Procesamiento del lenguaje Natural, nº 41 (2008), pp. 303-304 recibido 7-05-2008; aceptado 16-06-2008

TMILG (Tesouro Medieval Informatizado da Lingua Galega)

TMILG (Medieval Galician Computational Treasure)

António de Carlos Moura Barros
Angel López López
José Ramom Pichel Campos
imaxin|software
Rua Salgueiriños de abaixo N11 Local 6
Santiago de Compostela, Galiza
antonmoura@imaxin.com
angel@imaxin.com
jramompichel@imaxin.com

Resumen: El "Tesouro Medieval Informatizado da Lingua Galega" es un proyecto de investigación realizado en el Instituto da Lingua Galega (ILG) (a cargo de Xavier Varela y en convenio con la DXPL > SXPL de la Xunta de Galicia) que es visible en la Internet a través del corpus TMILG (http://ilg.usc.es/tmilg). En número, los documentos colectados son más de 12.500. El arco cronológico va del siglo XIII a principios del XVI. **imaxin**|software fue la encargada de realizar tanto los motores de indexación del corpus como el motor de búsquedas. Este recurso permite búsquedas variadas en la documentación gallega medieval: por fechas, géneros, tipología textual, por variantes de una misma palabra y por concordancias, además de por patrones y expresiones regulares. No tiene equivalente en ninguna de las lenguas románicas. Las obras que ofrece son muy variadas, y van desde la lírica profana o religiosa (Lírica trobadoresca gallego-portuguesa, Cantigas de Santa María) incluso la prosa técnica (Arte de Trovar, Tratado de Albeitaría), pasando por la prosa literaria (Crónica Troiana, Historia Troiana, Livro de Tristán), la prosa histórica (Crónica General y Crónica de Castilla, General Historia), la prosa religiosa (Miragres de Santiago, Crónica de Santa María de Iria) y la prosa jurídica (Flores de Derecho, fragmentos de la Partidas, Ordenamiento de Alcalá de Henares...). En lugar preferente está la prosa notarial con copiosas colecciones religiosas y civiles, entre las que destacan especialmente las monásticas.
Palabras clave: TMILG, Galego-portugués, Portugués, Galego, corpus, imaxin|software, ILG.

Abstract: The "Medieval Galician Computational Treasure" is a research project developed in the ILG (Institute of Galician Language) (coordinated by Xavier Varela and in agreement with the DXPL > SXPL -Linguistic Policy General Secretariat- of the Galician Government) and is accessible through the TMILG corpus (http://ilg.usc.es/tmilg). In total there are more than 12500 documents collected that date from the 13th century to the beginning of the 16th century. **imaxin**|software was in charge of developing both the corpus indexing and search engines. This feature allows the user to carry out customized searches within the Medieval Galician documents based on dates, genres, text typology, variants of the same word, agreement, regular expressions... There is no equivalent to this in any Romance Language. The corpus includes varied types of works: sacred or profane lyric poetry, technical prose, literary prose, historical prose, sacred prose and legal prose. One outstanding genre is the notarial prose, including substantial sacred and civil collections, monastic prose being the most prominent.
Keywords: TMILG, Galician-Portuguese, Portuguese, Galician, corpus, imaxin|software, ILG.

ISSN: 1135-5948

1 Codificación de los textos

La codificación de los textos demandó una atención muy especial, por las peculiaridades tanto de la scripta medieval gallego-portuguesa como de los hábitos modernos de edición. La especificidad de las grafías medievales los obligó a arbitrar soluciones poco convencionales en el asentamiento de los textos y en su tratamiento informático. La nasalidad sobre vocal o consonante (c, g, q, m) es el rasgo más característico. Otros elementos específicos son los tipos de (alto, de doble corva y corva y sigmática, el visigótico y el signo tironiano).

2 Indexadores

Se han desarrollado motores de indexación específicos para la migración de los documentos etiquetados en XML a formatos gestionables por sistemas de bases de datos como MySql. Por ahora no están lematizadas las palabras presentes en el corpus, cuestión que merece una investigación profunda, debido a la heterografía de todo corpus especialmente los medievales. Esta lematización y posterior agrupación de lemas facilitaría muchos trabajos posteriores diacrónicos del idioma gallego-portugués en la versión de Galicia.

3 Consultas

El corpus es de acceso libre, previo registro como usuario. El sistema de consulta permite buscar una o varias palabras, hacer buscas booleanas, utilizar comodines, un módulo estadístico para ver la frecuencia de aparición de palabras por tipología de documentos y por siglo, además de poder refinar las búsquedas haciendo restricciones cronológicas, por género, por subgénero o por obra. Estas consultas se realizan sobre bajo herramientas de gestión de bases de datos de software libre MySQL y también se ha desarrollado toda la programación en PHP 5.0.

4 Visualización de las consultas

El resultado de las consultas tiene cuatro formatos de salida “Listaxe”, “Tipoloxía textual”, “Formas” y “Concordancias”. En el primer formato de salida se mostrarán el resultado de las consultas en sus diferentes formas. En nuestro caso si buscamos la palabra “Galiza” aparece con dos formas “galiza” y “Galiza”. En el caso de elegir la “Tipoloxía textual” podremos ver la frecuencia de aparición de la palabra en función de la tipología textual y el siglo de aparición de esa palabra. Si pulsamos en cualquiera de estas dos frecuencias visualizaremos la palabra en formato KWIC (Key Word in Context), donde podremos ver la palabra en su contexto de la izquierda y el contexto de la derecha, además del Año, la obra, etc. Existe otro formato de visualización de las consultas llamado “Formas”, en el cual podremos ver la frecuencia por siglos de cada una de las formas superficiales que en nuestro caso son “galiza” y “Galiza”. Por último “Concordancias” que muestra directamente las palabras en sus diferentes formas en su contexto.

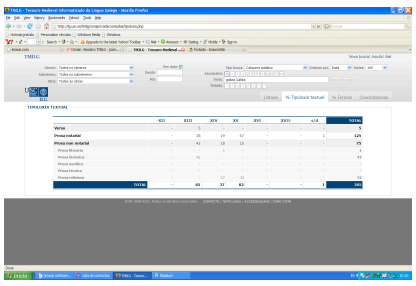

Figura 1: Ejemplo de visualización de los resultados.

5 Trabajos similares y futuros

imaxin|software ha desarrollado para el Instituto da Lingua Galega el TILG (Tesouro Informatizado da Lingua galega), un corpus contemporáneo gallego antecesor del medieval, Dicionario de dicionarios y Dicionario de dicionarios medieval. El trabajo futuro se centrará en crear el gran “Tesouro da lingua galega” que abarque desde la época medieval hasta la actualidad, lo cual facilitará el estudio diacrónico de la lengua gallega la investigación en lexicografía en el ámbito global de la lengua galego-luso-brasileira.