I Jornadas Filológicas Internacionales-UR

Corpus, bases de datos y diccionarios para la investigación de la Lengua y la Literatura

LIBRO DE RESÚMENES

30 septiembre - 2 octubre 2015

www.unirioja.es/fiur/

jornadas.fiur@gmail.com

Universidad de La Rioja - Cilengua

Título

Libro de resúmenes de las I Jornadas Filológicas Internacionales-UR: *Corpus, bases de datos y diccionarios para la investigación de la Lengua y la Literatura*.

Editoras

Sara Gómez Seibane

Rebeca Lázaro Niso

Zaida Vila Carneiro

Logroño, septiembre de 2015

ISBN **978-84-608-1721-5**

CORPUS ORAL INFORMATIZADO DA LINGUA GALEGA

Lucía Barreiro

Xosé Manuel Dopazo

Naír García

Xosé Luís Regueira

Instituto da Lingua Galega – Universidade de Santiago de Compostela

Carmen García Mateo

Rocío Varela

Marta Martínez

Roberto Seara

Grupo de Tecnoloxías Multimedia – Universidade de Vigo

El proyecto CORILGA (Corpus Oral Informatizado da Lingua Galega) se concibe como un corpus de la lengua gallega oral pensado para consultar en línea. En él se integran transcripciones ortográficas y fonéticas alineadas con el correspondiente archivo de audio. Además, incluye anotaciones léxicas y gramaticales y la clasificación por tipo de texto y tema.

El corpus está constituido por grabaciones de diferentes variedades de lengua gallega oral, conversaciones informales, entrevistas dirigidas, discursos formales y lecturas literarias, incluyendo lengua rural y urbana, variedades estándar y no estándar, de hablantes de diferentes generaciones y niveles sociales. Esta variedad de textos, junto con la posibilidad de filtrado de los mismos que ofrece el buscador, da al proyecto un carácter innovador con respecto a otros corpus del mismo ámbito. La integración de toda esta variedad textual es, en parte, posible gracias a la incorporación de herramientas automáticas de ayuda a la transcripción, alineación, análisis morfológico y búsqueda.

CORILGA permitirá a los investigadores disponer de un corpus amplio que permitirá estudiar la variación lingüística del gallego actual, tanto diatópica como diafásica y diastrática. Al disponer de grabaciones de diferentes generaciones desde mediados de los años 1960, este corpus permitirá el estudio del cambio lingüístico en gallego tanto en tiempo aparente como en tiempo real.

Los archivos de audio se transcriben y anotan con el programa ELAN (Max Planck Institute for Psycholinguistics, http://tla.mpi.nl/tools/tla-tools/elan/) (Wittenburg *et al.* 2006), y se gestionan a través de una base de datos. Para su máximo aprovechamiento, la interfaz de usuario del CORILGA cuenta con una estructura de filtrado en la que podemos acotar los siguientes parámetros:

- Por grabación: año, tipo de texto y tema.
- Por hablantes: sexo, tramo de edad, nivel de estudios, lugar de nacimiento, lugar de residencia, lengua inicial y lengua de grabación.

CORILGA pretende ser principalmente un corpus recurrente para la investigación filológica (lingüística basada en corpus, estudio de la variación y cambio lingüístico, estudios lingüísticos interdisciplinares y tecnologías del habla), pero también podría resultar de utilidad para otros ámbitos (educativo y antropológico), por lo que estará abierto al público de manera gratuita. Está prevista la apertura de una parte del corpus al público a finales de 2015.

Este proyecto ha sido desarrollado en colaboración entre investigadores del Instituto da Lingua Galega, de la Universidade de Santiago de Compostela, y el Grupo de Tecnoloxías Multimedia, de la Universidade de Vigo, en el marco de la red TecAnDaLi desde el año 2012.

BIBLIOGRAFIA

Wittenburg, P. *et al.* (2006): "ELAN: a Professional Framework for Multimodality Research", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006),* 1556-1559.