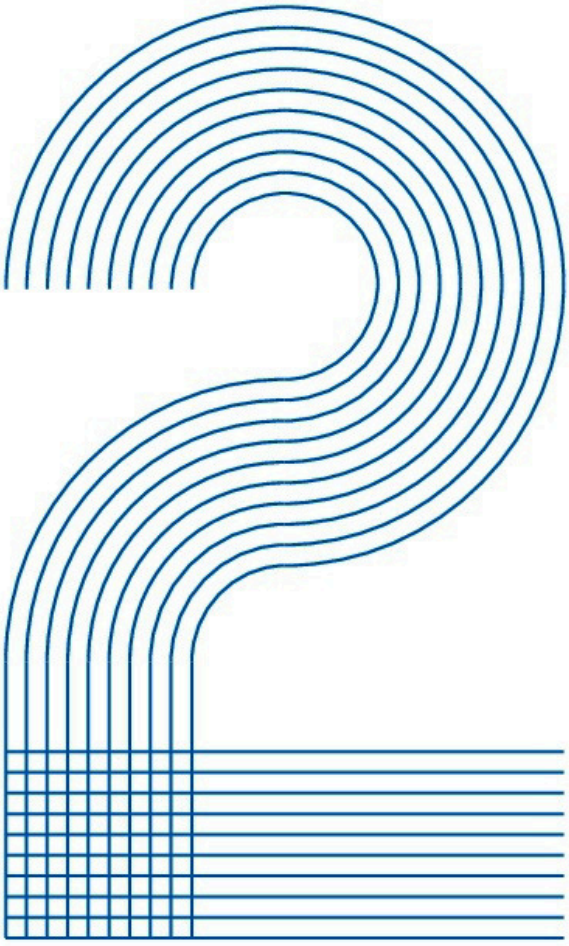
CUMIEIRA

Cadernos de investigación da nova Filoloxía Galega

Departamento de Filoloxía Galega e Latina
Vol. 2 - 2017



Xosé A. Fernández Salgado (ed.)

Universidade Vigo
Departamento de Filoloxía Galega e Latina

Cumieira : Cadernos de investigación da nova Filoloxía Galega. Vol. 2 / Xosé A. Fernández Salgado (ed.). -- Vigo: Universidade de Vigo, Departamento de Filoloxía Galega e Latina, 2017

156 p.; 24 cm

D.L. VG 687-2017

ISBN: 978-84-8158-757-9

1. Filoloxía Galega. I. Fernández Salgado, Xosé Antonio, ed. lit. II. Universidade de Vigo, Departamento de Filoloxía Galega e Latina, ed.

Revisión científica do volume 2:
Anxo Angueira (UVigo), Xavier G. Guinovart (UVigo), Ana Iglesias Álvarez (UVigo), Bieito Arias (UVigo), Gonzalo Navaza (UVigo) e Xosé A. Fernández Salgado (UVigo)

Maquetación: Xosé A. Fernández Salgado

Corrección: Antón Palacio

Deseño da portada: Tania Sueiro (Área de Imaxe da Universidade de Vigo)

Edición:
Departamento de Filoloxía Galega e Latina da Universidade de Vigo
Facultade de Filoloxía e Tradución
Campus das Lagoas-Marcosende - 36310 Vigo

Impresión: Tórculo Comunicación Gráfica, S.A.

Dep. Legal: VG 687-2017

ISBN: 978-84-8158-757-9

CUMIEIRA. Cadernos de investigación da nova Filoloxía Galega aparece recollida na base de datos DIALNET, no *Catálogo Italiano dei Periodici* ACNP (Università di Bologna) e en *Linguistic Bibliography* e Brillonline de BRILL.

Índice / *Index*

DESEÑO E CONSTRUCIÓN DUN CORPUS PARALELO ETIQUETADO SEMANTICAMENTE PARA O GALEGO

Design and elaboration of a parallel corpus semantically tagged for Galician language

Susana Brandariz Varela

Universidade de Vigo
susana.brandariz@gmail.com

Resumo: Este traballo describe o procedemento de deseño e construción dun corpus inglés-galego lematizado e desambiguado semanticamente con respecto aos sentidos das palabras definidas nunha base de datos léxica. Pártese dun conxunto de textos en inglés xa anotados coas etiquetas correspondentes aos nomes, verbos, adxectivos e adverbios; estes textos tradúcense ao galego e as palabras galegas anótanse co lema e o sentido léxico. O resultado, o corpus SensoGal, representa un recurso útil que calquera usuario pode consultar e reutilizar, ao tempo que facilita a presenza do idioma galego no ámbito das tecnoloxías. Nas seguintes seccións preséntase o proceso de elaboración en que se identifican as dificultades atopadas nas fases de tradución e anotación e se rexistran as decisións tomadas por se poden servir como referencia na esperable continuación do proxecto. Tamén se detalla o sistema de consultas e se fai unha reflexión sobre os resultados obtidos e o posible traballo futuro.

Palabras chave: Lingüística de corpus, corpus SemCor, corpus paralelo bilingüe SensoGal, desambiguación.

Abstract: This paper presents the design and elaboration of an English-Galician corpus lemmatized and semantically disambiguated with respect to the meanings of the words defined in a lexical database. To perform the task, we used a group of texts in English where nouns, verbs, adjectives and adverbs were already tagged; these texts were translated into Galician and the words tagged with their lemma and lexical meaning. The result is the corpus SensoGal: a useful resource for users and linguists that facilitates the presence of Galician language in the field of technology. In the following sections the elaboration process will be described in the phases of translation and labelling by registering the difficulties met and the decisions taken to serve as a reference in the foreseeable continuation of the project. The search system will be also explained. Finally, a reflection about the results and the future work will be done.

Keywords: Corpus linguistics, SemCor corpus, SensoGal bilingual parallel corpus, disambiguation.

1. Introdución[1]

O proxecto SensoGal que se presenta neste traballo empezou a desenvolverse grazas a unha bolsa de formación concedida pola Área de Normalización Lingüística da Universidade de Vigo no ano 2015. O proxecto está conectado con outro máis amplo que se centra na construción da versión galega da rede inglesa WordNet: un recurso léxico cuxo desenvolvemento para a nosa lingua está sendo levado a cabo polo Grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega) da Universidade de Vigo (Solla Portela / Gómez Guinovart 2015).

Durante o tempo de duración da bolsa —abril-novembro 2015 e o mesmo período de 2016—, o obxectivo principal consistiu en crear un corpus paralelo inglés-galego desambiguado semanticamente e aliñado a nivel de oración. Para isto contábase inicialmente cun corpus en inglés etiquetado semanticamente e desambiguado lexicamente que se vai traducir. Este corpus é o SemCor, un conxunto de textos de acceso libre utilizado en lingüística computacional con anterioridade en numerosas ocasións pola idoneidade do seu tamaño e pola fiabilidade que lle dá o feito de ter sido anotado de forma manual. A anotación que contén consiste en incluír para cada nome, verbo, adxectivo e adverbio unha marca que sirva para identificar o seu sentido, o cal permite traballar cos textos procesándoos por ordenador para levar a cabo experimentos tradutolóxicos, estatísticos ou extraccións terminolóxicas posteriores (Solla Portela / Gómez Guinovart 2017).

Coa axuda do programa de tradución asistida OmegaT, estes textos trasladáronse ao idioma galego e a desambiguación no momento de traducir levouse a cabo con respecto ao dicionario WordNet, un recurso de semántica léxica que permite enlazar interlingüisticamente os léxicos de seis linguas (inglés, español, catalán, vasco, portugués e galego) para utilizarse neste ámbito.

Deste xeito obtívose un conxunto de 30 textos de temas diversos (artes, historia, agricultura, empresa, sociedade...) e de aproximadamente 2.000 palabras cada texto, que foron traducidos do inglés ao galego e etiquetados en galego co seu lema e o seu sentido. Os textos paralelos resultantes almacénanse en formato TMX (Translate Memory eXchange), o estándar para a codificación en XML (eXtensible Markup

[1] Este proxecto non sería posible sen a tutela e a orientación de Xavier Gómez Guinovart e mais de Miguel Anxo Solla Portela. A eles e aos demais compañeiros o meu sincero agradecemento por atender con paciencia as miñas constantes preguntas e inseguridades. Grazas tamén á ANL da Universidade de Vigo pola oportunidade de aprendizaxe que supuxo o período como bolseira, e aos editores de *Cumieira* por contar co noso traballo para este volume.

Language) de memorias de tradución, o que permite o intercambio de información entre distintas aplicacións e engade a posibilidade de que o produto obtido se use noutros proxectos de tradución asistida por ordenador ou mesmo noutras aplicacións de procesamento lingüístico da linguaxe. Polo momento, a etiquetaxe semántica da tradución supuxo contribuír con 6.636 entradas, procedentes do aliñamento, ao crecemento da propia rede léxico-semántica de Galnet, o WordNet galego.

A diferenza deste corpus paralelo con outros corpus que codifican a equivalencia a nivel de oración na tradución (como o Corpus CLUVI[2], o Corpus COMPARA[3] ou o Corpus OPUS[4]) é que neste caso tamén se codifica o sentido de cada palabra con referencia ao dicionario WordNet. Como consecuencia desta desambiguación, a aliñación tamén se crea a nivel de palabra e, polo tanto, establécese unha equivalencia entre palabras que permite comparar a solución respectiva para un concepto en cada lingua (Solla Portela/ Gómez Guinovart 2015).

Aínda que o tamaño final dun corpus é un parámetro que con frecuencia serve para clasificar a súa utilidade, este non é o único, xa que tamén son factores determinantes a variedade e a precisión dos exemplos que inclúe. Por este motivo, a compilación de textos traducidos xa representa en si mesma un recurso de utilidade para comparar termos no seu contexto e para extraer información relevante a nivel léxico e tradutolóxico.

2. O corpus SensoGal

2.1. Os corpus lingüísticos

A *lingüística de corpus* é unha rama da lingüística que emprega conxuntos de textos para observar e analizar a linguaxe. Aínda que xa se empezan a compilar corpus dende a década de 1960 con este obxectivo, foi a partir da de 1980 cando a lingüística de corpus experimentou o maior avance, xa que o desenvolvemento das tecnoloxías favoreceu a dixitalización de textos e fixo máis fácil o manexo dos ordenadores.

[2] <http://sli.uvigo.gal/CLUVI/>.

[3] <http://www.linguateca.pt/COMPARA/>.

[4] <http://opus.lingfil.uu.se/>.

Un *corpus* é un conxunto de textos en formato electrónico, representativo dunha variedade lingüística; sobre estes textos pódese traballar procesándoos con medios informáticos para buscar padróns aplicables á investigación lingüística. Empréganse para facer investigación cuantitativa, elaborar gramáticas, extraer léxico xeral e terminoloxía para elaborar dicionarios... Tamén se poden usar en tarefas de procesamento da linguaxe natural, como recuperación de información, categorización de textos e resolución da ambigüidade léxica.

Os corpus pódense clasificar atendendo a diversos criterios. Segundo a procedencia do material que inclúen, encontramos corpus de texto escrito ou de transcricións orais. Segundo o contido dos textos, poden ser corpus de lingua xeral ou de sublinguaxes determinadas, máis especializadas. Se atendemos á perspectiva cronolóxica, poden ser diacrónicos, cando recompilan textos de diferentes épocas, e sincrónicos, se os exemplos que conteñen pertencen a un mesmo período.

Tamén é posible facer unha distinción en función do número de linguas que se inclúen no corpus. Así encontramos corpus monolingües, que estudan unha variedade lingüística, e plurilingües, formados por textos en dous ou máis idiomas. Unha especialización dos corpus plurilingües son os corpus paralelos que consisten nun texto e a súa tradución nunha ou máis linguas. Ademais, dise que un corpus paralelo ou de traducións está aliñado cando nel se identifican equivalencias de tradución entre os segmentos que o compoñen. Estes segmentos poden ser parágrafos, frases ou palabras. No caso que se presenta neste traballo son textos orixinais en inglés coa súa equivalencia en galego os que compoñen o corpus bilingüe paralelo aliñado SensoGal: dun lado o corpus SemCor inglés e do outro a versión galega anotada dese mesmo corpus.

Outra distinción posible dos corpus é a que se fai segundo conteñan ou non información lingüística adicional. Cando non a levan coñécense co nome de corpus crus ou limpos; cando a levan denomínanse corpus etiquetados. Nos corpus etiquetados a información que se engade en forma de etiqueta pode referirse a calquera nivel de análise da linguaxe —por exemplo: prosódica, discursiva, ortográfica, fonética, fonolóxica, semántica, morfosintáctica ou sintáctica—, e pódese incorporar de maneira manual, semiautomática ou completamente automática. Por exemplo, a anotación sintáctica, cando se fai de forma automática, non amosa bos resultados; na morfosintáctica, en cambio, o grao de acerto é alto. O proceso de etiquetaxe morfolóxica e semántica da versión galega do SemCor consiste en engadir a cada palabra léxica (nome, verbo, adxectivo ou adverbio) unha etiqueta con información sobre o sentido no dicionario WordNet e sobre o lema da palabra.

<wf ili="02003604a" lemma="aberto">abertas</wf>

As anotacións do sentido que contén o corpus SemCor especifican o sentido de cada concepto no texto; isto é, resolven a ambigüidade léxica dun termo polisémico con respecto ao contexto en que se atopa. Este proceso coñécese co nome de *desambiguación do sentido das palabras* ou DSP (en inglés *word sense disambiguation* ou WSD) e constitúe o paso previo en moitas das actividades do procesamento da linguaxe natural.

2.2. Corpus SemCor e rede léxico-semántica WordNet

Como xa se explicou en apartados anteriores, para construír o corpus de tradución bilingüe inglés-galego SensoGal utilizouse como fonte un corpus en inglés xa existente, o SemCor, que á súa vez é un subconxunto doutra recompilación de 500 textos feita en 1967 na Universidade Brown de Rhode Island e coñecida como corpus Brown.

O corpus SemCor foi creado nos anos 90 do século pasado na Universidade de Princeton e etiquetado semanticamente polo mesmo equipo que deseñou WordNet, o léxico que se usa para desambigualo semanticamente. Consiste en 352 exemplos de temas variados escritos por falantes nativos de inglés americano e que foran publicados nos anos 60; non son textos completos senón que en cada mostra se elixiu o inicio de forma aleatoria ata completar un tamaño aproximado de 2.000 palabras. En 186 dos textos que integran o corpus SemCor, todos os nomes, verbos, adxectivos e adverbios están enlazados co sentido correspondente á base de datos léxica WordNet. Destes 186 textos, incorporáronse por agora 30 ao corpus de tradución inglés-galego, nas categorías textuais de "Press: Reportage", "Press: Reviews", “Skill and hobbies”, “Popular Lore”, “Belles-Lettres”, "Miscellaneous: Government & House Organs", "Learned" e "Fiction: General", todos eles textos aparecidos na súa época en libros ou en publicacións periódicas de prensa.

WordNet, a base de datos coa que se enlazan as etiquetas do SemCor, é unha rede léxico-semántica concibida como un conxunto de nós relacionados entre si, cada nó é un concepto e está conectado aos outros nós por relacións semánticas. Cada un deses conceptos denomínase *synset* e está representado na base de datos por un grupo de lemas sinónimos que o poden expresar de forma equivalente. Eses sinónimos ou variantes léxicas dun mesmo concepto que aparecen no mesmo grupo desígnanse co nome de *variantes*. Ademais, os *synsets* poden levar ao lado unha definición do significado ou *glosa* e algunhas veces tamén inclúen exemplos de uso en contexto.

As relacións léxico-semánticas máis frecuentes entre os *synsets* que aparecen representadas no WordNet son as de hiponimia, hiperonimia, holonimia e meronimia no caso dos substantivos, de antonimia e cuasisinonimia no caso dos

adxectivos, as de antonimia e derivativas nos adverbios e as de implicación, hiperonimia/hiponimia, causatividade e oposición no caso dos verbos (Miller et al. 1990).

WordNet é un dos recursos léxicos máis utilizados en traballos de desambiguación porque, ademais de ser de acceso libre e gratuíto, é a base de datos máis grande deseñada para ser procesada por ordenadores (contén palabras de léxico, topónimos, nomes propios e léxico especializado) e presenta as relacións semánticas explícitas a través dos significados.

O WordNet orixinal para o inglés comezou en 1985 na Universidade de Princeton baixo a dirección do profesor George A. Miller. Na actualidade, a súa versión 3.0 contén 206.941 lemas (*variantes*) agrupados en 117.659 *synsets* ou grupos de sinónimos.

Despois do nacemento do WordNet de Princeton creáronse versións do WordNet en moitas linguas[5]. Unha destas versións é o proxecto EuroWordNet (Vossen 2002), que consiste nunha base de datos léxica multilingüe para varias linguas da Comunidade Europea (alemán, holandés, italiano, español, francés, checo e estonio). O proxecto empezou en 1994 e completouse en 1999, pero a súa importancia radica sobre todo en que serviu de modelo para o desenvolvemento doutros WordNet posteriores. EuroWordNet ten a mesma estrutura do WordNet de Princeton no relativo a *synsets* con relacións semánticas entre eles, mais os *synsets* das linguas están conectados entre si por un índice interlingüístico (ILI) que é único para cada concepto, e a través deste ILI están vinculados tamén cos *synsets* do WordNet inglés.

Este sistema de vinculación por ILI é o que segue tamén o Galnet, a versión galega de WordNet. O Galnet intégrase na plataforma Multilingual Central Repository (MCR) (González / Rigau 2013), que reúne na actualidade os léxicos WordNet de seis linguas (inglés, español, catalán, vasco, portugués e galego) enlazados polo ILI correspondente ao WordNet 3.0.

O proxecto Galnet comezou en 2009 e foi aumentando a súa cobertura léxica aproveitando outros traballos anteriores do Grupo TALG, como o Corpus Técnico do Galego (CTG), a base de datos terminolóxica Termoteca, o Dicionario CLUVI inglés-galego e algúns dos corpus paralelos CLUVI (Solla Portela / Gómez Guinovart 2015). Na actualidade, a versión 3.0.26 do Galnet contén 66.334 variantes pertencentes a 42.036 conceptos ou *synsets* e pódese descargar en forma de base de datos con licenza Creative Commons. Está dispoñible para consultas a través da interface

[5] Pódese consultar unha listaxe delas na páxina web da Global WordNet Association <http://www.globalwordnet.org>.

MCR[6], na páxina web do grupo de investigación SLI[7] e tamén na plataforma RILG de recursos integrados da lingua galega[8].

2.3. Proceso de elaboración

O corpus SemCor inglés-galego é un corpus paralelo formado por un conxunto de textos orixinais nunha lingua A, neste caso o inglés, e as súas traducións correspondentes nunha lingua B, o galego. O aliñamento das dúas partes faise durante o proceso asistido da tradución e os segmentos quedan emparellados a nivel de oración. Ao facer a tradución respectouse en todo momento a correspondencia 1:1 entre unidades polo que non hai omisións nin adicións a nivel de frase con respecto ao texto orixinal no que atinxe aos segmentos. Non ocorre o mesmo, como é lóxico, no nivel da palabra, onde a tradución biunívoca entre dúas linguas non é posible en todas as ocasións, e máis sendo o inglés un idioma caracterizado pola economía na linguaxe e por unha maior rixidez na orde dos constituíntes da oración do que o galego.

Tal como se detallará nos apartados que veñen a continuación, a creación do corpus consistiu en dous procesos ben diferenciados. Por un lado, levouse a cabo a tradución dos textos do corpus SemCor en inglés para o galego, tendo en conta as entradas recollidas na base de datos léxica WordNet; doutra parte, fíxose a anotación morfosintáctica co lema e o sentido léxico de cada nome, verbo, adxectivo e adverbio traducido.

2.3.1. Fase de tradución

Máis alá da complexidade das estruturas e do vocabulario técnico, a dificultade para traducir o corpus SemCor vén dada pola polisemia e pola escasa distancia entre os diferentes sentidos dun lema.

Como queda exemplificado na Figura 1 da páxina seguinte, despois de facer unha busca do lema *keep*, unha mesma forma léxica ten múltiples significados na lingua de orixe:

[6] <http://adimen.si.ehu.es/web/MCR>.

[7] <http://sli.uvigo.gal/galnet/>.

[8] <http://sli.uvigo.gal/RILG/>.

3-c01 (31)	But if you keep a calendar of events , as we do , you noticed a conflict .	Pero se levades un calendario de eventos, como facemos nós, detectariades un conflito.
4-c01 (77)	Also , perhaps , table-tennis and other indoor sports to keep them fit and contented .	Tamén, quizais, o tenis de mesa e outros deportes de interior para mantelos en forma e satisfeitos.
5-c01 (82)	When he needs money to buy something like , say , the Rolls-Royce he keeps near his vegetable_patch , he takes_a_flyer in the sale of surplus army supplies .	Cando necesita cartos para comprar algo como, digamos, o Rolls-Royce que garda preto da horta, arrisca na venda de provisións excedentes do exército.
6-e31 (95)	There is time left after cooking , and tent keeping , for the women to get_out and enjoy outdoor fun with their families .	Hai tempo sobro despois da cociña e o mantemento da tenda para que as mulleres saian e desfruten da diversión ao aire libre coas súas familias.
7-f16 (48)	The mate , Robert_Juet , who had kept the journal on the half_Moon , was experienced - but he was a bitter old_man , ready to complain or desert at any opportunity .	O primeiro oficial, Robert Juet, que escribira o diario da Half Moon, era un home experimentado pero era un vello amargado, listo para queixarse ou desertar en calquera ocasión.

Figura 1: Exemplo de resultados de busca no corpus SensoGal.

No caso do verbo *keep* inglés, no dicionario WordNet aparecen 22 entradas na categoría de verbo, sen contar as unidades pluriléxicas (locucións verbais e fraseoloxía). Identificar o significado apropiado vai requirir dun contexto, mesmo dunha busca de información adicional, que resolva a ambigüidade e confirme cal era sentido que tiña en mente o autor do texto orixinal (así, no exemplo 7 da Figura 1, Robert Juet *custodia* o diario ou *escríbeo*?)

Resolver a ambigüidade dunha palabra a nivel léxico é un dos principais problemas do procesamento da linguaxe natural (PLN). A ambigüidade léxica pódese abordar por medio de algoritmos que asocian automaticamente o significado apropiado da palabra no seu contexto, nunha tarefa que se coñece, como xa se dixo, co nome de desambiguación do sentido da palabra; pero en último caso é o tradutor o que debe comprobar e certificar que esa desambiguación é correcta.

Autores como Palmer (1998) e Gayral / Saint-Dizier (1999) xa fixeron notar a excesiva granularidade de WordNet nas tarefas de desambiguación automática e outros autores como Cucchiarelli / Velardi (1997) propuxeron reducir os sentidos aos máis esenciais, definir con menor nivel de detalle (Martí Antonín *et al*. 2003). Non é a intención deste traballo avaliar a adecuación de WordNet como método de desambiguación nin propoñer unha anotación distinta da que ofrece o sistema de ILI do SemCor inglés, senón aproveitar estes dous recursos para o experimento.

No SensoGal, os traballos de tradución e anotación realizáronse coa axuda do programa OmegaT[9] que ofrece suxestións de tradución a través de memorias previas e facilita a creación dun glosario para cada texto. As memorias que se utilizaron creáranse previamente coas traducións automáticas dos programas Apertium[10] e Google Translate[11].

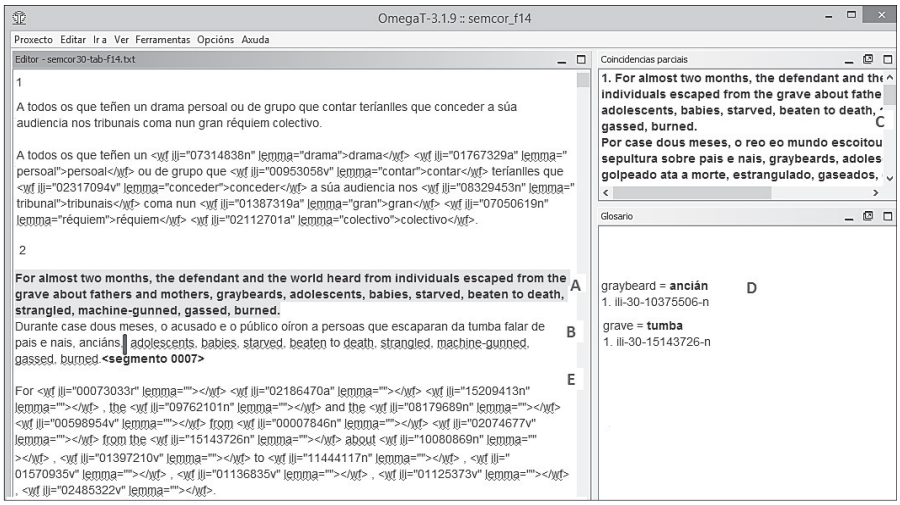


Figura 2a: Contorna de traballo co programa OmegaT.

A Figura 2a presenta a área de traballo de OmegaT onde o segmento 1 está finalizado e o segmento 2 activado e aínda en proceso de desenvolvemento. Na zona A aparece o texto orixinal en inglés; na zona B a equivalencia en galego. Ao traducir pódese facer uso das suxestións de tradución automática (C) e engadir no glosario

[9] <http://www.omegat.org/en/omegat.html>.

[10] <https://www.apertium.org/index.eng.html?dir=tat-kaz#translation>.

[11] <https://translate.google.com/?hl=gl>.

(D) novas variantes que se incorporarán á rede Galnet. A zona E corresponde á parte da anotación, onde haberá que cambiar tamén inglés por galego e completar o lema.

Ao tempo que se elabora o novo texto, explóranse no SemCor inglés as palabras enlazadas coa base de datos Galnet-WordNet consultando a glosa, as relacións, e a tradución se a houbese, do *synset*. Este proceso está representado na Figura 2b.

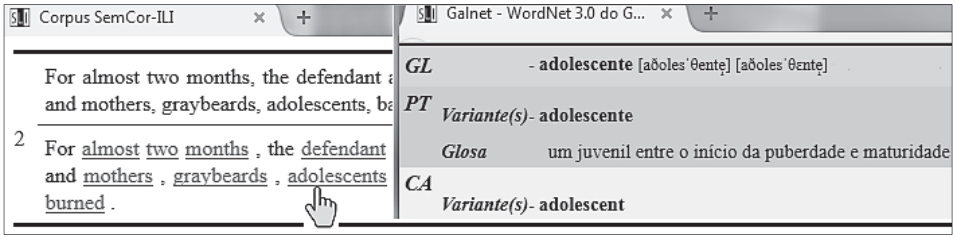


Figura 2b: Contorna de traballo con SemCor e Galnet.

Usando os recursos que acabamos de describir, o proceso de construción do corpus consiste en:

1) valorar se a proposta de sentido ofrecida polas anotacións do SemCor orixinal en inglés con respecto ao dicionario Galnet é correcta e aplicala á tradución cara ao galego (os problemas de anotación atopados explicaranse no apartado seguinte).

2) a) no caso de que non estivese rexistrada na base de datos Galnet unha tradución ao galego para ese concepto (o *synset* non ten versión en galego), propoñer unha nova entrada na base de datos; b) no caso de que xa houbese unha solución en galego para o concepto, utilizala na tradución ou propoñer unha variante nova se se considera que existe unha proposta equivalente de uso máis frecuente ou máis axeitada para ese caso.

As Figuras 3 e 4 son exemplos de *synsets* da base de datos Galnet que no momento da tradución non tiñan entrada en galego:

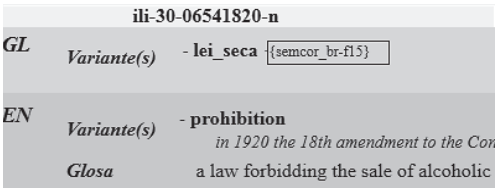


Figura 3: Exemplo entrada nova 1.

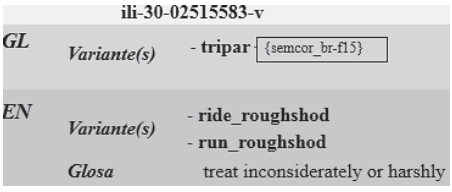


Figura 4: Exemplo entrada nova 2.

As Figuras 5 e 6 son exemplos de *synsets* que xa tiñan versión en galego e para os que se engadiron variantes novas; no primeiro caso, por adecuarse mellor ao contexto e no segundo, pola necesidade de contar con dous sinónimos para traducir a frase:

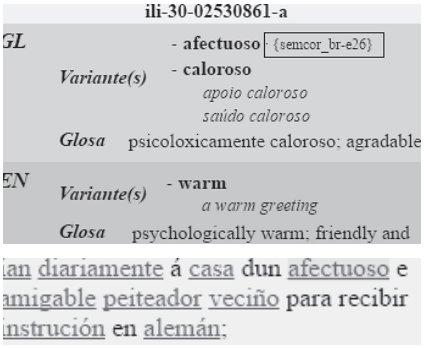


Figura 5: Exemplo variante nova 1.

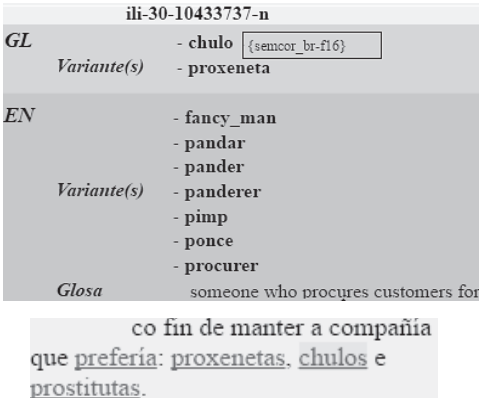


Figura 6: Exemplo variante nova 2.

A tradución dos textos realizouse seguindo os criterios lingüísticos adoptados pola Universidade de Vigo, que se recollen en Castro Figueiras / Rodríguez Ricart (2013): emprego de *ao* no canto da contracción *ó*, uso da segunda forma do artigo só nos casos en que é obrigatorio, uso dos sufixos *-aría* e *-ble*, emprego da preposición *ata* no canto de *até,* non contracción de conxunción comparativa *ca* cos artigos determinados, emprego dos signos de interrogación e de exclamación só ao remate das oracións.

Durante o proceso de tradución presentáronse algunhas dificultades recorrentes para as que se estableceron os criterios que agora se enumeran[12]:

- Os nomes de entidades (organismos, institucións, asociacións, empresas) escribíronse con todas as palabras en maiúscula e traducíronse só no caso de que existise unha versión en galego documentada (*Asociación Nacional do Rifle*), do contrario deixáronse en inglés (*American Motors Corporation*).

[12] O desenvolvemento pormenorizado e exemplificado das dificultades de tradución e de anotación constituíu a base dun traballo de fin de grao defendido en xullo de 2016 (Brandariz Varela 2016).

- Os títulos de obras literarias e pezas musicais escribíronse só con maiúscula na primeira palabra e sen cursiva para non entorpecer con etiquetas de formato o código das anotacións (*Concerto para piano nº 3*).
- Na configuración das cantidades, as unidades de millar separáronse con punto (*30.000*), os decimais con coma e engadíuselles un 0 á esquerda cando eran inferiores á unidade (o inglés *.75* pasa en galego a *0,75*).
- Segundo a convención internacional, os símbolos das unidades grafáronse sen punto: *msec* e *mm* (fronte ao inglés *msec. mm*.). Ademais, os símbolos de unidades que derivan de nomes propios deben ir en maiúscula (*Mc* non *mc*.).
- As medidas (*acres*, *polgadas*, *pés*) e as moedas (*dólares*, *centavos*, *libras*) traducíronse pero non se adaptaron ao sistema galego.
- Non se establece unha regra fixa para traducir o pronome *you*. Na maioría dos casos preferiuse a forma impersoal, pero nalgúns textos usouse o tratamento formal de *vostede*, sobre todo en preguntas directas.
- Intentouse evitar o uso excesivo dos posesivos por considerarse anglicismo: *lesionouse no ~~seu~~ xeonllo*.
- No que respecta aos verbos, os tempos compostos adaptáronse en galego case sempre na forma de pretérito perfecto (pretérito), algunha vez figuran na forma de presente. Só nos casos de carácter reiterativo se utilizou a perífrase *ter + participio*. Tamén se tivo conta de non utilizar o xerundio con valor de posterioridade na tradución.
- A pasiva ten un índice de frecuencia moi elevado en inglés. No SensoGal, cando na oración aparecía especificado o axente, utilizouse maioritariamente a construción activa; se o axente non estaba especificado, mantívose a pasiva ou cambiouse a pasiva reflexa.
- Os neoloxismos adaptáronse en acentuación ás normas do galego (*dous pinceis rígger*).
- Fíxose uso de linguaxe inclusiva sempre que non ralentice demasiado o texto (*un xogo de preguntas do tipo faino ti mesmo/a*).
- Con respecto á puntuación, é moi habitual en inglés colocar unha coma antes da conxunción, a maioría das veces non se considerou necesaria en galego e omitiuse na tradución (*tambores, xilófonos, castañolas e outros instrumentos de percusión*).
- Noutras ocasións modificouse o guión, cambiándoo por coma ou por puntos suspensivos, para conseguir un texto máis natural en galego.

- A palabra *etcétera* ou a súa abreviatura *etc*. non vai precedida de coma porque este latinismo significa 'e o demais', xa hai un *e* implícito.
- Por comodidade utilizáronse as comiñas altas ou inglesas (" ") e non as latinas (« »).
- Como se recolle na páxina do Servizo de Normalización Lingüística da USC[13] os signos de admiración e de interrogación poden ir dentro ou fóra das comiñas segundo o valor deles se circunscriba á parte entre comiñas ou se estenda a todo o fragmento. Aínda que nos textos orixinais do SemCor inglés os signos de admiración e de interrogación se colocan sempre fóra das comiñas, en galego fixéronse correspondencias diferentes segundo o caso.

2.3.2. Fase de anotación

Inicialmente o traballo de etiquetaxe morfolóxica e semántica levábase a cabo de forma totalmente manual ao mesmo tempo que se facía a tradución asistida co programa OmegaT. A Figura 7 exemplifica o proceso:

br-e22 47 His sense for rhythmic variety and timing is impeccable.	br-e22 47 O seu sentido da variedade rítmica e a medida do tempo é impecable.
His <wf ili="05807012n" lemma=""></wf> for <wf ili="02019021a" lemma=""></wf>	O seu <wf ili="05807012n" lemma="sentido">sentido</wf> da <wf ili="04751305n" lemma="variedade">variedade</wf> <wf ili="02019021a" lemma="rítmico">rítmica</wf> e a
<wf ili="04751305n" lemma=""></wf> and <wf ili="05046009n" lemma=""></wf>	<wf ili="05046009n" lemma="medida_do_tempo">medida do tempo</wf> <wf ili="02604760v"
<wf ili="02604760v" lemma=""></wf> <wf ili="01750847a" lemma=""></wf>	lemma="ser">é</wf> <wf ili="01750847a" lemma="impecable">impecable</wf>.

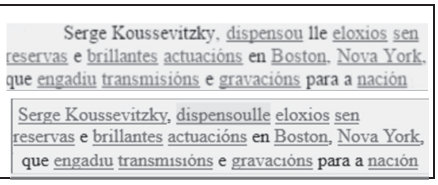
Figura 7: Fragmento do SemCor orixinal inglés anotado co ILI e a súa correspondencia en galego anotada co ILI e o lema.

[13] <http://www.usc.es/gl/servizos/snl/asesoramento/fundamentos/criterios/puntuacion2.html>.

No momento da anotación tivéronse que ter en conta as posibilidades de contracción dos pronomes átonos enclíticos e das preposicións. O criterio inicial de situalos fóra da etiqueta acabou desbotándose tras os textos iniciais para que non quedasen os enclíticos fóra da anotación da palabra:

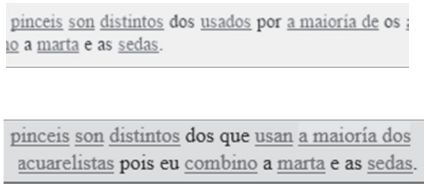
Incorrecto: <wf ili="02199590v" lemma="dispensar">dispensou</wf>**lle**

Correcto: <wf ili="02199590v" lemma="dispensar">dispensou**lle**</wf>



Incorrecto: <wf ili="01557120a" lemma="a_maioría_de">a maioría de</wf>**os**

Correcto: <wf ili="01557120a" lemma="a_maioría_de">a maioría **dos**</wf>>



Cando se anota a tradución galega, as etiquetas non sempre van coincidir coas do inglés. Unha das distincións que se debe facer vén orixinada por diferenzas de tipo cultural, xa que en galego para os séculos usamos o número cardinal, pero en inglés utilízase o ordinal: *20th Century→século XX*. O mesmo ocorre coas datas; por exemplo, en *April 10* o número no SemCor inglés remite ao *synset* ordinal '10th' pero en galego debe apuntar a '10' [13746512n], o cardinal.

Outras veces non hai máis opción que eliminar a etiqueta en galego: cando se fai unha tradución non literal, a nova palabra en galego queda sen etiquetar xa que non ten correspondencia coa inglesa; por exemplo, no segmento f16(120) *sea→onda* e en e28(10) *be on target→alcanzar obxectivos*. Tamén se debe quitar a etiqueta no texto galego cando no momento de traducir se fai un cambio de categoría: e22(38) *clearly→claridade*, e22(76) *children→infantís*, e25(17) *life→vital*. Son casos nos que se perde inevitablemente a aliñación entre palabras.

Nunha segunda etapa da construción do corpus fixéronse probas de etiquetaxe automática da tradución galega, seguida de revisión manual, con FreeLing[14] e UKB[15]

[14] <http://nlp.lsi.upc.edu/freeling/>.

[15] <http://ixa2.si.ehu.es/ukb/>.

(Aguirre / Soroa 2009) e tras varias melloras, o tempo de traballo en cada texto reduciuse a menos da metade do inicial.

A dificultade principal para poder seguir aforrando tempo no proceso, e tamén para un maior éxito da etiquetaxe automática, é que hai unha cantidade indeterminada de palabras no SemCor orixinal inglés que están mal anotadas con respecto á base de datos léxica Galnet. Ocorre sobre todo en formas que son pluriléxicas en inglés, pero están etiquetadas como monoléxicas: 'apothecary' 'shop' [10421470n] [04202417n] por 'apothecary's_shop' [03249342n]. No texto do SemCor inglés aparecen anotadas como dúas palabras independentes, pero é un concepto único en WordNet, como se pode ver na Figura 8:

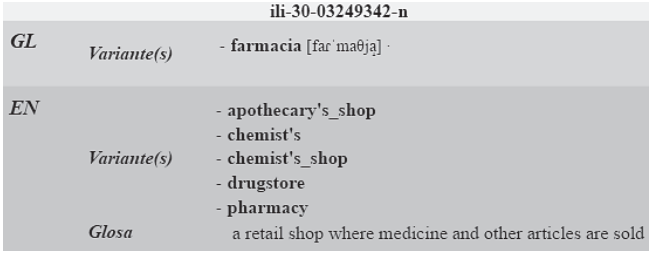


Figura 8: Exemplo de unidade pluriléxica en Galnet.

Outras formas recorrentes nos textos son 'make' 'sure' [00120316v] [00309740a] por 'make_sure' [02595234v] e 'more' and 'more' [01556355a] [01556355a] no canto de 'more_and_more' [00059854r].

O caso máis frecuente é o dos verbos con partícula: 'do' 'away' with [02560585v] [00235438r]→'do_away_with' [00471711v]. Con todo, non se deben confundir estes erros de anotación cos exemplos en que un *phrasal verb* separable ou unha expresión non levan o segundo termo incluído no enlace porque vai outra palabra no medio: 'give_up' en *give it up*, no segmento f08(82), 'take_advantage' en *take full advantage* e30(16). Aquí a anotación asignada é correcta aínda que só a primeira parte da unidade pluriléxica remite a ela.

En menor medida, tamén se encontraron unidades monoléxicas mal anotadas por interpretación equivocada do sentido: 'gay' co sentido de "homosexual" [01201937a] cando debería ter o de "showing merriment" [01367651a] en e22(2); 'march' referido a "month" [15210870n] cando debería enlazar con "music written for marching" [07058296n] en f22(40). Tamén aparecen algunhas malas anotacións en inglés por categoría equivocada: 'publishing' e26(40), 'spending' e30(41) e 'bathing' f08(42) están etiquetados como nome, pero son verbo nos tres casos;

'fight' en f16(66) indica verbo, pero é nome. Nestas últimas situacións haberá que valorar a decisión de corrixir a etiqueta en inglés tamén, o que permitirá que as palabras queden aliñadas nos dous idiomas.

Seguindo coas dificultades no proceso de anotación, atopamos que os prefixos en inglés poden aparecer como dúas unidades monoléxicas: 'Neo' 'Classicist' [01536276a] [09926519n], como unha palabra soa: 'neoclassicism' [06154464n] ou co prefixo unido por guión á base: 'neo_jazz' [07063921n]. En galego etiquetáronse sempre unidos directamente á palabra base: 'neoclasicista' [10352557n], 'neoclasicismo' [06154464n], 'neojazz' [07063921n].

Especial atención merecen tamén os comparativos de superioridade e os superlativos formados mediante sufixos, xa que algúns como 'higher', 'stronger', 'wider' 'widest' 'darkest' nos textos vistos remitiron sempre ao *synset* do adxectivo simple: 'high', 'strong', 'wide' 'dark' e polo tanto deixouse o adverbio *máis* fóra da etiqueta cando se fixo a anotación: *máis <forte>*, por exemplo no segmento e26(73). No entanto, 'bigger', 'smaller', 'larger' enlazaban ás veces co adxectivo simple e outras co comparativo: f16(74) *máis <grandes>*, pero e28(36) *<máis grandes>*.

Ademais, no SemCor orixinal inglés algúns *synsets* levan asignada unha etiqueta xenérica cando deberían levar a etiqueta propia que tamén aparece na base de datos (existen tres ILIS xenéricos que se asignan a calquera topónimo, antropónimo ou asociación que non teña entrada propia no WordNet).

Finalmente, apareceron termos sen anotar no SemCor inglés, como o substantivo *death*, talvez porque resultou difícil desambiguar o seu sentido, mesmo de forma manual e con contexto. Na Figura 9 lístanse todos os sentidos rexistrados en Galnet:

65 First came the cannon fodder, white clad civilians being driven into death as a massive human battering ram.

First came the cannon_fodder , white clad civilians being driven into death as a massive human battering_ram .

+ **death**	n 1 { alteration }	eng-30-07355491-n	*the event of dying or departure from life*
+ **death**	n 2 { organic_phenomenon }	eng-30-11444117-n	*the permanent end of all life functions in an organism*
+ **death**	n 3 { state }	eng-30-13962498-n	*the absence of life or state of being dead*
+ **death**	n 4 { point }	eng-30-15143477-n	*the time when something ends*
+ **death**	n 5 { point }	eng-30-15143276-n	*the time at which life ends; continuing until dead*
+ **Death**	n 6 { imaginary_being }	eng-30-09488259-n	*the personification of death*
+ **death**	n 7 { state }	eng-30-14562960-n	*a final state*
+ **death**	n 8 { change_of_state }	eng-30-00219575-n	*the act of killing*

Figura 9: Exemplo de resultados de busca na base de datos léxica Galnet.

Na medida do posible intentáronse corrixir todos estes aspectos engadindo a información que faltaba.

O resultado do código final en TMX, cando o texto xa está traducido ao galego, anotado e aliñado a nivel de oración co SemCor inglés, é un documento como o da Figura 10 dividido en unidades de tradución (<tu>).

```
<?xml version="1.0"?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
<header creationtool="TALG" creationtoolversion="1" segtype="sentence" o-tmf="TALG" adminlang="gl"
srclang="en" datatype="plaintext">
</header>
<body>
    <tu>
        <prop type="group">br-g43:13</prop>
        <tuv xml:lang="en"><seg>Scientists often turn out to be idiosyncratic, too.</seg></tuv>
        <tuv xml:lang="en-tag"><seg><![CDATA[<wf cmd="done" pos="NN" lemma="scientist" wnsn="1"
        lexsn="1:18:00::" ili="ili-30-10560637-n">Scientists</wf> <wf cmd="done" pos="RB"
        lemma="often" wnsn="1" lexsn="4:02:00::" ili="ili-30-00035058-r">often</wf> <wf cmd="done"
        pos="VB" lemma="turn_out" wnsn="1" lexsn="2:42:00::" ili="ili-30-02633881-v">turn_out</wf>
        <wf cmd="ignore" pos="TO">to</wf> <wf cmd="done" pos="VB" lemma="be" wnsn="1"
        lexsn="2:42:03::" ili="ili-30-02604760-v">be</wf> <wf cmd="done" pos="JJ"
        lemma="idiosyncratic" wnsn="1" lexsn="5:00:00:individual:00"
        ili="ili-30-00493820-a">idiosyncratic</wf> <punc>,</punc> <wf cmd="done" pos="RB"
        lemma="too" wnsn="2" lexsn="4:02:01::" ili="ili-30-00047534-r">too</wf> <punc>.</punc> ]]>
        </seg></tuv>
        <tuv xml:lang="gl"><seg>Ademais os científicos resultan idiosincráticos con frecuencia.
        </seg></tuv>
        <tuv xml:lang="gl-tag"><seg><![CDATA[<wf ili="00047534r" lemma="ademais">Ademais</wf> os
        <wf ili="10560637n" lemma="científico">científicos</wf> <wf ili="02634133v"
        lemma="resultar">resultan</wf> <wf ili="00493820a"
        lemma="idiosincrático">idiosincráticos</wf> <wf ili="00035058r" lemma="con_frecuencia">con
        frecuencia</wf>.]]></seg></tuv>
    </tu>

    <tu>
        <prop type="group">br-g43:14</prop>
        <tuv xml:lang="en"><seg>But still, the proposition is worth examination.</seg></tuv>
        <tuv xml:lang="en-tag"><seg><![CDATA[<wf cmd="ignore" pos="CC">But</wf> <wf cmd="done"
        pos="RB" lemma="still" wnsn="2" lexsn="4:02:04::" ili="ili-30-00027384-r">still</wf>
        <punc>,</punc> <wf cmd="ignore" pos="DT">the</wf> <wf cmd="done" pos="NN"
        lemma="proposition" wnsn="1" lexsn="1:10:00::" ili="ili-30-06750804-n">proposition</wf> <wf
        cmd="done" pos="VB" lemma="be" wnsn="1" lexsn="2:42:03::" ili="ili-30-02604760-v">is</wf>
        <wf cmd="done" pos="JJ" lemma="worth" wnsn="1" lexsn="5:00:00:worthy:00"
        ili="ili-30-02586206-a">worth</wf> <wf cmd="done" pos="NN" lemma="examination" wnsn="1"
        lexsn="1:04:00::" ili="ili-30-00635850-n">examination</wf> <punc>.</punc> ]]></seg></tuv>
        <tuv xml:lang="gl"><seg>Pero con todo, a proposición é digna de exame.</seg></tuv>
        <tuv xml:lang="gl-tag"><seg><![CDATA[Pero <wf ili="00027384r" lemma="con_todo">con
        todo</wf>, a <wf ili="06750804n" lemma="proposición">proposición</wf> <wf ili="02604760v"
        lemma="ser">é</wf> <wf ili="02586206a" lemma="digno">digna</wf> de <wf ili="00635850n"
        lemma="exame">exame</wf>.]]></seg></tuv>
    </tu>

    (...)
</body>
</tmx>
```

Figura 10: Código TMX do SemCor inglés e o SemCor galego aliñados.

Como se ve na Figura 10, cada unidade de tradución contén un segmento <tuv> (variedade de unidade de tradución) co texto orixinal en inglés, outro co texto orixinal anotado, outro coa tradución en galego e outro co texto galego anotado. As

<tuv> que conteñen as versións etiquetadas do texto almacénanse como cadeas de caracteres de tipo CDATA. Ao chegar a esta sección, o analizador interpreta os datos como cadeas de caracteres e non como contido etiquetado en XML e así non se producen erros de interpretación dos datos.

2.3.3. Consultas

O corpus SensoGal está aloxado na páxina do SLI (Seminario de Lingüística Informática)[16] da Universidade de Vigo.

Dende a súa interface, presentada na Figura 11, pódense realizar buscas por palabra, por lema ou por concepto (introducindo o ILI), tanto en galego coma en inglés. O resultado aparecerá nunha táboa na que cada fila contén unha das unidades atopadas posta en paralelo nas dúas linguas (véxase Figura 1). Segundo se progrese na construción do corpus e aumente a cantidade de textos aliñados, está previsto mellorar as buscas avanzadas con comodíns e outras opcións de selección.

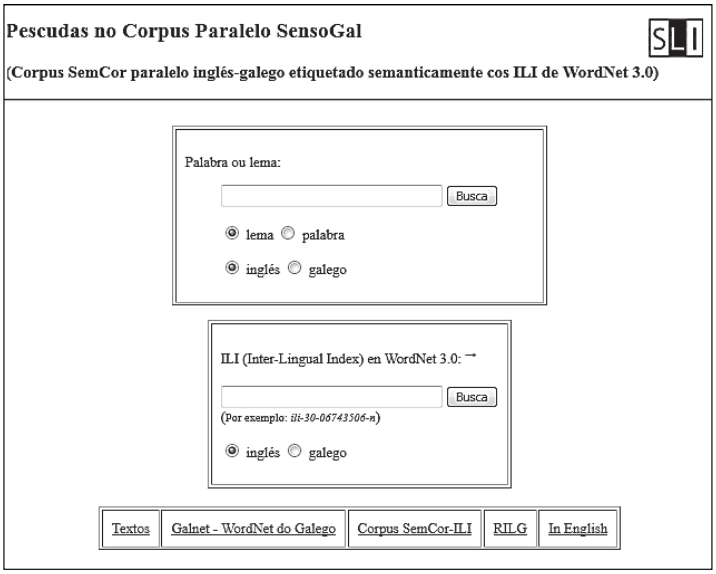


Figura 11: Interface SensoGal.

[16] http://sli.uvigo.gal/SensoGal/

A aplicación tamén dá acceso directo a outros recursos relacionados: a plataforma RILG (que contén varios corpus e dicionarios), os textos orixinais do SemCor inglés que xa foron traducidos, as estatísticas actualizadas do corpus SemCor inglés-galego; tamén aparece un enlace de acceso a Galnet dende onde se pode consultar unha listaxe das variantes en galego incorporadas á base de datos coa construción deste corpus (filtrando polo experimento semcor). O procedemento aparece representado na Figura 12

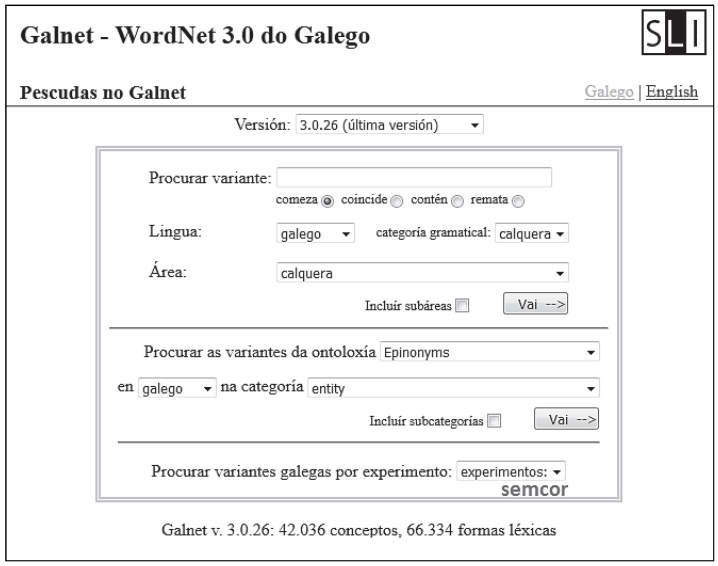


Figura 12: Interface Galnet.

3. Resultados e traballo futuro

Na actualidade, o corpus paralelo SensoGal está formado por 30 textos en inglés coas súas correspondencias aliñadas en lingua galega e segue en proceso de desenvolvemento. En total son máis de 60.000 palabras traducidas que se organizan en 2.734 unidades a nivel de oración, como se amosa na Táboa 1[17]:

[17] <Os datos actualizados pódense consultar en http://sli.uvigo.gal/SensoGal/corpus.html>.

Texto	Categoría textual	UT	Palabras EN	Palabras GL
a01	Press: Reportage	90	1987	2161
a11	Press: Reportage	80	2036	2208
c01	Press: Reviews	113	2111	2234
c02	Press: Reviews	97	2076	2093
e22	Skill and hobbies	84	2091	2111
e23	Skill and hobbies	75	2034	2075
e24	Skill and hobbies	115	2005	2086
e25	Skill and hobbies	87	2038	2108
e26	Skill and hobbies	83	2027	2010
e27	Skill and hobbies	109	1826	2003
e28	Skill and hobbies	103	1986	2066
e29	Skill and hobbies	88	2033	2096
e30	Skill and hobbies	136	2041	2217
e31	Skill and hobbies	99	2044	2255
f08	Popular lore	112	2071	2132
f10	Popular lore	101	2071	2155
f13	Popular lore	117	2053	2135
f14	Popular lore	64	2064	2060
f15	Popular lore	82	2022	2097
f16	Popular lore	122	2046	2079
f19	Popular lore	85	2080	2136
f22	Popular lore	113	2200	2185
g15	Belles lettres	72	2101	2062
g43	Belles lettres	75	2024	2031
h01	Government & House Organs	81	2032	2244
h17	Government & House Organs	99	2291	2284
j29	Learned	85	2031	2174
j34	Learned	103	2010	2027
k01	Fiction: General	159	2024	1982
k02	Fiction: General	134	2011	1968
		2734	61236	62577

Táboa 1: Tamaño e cobertura do corpus SensoGal.

Malia que o tamaño do corpus aínda é limitado permite xa observar o comportamento dun termo en contextos de temas moi diferentes grazas á diversidade de textos elixidos para traducir. O corpus tamén é útil para realizar estudos comparativos entre o inglés e o galego, para facer prácticas de tradución ou para buscar sinónimos por medio do ILI. Ao mesmo tempo, as traducións realizadas serviron para ampliar a base de datos léxica Galnet con 6.636 novos lemas no proceso de etiquetaxe semántica da tradución.

O recurso está accesible en Internet a través da web <http://sli.uvigo.gal/SensoGal> e permite buscar por lema, por palabra e por ILI tanto en galego coma en inglés.

Identificar os problemas resoltos e a evolución nas distintas etapas da creación do corpus permite facer previsións optimistas con respecto ao seu desenvolvemento futuro. Nas primeiras traducións houbo que establecer os criterios para anotar os clíticos e as contraccións da preposición, revisar continuamente a coherencia de tradución e de anotación, e lematizar o texto de xeito totalmente manual escribindo o lema de cada unha das palabras galegas e atendendo ás diferenzas de orde entre os dous idiomas. Nas fases seguintes a fluidez do traballo aumentou de forma considerable debido á experiencia adquirida e á semiautomatización do proceso de etiquetado dos textos en galego. Proba disto son os 19 textos traducidos e anotados no segundo ano da bolsa fronte aos 11 do primeiro. Finalmente, tamén se puideron extraer algúns padróns (relativamente numerosos tendo en conta que o ámbito analizado non está restrinxido a unha categoría gramatical ou a un dominio) que axudarán a cohesionar as traducións e facilitarán a etiquetaxe nos seguintes textos.

Esperamos, así mesmo, que as vindeiras ampliacións deste recurso contribúan a fomentar o uso da lingua galega dentro do sector das tecnoloxías e deste modo incidan na necesaria normalización desta lingua.

Referencias bibliográficas

Aguirre, E. / A. Soroa (2009): «Personalizing Pagerank for Word Sense Disambiguation», en *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (Atenas, 30-03-2009/3-04-2009*). Stroudsburg: Association for Computational Linguistics, 33-41.

Brandariz Varela, S. (2016): *Deseño e construción dun corpus paralelo etiquetado semanticamente para a lingua galega* (Traballo de Fin de Grao inédito). Vigo: Universidade.

Castro Figueiras, E. M.ª / M. C. Rodríguez Ricart (2013): *Na universidade en galego sen dúbida*. Vigo: Universidade.

Cucchiarelli, A. / P. Velardi (1997): «Automatic Selection of Class Label from a Thesaurus for an Effective Semantic Tagging of Corpora», en *Proceedings of the fifth conference on Applied Natural Language Processing* (Washington, 31-03-1997/3-04-1997). Stroudsburg: Association for Computational Linguistics, 380-387.

Gayral, F. / P. Saint-Dizier (1999): «Peut-on couper à la polysémie verbale?», en *Proceedings of the 6th conference on Traitement Automatique du Langage Naturel (Córcega, 12/17-07-1999).* París: Association pour le Traitement Automatique des Langues, 155-164.

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

Martí Antonín, M. A. / A. Fernández Montraveta / G. Vázquez García (2003): *Lexicografía computacional y semántica*. Barcelona: Universitat de Barcelona.

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

Palmer, M. (1998): «Consistent Criteria for Sense Distinctions», *Computers and the Humanities* 34, 217-222.

Solla Portela, M. A. / X. Gómez Guinovart (2015): «Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas», *Revista Galega de Filoloxía* 16, 169-201.

Solla Portela, M. A. / X. Gómez Guinovart (2017): «Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0», *Procesamiento de Lenguaje Natural* 59, 111-123.

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

Cumieira. Cadernos de investigación da nova Filoloxía Galega é unha publicación do Departamento de Filoloxía Galega e Latina da Universidade de Vigo que recolle traballos académicos dos novos investigadores e investigadoras no ámbito da Filoloxía Galega.

Neste vol. 2 poden lerse estudos lingüísticos sobre lingüística de corpus, sociolingüística, onomástica e léxico, e literarios sobre as vangardas e a lírica medieval.

Universidade Vigo