Integración de recursos lingüísticos na plataforma RILG

XAVIER GÓMEZ GUINOVART
Universidade de Vigo

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 plataforma RILG resultado do proxecto pode consultarse na rede no enderezo http://sli.uvigo.es/RILG/.

A finalidade da plataforma RILG (Recursos Integrados da Lingua Galega) é a integración, explotación conxunta e difusión dos recursos textuais e léxicos de tecnoloxía lingüística da lingua galega xerados en distintos proxectos realizados polo ILG e polo Grupo TALG. A plataforma RILG, de libre acceso en Internet, 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 do galego 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.

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. Por razóns históricas ligadas ao seu desenvolvemento, o TILG pode consultarse agora na web en dúas edicións, correspondentes aos anos de publicación de 2004 (http://www.ti.usc.es/TILG/) e de 2011 (http://sli.uvigo.es/TILG), e está en curso no ILG a elaboración da terceira edición ampliada do corpus. 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 TILG 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. Nesta segunda edición ampliada do TILG, accesíbel a través do RILG, 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.

2.2. Corpus Lingüístico da Universidade de Vigo (CLUVI)

O corpus CLUVI é un corpus de traducións do galego, directas e inversas, e en combinación con diversas linguas. O CLUVI 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 no CLUVI 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.es/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/).

2.3. Corpus Técnico do Galego (CTG) e Corpus Técnico Anotado do Galego (CTAG)

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, totalizando máis de 13 millóns de palabras (Gómez Guinovart 2008). Trátase dun corpus desenvolvido no Grupo TALG e pode consultarse libremente na web (http://sli.uvigo.es/CTG/). O CTAG é unha versión do CTG etiquetada gramaticalmente e lematizada (Gómez Guinovart e López Fernández 2009). Na web (http://sli.uvigo.es/CTAG/) pode consultarse unha sección do CTAG de máis de 2 millóns de palabras, correspondente a unha selección de textos do dominio das ciencias ambientais.

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

O TMILG é un corpus diacrónico do galego, de máis de vinte 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 a práctica totalidade 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.es/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 da Galicia medieval 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.

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 séculos 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, o *Dicionario de dicionarios*, 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 do *Dicionario de dicionarios*, con 392.768 entradas documentadas en 32 obras, pode ser consultada libremente desde 2006 como un recurso único (http://sli.uvigo.es/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 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 et al. 2006). e adaptado posteriormente á web para a súa libre consulta como recurso independente (http://sli.uvigo.es/DDGM/) ou integrado no RILG.

3.3. Galnet (WordNet)

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 ou adxectivos 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 léxico-semánticas máis frecuentes representadas no WordNet son as de hiperonimia/hiponimia e as de holonimia/meronimia; no caso dos adxectivos, as de antonimia; e no caso dos verbos, as de implicación, hiperonimia/hiponimia, causativa e oposición. Galnet (Gómez Clemente et al. 2013, Gómez Guinovart 2014) é a versión

galega do WordNet que está a ser elaborada polo Grupo TALG no marco de desenvolvemento do Multilingual Central Repository (González Agirre e Rigau 2013), unha plataforma web de libre consulta 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 30.000 palabras agrupadas en 20.000 conceptos, e está dispoñíbel na web para consulta (http://sli.uvigo.es/galnet/) e tamén para descarga (http://adimen.si.ehu.es/web/MCR/).

3.4. 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 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 CLUVI inglés-galego* 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 e Gómez Guinovart 2014). O acceso ao dicionario na web pode facerse consultando directamente o recurso (http://sli.uvigo.es/dicionario) ou a través da plataforma RILG. Tamén resulta posíbel descargar o recurso de modo directo (http://hdl.handle.net/10230/20053) ou a través da plataforma europea Meta-Share (http://metashare.elda.org/).

3.5. 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 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 da Termoteca 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 da Termoteca 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 ás súas variantes documentadas, tanto intralingüísticas (termos sinónimos, variantes ortográficas ou variantes dialectais) como interlingüísticas (traducións ou, con maior propiedade, equivalencias). A información especificada na Termoteca 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 da Termoteca 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.es/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/).

3.6. Neoteca

A Neoteca é o 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 e Rodríguez Guerra 2003). A Neoteca pódese consultar libremente na web como recurso independente (http://sli.uvigo.es/NEO/) ou integrado no 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 da lingua galega xerados de xeito independente polo Instituto dá 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 científica 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 uso gratuíto garante esta vocación de transferencia social do coñecemento compartida por todos os participantes no proxecto.

BIBLIOGRAFÍA

ÁLVAREZ LUGRÍS, A. e X. GÓMEZ GUINOVART (2014): "Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del Dicionario Moderno Inglés-Galego", in M. J. Domínguez Vázquez, X. Gómez Guinovart e C. Valcárcel Riveiro

(eds.): *Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva*. Berlín: Gruyter.

GÓMEZ CLEMENTE, X. M. e A. RODRÍGUEZ GUERRA (2003): *Neoloxía e lingua galega: teoría e práctica*. Vigo: Universidade de Vigo.

GÓMEZ CLEMENTE, X. M., X. GÓMEZ GUINOVART, A. GONZÁLEZ PEREIRA e V. TABOADA LORENZO (2013): "Sinonimia e rexistros na construción do WordNet do galego", *Estudos de lingüística galega* 5, pp. 27-42.

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

GÓMEZ GUINOVART, X. (2012): "A Hybrid Corpus-Based Approach to Bilingual Terminology Extraction", in I. Moskowich-Spiegel Fandiño e B. Crespo (eds.): *Encoding the Past, Decoding The Future: Corpora in the 21st Century*. Newcastle upon Tyne: Cambridge Scholar Publishing, pp. 147-175.

GÓMEZ GUINOVART, X. (2014): "Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do galego", in *Actas do XV Colóquio de Outono*. Braga: Centro de Estudos Humanísticos da Universidade do Minho.

GÓMEZ GUINOVART, X. e S. LÓPEZ FERNÁNDEZ (2009): "Anotación morfosintáctica do Corpus Técnico do Galego", *Linguamática* 1.1, pp. 61-71.

GÓMEZ GUINOVART, X., A. ÁLVAREZ LUGRÍS e E. DÍAZ RODRÍGUEZ (2012): *Dicionario moderno inglés-galego*. Ames: 2.0 Editora.

GONZÁLEZ AGIRRE, A. e G. RIGAU (2013): "Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository", *Linguamática* 5.1, pp. 13-28.

GONZÁLEZ SEOANE, E., M. ÁLVAREZ DE LA GRANJA e I. BOULLÓN AGRELO (eds.) (2006): *Dicionario de dicionarios do galego medieval*. Anexo 57 de *Verba (Anuario Galego de Filoloxía)*. Santiago de Compostela: Universidade de Santiago de Compostela.

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

MARTÍNEZ LEMA, P., R. DOURADO FERNÁNDEZ e C. OSORIO PELÁEZ (2010): "Un novo recurso para os estudos toponomásticos: o Inventario Toponímico da Galicia Medieval (ITGM)", in X. Sousa Fernández (ed.): *Toponimia e cartografí*a. Santiago de Compostela: Consello da Cultura Galega/Instituto da Lingua Galega, pp. 239-263

SANTAMARINA, A. (ed.) (2003): *Diccionario de diccionarios*, 3ª ed. A Coruña: Fundación Barrié de la Maza.

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

VARELA BARREIRO, X. (2004): "Un proxecto do ILG no abalo da gramática histórica da lingua galega", en R. Álvarez, F. Fernández Rei e A. Santamarina (eds.): *A lingua galega: historia e actualidade*. Santiago de Compostela: Instituto da Lingua Galega/Consello da Cultura Galega, vol. 2, pp. 649-684.