Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 375-376 recibido 1-05-2009; aceptado 5-06-2009

RILG: Recursos Integrados da Lingua Galega*

RILG: Integrated Language Resources for Galician

Xavier Gómez Guinovart
Seminario de Lingüística Informática
Universidade de Vigo
xgg@uvigo.es

Antón Santamarina
Instituto da Lingua Galega
Univ. de Santiago de Compostela
anton.santamarina@usc.es

Resumen: La finalidad del proyecto RILG es la integración, explotación conjunta y difusión de los recursos de tecnología lingüística de la lengua gallega generados en distintos proyectos previos llevados a cabo por el Instituto da Lingua Galega de la Universidade de Santiago de Compostela y por el Grupo TALG de la Universidade de Vigo.
Palabras clave: integración de recursos lingüísticos, tecnologías de la lengua gallega

Abstract: The purpose of this project is the integration, joint exploitation and diffusion of the resources of linguistic technology of the Galician language generated in different projects carried out by the Instituto da Lingua Galega of the University of Santiago of Compostela and by the TALG Group of the University of Vigo.
Keywords: integration of linguistic resources, Galician language technologies

1. Datos del proyecto

El proyecto RILG (Recursos Integrados da Lingua Galega) es un proyecto coordinado entre el Grupo TALG de la Universidade de Vigo y el Instituto da Lingua Galega de la Universidad de Santiago de Compostela, que ha obtenido financiación en convocatorias competitivas de los Planes Nacionales de I+D+i del Ministerio de Educación y Ciencia del Gobierno de España (2006-2009) y de la Consellaría de Innovación e Industria de la Xunta de Galicia (2008-2011). Los responsables del proyecto son Xavier Gómez Guinovart (investigador principal del proyecto coordinado y del subproyecto de la Universidade de Vigo) y Antón Santamarina (investigador principal del subproyecto de la Universidade de Santiago de Compostela). Los resultados del proyecto se difunden a través de la web, en la dirección http://sli.uvigo.es/RILG.

* Financiado por el Ministerio de Educación y Ciencia y el Fondo Europeo de Desarrollo Regional (FEDER), dentro del proyecto *Diseño e implementacion de un servidor de recursos integrados para el desarrollo de tecnologías de la lengua gallega (RILG)* del Plan Nacional de I+D+I, 2006-2009 (ref. HUM2006-11125-C02-01/FILO); y por la Consellaría de Innovación e Industria de la Xunta de Galicia, dentro del proyecto *Desenvolvemento e aplicación de recursos integrados da lingua galega* del Plan galego de investigación, desenvolvemento e innovación tecnolóxica (Incite), 2008-2011 (ref. INCITE08PXIB302185PR). Ambos son proyectos coordinados de la Universidade de Vigo (Grupo TALG) con la Universidade de Santiago de Compostela (Instituto da Lingua Galega).

2. Descripción del proyecto

El proyecto RILG tiene como objetivo ofrecer un portal web de servicios lingüísticos del gallego desde el que se pueda acceder de modo conjunto a los bancos de datos textuales y léxicos desarrollados por el Instituto da Lingua Galega (ILG) de la Universidade de Santiago de Compostela y por el Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) de la Universidade de Vigo (formado por los equipos de investigación del Seminario de Lingüística Informática y del Observatorio de Neoloxía de esta universidad).

Los recursos lingüístico-computacionales del gallego que está previsto integrar en la plataforma RILG (algunos de ellos disponibles hasta ahora sólo en versión CD-ROM) proceden, primordialmente del ámbito de la lingüística de corpus, pero también del de la lexicografía y de la terminología computacional y basada en corpus, y son fruto del trabajo de los equipos participantes en proyectos de investigación subvencionados previos. A continuación, se ofrece una breve descripción de los recursos más importantes a los que se podrá acceder a través de este servicio:

ISSN: 1135-5948 © 2009 Sociedad Española para el Procesamiento del Lenguaje Natural

- *Tesouro Informatizado da Lingua Galega* (http://www.ti.usc.es/TILG). Contiene casi todas las obras escritas en gallego entre 1612 y 1980, y una amplia representación de las publicadas hasta 2009. Excede los 20 millones de palabras, de las que en la actualidad están lematizadas y etiquetadas morfosintácticamente 12 millones (todas las palabras léxicas y parte de las gramaticales).
- *Tesouro Medieval Informatizado da Lingua Galega* (http://ilg.usc.es/tmilg). Contiene la totalidad de las obras medievales no notariales y un 80 % de las notariales. Está en proceso de anotación y excede los 9 millones de palabras.
- *Corpus Lingüístico da Universidade de Vigo (Corpus CLUVI)* (http://sli.uvigo.es/CLUVI). Conjunto de corpus paralelos de traducciones al/del gallego, de los ámbitos jurídico, informático, económico, literario, social y científico, alineado a nivel de oración. Totaliza 22 millones de palabras.
- *Corpus Técnico do Galego* (http://sli.uvigo.es/CTG). Colección de corpus especializados del gallego contemporáneo con textos publicados en los campos del derecho, de la informática, de la economía, de las ciencias ambientales, de las ciencias sociales y de la medicina que totalizan más de 13 millones de palabras.
- *Corpus Técnico Anotado do Galego* (http://sli.uvigo.es/CTAG). Versión categorizada y lematizada del Corpus Técnico do Galego. La anotación está en curso, pero ya se puede consultar en la web una sección del CTAG de más de 2 millones de palabras, correspondientes a textos de las ciencias ambientales.
- *Dicionario de Dicionarios* (CD-ROM, 3ª ed., Fundación Barrié de la Maza, 2003). Corpus lexicográfico del gallego moderno formado por 25 obras lexicográficas de los siglos XIX y XX. Contiene 345.742 entradas acumulativas, equivalentes a 136.164 lemas diferentes.
- *Dicionario de Dicionarios do Galego Medieval* (CD-ROM, 1ª ed., Universidade de Santiago, 2006). Contiene 13 obras lexicográficas del período medieval, totalizando 53.564 lemas.
- *Dicionario CLUVI Inglés-Galego* (http://sli.uvigo.es/dicionario). Diccionario elaborado a partir de los datos del Corpus CLUVI, con 20.000 entradas, 30.000 traducciones y 60.000 ejemplos documentados en el corpus.
- *Termoteca* (http://sli.uvigo.es/termoteca). Banco de conocimientos terminológicos basado en los datos de los corpus CLUVI y CTG, con información sobre más de 10.000 términos en los ámbitos de la economía, la ecología, la sociología, la medicina y el derecho.
- *Neoteca* (http://sli.uvigo.es/NEO). Banco de neologismos del gallego, con más de 10.000 registros documentados en el Corpus Neo de prensa gallega.
- *Dicionario Aquén de Toponimia Galega* (http://sli.uvigo.es/toponimia). Toponimia de los 315 ayuntamientos, 3.794 parroquias y 37.297 lugares de Galicia.
- *Inventario Toponímico da Galicia Medieval* (http://ilg.usc.es/ITGM). Banco de datos en curso con información lingüística georreferenciada extraída de un corpus toponomástico medieval.

3. Estado actual del proyecto

En la actualidad, el servidor RILG ofrece un acceso integrado a la consulta de la totalidad de los datos para el gallego recogidos en los corpus CLUVI, CTG y CTAG, y en los bancos de datos léxicos del Dicionario de Dicionarios (DdD), del Dicionario de Dicionarios do Galego Medieval (DdDGM), del Dicionario CLUVI Inglés-Galego, de la Termoteca, de la Neoteca y del Dicionario Aquén de Toponimia. En el caso de los dos corpus de diccionarios (el DdD y el DdDGM), distribuidos originalmente sólo en versión CD-ROM, el proyecto RILG sirvió también para ponerlos a disposición del público a través de la web. En ambos casos, se elaboró también además una interface web específica de consulta a cada una de estas dos obras. El servidor RILG ofrece por ahora un acceso sólo parcial a los datos del corpus del Tesouro Informatizado da Lingua Galega (TILG). En estos momentos, se está procediendo a una labor de conversión del TILG a XML, y de lematización y etiquetación morfosintáctica de los aproximadamente 8 millones de palabras del corpus pendientes de análisis.