HERRAMIENTAS INFORMÁTICAS PARA EL ESTUDIO DIACRÓNICO DEL GALLEGO

XOSÉ AFONSO ÁLVAREZ PÉREZ
Universidade de Santiago de Compostela

OBJETIVOS DEL PRESENTE TRABAJO

El propósito de esta intervención, de acuerdo con la petición cursada por las organizadoras que han tenido la amabilidad de invitarme, es presentar diferentes herramientas informáticas desarrolladas en Galicia en los últimos años y que son de interés para el estudio diacrónico de la lengua gallega, pero, obviamente, también pueden dar buenos frutos al hispanista o al romanista, como ya han puesto de manifiesto algunos trabajos (Santamarina 2004), y pueden servir –y, de hecho, sirven (Corrales y Corbella 2007)– de inspiración para el desarrollo de utilidades semejantes en otros ámbitos lingüísticos[1]. La escasez de espacio motiva que no me pueda detener demoradamente en cada una de las herramientas e impide una aproximación crítica a estas. Esta tarea recaerá, por tanto, en el lector y, para facilitarla, sobre todo con un público no necesariamente familiarizado con la lingüística gallega, procuraré ser prolijo en las referencias bibliográficas, incluyendo no sólo descripciones más o menos detalladas de las obras aquí reseñadas, sino también trabajos que deben mucho, para su elaboración, a estas herramientas. Del mismo modo, para las herramientas web no ofreceré ninguno de los gráficos usados en la presentación oral de esta intervención, ya que están a tiro ¿de piedra? en el ciberespacio.

Presentaré aquí una bibliografía informatizada, tres *corpora* que abarcan distinta tipología textual y un arco cronológico diferente y dos productos denominados *diccionario de diccionarios*, esto es, un tesoro que permite la consulta simultánea de múltiples obras lexicográficas, como se detallará en su momento.

1. BIBLIOGRAFÍA INFORMATIZADA DA LINGUA GALEGA (*BILEGA*)

La primera de las herramientas –descrita en detalle por López Martínez (1998) y, sobre todo, García Gondar (2003)– es una exhaustiva base de datos que recoge[2] 14645 registros de obras que, de un modo u otro, se ocupan de lingüística gallega. Este recurso, dirigido por el Prof. Francisco García Gondar, de la Universidade de Santiago de Compostela, se hospeda en la web http://www.cirp.es/bdo/bil, que cuelga del sitio

[1] Para un abordaje de este tema desde el punto de vista de la informática, véase Imaxín Software (en prensa).
[2] Todos los datos estadísticos de este apartado se refieren al 31 de diciembre de 2007.

web del Centro Ramón Piñeiro para a Investigación en Humanidades, dependiente de la Xunta de Galicia.

No consiste esta esencial utilidad en un simple repertorio bibliográfico, sino que la ficha de cada obra indica diferentes parámetros, que el motor de búsqueda puede utilizar como criterios para recuperar información. Por su interés, cabe señalar los siguientes campos:

a) Tema(s), entre un amplio tesauro clasificado y jerarquizado que comprende 529 posibilidades.
b) Período cronológico.
c) Variedades de lengua, entre 25 posibilidades, diatópicas, diastráticas y diafásicas.
d) Identificadores (nombres propios de obras, autores e instituciones citados en el documento).

Es importante señalar que más del noventa por ciento de los registros (para ser exactos, 13534, un 92'41%) incorpora una sinopsis del contenido de la obra fichada en *BILEGA*.

Casi tres mil registros (2851) incluyen la referencia de un sitio web en el que poder consultar el texto íntegro (o una porción considerable) de la obra incorporada a la base de datos, lo que convierte a *BILEGA* en un repositorio documental de primer orden para la lingüística gallega. Por último, es importante señalar que existe un campo que da cuenta de las reseñas existentes para cada uno de los trabajos incluidos; actualmente están catalogadas 4591 recensiones.

Parte de esta herramienta se ha ido publicando también en formato papel, según criterios cronológicos; en concreto, existe un volumen consagrado al estudio de las obras de lingüística gallega desde los inicios hasta el año 1994[3] y otro dedicado a la producción del año 2004[4]. Existen, además, varios trabajos que explotan las posibilidades de *BILEGA* dentro del campo de la historiografía lingüística[5].

2. CORPUS DOCUMENTALE LATINUM GALLAECIAE (*CODOLGA*)

Este recurso también se hospeda en el Centro Ramón Piñeiro para a Investigación en Humanidades (http://balteira.cirp.es/codolga); está dirigido por el Prof. José Eduardo López Pereira (Universidade da Coruña).

Esta herramienta[6] proporciona acceso a un amplio corpus de documentación medieval latina relacionada con Galicia desde el siglo VI hasta el XV, obtenida principalmente de cartularios y colecciones de centros monásticos, además de la documentación de las sedes catedralicias de Santiago de Compostela, Mondoñedo, Lugo, Tui, Ourense o Astorga.

El *CODOLGA* es un buscador que explota la base de datos *GALADOC*, que (datos de febrero de 2007) abarca 12000 documentos, obtenidos de 150 ediciones distintas en libros, tesis, revistas, etc., y supera los ocho millones de formas. El equipo de investigación ha emprendido también el proyecto de ofrecer online en su web

[3] García Gondar (dir.) (1995).
[4] García Gondar (dir.) (2006).
[5] Señalemos como ejemplo García Gondar (2000) y García Gondar (2005).
[6] Descrita en Díaz de Bustamante (2004a) y en López Alsina (2005). Díaz de Bustamante (2004b) repasa diferentes estrategias y condiciones para la edición electrónica de textos medievales y, hacia el final, se detiene en el *CODOLGA*. Es también útil Díaz de Bustamante (2005), que se centra en estudiar el manejo y gestión de bases de datos textuales, tratando también diferentes aspectos de la concepción y organización de la herramienta que describimos aquí.

colecciones documentales inéditas, y ya está disponible la del monasterio de Santa María de Meira[7].

El sistema de recuperación de información es muy sencillo: ofrece la posibilidad de búsqueda por palabra, parte de palabra o conjunto de palabras, pudiendo utilizarse también comodines y operadores lógicos. Es posible restringir el conjunto textual sobre el que se busca de acuerdo con una serie de parámetros:

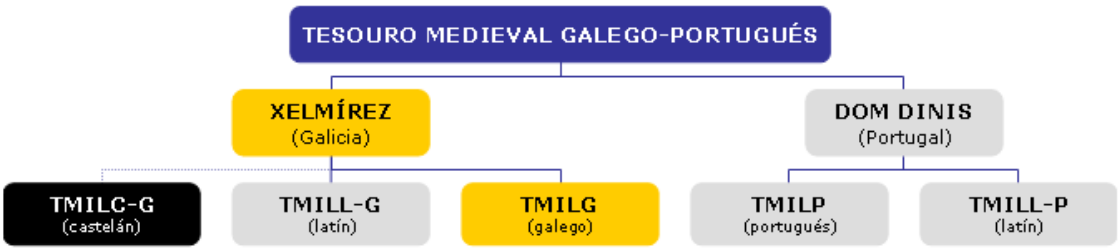
a) Otorgante (5 tipos: episcopal, particular, pontificio, real, sin especificar).
b) Soporte (5 tipos: códice o fragmento, libro impreso, papel suelto, pergamino suelto, sin especificar).
c) Procedencia.
d) Fecha del documento.

Una vez introducida la palabra o palabras que nos interesan, el programa nos presenta una estadística de los documentos que incorporan el texto que hemos buscado, señalando cuántas de ellas se inscriben dentro de las diferentes categorías que se han presentado en el párrafo anterior. En esa pantalla, el investigador decide con qué criterios quiere ordenar los resultados y accede, posteriormente, a la última pantalla, en la que aparece el texto en cuestión, con un breve contexto a izquierda y derecha, la edición de referencia e información sobre el documento de procedencia y la situación de la cita dentro de éste.

Couceiro (2005) es un ejemplo de un trabajo lingüístico-filológico realizado con el apoyo esencial del *CODOLGA*.

3. TESOURO MEDIEVAL INFORMATIZADO DA LINGUA GALEGA (*TMILG*)

El tercer recurso que presento es, como el anterior, un corpus de ámbito medieval, pero de textos gallegos. El *Tesouro Medieval Informatizado da Lingua Galega* (*TMILG*) está dirigido por el Prof. Xavier Varela (USC) y albergado en el sitio web del Instituto da Lingua Galega: http://ilg.usc.es/tmilg[8]. Podemos señalar que este recurso constituye el pilar más importante del *Proxecto Xelmírez*, que es, a su vez, una de las dos ramas del *Tesouro Medieval Galego-portugués*[9].



Esta herramienta recoge textos gallegos desde finales del siglo VIII hasta el año 1600, abarcando unas 16000 unidades textuales pertenecientes a 82 obras de diferente tipo, tanto literarias como no literarias. Como en el caso del *CODOLGA*, se pueden realizar búsquedas por palabra, parte de palabra o conjunto de palabras, pudiendo

[7] http://balteira.cirp.es/codolga/meira_portada.html
[8] Los antecedentes de esta herramienta han sido explicados por Varela Barreiro (2004); una descripción más actual del *TMILG* es la de Martínez Lema (en prensa).
[9] En la página http://ilg.usc.es/tmilg/proxecto.html pueden consultarse más detalles sobre este proyecto, por ahora mucho más desarrollado en la rama gallega que en la portuguesa.

utilizarse también comodines y operadores lógicos. Es posible restringir el conjunto textual sobre el que se busca de acuerdo con una serie de parámetros:

a) género (prosa o verso).
b) subgénero (9 tipos: prosa jurídica, lírica profana, etc.).
c) obra a la que pertenece el documento.
d) fecha.

Una vez introducida la cadena de texto que nos interesa recuperar y, en su caso, las restricciones oportunas, aparecemos en una primera pantalla que nos ofrece todas las formas posibles de acuerdo con las condiciones estipuladas. Allí, podemos seleccionar las voces que nos interesen y, a continuación, pasar a las ventanas de estadísticas (por tipología textual y por formas) y a la de concordancias, que nos ofrece la palabra en su contexto y diferentes datos (autor, obra, tipología, fecha, etc.).

4. TESOURO INFORMATIZADO DA LINGUA GALEGA (*TILG*)

La cuarta herramienta, aunque de aparición anterior, es la continuación cronológica de la tratada en el apartado precedente, puesto que recoge textos de todo género y registro desde 1612 a la actualidad. El *Tesouro Informatizado da Lingua Galega* (*TILG*), está dirigido por el Prof. Antón Santamarina (USC) y hospedado también en la página web de esta misma universidad: http://www.ti.usc.es/TILG.

Este corpus recoge documentos de todo tipo (incluso transcripciones de grabaciones orales), aunque, evidentemente, está condicionado por las circunstancias históricas de nuestra lengua, que explican, por ejemplo, la escasez de textos científicos o ensayísticos. En conjunto, se han vertido 1464 obras, aunque está previsto realizar próximamente un volcado de nuevos textos, por lo que en el momento en que esté impreso este trabajo, la cantidad será considerablemente mayor.

Se trata de una herramienta que supera los 11 millones de registros, agrupados en más de noventa mil lemas. Esta aplicación permite recuperar datos por lema, por palabra o por aparición próxima de dos palabras en un texto; se permite el uso de comodines. Se puede restringir la búsqueda según la fecha, los autores, las obras o la categoría gramatical de las formas investigadas. Aunque el contexto a izquierda y derecha que aparece en la pantalla principal de resultados es bastante pequeño (en la altura de las obras reseñadas anteriormente), un botón permite ampliar considerablemente el texto (se dan siete líneas). Es posible exportar los resultados a XML y XSL para trabajar en modo local con los datos.

5. DICCIONARIO DE DICCIONARIOS Y DICIONARIO DE DICIONARIOS DO GALEGO MEDIEVAL

Presentaré en este apartado dos herramientas comercializadas en CD-Rom, de concepción similar y que comparten una *interface* casi idéntica; son el *Diccionario de diccionarios*, dirigido por el Prof. Antón Santamarina (USC) y publicado por la Fundación Barrié, y el *Dicionario de dicionarios do galego medieval*, dirigido por el Profesor Ernesto González Seoane (USC) y publicado por la Universidade de Santiago de Compostela como anexo de la prestigiosa revista *Verba*.

Parafraseando al profesor Gutiérrez Cuadrado[10], estas herramientas vierten el vino añejo de las obras lexicográficas en odres nuevos mágicos, que permiten consultar simultáneamente varias decenas de trabajos, pero sin mezclarlos, dejando que el catador

[10] Gutiérrez Cuadrado (2006).

recupere el líquido de las botellas que le interesan. Es interesante señalar que, a diferencia de herramientas como el *Nuevo Tesoro Lexicográfico*, estos recursos ofrecen el acceso a las obras en formato texto, no como imagen, lo que abre la puerta a muchas más posibilidades de búsqueda y de trabajo con los datos.

5.1. *Diccionario de diccionarios*

La versión que se comenta[11], la última en el momento de acometer la escritura del trabajo, es la tercera, del año 2003, que integra las dos anteriores versiones y añade nuevos trabajos; en total, se incluyen obras de veinticuatro lexicógrafos o equipos de lexicógrafos, que pueden ir desde un solo diccionario hasta recopilaciones de decenas de artículos[12]. Se está trabajando en una cuarta edición y existe el proyecto, a medio plazo, de volcar en Internet toda esta información.

Actualmente, cuenta con unas 345742 entradas acumulativas –esto es, la suma de las entradas individuales de cada uno de los diccionarios o lista de voces incluidos–, que se contienen en 136164 lemas distintos. La pantalla principal permite la consulta mediante lemas (en la columna de la izquierda tenemos todo el listado, en la parte de la derecha, la fila superior nos permite acceder al diccionario deseado, cuyo texto se muestra en la parte inferior), pero no son todas las posibilidades del *DdD*, ya que es posible buscar también en el cuerpo de las entradas, a partir de referencias en gallego, castellano, afines, localidad, autor, ejemplos, refranes y poemas. Un código de colores indica las diferentes categorías de la entrada (lema, refranes, ejemplos, voces gallegas dentro del cuerpo de la definición, etc.).

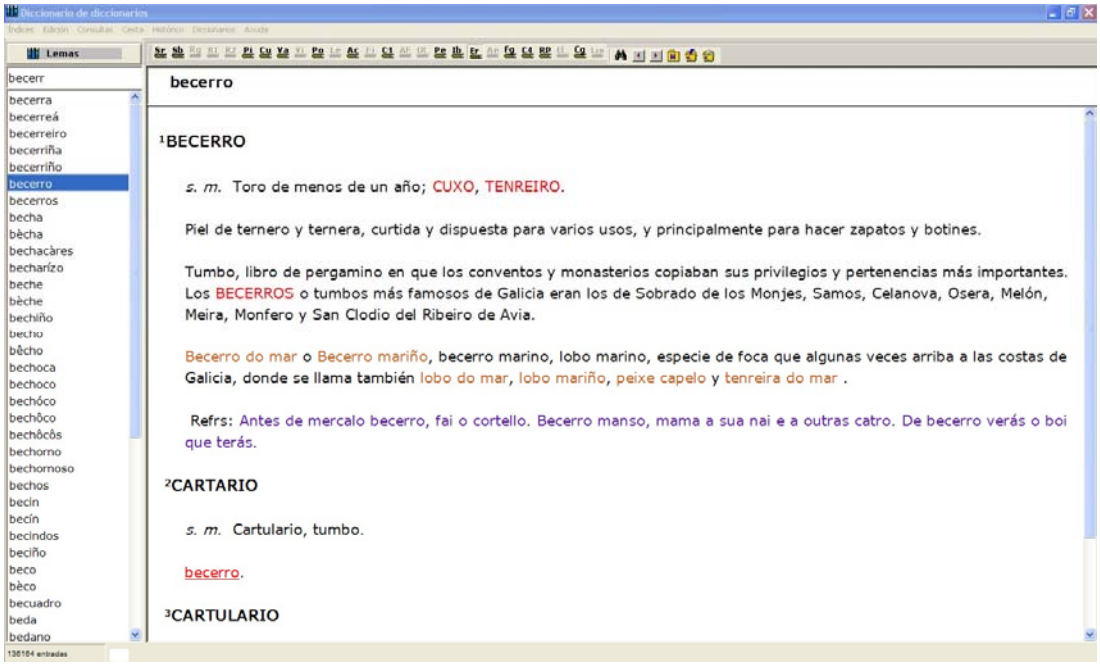


Gráfico 2: Pantalla principal del programa, con un lema y una obra seleccionados

[11] Una descripción desde un punto de vista más técnico es la de Fernández Cabezas y Pichel Campos (2000).

[12] Es inútil reseñar aquí todos los trabajos incluidos, que recogen los nombres clave de la lexicografía gallega: Sarmiento, Valladares, Aníbal Otero, Eladio Rodríguez, etc. En el propio CD-Rom se incluye un archivo que ofrece la relación completa de obras y una aproximación biobibliográfica a los diferentes autores. Son también recomendables otros archivos, versiones completas de trabajos del propio Santamarina sobre lexicografía gallega y conformación del modelo estándar de lengua.

5.2. *Dicionario de dicionarios do galego medieval*

Esta obra, dirigida por el Prof. Ernesto González Seoane, es una herramienta más reciente, publicada en el año 2006 como anexo de la revista *Verba*, con el número 57. Incluye 14 glosarios o vocabularios de textos gallegos o de la tradición común gallego-portuguesa; se está trabajando en una segunda edición que supera notablemente esta cifra[13].

Como se señalaba antes, mantiene la misma estructura del *Diccionario de diccionarios* reseñado en el apartado anterior, aunque, lógicamente, puesto que la estructura de las obras incluidas no es la misma, cambian los parámetros de recuperación de información: equivalencias/definiciones, étimos, antropónimos, topónimos, ejemplos gallego-portugueses o en otras lenguas, sintagmas y fuentes textuales o bibliográficas.

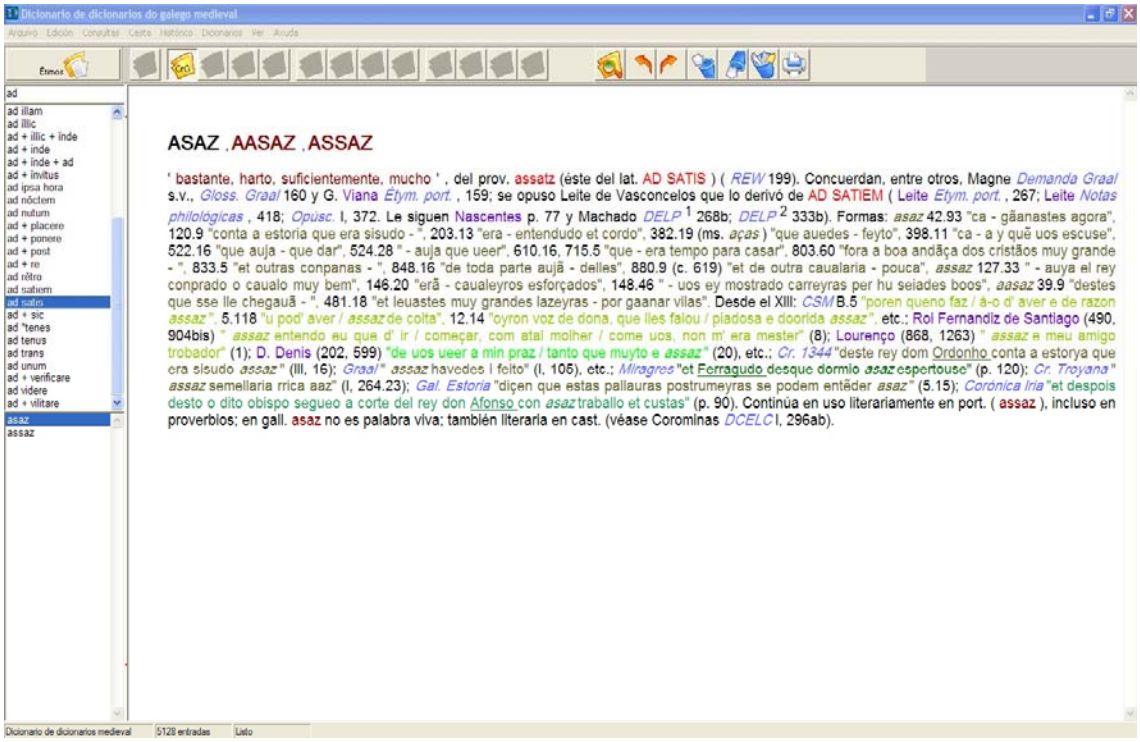


Gráfico 3: *Ejemplo de búsqueda por étimo*

6. PERSPECTIVAS FUTURAS

Para concluir, quisiera señalar que actualmente está en marcha un proyecto conjunto de las universidades de Vigo y Santiago para constituir el *RILGA*, *Recursos Informatizados da Lingua Galega*, que permitirá consultar conjunta o separadamente los dos diccionarios de los que he hablado, además de integrar diferentes herramientas[14]

[13] El propio CD-Rom contiene en formato PDF una completa guía de uso y descripción de las obras incluidas (más de cincuenta páginas). Una descripción de la herramienta, más centrada en el aspecto metodológico, es González Seoane (2005), mientras que González Seoane (en prensa) se centra más en la utilización y posibilidades de este programa informático.

[14] No he considerado oportuno detallarlas aquí al no estar estas herramientas tan orientadas al estudio diacrónico de la lengua. López Guinovart (en prensa) es un análisis de dos corpus lingüísticos, aunque los trabajos de este equipo abarcan, además, correctores, traductores automáticos, bases de datos de

promovidas por la Universidade de Vigo en el seno de su muy recomendable *Seminario de Lingüística Informática* (http://sli.uvigo.es).

REFERENCIAS BIBLIOGRÁFICAS

CORRALES, Cristóbal / CORBELLA, Dolores (2007): «El proyecto *Bilican*: nuevos datos y perspectivas». Josefa Dorta (ed.), *Temas de dialectología*. La Laguna – Tenerife: Instituto de Estudios Canarios.

COUCEIRO, Xosé Luís (2005): «Codolga na investigación do léxico hispánico primitivo». Ana Isabel Boullón Agrelo *et al.* (eds.), *As tebras alumeadas: Estudos filolóxicos ofrecidos en homenaxe a Ramón Lorenzo*. Santiago de Compostela, Universidade de Santiago de Compostela, 83-102.

DÍAZ DE BUSTAMANTE, José Manuel (2004a): «Noticia e presentación dunha nova ferramenta de investigación: Corpus Documentale Latinum Gallaeciae (CODOLGA)». *Compostellanum: revista de la Archidiócesis de Santiago de Compostela*, vol. 49, nº 3-4, 755-762.

DÍAZ DE BUSTAMANTE, José Manuel (2004b): «Los trabajos y los días: acerca de colecciones y ediciones de documentos latinos de la Edad Media». Centro de Estudios e Investigación «San Isidoro» (ed.), *Orígenes de las lenguas romances en el reino de León: siglos IX-XII*. León: Centro de Estudios e Investigación «San Isidoro», vol. I, 349-361.

DÍAZ DE BUSTAMANTE, José Manuel (2005): «Notas á xestión de bases de datos textuais: a documentación latina medieval do Reino de Galicia». Ana Isabel Boullón Agrelo *et al.* (eds.), *As tebras alumeadas: Estudos filolóxicos ofrecidos en homenaxe a Ramón Lorenzo*. Santiago de Compostela, Universidade de Santiago de Compostela, 115-125.

FERNÁNDEZ CABEZAS, Antonio / PICHEL CAMPOS, José Ramón (2000): «Diccionario de diccionarios electrónico da língua galega». *Procesamiento del lenguaje natural*, 26, 99-102.

GARCÍA GONDAR, Francisco (2000): «La presencia del gallego en la filología española (1914-1970): Análisis de algunas revistas». Marina Maquieira Rodríguez, Mª Dolores Martínez Gavilán y Milka Villayandre Llamazares (eds.), *Actas del II Congreso Internacional de la Sociedad Española de Historiografía Lingüística (León, 2-5 de marzo de 1999)*. Madrid: Arco Libros, 435-446.

GARCÍA GONDAR, Francisco (2003): «*De re bibliographica*: BILEGA entre os recursos sobre o galego e o portugués na Internet». *Revista Galega de Filoloxía*, 4, 59-95.

GARCÍA GONDAR, Francisco (2005): «A contribución de Harri Meier aos estudos etimolóxicos galegos: bosquexo dunha bibliografía analítica». Ana Isabel Boullón Agrelo *et al.* (eds.), *As tebras alumeadas: Estudos filolóxicos ofrecidos en homenaxe a Ramón Lorenzo*. Santiago de Compostela, Universidade de Santiago de Compostela, 141-164.

GARCÍA GONDAR, Francisco (dir.) (1995): *Repertorio bibliográfico da lingüística galega: Desde os seus inicios ata 1994 inclusive*, Santiago de Compostela: Xunta de Galicia, Centro de Investigacións Lingüísticas e Literarias Ramón Piñeiro.

GARCÍA GONDAR, Francisco (dir.) (2006): *Bibliografía analítica da lingua galega (2004)*. Santiago de Compostela: Xunta de Galicia, Centro Ramón Piñeiro para a Investigación en Humanidades.

GONZÁLEZ SEOANE, Ernesto Xosé (2005): «Aspectos metodolóxicos da elaboración do *Dicionario de dicionarios do galego medieval*». Ana Isabel Boullón Agrelo *et al.* (eds.), *As tebras alumeadas: Estudos filolóxicos ofrecidos en homenaxe a Ramón Lorenzo*. Santiago de Compostela, Universidade de Santiago de Compostela, 165-178.

GONZÁLEZ SEOANE, Ernesto Xosé (en prensa): «El *Dicionario de dicionarios do galego medieval*» *Actas del XIII Congreso Internacional Euralex. 25 años estudiando diccionarios (Barcelona, 15 - 19 Julio 2008)*.

GUTIÉRREZ CUADRADO, Juan (2006): «*Diccionario de Diccionarios*. Edición a cargo de Antón Santamarina [versión 3].- [A Coruña]: Fundación Pedro Barrié de la Maza, 2003.-1 disco CD-ROM + 1 guía, (Biblioteca Filolóxica Galega| Instituto da Lingua Galega). ISBN 84-9752-012-2». *Estudis Romànics*, 28, 370-377.

IMAXIN SOFTWARE (en prensa): «Tecnoloxías informáticas ao servizo da lexicografía». Antón Santamarina, Ernesto González Seoane y Xavier Varela (eds.), *A lexicografía galega do século XXI*. Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega.

LÓPEZ ALSINA, Fernando (2005): «La red y las fuentes documentales medievales: el ejemplo de *CODOLGA*». Departamento de Historia Medieval, Ciencias y Técnicas Historiográficas y Estudios Arabes e Islámicos (ed.), *Pescar o navegar : la Edad Media en la red*. Zaragoza: Universidad de Zaragoza.

---

neologismos, etiquetadores, etc. En la página web del grupo hay un completo listado de publicaciones, casi todas disponibles en línea, estando algunas de ellas dedicadas a la descripción de estas herramientas.

LÓPEZ GUINOVART, Xavier (en prensa): «*Córpora e dicionarios do Seminario de Lingüística Informática: Corpus Lingüístico da Universidade de Vigo (CLUVI) e Corpus Técnico do Galego (CTG)*». Antón Santamarina, Ernesto González Seoane y Xavier Varela (eds.), *A lexicografía galega do século XXI*. Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega. [Existe también edición electrónica] http://webs.uvigo.es/sli/arquivos/sli_ilg07.pdf [consulta: 18/04/2008]

LÓPEZ MARTÍNEZ, Marisol (1998): «A base de datos BILEGA (Bibliografía Informatizada da Lingua Galega)». Alexandre Rodríguez Guerra (dir., coord. e ed.), *Galicia dende Salamanca 2*. Salamanca: Universidad de Salamanca / Santiago de Compostela: Xunta de Galicia, 289-295.

MARTÍNEZ LEMA, Paulo (en prensa): «Os córpora do Instituto da Lingua Galega: o TMILG». Antón Santamarina, Ernesto González Seoane y Xavier Varela (eds.), *A lexicografía galega do século XXI*. Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega.

SANTAMARINA FERNÁNDEZ, Antón (2004): «Lenguas del entorno leonés. Los diccionarios gallegos como instrumento para el hispanista». Centro de Estudios e Investigación «San Isidoro» (ed.), *Orígenes de las lenguas romances en el reino de León: siglos IX-XII*. León: Centro de Estudios e Investigación «San Isidoro», vol. II, 27-68.

VARELA BARREIRO, Francisco Xavier (2004): «Un proxecto do ILG no abalo da *Gramática Histórica da Lingua Galega*». Rosario Álvarez Blanco, Francisco Fernández Rei y Antón Santamarina (eds.), *A Lingua Galega: historia e actualidade: Actas do I Congreso Internacional (Santiago de Compostela, 16-20 de setembro de 1996)*. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, vol. II: 649-684. [Existe también edición electrónica] http://www.consellodacultura.org/mediateca/pubs.pdf/galego_historia_2.pdf. [consulta: 18/04/2008]