A INVESTIGACIÓN EN LEXICOGRAFÍA E TERMINOLOXÍA NO *CORPUS LINGÜÍSTICO DA UNIVERSIDADE DE VIGO* (*CLUVI*) E NO *CORPUS TÉCNICO DO GALEGO* (*CTG*)*

Xavier Gómez Guinovart

Seminario de Lingüística Informática (Universidade de Vigo)

1. INTRODUCIÓN

Neste artigo presentamos os traballos en lexicografía e terminoloxía a partir de córpora que se están a desenvolver na Universidade de Vigo por parte dos equipos de investigación do Seminario de Lingüística Informática e do Observatorio de Neoloxía, que realizan un labor conxunto orientado á creación de recursos para a lingua galega no marco do grupo TALG (Tecnoloxías e Aplicacións da Lingua Galega). Neste traballo explicaremos as características dos córpora *CLUVI* e *CTG*, que constitúen a fonte destes traballos, a metodoloxía seguida para a elaboración do *Dicionario CLUVI Inglés-Galego* e do *Banco de Datos Terminolóxico da Universidade de Vigo*, así como os resultados obtidos ata o momento e as tarefas que estamos a realizar e que temos en perspectiva.

2. *CORPUS LINGÜÍSTICO DA UNIVERSIDADE DE VIGO*

O Corpus Lingüístico da Universidade de Vigo (*CLUVI*) é un conxunto de córpora textuais de traducións en ámbitos específicos da lingua galega contemporánea, accesibles para consulta na web desde setembro de 2003 no enderezo http://sli.uvigo.es/CLUVI/. Cunha extensión actual total superior aos 20 millóns de palabras, o *CLUVI* está formado por seis

* Este traballo foi financiado polo Ministerio de Educación y Ciencia e o Fondo Europeo de Desenvolvemento Rexional (FEDER), dentro do proxecto "Deseño e implementacion dun servidor de recursos integrados para o desenvolvemento de tecnoloxías da lingua galega (RILG)" do Plan Nacional de I+D+I, 2006-2009 (ref. HUM2006-11125-C02-01/FILO), proxecto coordinado da Universidade de Vigo (Seminario de Lingüística Informática e Observatorio de Neoloxía) e da Universidade de Santiago de Compostela (Instituto da Lingua Galega).

córpora paralelos principais pertencentes a catro rexistros especializados (dos ámbitos xurídico-administrativo, literario, da informática e de divulgación científica) e a cinco combinacións lingüísticas diferentes (bilingüe galego-español, bilingüe inglés-galego, bilingüe francés-galego, tetralingüe inglés-galego-francés-español e tetralingüe español-galego-catalán-euskara). Estes seis córpora, cos datos actuais sobre a súa extensión, son o *Corpus Lega* de textos xurídico-administrativos galego-español (6.329.655 palabras), o *Corpus Unesco* de divulgación científica inglés-galego-francés-español (3.724.620 palabras), o *Corpus Logaliza* de localización de software inglés-galego (1.979.687 palabras), o *Corpus Tectra* de textos literarios inglés-galego (1.476.020 palabras), o *Corpus Fega* de textos literarios francés-galego (1.267.119 palabras) e o *Corpus Consumer* español-galego-catalán-euskara de información sobre consumo (5.586.431 palabras) (figura 1).

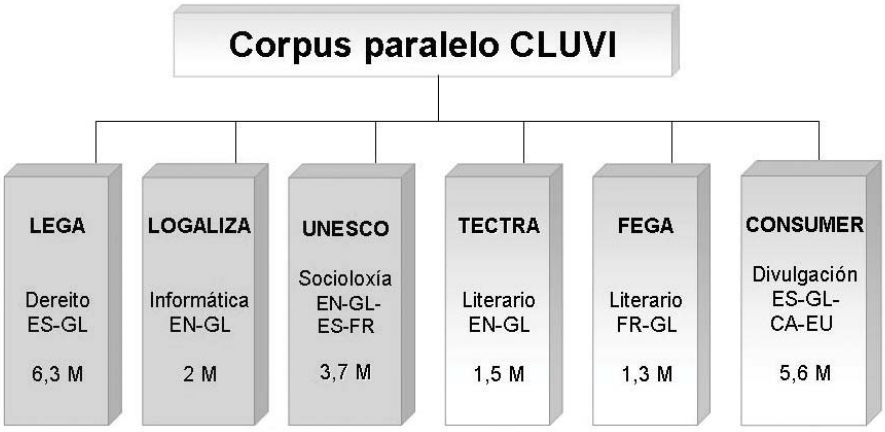

Figura 1. Composición do Corpus *CLUVI*

O corpus xurídico *Lega* galego-español contén material textual bilingüe de dous ámbitos especializados da linguaxe xurídica moi próximos, mais ben diferenciados: por unha banda, o ámbito administrativo, representado por 30 exemplares do *Diario Oficial de Galicia* publicados entre 2000 e 2005; e pola outra banda, o ámbito lexislativo, representado por un conxunto de 62 textos publicados entre 1978 e 2007 con lexislación diversa de ámbito estatal (publicada no *Boletín Oficial del Estado*) e supraestatal (*Constitución Europea*). Máis concretamente, as leis e regulamentos publicados no *BOE* pertencen a distintos eidos dentro do ámbito lexislativo: ao eido xudicial, ao ámbito da Constitución e dos Estatutos, ao eido económico, ao eido social, ao Dereito ambiental, ao Dereito informático e ao ámbito relacionado con sectores específicos (universidades, pesca, circulación, etc.). Os textos que proveñen do *DOG* suman un

total de 2.394.407 palabras, mentres que os textos lexislativos, procedentes maioritariamente do *BOE*, representan un total de 3.935.248 palabras. Por outra parte, o *Corpus Logaliza* de localización de software inglés-galego contén a localización ao galego do paquete ofimático *OpenOffice*, do sistema operativo *Windows XP* de Microsoft e do escritorio *Gnome* para *Linux*. O *Corpus Unesco* inglés-galego-francés-español de divulgación científica está constituído por 32 exemplares íntegros da revista mensual *The Unesco Courier / O Correo da Unesco / El Correo de la Unesco / Le Courier de l' Unesco* publicados entre 1998 e 2001. O *Corpus Consumer* español-galego-catalán-euskara de información sobre consumo inclúe 1.036 artigos da revista *Consumer Eroski* publicados entre 1998 e 2005. Finalmente, o corpus literario *Tectra* inglés-galego recompila 36 textos literarios en lingua inglesa coas súas traducións para o galego; e o corpus literario *Fega* francés-galego, 24 textos literarios en lingua francesa coas súas traducións para o galego. Os córpora *Tectra* inglés-portugués (735.529 palabras), e inglés-español (122.251 palabras), aínda en desenvolvemento, inclúen os textos literarios bilingües do *Tectra* inglés-galego na súa tradución ao portugués e ao español.

O *CLUVI* inclúe tamén outros cinco córpora paralelos en distintas fases de elaboración: o *Corpus Egal* de economía galego-español (718.642 palabras), o *Corpus Palop* de literatura poscolonial portugués-español (566.590 palabras), o *Corpus literario Dega* alemán-galego (76.364 palabras), o *Corpus Veiga* de subtitulación inglés-galego (71.618 palabras) e o *Corpus Turigal* de turismo portugués-inglés (373.126 palabras), este último aínda non dispoñible publicamente na web. Cómpre salientar que a través da interface do *CLUVI* se pode acceder tamén á consulta do *Corpus Lege-Bi* euskara-español de textos xurídico-administrativos (2.384.053 palabras) desenvolvido polo grupo DELi da Universidade de Deusto.

O aliñamento dos textos paralelos almacénase no *CLUVI* nunha adaptación do formato TMX (Translation Memory eXchange), o estándar para a codificación en XML de memorias de tradución, independentemente da aplicación utilizada. O concepto de memoria de tradución está relacionado coa tradución asistida por ordenador e, máis concretamente, cos escritorios de tradución como *Trados*, *DéjàVu*, *SDLX*, *Transit* ou *Passolo* (este último orientado á localización de software). Estes asistentes informáticos para a tradución integran nun único produto un procesador de textos especialmente deseñado para traducir, un conxunto de dicionarios bilingües, ferramentas para a xestión terminolóxica (creación e mantemento de glosarios, consulta automática de glosarios durante a tradución, extracción automática de terminoloxía...), e unha utilidade de memoria

de tradución. A memoria de tradución é unha base de datos onde se almacenan a versión orixinal e traducida de cada unha das frases que se traducen no marco da aplicación. Cando se está a traducir unha frase, o programa detecta automaticamente se esa mesma frase ou outra similar xa foi traducida con anterioridade, co obxecto de que se poida reutilizar a tradución sen necesidade de reescribila completamente, facendo as modificacións que se consideren máis axeitadas. En 1997 a industria creou e impulsou o estándar TMX para permitir o intercambio de memorias de tradución entre os distintos programas de tradución asistida. Con certas diferenzas, un corpus paralelo aliñado equivale a unha memoria de tradución e, na práctica, existe un número considerable de córpora paralelos aliñados codificados en TMX, coa vantaxe adicional de que os córpora así etiquetados poden ser empregados como memorias de tradución para alimentar os programas de tradución asistida. A xeito de ilustración das características xerais deste formato, amósase a seguir o aliñamento inglés-galego en TMX simplificado das tres primeiras frases de *A Perla* de Steinbeck, no orixinal inglés e mais na súa tradución ao galego. Obsérvese que, no formato TMX, tanto o orixinal, coma a súa tradución, coma a información sobre os aliñamentos, está todo incluído nun único ficheiro. Nos ficheiros TMX, os segmentos orixinais e traducidos discorren literalmente en paralelo, rodeados por etiquetas que explicitan a súa adscrición lingüística e as súas equivalencias.

```
<?xml version=”1.0” ?>
<!DOCTYPE tmx SYSTEM “tmx14simp.dtd”>
<tmx version=”1.4”>
<header creationtool=”TRANS Suite 2000” creationtoolversion=”1.4.2” segtype=”sentence” o-
tmf=”CTMTS2000” adminlang=”gl” srclang=”en” datatype=”empty”>
</header>
<body>
<tu>
<tuv xml:lang=”en”>
<seg>In the town they tell the story of  the great pearl --how it was found and how it was lost
again.</seg>
</tuv>
<tuv xml:lang=”gl”>
<seg>Na cidade cóntase a historia da gran perla, de como foi atopada e de como foi perdida de
novo.</seg>
</tuv>
</tu>
```

<tu>
<tuv xml:lang="en">
<seg>They tell of Kino, the fisherman, and of his wife, Juana, and of the baby, Coyotito.</seg>
</tuv>
<tuv xml:lang="gl">
<seg>Fálase de Kino, o pescador, e da súa muller, Juana, e do neno, Coyotito.</seg>
</tuv>
</tu>
<tu>
<tuv xml:lang="en">
<seg>And because the story has been told so often, it has taken root in every man's mind.</seg>
</tuv>
<tuv xml:lang="gl">
<seg>E como a historia foi contada tan a miúdo, acabou por botar raíces na mente de cada home.</seg>
</tuv>
</tu>
</body>
</tmx>

A unidade básica de segmentación para o aliñamento dos bitextos do corpus *CLUVI* é a frase ortográfica do texto orixinal. Xa que logo, a correspondencia entre o texto orixinal e a tradución vai ser sempre do tipo 1:n. Con frecuencia, a unha frase do orixinal correspóndelle unha frase da tradución (1:1). Porén, danse tamén casos nos que unha frase do orixinal non se traduce (1:0), ou nos que a unha frase do orixinal lle corresponde na tradución media frase (1:1/2) ou dúas frases (1:2), ou mesmo nos que unha frase da tradución non se corresponde con ningunha frase do orixinal (0:1). Alén diso, a tradución implica ás veces desprazamentos de frases enteiras, ou movementos de fragmentos de frases do orixinal a outras frases na tradución. Estes movementos reordénanse na sección de textos traducidos dos corpus paralelos do *CLUVI* para cumprir o requisito do aliñamento 1:n, que preserva a integridade e a orde das unidades de tradución do texto orixinal. Este criterio é crucial cando se aplica ao procesamento de córpora plurilingües de máis de dúas linguas, debido a que as frases do orixinal son as que, actuando a modo de intermediarias, nos permiten estabelecer as correspondencias entre as frases equivalentes das distintas linguas. A especificación TMX non ten en conta a

codificación destes aspectos das traducións, xa que foi deseñada para o almacenamento e intercambio de memorias de tradución, e non para a representación de segmentos equivalentes en córpora paralelos. O sistema de codificación do *CLUVI* está baseado no TMX, e utiliza unha versión adaptada dalgunhas das etiquetas que forman parte da especificación TMX 1.4b (Savourel 2005) para representar as correspondencias que non son 1:1 (por omisión ou por adición) e os reordenamentos codificados no corpus paralelo.

A ferramenta de busca e visualización a través dunha interface web de consulta, deseñada na linguaxe de programación PHP polo SLI, está concibida para realizar buscas bilingües en textos etiquetados conformes co formato TMX, incluídas as especificacións usadas no *CLUVI* para as omisións, as inserción e os reordenamentos. Esta aplicación PHP permite facer buscas simples e complexas (con comodíns) de palabras illadas ou de secuencias de palabras, e observar as equivalencias bilingües dos termos pescudados nos seus contextos de uso en traducións reais e documentadas. Os termos buscados poden corresponder a calquera das dúas linguas da tradución, sendo posible tamén realizar consultas autenticamente bilingües, isto é, consultas a partir de dous termos, un de cada lingua, ou mesmo consultas tetralingües, nos casos do corpus *Unesco* e do corpus *Consumer*.

3. *DICIONARIO CLUVI INGLÉS-GALEGO*

O *Dicionario CLUVI Inglés-Galego* (*CLIG*) é un dicionario baseado na colección de textos ingleses traducidos ao galego que forma parte do Corpus *CLUVI* e constitúe, ao noso entender, o primeiro dicionario baseado en córpora da lexicografía galega. Todas as palabras inglesas que aparecen nas súas entradas están documentadas nos textos en inglés traducidos ao galego recompilados no *CLUVI*. Alén diso, todas as traducións galegas recollidas no dicionario para esas palabras son traducións reais identificadas nas versións galegas dos textos ingleses do corpus. Finalmente, para cada tradución identificada, o dicionario fornece un exemplo real do seu uso tal como está documentado no corpus.

Desde o ano 2005 está dispoñible na web, no enderezo http://sli.uvigo.es/CLIG/, a primeira edición deste dicionario, que na súa versión 1.5 (2006) alcanza as 6.677 entradas e 10.807 traducións. Aínda que as entradas desta primeira edición están redactadas só na dirección de tradución inglés-galego, o sistema de busca implementado permite recuperar tamén as entra-

das a partir das súas traducións ao galego. Nestes momentos, o SLI esta a completar a segunda edición do dicionario, cuxa publicación está prevista para o último cuarto de 2008. Esta segunda edición no prelo incorpora ao dicionario un maior número de entradas e equivalencias tiradas do corpus (ao redor de 20.000 entradas e 40.000 traducións), ao tempo que amplía os datos lexicográficos contidos nas entradas con información sobre os americanismos do inglés e con notas de interese gramatical, tradutolóxico e normativo, co obxectivo de que a ferramenta resultante poida ser realmente útil tanto na docencia do inglés como na tradución inglés-galego.

A extracción de léxico bilingüe a partir do corpus *Tectra*, que serviu de base á xeración do dicionario tivo lugar en catro fases: anotación do corpus coas equivalencias de tradución entre frases, preparación do corpus para a extracción (fase de preedición), extracción léxica bilingüe automática, e edición manual dos resultados da extracción (fase de postedición). O problema central da extracción automática de léxico bilingüe consiste en converter un corpus paralelo anotado cos aliñamentos a nivel de oración (isto é, coas equivalencias oracionais de tradución) nun corpus etiquetado paralelo cos aliñamentos a nivel de palabra (isto é, coas equivalencias léxicas de tradución). Para acadar esta tarefa, existen diversos algoritmos baseados principalmente en medidas estatísticas relacionadas coa asociación mutua ou coa coaparición dos elementos léxicos nas frases bilingües aliñadas (Och / Ney 2003). Todos estes algoritmos presentan unha marxe de erro considerable nos resultados (Tiedemann 2003) por mor da natureza intrinsecamente "non literal" da tradución e doutras dificultades relacionadas coas características dos córpora, como a distancia lingüística entre as linguas implicadas, o tipo de textos ou o estilo da tradución. Para tentar superar as limitacións da extracción léxica baseada unicamente nos aliñamentos oracionais, codificamos no corpus paralelo a información tradutolóxica sobre asimetrías de tradución (aliñamentos non biunívocos e alteracións de orde na tradución), e preeditamos o corpus mediante a eliminación de diversos elementos que posúen unha incidencia directa nos erros da extracción (segmentos de texto marcados como omisións ou adicións, signos de puntuación agás os guións de unión de palabras compostas, díxitos e palabras gramaticais cun alto índice de frecuencia). A partir da versión preeditada do corpus paralelo, realizouse a extracción léxica bilingüe automática utilizando como ferramenta o programa de aliñamento léxico *NATools* (Simões / Almeida 2003). Este programa calcula o índice de correlación entre as coaparicións dos elementos léxicos nas oracións bilingües aliñadas e ofrece como saída da extracción un dicionario probabilístico inglés-galego consistente nunha lista bilingüe de todas

as palabras distintas que aparecen nos textos en inglés do corpus, cada unha delas acompañada da súa frecuencia absoluta no corpus e de ata oito palabras en galego consideradas polo aliñador como traducións máis probables. Para cada palabra galega do léxico bilingüe xerado indícase un índice estimativo da correlación entre a súa presenza nunha frase e a presenza da palabra inglesa orixinal na frase aliñada correspondente, é dicir, un estimativo da probabilidade de coaparición dos dous elementos léxicos (o inglés e o galego) nunha mesma unidade oracional de tradución. Por último, e coa finalidade de mellorar a calidade dos resultados do programa, elaboramos un "filtro de fiabilidade" para eliminar do dicionario bilingüe probabilístico xerado os candidatos de tradución menos fiables. Os estatísticos que se comproban na peneira do dicionario posterior á extracción léxica son a frecuencia absoluta do lema e a probabilidade da súa tradución máis probable. O valor concreto destes dous estatísticos é un heurístico calculado a partir da avaliación dos resultados en bruto do programa (Gómez Guinovart / Sacau 2005).

O dicionario probabilístico resultante da aplicación do filtro de fiabilidade á extracción léxica automática ten que ser editado manualmente co obxectivo de mellorar a súa precisión, eliminando as traducións erróneas que pasaran o primeiro filtrado automático, e engadindo correspondencias correctas documentadas no *CLUVI*, pero que non aparecen no dicionario xerado, ben por non formar parte do conxunto de traducións elixido (isto é, das seleccionadas polo aliñador *NATools* como primeira ou segunda opción), ben por seren palabras gramaticais frecuentes eliminadas no proceso de preedición do corpus. Nesta fase de postedición do dicionario, engadíronse as categorías gramaticais correspondentes á palabra de orixe, así como un exemplo para cada tradución coa súa referencia tirada do *CLUVI*. A primeira versión da primeira edición do *Dicionario CLIG* (a versión 1.0), que recollía un total de 5.324 entradas e 7.998 traducións, publicouse na web en maio de 2005. Nas versións posteriores a esta primeira edición, fóronse engadindo paulatinamente os lemas documentados nos textos paralelos inglés-galego do *CLUVI* (principalmente, nos córpora *Tectra* e *Unesco*, pero tamén nos córpora *Logaliza* e *Veiga*) que constaban en diversos vocabularios básicos da lingua inglesa mais que, por diversas razóns, non aparecían aínda recollidos como entradas do dicionario, ata chegar á listaxe da segunda edición (2008, no prelo) que alcanza as 20.000 entradas e 40.000 traducións.

O *Dicionario CLIG* está almacenado nun formato interno codificado en XML, consonte o cal cada entrada do dicionario inclúe ademais do lema en inglés un conxunto de informacións tradutolóxicas agrupadas en función

das posibles categorías gramaticais do lema. Cada un destes conxuntos (denominados *super_cat* na codificación utilizada) pode conter unha ou máis acepcións, dependendo da polisemia de cada lema en cada categoría gramatical. A información agrupada en cada acepción inclúe a tradución ao galego, un exemplo de uso documentado no *CLUVI* e, opcionalmente, a expresión plurilexemática da que forma parte o lema cando é o caso. Por último, cada exemplo consta dun fragmento textual do corpus *Tectra* en inglés, a súa tradución ao galego e a referencia da obra na que se documenta o exemplo. De maneira opcional, as entradas poden incluír información sobre os americanismos do inglés e notas á entrada; e as acepcións poden conter notas normativas ou de tradución. Deste xeito, cada entrada do dicionario pode incluír unha ou máis categorías gramaticais cunha tradución ou máis, sendo codificada internamente en XML, como se ilustra a seguir mediante un exemplo:

```
<entrada>
<lema>annexe</lema>
<super_cat>
<categoria>noun</categoria>
<acepcion>
<traducion>anexo</traducion>
<exemplo>
<en>The Café sur la Rue opened its Internet @annexe# last October_a kind of high-tech
office with a dozen computers.</en>
<gl>O "Café sur la rue" dotouse do seu @anexo# virtual en 1998: unha oficina equipada cuns
dez ordenadores.</gl>
<fonte>C04 (1300)</fonte>
</exemplo>
</acepcion>
</super_cat>
<super_cat>
<categoria>transitive verb</categoria>
<acepcion>
<traducion>anexar</traducion>
<exemplo>
<en>Jordan @annexes# the West Bank, while Egypt rules the Gaza Strip.</en>
<gl>Xordania @anexa# Cisxordania, Exipto administra a Franxa de Gaza.</gl>
<fonte>C14 (1542)</fonte>
</exemplo>
</acepcion>
</super_cat>
</entrada>
```

Este formato interno pódese consultar e converter a distintos formatos de presentación de acordo cos requisitos lexicográficos precisos en cada caso. Así, na versión para a web do dicionario, o dicionario en XML é procesado mediante un programa en PHP que permite a consulta interactiva do dicionario e a presentación dinámica dos resultados xerados en HTML para a súa visualización, como se pode comprobar realizando a consulta da calquera palabra inglesa na web do dicionario (http://sli.uvigo.es/CLIG/). Tamén se poden xerar presentacións das entradas aplicando follas de estilo XSL directamente sobre o formato XML, con resultados como o que se mostra a continuación:

annexe □ *noun*

anexo Δ *The Café sur la Rue opened its Internet annexe last October_a kind of high-tech office with a dozen computers*. O "Café sur la rue" dotouse do seu anexo virtual en 1998: unha oficina equipada cuns dez ordenadores. [C04 (1300)]

□ *transitive verb*

anexar Δ *Jordan annexes the West Bank, while Egypt rules the Gaza Strip*. Xordania anexa Cisxordania, Exipto administra a Franxa de Gaza. [C14 (1542)]

Neste exemplo de presentación impresa, as distintas categorías dunha entrada introdúcense mediante o cadrado, as acepcións inician parágrafo e van subliñadas, mentres que os exemplos van precedidos dun triángulo e, nun corpo de letra máis pequeno, conteñen a frase en inglés en cursiva, a súa tradución ao galego en redonda, e levan ao final entre corchetes a referencia bibliográfica abreviada da fonte (obra e número de frase). Nos restantes exemplos, ilustramos diferentes tipos de información lexicográfica codificada no dicionario: fraseoloxía, americanismos, notas ás entradas e notas ás acepcións. Todos os exemplos forman parte da segunda edición do dicionario:

angel □ *noun*
anxo Δ *Prue, a perfect angel with the others, and sometimes now, at night especially, she took one's breath away with her beauty*. Prue era un verdadeiro anxo cos outros, e ás veces, sobre todo polas noites, estaba tan fermosa que lle cortaba a un a respiración. [CAR (766)]
□ **guardian angel** anxo da garda Δ *What's more, the train's guardian angels bent over backwards trying to prevent it from turning into a dragon*. Polo demais, os anxos da garda do tren trataban por tódolos medios de impedir que se transformase nun dragón. [C30 (1258)]

amphitheatre (□ amphitheater) □ *noun*
anfiteatro Δ *Passing through the ravine, they came to a hollow, like a small amphitheatre, surrounded by perpendicular precipices, over the brinks of which impending trees shot their branches, so that you only caught glimpses of the azure sky and the bright evening cloud*. Pasaron a través do desfiladeiro e chegaron ata un val,

que era coma un pequeno anfiteatro, rodeado de precipicios verticais, con árbores próximas ó bordo, que proxectaban as súas pólas, polo que a penas se podía albisca-lo azul do ceo e as nubes brillantes da noitiña. [RIP (81)]

actual □ *adverb* □ O inglés *actual* nunca debe traducirse polo galego *actual*. Para este significado, o inglés emprega ou ben *current* ou *present*.
real Δ *And yet was he to accuse Miss Daisy Miller of actual or potential inconduite, as they said at Geneva?* E sen embargo, ¿ía el acusar a Miss Daisy Miller de inconduite real ou potencial, como dicían en Xenebra? [DAI (216)]
auténtico Δ *At most, by an alms given to a beggar whose blessing he fled from, he might hope wearily to win for himself some measure of actual grace*. Todo o máis, ao dar unha esmola a un mendigo de cuxa benzón fuxira, podería agardar, con canseira, conseguir certa medida de auténtica gracia. [RET (2146)]
verdadeiro Δ *The tortures endured, however, were indubitably quite equal for the time, to those of actual sepulture*. Con todo, as torturas sufridas naqueles momentos foran indubidablemente iguais ás dun verdadeiro enterro. [BUR (249)]

ache □ *noun*
dor □ A diferenza entre *ache* e *pain* consiste en que a primeira refírese a unha dor intensa e xeralizada (a cabeza, a espalda, as moas) mentres que a segunda supón unha dor máis localizada, p.ex., nun brazo. Δ *He felt only an ache of soul and body, his whole being, memory, will, understanding, flesh, benumbed and weary*. Sentía soamente dor de corpo e alma: todo o seu ser, memoria, vontade, intelecto, carne, atordoado e canso. [RET (2751)]
pena Δ *They sat with bowed heads, dead to all things but the ache at their hearts*. Estaban sentados coa cabeza gacha, insensibles a todo agás á pena dos seus corazóns. [LEG (668)]
□ *intransitive verb*
doer Δ *Her head began to ache, and the lights on the altar swayed before her eyes*. Empezoulle a doe-la cabeza, e a notar como as luces do altar oscilaban ante os seus ollos. [ESP (953)]

4. *CORPUS TÉCNICO DO GALEGO*

O *Corpus Técnico do Galego* (*CTG*) é unha colección de córpora do galego contemporáneo composta de textos monolingües especializados nos eidos do dereito, da informática, da economía, das ciencias ambientais, da socioloxía e da medicina, dispoñible desde 2006 para libre consulta no enderezo http://sli.uvigo.es/CTG/. Cunha extensión actual duns 12 millóns de palabras, o *CTG* está constituído polo *Corpus Galex* de textos xurídico-administrativos en galego (2.516.846 palabras), o *Corpus Xiga* de textos de informática e telecomunicacións en galego (2.027.816 palabras), o *Corpus Auga* de textos de ecoloxía e ciencias ambientais en galego (2.349.362 palabras), o *Corpus Achega* de textos de economía en galego (2.055.837 palabras), o *Corpus Sogal* de textos de socioloxía en galego (2.442.765 palabras) e o *Corpus Medigal* de textos de medicina en galego (en fase de construción) (figura 2).

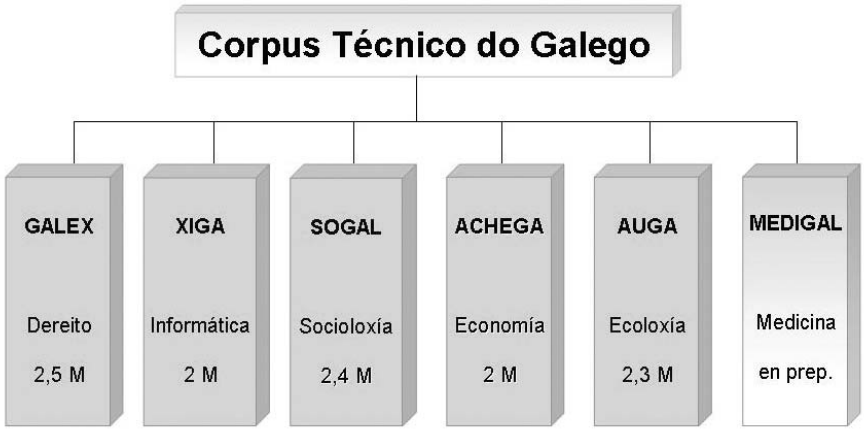

Figura 2. Composición do Corpus *CTG*

O corpus de dereito *Galex* recompila 72 textos lexislativos (leis, decretos, regulamentos) de ámbito autonómico, estatal e supraestatal, publicados entre 1998 e 2006. O corpus de informática *Xiga* contén 2450 textos sobre informática publicados entre 1997 e 2007, tirados de manuais, axudas, menús e mensaxes de programas (*Proxecto Xis, Trasno, OpenOffice, Windows*); de textos académicos e de divulgación (*Galipedia*, servizos informáticos universitarios, artigos en libros e revistas científicas); dos medios de comunicación especializados (*Dixitos, Fwwwrando-Vieiros, Código Cero, Océano Internet-Galicia Hoxe*), e de roldas, foros, grupos de novas e blogues. O corpus de economía *Achega* consta de 210 textos de economía publicados entre 2000 e 2007, tirados de convenios colectivos (banca, caixas de aforro, seguros), lexislación, libros, artigos (*Revista Galega de Economía, Terra e Tempo*), informes (Instituto Universitario de Estudos e Desenvolvemento de Galicia/IDEGA) e traballos académicos (teses de doutoramento, actas de congresos). O corpus *Auga* de ecoloxía e ciencias ambientais recolle 853 textos de ecoloxía publicados entre 1999 e 2007, e está formado por lexislación ambiental, libros, artigos, teses e traballos académicos, informes e guías, de fontes como a Federación Ecoloxista Galega, a Coordinadora para o Estudo dos Mamíferos Mariños/CEMMA, a Asociación para a Defensa Ecolóxica de Galiza/ADEGA, a Sociedade Galega de Historia Natural, a Consellaría de Medio Ambiente e a propia Universidade de Vigo. Por último, o corpus de socioloxía componse de 569 textos publicados entre 2000 e 2007 correspondentes a informes, artigos académicos e de opinión, e procedentes de diversas fontes, como a Xunta de Galicia, o Instituto Galego de Estatística, o Consello Económico e Social de Galicia, Altermundo.org, o Proxecto Galicia 2010, o Observatorio Galego dos Medios e a revista *Andaina*.

O *CTG* está almacenado no formato XML, anotado con información bibliográfica e temática, e segmentado en frases. A aplicación desenvolvida en PHP que realiza as buscas no *CTG* a través da interface web pública permite facer consultas de palabras ou grupos de palabras, utilizar comodíns para efectuar buscas complexas, e especificar o subconxunto específico do corpus ao que se desexa cinguir a pescuda. Na actualidade, o *CTG* está a ser anotado con información sobre o lema e a categoría morfosintáctica das palabras. O conxunto de etiquetas utilizado para indicar as categorías das palabras baséase nas etiquetas propostas polo grupo Eagles (Leech / Wilson 1996) para a anotación morfosintáctica de léxicos e córpora para todas as linguas europeas. Velaquí, a xeito de exemplo, un fragmento tirado do *CTG* na súa versión anotada e sen anotar[1]:

<frase>Galicia é a primeira Comunidade Autónoma pesqueira do Estado español, o sector pesqueiro representa o 8% do PIB e o 5% da poboación activa, estas cifras a pesar de estar en consonancia coa importancia do litoral a nivel mundial, o 40% da poboación do mundo vive nas zonas costeiras, presenta unhas cifras moi por enriba de calquera dos outros países comunitarios.</frase>

<frase>Galicia/Galicia_NP00000 é/ser_VIP3S00 a/o_AFS primeira/primeiro_NO0FS Comunidade/Comunidade_NCFS000 Autónoma/Autónomo_A0FS0 pesqueira/pesqueira_A0FS0 do/de_SPS00 ~/o_AMS Estado/estado_NCMS000 español/español_A0MS0 ,/,_Fc o/o_AMS sector/sector_NCMS000 pesqueiro/pesqueiro_A0MS0 representa/representar_VIP3S00 o/o_AMS 8/8_Z %/%_Ft do/de_SPS00 ~/o_AMS PIB/PIB_NCMS000 e/e_CC o/o_AMS 5/5_Z %/%_Ft da/de_SPS00 ~/o_AFS poboación/poboación_NCFS000 activa/activo_A0FS0 ,/,_Fc estas/este_DFP0 cifras/cifra_NCFP000 a pesar de/a pesar de_CS estar/estar_VN00000 en consonancia/en consonancia_R0 coa/con_SPS00 ~/o_AFS importancia/importancia_NCFS000 do/de_SPS00 ~/o_AMS litoral/litoral_A0CS0 a/a_SPS00 nivel/nivel_NCMS000 mundial/mundial_A0CS0 ,/,_Fc o/o_AMS 40/40_Z %/%_Ft da/de_SPS00 ~/o_AFS0 poboación/poboación_NCFS000 do/de_SPS00 ~/o_AMS mundo/mundo_NCMS000 vive/vivir_VIP3S00 nas/en_SPS00 ~/o_AFP zonas/zona_NCFP000 costeiras/costeiro_A0FP0 ,/,_Fc presenta/presentar_VIP3S00 unhas/un_IFP0 cifras/cifra_NCFP000 moi/moi_R0 por/por_SPS00 enriba/enriba_R0 de/de_SPS00 calquera/calquera_INS0 dos/de_SPS00 ~/o_AMP outros/outro_IMP0 países/país_NCMP000 comunitarios/comunitario_A0MP0 ./._Fp </frase>

5. *BANCO DE DATOS TERMINOLÓXICO DA UNIVERSIDADE DE VIGO*

O *Banco de Datos Terminolóxico da Universidade de Vigo* (*TUVI*) é unha base de datos terminolóxica baseada nos textos de especialidade mono-

1. Fragmento incluído na sección de ecoloxía e ciencias ambientais do *CTG* e pertencente á tese de doutoramento de Alfredo López Fernández, *Estatus dos pequenos cetáceos da plataforma de Galicia*, dirixida por Ángel Guerra Sierra e Graham J. Pierce, e presentada na Facultade de Bioloxía da Universidade de Santiago de Compostela en 2003.

lingües e paralelos recompilados nos córpora da Universidade de Vigo, isto é, no *Corpus Lingüístico da Universidade de Vigo* (*CLUVI*) e no *Corpus Técnico de Galego* (*CTG*). Esta base de datos terminolóxica, de libre acceso na web no enderezo http://sli.uvigo.es/TUVI, está mantida polo Seminario de Lingüística Informática e polo Observatorio de Neoloxía da Universidade de Vigo e conta, na actualidade, con 5.625 termos documentados nos córpora *CLUVI* e *CTG* pertencentes aos ámbitos do dereito (1411 entradas bilingües e monolingües), da socioloxía (954 entradas tetralingües e monolingües), da economía (1163 entradas monolingües) e da ecoloxía (1324 entradas monolingües). Todos os termos incluídos no *TUVI* están documentados nos córpora, estando en fase de realización os inventarios terminolóxicos dos eidos da informática e da medicina. A figura 3 representa en forma gráfica o proceso de baleirado terminolóxico dos textos que se segue para a elaboración da base de datos terminolóxica a partir dos córpora da Universidade de Vigo utilizados como fonte.

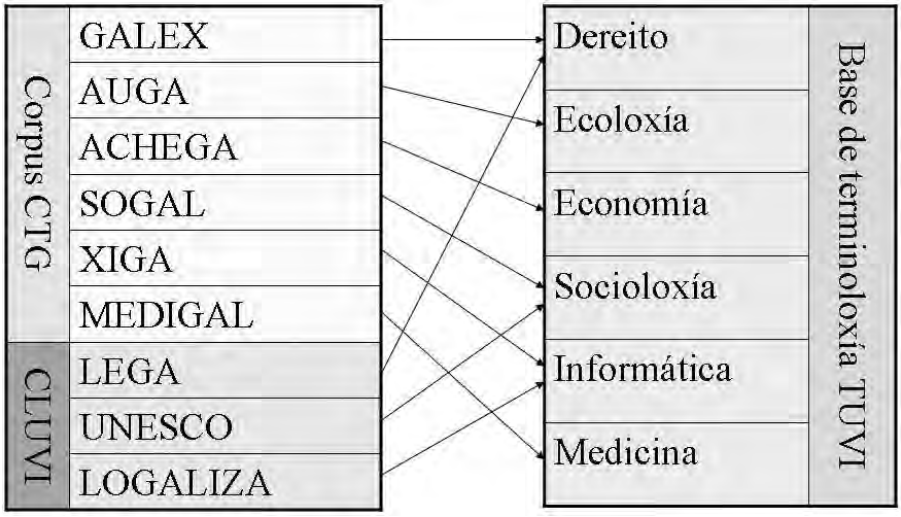

Figura 3. Fontes do Banco de Datos Terminolóxico *TUVI*.

Neste proceso de baleirado, a identificación no corpus das unidades terminolóxicas dun dominio realízase en dúas fases. Na primeira fase, analízanse as palabras e as secuencias de ata catro palabras máis frecuentes no corpus da especialidade. Mediante a revisión humana destas combinacións léxicas de maior frecuencia é posible identificar os termos máis prominentes nun ámbito. Por exemplo, analizando o *Corpus Auga* de ecoloxía e ciencias ambientais, pódense detectar termos frecuentes neste ámbito como *residuos* (5.900 veces no corpus), *especies* (2.464 veces), *impacto ambiental* (903 veces), *avaliación ambiental* (508 veces), *calidade do aire* (274 veces), *plan de xestión* (223 veces), *organismos modificados xe-*

neticamente (214 veces), *avaliación de impacto ambiental* (247 veces) ou *gases de efecto invernadoiro* (220 veces). Nesta primeira fase de baleirado, ademais das listas das frecuencias absolutas no corpus das combinacións léxicas analizadas, analízanse tamén as atraccións léxicas mediante o sistema estatístico Senta de extracción de candidatos a termos pluriléxicos, que calcula o grao de asociación entre unidades textuais contigua mediante un modelo probabilístico coñecido como expectación mutua (Dias / Guilloré / Lopes 2000).

Os resultados revisados da extracción estatística complétanse, na segunda fase da identificación das unidades terminolóxicas, comprobando a existencia no corpus dos termos recompilados no material bibliográfico de referencia ausentes da lista elaborada durante a primeira fase. Deste xeito, a selección das entradas do *TUVI* inclúe, ao carón das unidades terminolóxicas máis frecuentes no corpus e das que presentan un maior grao de asociación, os termos que, malia teren unha escasa frecuencia no corpus e/ou un baixo índice de asociación, son considerados fundamentais polo feito de estaren censados nun traballo terminolóxico de referencia no ámbito de estudo.

No *TUVI* a información terminolóxica está estruturada ao redor dos conceptos. Cada ficha do *TUVI* inclúe toda a información relativa a un concepto, expresado cun termo galego do que se poden recoller tamén variantes, tanto intralingüísticas (termos sinónimos, variantes ortográficas ou variantes dialectais) coma interlingüísticas (traducións ou, con maior propiedade, equivalencias). A información recollida para cada variante (incluíndo a variante común ou non marcada) inclúe o lema do termo, a súa categoría gramatical, a súa definición e un contexto de uso documentado no corpus. As fichas ou conceptos da base de datos están agrupadas segundo o seu campo temático, dentro da póla correspondente dunha árbore conceptual xerarquizada da materia. Alén diso, os conceptos da base de datos constitúen unha rede léxico-semántica pola que se pode navegar, e onde os nós conceptuais interrelacionan entre si en función das relacións semánticas (antonimia, hiperonimia, holonimia, etc.) que se establecen entre eles.

Formalmente, o *TUVI* almacénase internamente nunha estrutura XML que se ilustra deseguido mediante un exemplo simplificado da ficha para o concepto que recolle os termos *gas invernadoiro* e *gas de efecto invernadoiro*:

<cc>
<ic>666406</ic>
<rs tipo-rs=”caus”>66673</rs>
<rs tipo-rs=”hipo”>666405</rs>
<ct st=”Text”>AUGA.3.2.3.3</ct>
<lg xml:lang=”gl”>
<var tipo=”com”>
<lema>gas de efecto invernadoiro</lema>
<cat valor=”m”></cat>
<ex>
<texto_ex>l) Tonelada equivalente de dióxido de carbono: unha tonelada métrica de dióxido de carbono (CO2) ou unha cantidade de calquera outro |gas de efecto invernadoiro# recollido no anexo II cun potencial equivalente de quentamento do planeta.</texto_ex>
<fonte_ex>
<obra>A1018</obra>
<num>50618</num>
</fonte_ex>
</ex>
<frec>
<fab>5</fab>
<vcorpus>20.09.06</vcorpus>
<palcorpus>1604417</palcorpus>
</frec>
</var>
<var tipo=”morf”>
<lema>gas invernadoiro</lema>
<cat valor=”m”></cat>
<ex>
<texto_ex>O dióxido de carbono (C O2) é o principal |gas invernadoiro# e emítese, de forma inevitábel, ao queimarmos combustíbeis fóseis (carbón, petróleo e gas).</texto_ex>
<fonte_ex>
<obra>A031</obra>
<num>477</num>
</fonte_ex>
</ex>
<frec>
<fab>1</fab>
<vcorpus>20.09.06</vcorpus>
<palcorpus>1604417</palcorpus>
</frec>
</var>
</lg>
</cc>

Nesta estrutura de ficha terminolóxica, cada concepto (etiquetado como *cc*) está indexado mediante o seu índice conceptual (*ic*), un número único que identifica como un conxunto toda a información terminolóxica contida nunha ficha. No exemplo, o índice conceptual da ficha é o 666406. Un concepto pode establecer relacións semánticas (*rs*) con outros conceptos da base de datos. No exemplo, o concepto correspondente a esta ficha establece unha relación de causa-efecto co concepto correspondente ao índice conceptual 66673 (o da ficha terminolóxica que recolle o concepto correspondente ao termo *efecto invernadoiro*) e unha relación de hiponimia co concepto 666405 (*gas*).Cada concepto está asignado a unha póla da árbore conceptual correspondente á súa especialidade. No exemplo, o concepto 666406 está asignado ao campo temático dos gases contaminantes, que comparte con conceptos relacionados cos termos *monóxido de carbono* e *gas de vertedoiro*, entre outros. A ficha para cada concepto pode conter información sobre un ou máis termos sinónimos, nunha soa lingua ou en máis dunha lingua. A información terminolóxica relativa a cada lingua está agrupada dentro da etiqueta *lg* especificada mediante un atributo *xml:lang* que adopta o valor do código ISO de dúas letras da lingua en cuestión. Cada un dos termos nunha lingua representa unha variante lingüística do concepto nesa lingua, podéndose tratar dunha variante común, dunha variante por sinonimia, dunha variante (orto)gráfica ou dunha variante dialectal. Toda a información relativa a unha variante está contida dentro da etiqueta *var*, especificada mediante un atributo tipo que adopta o valor correspondente ao tipo concreto de variante. No exemplo, a información relativa ao galego contén dúas variantes, a variante común e unha variante morfolóxica desta. Cada variante inclúe o lema para o termo, a súa categoría gramatical e un contexto de uso documentado no corpus. No exemplo, a primeira variante para o galego, que é de tipo sinonímico, é para o lema *gas de efecto invernadoiro*, de categoría *m* (substantivo masculino). Os datos sobre o exemplo están separados en dúas partes: o (con)texto seleccionado (indicado coa etiqueta *texto_ex*) e a fonte do exemplo que, á súa vez, se divide en obra e número de frase. Os datos sobre a frecuencia de uso inclúen a frecuencia absoluta do termo (etiqueta *fab*) e a versión (*vcorpus*) e o tamaño (*palcorpus*) do corpus no momento do recoller esta información.

Como se explicou con anterioridade, este formato XML interno pódese converter a distintos formatos lexicográficos de presentación en función das necesidades. Así, na ferramenta de consulta na web, a base de datos terminolóxica é procesada por un programa informático en PHP que presenta os resultados da consulta en forma de táboa, como se mostra a continuación:

Ref. 666406	
Campo temático:	<AUGA.3.2.3.3/3.2.3.3. Gases contaminantes>
Relacións semánticas:	[caus => 66673] [hipo => 666405]
Termo GL:	**gas de efecto invernadoiro**
Categoría:	m
Variante:	com
Frecuencia relativa:	3.11639679709
Contexto de uso:	l) Tonelada equivalente de dióxido de carbono: unha tonelada métrica de dióxido de carbono (CO^2) ou unha cantidade de calquera outro *gas de efecto invernadoiro* recollido no anexo II cun potencial equivalente de quentamento do planeta. [A1018]
Ver contextos no CTG:	Procurar máis exemplos (algún dos exemplos do termo no corpus poden non corresponder ao concepto da ficha)
Termo GL:	**gas invernadoiro**
Categoría:	m
Variante:	morf
Frecuencia relativa:	0.623279359418
Contexto de uso:	O dióxido de carbono (CO^2) é o principal *gas invernadoiro* e emítese, de forma inevitábel, ao queimarmos combustíbeis fóseis (carbón, petróleo e gas). [A031]
Ver contextos no CTG:	Procurar máis exemplos (algún dos exemplos do termo no corpus poden non corresponder ao concepto da ficha)

As relacións semánticas entre termos codificados no *TUVI* son as de antonimia (*esfera pública/esfera privada*), hiponimia (*educación primaria/educación*), hiperonimia (*explotación/escravitude*), secuencialidade temporal anterior (*adolescencia/xuventude*), secuencialidade temporal posterior (*eleccións/campaña electoral*), axente (*debedor/débeda*), produto (*emigración/emigrante*), causa (*medios de comunicación/comunicación*), efecto (*quecemento do planeta/contaminación*), instrumento (*lei/xustiza*), fin (*saúde/medicina*), meronimia (*folklore/cultura*) e holonimia (*ideoloxía/crenza*). A través dos enlaces da presentación web das fichas é posible percorrer os nodos conceptuais seguindo as relacións semánticas indicadas. O *TUVI* pode concibirse, logo, como unha rede léxico-semántica a dous niveis formada por nós conceptuais que interrelacionan entre si en función da súa clasificación temática e das súas relacións semánticas (figura 4).

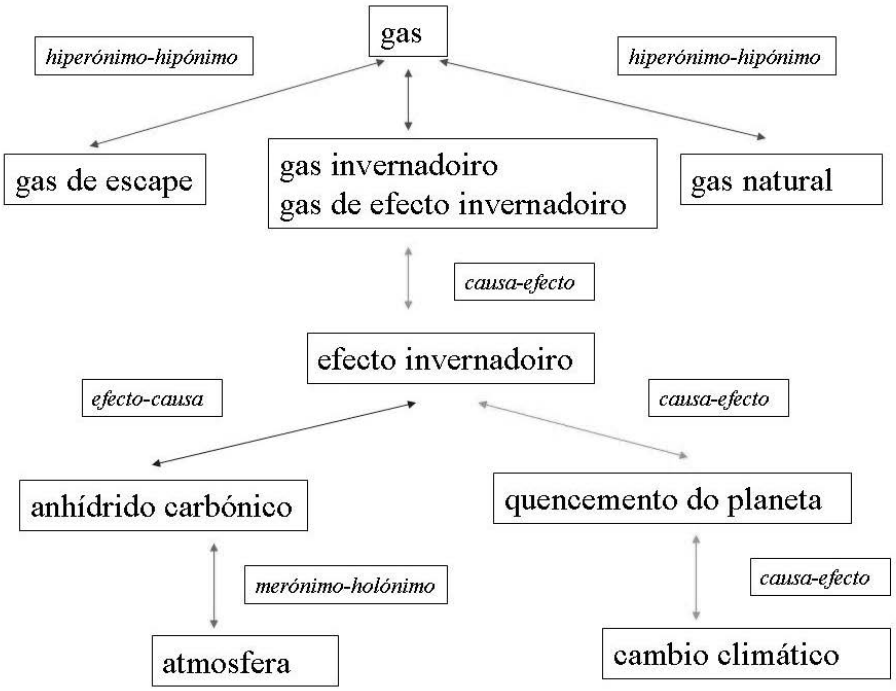

Figura 4. Relacións semánticas entre nós do *TUVI*

6. PERSPECTIVAS

Nestes momentos, os nosos esforzos de investigación no eido da lexicografía e a terminoloxía baseadas en córpora están centrados na ampliación dos córpora *CLUVI* e *CTG* (que supón a incorporación de novos textos, de novos campos especializados e de máis información lingüística), na segunda edición do *Dicionario CLUVI Inglés-Galego*, e na extensión da información terminolóxica do *TUVI* aos ámbitos da informática e da medicina.

Simultaneamente, estamos a traballar xunto co Instituto da Lingua Galega da Universidade de Santiago de Compostela no proxecto RILG (Recursos Integrados para o desenvolvemento de tecnoloxías da Lingua Galega), un proxecto encamiñado a ofrecer un portal web de servizos lingüísticos do galego desde o que se poida acceder de xeito conxunto aos bancos textuais e aos dicionarios desenvolvidos polo Seminario de Lingüística Informática e o Observatorio de Neoloxía da Universidade de Vigo e polo Instituto da Lingua Galega. Os recursos que se pretende integrar neste servizo inclúen o *Tesouro Informatizado da Lingua Galega* (*TILG*), o *Tesouro Medieval Informatizado da Lingua Galega* (*TMILG*), o Corpus *CLUVI*, o *Corpus Técnico do Galego* (*CTG*), o *Dicionario de Dicionarios*

(*DD*) (Santamarina 2003b), o *Dicionario de Dicionarios do Galego Medieval* (*DDGM*) (González Seoane 2006), o *Dicionario CLUVI Inglés-Galego* (*CLIG*), o *Banco de Datos Terminolóxico TUVI* e a *Neoloteca da Universidade de Vigo* (Gómez Clemente / Rodríguez Guerra 2003b; López Fernández et al. 2005) (figura 5). Esperamos poder ofrecer os primeiros resultados deste importante proxecto durante o ano 2008.

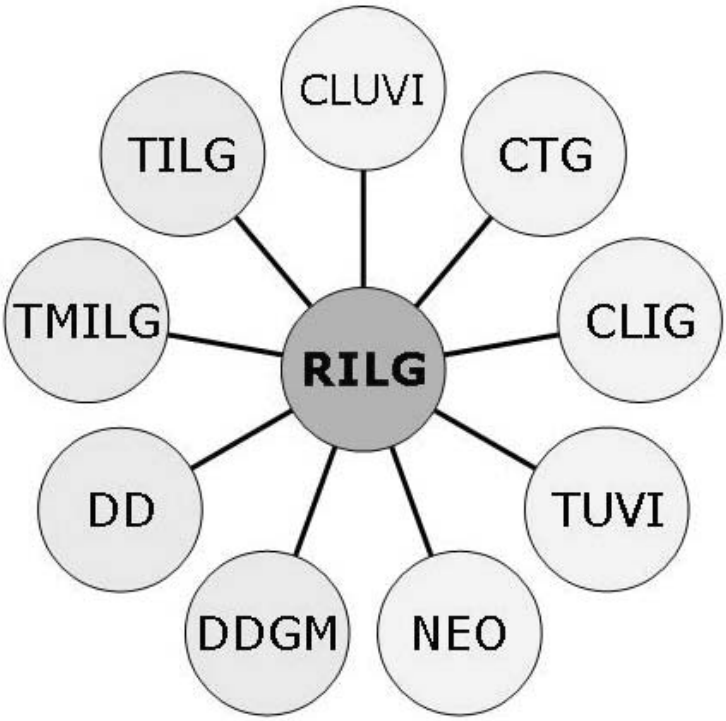

Figura 5. Integración de recursos no Proxecto RILG