídica.

De los exámenes anteriores —y de otros que no presentamos aquí— cabe deducir un doble patrón de comportamiento del cambio morfológico y sintáctico, según la respuesta al parámetro diatópico:

- (1) Los cambios que se producen de abajo a arriba en la escala social son sensibles al factor geográfico, pues no parece que un cambio empiece al unísono y se desarrolle con la misma cadencia y reglas internas en todos los ámbitos regionales ²⁹.

²⁸ *Por tanto* es considerado en la lengua moderna un «conector consecutivo». Se ha de notar que en lo antiguo no está claro su carácter “parentético”, es decir, que implique pausa inmediatamente anterior, y tal vez quepa considerarlo integrado en la oración a la que sigue, lo mismo que el medieval *por ende*, que traspasa con mucho la Edad Media, y que conserva comportamientos sintácticos del *ende*, p. ej., en la fórmula «non fagades ende ál», que aun encontramos en un documento del Consejo Real de Hacienda de Carlos II, dado en Alcalá de Henares en diciembre de 1674 (CODEA 818), donde es secuencia fosilizada.

²⁹ No puede descartarse, sin embargo, que un cambio de este tipo pueda ser multifocal.

- (2) Los cambios que se producen de arriba a abajo, es decir, los que empiezan en las capas culturalmente altas, que se suelen difundir, al menos inicialmente, a través de la escritura, no responden al factor geográfico.

5.4. Léxico

La parroquia de San Gil de Molina de Aragón (Guadalajara) cuenta con importantes fondos prácticamente desconocidos por los investigadores³⁰. Ofrecemos un segmento de una carta de trueque escrita por Álvaro Ruiz el 26 de agosto de 1370, y hasta ahora inédita.

Geografía lingüística histórica del léxico: *pieza y pieza de tierra*

[...] ot<or>gamos & conosçemos q<ue> fazemos/ troq<ue> en vno en la man<er>a q<ue> se sygue q<ue> nos el dicho cabylldo & arçipste & abad q<ue> ayamos co<n> vos los dichos/ frnd m<artine>z & don<n>a (6) marja vnas casas q<ue> nos el dicho cabylldo auemos aq<ui> en moljna q<ue> son en la cal de tras sant john de conçejo ygles<i>a de moljna [...] / Por **vna pieça de t<ie>rra de pan levar** de nos los dichos f<e>rr<a>nd m<aertine>z & don<n>a marja q<ue> auedes enlas charcjellas q<ue> dizen (9) del rregado/ E a por lynderas **dela vna p<ar>te pieça de mj<guel> p<ere>z & de vraca f<e>rrnd<a>ez** su muger/ & **dela otra parte pieça dela dicha conpanja de sant gil** & de parte de ençima (10) co<n> el açeq<ui>a q<ue> viene alas huertas de tras santa marya la vieja/ E pasa por delante dela dicha ygles<i>a/ E nos los dichos f<e>rr<a>nd m<artine>z e don<n>a marja ot<or>gamos q<ue> faze(11)mos co<n> vos los dichos/ cabylldo/ & arçip<re>ste /& abad el dicho troq<ue> **dela dicha n<uest>ra pieça** por las dichas v<uest>ras casas q<ue> son de vos el dicho cabylldo/ [...]

Parroquia de San Gil, 9 (11); Molina de Aragón (Guadalajara)
26 de agosto de 1370. Alvar Ruiz

Hemos marcado el término *pieza* y el sintagma *pieza de tierra*, empleados para el ‘concepto parcela de terreno agrícola’. Con el objeto de situar geográficamente este uso acudimos al CODEA. En otro lugar estudiamos la distribución espacial de ciertos términos en los dialectos centrales de la Península Ibérica (Sánchez-Prieto Borja 2011a), y de ahí extraemos estos datos sobre *pieza* y *pieza de tierra*, donde puede apreciarse cómo se concentran en Navarra y las provincias aragonesas.

³⁰ Agradezco a mi alumna Marián Sanz que me llamara la atención sobre estos fondos, me ayudara en la labor de fotografiado y transcribiera algunas piezas.

Pie(ç)a y pie(ç)a de tierra en el CODEA

	s. XIII	s. XIV	s. XV	
Burgos	1			1
Navarra	9	3	1	13
La Rioja	1		2	3
Teruel			1	1
Zaragoza	2	10	10	22
	13	13	14	40

pieza

	s. XV	
Navarra	2	2
Teruel	4	4
Zaragoza	6	6
	12	12

pieza de tierra

En efecto, Molina de Aragón muestra en este uso una coincidencia significativa con el área oriental, y se revela clave para el estudio de la «frontera» oriental de Castilla. Un mapa sinóptico general permite situar mejor las formas léxicas examinadas (el punto localiza a Molina de Aragón dentro de la actual provincia de Guadalajara):



Términos para 'parcela de terreno agrícola' en la Edad Media (AHN, Clero)

6. CONCLUSIONES

El CODEA se presenta como un corpus primario, con fines no específicos, abierto a investigadores de diversas parcelas (historiadores de la lengua, dialectólogos, lingüistas de cualquier orientación, paleógrafos, historiadores y aun a usuarios generales). Gracias a la aplicación de unos criterios de edición específicamente establecidos para fuentes

documentales, con triple presentación, y al proceso de edición mismo, en el que se han llevado a cabo hasta seis revisiones, el *CODEA* es un corpus citable directamente por los investigadores, a los que, en todo caso, se les proporcionan los instrumentos de comprobación necesarios (facsimiles).

La utilidad de un corpus como este no se acaba en la consulta. De acuerdo con una concepción empírica de la historia de la lengua, especialmente necesaria en sintaxis, la idea de consulta nos parece insuficiente, pero es, claro, la única vía para corpus a cuyos textos íntegros no puede accederse. No es este el caso del *CODEA*, ya que todos los textos se editan íntegramente, y así se ofrecen a los lectores, sin claves de acceso, registro previo, pagos ni restricción alguna.

La naturaleza de los textos editados (documentos archivísticos) permite deducciones basadas en las coordenadas espacio-temporales, de la grafía a la sintaxis, y de los orígenes a 1700. Los cambios desde abajo en la escala social son sensibles al componente geográfico; los que se han producido desde arriba (p. ej., atestiguados por la cancillería o el consejo real), no son sensibles a este componente o no lo muestran en el estado actual del corpus.

Por otra parte, el *CODEA*, como corpus en continuo desarrollo, está abierto a la incorporación de nuevos documentos que permitan completar su estado actual, llenando las lagunas geográficas y cronológicas. Al mismo tiempo, se incluirán desarrollos tecnolingüísticos (navegación y búsqueda compleja con varios parámetros de filtro, con presentación directa en mapa de los resultados de las búsquedas y presentación estadísticas de los resultados). Ello contribuirá a una conceptualización más fácil e inmediata de la variación espacial, geográfica y social de la lengua y a un acceso cualitativo/cuantitativo que hará posible formular hipótesis sobre la historia de las variedades del español.

Por último, en el campo de la enseñanza, los corpus son un campo de ensayo estupendo en el aula, una vía para aprender un método científico, y no solo para encontrar respuestas, sino para ayudarnos a plantear nuevas preguntas. El *CODEA* puede contribuir así a afianzar la idea de que el nuestro —más que un conjunto de certezas transmitidas— es un saber adquirible y contrastable, y a contrarrestar cierta falta de anclaje empírico que suele achacarse, y no siempre sin razón, a nuestros estudios³¹.

REFERENCIAS BIBLIOGRÁFICAS

- ARIAS ÁLVAREZ, Beatriz (2009): «Confección de un corpus para conocer el origen, la evolución y la consolidación del español en la Nueva España», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 55-75.
- CICA = *Corpus Informatitzat del Català Antic*, Joan Torruella Casañas (dir.), junt amb Manel Pérez Saldanya i Josep Martines. <http://lexicon.uab.cat/cica> [Consulta: 6/11/2011].
- CODEA = *Corpus de Documentos Españoles Anteriores a 1700*. <http://textoshispanicos.es> [Consulta: 6/11/2011].
- CORDE = *Corpus diacrónico del español*. <http://www.rae.es>.
- DÍEZ DEL CORRAL ARETA, Elena (en prensa): «La problemática de las tradiciones textuales en el estudio lingüístico del documento indiano», en *Actas del I Congreso internacional: Tradición e*

³¹ Los ensayos en el aula, y en trabajos de curso, con alumnos del Grado en Estudio Hispánicos de la Universidad de Alcalá nos han dado buen resultado. El *CODEA* puede contribuir a la adquisición de un método empírico por parte de los estudiantes en las materias de Historia de la lengua y la Dialectología.

- innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos (Madrid, 11-13 de noviembre de 2009).*
- ENRIQUE-ARIAS, Andrés (ed.) (2009): *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt: Iberoamericana/Vervuert.
- ENRIQUE-ARIAS, Andrés (2010): «La traducción del código Escorial 1.1.6 en el contexto de los romanceamientos bíblicos medievales», en Andrés Enrique-Arias (ed.), *La Biblia Escorial 1.1.6. Transcripción y estudios*. Logroño: Centro Internacional de Investigación de la Lengua Española (Cilengua), pp. 67-86.
- FERREIRO FERNÁNDEZ, Manuel (2010): «Os hapax como problema e como solución. Sobre a cantiga 493/18,11 [B 459/v 78] de Alfonso x», en Mariña Arbor Aldea e Antonio Fernández Guiadanes (ed.), *Estudos de edición crítica e lírica galego-portuguesa*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 239-262 (*Verba*, Anexo 67).
- GARCÍA GODOY, María Teresa (en prensa): «El cambio *vuestra merced* > *usted* desde la documentación archivística», en *Actas del II Congreso internacional: Tradición e innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos (Neuchâtel, 7, 8 y 9 de septiembre de 2011)*.
- GOEBL, Hans (1975): «Qu'est-ce que la scriptologie? », *Medioevo Romanzo*, II, pp. 3-43.
- HARRIS-NORTHALL, Ray (1991): «Apocope in Alfonsine Texts: A Case Study», en Ray Harris-Northall and Thomas D. Cravens (eds.), *Linguistic Studies in Medieval Spanish*. Madison: Hispanic Seminary of Medieval Studies, pp. 29-38.
- ISASI MARTÍNEZ, Carmen (2010): «Edición digital: retos nuevos en los nuevos recursos», en Mariña Arbor Aldea e Antonio Fernández Guiadanes (eds.), *Estudos de edición crítica e lírica galego-portuguesa*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 353-368 (*Verba*, anexo 67).
- KAWASAKI, Yosifumi (en prensa): «La determinación cronológica de cambios según el CODEA», en *Actas del II Congreso internacional: Tradición e innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos (Neuchâtel, 7, 8 y 9 de septiembre de 2011)*.
- LÓPEZ VILLALBA, José Miguel (1998): «Normas españolas para la transcripción y edición de colecciones diplomáticas», *Espacio, Tiempo y Forma, Serie III, H.^a Medieval*, t. 11, pp. 285-306.
- MATUTE MARTÍNEZ, Cristina (2004): *El sistema referencial de los pronombres personales en la documentación castellana medieval. Intento de reconstrucción histórica*. Madrid: Universidad Autónoma de Madrid. Tesis Doctoral dirigida por Inés Fernández-Ordóñez.
- MORENO BERNAL, Jesús (1993): «Les conditions de l'apocope dans les anciens textes castillans», en Maria Selig, Barbara Frank et Jörg Hartmann (eds.), *Le passage à l'écrit des langues romanes*. Tübinga: Narr, pp. 193-206.
- MORENO BERNAL, Jesús (2003): «Efectos de la síncope en los futuros románicos», en José Luis GIRÓN ALCONCHEL, Silvia Iglesias Recuero, Francisco Javier Herrero Ruiz de Loizaga, Antonio Narbona Jiménez (eds.), *Estudios ofrecidos al Profesor José Jesús de Bustos Tovar*, vol. I. Madrid: Universidad Complutense de Madrid, pp. 247-257.
- ORDUNA, Germán (1990): «La "edición crítica"», *Incipit*, X, pp. 17-43.
- PEREDES GARCÍA, Florentino (en prensa): «Factores condicionantes de la variación *otro/otre/otri/otrie* en español medieval», en *Actas del II Congreso internacional: Tradición e innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos (Neuchâtel, 7, 8 y 9 de septiembre de 2011)*.
- PASCUAL RODRÍGUEZ, José Antonio y Carlos DOMÍNGUEZ CINTAS (2009): «Un corpus para un nuevo diccionario histórico del español», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 79-93.

- PATO, Enrique (2009): «Notas aclaratorias sobre la historia del indefinido alguien: una aplicación directa del uso de corpus diacrónicos», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 400-416.
- PATO, Enrique y Elena FELÍU ARQUIOLA (2005): «Alternancia de formas, nivelación e inferencia semántica: el caso de los participios en -udo del español medieval», *Revue de Linguistique Romane*, vol. 69, ts. 275-276, pp. 437-463.
- PUÑAL FERNÁNDEZ, Tomás (2004): «Tipología documental de la Edad Media», en Pedro L. Lorenzo Cadarso y Agustín Vivas Moreno (eds.), *Lecciones de Archivística y Documentación del Patrimonio Histórico*. Sevilla: Universidad de Extremadura/Editorial @becedario, pp. 153-157.
- ROMERO CAMBRÓN, Ángeles (2005): «Sintaxis histórica y crítica textual: camino de ida y vuelta», José Luis Girón Alconchel y José Jesús de Bustos Tovar (eds.), *Actas del VI Congreso Internacional de Historia de la Lengua Española*, vol. II. Madrid: Arco/Libros, pp. 2051-2062.
- SÁNCHEZ-PRIETO BORJA, Pedro (2000): «La preposición *a* con valor 'lugar en donde' en castellano antiguo», en Annick Englebert, Michel Pierrard, Laurence Rosier et Dan Van Raemdonck (eds.), *Actes du xxiiie Congrès International de Linguistique et de Philologie Romanes, (Bruxelles, 23-29 juillet 1998)*, vol. II. Tübinga: Max Niemeyer, pp. 393-406.
- SÁNCHEZ-PRIETO BORJA, Pedro (2004): «La normalización del castellano escrito en el siglo XIII. Los caracteres de la lengua: grafías y fonemas», en Rafael Cano Aguilar (dir.), *Historia de la lengua española*. Barcelona: Ariel, pp. 423-448.
- SÁNCHEZ-PRIETO BORJA, Pedro (2011a): «Ensayo de geografía lingüística histórica: términos para 'parcela de terreno agrícola' en las fuentes documentales de la Edad Media», en Sara Gómez Seibane y José Luis Ramírez Luengo (eds.), *Maestra en mucho. Estudios filológicos en Homenaje a Carmen Isasi Martínez*. Buenos Aires: Voces del Sur, pp. 271-302.
- SÁNCHEZ-PRIETO BORJA, Pedro (2011b): *La edición de textos españoles medievales y clásicos*. San Millán de La Cogolla: Centro Internacional de Investigación de la Lengua Española (Cilengua).
- SÁNCHEZ-PRIETO BORJA, Pedro (en prensa a): «Un corpus para el estudio integral de fuentes documentales», *Actas del VIII Congreso Internacional de Historia de la Lengua Española (Santiago de Compostela, 14-18 de septiembre de 2009)*.
- SÁNCHEZ-PRIETO BORJA, Pedro (en prensa b): «Para una historia de la escritura en León, Castilla y Aragón: algunas claves interpretativas», *Medioevo Romano*.
- SÁNCHEZ-PRIETO BORJA, Pedro, Micaela CARRERA, Carmen ISASI y Paul SPENCE (en prensa): «El corpus de CHARTA», en *Actes del 26è Congrès Internacional de Lingüística i Filologia Romàniques (València, 6-10 de setembre de 2010)*.
- SHISHIDA, Michiyasu (2010): *Estudio diacrónico y contrastivo de los pronombres de cortesía en español con referencias al portugués*. Alcalá de Henares: Universidad de Alcalá. Tesis Doctoral dirigida por Pedro Sánchez-Prieto Borja.
- SIMÓN PARRA, María (2009): *El nombre de persona en la documentación medieval*. Alcalá de Henares: Universidad de Alcalá. Tesis Doctoral dirigida por Pedro Sánchez-Prieto Borja.
- SPENCE, Paul, Carmen ISASI, Elena PIERAZZO e Irene VICENTE (en prensa): «Cruzando la brecha: la marcación digital con criterios filológicos», en *Actas del I Congreso internacional: Tradición e innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos (Madrid, 11-13 de noviembre de 2009)*.
- TMILG = *Tesouro medieval Informatizado da Lingua Galega*. <http://ilg.usc.es/tmilg/usar.html> [Consulta: 6/11/2011].
- UEDA, Hiroto (en prensa): «La apócope extrema medieval en la fonética castellana. Observaciones en el 'Corpus de Documentos Españoles Anteriores a 1700' (CODEA)», en *Actas del II Congreso internacional: Tradición e innovación: nuevas perspectivas para la edición y el estudio de documentos antiguos (Neuchâtel, 7, 8 y 9 de septiembre de 2011)*.