DO DICIONARIO DE SINÓNIMOS Á REDE SEMÁNTICA: FONTES LEXICOGRÁFICAS NA CONSTRUCIÓN DO WORDNET DO GALEGO

Xavier Gómez Guinovart
DEPARTAMENTO DE TRADUCIÓN E LINGÜÍSTICA, UNIVERSIDADE DE VIGO
xgg@uvigo.es

1. A rede léxico-semántica

WordNet é unha base de datos léxica do 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 (Fellbaum, 1998; Miller et al., 1990). Os nós da rede están formados por nomes, verbos ou adxectivos agrupados pola súa sinonimia. Na terminoloxía asociada a WordNet, cada grupo de sinónimos é un *synset*, e cada sinónimo lematizado que forma parte dese grupo é unha *variant* (ou variante léxica dun mesmo concepto). Deste xeito, un *synset* representa un concepto lexicalizado único e agrupa o conxunto de variantes sinonímicas dese concepto. Como complemento de cada *synset*, WordNet pode incluír 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.

No modelo de representación do léxico de WordNet, todos os *synsets* 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. Na Figura 1, ofrécese unha visualización obtida co VisuGal[1] dun subconxunto desta arañeira en forma de grafo.

WordNet foi concibido orixinalmente 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. Na súa versión 3.0, o WordNet do inglés contén 206941 lemas ou variantes sinonímicas (155287 das cales son formas únicas non homógrafas) agrupadas en 117659 grupos de sinónimos ou *synsets*.

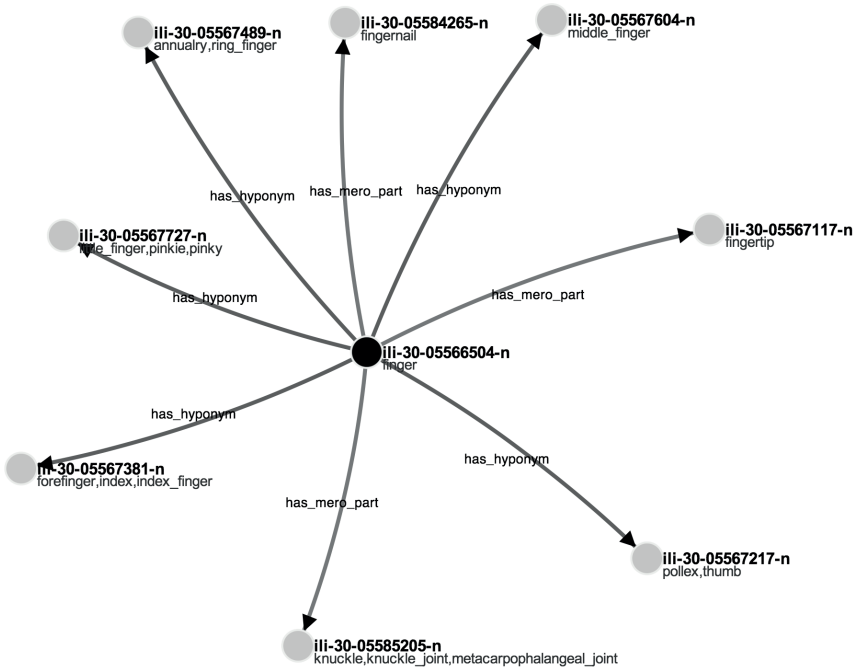

Figura 1. Sección do WordNet do inglés

WordNet constitúe, sen dúbida, o recurso de semántica léxica computacional máis importante na actualidade, especialmente, no ámbito do procesamento da linguaxe natural, onde é utilizado, por exemplo, en tarefas de

1 http://tec.citius.usc.es/VISUGAL/

desambiguación semántica automática (Agirre & Edmonds, 2006), de recuperación da información (Zhao *et al.* 2012), de clasificación automática de textos (Elberrichi et al., 2008) ou de resumo automático (Plaza et al., 2010).

Na actualidade existen versións do WordNet en distintas fases de desenvolvemento para moi diversas linguas, incluídas o hebreo (Ordan & Wintner, 2007), o italiano (Pianta et al., 2002), o xaponés (Isahara et al., 2008), o castelán (Fernández & Vázquez, 2010), o catalán (Oliver & Climent, 2011) e o euskera (Pociello et al., 2011). The Global WordNet Association mantén unha listaxe de léxicos WordNet desenvolvidos por linguas na súa páxina web[2]. Tamén se pode acceder a unha boa variedade de léxicos WordNet para distintas linguas a través da páxina do proxecto Open Multilingual Wordnet[3].

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 do WordNet da lingua propia están vinculados cos *synsets* do resto das linguas a través dun índice interlingüístico (*InterLingual Index* ou ILI) que é único para cada concepto e que principalmente está baseado nos synsets do WordNet inglés de referencia. Deste modo, 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 recuperación interlingüística da información (Agirre et al. 2007) ou a busca de respostas plurilingüe (Ferrández et al., 2007). 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 et al., 2002) e Top Concept Ontology (Álvez et al., 2008), o que permite un mellor aproveitamento do recurso en diversas aplicacións.

Neste artigo revisaremos o estado da cuestión do proxecto Galnet do Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, dirixido á construción da versión galega do WordNet. Trátase dun proxecto que se atopa na súa fase inicial de desenvolvemento, mais do que xa se obtiveron uns primeiros resultados que están dispoñíbeis para a consulta. Nos seguintes apartados describiremos os trazos xerais do proxecto, a metodoloxía seguida para a construción do recurso, e algúns dos resultados e conclusións obtidas nas primeiras etapas do labor lexicográfico.

---

2 http://www.globalwordnet.org

3 http://casta-net.jp/~kuribayashi/multi/

2. Galnet: WordNet do galego

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 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), responsábel da elaboración do Galnet.

O marco de desenvolvemento no que se integra o Galnet é o do Multilingual Central Repository[4] (MCR) (González et al., 2012; González & 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 polo ILI correspondente ao WordNet 3.0 e cos ILI categorizados na xerarquía de dominios IRST e nas ontoloxías SUMO e *Top Concept Ontology*. Na Figura 2, inclúese unha visualización con VisuGal dunha sección da rede semántica plurilingüe en construción no MCR.

Nas seguintes subseccións describiremos a metodoloxía e as ferramentas empregadas na construción do Galnet nas súas sucesivas etapas de desenvolvemento definidas a partir das fontes lexicográficas empregadas en cada fase.

3. WordNet como fonte lexicográfica

Os obxectivos da 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 cobertura léxica. A metodoloxía utilizada para levar a cabo o primeiro obxectivo consistiu na creación da versión galega dos synsets nominais e verbais pertencentes a un conxunto de conceptos básicos definidos para WordNet, os *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.

4 http://adimen.si.ehu.es/web/MCR/

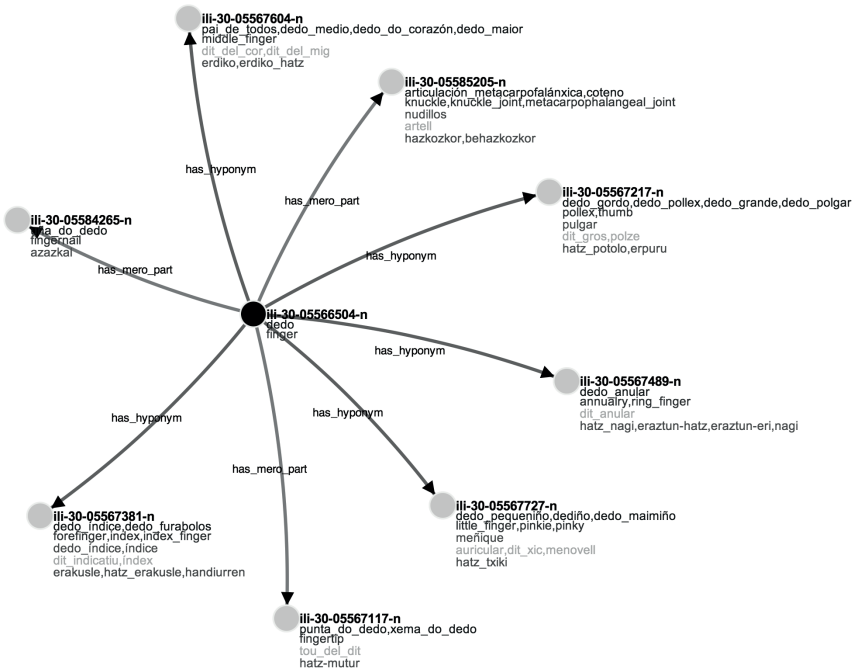

Figura 2. Sección do MCR inglés-galego-español-catalán-euskera

Os *Basic Level Concepts* (Izquierdo et al., 2007) son un conxunto seleccionado de conceptos do WordNet que representan un compromiso entre dous principios de caracterización contraditorios: representar o maior número posíbel de conceptos (ser conceptos abstractos) e representar o maior número posíbel 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[5] dos BLC do WordNet 3.0, sen incluír na adaptación nin as glosas nin os exemplos incluídos nos *synsets* correspondentes da lingua inglesa.

5 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[6], 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)[7] e na base de datos terminolóxica da Termoteca[8]. 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. Estes resultados corresponden a 649 *synsets* (1333 variantes léxicas) dos BLC de categoría nominal, 616 *synsets* (1416 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.

Táboa 1. Cobertura inicial de Galnet

	WordNet 3.0		**Galnet**	
	Vars	**Syns**	**Vars**	**Syns**
Nomes	146312	82115	9183	5646
Verbos	25047	13767	1416	616
Adxectivos	30002	18156	4864	3114
Adverbios	5580	3621	0	0
TOTAL	206941	117659	15463	9376

6 http://translate.google.com/toolkit/

7 http://sli.uvigo.es/CTG/

8 http://sli.uvigo.es/termoteca/

Tendo en conta os resultados obtidos en todas as categorías, a extensión do Galnet nesta primeira fase do proxecto atinxiu unha cobertura semántica 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, tomando como fontes lexicográficas a Wikipedia e un dicionario bilingüe inglés-galego.

4. Fontes lexicográficas bilingües inglés-galego

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 (denominada Galipedia na súa versión en lingua galega) e o Dicionario CLUVI inglés-galego (Gómez et al. 2012). As técnicas de extracción automática aplicadas 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 arduo 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*, isto é, 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.

A Táboa 2 recolle o estado final do proxecto Galnet acadado nesta segunda fase de desenvolvemento, ao carón dos datos fornecidos polo WordNet 3.0 da lingua inglesa. Cómpre salientar que a primeira distribución pública do Galnet, liberada en 2012 e dispoñíbel tanto para consulta[9] como para descarga[10], contén os datos do repertorio léxico neste estado de desenvolvemento do proxecto.

Táboa 2. Distribución actual de Galnet

	WordNet 3.0		**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

Na seguinte subsección, describiremos a metodoloxía seguida para a ampliación en curso do Galnet no ámbito do léxico xeral, utilizando como fonte lexicográfica un dicionario de sinónimos.

9 http://sli.uvigo.es/galnet/

10 http://adimen.si.ehu.es/web/files/mcr30/mcr30.zip

5. Do dicionario de sinónimos a Galnet

5.1. Preparación do dicionario

Sendo a sinonimia a relación semántica fundamental que vertebra WordNet, os dicionarios de sinónimos representan unha fonte potencial moi importante de enriquecemento deste recurso. No caso do galego, ao inicio do proxecto de elaboración de Galnet non contabamos con ningún dicionario de sinónimos, nin comercial nin libre, dispoñíbel en soporte dixital. Por esta razón, decidímonos a planificar a revisión, ampliación e conversión a formato XML TEI (TEI Consortium, 2014) dun dicionario de sinónimos tradicional do galego publicado en papel e xa descatalogado (Noia et al. 1997). Como base da conversión, contamos co conxunto de ficheiros MS-Word elaborados polos autores da obra orixinal e previos á corrección editorial e maquetación da obra de traballo.

A primeira tarefa na construción do novo dicionario electrónico de sinónimos consistiu en converter a información textual desestruturada dos ficheiros MS-Word nunha base de datos lexicográfica normalizada etiquetada en XML consonte o subconxunto de TEI deseñado para a representación de dicionarios, que ilustramos seguidamente mediante a representación en XML TEI da entrada xerada para o adxectivo *branco*:

```
<entry>
<form>
<orth>brando</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n="1"><syn><lemma>Débil</lemma></syn>, <syn><lemma>dondo</
lemma></syn>, <syn><lemma>feble</lemma></syn>, <syn><lemma>flexible</
lemma></syn>, <syn><lemma>fofo</lemma></syn>, <syn><lemma>frouxo</
lemma></syn>, <syn><lemma>lene</lemma></syn>, <syn><lemma>macio</
lemma></syn>, <syn><lemma>mol</lemma></syn>.</def>
<def n="2"><syn><lemma>Lene</lemma></syn>, <syn><lemma>suave</
lemma></syn>.</def>
<def n="3"><syn><lemma>Doce</lemma></syn>, <syn><lemma>tenro</
lemma></syn>.</def>
<def n="4"><syn><lemma>Afable</lemma></syn>, <syn><lemma>amable</
lemma></syn>, <syn><lemma>apracible</lemma></syn>,
```

```
<syn><lemma>benévolo</lemma></syn>, <syn><lemma>sereno</lemma></
syn>, <syn><lemma>transixente</lemma></syn>.</def>
<def n="5"><lbl>fig</lbl> <syn><lemma>Abaixado</lemma></syn>,
<syn><lemma>apoucado</lemma></syn>, <syn><lemma>coitado</lemma></
syn>, <syn><lemma>covarde</lemma></syn>, <syn><lemma>mexeriqueiro</
lemma></syn>, <syn><lemma>pusilánime</lemma></syn>.</def>
</sense>
</entry>
```

A consecución desta tarefa de conversión non estivo exenta de dificultades, debido principalmente os erros de formato e outras inconsistencias da edición orixinal. Moitos dos erros da conversión automatizada tiveron que ser revisados manualmente. Por fortuna, algúns deles puideron ser revisados dun modo asistido. Por exemplo, a aparición de espazos en branco espurios entre as letras dos lemas do dicionario nos textos orixinais provoca un erro frecuente na conversión a XML das entradas do dicionario. O tratamento deste erro non se pode automatizar totalmente, xa que o dicionario contén lemas con espazos en branco xenuínos, de maneira que a revisión da conversión debe facerse de xeito manual. Con todo, a corrección deste erro pode ser asistida mediante un mecanismo baseado en expresións regulares que identifique automaticamente as secuencias textuais candidatas a erro, guiando e facilitando o necesario labor de revisión humana (Gómez & Simões, 2013).

Co dicionario xa en formato XML, a segunda etapa na súa actualización consistiu na normalización da súa ortografía, morfoloxía e léxico consonte a normativa oficial vixente do galego estabelecida en 2003. A primeira ortografía oficial do galego foi aprobada en 1982 pola Real Academia Galega e o Instituto da Lingua Galega, e foi ratificada en 1983 polo goberno galego coa publicación da Lei de Normalización Lingüística. En xullo de 2003, a Real Academia Galega modificou a normativa, introducindo algúns cambios importantes na ortografía, morfoloxía e léxico (Real Academia Galega, 2004; González & Santamarina, 2004). O dicionario fonte co que traballamos foi redactado en 1997, seguindo a normativa de 1982. Por tanto, a preparación do dicionario implicou a corrección do texto consonte aos criterios normativos vixentes na actualidade.

Esta revisión normativa representa un volume de traballo moi elevado, xa que obriga a revisar o texto do dicionario na súa totalidade, identificando e corrixindo cada desvío lingüístico das normas actuais vixentes. Para facilitar esta revisión, deseñamos un programa informático que identifica no texto do dicionario fonte todas as "variantes históricas" do galego, isto é, todas as formas correctas de acordo coa normativa de 1982, pero incorrectas segundo as modificacións da normativa de 2003. Alén de identificalas, o programa tamén

substitúe a variante histórica pola súa forma correcta actual, etiquetando os cambios cunha marca que indica o tipo de normalización aplicado, seguindo unha tipoloxía elaborada para tal efecto.

Cada tipo de normalización aplicado polo programa implica un tipo distinto de postedición humana. Deste xeito, despois da normalización automática realizada polo programa, segue un proceso de postedición humana guiado polas marcas correspondentes aos tipos de normalización aplicados en cada caso (Gómez & Simões, 2013). Por exemplo, existe un tipo de normalización morfoléxica que implica a substitución dunha palabra por outra só en certos sensos do vocábulo. É o caso da palabra *vocal*, forma ortográfica instituída na normativa de 1982 para todas as súas acepcións, que pasou a escribirse como *vogal* na normativa de 2003 en todos os seus sentidos excepto cando significa "da voz ou relativo a ela". Por tanto, na súa normalización non se pode aplicar unha substitución ás cegas. O programa substitúe en todos os casos *vocal* por *vogal*, mais deixa unha marca específica deste tipo de normalización no lugar da substitución, marca que permite na postedición saber o tipo de verificación humana necesaria para a consolidación ou reversión do cambio proposto polo programa.

Unha vez rematada a normalización informática e lingüística do texto, preparamos unha interface web[11] para facilitar a súa consulta (Figura 3) e emprendemos unha dilatada fase de ampliación e revisión lexicográfica baseada na expansión dos sinónimos que dan lugar a pistas perdidas no dicionario e na resolución das remisións a outras entradas.

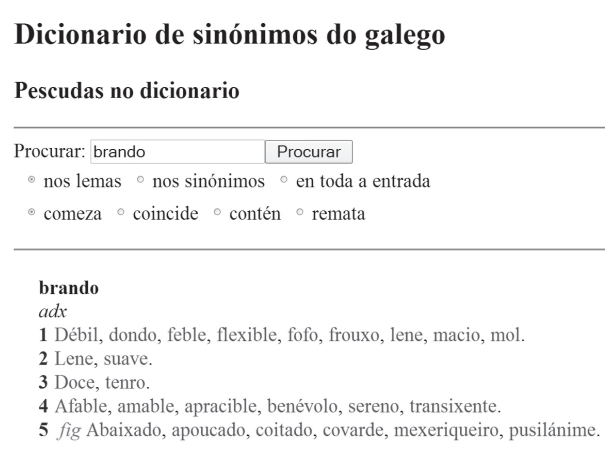

Figura 3. Dicionario de sinónimos na web

11 http://sli.uvigo.es/sinonimos/

5.2. *Solución das pistas perdidas*

A seguinte tarefa na construción do dicionario de sinónimos partiu do dicionario fonte revisado, resultado da normalización lingüística e informática descrita no apartado anterior. Sobre este dicionario, aplicamos un proceso de revisión das pistas perdidas empregando como ferramenta informática un programa capaz de identificar os sinónimos non recollidos como lemas no dicionario (isto é, as pistas perdidas) e de xerar unha entrada candidata nova coa información precisa para a súa solución mediante unha revisión humana.

A base deste proceso de revisión consiste, logo, en xerar de modo automático unha entrada candidata nova cada vez que se identifica un sinónimo que non está recollido no dicionario como lema dunha entrada. De acordo co protocolo estabelecido para o experimento, esta nova entrada candidata xerada automaticamente consta sempre dunha soa acepción e inclúe o sinónimo como lema, a categoría gramatical da entrada orixinal na que se identificou a pista perdida, e unha lista de sinónimos formada, por unha parte, polo lema da entrada orixinal e, por outra, polos posíbeis sinónimos asociados á pista perdida na entrada orixinal. Por exemplo, a partir da seguinte entrada para o adverbio *seriamente*:

```
<entry>
<form>
<orth>seriamente</orth>
</form>
<sense>
<gramGrp>adv</gramGrp>
<def n="1"><syn><lemma>Adustamente</lemma></syn>,
<syn><lemma>gravemente</lemma></syn>, <syn><lemma>rigorosamente</
lemma></syn>.</def>
<def n="2"><syn><lemma>Cabalmente</lemma></syn>,
<syn><lemma>honradamente</lemma></syn>, <syn><lemma>integramente</
lemma></syn>, <syn><lemma>responsablemente</lemma></syn>,
<syn><lemma>rigorosamente</lemma></syn>.</def>
</sense>
</entry>
```

e dado que o sinónimo *adustamente* non estaba recollido no dicionario como lema (tratándose, por tanto, dunha pista perdida), xeraríase automaticamente

unha entrada candidata nova con *adustamente* como lema e mais coa seguinte información lexicográfica:

```
<entry>
<form>
<orth>adustamente</orth>
</form>
<sense>
<gramGrp>adv</gramGrp>
<def n="1"><syn><lemma>Seriamente</lemma></syn>,
<syn><lemma>gravemente</lemma></syn>, <syn><lemma>rigorosamente</
lemma></syn>.</def>
</sense>
</entry>
```

A incorporación das novas entradas candidatas resultado dunha expansión automatizada destas características precisan dun custoso proceso de revisión humana que implica, en moitos casos, a revisión complementaria dos contidos das entradas orixinais debidas á identificación mediante este proceso de erros de diverso tipo (lapsus ortográficos, sinónimos dubidosos, formas incorrectas) no repertorio fonte. Na revisión humana das entradas candidatas son frecuentes as fusións de diversas entradas candidatas e os cambios de categoría na entrada xerada. Estes cambios son precisos, por exemplo, cando unha nova entrada xerada para un lema substantivo feminino se orixina a partir dunha pista perdida presente nunha entrada orixinal con lema substantivo masculino. Por outra banda, a fusión de entradas xeradas é necesaria cando a mesma pista perdida aparece en varias entradas do dicionario orixinal e, por tanto, dá lugar á xeración de varias novas entradas. Na revisión manual dos resultados da expansión cómpre fusionar todas esas entradas xeradas nunha soa, posibelmente, con varias acepcións ou categorías.

Así, a partir de tres entradas do dicionario fonte, para os lemas *frecuentemente*, *insistentemente* e *iterativamente*, nas que aparecía como sinónimo a pista perdida *reiteradamente*, xeráronse de modo automático as tres seguintes entradas candidatas que foron fusionadas durante a revisión humana na cuarta entrada:

```
<entry>
<form>
```

```
<orth>reiteradamente</orth>
</form>
<sense>
<gramGrp>adv</gramGrp>
<def n="1"> <syn><lemma>Acotío</lemma></syn>, <syn><lemma>a
miúdo</lemma></syn>, <syn><lemma>adoito</lemma></syn>,
<syn><lemma>correntemente</lemma></syn>, <syn><lemma>decote</
lemma></syn>, <syn><lemma>habitualmente</lemma></syn>,
<syn><lemma>normalmente</lemma></syn>, <syn><lemma>ordinariamente</
lemma></syn>, <syn><lemma>frecuentemente</lemma></syn>,
<syn><lemma>usualmente</lemma></syn>.</def>
</sense>
</entry>

<entry>
<form>
<orth>reiteradamente</orth>
</form>
<sense>
<gramGrp>adv</gramGrp>
<def n="1"> <syn><lemma>Continuamente</lemma></
syn>, <syn><lemma>pertinazmente</lemma></
syn>, <syn><lemma>insistentemente</lemma></syn>,
<syn><lemma>reiterativamente</lemma></syn>.</def>
</sense>
</entry>

<entry>
<form>
<orth>reiteradamente</orth>
</form>
<sense>
<gramGrp>adv</gramGrp>
<def n="1"> <syn><lemma>iterativamente</lemma></syn>,
<syn><lemma>repetidamente</lemma></syn>.</def>
</sense>
</entry>
```

```
<entry>
<form>
<orth>reiteradamente</orth>
</form>
<sense>
<gramGrp>adv</gramGrp>
<def n="1"><syn><lemma>Acotío</lemma></syn>,
<syn><lemma>a miúdo</lemma></syn>, <syn><lemma>adoito</
lemma></syn>, <syn><lemma>continuamente</lemma></syn>,
<syn><lemma>correntemente</lemma></syn>, <syn><lemma>decote</
lemma></syn>, <syn><lemma>frecuentemente</lemma></
syn>, <syn><lemma>habitualmente</lemma></syn>,
<syn><lemma>iterativamente</lemma></syn>, <syn><lemma>normalmente</
lemma></syn>, <syn><lemma>ordinariamente</lemma></syn>,
<syn><lemma>repetidamente</lemma></syn>, <syn><lemma>usualmente</
lemma></syn>.</def>
<def n="2"><syn><lemma>Insistentemente</lemma></syn>,
<syn><lemma>pertinazmente</lemma></syn>.</def>
</sense>
</entry>
```

Así mesmo, hai que ter en conta tamén que non sempre é posíbel darlle a volta completa as entradas dun dicionario de sinónimos. Con moita frecuencia, a causa desta asimetría radica en que a relación semántica codificada na entrada do dicionario é máis unha relación de hiperonimia que unha relación de equivalencia. Por exemplo, a entrada orixinal para o substantivo *climaterio* recolle como sinónimos os substantivos cohipónimos para "climaterio masculino" (*andropausa*) e para "climaterio feminino" (*menopausa*):

```
<entry>
<form>
<orth>climaterio</orth>
</form>
<sense>
<gramGrp>sm</gramGrp>
<def n="1"><syn><lemma>Andropausa</lemma></syn>,
<syn><lemma>menopausa</lemma></syn>.</def>
```

```
</sense>
</entry>
```

Ao non estar recollida a palabra *andropausa* no dicionario orixinal, a entrada xerada automaticamente, antes da súa corrección, era:

```
<entry>
<form>
<orth>andropausa</orth>
</form>
<sense>
<gramGrp>sm</gramGrp>
<def n="1"><syn><lemma>Climaterio</lemma></syn>,
<syn><lemma>menopausa</lemma></syn>.</def>
</sense>
</entry>
```

Alén da adaptación da categoría gramatical ao novo lema, a corrección manual desta entrada implica a eliminación do sinónimo *menopausa* de modo que a relación codificada na entrada se limite á estabelecida entre o lema hipónimo *andropausa* e o sinónimo hiperónimo *climaterio*:

```
<entry>
<form>
<orth>andropausa</orth>
</form>
<sense>
<gramGrp>sf</gramGrp>
<def n="1"><syn><lemma>Climaterio</lemma></syn>.</def>
</sense>
</entry>
```

Estes son algúns exemplos ilustrativos desta revisión e extensión do dicionario guiada polo procesamento das pistas perdidas e realizada a partir do resultado da normalización lingüística e informática do repertorio fonte, unha tarefa de procesamento, ampliación e corrección do material lexicográfico inicial dispoñíbel que, en termos meramente cuantitativos, podemos resumir cos datos da Táboa 3.

Táboa 3. Cobertura inicial do Dicionario de Sinónimos

	Dicionario normalizado	Resultado revisión
Entradas	24573	27176
Acepcións	41926	44541
Sinónimos	159794	172607

5.3. *Resolución das remisións*

Co punto de partida dos resultados do procesamento das pistas perdidas (Táboa 3), levamos a cabo unha segunda fase de expansión e revisión do material do dicionario focalizada no procesamento das remisións lexicográficas, isto é, das chamadas entre entradas do dicionario codificadas coa abreviatura *V* (para *Véxase*). No dicionario de sinónimos, este tipo de remisións serve para substituír o conxunto de sinónimos dunha acepción dunha entrada polo conxunto de sinónimos da acepción correspondente da entrada á que se remite. Por exemplo, na entrada orixinal para o substantivo *abadengo*, a única acepción indicada remite á entrada de *abacial*, lugar onde se atoparían os sinónimos correspondentes a *abadengo*, incluíndo como sinónimo deste o propio termo *abacial*:

```
<entry>
<form>
<orth>abadengo</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n=”1”>V <ref target=”#abacial”>abacial</ref>.</def>
</sense>
</entry>

<entry>
<form>
<orth>abacial</orth>
</form>
<sense>
```

```
<gramGrp>adx</gramGrp>
<def n="1"><syn><lemma>Abadengo</lemma></syn>,
<syn><lemma>conventual</lemma></syn>, <syn><lemma>monacal</lemma></
syn>, <syn><lemma>monástico</lemma></syn>.</def>
</sense>
</entry>
```

Co obxectivo de facilitar o exame e resolución das máis de 4000 remisións incluídas no material lexicográfico de partida, preparamos un programa informático para procesar o dicionario capaz de identificar as remisións nas entradas e substituílas polos sinónimos referidos. Así, para a entrada de *abadengo*, este proceso determina en que acepción da entrada referida *abacial* aparece *abadengo* como sinónimo, e substitúe a remisión por todos os sinónimos da acepción identificada, menos o propio *abadengo* e máis o lema *abacial*. Deste xeito, a nova entrada xerada automaticamente para o lema *abadengo*, coa chamada resolta, será:

```
<entry>
<form>
<orth>abadengo</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n="1"><syn><lemma>Abacial</lemma></syn>,
<syn><lemma>conventual</lemma></syn>, <syn><lemma>monacal</lemma></
syn>, <syn><lemma>monástico</lemma></syn>.</def>
</sense>
</entry>
```

De novo cómpre observar que, como en case todos os procesos automáticos aplicados no eido do procesamento lingüístico, os resultados non son completamente fiábeis e precisan, por tanto, dun traballo posterior de exame e corrección pormenorizada. Algúns dos problemas para a automatización deste procesamento xorden da presenza no dicionario de "remisións perdidas" a entradas inexistentes (concretamente, 35 casos) e das "remisións semiperdidas", onde a entrada que inclúe a remisión non está como sinónimo en ningunha acepción da entrada referenciada (431 casos).

O tratamento das remisións perdidas supón un traballo de revisión humana que implica a creación manual dunha nova entrada para a entrada inexistente e a resolución manual da remisión na entrada que a contén. Por exemplo, no dicionario fonte a entrada de *airoa*, inclúe unha chamada á entrada de *anguía*:

```
<entry>
<form>
<orth>airoa</orth>
</form>
<sense>
<gramGrp>sf</gramGrp>
<def n="1">V <ref target="#anguía">anguía</ref>.</def>
</sense>
</entry>
```

Porén, non existe unha entrada no texto fonte para o lema *anguía*. Para guiar a revisión manual deste problema específico, o programa engade unha etiqueta ("REF NOT FOUND") na entrada orixinal que contribúe a facilitar a postedición dos resultados, consistente na creación da entrada de *anguía* e na corrección da remisión na entrada de *airoa*:

```
<entry>
<form>
<orth>airoa</orth>
</form>
<sense>
<gramGrp>sf</gramGrp>
<def n="1">V <ref target="#anguía" note="REF NOT FOUND">anguía</
ref>.</def>
</sense>
</entry>
```

Como resultado deste traballo asistido de postedición, a entrada orixinal para *airoa* substitúese por unha entrada coa remisión corrixida, ao tempo que se crea unha entrada nova para o termo *anguía* e se introducen algunhas melloras en ambas, concretamente, a anotación da nome científico e a inclusión dun sinónimo adicional *eiroa*:

```
<entry>
<form>
<orth>airoa </orth>
</form>
<sense>
<cit type=”taxonomy”>Anguilla anguilla</cit><gramGrp>sf</gramGrp>
<def n=”1”> <syn><lemma>Anguía</lemma></syn>, <syn><lemma>eiroa</
lemma></syn>.</def>
</sense>
</entry>

<entry>
<form>
<orth>anguía</orth>
</form>
<sense>
<cit type=”taxonomy”>Anguilla anguilla</cit><gramGrp>sf</gramGrp>
<def n=”1”> <syn><lemma>Airoa</lemma></syn>, <syn><lemma>eiroa</
lemma></syn>.</def>
</sense>
</entry>

<entry>
<form>
<orth>eiroa</orth>
</form>
<sense>
<cit type=”taxonomy”>Anguilla anguilla</cit><gramGrp>sf</gramGrp>
<def n=”1”> <syn><lemma>Anguía</lemma></syn>, <syn><lemma>airoa</
lemma></syn>.</def>
</sense>
</entry>
```

No caso das remisións semiperdidas, onde se debe aplicar tamén unha postedición humana, o traballo posterior tamén pode ser guiado por unha etiqueta específica engadida polo programa durante o procesamento das entradas. Por exemplo, na entrada do lema *zazamelo* hai unha chamada á entrada de *zarabeto*:

```
<entry>
<form>
<orth>zazamelo</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n="1">V <ref target="#zarabeto">zarabeto</ref>.</def>
</sense>
</entry>
```

Non obstante, esta remisión non se pode resolver polo programa seguindo estritamente os criterios estabelecidos xa que, aínda que exista unha entrada para *zarabeto*, esta non inclúe o sinónimo *zazamelo* en ningunha das súas acepcións, o que podería implicar un erro na chamada:

```
<entry>
<form>
<orth>zarabeto</orth>
</form>
<sense>
<gramGrp>adx s</gramGrp>
<def n="1"><syn><lemma>Gago</lemma></syn>, <syn><lemma>tartamudo</
lemma></syn>, <syn><lemma>tatabexo</lemma></syn>,
<syn><lemma>tatexo</lemma></syn>, <syn><lemma>tato</lemma></syn>,
<syn><lemma>zarzallo</lemma></syn>.</def>
</sense>
</entry>
```

Igual que no caso das remisións perdidas a entradas inexistentes, o programa etiqueta as entradas xeradas que conteñen remisións semiperdidas cunha indicación ("REF LOST") para facilitar na postedición humana a revisión destas chamadas e a consideración da súa resolución:

```
<entry>
<form>
<orth>zazamelo</orth>
</form>
<sense>
```

```
<gramGrp>adx</gramGrp>
<def n="1">V <ref note="REF LOST" target="#zarabeto">zarabeto</
ref>.</def>
</sense>
</entry>
```

Con estes breves exemplos pretendemos ilustrar o traballo levado nesta fase de revisión e ampliación centrada no procesamento das remisións lexicográficas nas entradas do dicionario e tomando como material de partida os resultados do procesamento das pistas perdidas. A modo de resumo do labor realizado, a Táboa 4 recolle os datos cuantitativos obtidos nas sucesivas fases de construción deste dicionario.

Táboa 4. Cobertura actual do Dicionario de Sinónimos

	Edición normalizada	Revisión intermedia	Cobertura actual
Entradas	24573	27176	27104
Acepcións	41926	44541	44849
Sinónimos	159794	172607	203251

5.4. *Exportación a Galnet*

O caudal léxico do galego codificado así no dicionario de sinónimos serviunos de fonte lexicográfica para enriquecer os *synsets* de Galnet con novas variantes. A metodoloxía utilizada para esta extracción baseouse na coincidencia de formas léxicas entre as variantes dos *synsets* de Galnet e as variantes dos *synsets* do dicionario, considerando que un *synset* do dicionario é o conxunto de formas léxicas formado polos lema e os sinónimos contidos nunha acepción dunha entrada do dicionario. Deste modo, as variantes dun *synset* do dicionario poden converterse en variantes dun *synset* de Galnet se existe coincidencia formal entre algunha das variantes incluídas nestes dous *synsets*.

Consideremos a entrada para o adxectivo *aleuto* no dicionario, formada por unha soa acepción e catro sinónimos:

```
<entry>
<form>
<orth>aleuto</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
<def n="1"><syn><lemma>Agudo</lemma></syn>, <syn><lemma>espelido</
lemma></syn>, <syn><lemma>intelixente</lemma></syn>,
<syn><lemma>listo</lemma></syn>.</def>
</sense>
</entry>
```

O *synset* desta acepción estaría composto por cinco formas léxicas: *{aleuto, agudo, espelido, intelixente, listo}*. Por outra banda, consideremos o *synset* identificado como glg-30-00061885-a no Galnet formado por *{enxeñoso, aleuto}*. Tomados en conxunto, a forma coincidente nos dous *synsets* é *{aleuto}* e, en consecuencia, as novas variantes propostas para o *synset* de Galnet son *{agudo, espelido, intelixente, listo}*.

Para comprobar a eficacia deste procedemento no que consideramos a posibilidade de extracción máis produtiva, deseñamos un programa para extraer as propostas de novas variantes para Galnet cinguíndonos aos *hápax legómena*, isto é, as variantes documentadas unha soa vez, tanto no dicionario de sinónimos coma no Galnet. O programa busca no dicionario as variantes de frecuencia única (sexan sinónimos ou lemas) e comproba se esas formas aparecen tamén como variantes de frecuencia única no Galnet. Nese caso, comproba as coincidencias entre as variantes dos *synsets* correspondentes do dicionario e de Galnet e ofrece como proposta de ampliación de Galnet as variantes do dicionario non coincidentes. A vantaxe de cruzar os *synsets* do dicionario e de Galnet unicamente no caso de compartir un hápax permite limitar as propostas de extracción incorrectas debidas á polisemia. Os resultados ofrecidos polo programa están deseñados co obxectivo de facilitar a revisión humana das súas propostas automáticas, deste xeito:

```
Cruzamento f1*f1 #54: aleuto_2197*glg-30-00061885-a
- Synset do dicionario:
aleuto – agudo – espelido – intelixente - listo
- Synset do Galnet:
aleuto
- Candidatas a variante:
```

```
glg-30-00061885-a      agudo
glg-30-00061885-a      espelido
glg-30-00061885-a      intelixente
glg-30-00061885-a      listo
```

Aplicamos esta técnica á distribución actual de Galnet (Táboa 2) e á revisión intermedia do dicionario (Táboa 3). O número de hápax presentes no dicionario é de 14042, e de 20472 no Galnet. A cantidade de cruzamentos, isto é, de *synsets* do Galnet e do dicionario que comparten un hápax ascende a 1355, cun total de 5890 variantes (lemas ou sinónimos) nos *synsets* do dicionario implicados nos cruzamentos e de 2205 variantes nos respectivos *synsets* de Galnet. A partir destes cruzamentos, a cantidade de variantes do dicionario propostas polo programa como candidatas a nova variante do Galnet foi de 4283.

Na posterior revisión e avaliación humana dos resultados da extracción automática baseada no cruzamento dos hápax compartidos, foron aprobadas un 65% das propostas de variantes incluídas nun 82% dos cruzamentos. Isto indica que un 82% dos cruzamentos foron produtivos. Estes datos confirman, ao noso entender, a validez do experimento realizado e a relevancia dos dicionarios de sinónimos como fonte lexicográfica das redes semánticas no modelo de WordNet.

Co incentivo destes resultados, deseñamos un segundo experimento, aínda en fase de avaliación, consistente no cruzamento das variantes *dislegómena* do dicionario (as que se repiten dúas voltas) cos hápax legómena do Galnet. O total de variantes dislegómena no dicionario é de 8785 e os cruzamentos cos hápax do Galnet ascenden a 1030, cun total de 9078 variantes nos *synsets* do dicionario (6905 distintas) e de 1885 variantes nos *synsets* de Galnet implicados nos cruzamentos. Neste caso, o total de candidatos do dicionario a nova variante do Galnet é de 5570. A revisión destas propostas require aínda unha maior atención que as derivadas do cruzamento dos hápax, xa que neste segundo experimento a polisemia (do lado do dicionario) está garantida. Por exemplo, a entrada do adxectivo *aberrante* implica dous *synsets* no dicionario, un para cada acepción (*{aberrante, estraño, raro}* e *{aberrante, absurdo, erróneo}*):

```
<entry id="aberrante">
<form>
<orth>aberrante</orth>
</form>
<sense>
<gramGrp>adx</gramGrp>
```

```
<def n="1"><syn><lemma>Anormal</lemma></syn>, <syn><lemma>estraño</
lemma></syn>, <syn><lemma>raro</lemma></syn>.</def>
<def n="2"><syn><lemma>Absurdo</lemma></syn>, <syn><lemma>erróneo</
lemma></syn>.</def>
</sense>
</entry>
```

A forma *aberrante* non aparece en máis *synsets* do dicionario, de modo que constitúe un hápax dislegómena desta obra. A mesma forma aparece unha soa vez no Galnet, no synset glg-30-00193367-a formado polas variantes *{asombroso, abraiante, arrepiante, desacougante, espantoso, inhumano, monstruoso, terrible, aberrante}*. De novo, a saída do programa ofrece unha presentación das variantes candidatas que trata de facilitar o traballo posterior de selección por parte do equipo de lexicografía:

```
Cruzamento f2*f1 #4: aberrante_216/aberrante_217*glg-30-00193367-a
- Synsets do dicionario:
[216 aberrante] aberrante – anormal – estraño - raro
[217 aberrante] aberrante – absurdo - erróneo
- Synset do Galnet: asombroso – abraiante – arrepiante – desacou-
gante – espantoso – inhumano – monstruoso – terrible - aberrante
- Candidatas a variante:
glg-30-00193367-a      anormal
glg-30-00193367-a      estraño
glg-30-00193367-a      raro
glg-30-00193367-a      absurdo
glg-30-00193367-a      erróneo
```

Como se pode observar nestes exemplos, tanto no cruzamento dos hápax legómena coma no alicerzado nos hápax dislegómena do dicionario, o traballo posterior de selección das formas candidatas a variante de Galnet é vagaroso e esixe unha grande concentración e coñecemento da lingua. Porén, pensamos que o esforzo humano necesario paga a pena xa que nos permite, por unha banda, a ampliación controlada dun recurso imprescindíbel para o procesamento da lingua galega usando unha fonte lexicográfica de calidade e, por outra, a reflexión teórica sobre os diversos conceptos de sinonimia aplicados na elaboración dos repertorios léxicos tradicionais e modernos, nos dicionarios de sinónimos e nas redes semánticas (Gómez et al. 2013).

6. Conclusións

Neste artigo presentamos o estado actual do proxecto Galnet do Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, dirixido á construción da versión galega do WordNet. Repasamos os diferentes procesos de extracción e as distintas fontes lexicográficas utilizadas ata o momento na construción deste recurso, focalizando a exposición nas tarefas de procesamento dun dicionario de sinónimos.

A elaboración do Galnet atópase aínda nas súas primeiras fases. Porén, xa podemos ofrecer algúns resultados interesantes da investigación en curso. Así, o Dicionario de Sinónimos electrónico empregado como fonte lexicográfica de Galnet pode ser xa consultado na web do Grupo TALG[12] e están en fase de elaboración outras interfaces de consulta ao dicionario para dispositivos móbiles, alén de estar prevista a súa distribución aberta con licenza libre GPL.

Por outro lado, a distribución actual do Galnet pode explorarse tamén nas páxinas do Grupo TALG mediante unha interface web específica de consulta do Galnet[13] ou en conxunción con outros recursos léxicos e textuais do galego a través da plataforma RILG (Recursos Integrados da Lingua Galega)[14]. A mesma distribución do Galnet está incluída no WordNet plurilingüe accesíbel para consulta e descarga no Multilingual Central Repository[15], onde o Galnet convive co WordNet do inglés e coas versións española, catalá e vasca desta rede semántica. Así mesmo, como vimos anteriormente, a interface de consulta do VisuGal[16] permite unha visualización en forma de grafo do WordNet para estas cinco linguas. Por último, a interface de consulta do Open Multi-Lingual WordNet[17] permite a consulta conxunta da rede léxico-semántica de Galnet cos léxicos WordNet elaborados para outras 20 linguas.

12 http://sli.uvigo.es/sinonimos/
13 http://sli.uvigo.es/galnet/
14 http://sli.uvigo.es/RILG/
15 http://adimen.si.ehu.es/web/MCR/
16 http://tec.citius.usc.es/VISUGAL/
17 http://casta-net.jp/~kuribayashi/multi/

Bibliografía

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

Agirre, Eneko, Iñaki Alegria, German Rigau & Piek Vossen (2007), "MCR for CLIR", *Procesamiento del Lenguaje Natural*, vol. 38, pp. 3-15.

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Antoni Oliver & German Rigau (2008), "Consistent Annotation of EuroWordNet with the Top Concept Ontology", en *Proceedings of the 4th Global WordNet Conference*, Szeged: GWN, s. p.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini & Emanuele Pianta (2004), "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing", en *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, Xenebra, ACL, pp. 101-108.

Elberrichi, Zakaria, Abdelattif Rahmoun & Mohamed Amine Bentaalah (2008), "Using WordNet for Text Categorization", *The International Arab Journal of Information Technology*, vol. 5, nº 1, pp. 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 (ed.), *La lexicografía en su dimensión teórica*, Málaga: Universidad de Málaga, pp. 201-220.

Ferrández, Sergio, Antonio Ferrández, Sandra Roger & Pilar López–Moreno (2007), "Búsqueda de respuestas bilingüe basada en ILI, el sistema BRILI", *Procesamiento del Lenguaje Natural*, vol. 38, pp. 27-33.

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

Gómez Guinovart, Xavier, Eva Díaz Rodríguez & Alberto Álvarez Lugrís (2008), "Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego", *Viceversa: Revista Galega de Tradución*, vol. 14, pp. 71-87.

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

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

González, Manuel & Antón Santamarina (2004), *Vocabulario Ortográfico da Lingua Galega (VOLGa)*, A Coruña/Santiago: Real Academia Galega/Instituto da Lingua Galega.

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.

González Agirre, Aitor & German Rigau (2013), "Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository", *Linguamática*, vol. 5, nº 1, pp. 13-28.

Isahara, Hitoshi, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama & Kyoko Kanzaki (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, pp. 298-302.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller (1990), "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, vol. 3, nº 4, pp. 235-244.

Oliver, Antoni (2012), "WN-Toolkit: un toolkit per a la creació de WordNets a partir de diccionaris bilingües", *Linguamática*, vol. 4, nº 2, pp. 93-101.

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*, vol. 47, pp. 293-300.

Ordan, Noam & Shuly Wintner (2007), "Hebrew WordNet: a Test Case of Aligning Lexical Databases Across Languages", *International Journal of Translation*, vol. 19, nº 1, pp. 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, Luisa Bentivogli & Christian Girardi (2002), "MultiWordNet: Developing an Aligned Multilingual Database", en *Proceedings of the First International Conference on Global WordNet*, Mysore: GWN, pp. 21-25.

Plaza, Laura, Alberto Díaz & Pablo Gervás (2010), "Automatic summarization of news using WordNet concept graphs", *IADIS International Journal on Computer Science and Information Systems*, vol. 5, nº 1, pp. 45-57.

Pociello, Elisabete, Eneko Agirre & Izaskun Aldezabal (2011), "Methodology and Construction of the Basque WordNet", *Language Resources and Evaluation*, vol. 45, nº 2, pp. 121-142.

Real Academia Galega (2004), *Normas ortográficas e morfolóxicas do idioma galego*, Vigo: Galaxia.

TEI Consortium (ed.) (2014), *P5: Guidelines for Electronic Text Encoding and Interchange*, dispoñíbel no enderezo http://www.tei-c.org/release/doc/tei-p5-doc/en/html/, consultado o 24/1/2014.

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

Zhao, Feng, Fei Fang, Fengwei Yan, Hai Jin & Qin Zhang (2012), "Expanding approach to information retrieval using semantic similarity analysis based on WordNet and Wikipedia", *International Journal of Software Engineering and Knowledge Engineering*, vol. 22, nº 2, pp. 305-322.