Scriptum Digital
1 (2012), pp. 85-106

DOS PROBLEMAS EN EL USO DE CORPUS DIACRÓNICOS DEL ESPAÑOL: PERSPECTIVA Y COMPARABILIDAD*

ANDRÉS ENRIQUE-ARIAS
Universitat de les Illes Balears

RESUMEN:
Este trabajo aborda una serie de problemas metodológicos relacionados con la investigación de la variación lingüística a partir de corpus informatizados. A través de una serie de estudios de caso con datos de Biblia Medieval (un corpus paralelo de traducciones de la Biblia en español antiguo) se muestra cómo este tipo de investigación puede beneficiarse del uso de textos paralelos. Para empezar, la metodología de los corpus paralelos ofrece una perspectiva más abierta, ya que es posible analizar todas las formas utilizadas para expresar contenidos del idioma de origen. Del mismo modo, los textos paralelos ofrecen comparabilidad directa de ejemplos concretos a través de diferentes períodos históricos, pues los equivalentes de traducción suelen insertarse en contextos de ocurrencia sintáctica, semántica y pragmática que son idénticos o muy similares. Por último, en un corpus paralelo es posible analizar la variación estilística de forma más controlada examinando cómo el mismo traductor selecciona diferentes opciones lingüísticas en función del género de cada texto.
PALABRAS CLAVE: lingüística de corpus, español medieval, corpus paralelos, traducciones de la biblia en español antiguo

ABSTRACT: This paper addresses a number of methodological problems related to corpus-based research in language variation. It is shown through a number of case studies using data from *Biblia medieval* (a parallel corpus of Old Spanish bible translations) how this kind of research can profit from parallel texts. To begin with, the perspective afforded by parallel corpus methodology is more open as it is possible to analyze all the forms used to express contents in the source language. Likewise, parallel texts offer direct comparability of concrete examples across different historical periods, as translation equivalents are likely to be inserted in the same or very similar syntactic, semantic and pragmatic contexts of occurrence. Finally, in a parallel corpus it is possible to analyze stylistic variation in a more controlled manner by examining how the same translator selects different linguistic options depending on the genre of each text.
KEY WORDS: Corpus linguistics, Medieval Spanish, Parallel corpora, Old Spanish Bible translations

1. INTRODUCCIÓN

Las últimas décadas del siglo xx han visto un resurgir de la lingüística histórica al que han contribuido dos importantes innovaciones metodológicas: la disponibilidad de herramientas informáticas que permiten analizar extensas colecciones de textos históricos y la aplicación de métodos de estudio de la variación lingüística (Joseph 2008: 182). La investigación en diacronía del español no ha quedado ajena a estos avances; en los estudios que se llevan a cabo hoy día es constante el empleo de datos extraídos de los corpus diacrónicos disponibles para esta lengua. No obstante, la relevancia que ha adquirido el uso de grandes bases de datos textuales en investigaciones de orientación diacrónica no se ha visto acompañada de un interés por analizar críticamente las posibilidades que nos ofrecen los corpus informatizados. En muchos casos el uso de estas herramientas se hace de manera irreflexiva, motivada principalmente por la facilidad del acceso a los datos, y no tanto como resultado

* Esta investigación se ha llevado a cabo en el marco del proyecto financiado del Ministerio de Economía y Competitividad de España FFI2010-18214, cofinanciado con fondos FEDER.

Fecha de recepción: 07/05/2012
Fecha de aceptación: 21/06/2012 ISSN: 2014-640X

de la selección consciente de las opciones metodológicas más apropiadas para la investigación que se pretende realizar[1]. Asimismo, existe la percepción un tanto simplista de que la adopción de medios electrónicos es en sí mismo una innovación metodológica que nos permite una mejor interpretación cualitativa y cuantitativa de los datos de la historia del español.

No obstante, si nos detenemos a examinar cómo se emplean estos nuevos recursos informáticos, resulta evidente que no siempre podemos hablar de innovación metodológica. En su empleo más básico los corpus permiten establecer concordancias y, a partir de las mismas, realizar sencillos análisis de coocurrencias; se trata, por tanto, de hacer de manera informatizada (es decir, más cómoda, rápida y a mayor escala) lo que antes se hacía manualmente leyendo cientos de páginas de texto y anotando en fichas todas las ocurrencias del fenómeno estudiado. La posibilidad de rastrear en un instante a lo largo de millones de palabras de textos históricos es una ventaja práctica del empleo de los corpus informatizados que debe valorarse positivamente, pero la lingüística computacional no se conforma con hacer de forma automática y a mayor escala lo que ya se hacía manualmente sino que aspira a desarrollar herramientas que permitan un análisis más sofisticado de la variación y de este modo constituir una nueva disciplina dentro de la lingüística con metodología y presupuestos teóricos propios.

El objetivo de este artículo es precisamente mostrar algunos problemas metodológicos del empleo de los corpus informatizados que han pasado a ser fuente habitual de datos en los estudios de la variación y el cambio lingüísticos en la diacronía del español y proponer soluciones creativas para aliviar estos problemas. En primer lugar presento una descripción general de los corpus diacrónicos del español de uso más frecuente para a continuación identificar una serie de limitaciones relacionadas con los parámetros de perspectiva y comparabilidad. Seguidamente discuto algunos ejemplos concretos de investigaciones de fenómenos de la historia del español a partir de datos extraídos de un corpus paralelo de traducciones bíblicas para mostrar cómo la comparación de textos paralelos puede ayudar a superar algunas de las limitaciones de los corpus convencionales. Por último, presento las conclusiones teóricas y metodológicas relevantes.

2. CORPUS CONVENCIONALES Y CORPUS PARALELOS

La investigación en lingüística diacrónica del español es a día de hoy prácticamente inconcebible sin recurrir a los corpus diacrónicos informatizados. En los últimos quince años ha ido apareciendo una serie de bases de datos de uso libre en la red entre las que destacan

[1] Son todavía relativamente pocos los estudios que se detienen a examinar cuestiones relacionadas con la solidez metodológica de los corpus diacrónicos del español (para una colección de estudios que analizan tanto aspectos filológicos de la selección y presentación de los textos como cuestiones de lematización, anotación gramatical y arquitectura de la base de datos, véase Enrique-Arias 2009a).

el *Corpus Diacrónico del Español* (*CORDE*), el *Corpus del Español* (*CE*), el *Corpus de Textos Españoles Anteriores a 1700* (*CODEA*) o el corpus *Biblia Medieval* (*BM*)[2]. El *CORDE*, creado a finales de la década de 1990, fue el primer gran corpus diacrónico del español de acceso libre en la red. En la actualidad se compone de aproximadamente 250 millones de palabras de textos que comprenden desde las primeras manifestaciones escritas del español hasta finales del siglo XX. Por su parte el *CE* se terminó un poco más tarde, en 2002, y contiene alrededor de 100 millones de palabras de los mismos periodos reflejados en el *CORDE* si bien, a diferencia de esta último, el *CE* está lematizado y tiene anotación gramatical (para una revisión de las ventajas metodológicas que aporta la anotación en el *CE*, véase Davies 2009). Ambos corpus se componen de textos de una variedad de géneros —poesía, prosa literaria, textos jurídicos, expositivos, técnico-científicos, etc.— con buena representación de la época medieval (23 y 18 millones de palabras en *CORDE* y *CE*, respectivamente). Junto a estos dos grandes corpus hay otro de creación más reciente y ámbito más especializado, el *CODEA*. Este corpus contiene más de 1500 documentos antiguos editados con el sistema de triple presentación de la red *CHARTA*: facsímil, transcripción paleográfica y presentación crítica. Se trata de documentos procedentes de distintos archivos españoles que abarcan desde la época de los orígenes de la lengua hasta 1700. Por último, en las páginas que siguen me referiré al corpus *Biblia medieval* (*BM*), que salió a la luz en 2009. *BM* es una herramienta de libre acceso en línea que permite consultar en paralelo versiones medievales españolas de la Biblia, compararlas con su fuente hebrea o latina y ver imágenes digitales de los originales. Este corpus, que solamente cubre el periodo medieval, cuenta con unos cinco millones de palabras.

Aunque, como ya se ha expuesto, *CORDE*, *CE* y *CODEA* tienen características particulares que los diferencian entre sí, son ejemplos claros de lo que puede considerarse un corpus *convencional*, que es a su vez el tipo de corpus de empleo más usual en las investigaciones diacrónicas del español. Tales corpus constan de una base de datos informatizada que contiene textos históricos de diferentes épocas y una herramienta de búsqueda para recuperar información de los textos. Con el fin de acceder a los datos, los usuarios necesitan introducir una palabra o frase en un cuestionario de consulta y la aplicación de búsqueda crea una concordancia que muestra todos los ejemplos del texto buscado en el corpus junto a su contexto de aparición, con información básica sobre el texto de origen como título, autor y fecha de composición.

[2] Me referiré principalmente a estos corpus por ser representativos del tipo de recursos que tienen hoy día una difusión más amplia entre los investigadores de la diacronía del español. Existen otros corpus diacrónicos importantes para el español, como las colecciones de textos del *Hispanic Seminary of Medieval Studies* (en adelante HSMS) o el *Archivo Digital de de Manuscritos y Textos Españoles* (ADMYTE); no obstante estas iniciativas no han tenido una repercusión comparable a la de los grandes corpus de acceso libre en la red (para una visión general del desarrollo de los corpus diacrónicos del español, véase Enrique-Arias 2009b: 13-14).

A diferencia de las bases de datos textuales que acabo de describir, *BM* es un corpus *paralelo*, es decir, una colección de textos originales y sus equivalentes de traducción[3]. En los corpus paralelos los textos están alineados de tal forma que es posible identificar palabras o frases en el texto original y emparejarlos con la expresión correspondiente en las demás versiones paralelas. En el caso de *BM* el corpus está compuesto por la Biblia hebrea y la Vulgata Latina, que son los textos originales, y las versiones en español medieval[4]. Así, cuando el usuario introduce una consulta para cualquiera de las versiones paralelas en el corpus, ya sea en el texto original, o en cualquiera de las trece versiones en español medieval que contiene, la aplicación de búsqueda muestra todas las ocurrencias de la consulta en la versión correspondiente al lado de los equivalentes de traducción en todas las demás versiones.

3. LIMITACIONES DE LOS CORPUS CONVENCIONALES

El empleo de los corpus convencionales para estudiar la variación y el cambio lingüísticos en perspectiva diacrónica se encuentra con al menos dos limitaciones significativas que examinaremos en las páginas que siguen. Una de ellas afecta a la noción de *perspectiva*, entendida como la forma en que se accede a la información en estas colecciones de textos antiguos. En los corpus convencionales los usuarios introducen en la casilla de consulta las formas que son relevantes para rastrear el fenómeno que desean investigar. Esto significa que es necesario conocer tales formas de antemano por medio de gramáticas, diccionarios históricos o estudios monográficos. La principal desventaja de este proceder es que, por muy bien que hagamos nuestro trabajo previo de investigación de materiales de referencia, siempre existe el riesgo de pasar por alto alguna forma relevante por no haber sido estudiada con anterioridad: es decir, el valor heurístico del corpus queda severamente limitado pues la vía de acceso a los datos solamente nos permite rastrear lo que ya nos es conocido. Un inconveniente añadido es que, como la metodología de la lingüística a base de corpus gravita necesariamente hacia la búsqueda y observación de marcadores explícitos, en nuestras investigaciones solamente podemos considerar un número limitado de opciones; no podemos buscar de manera automática todas las maneras posibles de expresar una función pues los corpus convencionales no permiten incluir en la casilla de búsqueda formas

[3] El empleo de corpus paralelos de equivalentes de traducción como método para estudiar de manera más controlada la variación inter e intra lingüística es un planteamiento metodológico con amplia tradición que en los tiempos recientes ha llegado a constituir por sí mismo un paradigma completo y coherente dentro de la lingüística de corpus (véase, por ejemplo, McEnery y Xiao 2007).

[4] El recurso a traducciones de la Biblia a la hora de crear un corpus paralelo de la diacronía del español se justifica por ser el único texto para el que tenemos versiones compuestas en todas las etapas de la historia del español. Varios estudios avalan la solidez metodológica de la comparación de versiones bíblicas paralelas considerando parámetros como representatividad, comparabilidad, calidad, diversidad o perspectiva (Resnick *et al.* 1999, Kaiser 2005, De Vries 2007, Enrique-Arias 2008, 2009c, 2012).

como el cero morfológico, o el cambio en el orden de palabras. El procedimiento para acceder a los datos en los corpus convencionales puede ser apropiado cuando se quiere buscar elementos de clase cerrada, o cuando se conoce la lista exhaustiva de las formas posibles relacionadas con el fenómeno que se desea estudiar, pero es inadecuado cuando estamos investigando estructuras que se puede expresar con elementos de clase abierta o para los que no podemos saber de antemano todas las formas posibles de expresión.

El otro problema al que me referiré en las páginas que siguen es el de la c*omparabilidad*. Es un hecho aceptado en prácticamente todos los modelos teóricos de la lingüística histórica que los cambios lingüísticos se dan en tres etapas: un estadio original anterior al cambio, una fase en la que triunfa la nueva estructura y una etapa intermedia en la que coexisten el sistema innovador y el original. Es precisamente este momento intermedio, caracterizado por la variación, el más interesante para el investigador, ya que el estudio de los contextos que favorecen la aparición de la forma innovadora permite obtener información sobre los factores que motivan el cambio, los contextos en los que se ha originado y los canales por los que se ha extendido. En consecuencia, los métodos cuantitativos —sobre todo los análisis de coocurrencias de las variantes lingüísticas que compiten en los mismos contextos de aparición— se han convertido en una herramienta esencial en el empleo de los corpus para investigar cambios lingüísticos. Por ello, a la hora de desarrollar análisis de este tipo los investigadores deben hacer todo lo posible por asegurarse de que la base empírica sobre la que construyen sus teorías es tal que garantiza el más alto grado posible de comparabilidad entre las muestras estudiadas de cada corte sincrónico seleccionado, es decir, tenemos que asegurarnos de que los datos que obtenemos de textos pertenecientes a diferentes periodos, registros y dialectos, y que se insertan en diferentes contextos de ocurrencia, están en una relación de equivalencia entre sí, y es por lo tanto pertinente establecer una comparación entre los mismos.

El problema es que, cuando se trabaja con un corpus convencional, no siempre es fácil identificar y controlar el amplio número de factores que condicionan la variación en los textos escritos. Aparte de los factores estructurales (sintácticos, semánticos, discursivos) cabe tener en cuenta los numerosos factores contextuales que influyen en las opciones lingüísticas de los que escriben (caracterización social del autor y los destinatarios, el género y el registro, rasgos de oralidad, procedencia dialectal, tradiciones de escritura, o el hecho de que el texto sea traducción de otra lengua). Como los textos medievales nos han llegado muchas veces en testimonios que no son originales y sin información sobre fecha precisa de composición, autoría, procedencia geográfica y perfil de los destinatarios, el lingüista se expone a cometer errores de interpretación y análisis de los datos procedentes de los textos.

En lo que sigue examinamos algunos problemas metodológicos específicos que suscita el empleo de datos de corpus para mostrar algunas de las limitaciones del empleo de los corpus convencionales en lo que respecta a los parámetros de perspectiva y de comparabilidad.

4. PERSPECTIVA

El diseño y la arquitectura de un corpus condicionan la manera en que los estudiosos acceden a los datos lingüísticos. Como ya se ha señalado, cuando se utiliza un corpus convencional, los usuarios introducen en la casilla de búsqueda las formas que supuestamente son relevantes para el fenómeno que desean estudiar y obtienen un listado de ocurrencias de tales formas en su contexto inmediato. A partir de la observación de estos ejemplos el lingüista puede deducir el significado y función de las formas estudiadas. En los corpus convencionales el análisis procede en la dirección *forma* → *función*, con los inconvenientes que ya se han mencionado: el conocimiento deficiente o insuficiente de las formas relevantes se traducirá en un análisis incompleto del fenómeno estudiado. Por el contrario, en la metodología de los textos paralelos partimos de ejemplos específicos integrados en su contexto y observamos las formas utilizadas en los equivalentes de traducción en las versiones paralelas, lo cual, como explico a continuación, ofrece una serie de ventajas.

4.1. Función heurística de las versiones paralelas

La primera ventaja es la función heurística de los textos paralelos, que no tiene equivalente en otras fuentes de datos. Por ejemplo, supongamos que queremos estudiar la evolución histórica de los elementos exceptivos (es decir, los recursos lingüísticos utilizados para expresar excepción). Si queremos usar un corpus convencional, primero tenemos que consultar materiales de referencia y elaborar una lista de elementos que puedan expresar esta función (por ejemplo, *excepto*, *salvo*, *menos*, *fueras*, etc.). A continuación, realizamos búsquedas de estas formas y utilizamos, finalmente, los resultados para examinar ejemplos específicos en su contexto funcional. Al proceder de esta forma estamos adoptando una perspectiva del tipo *forma* → *función*; la consecuencia inmediata es que no hay manera de saber si el corpus contiene otros elementos que pueden ser utilizados con la misma función y en los mismos contextos, pues solamente exploramos las formas que conocemos de antemano. Por el contrario, en un corpus paralelo como *BM* no es necesario partir de una lista exhaustiva de las formas relevantes pues la exploración del corpus y las comparaciones con las versiones paralelas nos guiarán en la búsqueda de las unidades de expresión posibles para la estructura que se está investigando. En *BM* tenemos varias vías para extraer los pasajes que contienen los elementos que son relevantes para nuestra investigación. Podemos, por ejemplo buscar en el original latino formas exceptivas conocidas como *absque*, *praeter*, *nisi*, *non* ... *sed*, o hacer lo propio con cualquiera de las palabras relevantes en la versión hebrea, o buscar las formas que conozcamos en cualquiera de los textos en español, y luego observar las formas que se utilizan en el mismo contexto y con las mismas

funciones en las versiones paralelas. A su vez, podemos buscar las formas que encontramos en estos escarceos, lo cual dará lugar a más formas que pueden ser utilizadas para nuevas búsquedas. Esta perspectiva, en que partimos de determinados contenidos incorporados en el texto y observamos las formas empleadas en las versiones paralelas (es decir, *función → forma*), facilita la observación de elementos que de otro modo habrían sido ignorados.

Es evidente, sin embargo, que un corpus paralelo como *BM* nunca debe ser la única fuente de información en un estudio diacrónico. Otras fuentes, tales como diccionarios, gramáticas, estudios, y sobre todo grandes corpus convencionales, como *CORDE* o *CE*, son fuentes indispensables para asegurarse de que las formas que descubrimos gracias al corpus paralelo no son solo palabras empleadas en traducción bíblica, sino que tienen empleo en otros géneros. Por ejemplo, en su estudio de partículas exceptivas derivadas de participios latinos en la historia del español, Sánchez López (en prensa) consultó en el corpus *BM* formas latinas que expresan excepción y al examinar los equivalentes castellanos pudo encontrar, junto a numerosos casos de los consabidos *excepto*, *salvo*, *excluso*, *menos*, etc. 53 ocurrencias de *salvante*, una forma que no se habían registrado en los materiales de referencia anteriores. Una búsqueda en *CORDE* demuestra que *salvante* no es una forma limitada al lenguaje bíblico, pues aparece 33 veces en 24 textos de diferentes géneros, fechados entre 1380 y 1758.

4.2. Perspectiva abierta de los corpus paralelos

Otra ventaja de la perspectiva *función → forma* empleada en la metodología de los corpus paralelos de equivalentes de traducción es que, por ser mucho más abierta que la de los corpus convencionales, permite analizar cualquier forma de expresar un contenido de la lengua fuente. Como ya se ha comentado más arriba, en la metodología de los corpus paralelos las búsquedas no están limitadas a marcadores explícitos ni a un número limitado de formas. En un corpus paralelo se trabaja a partir de equivalentes de traducción, lo cual permite considerar sin restricciones todas las formas de expresar un contenido de la lengua fuente.

Por ejemplo, si nos proponemos estudiar las formas de expresar preguntas retóricas en los textos medievales nos encontramos con el problema de que no hay una forma fácil de rastrearlas automáticamente ya que se expresan de múltiples maneras y no siempre con un marcador explícito (Enrique-Arias y Burguera 2010). El acceso a las fuentes es de gran ayuda en la localización de ocurrencias del fenómeno estudiado: un rastreo de la partícula interrogativa *hă-* en la versión hebrea o *numquid* en la latina nos permite localizar automáticamente un gran número de casos de preguntas retóricas y examinar cómo están

expresadas en las versiones castellanas, como se puede apreciar en las traducciones de *Job* 8:3 en el corpus[5]:

(1) *Job* 8:3

[Hebreo]	ha'el ye 'avet mišpat
[Vulgata]	numquid Deus subplantat iudicium
[E8]	¿E Dios, *tienes que* engaña el juizio?
[GE]	¿*Si non* derriba Dios el tu juizio?
[E3]	¿*Si* Dios atuerce el juizio?
[E5]	¿*Quiçá* Dios estuerce el derecho?
[Arragel]	*Nunca* el Señor Dios atorció juizio

Si quisiéramos estudiar cómo se formaliza la pregunta retórica en el español medieval a partir de un corpus convencional tendríamos que consultar primero diccionarios y gramáticas históricas y compilar una lista de elementos que pueden expresar esta función (¿*acaso*...? ¿*por ventura*...? etc.), a continuación hacer búsquedas con estas palabras, y finalmente utilizar los resultados para observar ejemplos específicos en su contexto funcional. El problema es que en un corpus convencional no hay manera de saber si existen otros elementos que pueden usarse con la misma función en estos mismos contextos. Por el contrario, en un corpus paralelo de equivalentes de traducción como *BM* podemos extraer los pasajes que contienen estos elementos (por ejemplo buscando *numquid* en el original latino) y observar qué palabras se usan en el mismo contexto en cada una de las versiones paralelas. De este modo es fácil localizar expresiones que de otro modo podrían haber pasado desapercibidas. En el ejemplo de *Job* 8:3 apreciamos la rica variedad de expresiones utilizadas para formalizar la interrogativa retórica: un verbo de entendimiento en E8 (*tienes que* significa aquí 'crees que' o 'piensas que'); *si non* en GE, *si* en E3; el marcador epistémico de duda *quiçá* en E5. El caso de la traducción de Arragel es peculiar porque ha optado por emplear una aserción con cambio de polaridad (la pregunta retórica es una aserción encubierta: '¿acaso tuerce Dios el derecho?' se interpreta como 'Dios nunca tuerce el derecho'). Vemos, por tanto, que la perspectiva de los corpus paralelos es tan abierta que nos permite observar cualquier tipo de equivalente para la interrogativa retórica del original, incluso cuando ese contenido no se expresa mediante una oración interrogativa sino con una aserción.

4.3 Análisis de formas en competición

[5] Todos los ejemplos de traducciones bíblicas medievales en castellano proceden del corpus *BM* y han sido normalizados siguiendo en términos generales las normas de presentación crítica de la red *CHARTA* (*Corpus Hispánico y Americano en la Red: Textos Antiguos*), disponibles en www.charta.es. Para información completa sobre los manuscritos que han transmitido traducciones de la Biblia y las abreviaturas que empleo para referirme a ellos, consúltese la página del proyecto *Biblia medieval* (www.bibliamedieval.es).

Otra ventaja de la perspectiva metodológica de los corpus paralelos es que nos permite, a la hora de estudiar el significado y funciones de una estructura lingüística, localizar las diferentes formas que entran en competición a la hora de expresar una misma función y obtener así una información crucial para estudiar fenómenos de variación y cambio lingüísticos. Consideremos, por ejemplo, el estudio del significado y las funciones del marcador discursivo *he* (y sus variantes *afé*, *ahé*) en castellano medieval, un elemento que se emplea para llamar la atención sobre algo (como por ejemplo en *helo aquí*). Si queremos estudiar en un corpus convencional cómo se expresa esta función en español antiguo nos tenemos que limitar a realizar búsquedas de aquellos elementos explícitos que, suponemos, se podrían utilizar en estos contextos (por ejemplo *ahé* y sus variantes, *evás,* mandatos del tipo *ved*, *mirad*, etc.). Por el contrario, en el empleo de un corpus como *BM* la metodología es mucho más abierta pues podemos localizar cualquier elemento utilizado para expresar esta función: simplemente buscamos en los textos de origen todas las apariciones de los marcadores *hinné* del hebreo o *ecce* del latín y observamos cómo se traducen en las versiones en español. Como se puede apreciar en las traducciones de *Deuteronomio* 31:16 a continuación, el corpus permite captar la amplia gama de expresiones que los traductores medievales utilizan para transmitir el significado y la función de este marcador:

(2) *Deuteronomio* 31:16

[Hebreo]	vayó' mer yhvh 'el mošé *hine*já šojeb 'im 'aboteja
[Vulgata]	dixitque Dominus ad Moysen: *ecce* tu dormies cum patribus tuis
[Fazienda]	E dixo a Moisén: *e* tú izrás con tos parientes
[E8]	Et dixo Dios a Moisén: *evás* que tú dormirás con tus padres
[E3]	E dixo Dios a Muisén: *cata* que tú yacerás con tus parientes
[E4]	E dixo el Señor a Moisén: *hete* que dormirás con tus padres
[E7]	E dixo el Señor a Muisén: *ya tú* vas a yazer con tus parientes
[E19]	E dixo Dios a Muisén: *aquí* tú yacerás con tus parientes
[Arragel]	Dixo el Señor a Moisén: *sepas* que así como tú yoguieres con tus parientes

La comparación de las diferentes versiones nos permite observar una rica variedad de equivalentes; además de los marcadores *he* en E4 y *evás* en E8, los traductores emplean verbos de percepción (*cata que* en E3) o conocimiento (*sepas que* en *Arragel*), deícticos de tiempo (*ya* en E7) y lugar (*aquí* en E19), o también se deja la expresión sin traducir como en la *Fazienda*.

En un estudio reciente (cf. Enrique-Arias y Camargo en prensa) hemos aprovechado la facilidad de acceso a los equivalentes de traducción de *hinné* en *BM* para analizar 630 expresiones empleadas para expresar ese significado y función. Después de examinar el amplio número de ejemplos al que hemos tenido acceso gracias al corpus paralelo hemos clasificado cinco significados o funciones para *he*, ampliando así las descripciones tradicionales que se limitaban a caracterizarlo como un "adverbio demostrativo" que se usa "con un nombre del objeto señalado" (Menéndez Pidal 1911: 685-686) o como "un adverbio

que [...] sirve para mostrar una persona o cosa" (Corominas y Pascual 1984-1991, s.v. *he*). Los ejemplos que ilustran cada función son de la Biblia E3:

(3) Funciones de *ahé*
a. Locativo. Señala entidades concretas (personas o cosas) próximas en el espacio
E dixeronle: «¿Dó es Çara tu muger?» E dixo: «Hela en la tienda» [*Génesis* 18:9]
b. Eventivo. Señala una acción o evento próximo en el tiempo.
E fue a otro dia, entró Muisén a tienda de plazo, e hé que floresció la vara de Aharón [*Números* 17:23]
c. Citativo. Señala una alocución que contiene información relevante para el oyente.
E aparesçiose el ángel del Señor a la muger e díxole: «Ahé agora eres tú mañera e non pares, e empreñar te has e parirás un fijo» [*Jueces* 13:3]
d. Refuerzo argumentativo. Contribuye a establecer una relación entre dos proposiciones que aparecen contiguas en el discurso.
E si non lo quisieres enviar, ahé yo plagaré todo tu término con ranas [*Éxodo* 7:27]
e. Intensificador. Énfasis (sin significado deíctico)
hete fermosa, mi conpañera, hete fermosa [*Cantar* 1:15]

Una característica útil del corpus *BM* para el estudio de los usos y funciones de *he* es que podemos analizar en las versiones paralelas los casos en que los traductores estimaron que, a la hora de traducir *hinné* o *ecce* del original, era oportuno emplear otra expresión equivalente castellana o incluso dejarlo sin traducir, es decir, para cada función hemos podido observar qué formas entran en competición en castellano medieval. Un hecho destacable es que la distribución de las formas competidoras de *he* tiene una relación evidente con el tipo de función que expresan en cada caso: si excluimos los casos de *evás* y los que dejan la expresión sin traducir, tenemos que la alternativa mayoritaria a *he* cuando funciona como intensificador (96%) es precisamente una exclamación (*¡cuán!*, *¡qué!*, *¡oh!*). Para el citativo, la alternativa más corriente (33%) es el uso de estilo indirecto para introducir la cita; en el refuerzo argumentativo (71%), una conjunción (*ca*, *porque*, *pues que*, *por ende*, *mas*, *pero*, *que*) y en el eventivo (61%), verbos de percepción (*ver*, *fallar*, *mirar*, *entender*, *catar*).

La variedad de equivalentes de traducción de *he* demuestra, asimismo, que los traductores no actuaban de forma automática calcando el *hinné* o el *ecce* del original, sino que hacían una traducción interpretativa en la que trataban de seleccionar entre los recursos de la lengua del momento las expresiones más apropiadas para acercarse al sentido del texto.

4.4. Vías de acceso

Otro aspecto que podemos subsumir dentro del parámetro de *perspectiva* (aunque también tiene relación con otros, como *acceso* y *calidad*) tiene que ver con los recursos de los que

dispone el usuario para llegar a una mejor comprensión de los textos de un corpus (por ejemplo, para examinar los elementos escriptológicos o las opciones gráficas del manuscrito original, resolver lecturas dudosas o detectar errores). En lo que respecta a esta faceta, *CORDE* y *CE* nos dan una información mínima, pues contienen textos en una única versión, no siguen un criterio unificado para su presentación y no permiten acceso al facsímil del original. Quiere ello decir que, ante una lectura dudosa el usuario no cuenta con elementos para hacer las comprobaciones pertinentes. Por su parte el *CODEA* utiliza el sistema de triple presentación de la red *CHARTA* con posibilidad de consulta de una transcripción paleográfica que reproduce las opciones gráficas del original o una presentación crítica con grafías normalizadas que eliminan la variación gráfica superflua[6]. Además es posible acceder a imágenes digitales de los facsímiles. Este sistema de triple presentación permite hacer búsquedas enfocadas hacia fenómenos que operan en un nivel concreto de análisis, pues la transcripción paleográfica nos facilita el estudio gráfico-fónico, y la presentación crítica nos ayuda a rastrear aspectos de la morfosintaxis o del léxico. En lo que respecta a *BM*, en el momento actual cuenta con textos en transcripción paleográfica siguiendo unos criterios unificados para todo el corpus y acceso a los facsímiles de los originales mientras que en el desarrollo futuro del corpus está previsto ofrecer una versión normalizada con lematización y anotación gramatical.

Una ventaja evidente de *BM* frente a los corpus convencionales es que, al tratarse de un corpus paralelo de equivalentes de traducción con acceso al facsímil, ofrece una información más completa para facilitar la interpretación de las estructuras que contiene. Ante una lectura cuestionable o de dudosa interpretación el investigador puede, en primer lugar, consultar el facsímil para asegurarse de que la transcripción es correcta; si una vez descartado un error de edición la lectura todavía suscita dudas se puede aclarar su significado consultando el texto subyacente y las versiones paralelas. Por ejemplo, en *BM* resulta inmediatamente evidente que la lectura *estruirá* en la versión de *Isaías* 10:19 de E4 (*una criatura los estruirá*) es un error de copia por *escrevirá*, pues en el original hebreo tenemos *yi<u>k</u>tə<u>b</u>ēm* y en el latino *scribet* ‘escribirá’. Además en las versiones paralelas de RAH, Arragel, E3, BNM y *General Estoria* encontramos las formas *escrevirá* o *escrivirá*. Esta característica del corpus es una ventaja evidente sobre corpus convencionales en lo que no hay referentes semejantes con los que comparar.

5. COMPARABILIDAD

Como ya se ha señalado, uno de los problemas metodológicos centrales del estudio de la variación y el cambio lingüísticos mediante corpus diacrónicos es el de la comparabilidad. Como la variación lingüística está sometida a un amplio número de factores estructurales

[6] Véase al respecto la nota 5.

(semánticos, morfosintácticos) y contextuales (género, dialecto, destinatarios, etc.) siempre será difícil extraer ejemplos de un mismo fenómeno de textos de diferentes épocas con seguridad absoluta de que son realmente comparables.

Los problemas de comparabilidad se pueden atenuar considerablemente mediante la metodología de los corpus paralelos, es decir, comparando textos que son traducción de un mismo original y, por tanto, tienen el mismo contenido y han sido influidos por convenciones textuales semejantes. Dicho de otro modo, el objetivo del historiador de la lengua es observar cómo se expresa el mismo enunciado en la lengua de los periodos históricos A, B y C, y la forma más sistemática y directa de obtener la respuesta es comparando equivalentes de traducción de diferentes épocas (Goyens y van Hoecke 1992). Presentamos a continuación varios ejemplos de investigaciones diacrónicas que ilustran las posibilidades de un corpus paralelo ante los típicos problemas de comparabilidad a los que se enfrenta el estudioso de la historia del español.

5.1. Control de factores estructurales y contextuales

Una paradoja de la composición de los corpus diacrónicos es que, por una lado, deben ser heterogéneos (tienen que incluir textos de diferentes autores, épocas, géneros, registros, dialectos) y a la vez deben ser homogéneos (es decir, los diferentes cortes sincrónicos representados en el corpus tienen que ser comparables entre sí). A la hora de compilar un corpus cabe preguntarse si, por poner un ejemplo, es metodológicamente acertado comparar datos lingüísticos del siglo XIII extraídos de crónicas medievales con los de novelas del siglo XVI. A pesar de que en ambos casos se trata de discurso narrativo, nos encontramos ante obras con convenciones textuales muy diferentes, y en las que la distribución de narración, descripción y diálogo puede diferir significativamente. Es decir, a la hora de diseñar un corpus debemos asegurarnos de que, para cada periodo representado, estamos caracterizando estados de la lengua y no meras tipologías textuales.

Por estas razones, el estudio de la variación estilística en los textos medievales es una tarea que apareja ciertos riesgos. Para poder hacer un análisis con las debidas garantías sería necesario comparar obras de diferentes géneros compuestas por autores del mismo lugar de procedencia en fechas no muy distantes; en caso contrario corremos el riesgo de que las diferencias atribuibles al registro en realidad lo sean a otros factores, como procedencia dialectal diversa, características del autor (edad, condición social), o diferencias en la fecha de composición de los textos.

Una ventaja del corpus *BM* es que nos permite consultar colecciones de libros de diferentes géneros que son fruto de una misma labor romanceadora. Hay que tener en cuenta que la Biblia no es un libro sino una antología que engloba textos de variada tipología textual (narrativos, legislativos, líricos, sapienciales, proféticos, epistolares). El análisis de la lengua de las biblias medievales permite estudiar la variación estilística de manera más

controlada pues es posible examinar cómo un mismo traductor o compilador selecciona diferentes opciones lingüísticas apropiadas para cada uno de los géneros que representan los diferentes textos. Esta característica es especialmente útil cuando tenemos que estudiar fenómenos cuya variación se halla sujeta simultáneamente a factores internos (estructurales) y externos (contextuales).

Consideremos, por ejemplo, la variación en la distribución de *artículo + posesivo* (*la mi casa*) frente a posesivo sin determinar (*mi casa*), un fenómeno que, como demuestran numerosos estudios, obedece a una compleja combinación de factores estructurales. En (2) mostramos algunos de los que han sido identificados (datos adaptados de Wanner (2005) y de mis propias observaciones):

(4) Contextos que favorecen *artículo + posesivo* en español antiguo[7]
a. Rasgos del poseedor:
- 1ª y 2ª persona > 3ª persona
- singular > plural
b. Rasgos de la entidad poseída
- inanimado > (animado > términos de parentesco)
- partes del cuerpo > otros sustantivos
c. Función sintáctica del sintagma que contiene la estructura posesiva
- sujeto > objeto
- solo > con preposición (contra Wanner 2005)

El estudio del posesivo precedido de artículo se complica todavía más por ser un fenómeno fuertemente condicionado por factores de tipo estilístico. Al ser una estructura que enfatiza la posesión, es utilizada por los autores con valores estilísticos como expresividad, solemnidad, énfasis o reverencia. Ante tal heterogeneidad de factores y la forma tan compleja en que interactúan, el lingüista se enfrenta al problema de identificar claramente qué mecanismos están condicionando la variación y separar los asociados con la variable registro de los demás. En un corpus paralelo de equivalentes de traducción es relativamente más sencillo controlar los diferentes factores y estudiarlos de manera más atenta pues podemos comparar la distribución de *artículo + posesivo* en diferentes géneros textuales, como ejemplifican los datos que presentamos a continuación en los que consideramos pasajes líricos, sapienciales y narrativos en tres traducciones bíblicas medievales (ver Figura 1)[8].

[7] El vector (>) indica que la categoría que aparece a la izquierda se expresa con posesivo precedido de artículo con más frecuencia que la de la derecha.
[8] Los textos analizados para cada uno de los géneros son: lírico (*Cantar de los cantares*), sapiencial (*Proverbios* 1-10) y narrativo (*Macabeos I* 1-5 para E6 y *General Estoria* y *Génesis* 1-8 para *Arragel*).

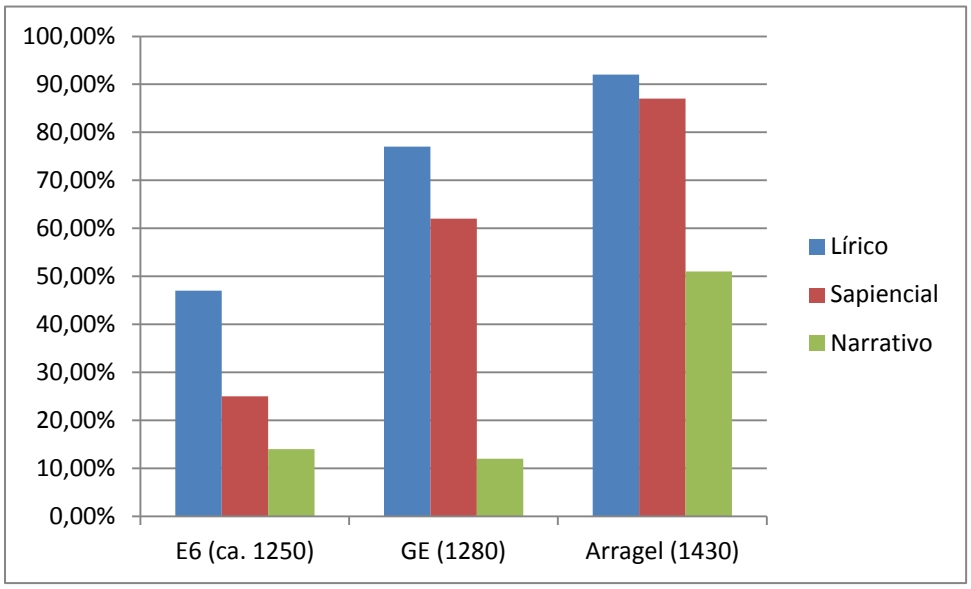

Figura 1. Porcentaje de *artículo + posesivo* (frente a posesivo solo) según género textual

La distribución observada nos indica claramente que el empleo de esta estructura tenía un decidido valor estilístico en la lengua medieval. El porcentaje de *artículo + posesivo* en el corpus seleccionado varía de forma considerable, con un amplio rango que va desde el 12% en la muestra narrativa de la GE hasta el casi 92% de la muestra lírica en la Biblia de Arragel. Junto a las grandes diferencias que encontramos entre los diferentes textos se aprecia claramente un patrón regular que se repite en cada uno de los romanceamientos examinados: el texto narrativo siempre es el que tiene menos proporción de *artículo + posesivo* y el lírico el que más. Los datos vienen a confirmar la observación de Lapesa (2000 [1971]: 422) de que *artículo + posesivo* suele escasear en pasajes «impersonalmente narrativos, pero aumenta en frecuencia y hasta predomina a veces en los fragmentos de carácter poético, retórico o donde hay proximidad afectiva del narrador».

La variación estilística que observamos en la Figura 1 queda demostrada de manera más controlada que si hubiéramos utilizado un corpus convencional pues no se están comparando textos de variada procedencia sino, para cada época, partes de una Biblia, es decir, una antología de textos que, al ser fruto de una misma labor romanceadora, están cercanos en el tiempo y el espacio.

5.2. Aumento de frecuencia textual

Una observación que aparece de forma recurrente en los estudios sobre el cambio lingüístico es que los patrones que se utilizan con frecuencia en el discurso terminan por convencionalizarse y convertirse en estructuras que expresan un significado gramatical (Bybee 2007). Por ello, la frecuencia textual es un parámetro de gran importancia en los estudios lingüísticos de orientación diacrónica. Pero de nuevo, al hacer estudios de frecuencia a partir de textos de diferentes épocas, nos encontramos con los desafíos

metodológicos relacionados con la noción de comparabilidad. Para poder postular que se ha dado un cambio en la frecuencia de una estructura necesitamos asegurarnos de que las muestras que escogemos para cada corte sincrónico son en efecto comparables.

Consideremos, por ejemplo, la evolución de los clíticos pronominales del español, los cuales, según el análisis que ya se acepta ampliamente, han experimentado un proceso de morfologización y han pasado a ser afijos verbales de concordancia objetiva en el español actual (Enrique-Arias 2005). Es bien sabido que los cambios de este tipo normalmente acarrean un aumento de frecuencia textual, pero tal desarrollo es difícil de documentar en los corpus disponibles. El problema es que, en textos de contenido diferente, el número y función de los pronombres átonos variará dependiendo de circunstancias idiosincráticas de cada texto: por ejemplo, la proporción de narración, descripción y diálogo, o el punto de vista (primera persona frente a tercera), o el tipo de contenidos, son elementos que determinan la mayor o menor presencia de ciertos tipos de verbos (transitivos, reflexivos, impersonales) y a su vez condicionan el número y distribución de los diferentes tipos de clíticos. Para remediar este problema es necesario hacer búsquedas en un corpus lo suficientemente grande como para que la abundancia de textos ayude a igualar las características de cada corte sincrónico. Como los clíticos medievales presentan considerable variedad formal (proclíticos, enclíticos, apocopados, interpolados) no es posible localizarlos haciendo búsquedas automáticas en los grandes corpus disponibles en línea. Además tampoco existe la posibilidad de descargarse los textos para que el usuario pueda implementar búsquedas complejas en su propio ordenador.

Una de las ventajas del corpus paralelo *BM* es que, al tratarse de textos con el mismo contenido, es posible comparar de manera inmediata ocurrencias de ejemplos lingüísticos concretos con un alto grado de equivalencia semántica, sintáctica y pragmática y así analizar la variación de manera más controlada. De este modo, es posible apreciar mediante el análisis de una muestra relativamente pequeña los cambios que tienen como consecuencia el aumento de frecuencia textual de una estructura determinada. En un estudio reciente (Enrique-Arias y Bouzouita 2013) hemos aprovechado estas ventajas metodológicas para abordar la cuestión de la frecuencia textual de los clíticos de objeto. Para ello hemos hecho un cómputo exhaustivo del número y función de los pronombres átonos que aparecen en fragmentos bíblicos en versiones compuestas en diferentes épocas y con representación de varios géneros (para ampliar el marco cronológico complementamos el corpus medieval con versiones del siglo XVI y XX). Los resultados aparecen resumidos en la Tabla 1.

FUNCIÓN GRAMATICAL	E6/8 (CA. 1250)	OSO (1569)	JERUSALÉN (1966)	VARIACIÓN
No reflexivo	412 (76%)	424 (63,3%)	491 (60%)	-21,1%
Reflexivo	130 (24%)	247 (36,7%)	328 (40%)	+66,7%
TOTAL	**542**	**670**	**819**	**+51,1%**

Tabla 1. Clíticos reflexivos y no reflexivos en tres cortes sincrónicos

Como puede apreciarse en los resultados de la Tabla 1, la comparación de las diferentes versiones nos permite comprobar que en efecto se ha producido un aumento de frecuencia textual de los clíticos del español. El cómputo de los clíticos empleados en los fragmentos del XIII arroja un total de 542 ocurrencias. En la versión del XVI, que, recordemos, son textos de la misma extensión que expresan los mismos contenidos, obtenemos 670 clíticos, y en la del XX, 819; es decir, se da un aumento de frecuencia textual de algo más del 50%. Los datos del estudio nos indican también que esta evolución se debe en gran medida al aumento proporcional de los diferentes tipos de reflexivos, que pasan de representar el 24% de los clíticos (130/542) en el siglo XIII a ser el 40% en el XX (328/819).

Haciendo un estudio pormenorizado de la frecuencia textual de los diferentes tipos de clíticos detectamos que los datos reflejan cabalmente los procesos que han sido descritos en la evolución de los diferentes significados y funciones de los clíticos de objeto del español. Los desarrollos semánticos del OI y de los diferentes tipos de estructuras reflexivas desde significados más referenciales a otros más metafóricos se han visto acompañados de aumentos en la frecuencia textual. En lo que respecta a los clíticos no reflexivos, la función de OD prácticamente no ha variado su frecuencia mientras que las funciones del OI han aumentado en torno al 50%. Los diferentes tipos de reflexivos han aumentado por encima del 100% y, en el caso de los desarrollos más recientes hacia la pasiva impersonal, han pasado de no registrarse en los textos medievales de nuestro corpus a tener una presencia significativa en los textos posteriores. Mientras el aumento de los clíticos no reflexivos es inferior a un 20%, el de los reflexivos supera el 150%.

Nuestra investigación pone de manifiesto la utilidad de la metodología de los corpus paralelos para obtener de manera sencilla y directa un gran número de estructuras lingüísticas insertas en el mismo contexto de ocurrencia. De este modo es posible observar de manera más controlada cambios diacrónicos de fenómenos morfosintácticos cuya variación está regida por una combinación compleja de factores. Al tratarse de textos con el mismo contenido, la comparación de versiones bíblicas compuestas en diferentes periodos históricos permite apreciar mejor que en los corpus convencionales los cambios que tienen como consecuencia el aumento de frecuencia textual de una estructura determinada y hacerlo además con una selección de texto relativamente pequeña.

5.3. Variable *origen geográfico*

Un parámetro que ha adquirido una renovada relevancia en estudios lingüísticos de orientación diacrónica en los últimos tiempos es el de la variación diatópica. Investigaciones como las recogidas en Inés Fernández-Ordóñez (2011) nos demuestran que las innovaciones lingüísticas se transmiten a través del espacio geográfico y por tanto es importante controlar en lo posible el factor diatópico en la caracterización de los textos de los corpus diacrónicos. Los grandes corpus convencionales (*CORDE*, *CE*) no nos dan información sobre el origen geográfico de los textos (y difícilmente lo podrían hacer pues para muchos textos medievales ni siquiera se conoce con certeza, o la intermediación de copias en la transmisión textual ha desdibujado los rasgos informativos de la variedad del autor de la obra). En el caso de *BM* se está procediendo a estudiar la caracterización dialectal de los textos del corpus (véase por ejemplo el trabajo de Matute y Pato 2010 para la biblia E6 o Enrique-Arias y Matute 2010 para la de Arragel) pero esta labor dista de estar concluida.

En realidad nos encontramos ante una carencia general de la filología española, que no ha desarrollado todavía una dialectología histórica que nos permita conocer con detalle los rasgos de las variedades diatópicas del español a lo largo de la historia de la lengua. Aparte de que se trata de una información con un interés intrínseco, pues nos permite conocer mejor el origen y la difusión de las innovaciones en la historia del español, esta laguna tiene serias consecuencias para la comparabilidad de los textos que empleamos en los estudios de corpus. Si no conocemos la adscripción dialectal de las muestras de textos de diferentes épocas que consideramos en nuestros análisis nos exponemos a cometer errores de interpretación, pues podemos acabar atribuyendo a otros factores diferencias que en realidad se deben a la variación dialectal.

Pensemos por ejemplo en la distribución de las formas largas y breves de los pronombres y adjetivos demostrativos (los pares como *este* y *aqueste*) que conviven a lo largo del español medieval y clásico. Una búsqueda de *estos* y *aquestos* en el *CORDE* y el *CE* (ver Figura 2) nos da lo que podríamos denominar una *curva de cambio fallido*, es decir, una curva ascendente que representa un aumento de frecuencia de la variante innovadora a lo largo del tiempo seguida de un descenso pronunciado que indica que la variante en cuestión no llega a generalizarse y acaba por disminuir y desaparecer de forma abrupta[9]. Tanto en un corpus como en el otro observamos que la forma larga, que siempre ha sido minoritaria, aumenta su frecuencia relativa alcanzando su momento de mayor presencia hacia finales de la Edad Media (en el siglo XIV, según el *CORDE*, o el XV, según el *CE*) para declinar y prácticamente desaparecer en el XVI.

[9] Hemos rastreado el contraste entre *estos* y *aquellos* por ser *estos* la única forma del demostrativo que permitía una búsqueda automática al no presentar problemas de homofonía con formas del verbo *estar* (*esto*/*estó*, *esta*/*está*, *estas*/*estás*, *este*/*esté*).

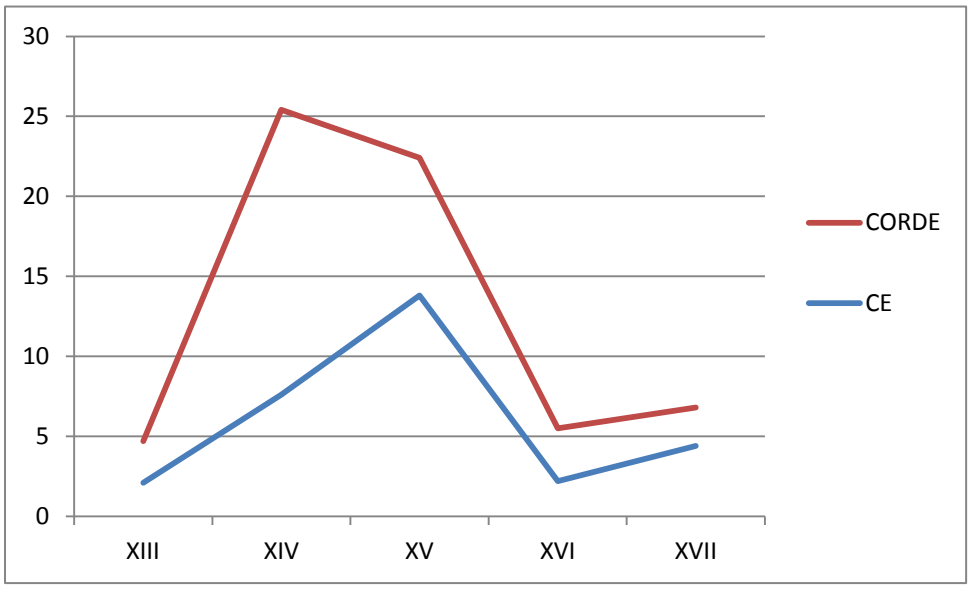

Figura 2. Porcentaje de *aquestos* frente a *estos* en *CORDE* y *CE*

Al mismo tiempo vemos discordancias importantes entre los datos de los dos corpus. En el *CORDE* el momento de mayor frecuencia de la forma larga es el siglo XIV, donde llega al 25%, mientras en el *CE* el pico se produce más tardíamente, en el XV, con valores que no llegan al 15%.

Los estudios que se han ocupado del asunto han propuesto una serie de factores que favorecen el empleo de las formas largas (resumidos en Ranson 2005), como la mayor aparición en verso que en prosa y en diálogo frente a narración, el empleo más frecuente cuando hay presencia física del referente, y los valores pragmáticos de uso contrastivo, de aserción contraria a lo esperado, o para señalar un cambio de estatus informacional a tópico. Al mismo tiempo ningún estudio considera la posibilidad de una adscripción geográfica concreta para el fenómeno ni controla la procedencia geográfica de los textos que analizan. Se trata, no obstante, de un aspecto potencialmente relevante: como las formas largas se han generalizado en el catalán cabría pensar que tuvieran una mayor presencia en textos de procedencia oriental.

El *CODEA* y los demás corpus que se están creando en el ámbito de la red *CHARTA* constituyen un recurso insustituible para poder estudiar el aspecto dialectal de los cambios lingüísticos de forma sistemática pues se trata de textos datables con asignación geográfica contrastada. Con el objeto de observar la distribución geográfica de *este* y *aqueste* en el español medieval he seleccionado los textos del *CODEA* con presencia de formas largas y he codificado los demostrativos que contienen. Gracias a la información de lugar de emisión de los documentos he podido separar los de Castilla y León de los orientales (Navarra y Aragón). Los resultados aparecen en la Figura 3:

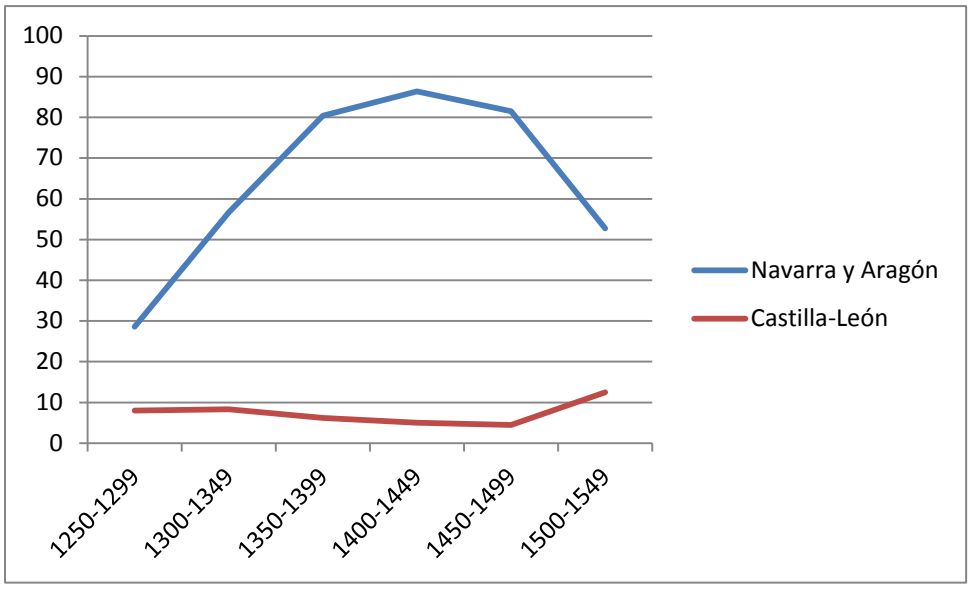

Figura 3. Distribución de formas tipo *este* y *aqueste* en el *CODEA* según región

La visión que nos ofrecen los textos del *CODEA* es ciertamente reveladora. Si nos fijamos solamente en los textos castellanos y leoneses vemos que las formas reforzadas han tenido una presencia minoritaria, por debajo del 10%, y que no se da la curva de cambio fallido que apreciamos en el *CORDE* y el *CE*. Al mismo tiempo, en los documentos navarroaragoneses es mucho mayor la presencia de formas largas, que es mayoritaria a lo largo de gran parte del periodo considerado y sí que se ha producido la curva de cambio fallido. La cuestión merece un estudio más detallado, pero podemos aventurar, a la vista de los datos, que la distribución que encontramos en el *CORDE* y el *CE* se debe a la mezcla de textos de diferentes áreas geográficas, que dan una impresión de cambio fallido, pero que posiblemente no es exacta en el caso de los textos castellanos y sí lo es, pero con porcentajes mucho más altos de formas largas, para los orientales. Asimismo, las discrepancias que encontramos entre el *CORDE* y el *CE* se podrían deber a la diferente datación de los textos (*CORDE* lo hace por fecha de composición y *CE* por fecha del testimonio) y a la distribución de textos de diferentes áreas geográficas. En cualquier caso, los datos de diferentes corpus que presentamos aquí ponen de manifiesto la necesidad de controlar la procedencia geográfica de los textos de los corpus diacrónicos para poder asegurar la comparabilidad de las muestras de cada periodo histórico representado en los mismos. Por todo ello, una de las labores prioritarias en el desarrollo del corpus *BM* es la caracterización dialectal de los textos que lo componen.

6. CONCLUSIONES

El estudio del cambio lingüístico en español se ha visto beneficiado por la aparición de bases de datos textuales de uso libre en la red que nos permiten hacer en un instante búsquedas

en textos históricos que suman millones de palabras. Junto a las ventajas innegables que nos ofrecen estos nuevos recursos metodológicos, los estudios de caso de las páginas precedentes sirven para poner de relieve sus limitaciones en lo que atañe a los parámetros de perspectiva y comparabilidad. Hemos mostrado también cómo *BM*, un corpus paralelo de traducciones bíblicas medievales, puede aportar soluciones para remediar tales deficiencias.

El corpus *BM* abre nuevas perspectivas en el estudio histórico de la variación y el cambio lingüísticos en el español antiguo al constituir un complemento útil y valioso a los corpus existentes. Con la ayuda del corpus bíblico es posible la aplicación de análisis cuantitativos y cualitativos con algunas ventajas claras sobre los corpus convencionales. En primer lugar, la metodología de los corpus paralelos es más abierta, ya que es posible analizar todas las formas utilizadas para expresar un contenido. Del mismo modo, los textos paralelos ofrecen comparabilidad directa de ejemplos concretos a través de diferentes períodos históricos, ya que los equivalentes de traducción se insertan en contextos de ocurrencia idénticos o muy semejantes. Además, en un corpus paralelo es posible analizar la variación estilística de forma más controlada analizando cómo el mismo traductor selecciona diferentes opciones lingüísticas en función del género de cada texto. Por último, en un corpus paralelo es más fácil rastrear los cambios que conllevan un aumento de frecuencia textual.

Es de esperar que la disponibilidad de esta nueva herramienta facilite la aplicación del método de los corpus paralelos en los estudios de historia de la lengua y permita enriquecer desde una perspectiva teórica la comprensión de los fenómenos de cambio y variación del español en perspectiva diacrónica.

REFERENCIAS BIBLIOGRÁFICAS

BM = ENRIQUE-ARIAS, Andrés (dir.) (2009-): *Biblia Medieval*. http://bibliamedieval.es.

BYBEE, Joan L. (2007): *Frequency of use and the organization of language*. Oxford: OUP.

CE = DAVIES, Mark (2002-): *Corpus del Español*. http://www.corpusdelespanol.org.

CODEA = SÁNCHEZ-PRIETO BORJA, Pedro (dir.) (2010-): *Corpus de Textos Españoles Anteriores a 1700*. http://www.textoshispanicos.es.

CORDE = REAL ACADEMIA ESPAÑOLA (1998-): *Corpus Diacrónico del Español*. http://www.rae.es.

COROMINAS, Joan y José A. PASCUAL (1980-1991): *Diccionario crítico etimológico castellano e hispánico*. Madrid: Gredos.

DAVIES, Mark (2009): «Creating useful historical corpora: A comparison of *CORDE*, the *Corpus del español*, and the *Corpus do português*», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Frankfurt/Madrid: Vervuert/Iberoamericana, pp. 137-166.

DE VRIES, Lourens (2007): «Some remarks on the use of Bible translations as parallel texts in linguistic research», en Michael Cysow y Bernhard Wälchli (eds.), *Parallel Texts: Using translational*

equivalents in linguistic typology. Número especial de *Sprachtypologie und Universalienforschung* (*STUF*), 60, pp. 95-99.

ENRIQUE-ARIAS, Andrés (2005): «When clitics become affixes, where do they come to rest? A case from Spanish», en Wolfgang Dressler, Dieter Kastovsky, Oskar Pfeiffer y Franz Rainer (eds.), *Morphology and its demarcations*. Filadelfia/Amsterdam: John Benjamins, pp. 67-79.

ENRIQUE-ARIAS, Andrés (2008): «Biblias romanceadas e historia de la lengua», en Concepción Company y José Moreno de Alba (eds.), *Actas del VII Congreso Internacional de Historia de la Lengua Española*, vol. II. Madrid: Arco/Libros, pp. 1781-1794.

ENRIQUE-ARIAS, Andrés (ed.) (2009a): *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Frankfurt/Madrid: Vervuert/Iberoamericana.

ENRIQUE-ARIAS, Andrés (2009b): «Lingüística de corpus y diacronía de las lenguas iberorromances», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Frankfurt/Madrid: Vervuert/Iberoamericana, pp. 11-21.

ENRIQUE-ARIAS, Andrés (2009c): «Ventajas e inconvenientes del uso de *Biblia medieval* (un corpus paralelo y alineado de textos bíblicos) para la investigación en lingüística histórica del español», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Frankfurt/Madrid: Vervuert/Iberoamericana, pp. 269-283.

ENRIQUE-ARIAS, Andrés (2012): **«**On the usefulness of using parallel texts in diachronic investigations: insights from a parallel corpus of Spanish medieval Bible translations**»**, en Paul Bennett, Martin Durrell, Silke Scheible y Richard J. Whitt (eds.), *New Methods in Historical Corpus Linguistics*. Tubinga: Gunter Narr (*Corpus linguistics and Interdisciplinary perspectives on language*, vol. 3).

ENRIQUE-ARIAS, Andrés y Miriam BOUZOUITA (en prensa): «La frecuencia textual en la evolución histórica de los clíticos pronominales en español**»,** *Iberoromania.*

ENRIQUE-ARIAS, Andrés y Joan BURGUERA (2010): **«**Variación y cambio en la formalización de la interrogación retórica en la historia del español». Comunicación leída en el *XXVI Congreso Internacional de Lingüística y Filología Románicas* (Universidad de Valencia, 6-11 de septiembre 2010).

ENRIQUE-ARIAS, Andrés y Laura CAMARGO (en prensa): **«**Problemas en torno a la caracterización de un marcador del discurso en español medieval: el caso de *he*», en Amalia Rodríguez Somolinos, Margarita Borreguero Zuloaga y Sonia Gómez-Jordana Ferary (eds.), *Marqueurs discursifs dans les langues romaines*. Limoges: Lambert Lucas.

ENRIQUE-ARIAS, Andrés y Cristina MATUTE MARTÍNEZ (2010): «El estudio morfosintáctico de la lengua de la *Biblia de Alba*: un acercamiento a la variación discursiva y dialectal del español del siglo XV», en Maria Iliescu, Heidi M. Siller-Runggaldier, Paul Danler (eds.), *Actes du XXV Congrès International de Linguistique et de Philologie Romanes* (Innsbruck, 3-8 septiembre 2007), vol. VI. Berlin: Walter de Gruyter, pp. 115-123.

FERNÁNDEZ-ORDÓÑEZ, INÉS (2011): *La lengua de Castilla y la formación del español*. Discurso leído el 13 de febrero de 2011 en su recepción pública por la Excma. Sra. D.ª Inés Fernández-Ordóñez y contestación del Excmo. Sr. D. José Antonio Pascual. Madrid.

GOYENS, Michèle y Willy van HOECKE (1992): «La traduction comme témoin de l'évolution linguistique», en Ramón Lorenzo (ed.), *Actas do XIX Congreso Internacional de Lingüística e Filoloxía Románicas*, vol. V. A Coruña: Fundación Barrié de la Maza, pp. 13-32.

JOSEPH, Brian (2008): «Historical linguistics in 2008. The state of the art», en Piet van Sterkenburg (ed.), *Unity and diversity of languages*. Amsterdam/Filadelfia: John Benjamins, pp. 175-188.

KAISER, Georg A. (2005): «Bibelübersetzungen als Grundlage für empirische Sprachwandeluntersuchungen», en Claus D. Pusch, Johannes Kabatek y Wolfgang Raible (eds.), *Romance Corpus Linguistics II. Corpora and Diachronic Linguistics*. Tübingen: Gunter Narr Verlag, pp. 71-83.

LAPESA, Rafael (2000 [1971]): «Sobre el artículo ante posesivo en castellano antiguo», en Rafael Cano y M. Teresa Echenique (eds.), *Estudios de morfosintaxis histórica del español*. Madrid: Gredos, pp. 413-435.

MATUTE MARTÍNEZ, Cristina y Enrique PATO (2010): «Morfología y sintaxis en el códice Escorial I.I.6», en Andrés Enrique-Arias (ed.), *La Biblia Escorial I.I.6. Transcripción y estudios*. Logroño: Centro Internacional de Investigación de la Lengua Española (Cilengua), pp. 45-65.

MENÉNDEZ PIDAL, Ramón (1911): *Cantar de Mio Cid*. *Texto, gramática y vocabulario*. Madrid: Bailly/Ballière.

MCENERY, Tony y Zhonghua XIAO (2007): «Parallel and comparable corpora: The state of play», en Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori y Yoichiro Tsuruga (eds.): *Corpus-Based Perspectives in Linguistics*. Amsterdam: John Benjamins, pp. 131-145.

RANSON, Diana L. (2005): «Variation of the Spanish demonstratives *aqueste* and *este*», en Roger Wright y Peter Ricketts (eds.), *Studies on Ibero-Romance linguistics dedicated to Ralph Penny*. Newark: Juan de la Cuesta, pp. 187-214.

RESNIK, Philip, Mari B. OLSEN y Mona DIAB (1999): «The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'», *Computers and the Humanities*, 33, 1-2, pp. 129-153.

SÁNCHEZ LÓPEZ, Cristina (en prensa): «Preposiciones, conjunciones y adverbios derivados de participios», en Concepción Company (ed.), *Sintaxis histórica de la lengua española. Tercera parte*. México: Fondo de Cultura Económica.

WANNER, Dieter (2005): «The corpus as a key to diachronic explanation», en Johannes Kabatek, Claus D. Pusch y Wolfgang Raible (eds.), *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tübingen: Gunther Narr, pp. 31-44.