Rev Gal Filo, 2015, 16: 169-201
DOI 10.17979/rgf.2015.16.0.1383

Revista GaLEGa de filoloxía

Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas

Miguel Anxo Solla Portela
Xavier Gómez Guinovart
Universidade de Vigo

Data de recepción: 20/06/2015 | Data de aceptación: 29/07/2015

Resumo:

Neste artigo amosaremos as diferentes metodoloxías e os recursos que se utilizaron durante o proceso de elaboración de Galnet, a versión galega de WordNet. Presentarase tamén a ferramenta Termonet, desenvolvida para a consulta e verificación en corpus de léxicos de especialidade extraídos de WordNet. Por último, describirase a experimentación orientada ao deseño automático de áreas semánticas mediante a explotación das relacións léxico-semánticas de WordNet.

Palabras chave:

WordNet, lexicografía computacional, adquisición de información léxica, terminoloxía computacional, recursos lingüísticos.

Sumario:

1. Introdución. 2. O proxecto Galnet. 2.1. Alicerces. 2.2. Primeira distribución. 2.3. Novas ampliacións coa ferramenta WN-Toolkit. 2.4. Ampliación a partir do dicionario de sinónimos. 2.5. Ampliación fraseolóxica e terminolóxica. 2.6. Estado actual de Galnet. 3. Relacións léxico-semánticas e terminoloxía. 3.1. Termonet. 3.1.1. Recursos. 3.1.2. Extracción de variantes nun ámbito de especialidade. 3.1.3. Verificación en corpus especializados. 3.2. Epinónimos. 3.2.1. Categorización de synsets nominais epinónimos. 3.2.2. Emparellamento de cada synset de WordNet co epinónimo da área semántica. 4. Conclusións e liñas futuras de investigación. Referencias bibliográficas.

Galnet: the Galician WordNet. Applications in the fields of lexicology and terminology

Abstract:

In this article we will present the different methodologies and resources which were used during the elaboration process of GalNet, the Galician version of WordNet. We will also present Termonet, a tool developed to consult and verify specialty lexicons extracted from WordNet in corpora. Finally, we will describe an experiment oriented to the automatic shaping of semantic areas by using lexical-semantic relationships in WordNet.

Key words:

WordNet, computational lexicography, lexical acquisition, computational terminology, language resources.

Contents:

1. Introduction. 2. The Galnet project. 2.1. Foundation. 2.2. First distribution. 2.3. Further expansion with the WN-Toolkit. 2.4. Expansion from a synonyms dictionary. 2.5. Expansion in phraseology and terminology. 2.6. Current state of Galnet. 3. Lexical-semantic relations and terminology. 3.1. Termonet. 3.2. Epinonyms. 3.2.1. Categorization of nominal synsets epinonyms. 3.2.2. WordNet synsets pairing with their semantic area epinonyms. 4. Conclusion and future work. References.

1. Introdución

Neste artigo[1] revisaremos o avance do proxecto Galnet do Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo, centrado na construción da versión galega de WordNet. Trátase dun proxecto que se atopa aínda en fase de desenvolvemento, mais do que xa se obtiveron resultados interesantes e útiles nos ámbitos da lexicoloxía, da semántica e do tratamento automático da lingua galega. Nos seguintes apartados describiremos os trazos xerais do proxecto, a metodoloxía seguida até o momento para a construción do recurso, e algunhas das súas aplicacións no ámbito de investigación en ontoloxías e no traballo terminolóxico.

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 e Miller et al. 1990). Os nós da rede están formados por nomes, verbos, adxectivos e adverbios 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, os *synsets* están conectados por relacións léxico-semánticas. No caso dos substantivos, algunhas das relacións máis frecuentes representadas no WordNet son as de hiperonimia/hiponimia e as de holonimia/meronimia; no caso dos adxectivos, as de antonimia e as de cuasisinonimia; no caso dos adverbios, as de antonimia e as derivativas; e no caso dos verbos, as de implicación, hiperonimia/hiponimia, causatividade e oposición.

WordNet foi concibido orixinalmente para a lingua inglesa e, aínda que hoxe existen versións do WordNet en moitas linguas, o do inglés segue sendo arestora a versión de referencia e a máis desenvolvida. 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 206.941 lemas ou variantes sinonímicas (155.287 das cales son formas únicas non homógrafas) agrupadas en 117.659 grupos de sinónimos ou *synsets*. Na Figura 1, pódese consultar unha representación textual dun fragmento do WordNet do inglés obtida coa utilidade de consulta ofrecida pola Universidade de

1 Esta investigación realízase no marco 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)" financiado polo Ministerio de Economía y Competitividad, TIN2012-38584-C06-04.

Princeton[2]; e na Figura 2, unha visualización obtida co VisuGal[3] dunha parte dese fragmento en forma de grafo.

- S: (n) **finger** (any of the terminal members of the hand (sometimes excepting the thumb)) *"her fingers were long and thin"*
  - *direct hyponym / full hyponym*
    - S: (n) thumb, pollex (the thick short innermost digit of the forelimb)
    - S: (n) index, index finger, forefinger (the finger next to the thumb)
    - S: (n) ring finger, annualry (the third finger (especially of the left hand))
    - S: (n) middle finger (the second finger; between the index finger and the ring finger)
    - S: (n) little finger, pinkie, pinky (the finger farthest from the thumb)
  - *part meronym*
    - S: (n) pad (the fleshy cushion-like underside of an animal's foot or of a human's finger)
    - S: (n) fingertip (the end (tip) of a finger)
    - S: (n) fingernail (the nail at the end of a finger)
    - S: (n) knuckle, knuckle joint, metacarpophalangeal joint (a joint of a finger when the fist is closed)
  - ***direct hypernym** / inherited hypernym / sister term*
  - *part holonym*
    - S: (n) hand, manus, mitt, paw (the (prehensile) extremity of the superior limb) *"he had the hands of a surgeon"; "he extended his mitt"*
  - *derivationally related form*
- S: (n) **finger**, fingerbreadth, finger's breadth, digit (the length of breadth of a finger used as a linear measure)
- S: (n) **finger** (one of the parts of a glove that provides covering for a finger or thumb)

Figura 1. Representación textual dun fragmento do WordNet do inglés

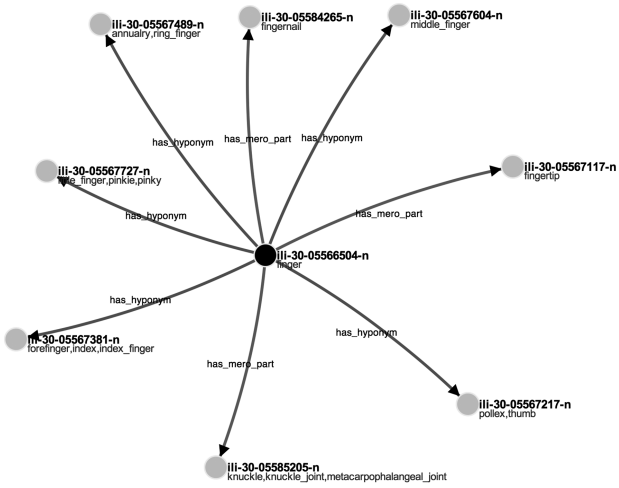

Figura 2. Representación visual dun fragmento do WordNet do inglés

2 http://wordnet.princeton.edu

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

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 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 euskara (Pociello et al. 2011). The Global WordNet Association mantén unha listaxe de léxicos WordNet desenvolvidos por linguas na súa páxina web[4]. Tamén se pode acceder a unha boa variedade de léxicos WordNet para distintas linguas a través da páxina do proxecto Open Multilingual Wordnet[5].

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 grande utilidade en aplicacións das tecnoloxías lingüísticas que precisan o procesamento plurilingüe da linguaxe, como a tradución automática (Vintar et al. 2012), 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 EuroWordNet 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.

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

4 http://www.globalwordnet.org

5 http://compling.hss.ntu.edu.sg/omw/

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 proxecto Galnet está tamén coordinado co desenvolvemento do WordNet do portugués que se está a levar a cabo no proxecto PULO (Ontologia Lexical Unificada para o Português)[6] no Centro de Estudos Humanísticos da Universidade do Minho (Simões / Gómez Guinovart 2014).

O marco de desenvolvemento no que se integra o Galnet é o do Multilingual Central Repository[7] (MCR) (González et al. 2012 e González / Rigau, 2013), unha plataforma 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 *synsets* categorizados na xerarquía de dominios IRST e nas ontoloxías SUMO e *Top Concept Ontolog*y. Na Figura 3, inclúese a modo de exemplo unha visualización con VisuGal dunha sección da rede semántica plurilingüe en construción no MCR.

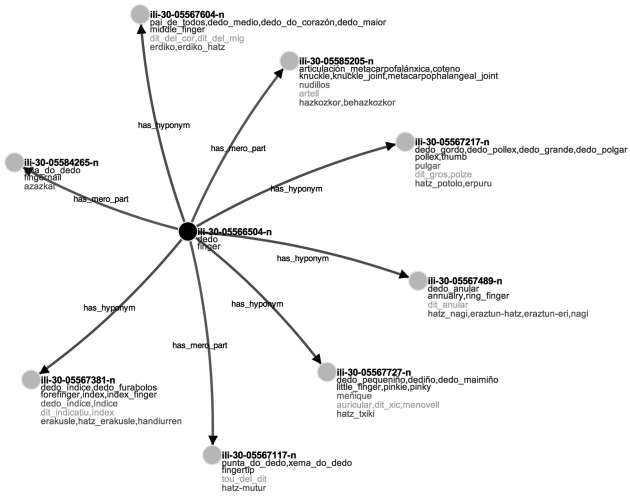

Figura 3. Representación visual dun fragmento do MCR

6 http://wordnet.pt/

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

A interface específica deseñada para a consulta de Galnet[8], ilustrada na Figura 4, amplía as funcionalidades do MCR coa consulta das solucións portuguesas xeradas no proxecto PULO e cunha nova categorización semántica de orientación terminolóxica baseada en *epinónimos* que será explicada máis adiante na sección 3 deste artigo.

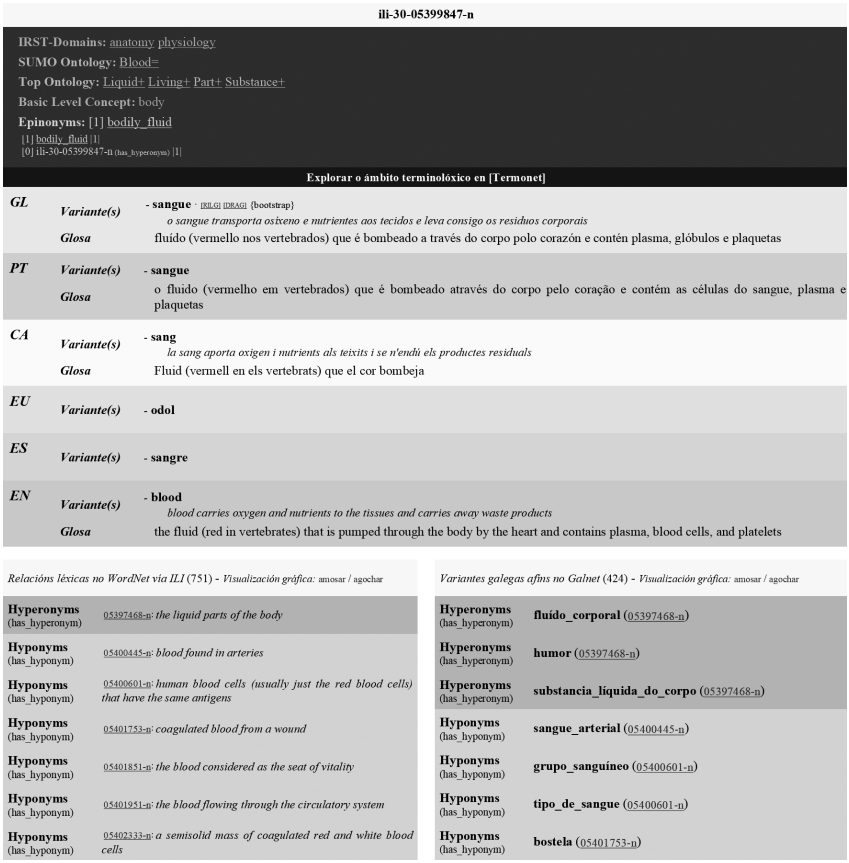

Figura 4. Interface de consulta de Galnet

Nos seguintes apartados desta sección describiremos a metodoloxía e as ferramentas empregadas na construción do Galnet nas sucesivas etapas da construción do recurso.

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

2.1. Alicerces

Os obxectivos da primeira fase na construción do Galnet foron, en primeiro lugar, elaborar un conxunto de *synsets* básicos para a operatividade do recurso en 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.

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[9] dos BLC do WordNet 3.0, sen incluír na adaptación nin as glosas nin os exemplos incluídos nos *synsets* correspondentes da lingua inglesa.

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[10], 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)[11] e na base de datos terminolóxica da Termoteca[12].

9 http://adimen.si.ehu.es/web/BLC/

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

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

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

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* (1.333 variantes léxicas) dos BLC de categoría nominal, 616 *synsets* (1.416 variantes) dos BLC de categoría verbal, 2.014 *synsets* (3.550 variantes) do ficheiro lexicográfico de nomes relacionados coas partes do corpo, 2.983 *synsets* (4.300 variantes) do ficheiro lexicográfico de nomes de substancias, e 3.114 *synsets* (4.864 variantes) do conxunto de adxectivos de tipo xeral incluídos en WordNet 3.0. As variantes galegas procedentes desta fase inicial do proxecto poden ser examinadas seleccionando *bootstrap* como experimento no formulario da interface web de consulta pública de Galnet.

	WordNet 3.0		**Galnet (*bootstrap*)**	
	variantes	***synsets***	***variantes***	***synsets***
Nomes	146.312	82.115	9.183	5.646
Verbos	25.047	13.767	1.416	616
Adxectivos	30.002	18.156	4.864	3.114
Adverbios	5.580	3.621	0	0
TOTAL	206.941	117.659	15.463	9.376

Táboa 1. Cobertura inicial de Galnet (versión *bootstrap*)

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% respecto da 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.

2.2. Primeira distribución

Na segunda fase de desenvolvemento do proxecto Galnet, utilizamos a ferramenta WN-Toolkit (Oliver 2012) para amplialo 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 Guinovart et al. 2012 e Álvarez Lugrís / Gómez Guinovart 2014). 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 equipo de revisores. O resultado da extracción automática, revisado manualmente, serviu para ampliar o Galnet con 11.677 novas variantes e 9.936 novos *synsets*, isto é, ao duplo da extensión obtida na primeira fase.

As técnicas de extracción aplicáronse de xeito secuencial e ordenado, dándolle 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:

- 2.945 variantes nominais pluriléxicas do inglés coas iniciais de todas as palabras en maiúscula e que figuran na Wikipedia (variantes marcadas con *capitals* como nome de experimento);
- 2.483 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 (experimento *dic-moneng-1trad-uni*);
- 1.529 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 (experimento *dic-moneng-1trad-mul*);
- 1.818 variantes nominais e adxectivas do dicionario cuxo lema inglés aparece nun único *synset* do WordNet e que teñen como tradución máis dunha palabra galega (experimento *dic-moneng-multitrad*);
- 2.971 variantes nominais enlazadas do galego ao inglés na Galipedia e que non estaban no Galnet (experimento *galipedia*).

Todas as variantes galegas procedentes desta segunda fase de proxecto poden visualizarse seleccionando *capitals*, *dic-moneng* e *galipedia* na interface web de consulta pública do 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 para descarga como parte do MCR 3.0[13], contén os datos do repertorio léxico neste estado de desenvolvemento do proxecto.

	Galnet no MCR 3.0	
	variantes	***synsets***
Nomes	18.949	14.285
Verbos	1.416	612
Adxectivos	6.773	4.415
Adverbios	0	0
TOTAL	27.138	19.312

Táboa 2. Distribución de Galnet no MCR 3.0

2.3. Novas ampliacións coa ferramenta WN-Toolkit

A partir desta primeira distribución do Galnet no MCR 3.0, realizamos dous novos experimentos de expansión coa ferramenta do WN-Toolkit (Oliver 2012). No primeiro (Gómez Guinovart / Oliver 2014), usamos como fontes para a extracción de novas variantes o dicionario inglés-galego do tradutor Apertium[14], o Galizionario (versión galega do Wiktionary)[15], o dicionario inglés-galego de Babelnet 2.0[16], un subconxunto dos corpus paralelos CLUVI[17] (concretamente, o corpus Unesco galego-español de divulgación científica, o corpus xurídico Lega galego-español, o Corpus Consumer Eroski español-galego de información sobre consumo e o corpus literario Tectra inglés-galego) e o Corpus SemCor[18] inglés-galego traducido automaticamente con Google Translate (Oliver / Climent 2014). Como resultado deste experimento, incorporáronse no Galnet 4.499 novas variantes en 4.343 *synsets* que se poden examinar na interface de Galnet mediante a consulta polo experimento *wnt7* e cuxa distribución por categorías se recolle na Táboa 3.

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

14 http://sourceforge.net/projects/apertium/

15 http://gl.wiktionary.org

16 http://babelnet.org

17 http://sli.uvigo.es/CLUVI/

18 http://www.gabormelli.com/RKB/SemCor_Corpus

	Expansión WN-Toolkit/*wnt7*	
	variantes	***synsets***
Nomes	3.029	2.934
Verbos	542	520
Adxectivos	664	653
Adverbios	264	236
TOTAL	4.499	4.343

Táboa 3. Resultados da expansión co WN-Toolkit/*wnt7*

A partir desta ampliación, coa mesma ferramenta para a extensión do recurso, realizamos un segundo experimento usando agora como fontes para a extracción de variantes seis recursos léxicos bilingües co galego procedentes do dicionario plurilingüe OmegaWiki[19], a base de datos de topónimos GeoNames[20], o inventario de especies Wikispecies[21] e as versións actualizadas do Galizionario, do dicionario de Apertium e da Wikipedia. Como resultado, engadíronse ao Galnet 3.450 novas variantes en 2.552 *synsets* que se poden examinar na interface de Galnet mediante a procura polo experimento *wnt6dic* e das que se recolle a distribución por categorías na Táboa 4.

	Expansión WN-Toolkit/*wnt6dic*	
	variantes	***synsets***
Nomes	3.033	2.207
Verbos	136	112
Adxectivos	136	115
Adverbios	145	118
TOTAL	3.450	2.552

Táboa 4. Expansión co WN-Toolkit/*wnt6dic*

2.4. Ampliación a partir do dicionario de sinónimos

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 un formato dixital normalizado dun dicionario de sinónimos tradicional do galego publicado en papel e xa descatalogado (Noia et al. 1997). Como

19 http://www.omegawiki.org

20 http://www.geonames.org

21 http://species.wikimedia.org

base da conversión, contamos co conxunto de ficheiros MS-Word elaborados polos autores da obra orixinal previos á corrección editorial e maquetación da obra.

Os traballos de elaboración deste novo dicionario electrónico de sinónimos efectuáronse en tres fases. A primeira tarefa consistiu en converter a información textual desestruturada dos ficheiros de Word nunha base de datos lexicográfica normalizada (Gómez Guinovart / Simões 2013). A consecución desta tarefa de conversión non estivo exenta de dificultades, debido principalmente aos erros de formato e outras inconsistencias atopadas nos ficheiros orixinais. Moitos dos erros da conversión automatizada tiveron que ser revisados manualmente. Co dicionario xa regularizado en formato dixital, 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. O dicionario fonte co que traballamos foi redactado en 1997, seguindo a normativa de 1982. Por tanto, a preparación do dicionario comportou a corrección do texto segundo os criterios normativos vixentes na actualidade. Esta revisión normativa representou un volume de traballo moi elevado, xa que nos obrigou a revisar o texto do dicionario na súa totalidade. Unha vez rematada a normalización informática e lingüística do texto, preparamos unha interface web para facilitar a súa consulta pública e emprendemos unha dilatada fase de ampliación e revisión lexicográfica que incluíu a expansión dos sinónimos que dan lugar a pistas perdidas no dicionario e a comprobación e resolución das remisións lexicográficas presentes nas entradas (Gómez Guinovart 2014). A modo de resumo cuantitativo do labor realizado, a Táboa 5 recolle os datos por entradas, acepcións e sinónimos deste dicionario.

	Edición normalizada	**Extensión actual**
	variantes	***synsets***
Entradas	24.573	27.104
Acepcións	41.926	44.849
Sinónimos	159.794	203.251

Táboa 5. Extensión do *Dicionario de sinónimos do galego*

A primeira versión do novo *Dicionario de sinónimos do galego* publicouse na web do Seminario de Lingüística Informática[22] en 2013, tratándose do primeiro e único dicionario electrónico do galego dentro desta categoría de dicionarios. O dicionario dispón tamén de aplicación para dispositivos móbiles, dispoñíbel desde finais de 2014 no Google Play[23] para Android e na App Store[24] para Apple iOS.

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

23 https://play.google.com/store/apps/details?id=net.ayco.sinonimosgal

24 https://itunes.apple.com/us/app/sinonimos-do-galego/id940045971?l=es&ls=1&mt=8

O caudal léxico do galego codificado 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 lemas e os sinónimos contidos nunha acepción dunha entrada do dicionario. Deste xeito, e con moitas matizacións, as variantes dun *synset* do dicionario poden converterse en variantes dun *synset* de Galnet, tras unha revisión e validación manual, se existe coincidencia formal entre algunha das variantes incluídas nestes dous *synsets*.

Porén, axiña descubrimos as dificultades de acadar mediante esta metodoloxía uns índices de precisión que permitisen a validacion manual dos resultados da extracción nun tempo razoábel. As causas desta dificultade radican principalmente no distinto concepto de sinonimia utilizado por cada un destes dous recursos, moito máis estrito e delimitado pola glosa no WordNet (Gómez Clemente et al. 2013), moito máis laxo e abranguente doutras relacións semánticas no dicionario de sinónimos tradicional.

Deseñáronse varios experimentos de extracción de variantes do dicionario de sinónimos co obxectivo de optimizar a súa precisión e minimizar a fase de revisión previa á súa incorporación a Galnet. Na primeira aproximación ao problema (Gómez Guinovart 2014), para comprobar a eficacia deste método 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, ás 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. Na posterior revisión e avaliación humana dos resultados da extracción automática baseada no cruzamento dos hápax compartidos, foron aprobadas un 65% das 4.283 propostas de novas variantes, o que confirmaría a validez do experimento realizado e a relevancia dos dicionarios de sinónimos como fonte lexicográfica das redes semánticas no modelo de WordNet.

Con todo, co fin de probar a diminuír a carga da revisión lexicolóxica aumentando a precisión da extracción automática, deseñouse un segundo experimento (Solla Portela/Gómez Guinovart 2014) cruzando as acepcións do dicionario que compartisen tres sinónimos ou máis con tres variantes no mesmo *synset* de Galnet,

o que forneceu 6.335 candidaturas. Unha avaliación prospectiva de 100 destas candidatas a variantes confirmou a baixa precisión desta metodoloxía, pois só se consideraron válidas para Galnet un 35% das formas analizadas. Así mesmo, durante a revisión das formas candidatas, detectouse que a precisión diminuía conforme se incrementaba o índice de dispersión semántica; isto é, que cando existe un número elevado de sinónimos na mesma acepción do dicionario, as candidaturas propostas para incorporarse a Galnet son menos acertadas. Como froito desta observación repetiuse o experimento eliminando dos cruzamentos as acepcións do dicionario con cinco ou máis sinónimos. O resultado foron 856 formas candidatas a variantes das que foron aprobadas un 60%, cunha precisión que consideramos asumíbel para una revisión humana eficaz.

A partir destes dous experimentos, tratamos de axustar aínda máis con novos parámetros a metodoloxía de extracción utilizada de xeito que nos permitise maximizar os esforzos (sempre demasiado escasos) dedicados á revisión humana dos resultados, aumentando o máis posíbel a precisión dos resultados sen diminuír a súa cobertura a límites inútiles (Gómez Guinovart / Solla Portela 2014). Como resultado parcial destes experimentos, engadíronse até agora ao Galnet 1.587 novas variantes en 665 *synsets* que se poden examinar na interface de Galnet consultando polos experimentos *dicsingal* e *singalnet* e dos que recollemos a súa distribución por categorías na Táboa 6.

	Expansión con sinónimos	
	variantes	***synsets***
Nomes	620	296
Verbos	384	178
Adxectivos	568	186
Adverbios	15	5
TOTAL	1.587	665

Táboa 6. Expansión co *Dicionario de sinónimos do galego*

2.5. Ampliación fraseolóxica e terminolóxica

No ámbito da fraseoloxía, ampliamos a riqueza de Galnet, fundamentalmente na categoría verbal, a partir do repertorio de 850 locucións verbais recollidas no traballo de Álvarez de la Granja (2003). Estas locucións foron revisadas individualmente e aceptadas ou rexeitadas considerando a súa asignación a un *synset* de Galnet. As locucións seleccionadas, na medida do posíbel, foron complementadas con novas locucións de carácter sinonímico procedentes de fontes lexicográficas. Como

resultado desta ampliación, incorporáronse a Galnet 1.773 variantes repartidas en 351 *synsets*, que poden ser consultadas mediante a selección do experimento *fralnet* na interface web de consulta de Galnet.

No tocante ao léxico terminolóxico, tratamos de completar con variantes galegas os synsets de Galnet pertencentes ao dominio da medicina empregando como fontes de referencia o Corpus Técnico do Galego, a base de datos terminolóxica da Termoteca e os repertorios lexicográficos da Real Academia de Medicina e Cirurxía de Galicia (2002) e do Servizo de Normalización Lingüística da USC (Rodríguez Río 2008). Os resultados desta ampliación ascenden a 2.273 variantes en 1.439 *synsets* nominais, que poden ser revisadas na interface de Galnet seleccionando o experimento *medicalnet* na consulta.

2.6. Estado actual de Galnet

Outros experimentos de ampliación en curso, en fases menos avanzadas de execución, inclúen a incorporación ao Galnet das variantes do dominio terminolóxico da economía (con 162 variantes xa introducidas no momento actual), a adaptación ao galego das variantes portuguesas procedentes do proxecto PULO sen versión galega (con 958 variantes xa adaptadas), a extracción e verificación das propostas galegas recompiladas no Extended Open Multilingual Wordnet[25] (421 variantes xa incorporadas) e a inclusión das variantes galegas novas procedentes da elaboración do corpus SensoGal[26] (768 variantes xa engadidas ao Galnet), un corpus paralelo inglés-galego en desenvolvemento, anotado semanticamente con referencia a Galnet e aliñado a nivel de frase e de palabra co corpus SemCor da lingua inglesa. Todas estas novas variantes galegas obtidas nestes experimentos en curso poden ser consultadas indicando como experimento *econonet*, *porgalnet*, *xomwn* e *semcor*, respectivamente, na interface de consulta do recurso.

A Táboa 7 recolle o estado actual de desenvolvemento de Galnet, na súa versión 3.0.15, dispoñíbel para consulta a través da interface web[27] do noso grupo de investigación. Cómpre indicar que a distribución oficial do recurso, sendo de vital importancia para a súa difusión e uso, non deixa de ser unha versión "conxelada" dos datos fornecidos para o galego, debéndose acudir sempre á interface propia de Galnet na procura dos datos máis actualizados.

25 http://compling.hss.ntu.edu.sg/omw/summx.html

26 http://sli.uvigo.es/SensoGal/

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

	WordNet 3.0		Galnet 3.0.15	
	variantes	***synsets***	***variantes***	***synsets***
Nomes	146.312	82.115	28.911	21.220
Verbos	25.047	13.767	4.839	1.900
Adxectivos	30.002	18.156	8.472	5.274
Adverbios	5.580	3.621	626	459
TOTAL	206.941	117.659	42.848	28.853

Táboa 7. Extensión actual de Galnet (versión 3.0.15)

Galnet, na súa versión máis recente, tamén está dispoñíbel para consulta, xunto con outros importantes recursos léxicos e textuais, na plataforma RILG de recursos integrados da lingua galega[28].

3. Relacións léxico-semánticas e terminoloxía

> *Un concepte forma part d'un conjunt estructurat de conceptes en referència als quals adquireix el seu valor. Així doncs, un concepte només ho és en relació a un determinat camp conceptual* (Cabré i Castellví 1992:192).

WordNet concibiuse orixinariamente desde a perspectiva da psicolexicoloxía e estruturouse a través das relacións semánticas entre os *synsets* en función das diferentes categorías gramaticais. En palabras dos seus impulsores: "WordNet is organized by semantic relations" (Miller et al. 1990). Manifesta, polo tanto, moitas concomitancias con aspectos metodolóxicos da terminoloxía no que respecta á tipoloxía e á estruturación dos conceptos (Sager 1990). Relacións semánticas semellantes utilízanse en multitude de repertorios terminolóxicos na actualidade, como SNOMED Clinical Terms[29] ou o propio banco de datos terminolóxicos da Universidade de Vigo, a Termoteca.

A partir da observación destas características similares, xermolou a idea de reorientar as relacións presentes en WordNet cara a estratexias de exploración terminolóxica da rede léxico-semántica. Comezamos daquela unha revisión das relacións léxico-semánticas do Galnet para estudar as posibilidades de tratar de tecer un *campo xerárquico* (Cabré i Castellví 1992) entre os nós conceptuais de WordNet, conscientes da dificultade de partirmos dunha extensión de *synsets* moito máis xenérica do que é habitual no labor terminolóxico, onde se adoita traballar cun conxunto de conceptos máis precisos.

28 http://sli.uvigo.es/RILG/

29 http://www.ihtsdo.org/snomed-ct/

En WordNet, ademais, conviven *synsets* de catro categorías gramaticais diferentes (83.246 nominais, 18.156 adxectivais, 13.885 verbais e 3.621 adverbiais) e, pese á cantidade de substantivos, o número de *synsets* que a priori non encaixan como conceptos terminolóxicos é considerábel. De forma xenérica, o tratamento que se ideou para os adxectivos e adverbios foi intentar vinculalos, sempre que for posíbel, cun *synset* terminoloxicamente conceptual a través de relacións léxicas transcategoriais; por exemplo, vincular os conceptos de *cirúrxico* e *cirurxicamente* co de *cirurxía*. Este procedemento presenta abondosas limitacións, primordialmente porque en moitos casos WordNet non contén esta relación léxica, mais tamén porque cando estas relacións de raíz morfolóxica están codificadas en WordNet son relacións entre as variantes en lingua inglesa, e non entre os nós conceptuais. Cando este tratamento non for posíbel, empregaríanse outras relacións semánticas, como a cuasisinonimia ou a antonimia pese á súa toxicidade, pois optouse por unha estratexia que incluíse os nós conceptuais da rede léxico-semántica de calquera categoría gramatical.

Neste empeño, desde un primeiro momento asumiuse tamén a necesidade de ampliar o enfoque cognitivo do estudo das relacións cunha verificación empírica dos resultados, examinando, desde a perspectiva da terminoloxía comunicativa, a presenza das variantes en corpus textuais de linguaxes de especialidade.

Con estas premisas, revisáronse as relacións procedentes de WordNet e reagrupáronse en Galnet do modo que se ilustra na Táboa 8 (onde *idg* representa o identificador do grupo e *idr* o identificador da relación) e se exemplifica na Figura 5.

grupo	idg	relación	relación inversa	idr
Antonyms	0	near_antonym	near_antonym	34
Synonyms	1	near_synonym	near_synonym	33
Hyperonyms	2	has_hyponym	has_hyperonym	12
Hyponyms	3	has_xpos_hyponym	has_xpos_hyperonym	21
Holonyms Meronyms	4 5	has_holo_part	has_mero_part	8
		has_holo_member	has_mero_member	7
		has_holo_madeof	has_mero_madeof	6
Related	6	see_also_wn15		49
		related_to	related_to	64
Verbs	7	causes	is_caused_by	2
		has_subevent	is_subevent_of	19
		verb_group	verb_group	52
Domain	8	category_term	category	63
		region_term	region	66
		usage_term	usage	68
Glosses	9	rgloss	gloss	61

Táboa 8. Grupos e relacións léxico-semánticas en Galnet

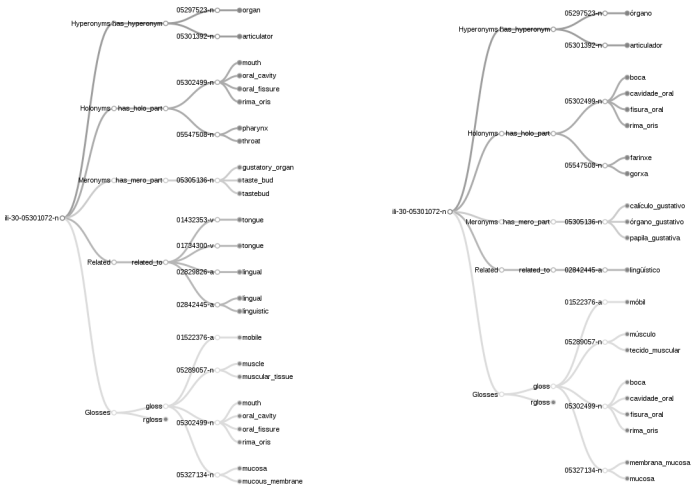

Figura 5. Exemplo gráfico de agrupacións e relacións

Da experimentación cos novos agrupamentos das relacións léxico-semánticas entre os synsets de WordNet, emanaron unha serie de termos precisos para a análise do seu comportamento e para a descrición das súas características que indicamos deseguido:

- *Reversibilidade*, *dirección* (*horizontal* / *vertical*) e *sentido* (*ascendente* / *descendente*) da relación. Internamente, a tipoloxía das relacións léxico-semánticas en Galnet é bastante variada e require unha caracterización pormenorizada. Por unha banda, a maior parte das relacións permiten unha lectura *reversíbel*, de tal xeito que a mesma relación entre dous *synsets* pode reflectir, por exemplo, que un destes nós conceptuais é hiperónimo do outro, mais a lectura desa mesma relación en *sentido* inverso (*ascendente*) infórmanos de que este segundo *synset* é hipónimo do primeiro. A única relación *non-reversíbel* en Galnet é ***see_also_wn15***, que só permite a lectura nun *sentido* e en *dirección horizontal*. Por outra banda, a consulta das relacións reversíbeis pódese considerar en *sentido* de *ascendencia* / *descendencia* (os casos menos controvertíbeis serían a hiperonimia / hiponimia e a holonimia / meronimia, mais en WordNet débense incluír tamén as relacións do grupo Domain). As relacións dos grupos Antonyms, Synonyms, Related e Verbs manexáronse en *dirección horizontal*.
- *Distancia* e *ruta*. A *distancia* enténdese como a cantidade mínima de relacións léxico-semánticas que se precisan para chegar desde un *synset* a outro nó conceptual co que existe unha *ruta* de conexión a través das relacións, xa sexa directamente ou mediante *synsets* intermedios.

- *Toxicidade*. É o termo co que designamos a desviación do campo conceptual, o afastamento do *synset* de destino dunha relación do ámbito conceptual do *synset* de orixe. O grao de *toxicidade* intrínseca das relacións ***usage*** / ***usage_term*** e ***region*** / ***region_term*** impediu que se puidesen utilizar durante os experimentos en ningún caso. As relacións do grupo Glosses utilízanse unicamente en casos nos que se carece doutra alternativa, xa que non teñen a súa orixe nunha relación semántica, senón que se elaboraron a partir do corpus de glosas etiquetadas semanticamente[30] do WordNet de Princeton. Co resto das relacións, a *toxicidade* ponderouse en diferentes graos segundo a adecuación á *dirección* e ao *sentido* da exploración nunha situación determinada. Por exemplo, cando quixermos explorar a *ascendencia* dun *synset*, a hiperonimia considerouse exenta de *toxicidade*; no entanto, a hiponimia ponderouse como altamente *tóxica*.

3.1. Termonet

Durante a fase de revisión e remodelación das relacións léxico-semánticas de Galnet experimentouse asemade coa posibilidade de filtrar os *synsets* relacionados de xeito que gardasen coherencia terminolóxica; isto é, deseñouse unha metodoloxía para a navegación parametrizábel polo léxico dunha área de especialidade a partir dun *synset* que representase ese ámbito de especialidade. Ideouse, ademais, un método de verificación da presenza empírica dos conceptos en corpus especializados en lingua galega. A navegación terminolóxica a partir dun *synset* e a verificación en corpus especializados da área do léxico seleccionada foi implementada nunha aplicación web de libre consulta ligada a interface Galnet e denominada Termonet[31].

3.1.1. Recursos

As funcionalidades de Termonet alicérzanse en dous recursos básicos: o léxico WordNet e un corpus textual lematizado e desambiguado semanticamente consonte os sentidos de WordNet. Na implementación actual de Termonet, deseñada para a súa aplicación en tarefas terminolóxicas nos ámbitos da medicina, da ecoloxía e da economía, estes dous recursos básicos son o léxico do Galnet e o Corpus Técnico do Galego (CTG) con anotación semántica.

Termonet utiliza a versión de desenvolvemento de Galnet, na actualidade a versión 3.0.15. Pola súa banda, o CTG é un corpus de orientación terminolóxica de 15 millóns de palabras, formado por textos especializados do galego contemporáneo nos ámbitos do dereito, informática, economía, ciencias ambientais, ciencias sociais

30 http://wordnet.princeton.edu/glosstag.shtml

31 http://sli.uvigo.es/galnet/termonet.php

e medicina. As seccións do CTG anotadas semanticamente que se utilizan na implementación actual de Termonet son o corpus de medicina Medigal de 3.823.232 palabras, o corpus de ecoloxía Auga de 2.691.036 palabras e o corpus de economía Achega de 2.055.837 palabras. A etiquetación semántica destes corpus foi realizada con Freeling[32] e UKB (Agirre / Soroa 2009), usando Galnet como léxico para a desambiguación semántica do corpus.

3.1.2. Extracción de variantes nun ámbito de especialidade

A función principal de Termonet consiste en facilitar a extracción de variantes de WordNet relacionadas cun ámbito de especialidade. Con este fin, Termonet ofrece un formulario de consulta que permite elixir un *synset* da rede léxico-semántica e, a partir del, realizar unha extracción dos termos relacionados en función da configuración das relacións semánticas que se seleccione. Aínda que Termonet permite a extracción desde calquera *synset* de WordNet, dada a súa orientación termonolóxica, a aplicación trata de suxerir sempre as variantes nominais máis próximas cando se propón iniciar a exploración desde un *synset* que non sexa nominal.

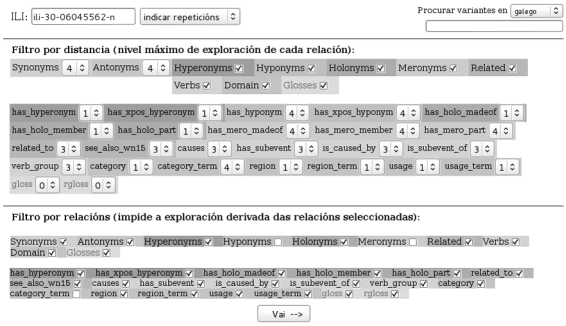

Figura 6. Consulta en Termonet

Como se ilustra na parte superior da Figura 6, Termonet permite indicar o *synset* de orixe que definirá o ámbito de extracción terminolóxica, e seleccionar logo o conxunto de relacións semánticas que se utilizarán para a identificación dos termos dese ámbito, así como a distancia ou nivel de profundidade até onde se desexa explorar cada tipo de relación. Deste xeito, Termonet despregará a árbore de relacións desde o *synset* de orixe a través desa relación até acadar o nivel de profundidade determinado. Véxase na Figura 7, por exemplo, a relación de hiponimia despregada

32 http://nlp.lsi.upc.edu/freeling/

até o nivel 4 de profundidade na terminoloxía do ámbito da medicina construída a partir do *synset* representado pola variante inglesa *medical_science* cos parámetros que se ilustran na Figura 6.

Figura 7. Extracción terminolóxica

A aplicación conta tamén cun subformulario (parte inferior da Figura 6) que permite restrinxir a extracción terminolóxica impedindo a exploración derivada das relacións seleccionadas. Mediante este filtro, en sintonía coa distancia de exploración de cada relación, trátase de limitar a *toxicidade* para a selección dos termos dun ámbito de especialidade; é dicir, de reducir o impacto das relacións que introducen *synsets* que se desvían do campo conceptual. Segundo este criterio, por exemplo, a hiperonimia considérase unha relación *tóxica*, xa que amplía a cobertura semántica inicial e tende a introducir termos de campos conceptuais máis amplos ca os de partida.

Pese a que a ferramenta de extracción terminolóxica se atopa aínda en fase de probas, permite xa, con configuracións dos parámetros de exploración relativamente simples, a obtención de conxuntos de resultados congruentes e cuantitativamente significativos. O procedemento de extracción é idéntico para ámbitos conceptuais amplos, como por exemplo a bioloxía, e para campos máis concisos, como a microbioloxía.

3.1.3. Verificación en corpus especializados

Como xa se mencionou con anterioridade, Termonet permite verificar os resultados da extracción nun corpus textual lematizado e desambiguado respecto dos sentidos de WordNet. Na actualidade permite contrastar os termos galegos que se identifican nos corpus Medigal, Auga e Achega do CTG etiquetados con Freeling e UKB. Porén, está previsto que se engadan progresivamente os restantes subcorpus do CTG (Galex, do ámbito do dereito; Xiga, de informática; e Sogal, de socioloxía).

O corpus desambiguado facilita o desenvolvemento de estratexias de verificación con base semántica para as variantes monoléxicas procedentes de Galnet, mais presenta certas dificultades para as formas pluriléxicas, que non contan con etiquetaxe semántica debido ás características da lematización do corpus con Freeling. Co fin de comprobar a presenza no corpus deste último tipo de variantes, Termonet identifica as palabras léxicas da variante en lemas sucesivos do corpus e calcula a súa frecuencia.

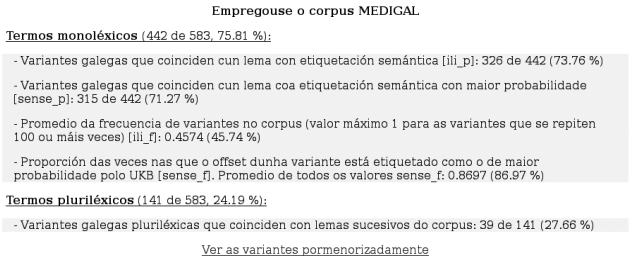

Figura 8. Verificación en corpus

Termonet avalía a presenza de cada termo monoléxico no corpus aplicándolle catro criterios cuantificados de 0 a 1, e finalmente combina os resultados obtidos por todos os termos nun índice xeral para cada criterio. Os catro parámetros que se avalían son os seguintes:

- A variante está presente (1) ou non (0) como lema do corpus e coa etiqueta semántica do *synset* correspondente.
- A variante está presente como lema do corpus e coa etiqueta semántica máis probábel (1) ou non (0) segundo UKB.
- Frecuencia absoluta da variante no corpus, ponderando o valor máximo (1) para as variantes etiquetadas semanticamente que se repiten 100 veces ou máis, e o valor mínimo (0) para as variantes que non están presentes no corpus.

- Frecuencia coa que UKB lle atribúe a maior probabilidade á etiqueta do *synset* da variante, asignando o valor máximo (1) para a totalidade das veces e o mínimo (0) para ningunha.

atropina 02754756-n 29 *a poisonous crystalline alkaloid extracted from the nightshade family; used as an antispasmodic and to dilate the eye pupil; also administered in large amounts as an antidote for organophosphate nerve agents or organophosphate insecticides*

- ili_p: 1
- sense_p: 1
- ili_f: 0.29
- sense_f: 1

auricular 05231592-n 21 *the craniometric point at the center of the opening of the external acoustic meatus*

- ili_p: 1
- sense_p: 1
- ili_f: 0.21
- sense_f: 1

autismo 05896998-n 159 *(psychiatry) an abnormal absorption with the self; marked by communication disorders and short attention span and inability to treat others as people*

- ili_p: 1
- sense_p: 1
- ili_f: 1
- sense_f: 1

autoenxerto 05583158-n 1 *tissue that is taken from one site and grafted to another site on the same person*

- ili_p: 1
- sense_p: 1
- ili_f: 0.01
- sense_f: 1

autopsia 00141396-n 91 *an examination and dissection of a dead body to determine cause of death or the changes produced by disease*

- ili_p: 1
- sense_p: 1
- ili_f: 0.91
- sense_f: 1

Figura 9. Avaliación dos termos

Na Figura 8 amósanse os índice globais que se obtiveron coa terminoloxía construída a partir do *synset* representado pola variante *medical science* cos parámetros que se ilustran na Figura 6. A partir da análise pormenorizada das variantes (Figura 9), Termonet ofrece a posibilidade de comprobar os seus contextos de uso no corpus especializado (Figura 10), permitindo a adquisición dunha valiosa información terminolóxica sobre o uso real dos termos.

O lema atropina|02754756-n documéntase en 29 ocasión/s (en 27 frase/s):

1 [CTG 195/17220] - **En** en SPS00 1 - **casos** caso NCMP000 1 13943400-n:0.0113659/07308889-n:0.0106814 **severos** severo AQ0MP0 1 02448437-a:0.0220908 **,** , Fc 1 - **prolongados** prolongar VMP00PM 0.692458 00317888-v:0.0110262/01439155-a:0.0101206 **ou** ou CC 1 - **con** con SPS00 0.999941 - **bradicardia** bradicardia NCFS000 1 14362510-n:0.0191861 **asociada** asociado AQ0FS0 0.242263 - **,** , Fc 1 - **atropina** atropina NCFS000 1 02754756-n:0.0198199 **0,5-1mg** 0,5-1mg Z 1 - **SC** sc NP00000 0.963257 - **.** . Fp 1 - **/** / Fh 1 - **IV** iv NP00000 1 - **.** . Fp 1 -

2 [CTG 198/19420] - **-** - Fg 1 - **Reacción** reacción NCFS000 1 - **vagal** vagal AQ0CS0 0.291805 - **:** : Fd 1 - **poñer** poñer VMN0000 0.533333 01494310-v:0.0091453/01570108-v:0.00736441 **a** o DA0FS0 0.696141 - **paciente** paciente NCMS000 0.219037 10405694-n:0.0183271 **en** en SPS00 1 - **decúbito** decúbito NCMS000 1 14212126-n:0.0141073 **con** con SPS00 1 - **as** o DA0FP0 1 - **pernas** perna NCFP000 1 05560787-n:0.0161418 **elevadas** elevado AQ0FP0 0.980769 02472563-a:0.0148963 **e** e CC 0.995045 - **,** , Fc 1 - **se** se CS 0.0701273 - **non** non RN 0.999905 00024073-r:0.018041 **cede** ceder VMIP3S0 0.99802 02358655-v:0.00335149/01989053-v:0.00326679/02703289-v:0.00310862/00804476-v:0.00308656 /02199590-v:0.00293483 **,** , Fc 1 - **administrar** administrar VMN0000 1 - **lle** lle PP3CSD00 1 - **atropina** atropina NCFS000 1 02754756-n:0.0144155 **.** . Fp 1 -

Figura 10. Termos en contexto

3.2. Epinónimos

Unha vez que Galnet contaba cunha versión revisada das relacións léxico-semánticas e gozaba duns agrupamentos que se consideraron satisfactorios, decidiuse emprender unha nova fase experimental que explotase as implicacións terminolóxicas das relacións entre *synsets* sen que fose preciso a intervención humana na selección do punto de partida nin na configuración da exploración. En certa medida, a idea inicial foi a de estabelecer un percorrido en sentido contrario ao que utiliza Termonet para explorar un ámbito a partir dun *synset*, de tal xeito que cada *synset* buscase o seu propio camiño entre as relacións para chegar a un *synset* nominal *epinónimo* que representase a área semántica na que incluírse de forma totalmente automática. Enténdese, daquela, un epinónimo como un *synset* nominal que representa a categoría dunha área semántica á que se adscribirán automaticamente outros *synsets* mediante algoritmos que avaliarán a súa proximidade a través do tratamento terminolóxico das relacións léxico-semánticas.

A exploración destas implicacións terminolóxicas das relacións entre *synsets* realizouse en dúas fases diferenciadas:

- A selección de *synsets* nominais epinónimos que se erixisen en categorías para representar as diferentes áreas semánticas e agrupar os *synsets* que comparten a adscrición a cada un destes ámbitos.
- O deseño dun conxunto de algoritmos para calcular, desde cada un dos *synsets* de WordNet, a súa ruta a través das relacións léxico-semánticas até a categoría ou categorías epinonímicas máis próximas.

3.2.1. Categorización de *synsets* nominais epinónimos

Co fin de configurar unha metodoloxía coa que categorizar os epinónimos, estabelecéronse dous criterios para sérenlles aplicados aos 83.246 *synsets* nominais, de tal xeito que unicamente aqueles que os cumprisen se autoerixisen en epinónimos. Estes dous criterios son comprobados mediante un algoritmo que percorre os *synsets* nominais en sentido descendente a través da relación de hiponimia; é dicir, parte do *synset* que se corresponde coa variante galega *entidade*, que ocupa o nivel superior (ou nivel 0) de hiperonimia entre os *synsets* nominais en WordNet, e desprázase a través das relacións de hiponimia (aumentando, segundo a distancia con *entidade*, o seu nivel de hiponomia). Os parámetros que este algoritmo ten en conta na selección dos hipónimos son:

- O cómputo de relacións do propio *synset* en relación co seu nivel de hiponimia. Para calcular esta cantidade non se teñen en conta todas as relacións: as dos grupos Synonyms e Antonyms omítense por seren propias dos adxectivos, as relacións dos grupos Hyperonyms e Holonyms que-

dan fóra tamén do cómputo polo seu carácter ascendente e as relacións ***see_also_wn15***, ***region*** / ***region_term***, ***usage*** / ***usage_term*** e ***gloss*** / ***rgloss*** exclúense polo seu alto grao de toxicidade intrínseca.

- A ponderación da toxicidade. Consiste na media aritmética entre os valores asignados a cada unha das relacións computadas no apartado anterior segundo o seguinte baremo: ás relacións do grupo Hyponyms asígnaselles 1; ás do grupo Meronyms, 0,9; ás relacións ***category_term***, 1,5; ás ***category***, 0,1; e ás demais, 0,5.

Na versión de desenvolvemento actual, seleccionáronse como epinónimos os *synsets* cun cómputo (*c*) superior a 15 no primeiro apartado e unha ponderación (*p*) maior ca 0,85 no segundo e que, ademais, obtivesen un valor superior a tres no resultado de dividir o produto do cómputo pola ponderación entre o nivel de hiponimia (*h*). Cómpre indicar que o conxunto de resultados pódese constrinxir ou aumentar lixeiramente reducindo ou incrementando os valores utilizados, mais en xeral apenas se aprecia variabilidade na selección dos *synsets* epinónimos salvo con modificacións moi bruscas dos parámetros. Nos experimentos máis recentes, dispoñíbeis para consulta na interface de Galnet, obtivéronse 927 epinónimos que supoñen, respecto ao cómputo total de *synsets* nominais e á totalidade dos *synsets* de WordNet, un 1,11% dos *synsets* nominais (83.246) e un 0,78% do total de *synsets* de WordNet (118.868).

O deseño do algoritmo, ao comezar polo nivel superior de hiperonimia, permite manter a herdanza na árbore de hiponimia de WordNet para poder así reconstruír as relacións entre os epinónimos seleccionados nunha arborescencia de subcategorías cuxas pólas principais pódense observar na Figura 11[33].

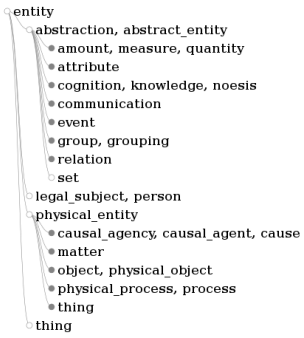

Figura 11. Pólas principais da árbore dos epinónimos

33 Pódese consultar a árbore completamente despregada en http://sli.uvigo.es/galnet/hierarchy.php?version=dev&ontology=epinonyms&category=entity#ili-30-00001740-n

A selección de epinónimos ofrece unha perspectiva xeral das áreas semánticas de WordNet, consecuente coa súa propia estrutura interna e obtida automaticamente a partir da exploración das relacións desde unha perspectiva terminolóxica; exenta, polo tanto, de calquera modelado da rede léxico-semántica para encaixar en concepcións categoriais preconcibidas.

3.2.2. Emparellamento de cada *synset* de WordNet co epinónimo da área semántica

Unha vez que contabamos cun conxunto de *synsets* nominais representativo, deseñouse unha metodoloxía para ligar, con criterios terminolóxicos baseados nas características das relacións léxico-semánticas, cada *synset* de WordNet co seu epinónimo ou epinónimos máis próximo(s).

O algoritmo para asignar epinónimo execútase desde cada un dos *synsets* de WordNet e trata de atopar o *synset* epinónimo que se encontra a menor distancia. A estratexia é, en certa medida, similar á que se utilizou para a selección de epinónimos mais en sentido oposto; isto é, orientada en sentido ascendente. Ademais, execútase recursivamente cada vez que se incrementa a distancia co *synset* de orixe; isto é, cada vez que se exploran sen éxito todas as posibilidades nunha distancia dada. A primeira comprobación do algoritmo consiste en determinar se no nivel de distancia no que se atopa nese momento existe algún *synset* que estea categorizado como epinónimo. En caso afirmativo, devolve a distancia, o(s) epinónimo(s) correspondente(s) e, cando a distancia é superior a 0, a ruta. En caso contrario, comeza a buscar rutas para incrementar un nivel de distancia até chegar ao(s) epinónimo(s) que se atopen á menor distancia posíbel cunha serie de pasos encadeados para optimizar o percorrido:

1. Restrínxese a busca para que só se utilicen as relacións dos grupos Hyperonyms, Holonyms ou as relacións ***category***, ***category_term*** e ***related_to***. Ademais engadíuselle un filtro para que se prefira a hiperonimia fronte ás outras relacións, no caso de que existan diferentes alternativas neste contexto, e para que abandone as rutas que utilizan a relación ***category*** a unha distancia superior a 1 desde o *synset* de orixe. Se non atopa ningún *synset* de destino con estas características para engadir á ruta, volve intentalo no seguinte paso.
2. Exclúense da exploración as relacións dos grupos Antonyms, Synonyms e Hyponyms e as relacións ***see_also_wn15***, ***region***, ***region_term***, ***usage***, ***usage_term***, ***gloss*** e ***rgloss***. Igual ca no caso anterior, só continúa ao paso seguinte de non atopar ningunha relación para estabelecer a ruta.
3. Explóranse unicamente as relacións ***gloss*** e ***rgloss***. No caso de ***gloss*** omítense os *synsets* de destino adxectivais e adverbiais. A exploración con ***rgloss*** só se

produce se foi infrutuosa a anterior e unicamente no caso de que o *synset* de destino teña a mesma categoría gramatical que o *synset* de orixe. Unha vez máis, se aínda non hai ningún *synset* de destino para a ruta, compróbase o seguinte e último paso.

4. Procúrase algún *synset* de destino a través de calquera das relacións salvo ***region***, ***region_term***, ***usage***, ***usage_term***, ***gloss*** e ***rgloss***. Se o algoritmo non atopar ningún *synset* de destino abandónase o *synset* de orixe.

En todos os casos anteriores engádese tamén unha asignación inversamente proporcional á toxicidade para cada tránsito por unha relación. Deste xeito, ás relacións do grupo Hyperonyms asígnaselles un valor de ponderación de 1; á relación ***category***, 0,9; ás relacións ***near_antonym***, ***near_synonym***, ***has_hyponym***, ***has_xpos_hyponym***, ***has_meronym***, ***gloss*** e ***rgloss*** asígnaselles 0,1; e a todas as demais, 0,5. A media aritmética destas asignacións na ruta até un epinónimo estabelece o criterio de desempate nos casos nos que se atopan diferentes epinónimos á mesma distancia.

Con esta metodoloxía estabelecéronse 128.986 emparellamentos, cantidade lixeiramente superior ao cómputo de *synsets* en WordNet (108,51%), dado que algúns *synsets* quedaron ligados con máis dun epinónimo. Ademais, conseguiuse que o 99,80% dos *synsets* de WordNet fosen quen de atribuírse automaticamente a(s) súa(s) propia(s) área(s) semántica(s) e unicamente 239 *synsets* (225 adverbiais, 8 verbais e 6 adxectivais) quedaron sen acadar ningún epinónimo, como se reflicte na Táboa 10.

	synsets **emparellados**	***synsets*** **en WordNet**	**porcentaxe**
Nomes	83.246	83.246	100%
Verbos	18.150	18.156	99,97%
Adxectivos	13.837	13.845	99,94%
Adverbios	3.396	3.621	93,79%
TOTAL	118.629	118.868	99,80%

Táboa 10. Cobertura dos emparellamentos de *synsets* con epinónimos

Dado o marcado carácter ontolóxico-relacional dos resultados obtidos, engadíronse, á par das ontoloxías xa presentes con anterioridade na interface de Galnet, os métodos de consulta precisos para explorar o deseño de áreas semánticas de WordNet que representan os epinónimos. Ademais, de xeito semellante ao que se describiu no apartado relativo a Termonet, implementouse unha aplicación web[34] que permite a

34 http://sli.uvigo.es/galnet/category.php

verificación de variantes galegas en corpus especializados a partir da selección de todas as variantes dunha categoría de calquera das ontoloxías.

4. Conclusións e liñas futuras de investigación

Neste artigo tratamos de recoller as diferentes metodoloxías que se utilizaron durante o proceso de elaboración do WordNet para a lingua galega até o momento actual. Repasamos os diferentes procesos de extracción e as distintas fontes lexicográficas e textuais utilizadas na construción deste recurso, e describimos algunhas das súas aplicacións na investigación en ontoloxías e no labor terminolóxico.

A nosa intención é continuar dedicando esforzos á extensión da cobertura léxica de Galnet, tanto en número de conceptos coma de variantes, para facer deste recurso unha ferramenta útil de normalización da nosa lingua no campo das tecnoloxías lingüísticas. Pretendemos seguir incorporando material léxico a Galnet nos experimentos en curso anteditos, e teimar na ampliación da súa cobertura nas áreas do léxico terminolóxico e da fraseoloxía. Así mesmo, planificamos a medio prazo complementar este recurso léxico do galego cun corpus específico, o corpus SensoGal, un corpus paralelo inglés-galego anotado semanticamente con referencia a Galnet e aliñado a nivel de frase e de palabra co corpus SemCor da lingua inglesa.

Pola súa banda, Termonet constitúe unha ferramenta orientada á extracción terminolóxica por dominios de especialidade que se atopa en fase de probas, aínda que xa é plenamente funcional. A verificación dos termos en lingua galega en corpus de especialidade desambiguados permite adquirir unha valiosa información sobre o seu uso real e constitúe unha fonte de coñecemento moi relevante para a expansión do propio Galnet guiada por campos conceptuais. As súas funcionalidades serán ampliadas progresivamente mediante a incorporación da totalidade dos subcorpus do Corpus Técnico do Galego.

O experimento de obtención de áreas semánticas arredor dos epinónimos atópase aínda en pleno desenvolvemento e é previsíbel que en fases sucesivas se realicen lixeiras correccións para mellorar a integridade dos resultados. Porén, na actualidade xa están previstas novas fases experimentais estruturando os seus resultados cara a tarefas de recuperación de información e categorización textual, e cara á desambiguación semántica de textos, co fin de determinar se un grafo de desambiguación que integre a estrutura dos epinónimos pode contribuír á mellora na eficacia dos algoritmos actuais.

Revista Galega de Filoloxía, 2015, 16: 169-201
ISSN 1576-2661 / e-ISSN 2444-9121 © UDC

Referencias bibliográficas

Agirre, E. / Edmonds, P. (2006): *Word Sense Disambiguation* (Berlín: Springer).

Agirre, E. / Alegria, I. / Rigau, G. / Vossen, P. (2007): "MCR for CLIR", *Procesamiento del Lenguaje Natural* 38: 3-15.

Agirre, E. / Soroa, A. (2009): "Personalizing PageRank for Word Sense Disambiguation", en *Proceedings of the 12th Conference of the European Chapter of the ACL*, 33-41.

Álvarez de la Granja, M. (2003): *As locucións verbais galegas* (Santiago de Compostela: Universidade de Santiago de Compostela).

Álvarez Lugrís, A. / Gómez Guinovart, X. (2014): "Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del *Dicionario Moderno Inglés-Galego*", en Domínguez Vázquez, M. J. / Gómez Guinovart, X. / Valcárcel Riveiro, C. (eds.), *Lexicografía de las lenguas románicas: Aproximaciones a la lexicografía moderna y contrastiva*, 31-48 (Berlín: De Gruyter Mouton).

Álvez, J. / Atserias, J. / Carrera, J. / Climent, S. / Oliver, A. / Rigau, G. (2008): "Consistent Annotation of EuroWordNet with the Top Concept Ontology", en *Proceedings of the 4th Global WordNet Conference*, s. p. (Szeged: GWN).

Bentivogli, L. / Forner, P. / Magnini, B. / Pianta, E. (2004): "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing", en *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, 101-108 (Xenebra: ACL).

Cabré i Castellví, M. T. (1992): *La terminologia. La teoria, els mètodes, les aplicacions* (Barcelona: Empùries).

Elberrichi, Z. / Rahmoun, A. / Bentaalah, M. A. (2008): "Using WordNet for Text Categorization", *The International Arab Journal of Information Technology* 5:1, 16-24.

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

Fernández Montraveta, A. / Vázquez, G. (2010): "La construcción del WordNet 3.0 en español", en Castillo, M. A. / García Platero, J. M. (eds.), *La lexicografía en su dimensión teórica*, 201-220 (Málaga: Universidad de Málaga).

Ferrández, S. / Ferrández, A. / Roger, S. / López–Moreno, P. (2007): "Búsqueda de respuestas bilingüe basada en ILI, el sistema BRILI", *Procesamiento del Lenguaje Natural* 38: 27-33.

Gómez Clemente, X. M. / Gómez Guinovart, X. / González Pereira, A. / Taboada Lorenzo, V. (2013): "Sinonimia e rexistros na construción do WordNet do galego", *Estudos de lingüística galega* 5: 27-42.

Gómez Guinovart, X. (2014): "Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do galego", en Macedo, A. G. / Mendes de Sousa, C. / Moura, V. (eds.), *XV Colóquio de Outono. As humanidades e as ciências: disjunções e confluências*, 331-358 (Braga: CEHUM-Universidade do Minho).

Gómez Guinovart, X. (coord.) / Álvarez Lugrís, A. / Díaz Rodríguez, E. (2012): *Dicionario moderno inglés-galego* (Ames: 2.0 Editora).

Gómez Guinovart, X. / Oliver, A. (2014): "Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit", *Procesamiento del Lenguaje Natural* 53: 43-50.

Gómez Guinovart, X. / Simões, A. (2013): "Retreading Dictionaries for the 21st Century", en Leal, J. P. / Rocha, R. / Simões, A. (eds.), *2nd Symposium on Languages, Applications and Technologies*, 115-126 (Saarbrücken: Dagstuhl Publishing).

Gómez Guinovart, X. / Solla Portela, M. A. (2014): "O dicionario de sinónimos como recurso para a expansión de WordNet", *Linguamática* 6.2: 69-74.

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

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

Isahara, H. / Bond, F. / Uchimoto, K. / Utiyama, M. / Kanzaki, K. (2008): "Development of the Japanese WordNet", en *Proceedings of the Sixth International Language Resources and Evaluation*, s. p. (Marrakech: ELRA).

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

Miller, G. A. / Beckwith, R. / Fellbaum, C. / Gross, D. / Miller, K. (1990): "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography* 3.4: 235-244.

Noia, C. / Gómez Clemente, X. M. / Benavente, P. (coords.) (1997): *Diccionario de sinónimos da lingua galega* (Vigo: Galaxia).

Oliver, A. (2012): "WN-Toolkit: un toolkit per a la creació de WordNets a partir de diccionaris bilingües", *Linguamática* 4.2: 93-101.

Oliver, A. / Climent, S. (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, A. / Climent, S. (2014): "Automatic creation of WordNets from parallel corpora", en *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1112-1116 (Reykjavik: ELRA).

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

Pease, A. / Niles, I. / Li, J. (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*, s. p. (Edmonton: AAAI).

Pianta, E. / Bentivogli, L. / Girardi, C. (2002): "MultiWordNet: Developing an Aligned Multilingual Database", en *Proceedings of the First International Conference on Global WordNet*, 21-25 (Mysore: GWN).

Plaza, L. / Díaz, A. / Gervás, P. (2010): "Automatic summarization of news using WordNet concept graphs", *IADIS International Journal on Computer Science and Information Systems* 5:1: 45-57.

Pociello, E. / Agirre, E. / Aldezabal, I. (2011): "Methodology and Construction of the Basque WordNet", *Language Resources and Evaluation* 45.2: 121-142.

Real Academia de Medicina e Cirurxía de Galicia (2002): *Diccionario galego de termos médicos* (Santiago de Compostela: Xunta de Galicia).

Rodríguez Río, X. A. (coord.) (2008): *Vocabulario de medicina: galego-español-inglés-portugués* (Santiago de Compostela: Universidade de Santiago de Compostela).

Sager, J. C. (1990): *A Practical course in terminology processing* (Amsterdam: John Benjamin).

Simões, A. / Gómez Guinovart, X. (2014): "Bootstrapping a Portuguese WordNet from Galician, Spanish and English wordnets", en Navarro Mesa, J. L. et al. (eds.), *Advances in Speech and Language Technologies for Iberian Languages*, 239-248 (Berlín: Springer).

Solla Portela, M. A. / Gómez Guinovart, X. (2014): "Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos", en Ureña López, L. A. *et al*. (eds.), *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal*, CEUR Workshop Proceedings, vol. 1199, 29-32 (Aachen: Sun SITE Central Europe).

Solla Portela, M. A. / Gómez Guinovart, X. (2015): "Termonet: Construcción de terminologías a partir de WordNet y corpus especializados", *Procesamiento del Lenguaje Natural* 55: 335-338.

Vintar, Š. / Fišer, D. / Vrščaj, A. (2012): "Were the clocks striking or surprising?: using WSD to improve MT performance", en *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (EACL 2012)*, 87-92 (Stroudsburg: ACL).

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

Zhao, F. / Fang, F. / Yan, F. / Jin, H. / Zhang, Q. (2012): "Expanding approach to information retrieval using semantic similarity analysis based on WordNet and Wikipedia", *International Journal of Software Engineering and Knowledge Engineering* 22.2: 305-322.