Galego 3.0

Novas oportunidades e desafíos para a investigación lingüística

XAVIER GÓMEZ GUINOVART

A investigación lingüística chega aos inicios do século XXI cunha visión altamente fragmentada e interdisciplinaria do seu obxecto de estudo. A aproximación estruturalista ao estudo da linguaxe como un sistema abstracto, autónomo e illado retrocedeu nos últimos decenios fronte a outras perspectivas da ciencia que volven máis complexo o fenómeno lingüístico integrando os seus aspectos cognitivos e sociais, ao tempo que favorecen as investigacións empíricas, aplicadas e pluridisciplinarias. A implantación da sociedade da información obriga a resituar a disciplina nuns parámetros determinados polas necesidades lingüísticas e polas tecnoloxías e servizos propios deste contexto. Trátase, sen dúbida, dunha oportunidade e dun desafío de futuro que debe servir para impulsar e anovar a investigación sobre o galego e as súas tecnoloxías lingüísticas.

No último lustro, a crecente penetración e imbricación social da Internet, epítome absoluto da sociedade da información, desembocou na chamada *web 2.0* ou web *social*, caracterizada polo crecemento dos proxectos colaborativos de tipo *wiki*, da publicación de *blogs*, das redes sociais virtuais e do compartimento de obxectos culturais. Na actualidade, algúns fenómenos máis recentes, ou que recentemente acadaron unha meirande transcendencia, como a conexión ubicua á Internet, os dispositivos móbiles de acceso á rede, o almacenamento en liña dos datos e programas, a portabilidade da identidade dixital ou o tratamento da información con maior atención ao seu significado, apuntan en conxunto cara a unha nova configuración da web que, desde diversas perspectivas, é denominada web *intelixente*, web *semántica* ou, simplemente, *web 3.0*, enfatizando así a importancia dos avances realizados respecto ás características do modelo anterior. Cómpre salientar aquí que disciplinas como a lingüística computacional e o procesamento da linguaxe natural constitúen pezas fundamentais neste novo paradigma da web semántica, fornecendo moitos dos recursos e das técnicas necesarias para a incorporación de significado na información textual manexada.

Se as ciencias da linguaxe pretenden seguir sendo unha disciplina fulcral dentro deste contexto social tecnoloxizado, postindustrial e posmoderno, deberemos ser quen de xerar un coñecemento lingüístico útil e actualizado, e de facelo accesible á sociedade, de xeito que poida contribuír ao seu avance. Na lingüística galega déronse xa moitos pasos orientados nesa dirección, incluíndo numerosos exemplos de estudos lingüísticos sobre o galego que aproveitan as ferramentas tecnolóxicas para a investigación e a difusión dos seus resultados, e de aplicacións de tecnoloxía lingüística desenvolvidas para o procesamento da lingua galega a nivel oral e escrito. A sociedade galega dispón xa dun variado repertorio de recursos lingüísticos actualizados para o coñecemento e uso do galego que, malia non ser suficiente en moitas áreas, nalgunhas outras iguala ou mesmo supera con mérito aos doutras linguas con moitos máis falantes e apoio institucional.

Un excelente exemplo recente de investigación lingüística levada a termo neste ámbito é o traballo realizado polo equipo coordinado por Xosé Luís Regueira no Instituto da Lingua Galega para a elaboración e divulgación do *Dicionario de pronuncia da lingua galega* (A Coruña: RAG, 2010). Neste dicionario, recóllese a pronuncia

estándar dos máis de 50.000 vocábulos contidos no *Vocabulario ortográfico da lingua galega (VOLGa)* (A Coruña / Santiago de Compostela: RAG / ILG, 2004), codificada en Alfabeto Fonético Internacional (AFI). Un dicionario fonético normativo destas características non existía previamente, malia se tratar dun material lingüístico imprescindible en ámbitos tan dispares como os do ensino da lingua, a locución xornalística ou a dobraxe. Porén, a utilidade e difusión pública desta ferramenta achábase algo limitada a causa do seu formato de publicación como libro impreso e mais pola esixencia de coñecer o AFI, isto é, o sistema de codificación fonética empregado. A transformación do *Dicionario de pronuncia da lingua galega* nunha aplicación web de libre consulta permitiu superar as dúas limitacións, dándolle unha nova dimensión ao traballo realizado. Esta aplicación, dispoñible desde hai uns poucos meses na web do ILG (http://ilg.usc.es/pronuncia/), permite atopar directamente a palabra buscada e escoitar sen pexa ningunha a súa locución estándar, gravada na coidada pronuncia do actor galego de dobraxe Luís Iglesia Besteiro.

A nova versión electrónica do dicionario de Regueira demostra con clareza meridiana a importancia da aplicación das novas tecnoloxías aos estudos lingüísticos. Mediante a adaptación do contido do dicionario ao formato web, a implementación dun sistema de busca e recuperación das entradas, e a inclusión das gravacións sonoras de voz con tecnoloxía *streaming*, incorpóranse no dicionario valores engadidos fundamentais que contribúen a valorizar e rendibilizar en gran medida os esforzos investidos na elaboración, presentación e difusión dos resultados da investigación lingüística realizada.

A hiperconectividade e ultrarrapidez asociada coa sociedade da información favorece sen dúbida que os dicionarios e obras de referencia de consulta libre na web sexan moito máis consultados que as obras en papel ou en disco. De aí a oportunidade da adaptación ao formato web do *Dicionario de dicionarios*, 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 (Fundación Pedro Barrié de la Maza: A Coruña, 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 Seminario de Lingüística Informática (SLI) do Grupo de Investigación en Tecnoloxías e Aplicacións da Lingua Galega (Grupo TALG) da Universidade de Vigo fixo posible 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 nas páxinas do SLI (http://sli.uvigo.es/DdD/), 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 innegable como dicionario da lingua, aínda non superado en extensión como conxunto por ningún outro.

Froito desta mesma colaboración interuniversitaria entre Vigo e Compostela, levouse a cabo 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 (Universidade de Santiago de Compostela: Santiago de Compostela, 2006) e adaptado posteriormente á web para a súa libre consulta nas páxinas do SLI da Universidade de Vigo (http://sli.uvigo.es/DDGM/).

Coa distribución a través da web destas dúas compilacións lexicográficas históricas, acadouse o obxectivo de poñer ao dispor do público interesado unha parte moi importante do patrimonio lexicográfico galego, permitindo o seu uso como ferramenta de consulta e alicerce filolóxico de estudos lingüísticos e culturais mediante unha interface de acceso libre, de fácil manexo, acceso ubicuo e consulta versátil. Trátase sen dúbida dunha boa demostración das posibilidades que ofrece a aplicación das ferramentas tecnolóxicas dispoñibles para a preservación do patrimonio cultural e a construción e difusión de recursos lingüísticos.

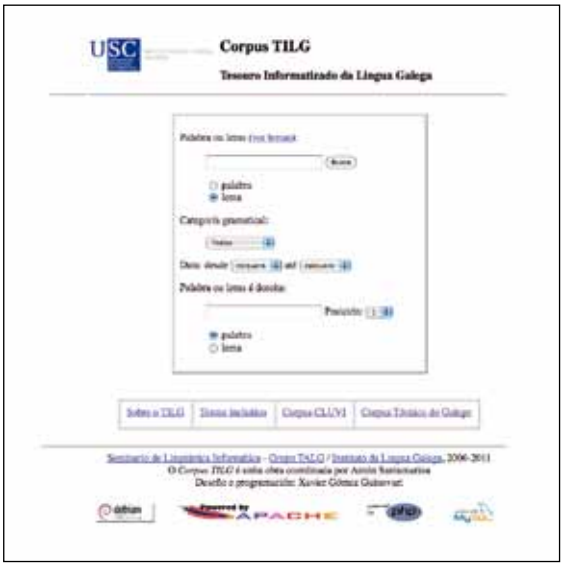
Outro dos campos nos que a lingüística galega aplicada de orientación tecnolóxica manifesta un desenvolvemento salientable é na compilación en soporte dixital e análise de grandes volumes de datos textuais representativos das variedades da lingua estudadas, un terreo de investigación que adoito recibe a denominación de *lingüística de corpus* e no que as tecnoloxías informáticas desempeñan un papel fundamental. Desde o punto de vista da investigación lingüística, a lingüística de corpus representa un avance considerable na precisión das descricións lingüísticas e na súa adecuación á realidade, fornecendo os datos empíricos necesarios para a xeración de hipóteses plausibles sobre a estrutura e funcionamento das linguas. A dispoñibilidade pública na web destes corpus textuais dixitalizados de grandes magnitudes (medidas en millóns de palabras) a través de interfaces de consulta axeitadas permite que calquera persoa interesada poida explorar doadamente, de modo empírico e desde diversos enfoques, os textos reais nas variedades lingüísticas representadas.

A lingua galega dispón dun abano de corpus textuais de acceso público relativamente completo, o que permite levar a cabo investigacións textuais empíricas en diversos eidos de estudo mediante as ferramentas tecnolóxicas e conceptuais da lingüística de corpus. Así, dispoñemos dun excelente corpus de galego antigo, o Tesouro Medieval Informatizado da Lingua Galega (TMILG), de máis de nove millóns de palabras, desenvolvido no ILG baixo a dirección de Xavier Varela Barreiro. O TMILG contén a totalidade das obras non notariais (literarias, históricas, relixiosas, xurídicas e técnicas) publicadas na Galicia medieval e o 80 por cento das obras notariais publicadas, e pode ser consultado na web do ILG (http://ilg.usc.es/tmilg/) a través dunha boa interface de consulta desenvolvida por Imaxin Software. O TMILG é a parte máis avanzada do proxecto do Corpus Xelmírez, un importante corpus textual que pretende recoller a produción escrita en Galicia en galego, castelán ou latín durante a época medieval, e que se atopa dispoñible para libre consulta nas páxinas do SLI (http://sli.uvigo.es/xelmirez/).

Así mesmo, dispoñemos tamén dun corpus de traducións do galego en combinación con diversas linguas, o Corpus Lingüístico da Universidade de Vigo (CLUVI), que recolle un conxunto textual de máis de 23 millóns de palabras formado por textos orixinais e as súas traducións. Desde un punto de vista temático, os textos recompilados no Corpus 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 castelán, o inglés, o francés, o alemán, o catalán, o portugués e o éuscaro. Este corpus paralelo, pioneiro a nivel mundial nos estudos de tradución baseados en corpus, atópase dispoñible para consulta pública na web desde setembro de 2003 nas páxinas do SLI (http://sli.uvigo.es/CLUVI/) e constitúe o alicerce empírico dun bo número de traballos académicos de investigación nos eidos da estilística da tradución, da didáctica do ensino de idiomas, da lingüística comparada, da terminoloxía e da lexicografía plurilingüe.

O mesmo grupo de investigación responsable do CLUVI, o Grupo TALG da Universidade de Vigo, levou a cabo a constitución do Corpus Técnico do Galego, un corpus de rexistros especializados do galego contemporáneo, con textos publicados nos campos do dereito, da informática, da economía, das ciencias ambientais, das ciencias sociais e da medicina que totalizan máis de 15 millóns de palabras, e que pode ser consultado libremente na web do SLI (http://sli.uvigo.es/CTG/).

O galego contemporáneo é recollido tamén, cunha maior amplitude temática, no Corpus de Referencia do Galego Actual (CORGA), unha colección textual de 25 millóns de palabras que comprende documentos publicados desde 1975 ata a actualidade. O CORGA está a ser desenvolvido no Centro Ramón Piñeiro para a Investigación en Humanidades (CRPIH) baixo a dirección

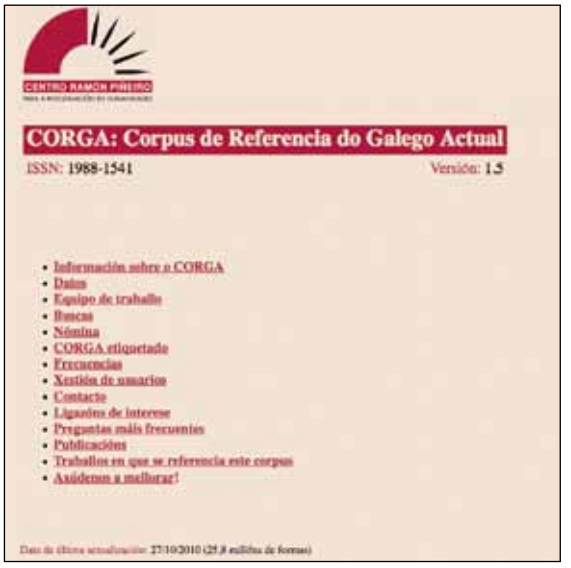


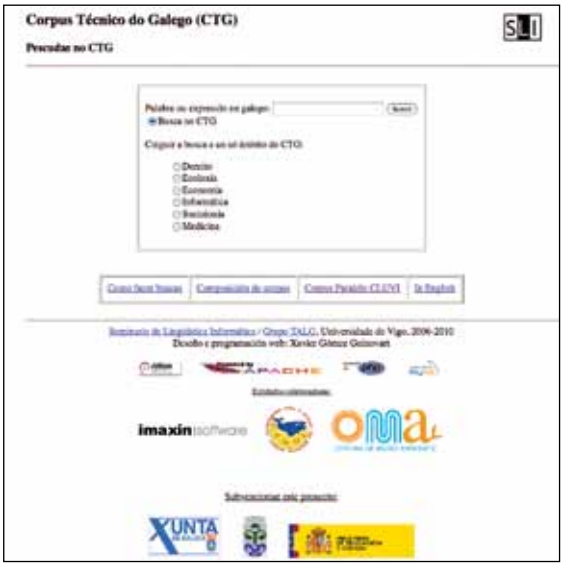
do profesor Guillermo Rojo, e pode ser consultado, previa alta no sistema, nas páxinas do CRPIH (http://corpus.cirp.es/corga/).

Con todo, o corpus textual filoloxicamente máis relevante para o estudo do galego moderno e contemporáneo é, sen ningunha dúbida, o Tesouro Informatizado da Lingua Galega (TILG). Este corpus foi desenvolvido no ILG baixo a dirección do profesor Antón Santamarina e inclúe a práctica totalidade das obras en galego entre 1612 e 1980, e unha ampla representación das publicadas desde 1980 ata a actualidade. Por razóns históricas, o TILG pódese consultar arestora 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). 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 SLI, constitúe unha edición revisada e ampliada deste corpus, tanto no seu número de textos, como no seu nivel de anotación lingüística. Nesta segunda edición ampliada do TILG, de libre acceso nas páxinas do SLI, o número de textos ascende a 1.897, incluíndo textos publicados ata o ano 2010 e totalizando máis de 25 millóns de palabras, todas elas lematizadas e anotadas gramaticalmente.

Por suposto, a simple acumulación en soporte dixital de transcricións textuais máis ou menos enriquecidas lingüisticamente non supón en si mesma unha grande axuda para a investigación filolóxica, precisándose do auxilio de ferramentas informáticas para a súa consulta e análise. A interface de consulta deseñada para a segunda edición do TILG, por exemplo, permite explorar os textos do corpus indicando a palabra, o lema, a categoría gramatical, as datas de publicación ou o contexto léxico, permite visualizar a lista de lemas documentados nos textos para elaborar a consulta, e permite comprobar a categoría morfosintáctica anotada no corpus para cada palabra do texto recuperado. Trátase de facilitar, mediante ferramentas de consulta axeitadas, a mellor explotación posible da información lingüística contida nos textos.

O conxunto de corpus textuais dispoñibles do galego constitúen un recurso lingüístico informatizado básico, cuxa consulta directa nos permite observar sen intermediarios o uso real dunha palabra nos textos, os seus contextos, os seus sentidos, as súas traducións, as súas construcións gramaticais e, moi especialmente, documentarnos sobre palabras que non están nos dicionarios, sobre sentidos, acepcións ou traducións que non se recollen nas obras lexicográficas, ou sobre equivalencias bilingües de tradución en pares de linguas que non dispoñen de (bos) dicionarios bilingües pero si de corpus paralelos. Así mesmo, estes corpus textuais representan unha base empí-





rica dispoñible como ferramenta de consulta para a codificación ortográfica, léxica e gramatical da lingua, permitindo que as institucións responsables da elaboración da normativa poidan ter en conta, para fundamentar as súas decisións, os usos lingüísticos documentados nos textos.

Os corpus tamén son compoñentes esenciais na elaboración de ferramentas básicas para o procesamento lingüístico da linguaxe. Etiquetadores gramaticais, ferramentas para a identificación de nomes propios, analizadores sintácticos probabilísticos, etiquetadores semánticos e programas de desambiguación do significado léxico son algunhas das utilidades baseadas en corpus sen as que non se poderían construír as aplicacións máis complexas das tecnoloxías da lingua, desde a tradución automática ao recoñecemento da fala. Por outra banda, moitas das aplicacións máis visibles destas tecnoloxías, como as aplicacións de texto preditivo, os sistemas de ditado, os correctores ortográficos, os programas de resumo de textos ou os sistemas de tradución automática estatística (do estilo de Google Translate), basean os seus algoritmos de funcionamento directamente nos datos deducidos dos corpus. Deste modo, os corpus lingüísticos están na base de moitas das tecnoloxías implicadas no uso da lingua en equipamentos característicos da sociedade da información, como poden ser o computador, o teléfono móbil, a axenda electrónica, o reprodutor MP3/MP4 ou a videoconsola, contextos de uso nos que a lingua demostra a súa utilidade social, ao tempo que gaña prestixio entre un amplo sector de poboación de interese estratéxico para a planificación lingüística.

Os retos que a sociedade da información presenta para a lingüística galega supoñen, ao mesmo tempo, unha oportunidade e un desafío: a oportunidade de imprimir un impulso renovador á disciplina e o desafío de realizar un traballo socialmente útil e valioso. Na súa andaina de futuro pola nova *sociedade 3.0*, a lingüística galega debe participar plenamente no espazo global creado polas tecnoloxías da información, adoptando a arañeira da Internet como plataforma de difusión mundial e traballo colaborativo, asegurando a presenza da lingua e das súas tecnoloxías nos equipamentos e servizos propios deste tipo de sociedade, e contribuíndo desde a investigación sobre o galego ao desenvolvemento da web semántica. Como demostran os exemplos mencionados ao longo do artigo, unha parte deste camiño está xa percorrido. Porén, precisaranse aínda moitas vontades e esforzos para non perder o rumbo ■