Gallæcia

Estudos de lingüística portuguesa e galega

CURSOS E CONGRESOS DA
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
N.º 242

INSTITUTO DA LINGUA GALEGA

Gallæcia
Estudos de lingüística portuguesa e galega

Edición ao coidado de
MARTA NEGRO ROMERO
ROSARIO ÁLVAREZ
EDUARDO MOSCOSO MATO

2017
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Este libro publícase coa axuda financeira da Fundação Calouste Gulbenkian (Programa Gulbenkian de Língua e Cultura Portuguesas 2015) e da Secretaría Xeral de Universidades (Xunta de Galicia - Fondo Europeo de Desenvolvemento Rexional) ao grupo de investigación *Filoloxía e Lingüística galega* (USC).

Deseño de cuberta
Servizo de Publicacións da Universidade de Santiago de Compostela

Maquetación
Raquel Vila-Amado

Edita
Servizo de Publicacións
Campus Vida
15782 Santiago de Compostela
usc.es/publicacions

DOI http://dx.doi.org/10.15304/cc.2017.1080

A Antón Santamarina e Ramón Lorenzo, que nos agasallan, día a día, coa súa presenza, o seu compromiso, o seu maxisterio e a súa amizade. Por moitos anos.

ÍNDICE

Gallæcia. Estudos de lingüística portuguesa e galega
Universidade de Santiago de Compostela, 2017.
DOI http://dx.doi.org/10.15304/cc.2017.1080.61, pp. 1037-1048

Recursos integrados da lingua galega para a investigación lingüística

Xavier Gómez Guinovart
Grupo TALG - Universidade de Vigo

Integrating Galician resources for linguistic research

Resumo
Neste artigo revisaremos as principais características dos recursos textuais e léxicos máis importantes incluídos na plataforma *RILG* (*Recursos Integrados da Lingua Galega*), que ten como obxectivo a integración, explotación conxunta e difusión dos recursos textuais e léxicos de tecnoloxía lingüística da lingua galega xerados en distintos proxectos realizados polo Instituto da Lingua Galega da Universidade de Santiago de Compostela e polo Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo.

Palabras-chave
Tecnoloxías lingüísticas, recursos léxicos, galego, corpus, dicionarios

Abstract
This paper presents the main features of the most important textual and lexical resources included in the RILG platform (*Recursos Integrados da Lingua Galega*), the objective of which is the integration, collection employment and dissemination of the textual and lexical 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 (Galician Language Tecnologies and Applications) of the University of Vigo.

Keywords
Language technologies, lexical resources, Galician, corpora, dictionaries

Sumario

Contents

1. Introdución

A plataforma *RILG* (*Recursos Integrados da Lingua Galega*) é o resultado dun proxecto de investigación coordinado entre o Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo e o Instituto da Lingua Galega (ILG) da Universidade de Santiago de Compostela, que obtivo financiamento en convocatorias competitivas dos *Planes Nacionales de I+D+I* do *Ministerio de Educación y Ciencia* do Goberno de España (2006-2009) e da Consellaría de Innovación e Industria da Xunta de Galiza (2008-2011). Os responsábeis da dirección do proxecto foron Xavier Gómez Guinovart (investigador principal do proxecto coordinado e do subproxecto da Universidade de Vigo) e Antón Santamarina (investigador principal do subproxecto da Universidade de Santiago de Compostela).

A finalidade da plataforma é a integración, explotación conxunta e difusión dos recursos de tecnoloxía lingüística do galego xerados en distintos proxectos realizados polo ILG e polo Grupo TALG. De libre acceso en Internet no enderezo http://sli.uvigo.gal/RILG/, ofrece un portal web de servizos lingüísticos do galego desde o que se pode acceder dun modo conxunto aos bancos de datos textuais e léxicos desenvolvidos por estes dous grupos de investigación, permitindo realizar pescudas nun abano de corpus textuais de referencia, que totalizan uns 75 millóns de palabras, e nunha variedade de obras lexicográficas do galego, que reúnen máis de 500.000 entradas. Os bancos de datos textuais e léxicos integrados no *RILG* abranguen un período temporal que arrinca nas orixes do idioma e remata no período contemporáneo, e variedades lingüísticas da práctica totalidade dos ámbitos e rexistros, alén de correspondencias lingüísticas con outros idiomas do noso contorno xeográfico e cultural. Desde a súa posta en marcha en 2006, a plataforma segue sendo obxecto de ampliación e mellora mediante a incorporación de novos recursos.

Neste artigo revisaremos as principais características dos recursos textuais e léxicos máis importantes incluídos na plataforma *RILG*, co obxecto de ofrecer unha visión panorámica das súas posibilidades como ferramenta de consulta lingüística e filolóxica.

2. Corpus textuais

2.1. *Tesouro Informatizado da Lingua Galega* (*TILG*)

Este corpus textual, desenvolvido no ILG baixo a dirección do profesor Antón Santamarina cunha orientación lexicográfica (Santamarina 2008), inclúe practicamente todas as obras publicadas en galego entre 1612 e 1980, e una representación moi ampla das publicadas desde 1980 ata a actualidade. Historicamente, o *TILG* desenvolveuse en

tres etapas. A primeira edición (2004) contén a transcrición de 1.464 textos publicados ata o ano 2002, totalizando uns 20 millóns de palabras, das que máis de 12 millóns (todas as palabras léxicas e parte das gramaticais) están lematizadas e anotadas desde o punto de vista morfosintáctico. A edición posterior do ano 2011, realizada en colaboración co Grupo TALG, constitúe unha edición revisada e ampliada deste corpus, tanto no número de textos, coma no nivel de anotación lingüística (Gómez Guinovart / Santamarina 2009). Nesta segunda edición ampliada, o número de textos ascende a 1.897, incluíndo textos publicados ata o ano 2010 e totalizando máis de 25 millóns de palabras completamente lematizadas e anotadas gramaticalmente. Na súa versión actual, accesíbel desde 2014 na súa propia web (http://ilg.usc.es/TILG/) e tamén a través do *RILG*, inclúe 1.958 obras de 704 autores e autoras publicadas entre 1612 e 2013, acadando os 26 millóns de palabras correspondentes a 95.409 lemas diferentes.

2.2. *Tesouro Medieval Informatizado da Lingua Galega* (*TMILG*) e *Corpus Xelmírez*

O *TMILG* é un corpus diacrónico do galego, de máis de nove millóns de palabras, elaborado no ILG baixo a dirección do profesor Xavier Varela (Varela Barreiro 2004). Este corpus medieval do galego, que contén a totalidade das obras non notariais publicadas da Galicia medieval (literarias, históricas, relixiosas, xurídicas e técnicas) e o 80% das obras notariais publicadas, está dispoñíbel na web (http://ilg.usc.es/tmilg/) para a libre consulta, previa alta no sistema.

O Corpus Xelmírez, accesíbel quer de xeito independente (http://sli.uvigo.gal/xelmirez/) quer a través do *RILG*, inclúe os textos do *TMILG* xunto aos correspondentes ao *Tesouro Medieval Informatizado da Lingua Latina* (Galicia) (*TMILL-G*) e ao *Tesouro Medieval Informatizado da Lingua Castelá* (Galicia) (*TMILC-G*), permitindo a recuperación de información de textos da Galicia medieval redactados en calquera destas tres linguas. Este corpus lingüístico constitúe o alicerce do *Inventario Toponímico da Galicia Medieval* (Martínez Lema *et al*. 2010), un recurso toponomástico dispoñíbel na web (http://ilg.usc.es/itgm/) baseado nos datos do *Corpus Xelmírez*.

2.3. *Corpus Técnico do Galego* (*CTG*)

O *CTG* é un corpus textual de orientación terminolóxica que recolle documentos publicados pertencentes a rexistros especializados do galego contemporáneo. Contén textos publicados nos campos do dereito, da informática, da economía, das ciencias ambientais, das ciencias sociais e da medicina, cunha extensión total de 18 millóns de palabras (Gómez Guinovart 2008). Trátase dun corpus desenvolvido no Grupo

TALG e pode consultarse libremente na web (http://sli.uvigo.gal/CTG/). O *CTG* está lematizado, etiquetado con categorías gramaticais e desambiguado con anotacións sobre semántica léxica baseadas nos datos do Galnet (*vid*. infra) e enlazado con este recurso léxico a nivel de lema.

2.4. *Corpus Lingüístico da Universidade de Vigo* (*CLUVI*)

O *CLUVI* é un corpus de traducións do galego, directas e inversas, en combinación con diversas linguas, que abrangue un conxunto textual de máis de 23 millóns de palabras, formado polos textos orixinais e mais as súas traducións. Desde un punto de vista temático, os textos recompilados pertencen aos ámbitos xurídico, informático, económico, literario, social e científico, en tanto que as linguas de tradución incluídas en relación de tradución co galego son o español, o inglés, o francés, o alemán, o catalán, o portugués e o euskera. Este corpus paralelo aliñado a nivel de oración está dispoñíbel para consulta na web desde setembro de 2003 (http://sli.uvigo.gal/CLUVI/), constituíndo o alicerce empírico dun variado conxunto de traballos académicos de investigación nos campos da estilística da tradución, da didáctica do ensino de idiomas, da lingüística comparada, da terminoloxía e da lexicografía plurilingüe (Gómez Guinovart 2008). A sección de traducións xurídico-administrativas do corpus paralelo español-galego, duns 6 millóns de palabras, está dispoñíbel tamén para descarga (http://hdl.handle.net/10230/20051) a través da plataforma europea Meta-Share (http://metashare.elda.org/).

2.5. *Corpus Paralelo SensoGal*

SensoGal é un corpus paralelo inglés-galego en desenvolvemento anotado semanticamente con referencia a *Galnet* (*vid*. infra) e aliñado a nivel de frase e de palabra co corpus *SemCor* da lingua inglesa. O *SemCor* (Miller *et al*. 1993) é un subconxunto do *English Brown Corpus* de 360.000 palabras que constitúe na actualidade o corpus máis extenso con anotación semántica sobre os sentidos das palabras. Os lemas de *SemCor* están desambiguados con referencia ao *WordNet* do inglés. Do total de 352 textos anotados no corpus *SemCor*, 186 conteñen todas as palabras léxicas etiquetadas con categoría gramatical, lema e sentido en *WordNet* (192.639 nomes, verbos, adxectivos e adverbios), mentres que nos restantes 166 textos só están anotados os verbos. O obxectivo do corpus *SensoGal*, que se pode consultar tamén de modo independente (http://sli.uvigo.gal/SensoGal/), é completar o aliñamento entre os 352 textos en inglés do *SemCor* que conteñen todas as palabras léxicas anotadas e as súas versións traducidas ao galego igualmente anotadas.

3. Repertorios léxicos

3.1. *Dicionario de dicionarios*

O *Dicionario de dicionarios* é un exemplo ilustre da confluencia harmoniosa de tradición e modernidade na lexicografía galega. Este dicionario é, en realidade, unha colección de obras lexicográficas dos sécs. XIX e XX, recompiladas e transcritas baixo a coordinación do profesor Antón Santamarina no ILG. Todos os textos foron anotados para facilitar as consultas por lemas, por sinónimos, por voces en castelán, por localidades ás que se adscriben, pola súa presenza en refráns ou en poemas citados etc. Publicado orixinalmente en formato CD-ROM, na súa terceira edición (Santamarina 2003), recollía 345.742 entradas (equivalentes a 136.164 lemas diferentes) correspondentes a 25 obras lexicográficas, incluídas todas as obras históricas da lexicografía galega (Rodríguez, Carré, Eladio, Real Academia...). A colaboración entre o ILG e o Grupo TALG fixo posíbel a publicación na web deste dicionario a partir dunha versión ampliada da súa edición en CD-ROM. Como resultado, a primeira edición web, con 392.768 entradas documentadas en 32 obras, pode ser consultada libremente desde 2006 como un recurso único (http://sli.uvigo.gal/DdD/) ou integrado no *RILG*, facendo que o acceso a este valioso material lingüístico sexa moito máis doado e directo do que era desde disco. O *Dicionario de dicionarios* de Antón Santamarina representa unha contribución fundamental á historia da lexicografía e á cultura galega, e ten tamén unha utilidade práctica innegábel como dicionario da lingua, aínda non superado en extensión como conxunto por ningún outro.

3.2. *Dicionario de sinónimos do galego*

O *Dicionario de sinónimos do galego* publicouse na páxina web do Grupo TALG en 2013, tratándose do primeiro e único dicionario electrónico do galego dentro desta tipoloxía de repertorios léxicos. Tamén é o primeiro publicado no formato de libro electrónico (Gómez Clemente *et al*. 2015) e o primeiro que se pode consultar no móbil mediante unha aplicación, que se pode descargar desde 2014 tanto para dispositivos móbiles con Android[1] coma para os dispositivos de Apple con iOS[2]. Este dicionario é o resultado da revisión, actualización, ampliación e conversión a formato dixital estruturado (Gómez Guinovart / Simões 2013; Gómez Guinovart 2014) dun excelente dicionario de sinónimos tradicional do galego publicado en papel e xa descatalogado, concretamente, do publicado por Galaxia en 1997 baixo a coordinación de Camiño Noia, Xosé María Gómez Clemente e Pedro Benavente, e que contou coa

1 https://play.google.com/store/apps/details?id=net.ayco.sinonimosgal
2 https://itunes.apple.com/us/app/sinonimos-do-galego/id940045971?l=es&ls=1&mt=8

participación de Gonzalo Constela, Xosé Henrique Costas e Valentín Arias na súa redacción (Noia *et al*. 1997). Na súa versión electrónica actual contén máis de 200.000 sinónimos agrupados nunhas 30.000 entradas, e pode ser consultado na súa propia web (http://sli.uvigo.gal/sinonimos/) ou a través da interface de consulta do *RILG*.

3.3. *Galnet*

WordNet é unha base de datos léxica, orixinalmente concibida para o inglés, configurada como unha rede semántica onde os nós son os conceptos (representados como grupos de sinónimos) e as ligazóns entre os nós son as relacións semánticas entre os conceptos léxicos. Os nós da rede están formados por nomes, verbos, adxectivos e adverbios agrupados pola súa sinonimia. Deste xeito, cada nó desta rede léxico-semántica representa un concepto lexicalizado único e agrupa o conxunto de variantes sinonímicas dese concepto. No modelo de representación do léxico de WordNet, todos os nós están conectados por relacións semánticas. No caso dos substantivos, algunhas das relacións máis frecuentes representadas no *WordNet* son as de hiperonimia/hiponimia e as de holonimia/meronimia; no caso dos adxectivos, as de antonimia e as de cuasisinonimia; no caso dos adverbios, as de antonimia e as derivativas; e no caso dos verbos, as de implicación, hiperonimia/hiponimia, causatividade e oposición.

Galnet (Gómez Clemente *et al*. 2013; Gómez Guinovart 2014; Solla Portela / Gómez Guinovart 2015) é a versión galega do *WordNet* que está a ser elaborada polo Grupo TALG no marco de desenvolvemento do Multilingual Central Repository (González Agirre / Rigau 2013), unha plataforma que abrangue na actualidade os léxicos *WordNet* de cinco linguas (inglés, español, catalán, vasco e galego) enlazados interlingüisticamente e categorizados por diversas ontoloxías. Na versión actual, en constante actualización, Galnet inclúe máis de 45.000 palabras agrupadas en máis de 30.000 conceptos, e está dispoñíbel na web para consulta (http://sli.uvigo.gal/galnet/) na súa última versión. A súa descarga pode realizarse de modo directo (http://hdl.handle.net/10230/22921), a través da plataforma europea Meta-Share (http://metashare.elda.org/) ou a través do MCR (http://adimen.si.ehu.es/web/MCR/).

3.4. *DBpedia do galego*

A *DBpedia* é un proxecto internacional de creación dunha versión estruturada dos contidos da Wikipedia e da súa libre dispoñibilización en Internet entrelazada con moitas outras bases de coñecementos que constitúen a web semántica (Auer *et al*. 2007). Permite realizar consultas complexas a partir do conxunto de datos derivados da Wikipedia e permite enlazar outros conxuntos de datos que hai na web (como os

datos sobre libros dixitais ofrecidos polo Project Gutenberg[3], os datos estatísticos sobre Europa dispoñibilizados por Eurostat[4] ou os datos do censo de Estados Unidos[5]) cos datos da Wikipedia, seguindo as especificacións para os datos enlazados abertos (Linked Open Data)[6] establecidas polo W3C (World Wide Web Consortium). A *DBpedia do galego*, desenvolvida e mantida polo Grupo TALG, contén 11 millóns de tripletes semánticos tirados a partir de toda a información contida na Galipedia (http://gl.wikipedia.org) e está aloxada no subdominio oficial de dbpedia.org correspondente á lingua galega (http://gl.dbpedia.org). Así mesmo, está accesíbel como un recurso léxico máis, na plataforma *RILG*, neste caso como un recurso enciclopédico. Os seus contidos poden consultarse e visualizarse tamén mediante as aplicacións Lodview (http://sli.uvigo.gal/dbpedia/lodview/) e LodLive (http://sli.uvigo.gal/dbpedia/lodlive/), ou a través do punto de acceso SPARQL aos datos estruturados (http://gl.dbpedia.org/sparql/).

3.5. *Dicionario de Dicionarios do Galego Medieval*

A mesma colaboración interuniversitaria entre Vigo e Compostela que permitiu levar o *Dicionario de dicionarios* do CD-ROM á web, facilitou tamén a edición web do *Dicionario de dicionarios do galego medieval*, unha obra complementaria á anterior e inspirada nela, que recompila as entradas de 13 obras lexicográficas do período medieval, cun total de 53.564 lemas. O repertorio, que foi compilado, transcrito e anotado no ILG baixo a dirección de Ernesto González Seoane, foi publicado orixinalmente só en CD-ROM (González Seoane / Álvarez de la Granja / Boullón Agrelo 2006). Esta versión foi actualizada e adaptada posteriormente á web para a súa libre consulta como recurso independente (http://sli.uvigo.gal/DDGM/) ou integrada no *RILG*, acadando nesta versión ampliada un total de 62.293 lemas documentados en 22 obras.

3.6. *Dicionario CLUVI inglés-galego*

O *Dicionario CLUVI inglés-galego* é un dicionario bilingüe baseado na colección de textos ingleses traducidos ao galego que forma parte do Corpus CLUVI e constitúe, ao noso entender, o primeiro dicionario baseado en corpus da lexicografía galega. Todas as palabras inglesas que aparecen nas súas entradas están documentadas nos textos en inglés traducidos ao galego recompilados no corpus paralelo CLUVI. Alén

[3] https://www.gutenberg.org
[4] http://ec.europa.eu/eurostat/
[5] http://www.census.gov
[6] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

diso, todas as traducións galegas recollidas no dicionario para esas palabras son traducións reais identificadas nas versións galegas dos textos ingleses do corpus. Finalmente, para cada tradución identificada, o dicionario fornece un exemplo real do seu uso tal como está documentado no corpus.

O dicionario está accesíbel na web do Grupo TALG para libre consulta desde 2005. A súa segunda edición electrónica, publicada en setembro de 2008, consta de 20.000 entradas con 30.000 traducións e 60.000 exemplos, ao tempo que amplía os datos lexicográficos contidos nos artigos da primeira edición con información sobre americanismos e variantes ortográficas e con notas de interese gramatical, tradutolóxico e normativo. O obxectivo destes engadidos é que a ferramenta resultante poida ser realmente útil tanto na docencia do inglés como na tradución inglés-galego. Aínda que as entradas desta obra están redactadas só na dirección de tradución inglés-galego, o sistema de busca implementado permite recuperar tamén as entradas a partir das súas traducións ao galego, converténdose así tamén nun dicionario galego-inglés.

O *Dicionario moderno inglés-galego*, publicado en versión impresa no 2012 (Gómez Guinovart *et al*. 2012), constitúe unha edición revisada e adaptada ao formato papel desta segunda edición do *Dicionario CLUVI* (Álvarez Lugrís / Gómez Guinovart 2014). O acceso ao dicionario na web pode facerse consultando directamente o recurso (http://sli.uvigo.gal/dicionario) ou a través da plataforma *RILG*. Tamén resulta posíbel descargalo de modo directo (http://hdl.handle.net/10230/20053) ou a través da plataforma europea Meta-Share (http://metashare.elda.org/).

3.7. *Termoteca*

A *Termoteca* é un banco de datos terminolóxico para o galego baseado nos textos de especialidade monolingües e paralelos recompilados, respectivamente, no *Corpus Técnico do Galego* (*CTG*) e no Corpus CLUVI. A información terminolóxica extraída dos corpus inclúe, en primeiro lugar, os propios termos, xunto cos seus contextos, variantes formais intralingüísticas e interlingüísticas coas súas frecuencias de uso; en segundo lugar, a súa definición ou definicións, cando se poden documentar nos corpus; e, por último, as relacións semánticas que estabelecen con outros termos do corpus, cando aparecen explicitamente codificadas nos textos. As técnicas utilizadas para tirar toda esta información son de tipo lingüístico-computacional e estatístico, e os seus resultados son sempre revisados e complementados por especialistas (Gómez Guinovart 2012).

A base de datos terminolóxica conta, na actualidade, cuns 8.000 rexistros con información sobre 16.120 termos documentados no *CLUVI* ou no *CTG* pertencentes aos ámbitos do dereito (termos en galego e español en rexistros bilingües e monolingües da *Termoteca*), da socioloxía (termos en galego, español, francés e inglés en rexistros tetralingües e monolingües), da economía (termos en galego e español

en rexistros monolingües e bilingües), da ecoloxía e ciencias ambientais (termos en galego en rexistros monolingües), da medicina (termos en galego en rexistros monolingües) e da informática (termos en galego e inglés en rexistros monolingües e bilingües), a partir dos datos das seccións especializadas correspondentes destes dous corpus. Cada rexistro inclúe toda a información relativa a un concepto especializado, expresado cun termo galego documentado nos corpus, e do que se poden recoller tamén no mesmo rexistro as súas variantes documentadas, tanto intralingüísticas (termos sinónimos, variantes ortográficas ou variantes dialectais) como interlingüísticas (traducións ou, con maior propiedade, equivalencias).

A información especificada para cada variante, incluída a variante común ou non marcada, abrangue o lema do termo, a súa categoría gramatical como conxunto, a análise morfosintáctica dos seus compoñentes, a súa definición, a súa frecuencia de aparición e un contexto de uso documentado no corpus. Todos os rexistros están catalogados, ademais, segundo o seu campo temático, en referencia a unha árbore conceptual xerarquizada da materia, e poden incluír información sobre as relacións semánticas (antonimia, hiperonimia, holonimia etc.) que gardan con outros rexistros do banco de datos.

A *Termoteca* é un recurso de libre consulta na web (http://sli.uvigo.gal/termoteca/) e no *RILG*, e está dispoñíbel tamén para descarga (http://hdl.handle.net/10230/17104) a través da plataforma europea Meta-Share (http://metashare.elda.org/).

3.8. *Neoteca*

A *Neoteca* é un banco de datos sobre neoloxía do galego desenvolvido polo Observatorio de Neoloxía do Grupo TALG sobre o que se elaborou o seu dicionario de neoloxismos (López Fernández *et al*. 2005). Na versión actual, contén máis de 10.000 rexistros neolóxicos identificados e documentados nun corpus de prensa galega publicada desde 1997 (Gómez Clemente / Rodríguez Guerra 2003). Este banco de datos pódese consultar libremente na web como recurso independente (http://sli.uvigo.gal/NEO/) ou integrado no *RILG*.

3.9. *Aquén - Toponimia galega*

Aquén é unha ferramenta de divulgación e consulta desenvolvida no Grupo TALG que permite coñecer, localizar xeograficamente e visualizar cuantitativamente os topónimos oficiais dos 315 concellos, 3.794 parroquias e 37.297 lugares de Galiza, tal como están estabelecidos na lexislación vixente de acordo cos ditames da Comisión

de Toponimia. A base de datos do *Aquén* baséase, por tanto, no Nomenclátor oficial da Xunta de Galiza. As pescudas nesta ferramenta permiten identificar e documentar os topónimos galegos a partir do seu nome ou dunha parte del. Unha vez identificado o topónimo, amosará a súa adscrición territorial e ofrecerá a posibilidade de xeolocalizalo no Google Maps e de consultar a súa frecuencia na toponimia galega mediante unha visualización gráfica en forma de nube de datos. O *Aquén* pode consultarse na súa propia páxina web (http://sli.uvigo.gal/toponimia) ou a través do *RILG*.

4. Conclusión

A integración dos recursos existentes nos centros de investigación é un obxectivo prioritario no campo das Humanidades, como en calquera campo científico. A integración nunha plataforma informática común dos recursos de tecnoloxía lingüística do galego xerados de xeito independente polo Instituto da Lingua Galega (ILG) da Universidade de Santiago de Compostela e polo Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, sen dúbida contribúe tanto ao avance da investigación e do coñecemento científico da lingua, como á divulgación e valorización do labor filolóxico realizado durante anos de traballo polos grupos de investigación destas dúas entidades. A implementación da plataforma *RILG* nun servidor web de acceso libre e de uso gratuíto garante esta vocación de transferencia social do coñecemento compartida por todos os participantes no proxecto.

Referencias bibliográficas

Álvarez Lugrís, Alberto / Xavier Gómez Guinovart (2014): "Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del Dicionario Moderno Inglés-Galego", en María José Domínguez Vázquez / Xavier Gómez Guinovart / Carlos Valcárcel Riveiro (eds.), *Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva*. Berlín / Boston: De Gruyter Mouton, 31-48.

Auer, Sören *et al*. (2007): "DBpedia: A Nucleus for a Web of Open Data", en Aberer *et al*. (eds.), *Proceedings of the 6th International Semantic Web Conference*. Berlín: Springer, 722-735.

Gómez Clemente, Xosé María / Alexandre Rodríguez Guerra (2003): *Neoloxía e lingua galega: teoría e práctica*. Vigo: Universidade de Vigo.

Gómez Clemente, Xosé María / Xavier Gómez Guinovart / Andrea González Pereira / Verónica Taboada Lorenzo (2013): "Sinonimia e rexistros na construción do WordNet do galego", *Estudos de lingüística galega*, 5, 27-42.

Gómez Clemente, Xosé María / Xavier Gómez Guinovart / Alberto Simões (2015): *Dicionario de sinónimos do galego*. Vigo: Xerais.

Gómez Guinovart, Xavier (2008): "A investigación en lexicografía e terminoloxía no Corpus Lingüístico da Universidade de Vigo (CLUVI) e no Corpus Técnico do Galego (CTG)", en Ernesto González Seoane / Antón Santamarina / Xavier Varela Barreiro (eds.), *A lexicografía galega moderna. Recursos e perspectivas*. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, 209-228.

Gómez Guinovart, Xavier (2012): "A Hybrid Corpus-Based Approach to Bilingual Terminology Extraction", en Isabel Moskowich-Spiegel Fandiño / Begoña Crespo (eds.), *Encoding the Past, Decoding The Future: Corpora in the 21st Century*. Newcastle upon Tyne: Cambridge Scholar Publishing, 147-175.

Gómez Guinovart, Xavier (2014): "Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do Galego", en Ana Gabriela Macedo *et al*. (eds.), *XV Colóquio de Outono. As humanidades e as ciências: disjunções e confluências*. Braga: CEHUM-Universidade do Minho, 331-358.

Gómez Guinovart, Xavier / Antón Santamarina (2009): "RILG: Recursos Integrados da Lingua Galega", *Procesamiento del Lenguaje Natural*, 43, 375-376.

Gómez Guinovart, Xavier / Alberto Álvarez Lugrís / Eva Díaz Rodríguez (2012): *Dicionario moderno inglés-galego*. Ames: 2.0 Editora.

Gómez Guinovart, Xavier / Alberto Simões (2013): "Retreading Dictionaries for the 21st Century", en José Paulo Leal / Ricardo Rocha / Alberto Simões (eds.), *2nd Symposium on Languages, Applications and Technologies*, vol. 29. Saarbrücken: Dagstuhl Publishing, 115-126.

González Agirre, Aitor / German Rigau (2013): "Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository". *Linguamática*, 5.1, 13-28.

González Seoane, Ernesto / María Álvarez de la Granja / Isabel Boullón Agrelo (eds.) (2006): *Dicionario de dicionarios do galego medieval*. Santiago de Compostela: Universidade de Santiago de Compostela (*Verba*, Anexo 57).

López Fernández, Susana *et al*. (2005): *Novas palabras galegas. Repertorio de creacións léxicas rexistradas na prensa e en Internet*. Vigo: Universidade de Vigo.

Martínez Lema, Paulo / Rocío Dourado Fernández / César Osorio Peláez (2010): "Un novo recurso para os estudos toponomásticos: o Inventario Toponímico da Galicia Medieval (ITGM)", en Xulio Sousa Fernández (ed.), *Toponimia e cartografí*a. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, 239-263.

Miller, George A. / Claudia Leacock / Randee Tengi / Ross T. Bunker (1993): "A Semantic Concordance", en *Proceedings of the Workshop on Human Language Technology*. Stroudsburg: Association for Computational Linguistics, 303-308.

Noia, Camiño / Xosé María Gómez Clemente / Pedro Benavente (coords.) (1997): *Diccionario de sinónimos da lingua galega*. Vigo: Galaxia.

Santamarina, Antón (ed.) (2003[3]): *Diccionario de diccionarios*. A Coruña: Fundación Barrié de la Maza.

Santamarina, Antón (2008): "Os dicionarios históricos. Trazos dun dicionario histórico galego e consideracións sobre a súa viabilidade", en Ernesto González Seoane / Antón Santamarina / Xavier Varela Barreiro (eds.), *A lexicografía galega moderna. Recursos e perspectivas*. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, 343-392.

Solla Portela, Miguel Anxo / Xavier Gómez Guinovart (2015): "Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas", *Revista Galega de Filoloxía*, 16, 169-201.

Varela Barreiro, Xavier (2004): "Un proxecto do ILG no abalo da gramática histórica da lingua galega", en Rosario Álvarez / Francisco Fernández Rei / Antón Santamarina (eds.), *A lingua galega: historia e actualidade*, 2. Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega, 649-684.