Estud. lingüíst. galega 5 (2013): 27-42

Sinonimia e rexistros na construción do WordNet do galego

Xosé María Gómez Clemente
Universidade de Vigo (Grupo TALG)
xgomez@uvigo.es
Xavier Gómez Guinovart
Universidade de Vigo (Grupo TALG)
xgg@uvigo.es
Andrea González Pereira
Universidade de Vigo (Grupo TALG)
andreagonzalezp@uvigo.es
Verónica Taboada Lorenzo
Universidade de Vigo (Grupo TALG)
veronicataboadal@uvigo.es

Recibido o 06/02/2013. Aceptado o 04/04/2013

Synonymy and registers in the construction of the Galician WordNet

Resumo

Neste artigo presentamos o tratamento lexicográfico do rexistro no proxecto Galnet do Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, dirixido á construción da versión galega do WordNet 3.0. Trátase dun proxecto que se atopa na súa fase inicial de desenvolvemento, mais do que xa se obtiveron uns primeiros resultados que están dispoñibles para a consulta. Ao longo deste artigo describiremos os trazos xerais do proxecto, a aproximación seguida para o tratamento do rexistro e algúns dos resultados e conclusións obtidas nesta primeira etapa do traballo lexicográfico.

Palabras chave

Lexicografía, semántica, lingüística computacional

Abstract

In this paper, we present the lexicographical treatment of register in the Galnet project of the TALG Group (Galician Language Tecnologies and Applications) of the University of Vigo, focused on the construction of the Galician version of WordNet 3.0. This is a project that is in its early stage of development, but which has already obtained some initial results available for consultation. In this paper we will describe the general characteristics of the project, the approach followed for the treatment of register and some results and conclusions obtained in this first stage of lexicographical work.

Keywords

Lexicography, semantics, computational linguistics

Este traballo foi financiado polo Ministerio de Economía y Competitividad, dentro do proxecto *Adquisición de escenarios de conocimiento a través de la lectura de textos: desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)* (ref. TIN2012-38584-C06-04); e pola Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia, grazas á convocatoria de Axudas para a consolidación e estruturación de unidades de investigación competitivas do Sistema Universitario de Galicia, dentro da *Rede de Lexicografía (Relex)* (ref. CN 2012/290) e da *Rede de Tecnoloxías e análise dos datos lingüísticos* (ref. CN 2012/179).

1. Introdución

WordNet (Fellbaum 1998, Miller *et al.* 1990) é unha base de coñecementos léxicos estruturada en forma de rede semántica. Nesta rede léxico-semántica, cada nó é un concepto, e os fíos que conectan estes nós son as relacións semánticas (hiponimia, meronimia...) que se establecen entre eles. Cada concepto na rede está representado polo grupo de lemas sinónimos que poden expresar ese concepto. Na terminoloxía asociada a WordNet, cada grupo de sinónimos é un *synset*, e cada sinónimo que forma parte dese grupo é unha *variant* (ou variante léxica dun mesmo concepto). WordNet inclúe, ao carón de cada *synset*, unha breve definición distintiva (ou glosa) do significado compartido por todas as variantes do *synset* e, en certos casos, exemplos de uso das variantes en contexto.

WordNet foi orixinalmente concibido para a lingua inglesa e, aínda que hoxe existen versións do WordNet en moitas linguas, o WordNet do inglés segue sendo arestora a versión máis desenvolvida e a de referencia. Os traballos do WordNet para esta lingua lévanse a cabo desde 1985 na Universidade de Princeton baixo a dirección do profesor George A. Miller. Na versión actual, o WordNet 3.0 do inglés contén 155287 lemas (variantes) agrupados en 117659 grupos de sinónimos (*synsets*).

WordNet constitúe, sen dúbida, o recurso de semántica léxica computacional máis importante na actualidade, especialmente, no ámbito do procesamento da linguaxe natural (PLN), onde é utilizado, por exemplo, en tarefas de desambiguación semántica automática (Agirre / Edmonds 2006), de recuperación da información (Varelas *et al.* 2005), de clasificación automática de textos (Elberrichi *et al.* 2008) ou de resumo automático (Barzilay / Elhadad 1997).

Na actualidade existen versións do WordNet en distintas fases de desenvolvemento para moi diversas linguas[1], incluídas o hebreo (Ordan / Wintner 2007), o italiano (Pianta *et al.* 2002), o xaponés (Isahara *et al.* 2008), o castelán (Fernández / Vázquez 2010, Fernández / Vázquez / Fellbaum 2008), o catalán (Oliver / Climent 2011) e o euskera (Pociello / Agirre / Aldezabal 2011).

A maioría das versións en linguas distintas do inglés seguen o modelo de deseño de EuroWordNet (Vossen 2002), no que os *synsets* que forman parte doWordNet da lingua propia están vinculados cos *synsets* do resto das linguas a través dun ILI (*InterLingual Index*, isto é, índice interlingüístico) que é único para cada concepto e que principalmente está baseado nos synsets do WordNet inglés de referencia. Deste xeito, o conxunto de léxicos WordNet nos distintos idiomas permiten a conexión entre os *synsets* de calquera par de linguas a través do ILI, constituíndo así un recurso de gran utilidade en aplicacións das tecnoloxías lingüísticas que precisan o procesamento plurilingüe da linguaxe, como a tradución automática ou a recuperación interlingüística da información. Cómpre salientar tamén que os conceptos que forman parte do ILI están catalogados en xerarquías de dominios e ontoloxías, como a xerarquía de dominios IRST (Bentivogli *et al.* 2004) ou as ontoloxías SUMO (Pease / Niles / Li 2002) e Top Concept Ontology (Álvez *et al.* 2008), o que permite un mellor aproveitamento do recurso en diversas aplicacións.

Neste artigo[2] presentamos o tratamento lexicográfico do rexistro no proxecto Galnet do Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, dirixido á construción da versión galega do WordNet 3.0. Trátase dun proxecto que se atopa na súa fase inicial de desenvolvemento, mais do que xa se obtiveron uns primeiros resultados que están dispoñibles para a consulta. Nos seguintes apartados describiremos os trazos xerais do proxecto, a aproximación seguida para o tratamento do rexistro e algúns dos resultados e conclusións obtidas nesta primeira etapa do traballo lexicográfico.

[1] The Global WordNet Association mantén unha listaxe de léxicos WordNet desenvolvidos por linguas na súa páxina web (http://www.globalwordnet.org). Tamén se pode acceder a unha boa variedade de léxicos WordNet para distintas linguas a través da páxina do proxecto Open Multilingual Wordnet (http://casta-net.jp/~kuribayashi/multi/).

[2] Queremos agradecer aquí sinceramente as valiosas contribucións á mellora do artigo das dúas persoas expertas que anonimamente realizaron a revisión previa á súa aceptación por parte da revista.

2. O proxecto Galnet

O obxectivo do proxecto Galnet consiste na construción dun WordNet para o galego aliñado co ILI xerado a partir do WordNet 3.0 do inglés. Este proxecto está incorporado noutro máis amplo encamiñado á integración coordinada das versións castelá, catalá, galega e vasca do WordNet 3.0, no que participan os grupos de investigación do IXA (da Euskal Herriko Unibertsitatea - Universidade do País Vasco), TALP (Universitat Politècnica de Catalunya), GRIAL (Universitat Autònoma de Barcelona, Universitat de Barcelona, Universitat de Lleida e Universitat Oberta de Catalunya), IULATERM (Universitat Pompeu Fabra) e TALG (Universidade de Vigo).

O marco de desenvolvemento no que se integra o Galnet é o do Multilingual Central Repository (MCR) (Atserias *et al.* 2004, González / Laparra / Rigau 2012), unha plataforma web de libre consulta[3] desenvolvida ao abeiro do proxecto europeo Meaning (IST-2001-34460) e dos proxectos de financiamento estatal KNOW (TIN2006-15049-C03) e KNOW2 (TIN2009-14715--C04-01), que contou coa colaboración para o galego do Grupo TALG da Universidade de Vigo a través do proxecto *Multilingual Central Repository 2.0: TALG* (FFI2009-08317-E/FILO) do Subprograma 2009 de *Acciones Complementarias* financiado polo Ministerio de Ciencia e Innovación, e na que participamos actualmente a través do proxecto vixente *Adquisición de escenarios de conocimiento a través de la lectura de textos: desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)* (TIN2012-38584-C06-04) do Ministerio de Economía y Competitividad. O MCR abrangue na actualidade os léxicos WordNet de cinco linguas (inglés, español, catalán, vasco e galego) enlazados interlingüisticamente polo ILI correspondente ao WordNet 3.0 e cos ILI categorizados na xerarquía de dominios IRST e nas ontoloxías SUMO e Top Concept Ontology.

Nas seguintes subseccións describiremos a metodoloxía e as ferramentas empregadas na construción do Galnet nas súas dúas primeiras etapas de desenvolvemento.

2.1. Primeira fase

Os obxectivos desta primeira fase na construción do Galnet foron, en primeiro lugar, elaborar un conxunto de *synsets* básicos para a operatividade do recurso na lingua galega e, en segundo lugar, fornecer un conxunto suficiente de entradas que servise para ilustrar a utilidade do recurso e ampliar a súa operatividade. A metodoloxía utilizada para levar a cabo o primeiro obxectivo consistiu na creación da versión galega dos *synsets* nominais e verbais pertencentes aos Basic Level Concepts (BLC). Como segundo obxectivo, elaboramos as entradas galegas para os ficheiros lexicográficos do WordNet correspondentes aos nomes relacionados coas partes do corpo e coas substancias, e para unha parte dos correspondentes aos adxectivos de tipo xeral.

Os Basic Level Concepts (Izquierdo / Suárez / Rigau 2007) son un conxunto seleccionado de conceptos do WordNet que representan un compromiso entre dous principios de caracterización contraditorios: representar o maior número posible de conceptos (ser conceptos abstractos) e representar o maior número posible de trazos distintivos (ser conceptos concretos). Así, os BLC aparecen tipicamente na parte media das relacións semánticas xerárquicas de WordNet, sendo deste modo frecuentes e destacados, nin claramente xerais nin demasiado específicos. A primeira tarefa do proxecto Galnet consistiu en elaborar manualmente a versión galega dos BLC (649 *synsets* nominais e 616 *synsets* verbais) recollidos no apartado *freqmin20/all* da distribución oficial[4] dos BLC do WordNet 3.0, sen incluír na adaptación nin as glosas nin os exemplos incluídos nos *synsets* correspondentes da lingua inglesa.

[3] http://adimen.si.ehu.es/web/MCR/

[4] http://adimen.si.ehu.es/web/BLC/

Unha vez elaborado o núcleo inicial de *synsets* do Galnet, continuamos a ampliación do recurso a partir da tradución asistida dos ficheiros lexicográficos do WordNet para os nomes relacionados coas partes do corpo e coas substancias, e para unha parte dos adxectivos de tipo xeral. A ferramenta empregada nesta tarefa foi Google Translator Toolkit[5], unha ferramenta colaborativa en liña que nos permitiu a postedición asistida das propostas de tradución automática do tradutor de Google.

A selección dos ficheiros lexicográficos relacionados coas partes do corpo e coas substancias veu motivada pola nosa vontade de aproveitar o material textual e terminolóxico elaborado en traballos previos do grupo e recollidos no Corpus Técnico do Galego (CTG)[6] e na base de datos terminolóxica da Termoteca[7]. A incorporación dos adxectivos xustificouse en virtude dunha maior cobertura lingüística dos resultados nesta fase inicial do traballo. Na Táboa 1 preséntanse, agrupados en categorías (nomes, verbos, adxectivos e adverbios) e diferenciando entre *synsets* e variantes, os resultados acadados desde un punto de vista cuantitativo nesta primeira xeira do desenvolvemento do proxecto Galnet[8]. Estes resultados corresponden a 649 *synsets* (1333 variantes léxicas) dos BLC de categoría nominal, 616 *synsets* (1414 variantes) dos BLC de categoría verbal, 2014 *synsets* (3550 variantes) do ficheiro lexicográfico de nomes relacionados coas partes do corpo, 2983 *synsets* (4300 variantes) do ficheiro lexicográfico de nomes de substancias, e 3114 *synsets* (4864 variantes) do conxunto de adxectivos de tipo xeral incluídos en WordNet 3.0.

	WN30		**Galnet**	
	Vars	Syns	Vars	Syns
N	117798	82115	9183	5646
V	11529	13767	1414	616
Adx	21479	18156	4864	3114
Adv	4481	3621	0	0
TOTAL	155287	117659	15461	9376

Táboa 1. Resultados iniciais (primeira fase)

Tendo en conta os resultados obtidos en todas as categorías, o crecemento lexicográfico do Galnet nesta primeira fase do proxecto atinxiu unha cobertura semántica moi próxima ao 10 % con relación a cobertura de referencia do WordNet 3.0 en lingua inglesa. Na subsección seguinte, describiremos as estratexias seguidas para a ampliación do Galnet na súa segunda etapa de desenvolvemento.

2.2. Segunda fase

Na segunda fase de desenvolvemento do proxecto Galnet, utilizamos a ferramenta WN-Toolkit (Oliver 2012) para ampliar o recurso a partir de dous recursos bilingües inglés-galego xa existentes: a Wikipedia[9] (denominada Galipedia na súa versión en lingua galega) e o Dicionario CLUVI inglés-galego[10] (Gómez / Álvarez / Díaz 2012). As técnicas de extracción automática aplicadas

[5] http://translate.google.com/toolkit/

[6] http://sli.uvigo.es/CTG/

[7] http://sli.uvigo.es/termoteca/

[8] Os datos cuantitativos completos sobre o WordNet 3.0 do inglés poden consultarse na web da Universidade de Princeton, en http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[9] http://wikipedia.org

[10] http://sli.uvigo.es/dicionario/

a estes dous recursos léxicos bilingües tiveron dous obxectivos diferenciados: por unha banda, ampliar o Galnet cos nomes propios que teñen unha forma ortográfica idéntica en inglés e en galego a partir do material fornecido pola Wikipedia; e por outra banda, ampliar o Galnet coas variantes galegas recollidas na Wikipedia e no Dicionario CLUVI como tradución de palabras inglesas incluídas nos *synsets* do WordNet (e non codificadas aínda no Galnet).

Debido á dificultade da tarefa, as técnicas de extracción automática aplicadas foron complementadas por un demorado proceso de revisión humana, no que as variantes candidatas identificadas polo programa de extracción foron aprobadas ou rexeitadas unha a unha por un revisor humano. O resultado da extracción automática, revisado manualmente, serviu para ampliar o Galnet con 11677 novas variantes e 9936 novos *synsets*, é dicir, ao duplo da extensión obtida na primeira fase.

As técnicas de extracción aplicáronse de xeito secuencial e ordenado, dando prioridade á información léxica sobre os lemas simples fornecida polo dicionario e á información sobre os nomes propios proporcionada pola Wikipedia. Deste modo, desde un punto de vista cuantitativo, os resultados da ampliación obtidos en cada unha das etapas da extracción léxica foron os seguintes:

2945 variantes nominais pluriléxicas do inglés coas iniciais de todas as palabras en maiúscula e que figuran na Wikipedia;

2483 variantes nominais e adxectivas do Dicionario cuxo lema inglés aparece nun único *synset* do WordNet e que teñen como tradución unha única palabra galega que non aparece como tradución noutros lemas ingleses;

1529 variantes nominais e adxectivas do Dicionario cuxo lema inglés aparece nun único *synset* do WordNet e que teñen como tradución unha única palabra galega que aparece tamén como tradución noutros lemas ingleses;

1818 variantes nominais e adxectivas do Dicionario cuxo lema inglés aparece nun único *synset* do WordNet e que teñen como tradución máis dunha palabra galega;

2971 variantes nominais enlazadas do galego ao inglés na Galipedia e que non estaban no Galnet.

2.3. Estado actual

A Táboa 2 recolle o estado actual do proxecto Galnet acadado trala súa segunda fase de desenvolvemento, ao carón dos datos fornecidos polo WordNet 3.0 da lingua inglesa.

	WN30		**Galnet**	
	Vars	Syns	Vars	Syns
Nomes	117798	82115	18949	14285
Verbos	11529	13767	1416	612
Adxectivos	21479	18156	6773	4415
Adverbios	4481	3621	0	0
TOTAIS	155287	117659	27138	19312

Táboa 2. Estado actual

Aínda que obviamente o Galnet é aínda un proxecto en curso, os seus resultados prácticos na fase de traballo actual xa se poden consultar libremente en internet mediante distintas interfaces web de consulta, entre as que cómpre salientar a interface de EuroWordNet[11], compartida co resto de grupos de investigación cos que traballa o noso grupo a nivel estatal, e a interface do proxecto xaponés[12] Open Multilingual WordNet, na que o galego se pode consultar na súa

[11] http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl

[12] http://casta-net.jp/~kuribayashi/multi/

relación de equivalencia co inglés, indonesio, malaio, xaponés, tai, farsi, árabe, hebreo, italiano, portugués, noruegués (bokmal e ninorsk), danés, francés, finés, catalán, español, vasco e albanés. Tamén se pode consultar a través das páxinas do SLI e na plataforma RILG de Recursos Integrados da Lingua Galega[13].

Así mesmo, os resultados das sucesivas versións do Galnet xeradas ao longo do proxecto pódense descargar libremente de internet en forma de base de datos. Na súa distribución pública actual, o Galnet distribúese con licenza Creative Commons (CC BY 3.0)[14] como parte do Multilingual Central Repository, actualmente dispoñíble na súa versión 3 (MCR 3.0)[15] de maio de 2012. O Galnet descárgase, así, en formato SQL nunha distribución pública que inclúe o WordNet 3.0 do inglés, do vasco, do español, do catalán e do galego, en diferentes fases de desenvolvemento.

3. A sinonimia no Galnet

A sinonimia é unha das relacións semánticas máis complexas e nela está baseada a información léxica representada no WordNet. Enténdese, en liñas xerais, que hai sinonimia cando para un significado existen na lingua varios significantes e que dúas palabras son sinónimas cando os trazos semánticos que definen os seus significados son os mesmos. No Galnet, o WordNet do galego, igual que nos léxicos WordNet das outras linguas, inclúense varias relacións semánticas (como, por exemplo, a antonimia, a hiponimia, a meronimia, a troponimia ou a implicación), aínda que a relación principal, maioritariamente representada e vertebradora da estrutura interna do léxico, é a sinonimia. De feito, a base organizativa de WordNet son os *synsets* que inclúen sinónimos que funcionan como tales se un deles aparece nun contexto e pode ser substituído polos outros, sen que cambie o seu contido semántico. Estamos perante un concepto de sinonimia amplo, o único que permite a construción dos *synsets*. En palabras de Miller (1998: 23-24):

> The basic relation in WordNet is synonymy. Sets of synonyms (synsets) form the basic building blocks. [...] The notion of synonymy used in WordNet does not entail interchangeability in all contexts; by this criterion, natural languages have few synonyms. The most modern claim is that WordNet synonyms can be interchanged in some contexts. To be careful, therefore, one should speak of synonymy relative to a context, but in order to facilitate the discussion this qualification will usually be presupposed, not asserted.

3.1. Rexistros e *synsets*

Na elaboración do Galnet do galego consideramos a pertinencia de caracterizar os rexistros aos que pertencen os elementos (variantes) das series sinonímicas (*synsets*). A consignación desta información ten o obxecto de mellorar a comprensión da relación de sinonimia que se dá entre as distintas variantes compoñentes dun *synset*. Esta práctica non é o habitual que cumpriría nas obras lexicográficas; porén, é a nosa intención incluír sistematicamente no Galnet esta información.

Os rexistros conforman as variedades diafásicas. Enténdese que una persoa utiliza a lingua de maneira diferente segundo sexa a situación comunicativa na que interactúa. A lexicógrafa francesa Rey-Debove considera que cómpre estudar os niveis de lingua como un fenómeno social e tamén individual, e chega a afirmar que "est bien difficile de décider de la synonymie si on ne connaît pas l'idiolecte de l'énonciateur" (1997: 98). Estamos perante valores que non sempre

[13] http://sli.uvigo.es/RILG/

[14] http://creativecommons.org/licenses/by/3.0

[15] http://adimen.si.ehu.es/web/files/mcr30/mcr30.zip

son codificables por pertenceren á fala individual (Gutiérrez 1992: 118-124), é dicir, a percepción do rexistro pode estar condicionada pola experiencia propia do falante.

A relación de sinonimia en WordNet está baseada na posibilidade de intercambiar os constituíntes dos *synsets* entre si nalgún contexto determinado. Xa que logo, cómpre axustar ao máximo a información que se dea de cada palabra que forma o conxunto dos *synsets* para tentar que funcione esa intercambiabilidade. A información que se dea sobre os rexistros facilitará a determinación do contexto específico onde se produce a sinonimia. Esta información é fundamental para todos os usuarios pero, sobre todo, para os que aprenden a lingua. É moi evidente que sinónimos que teñan distinto rexistro non van poder ser intercambiables dunha maneira automática. Vexámolo en dous exemplos de *synsets* tirados de Galnet: (1) formado cunha serie de sinónimos intercambiables e (2) con problemas de intercambiabilidade.

(1) glg-30-00262792-a: *valente, afouto, bravío, valoroso, varudo, decidido, bravo* (que posúe ou expón a súa coraxe; capaz de afrontar e manexar o perigo ou o medo sen vacilar).
(2) glg-30-00266985-a: *asustadizo, medroso, cagón, covarde* (falto de coraxe ou de vitalidade).

No caso de (2) o feito de non seren intercambiables está motivado por existir un sinónimo *cagón* cun nivel de lingua diferente e que se pode cualificar de coloquial. Xa que logo, como se pode ver, no WordNet non se tivo en conta o problema da intercambialidade dos sinónimos adscritos a distintos rexistros. Partiuse da idea teórica de que a sinonimia se produce nalgúns contextos e noutros non, e deixouse a determinación dos contextos efectivos de sinonimia á experiencia que teña o usuario no uso dunha determinada lingua.

Á diferenza do traballo feito ata agora, a nosa intención no desenvolvemento do Galnet é fornecer esa información, sempre dentro das posibilidades limitadas que ten a fixación exacta dos rexistros. Trátase dun problema non totalmente resolto na lexicografía, sobre todo no que atinxe á determinación dos niveis que están baixo o nivel estándar e, se cadra, á fixación das estremeiras entre o nivel culto/especializado e o estándar, nun momento en que se producen grandes transvasamentos de información entre niveis de lingua por mor dos procesos de vulgarización e terminoloxización en curso.

4. A DETERMINACIÓN DO REXISTRO

O obxectivo da investigación presentada neste artigo foi tentar establecer unha metodoloxía cuxa aplicación nos permitise a adscrición dun rexistro a cada variante do Galnet sen estarmos condicionados por un uso persoal do léxico, naqueles casos onde as indicacións lexicográficas non o orientaban. Esta metodoloxía supón a comprobación sistemática en corpus das ocorrencias das unidades léxicas investigadas para comprobar se se pode variar a adscrición inicial de rexistro tomada do *Diccionario da Real Academia Galega* (de agora en diante DRAG).

As variantes seleccionadas para a investigación pertencen ao ficheiro de partes do corpo de WordNet, formado por 2014 *synsets* nominais relacionados tematicamente. Concretamente, as 419 variantes obxecto inicial do estudo son as que constitúen os 206 *synsets* dese conxunto, caracterizados por incluír unha ou máis variantes pertencentes ao léxico patrimonial do galego. Desas 419 variantes, 67 son unidades léxicas polisémicas, como *queixo*, *bigornia*, *martelo*, *campaíña*, *polpa* etc., que foron eliminadas da análise final polas dificultades para comprobar nos corpus as ocorrencias que correspondían ás acepcións pertinentes, de modo que o número das variantes analizadas neste estudo é de 352, de acordo coa metodoloxía que se explica nesta sección.

4.1. Determinación do rexistro baseada en dicionarios

A adscrición inicial a un rexistro de cada variante seleccionada realizouse en dúas fases. En primeiro lugar, comprobamos no DRAG a determinación de rexistro asignada a cada variante analizada. Nos compostos sintagmáticos comprobouse no dicionario cada unha das bases que entraban no proceso de formación. Se o núcleo ou o modificador do composto non estaban asociados a ningunha marca no dicionario, o termo foi clasificado como estándar; pero se un dos elementos estaba asociado a algunha marca foi clasificado por esta. Así, asignamos inicialmente a variante *substancia cincenta* (glg-05483388-n)[16] ao rexistro estándar, xa que ningún dos seus constituíntes está marcado no DRAG, malia recoñecermos a súa alta probabilidade de ser un termo especializado. En contraste, a variante *antro etmoidal* (glg-05253165-n) foi clasificada como especializada, dado que *antro* tiña unha marca no DRAG que asociaba esa unidade á especialidade de anatomía.

Cando no DRAG non había documentación sobre un termo monoléxico ou pluriléxico que consideramos que podía estar adscrito inicialmente ao nivel especializado, fixemos unha comprobación na base de datos do Buscatermos, do Servizo de Normalización Lingüística da Universidade de Santiago de Compostela, para confirmar ou descartar a súa atribución inicial a ese rexistro. Este foi o caso de 43 variantes analizadas como por exemplo *valvuleta* (glg-05395548-n), *antisoro* (glg-05403702-n), *forame* (glg-05545439-n), *sura* (glg-05574332-n), *malleus* (glg-05325943-n) e outros latinismos.

Para a marcaxe inicial das variantes analizadas baseada nas indicacións do DRAG, realizamos unha adaptación a un sistema de 5 marcas a partir das 8 marcas de rexistro codificadas nas entradas deste dicionario. Na Táboa 3 recollemos as adaptacións aplicadas no estudo sobre o rexistro descrito neste artigo.

Marca de rexistro no DRAG	Marca no Galnet
sen indicación	*estándar*
lit. (literario)	*culto*
máis formal (sic)	*estándar*
ámbito técnico de uso, como *Anat.*	*especializado*
pop. (popular)	*coloquial*
fam. (familiar)	*coloquial*
ext. (por extensión)	*coloquial*
vulg. (vulgar)	*vulgar*
pex. (pexorativo)	*vulgar*

Táboa 3. Adaptación das marcas de rexistro do DRAG

Así, como se pode ver, reducimos as clasificacións de *familiar*, *por extensión*, *popular* e *coloquial* a unha única categoría de *coloquial* por entendermos que é difícil situar as estremeiras entre cada unha delas; as de *vulgar* e *pexorativo* a *vulgar*, xa que o "pexorativo" sempre está asociado a connotacións vulgares; a de *literario* convertémola en *culto* por interpretala como referida a usos restrinxidos dun alto nivel de lingua. Cando non había no DRAG ningunha indicación de rexistro, entendemos que estabamos perante o nivel estándar, e o mesmo fixemos cando aparecía a indicación *máis formal* no interior dunha entrada, xa que con esta indicación se caracteriza unha unidade como dun nivel superior (sempre asociado ao estándar) a un nivel necesariamente "menos formal". Con estas adaptacións, a asignación inicial das variantes analizadas ás etiquetas de rexistro baseadas no DRAG (e, como se explicou previamente, no Buscatermos) ficou coa distribución recollida na Táboa 4.

[16] Especificamos entre parénteses trala variante o número identificador do seu *synset* para precisar a acepción concreta á que nos referimos en cada caso.

Rexistro inicial	Variantes por rexistro
0. Vulgar	7 (1,98 %)
1. Coloquial	34 (9,65 %)
2. Estándar	195 (55,39 %)
3. Especializado	73 (20,73 %)
4. Culto	43 (12,21 %)

Táboa 4. Distribución das variantes analizadas por rexistro inicial

4.2. Determinación do rexistro baseada en corpus

Coa finalidade de tratar de determinar dunha maneira máis completa a asignación de rexistros ás variantes analizadas, fixemos unha comprobación sistemática das súas ocorrencias en corpus textuais, concretamente no *Corpus Técnico do Galego* (CTG)[17], no *Corpus de Referencia do Galego Actual* (CORGA)[18] e no *Tesouro Informatizado da Lingua Galega* (TILG)[19].

Dentro do CORGA, e a efectos da caracterización do rexistro das ocorrencias das variantes, distinguimos un subcorpus cun nivel medio-baixo de especialización formado polos textos do CORGA pertencentes ás categorías "Economía e Política" (CORGA-EP), "Cultura e Artes" (CORGA--CA) e "Ciencias Sociais" (CORGA-CS); un segundo subcorpus cun nivel medio-alto de especialización, constituído polos textos da sección de "Ciencias e Tecnoloxía" (CORGA-CT); e un terceiro subcorpus sen nivel de especialización, formado polos textos da categoría "Ficción" (CORGA-F) ou sen clasificar (CORGA-OU) .

Por outra parte, caracterizamos todos os subcorpus do CTG cun nivel medio de especialización, nos ámbitos do Dereito (CTG-DER), Ecoloxía (CTG-ECX), Economía (CTG-ECM), Informática (CTG-INF) e Socioloxía (CTG-SOC), agás o subcorpus do CTG de Medicina (CTG-MED) que se considera o máis especializado de todos os subcorpus utilizados.

Para a asignación de rexistro baseada en corpus, partimos das seguintes hipóteses correspondentes a cada un dos cinco niveis determinados en Galnet:

Rexistro vulgar (nivel 0): unha unidade léxica será do rexistro vulgar se se rexistra en corpus xerais e non nos especializados. O número de ocorrencias podería non ser moi elevado. Será do mesmo rexistro se non hai documentación no corpus, por seren palabras sometidas a unha forte interdición social. En principio, non deberá existir unha grande dispersión das ocorrencias entre os corpus, é dicir, aparecerán naqueles en que haxa textos que reflictan a fala popular e precisen dunha alta expresividade.

Rexistro coloquial (nivel 1): unha unidade léxica será do rexistro coloquial se se rexistra en corpus xerais, con maior dispersión ca no nivel anterior, e excepcionalmente en corpus especializados. A maior dispersión está xustificada pola ausencia de tabús sociais.

Rexistro estándar (nivel 2): unha unidade léxica será do rexistro estándar se se rexistra en todo tipo de corpus, cunha ampla dispersión.

Rexistro especializado (nivel 3): unha unidade léxica será especializada se se rexistra principalmente en corpus especializados, aínda que tamén poida aparecer nos xerais polos fenómenos de vulgarización terminolóxica.

[17] http://sli.uvigo.es/CTG/

[18] http://www.cirp.es/corga/

[19] http://sli.uvigo.es/TILG/

Rexistro culto (nivel 4): unha unidade léxica será culta se só se rexistra en corpus especializados, con pouca dispersión, ou non se rexistra por ser unha palabra de uso restrinxido.

5. Análise dos resultados

Da revisión das ocorrencias de cada unha das variantes en cada un dos 13 subcorpus consultados (CORGA-EP, CORGA-CA, CORGA-CS, CORGA-CT, CORGA-F, CORGA-OU, CTG-DER, CTG-ECX, CTG-ECM, CTG-INF, CTG-SOC, CTG-MED e TILG) despréndense os seguintes resultados: hai 93 variantes que non teñen ningunha ocorrencia nos 13 subcorpus; 91 rexístranse menos de 10 veces (22 só 1 vez); 58 entre 10 e 50 veces; 29 entre 50 e 100 veces; 55 entre 100 e 500; 11 entre 500 e 1000 e, por último, 23 entre 1000 e 5730 veces. No que atinxe á súa distribución polos subcorpus, 22 variantes aparecen nun único corpus, 29 en 2, 36 en 3, 27 en 4, 28 en 5, 14 en 6, 16 en 7, 14 en 8, 14 en 9, 14 en 10, 5 en 11, 10 en 12 e, por último, 2 en 13.

Aplicando as hipóteses descritas anteriormente para a asignación de rexistro baseada na presenza nos corpus, comprobouse se existía a posibilidade de cambiar a adscrición inicial de rexistro ou, se polo contrario, a adscrición inicial seguía sendo válida á luz dos novos datos. A comprobación fíxose rexistro por rexistro, e a continuación analizamos deste modo os seus resultados.

5.1. Rexistro vulgar

No rexistro vulgar incluímos inicialmente sete variantes (1, 98 %): *pelos da cona* (glg-05263732-n), *pirola* (glg-05526713-n), *cona* (variante asociada a dous *synsets*: glg-05514410-n e glg-05263732-n), *collón* (glg-05524615-n), *carallo* (glg-05526713-n) e *esgarro* (glg-05415815-n). Todas elas, agás a primeira que só se recolle unha vez, teñen unha especial presenza en corpus non especializados como o CORGA-F, aínda que son relevantes as ocorrencias en CORGA-CS, CORGA-CA, CORGA-EP que consideramos que teñen unha certa especialización. O composto *pelos da cona*, que só se rexistra unha vez en CORGA-F, semella estar afectado por un tabú social moito maior. *Esgarro* mesmo aparece recollido catro veces en CTG-MED.

A análise dos corpus permitiríanos, en principio, considerar como coloquiais *pirola*, *cona*, *esgarro*, *collón* e *carallo* por teren todos eles un número elevado de ocorrencias e unha relativa dispersión nos corpus, xa que aparecen a que menos *pirola*, en catro corpus, e a que máis, *carallo*, en oito.

5.2. Rexistro coloquial

No rexistro coloquial incluímos inicialmente 34 unidades (9,65%). Seis variantes non teñen ningunha documentación no corpus. Nos casos de *peseta* (glg-05521514-n)], *bólas* (glg-05524615-n) e *trastes* (glg-05524615-n), a ausencia de documentación estaría motivada por unha teórica interdición sexual ou polo feito de non ser denominacións especialmente frecuentes, sobre todo *trastes*. A ausencia de documentación podería provocar que se cambiase o rexistro para o nivel vulgar, pero non se debe entender que unicamente por esta razón *peseta* sexa vulgar e *cona* popular, tal e como argumentamos antes. *Óso da alegría* (glg-05580662-n), *nariz de falcón* (glg-05598982-n) e *óso tolo* (glg-05580662-n) son creacións metafóricas e expresivas que teñen moi pouca relevancia nos nosos corpus por teren unha autoría determinada ou unha distribución dialectal moi específica. Non se xustifica un cambio de nivel para ningunha das variantes mencionadas.

Hai outro grupo de variantes que ten poucas ocorrencias, menos de 100 en todos os corpus analizados, centradas en corpus determinados de tipo xeral e ausentes nos especializa-

dos. Son *faceira* (glg-05599769-n), *crica* (glg-05521514-n), *paxara* (glg-05521514-n), *ollo do cu* (glg-05538215-n), *perrecha* (glg-05521514-n), *cintura de avespa* (glg-05555840-n), *michelíns* (glg-05556204-n) e *pai de todos* (glg-05567604-n). As cinco primeiras rexístranse en tres ou en menos de tres corpus, cun número de rexistros moi baixo. Isto pode indicarnos que están sometidas a unha interdición sexual ou social. Deberían ser adscritas ao nivel vulgar. As tres restantes rexístranse en cinco corpus ou en menos e, en principio, non hai ningunha razón para cambiar a adscrición de rexistro.

Outro grupo ten un número significativo de ocorrencias que seguen sen rexistrarse en corpus especializados. *Cacho* (glg-05539454-n), *dedo furabolos* (glg-05567381-n), *testa* (glg-05539454--n), *inzo* (glg-05404336-n), *cachola* (glg-05539454-n) aparecen nun mínimo de seis corpus e nun máximo de 12. O número de ocorrencias e a dispersión non permiten cambiarlles a adscrición inicial de rexistro.

Un terceiro grupo fórmano aquelas variantes que teñen ocorrencias máis ou menos frecuentes pero con presenza en corpus especializados. Sobre as variantes *flegma* (glg-05415815--n), *cocote* (glg-05539595-n), *couquizo* (glg-05539595-n), *panza* (glg-05556071-n), *fociño* (glg-05598707-n), *gorxa* (glg-05547508-n) e *bagulla* (glg-05405324-n) entendemos que a súa presenza nos corpus, cun número de ocorrencias relativamente importante, a súa dispersión, cada unha delas en máis de seis corpus, e sobre todo a súa presenza en corpus especializados fai que debamos consideralas estándar e non coloquiais.

5.3. Rexistro estándar

No rexistro estándar incluímos inicialmente 195 variantes (55,39 %). Hai un primeiro subgrupo de variantes (26) que non teñen documentación nos corpus. Este feito non motiva que haxa un cambio na adscrición inicial do rexistro cara aos niveis inferiores, coloquial e vulgar, pero si hai unha serie de palabras como, entre outras, *sangue do cordón umbilical* (glg-05402472-n), *zume dixestivo* (glg-05405946-n), *moco seco* (glg-05416128-n), *substancia cincenta* (glg-05483388-n), *dente temporal* (glg-05306894-n), *moa cordal* (glg-05307952-n), *fisura de Sylvius* (glg-05224080--n), *fisura calcarina* (glg-05224585-n), *fisura de Rolando* (glg-05223823-n), *canle de nacemento* (glg-05226937-n) ou *canle de Schlemm* (glg-05251789-n) que son claramente especializadas. A inexistencia de documentación en ningún dos 13 subcorpus (xerais e especializados) non deixa de sorprender, xa que estamos perante compilacións textuais formadas por un número moi elevado de palabras. Tampouco podemos falar de que sexan denominacións extremadamente especializadas que non se poderían rexistrar máis ca en corpus de altísima especialización. Sexa cal for a razón desta situación, é moi evidente que a adscrición inicial de rexistro debe ser cambiada para o nivel 3 (especializado).

Hai outro conxunto de variantes, formado por *barbela dos Habsburgo* (glg-05600030-n), *pico de viúva* (glg-05256562-n), *cardado colmea* (glg-05257393-n), *costela verdadeira* (glg-05591999--n) e *óso nú* (glg-05271607-n), tamén sen documentación, que semellan creacións expresivas infrecuentes, ou inexistentes, en galego. Non hai ningunha razón para cambiar a adscrición inicial.

Outro grupo só se rexistra unha vez pero sempre en corpus especializados. Son variantes como *cuncha nasal* (glg-05229341-n), *dente da teta* (glg-05306894-n), *dente de mama* (glg-05306894-n), *orificio anatómico* (glg-05545439-n), *óso do peito* (glg-05281189-n), *óso nu* (glg-05271607-n), *dente molar* (glg-05306476-n), *tecido areolar* (glg-05268255-n), *padal duro* (glg-05309591-n), *dente posterior* (glg-05306476-n), *corneto* (glg-05229198-n), *bordo alveolar* (glg-05310351-n), *bóveda cranial* (glg-05540407-n), *enxerto de pel* (glg-05239437-n), *esputo* (glg-05415815-n) ou *medula ósea* (glg-05285623-n) que, en principio, debían ser consideradas todas como especializadas. Tamén se poden engadir a este grupo as variantes *vea xugular* (glg-05370918-n), *padal brando* (glg-05309392-n), *xenitais femininos* (glg-05514410-n), *estrutu-*

ra muscular (glg-05461816-n), *vidalla* (glg-05602683-n), *dedo maior* (glg-05567604-n), *secreción mucosa* (glg-05415395-n), *tecido intersticial* (glg-05268797-n), *zume gástrico* (glg-05406128-n), *soro sanguíneo* (glg-05403149-n) *substancia gris* (glg-05483388-n), *escápula* (glg-05279688-n), *testículo* (glg-05524615-n), *tecido adiposo* (glg-05268965-n) e *medula* (glg-05285623-n), que se rexistran en corpus xerais pero moito menos ca nos especializados. Todas estas variantes deberían pasar ao rexistro especializado.

Outro subgrupo fórmano as variantes pouco documentadas, rexistradas en moi poucos corpus e sen presenza en corpus especializados, como *óso espido* (glg-05271607-n), *óso do ombreiro* (glg-05279688-n), *dente de coello* (glg-05306390-n), *nariz ganchudo* (glg-05599501-n), *liña do pelo* (glg-05256220-n), *barriga da perna* (glg-05574332-n), *barba de chibo* (glg-05263029-n), *barbarote* (glg-05599617-n), *virilla* (glg-05597734-n), *carocha* (glg-05401753-n), *dediño* (glg-05567727--n), *cachela* (glg-05401753-n), *punto cego* (glg-05456082-n), *cara de porco* (glg-05601662-n), *barbada* (glg-05599769-n), *guedella* (glg-05258051-n), *dente de leite* (glg-05306894-n) e *queixelo* (glg-05599617-n). Se aplicamos a nosa hipótese inicial, estas unidades deberían pasar ao nivel coloquial xa que non están suficientemente representadas nos corpus, están presentes en moi poucos deles e non aparecen nos corpus máis especializados.

Por último, hai un grupo numeroso de variantes que non deberían cambiar de rexistro por estaren moi representadas nun gran número de corpus. Malia algunhas teren certo carácter especializado, a documentación fainos pensar que están suficientemente vulgarizadas para mantelas no nivel estándar. Son as seguintes: *seme* (glg-05404336-n), *válvula* (glg-05222591--n), *graxa* (glg-05268965-n), *esperma* (glg-05268965-n), *pelo* (glg-05254393-n), (glg-05254795--n), (glg-05255692-n) *cabelo* (glg-05254393-n), (glg-05256085-n) *peiteado* (glg-05256862--n), *vea* (glg-05418717-n), *bágoa* (glg-05405324-n), *estómago* (glg-05556943-n), *pescozo* (glg-05546540-n), *pupila* (glg-05320183-n), *pene* (glg-05526384-n), *ventre* (glg-05555917--n), *moa* (glg-05307773-n), *cairo* (glg-05307091-n), *media lúa* (glg-05582038-n), *musculatura* (glg-05582038-n), *abdome* (glg-05556943-n), *padal* (glg-05309725-n), *saliva* (glg-05416198-n), *labio* (glg-05305806-n), *caluga* (glg-05546540-n), *cella* (glg-05313535-n), *barriga* (glg-05555917--n), *rizo* (glg-05258299-n), *cabeleira* (glg-05256085-n), *lágrima* (glg-05405324-n), *sebo* (glg-05416979-n), *papo* (glg-05599769-n), *baba* (glg-05416678-n) , *bandullo* (glg-05556071-n), *pus* (glg-05417472-n), *moco* (glg-05416048-n) (glg-05415395-n), *ampola* (glg-05254197-n), *caspa* (glg-05401753-n), *queixada* (glg-05546298-n), *ceo da boca* (glg-05309725-n) etc.

5.4. Rexistro especializado

No nivel especializado incluímos inicialmente 73 variantes (20,73 %). Delas 35 non teñen ningunha documentación nos corpus. Este número elevado de termos non rexistrados en ningún dos corpus, nin sequera no do dominio da medicina, pode indicar un nivel alto de especialización que os nosos corpus non acadan. Son palabras como *antro etmoidal* (glg-05253165-n), *arteria do labirinto* (glg-05349445-n), *células cono* (glg-05456257-n), *retina cono* (glg-05456257-n), *óso do carpo* (glg-05271814-n), *suco lateral cerebral* (glg-05224080-n), *valvuleta* (glg-05224080-n), *vea xugular interna* (glg-05371482-n) etc., que non poden ser cambiadas a un rexistro inferior ou superior.

Hai un grupo de 20 variantes que só se rexistran nos corpus de especialidade, principalmente en CTG-MED. Son casos como *antisoro* (glg-05403702-n), *vaíña de mielina* (glg-05464685-n), *seo transverso* (glg-05252834-n), *vea xugular externa* (glg-05371301-n), *seo recto* (glg-05252705--n), *feblectasia* (glg-05421586-n), *fluído seminal* (glg-05404336-n) etc., que tampouco ven afectada a adscrición a este nivel.

As variantes *disco óptico* (glg-05456082-n), *materia gris* (glg-05483388-n), *albedo* (glg-05483388-n), *úvula* (glg-05309245-n), *colo do útero* (glg-05303232-n), *omoplata*

(glg-05279688-n), *axila* (glg-05549576-n), *fascículo* (glg-05475681-n), *farinxe* (glg-05547508-n), *encéfalo* (glg-05481095-n), *cartilaxe* (glg-05288091-n), *enxiva* (glg-05304932-n), *sistema nervioso* (glg-05462315-n) e *cerebro* (glg-05481095-n) rexístranse en máis de catro subcorpus, xerais e especializados, e teñen ocorrencias superiores a 10 (ata as 2.416 de *cerebro*). Son casos de vulgarización terminolóxica, producida pola difusión dos dominios especializados en textos que non son de especialidade. Poderíase propoñer consideralas como estándar porque, malia manteren o seu carácter de unidades de valor especializado, entran de cheo na comunicación non especializada.

5.5. Rexistro culto

No nivel culto incluímos inicialmente 43 unidades (12,21%), principalmente denominacións latinas. Delas 28 non teñen ningún tipo de documentación no corpus, feito que se explica pola inexistencia da tradición de denominar con cultismos latinos partes do corpo. En Galnet, deixamos esta posibilidade máis recoñecemos que case non ten representación no uso. Son variantes como *oculus dexter* (glg-05312149-n) *os scaphoideum* (glg-05272276-n), *dedo pollex* (glg-05567217-n), *macula lutea* (glg-05455690-n), *genu varum* (glg-05561834-n), *cervix uteri* (glg-05303232-n), etc.

As variantes *fasciculus* (glg-05475681-n), *malleus* (glg-05325943-n), *palpebra* (glg-05313822--n), *sura* (glg-05574332-n), *sinus coronarius* (glg-05252402-n) e *serum* (glg-05403149-n) só se rexistran en CTG-MED. *Supercilium*, *cervix*, *macula* rexístranse en CTG-MED e unha vez cada unha delas en corpus non especializados. Con todo non podemos pensar que esteamos perante un proceso de vulgarización terminolóxica. Os casos de *sinus*, *falo* e *ventre* están representados por un número significativo de ocorrencias en máis de cinco corpus. Nestes casos si que poderiamos falar de vulgarización terminolóxica e consideralos pertencentes ao rexistro estándar.

6. Conclusións

A metodoloxía empregada para establecer un criterio que nos permita unha adscrición do rexistro baseada no uso revelouse como interesante. As hipóteses iniciais confírmanse nunha porcentaxe relativamente elevada. No rexistro vulgar (nivel 0) cambiouse a adscrición inicial en 5 variantes o que supón un 71,4 % do total; no rexistro coloquial (nivel 1) fíxose en 12 casos, un 35,29 %; no estándar (nivel 2) en 52, un 26,6 %; no especializado (nivel 3) en 14, un 19,17 % e, por último, no culto (nivel 4) en 3, un 6,97 %. Pódese ver o sentido destas alteracións polo miúdo na Táboa 5.

	> 0	**> 1**	**> 2**	**> 3**	**> 4**
0	-	5	0	0	0
1	5	-	7	0	0
2		20	-	32	0
3		0	14	-	0
4			3		0

Táboa 5. Desprazamentos de rexistro inicial baseados no uso

Tendo en conta todos os rexistros analizados, a porcentaxe global das variantes que cambiaron a súa adscrición inicial mediante a aplicación da metodoloxía para a determinación do rexistro baseada en corpus presentada neste artigo foi do 24,43 %. A partir destes datos, hai que recoñecer que o éxito, o relativo éxito ou o fracaso da metodoloxía empregada, está relacionado coa estrutura dos corpus, coa posibilidade real de documentar as variantes buscadas e coa posibilidade de aplicala a un conxunto máis amplo delas.

Un dos aspectos metodolóxicos desta investigación que cumprirá mellorar para podermos incorporar información sobre o rexistro na totalidade do léxico de Galnet é o tratamento das variantes polisémicas, intencionadamente apartadas deste estudo coa finalidade de centrármonos na experimentación desta metodoloxía baseada en corpus. Coa mesma finalidade de permitir unha aplicación a maior escala da metodoloxía proposta, cumprirá axustar os criterios de adscrición inicial das variantes pluriléxicas non recollidas nos dicionarios. Nesta tarefa, podería ser de moita axuda a incorporación de novos repertorios léxicos de consulta, como o *Gran Dicionario Xerais da Lingua* (2009) ou o *Dicionario de galego* da Editorial Ir Indo (2004), que complementasen a información fornecida polo *DRAG* e polo *Buscatermos*, que constituíron a base da investigación actual.

Así mesmo, hai que ter en conta que a escalabilidade da metodoloxía proposta, e a súa aplicación á totalidade do léxico contido no Galnet, dependen en grande parte da posibilidade de automatizar algúns dos procesos requiridos e, moi especialmente, da posible automatización da comprobación nos corpus da presenza das variantes. As posibilidades deste tratamento automático vense enormemente dificultadas polo feito de non contarmos aínda cun corpus moderno de referencia do galego desambiguado semanticamente. A falta deste recurso ten unha incidencia negativa bastante importante, xa que unha boa porcentaxe das variantes son polisémicas, polo que non abonda para a súa documentación nos corpus unha mera comprobación formal.

Finalmente, cómpre observar tamén, a este respecto, a imposibilidade de automatizar por completo as tarefas relacionadas coa análise dos contextos de uso nos corpus das variantes documentadas, necesarias para establecermos os posibles desprazamentos dos rexistros iniciais, aínda que se poida deseñar algunha estratexia para un tratamento automático parcial do problema. Hai que recoñecer que a solución non é doada pero hai aberto un camiño que se podería explorar inspirado nos sistemas de extracción automática de candidatos a termos. Os máis desenvolvidos (Gómez Guinovart 2012, Nazar / Cabré 2011) utilizan a análise estatística de elementos léxicos, morfolóxicos e sintácticos que permite a selección de unidades de valor especializado, unidades que conformarían, para nós, o nivel especializado. Para os niveis vulgar e coloquial, sería posible adestrar o ordenador para que recoñecese a coaparición da variante analizada con outras variantes previamente catalogadas neses niveis. O nivel estándar podería detectarse pola ausencia total das variantes vulgares e coloquiais e pola ausencia parcial das especializadas nun contexto determinado.

Na actualidade, as nosas liñas de traballo en relación coa elaboración do Galnet como recurso léxico para o galego, pasan pola ampliación deste no ámbito do léxico especializado, do léxico xeral e da fraseoloxía, e na incorporación da información sobre rexistros baseada no uso en todas as unidades léxicas deste repertorio.

Referencias bibliográficas

Agirre, Eneko / Philip Edmonds (2006): *Word Sense Disambiguation*. Berlín: Springer.

Álvez Javier *et al.* (2008): "Consistent Annotation of EuroWordNet with the Top Concept Ontology", en *Proceedings of the 4th Global WordNet Conference*. Szeged: GWN, s. p.

Atserias, Jordi *et al.* (2004): "The MEANING Multilingual Central Repository", en *Proceedings of the 2nd Global WordNet Conference*. Brno: GWN, 23-30.

Barzilay, Regina / Michael Elhadad (1997): "Using Lexical Chains for Text Summarization", en *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*. Madrid: ACL/EACL, 10-17.

Bentivogli, Luisa *et al.* (2004): "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing", en *Proceedings of COLING Workshop on Multilingual Linguistic Resources*. Xenebra: ACL, 101-108.

Buscatermos = Servizo de Normalización Lingüística: *bUSCatermos*. Santiago de Compostela: Universidade de Santiago de Compostela. <http://www.usc.es/buscatermos/>

Carballeira, Xosé Mª *et al.* (2009): *Gran Dicionario Xerais da Lingua*. Vigo: Ed. Xerais.

CORGA = Guillermo Rojo (dir.): *Corpus de referencia do galego actual.* Santiago de Compostela: Centro Ramón Piñeiro para a Investigación en Humanidades. Santiago. <http://corpus.cirp.es/corga/>

CTG = Xosé María Gómez Clemente / Xavier Gómez Guinovart (dirs.): *Corpus técnico do galego.* Vigo: Universidade de Vigo. <http://sli.uvigo.es/CTG/>

Domínguez Dono, Xesús *et al.* (2004): *Dicionario de galego.* Vigo: Ir Indo Edicións.

DRAG = González González, Manuel / Constantino García (dirs.) (1997): *Diccionario da Real Academia Galega.* A Coruña: Real Academia Galega.

Elberrichi, Zakaria *et al.* (2008): "Using Wordnet for Text Categorization", *The International Arab Journal of Information Technology* 5(1), 16-24.

Fellbaum, Christiane (ed.) (1998): *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Fernández Montraveta, Ana / Gloria Vázquez (2010): "La construcción del WordNet 3.0 en español", en María Auxiliadora Castillo / Juan Manuel García Platero (eds.), *La lexicografía en su dimensión teórica*. Málaga: Universidad de Málaga, 201-220.

Fernández Montraveta, Ana / Gloria Vázquez / Christiane Fellbaum (2008): "The Spanish Version of WordNet 3.0", en Angelika Storrer *et al.* (eds.), *Text Resources and Lexical Knowledge*. Berlin: Gruyter, 175-182.

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

González Agirre, Aitor / Egoitz Laparra / German Rigau (2012): "Multilingual Central Repository Version 3.0: Upgrading a Very Large Lexical Knowledge Base", en *Proceedings of the Sixth International Global WordNet Conference*. Matsue: GWN, s. p.

Gutiérrez Ordóñez, Salvador (1992): *Introducción a la Semántica Funcional*. Madrid: Síntesis.

Isahara, Hitoshi *et al.* (2008): "Development of the Japanese WordNet", en *Proceedings of the Sixth International Language Resources and Evaluation*. Marrakech: ELRA, s. p.

Izquierdo, Rubén / Armando Suárez / German Rigau (2007): "Exploring the Automatic Selection of Basic Level Concepts", en *Proceedings of the International Conference on Recent Advances on Natural Language Processing*. Shoumen: INCOMA, 298-302.

Miller, George A. *et al.* (1990): "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography* 3(4), 235-244.

Miller, George A. (1998): "Nouns in WordNet", en Christiane Fellbaum (ed.), *WordNet. An Electronic Lexical Database.* Cambridge, Massachusetts/London: The MIT Press, 23-46.

Nazar, Rogelio / Teresa Cabré (2011) "Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados", *Debate Terminológico* 07, 36-55.

Oliver, Antoni / Salvador Climent (2011): "Construcción de los WordNets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente", *Procesamiento del Lenguaje Natural* 47, 293-300.

Oliver, Antoni (2012): "WN-Toolkit: un toolkit per a la creació de WordNets a partir de diccionaris bilingües", *Linguamática* 4(2), 93-101.

Ordan, Noam / Shuly Wintner (2007): "Hebrew WordNet: a Test Case of Aligning Lexical Databases Across Languages", *International Journal of Translation* 19(1), 39-58.

Pease, Adam / Ian Niles / John Li (2002): "The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications", en *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*. Edmonton: AAAI, s. p.

Pianta, Emanuele *et al.* (2002): "MultiWordNet: Developing an Aligned Multilingual Database", en *Proceedings of the First International Conference on Global WordNet*. Mysore: GWN, 21-25.

Pociello, Elisabete / Eneko Agirre / Izaskun Aldezabal (2011): "Methodology and Construction of the Basque WordNet", *Language Resources and Evaluation* 45(2), 121-142.

Rey-Debove, Josette (1997): "La synonymie ou les échanges de signes comme fondement de la sémantique", en A. Balibar-Mrabti *et al.*, *La synonymie.* Langages 128, 1997, 91-104.

RILG = Gómez Guinovart, Xavier (dir.) (2006-2013): *Recursos integrados da lingua galega.* Vigo: Universidade de Vigo, Santiago de Compostela: Instituto da Lingua Galega <http://sli.uvigo.es/RILG/>.

Termoteca = Xosé María Gómez Clemente / Xavier Gómez Guinovart (dirs.): *Termoteca, Banco de Datos Terminolóxico da Universidade de Vigo.* Vigo: Universidade de Vigo. <http://sli.uvigo.es/termoteca/>

TILG = Antón Santamarina (dir.): *Tesouro informatizado da lingua galega*. Santiago de Compostela: Instituto da Lingua Galega. <http://sli.uvigo.es/TILG/>

Varelas, Giannis *et al.* (2005): "Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web", en *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*. Nova York: ACM Press, 10-16.

Vossen, Piek (2002): "WordNet, EuroWordNet and Global WordNet", *Revue française de linguistique appliquée* 7, 27-38.