Técnicas para o desenvolvemento de dicionarios de tradución a partir de córpora aplicadas na xeración do Dicionario CLUVI Inglés-Galego

Xavier Gómez Guinovart
Elena Sacau Fontenla
Seminario de Lingüística Informática
Universidade de Vigo

Resumo: *Neste artigo presentamos o Dicionario CLUVI Inglés-Galego (CLIG), un dicionario de traducións baseado en córpora paralelos que se está a desenvolver no Seminario de Lingüística Informática (SLI) da Universidade de Vigo, a partir dos datos dispoñibles no Corpus Lingüístico da Universidade de Vigo (CLUVI).*

Abstract: *This paper presents the English-Galician CLUVI Dictionary (CLIG), a parallel corpus-based dictionary of translations which is being elaborated at the Computational Linguistics Group (SLI) of the University of Vigo, using data available in the Linguistic Corpus of the University of Vigo (CLUVI).*

1. Introdución

Os córpora paralelos son coleccións dixitalizadas de textos almacenados na súa versión orixinal e traducida. As posibilidades de explotación destes corpus son moi amplas, sendo de grande importancia os estudos relacionados coa tradución e a lexicoloxía nos campos da tradución automática estatística (Och e Hermann, 2000), das memorias de tradución e da tradución automática baseada en exemplos (Turcato e Popowich, 2001), da extracción léxica para a recuperación de información multilingüe (Brown et al., 2000), da extracción de terminoloxía bilingüe (Vintar, 2001) e da extracción de léxico bilingüe (Tiedemann, 2003). Neste artigo imos presentar unha investigación sobre extracción de léxico bilingüe que se está a desenvolver no Seminario de Lingüística Informática (SLI) da Universidade de Vigo, orientada á xeración de dicionarios bilingües baseados en equivalencias léxicas de tradución extraídas de corpus paralelos. A nosa presentación vaise centrar nas técnicas utilizadas no desenvolvemento do Dicionario CLUVI Inglés-Galego (CLIG) (Gómez Guinovart et al., 2005) a partir das traducións recompiladas no Corpus Lingüístico da Universidade de Vigo (CLUVI)[1].

O Corpus CLUVI (de libre consulta na web no enderezo http://sli.uvigo.es/CLUVI) está formado por un conxunto de córpora textuais de rexistros especializados de lingua galega contemporánea[2]. A súa sección de córpora paralelos, que conta na actualidade

[1] A investigación sobre o Corpus CLUVI está financiada polo Ministerio de Ciencia y Tecnología (MCYT) e o Fondo Europeo de Desenvolvemento Rexional (FEDER), dentro do proxecto "Procesamiento lingüístico-computacional del Corpus Lingüístico de la Universidad de Vigo (CLUVI)" (ref. BFF2002-01385), cofinanciado pola Dirección Xeral de I+D da Xunta de Galicia e pola Universidade de Vigo. Máis información en http://webs.uvigo.es/sli.

[2] Véxase neste mesmo volume o artigo de Xavier Gómez Guinovart e Ánxeles Torres Padín, e Gómez Guinovart e Sacau Fontenla (2004a) para unha descrición xeral do proxecto.

con máis de doce millóns de palabras aliñadas a nivel de oración, está constituída principalmente polo Corpus LEGA de textos xurídico-administrativos galego-español, o Corpus UNESCO de divulgación científica inglés-galego-francés-español, o Corpus TECTRA de textos literarios inglés-galego e o Corpus FEGA de textos literarios francés-galego. A extracción de léxico bilingüe que serviu de base á xeración do Dicionario CLIG (accesible na web no enderezo http://sli.uvigo.es/CLIG) realizouse, concretamente, a partir do Corpus TECTRA (1.476.020 palabras), e 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). No que resta do artigo, imos presentar en catro apartados as catro fases de desenvolvemento do CLIG (anotación do corpus paralelo, preedición, extracción automática das equivalencias léxicas e postedición dos resultados), para rematar con algunhas conclusións e liñas futuras de traballo.

2. Anotación do corpus paralelo

Un dos principais problemas para a extracción léxica bilingüe automática a partir dun corpus paralelo das características do TECTRA son as asimetrías de tradución. Estas asimetrías supoñen correspondencias de tradución non biunívocas e alteracións na orde dos segmentos traducidos. No Corpus TECTRA de textos literarios inglés-galego aliñado a nivel de oración que forma parte do Corpus CLUVI o máis común é que a cada frase do texto orixinal lle corresponda unha única frase no texto traducido (aliñamento 1:1). Porén, abundan os casos nos que unha frase orixinal non é traducida (1:0), nos que a unha frase do orixinal lle corresponde na tradución a metade dunha frase (1:1/2) ou máis dunha frase (1:2, 1:3, ...), ou mesmo nos que unha frase na tradución non ten correspondencia no orixinal (0:1). Así mesmo, a tradución pode implicar alteracións na orde orixinal, isto é, desprazamentos de frases enteiras ou movementos de fragmentos de frases do orixinal a outras frases na tradución. Todas estas asimetrías inciden negativamente na precisión da extracción automática de equivalencias léxicas bilingües, debido a que o establecemento automático das equivalencias candidatas baséase na coaparición simultánea dos elementos léxicos bilingües candidatos en oracións entre as que se identificara previamente una equivalencia de tradución.

No Corpus CLUVI, o sistema de codificación dos aliñamentos oracionais dos textos paralelos é o formato TMX (Translation Memory eXchange), estándar para a codificación en XML (eXtensible Markup Language) de memorias de tradución, e utiliza como unidade básica de segmentación a frase ortográfica do texto orixinal. O formato de anotación empregado no CLUVI utiliza unha versión adaptada dalgunhas das etiquetas incluídas na especificación TMX 1.4 (Savourel, 2004) para indicar as correspondencias de tradución que non son biunívocas e mais os casos de reordenamentos. Tal adaptación foi necesaria dado que a especificación TMX non ten en conta a codificación das asimetrías de tradución, xa que foi deseñada para o almacenamento e intercambio de memorias de tradución e non para a representación das equivalencias entre os segmentos nos corpus paralelos. Deste xeito, adaptáronse do

TMX as etiquetas dos elementos <hi> e <ph> co obxectivo de marcar os tres tipos de asimetrías: omisións, adicións e reordenamentos.

O fenómeno da omisión dáse nos casos nos que unha frase ou parte dunha frase non é traducida. Isto implica que un fragmento do texto de partida non ten correspondencia no texto de chegada. No Corpus CLUVI o elemento <hi> marca no texto de partida o elemento que se omite no texto de chegada mediante un atributo *type* co valor de "supr". Por exemplo, as frases aliñadas inglés-galego de (1) serían anotadas no CLUVI como (2) (onde, seguindo as convencións do TMX, *tu* equivale a *translation unit* ou unidade de tradución, *tuv* a *translation unit variant* ou variante da unidade de tradución, e *seg* a segmento):

(1) [en] ‘Hello’, I said.
[gl] -Ola.

(2) <tu> <tuv xml:lang=”en”> <seg>’Hello’, <hi type=”supr”>I said.</hi></seg> </tuv> <tuv xml:lang=”gl”> <seg>-Ola.</seg> </tuv> </tu>

No caso da adición, insírese un fragmento no texto de chegada que non ten correspondencia no texto de partida. A adición codifícase tamén co elemento <hi> que indica o segmento inserido na tradución, e distínguese da codificación das omisións por medio do atributo *type*, que neste caso ten o valor de "incl". Se o fragmento engadido é parte dunha frase, incorpórase á unidade de tradución na que está inserido. Pola contra, cando o fragmento inserido é unha oración ou secuencia de oracións engádese quer á unidade de tradución anterior, quer á posterior, dependendo do contexto. A continuación amósase un exemplo do uso desta etiqueta:

(3) [en] ‘Hello.’
[gl] -Ola - dixen.

(4) <tu> <tuv xml:lang=”en”> <seg>’Hello.’</seg> </tuv> <tuv xml:lang=”gl”> <seg>-Ola <hi type=”incl”>- dixen.</hi> </tuv> </tu>

Por último, o reordenamento implica un cambio na posición dos elementos da tradución con respecto ao texto orixinal. Pode ocorrer que se despracen frases enteiras, ou que se dean movementos de fragmentos de frases do orixinal a outras frases na tradución. Estes movementos son reordenados no texto traducido para que frases que son tradución unha da outra se atopen na mesma unidade de tradución e facilitar deste xeito a extracción léxica bilingüe. Para a codificación dos reordenamentos combínanse dous elementos: <hi> e <ph>. Desta forma, anotamos o elemento movido, xa sexa un fragmento ou unha oración enteira, mediante un elemento <hi> que inclúe dous tipos de atributos: un *type* con valor de "reord", e un *x* con valor numérico que actúa como un índice. Por outra banda, utilízase unha etiqueta baleira <ph> para indicar no texto traducido o lugar que ocupaba orixinariamente o elemento desprazado, mediante un atributo *x* que indica a relación entre o elemento desprazado e o lugar orixinario dese elemento e que comparte valor co índice codificado no elemento <hi> do segmento movido. Co obxectivo de evitar incoherencias entre as distintas persoas que participan na codificación dos aliñamentos no Corpus CLUVI, adoptouse o criterio de que os segmentos reordenados

na codificación do texto meta sempre sexan desprazados en dirección ao inicio do texto, tendo como resultado a inexistencia de secuencias do tipo de (5), xa que a secuencia de elementos é sempre como a de (6).

(5) [...] <ph x=”n”/> [...] <hi type=”reord” x=”n”>elemento reordenado</hi> [...]

(6) [...] <hi type=”reord” x=”n”>elemento reordenado</hi> [...] <ph x=”n”/> [...]

Velaquí un exemplo de anotación dos reordenamentos no CLUVI:

(7) [en] 'The front door!’ she said in this loud whisper. ‘It’s them!’
[gl] -A porta de fóra. ¡Son eles! - murmurou bastante alto.

(8) <tu> <tuv xml:lang=”en”> <seg>’The front door!’ she said in this loud whisper.</seg> </tuv> <tuv xml:lang=”gl”> <seg>-A porta de fóra.<hi type=”reord” x=”1”>- murmurou bastante alto.</hi></seg> </tuv> </tu> <tu> <tuv xml:lang=”en”> <seg>It’s them.</seg> </tuv> <tuv xml:lang=”gl”> <seg>¡Son eles! <ph x=”1”/> </seg> </tuv> </tu>

De houber reordenamentos adicionais estes serían codificados consecutivamente cos atributos <x=”2”>, <x=”3”>, ..., <x=”n”>, como se amosa a seguir:

(9) [en]‘Leave him alone, hey’ Sunny said.‘C’mon, hey. We got the dough he owes us. Let’s go.’
[gl] -Déixao. Imos logo. Xa témo-lo que nos debe - dicía Sunny.

(10) <tu> <tuv xml:lang=”en”> <seg>’Leave him alone, hey’ Sunny said.</seg> </tuv> <tuv xml:lang=”gl”> <seg>-Déixao. <hi type=”reord” x=”1”>- dicía Sunny.</hi></seg> </tuv> </tu> <tu> <tuv xml:lang=”en”> <seg><hi type=”supr”>’C’mon, hey.</hi></seg> </tuv> <tu xml:lang=”gl”> <seg/> </tuv> </tu> <tu> <tuv xml:lang=”en”> <seg>We got the dough he owes us.</seg> </tuv> <tuv xml:lang=”gl> <seg><hi type=”reord” x=”2”>Xa témo-lo que nos debe <ph x=”1”/></hi></seg> </tuv> </tu> <tu> <tuv xml:lang=”en”> <seg>Let’s go.’</seg> </tuv> <tuv xml:lang=”gl”> <seg>Imos logo.<ph x=”2”/></seg> </tuv> </tu>

3. Preedición do corpus

Na fase de preedición do corpus, posterior ao seu aliñamento e anotación, e co obxectivo de mellorar os resultados da extracción léxica bilingüe automática, creamos unha versión reducida do corpus paralelo eliminando aqueles elementos que complican innecesariamente o proceso de extracción. En concreto, elimináronse desta versión do corpus paralelo os segmentos de texto marcados como omisións ou adicións, posto que indican unidades sen correspondencia de tradución; os signos de puntuación, excluíndo os guións de unión de palabras compostas; os díxitos; e, finalmente, certas palabras gramaticais cun alto índice de frecuencia.

A elección dos elementos gramaticais suprimidos na preedición do corpus é dependente da lingua. Así, para a lingua inglesa foron eliminados os determinantes (*the, a, an*), pronomes persoais (*I, you, he, she, it, we, you, they, me, you, him, her, you*), posesivos (*my, his, her*), demostrativos (*this, that*), conxuncións (*and, but, if, or*), preposicións (*to, of, in, at, on, with, out, around, about*), partículas negativas (*no, not*), pronomes indefinidos (*all*), verbos auxiliares (*do, does, did, is, are, was, were, has, had*) e a marca de xenitivo saxón.

Por outa banda, para o galego eliminároнse artigos (*o, a, os, as*), indefinidos (*un, uns, unha, unhas*), pronomes persoais tónicos (*eu, ti, el, ela, nós, vós, eles, eles, me, se, nos*), posesivos (*meu, meus, seu, seus*), preposicións (*a, con, de, en, para, por*), contraccións de preposición con artigo (*ó, ao, á, ós, aos, ás, co, coa, cos, coas, do, da, dos, das, no, na, nos, nas, polo, pola, polos, polas*), conxuncións (*que, e, se, nin, ou, pero*), verbos de tipo auxiliar (*é, era*) e partículas negativas (*non*).

4. Extracción léxica bilingüe automática

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 e 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 a outras 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, díxitos e palabras gramaticais cun alto índice de frecuencia), tal como se explica nos anteriores apartados.

A partir da versión preeditada do corpus paralelo, realizouse a extracción léxica bilingüe automática utilizando como ferramenta o programa de aliñamento léxico NATools (Simões e Almeida, 2003), que á súa vez aplica unha versión mellorada do algoritmo de Twente (Hiemstra, 1998). 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. O resultado en bruto da extracción levada a cabo polo programa NATools pódese ver nos seguintes tres exemplos:

(11) windows_15 ->
fiestras_0.84175086,
cristais_0.07912458,
garaxe_0.07912458

(12) longing_3 ->
morriña_0.36656892,
señardade_0.15835778,
francia_0.15835778,
formara_0.15835778,
período_0.15835778

(13) bed_96 ->
cama_0.83687806,
(null)_0.04077505,
leito_0.03240258,
deitar_0.02170320,
entre_0.01576635,
dei_0.01096033,
durmir_0.01087335,
sentárase_0.01054715

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 automaticamente 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.

A precisión dos resultados da extracción foi avaliada tomando como referencia as traducións correctas ou parcialmente correctas que o aliñador NATools ofrece como primeira ou segunda opción, entendendo como parcialmente correctas aquelas traducións que formarían parte dunha expresión pluriléxica (locución, perífrase, etc.), como amosa o exemplo (14), onde *again* pode ser traducido por "de *novo*", "*outra* vez", "*volvín/volveu/volve* a", etc.

(14) again_182 ->
novo_0.47650307,
outra_0.35298964,
volvín_0.04795518,
volveu_0.04691147,
repetiu_0.01373287,

volve_0.00785520,
estábame_0.00679777,
virou_0.00560301

As conclusións da avaliación indican que, como regra xeral, o mellor criterio de selección para a xeración de dicionarios a partir da extracción léxica automática é un filtro que combina a frecuencia absoluta do lema (superior a 4) coa probabilidade da súa tradución máis probable (maior ou igual que 0,3, pero diferente de 0,5). Aplicando este criterio, que combina precisión e cobertura, a fiabilidade das entradas xeradas alcanza o 91,4% (co 54,7% de cobertura) (Gómez Guinovart e Sacau Fontenla, 2004b).

5. Fase de postedición

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 (T1 e T2), 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 do Dicionario CLIG recolle un total de 5.324 entradas e 7.998 traducións e está dispoñible na web desde maio de 2005 no enderezo http://sli.uvigo.es/CLIG.

O Dicionario CLIG está almacenado nun formato interno codificado en XML, de acordo coa seguinte definición de tipo de documento (DTD):

(15)
```
<!ELEMENT clig (entrada+)>
<!ELEMENT entrada (lema, super_cat+)>
<!ELEMENT lema (#PCDATA)>
<!ELEMENT super_cat (categoria, acepcion+)>
<!ELEMENT categoria (#PCDATA)>
<!ELEMENT acepcion (plurilex?, traducion, exemplo)>
<!ELEMENT traducion (#PCDATA)>
<!ELEMENT plurilex (#PCDATA)>
<!ELEMENT exemplo (en, gl, fonte)>
<!ELEMENT en (#PCDATA)>
<!ELEMENT gl (#PCDATA)>
<!ELEMENT fonte (#PCDATA)>
```

Consonte esta definición, 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 DTD) 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 en 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. Deste xeito, cada entrada do dicionario pode incluír unha ou máis categorías gramaticais con unha ou máis traducións, sendo codificada internamente como se ilustra a seguir mediante un exemplo:

(16)

<entrada>
<lema>peal</lema>
<super_cat>
<categoria>intransitive verb</categoria>
<acepcion>
<traducion>repinicar</traducion>
<exemplo> <en>The city is shaken with the firing of shells; the bells of the cathedral clash and @peal#.</en> <gl>A cidade é sacudida polos disparos de proxectís; as campás da catedral abouxan e @repinican#.</gl> <fonte>GAL (619)</fonte> </exemplo>
</acepcion>
</super_cat>
<super_cat>
<categoria>noun</categoria>
<acepcion>
<traducion>estrondo</traducion>
<exemplo> <en>As they ascended, Rip every now and then heard long rolling @peals#, like distant thunder, that seemed to issue out of a deep ravine, or rather cleft, between lofty rocks, toward which their rugged path conducted.</en> <gl>A medida que ían ascendendo, Rip sentía cada pouco tempo uns longos @estrondos#, coma tronos na distancia, que parecían vir dun profundo desfiladeiro, ou máis ben dunha greta entre rochas elevadas cara ás que conducía aquel camiño esgrevio.</gl> <fonte>RIP (79)</fonte> </exemplo>
</acepcion>
<acepcion>
<plurilex>peal of laughter</plurilex>
<traducion>gargallada</traducion>
<exemplo> <en>This welcome ended in a soft @peal of# mirthless @laughter# as Heron salaamed and then began to poke the ground with his cane.</en> <gl>Esta benvida rematou nunha @gargallada# sen ledicia mentres Heron saudaba cerimoniosamente e remexía no chan cun bastón.</gl> <fonte>RET (1599) </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 amosa nas seguintes ilustracións de uso da interfaz de consulta:

6. Conclusións

O Dicionario CLUVI Inglés-Galego (CLIG), elaborado no Seminario de Lingüística Informática (SLI) da Universidade de Vigo, é unha obra con características propias dentro da tradición lexicográfica galega. O CLIG está baseado nun corpus representativo de textos ingleses traducidos ao galego que forma parte do Corpus Lingüístico da Universidade de Vigo (CLUVI). Todas as palabras inglesas que aparecen como lemas das entradas do CLIG están documentadas nos textos en inglés traducidos ao galego recompilados no CLUVI. Alén diso, todas as traducións galegas que se recollen no CLIG para esas palabras son traducións reais identificadas nas versións galegas dos textos ingleses incluídas no CLUVI. Para cada tradución seleccionada, o CLIG fornece un exemplo real de uso documentado no CLUVI. A primeira versión do

CLIG recolle un total de 5.324 entradas e 7.998 traducións, e está dispoñible na web no enderezo http://sli.uvigo.es/CLIG. As persoas interesadas en consultar máis exemplos de uso dunha tradución, poden utilizar a interfaz web ao CLUVI dispoñible no enderezo http://sli.uvigo.es/CLUVI. Esta utilidade permite facer buscas simples e complexas (con comodíns) de palabras illadas ou de secuencias de palabras, e observar as equivalencias plurilingües dos termos pescudados nos seus contextos de uso en traducións reais e documentadas

Neste artigo presentamos o proceso que seguimos para a creación da versión 1.0 do CLIG (2005), desde o Corpus CLUVI ata o resultado final do dicionario posto a disposición do público na rede. Polas súas dimensións reducidas, en canto ao número de lemas e á cantidade de traducións, esta primeira versión do dicionario está moi limitada como ferramenta de consulta na tradución profesional. Con todo, esperamos que a súa presentación e divulgación resulte de utilidade desde o principio, tanto polas características novidosas da súa concepción e deseño, coma pola ausencia dun dicionario inglés-galego válido para o uso profesional no panorama de lexicografía galega dos nosos días. O traballo do SLI no CLIG a partir deste momento hase de centrar na revisión e ampliación das entradas e equivalencias do dicionario, ao tempo que se mellora e amplía a sección de córpora paralelos inglés-galego do Corpus CLUVI que supón a base empírica de coñecemento léxico na que se alicerza o dicionario.

7. Bibliografía

Brown, R.D., J.G. Carbonell e Y. Yang (2000). "Automatic dictionary extraction for cross-Language Information Retrieval". En J. Véronis (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, Dordrecht, pp. 275-298.

Gómez Guinovart, X. e E. Sacau Fontenla (2004a). "Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo)". En Teresa Lino et al. (ed.), *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1179-1182.

Gómez Guinovart, X. e E. Sacau Fontenla (2004b). "Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos". *Procesamiento delLenguaje Natural*, 33, pp. 133-140.

Gómez Guinovart, X. (coord.), Sacau Fontenla, E., Torres Padín, A., Díaz Rodríguez, E. e Álvarez Lugrís, A. (2005). *Dicionario CLUVI inglés-galego, versión 1.0*. Seminario de Lingüística Informática, Universidade de Vigo. [http://sli.uvigo.es/CLIG/]

Hiemstra, D. (1998). "Multilingual Domain Modeling in Twenty-One. Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus". En *Proceedings of the 8th CLIN Meeting*, pp. 41-58.

Och, F.J. e N. Hermann (2000). "Improved Statistical Alignment Models". *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447.

Och, F.J. e H. Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistic*s, 29(1), pp. 19-51.

Savourel, Y. (ed.) (2004). *TMX 1.4b Specification*. Localisation Industry Standards Association. [http://www.lisa.org/tmx/tmx.htm]

Simões, A.M. e J.J. Almeida (2003). "NATools: A Statistical Word Aligner Workbench". *Procesamiento del Lenguaje Natural*, 31, pp. 217-224.

Tiedemann, J. (2003). *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Acta Universitatis Upsaliensis, Upssala.

Turcato, D. e F. Popowich (2001). "What is Example-Based Machine Translation?". En *Proceedings of the Workshop on Example-Based Machine Translation (MT Summit VIII)*.

Vintar, Š. (2001). "Using Parallel Corpora for Translation-oriented Term Extraction". *Babel Journal*, 47(2), pp. 121-132.