TÉCNICAS DE PROCESAMENTO LINGÜÍSTICO-COMPUTACIONAL DE CORPUS PARALELOS NO *CLUVI* (CORPUS LINGÜÍSTICO DA UNIVERSIDADE DE VIGO)

XAVIER GÓMEZ GUINOVART
ELENA SACAU FONTENLA
Seminario de Lingüística Informática, Universidade de Vigo

1. INTRODUCIÓN[1]

O Corpus CLUVI (Corpus Lingüístico da Universidade de Vigo) é un corpus textual aberto de rexistros especializados de lingua galega contemporánea. No seu estado actual de desenvolvemento, os textos da sección principal do CLUVI pertencen a catro rexistros especializados (dos ámbitos xurídico-administrativo, xornalístico, informático e literario) e a catro combinacións lingüísticas en relación ao galego (monolingüe galego, tradución galego-español, tradución francés-galego e tradución inglés-galego), e posúen unha extensión total aproximada de 6 millóns de palabras. Este conxunto de textos do CLUVI atópase repartido en cinco subcorpus, cada un deles de arredor dun millón de palabras: o corpus paralelo TECTRA de textos literarios inglés-galego, o corpus paralelo FEGA de textos literarios francés-galego, o corpus paralelo LEGA de textos xurídico-administrativos galego-español, o corpus monolingüe XIGA de textos sobre informática en galego e o corpus monolingüe MEGA de linguaxe dos medios de comunicación social. A ampliación do CLUVI cos textos paralelos tetralingües inglés-galego-francés-español de *O Correo da Unesco*, e con textos paralelos cinematográficos inglés-galego está en fase de elaboración.

Nesta exposición, presentamos a metodoloxía desenvolvida no SLI (Seminario de Lingüística Informática da Universidade de Vigo) para a construción e procesamento do CLUVI, mostrando en concreto as solucións adoptadas para a anotación dos seus corpus paralelos TECTRA, FEGA e LEGA, e para a súa dispoñibilización na web (no enderezo http://sli.uvigo.es/CLUVI/).

A anotación dos corpus paralelos do CLUVI ofrece dúas vertentes diferenciadas: a etiquetaxe morfosintáctica e lematización do léxico, e a codificación dos aliñamentos de tradución. Na lematización e etiquetaxe morfosintáctica dos textos empregamos o estándar XML e os etiquetarios morfosintácticos propostos por EAGLES (Expert Advisory Group on Language Engineering Standards). Na etiquetaxe dos textos en galego empregamos o etiquetario morfosintáctico elaborado polo SLI consonte as directrices estándares europeas elaboradas por EAGLES.

O aliñamento dos textos paralelos almacénase en formato TMX, o estándar para a codificación en XML de memorias de tradución independentemente da aplicación

[1] Este traballo foi financiado pola Xunta de Galicia, dentro do proxecto "Estudio e adquisición de recursos básicos de lingüística computacional do galego para a elaboración e mellora de aplicacións informáticas de tecnoloxía lingüística" (ref. PGIDT01TICC06E); e mais 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), proxecto cofinanciado pola Dirección Xeral de I+D da Xunta de Galicia e pola Universidade de Vigo. Máis información en http://webs.uvigo.es/sli.

utilizada. Na exposición, amosaremos as solucións adoptadas no Corpus CLUVI para a codificación en TMX das equivalencias de tradución cando a correspondencia entre orixinal e tradución non é directa por mor da omisión, adición ou reordenamento de frases na tradución. Amosaremos tamén a solución adoptada para unir na codificación en TMX a información morfosintáctica e a información sobre as equivalencias de tradución; e as técnicas lingüístico-computacionais empregadas para a xeración de dicionarios bilingües baseados nos datos do corpus.

Finalmente, presentaremos a aplicación web deseñada polo SLI para a consulta pública dos corpus paralelos do CLUVI. A consulta pública dos corpus paralelos do CLUVI, a través da web do SLI (dispoñíbel en http://sli.uvigo.es/CLUVI/), permite examinar as equivalencias bilingües en textos reais con finalidades académicas de investigación e docencia, e tamén como ferramenta para a tradución.

2. ANOTACIÓN DAS EQUIVALENCIAS DE TRADUCIÓN

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). A máis, 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 corpus 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 corpus 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 1.4 (Savourel, 2002) para representar as correspondencias que non son 1:1 e os reordenamentos codificados no corpus paralelo. Os aspectos traductolóxicos codificados no Corpus CLUVI pódense agrupar en tres categorías -omisións, adicións e reordenamentos-, e son etiquetados mediante unha versión adaptada dos elementos `<hi>` e `<ph>`, parte do TMX 1.4.

2.1. Omisión

Na omisión, hai un anaco do texto de partida que non ten correspondencia no texto de chegada, isto é, unha frase ou parte dunha frase non é traducida. A omisión codifícase nos corpus paralelos do CLUVI co elemento `<hi>`. Consonte coa

especificación TMX 1.4, o elemento <hi> (de nome derivado do inglés “highlight”) “delimits a section of text that has special meaning, such as a terminological unit, a proper name, an item that should not be modified, etc.” (Savourel, 2002). Na especificación do CLUVI, baseada na TMX, o elemento <hi> marca no texto de partida o elemento que se omite no texto de chegada. Indicamos este uso da etiqueta <hi> mediante un atributo type caracterizado co valor de "supr". Por exemplo, as seguintes frases aliñadas inglés-galego serían anotadas do seguinte xeito:

‘Hello’, I said.
-Ola.

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

2.2. Adición

A adición na tradución implica unha inserción de fragmentos no texto de chegada que non teñen correspondencias no texto de partida. A adición tamén se codifica no CLUVI co elemento <hi>, facendo que este indique o fragmento inserido na tradución. Este uso da etiqueta <hi> distínguese mediante un atributo type caracterizado co valor de "incl". O fragmento engadido incorpórase á unidade de tradución na que está inserido. Cando o novo fragmento é unha oración (ou unha secuencia de oracións), incorpórase quer á unidade de tradución anterior, quer á seguinte, consonte co seu contexto, respectando así o criterio de aliñamento 1:1. Véxase a seguir un exemplo:

‘Hello.’
-Ola - dixen.

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

2.3. Reordenamento

O reordenamento implica 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. O reordenamento codifícase no CLUVI mediante unha combinación dos elementos <hi> e <ph>. Anotamos o fragmento ou a oración movida mediante un elemento <hi> que inclúe un atributo type con valor de "reord" e un atributo x cun valor numérico que actúa de índice. Por outra banda, indicamos cun elemento <ph> o lugar no texto que ocupaba orixinalmente o elemento desprazado. Segundo a especificación TMX 1.4, o elemento <ph> (ou "placeholder") utilízase "to delimit a sequence of native standalone codes in the segment. Standalone codes are codes that are not opening or closing of a pair, for example empty elements in XML" (Savourel, 2002). Na especificación do CLUVI, baseada na TMX, o elemento adaptado <ph> indica o punto de partida do movemento, mentres que e a relación entre o elemento desprazado e o lugar de partida é codificada no elemento <ph> mediante un atributo x que comparte valor co índice codificado no elemento <hi> do segmento movido. Obviamente, a etiqueta que indica o lugar de orixe sempre é unha etiqueta baleira. Como criterio de etiquetaxe na codificación do CLUVI, e coa finalidade de evitar incoherencias entre as distintas persoas que participan na codificación do corpus, os segmentos reordenados sempre son desprazados en dirección ao inicio do texto. En consecuencia, no CLUVI non hai ningunha secuencia semellante a <ph x="n"/> [...] <hi type="reord" x="n">Reordered element</hi>; no canto diso, as secuencias son sempre así: <hi type="reord" x="n">Reordered element</hi> [...] <ph x="n"/>. Velaquí un exemplo sinxelo de codificación dun reordenamento:

'The front door!' she said in this loud whisper. 'It's them!'
-A porta de fóra. ¡Son eles! - murmurou bastante alto.

```
<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 codificaríanse consecutivamente cos atributos <x="2">, <x="3">, ..., <x="n">, como se mostra no seguinte exemplo:

'Leave him alone, hey' Sunny said. 'C'mon, hey. We got the dough he owes us. Let's go.'
-Déixao. Imos logo. Xa témo-lo que nos debe - dicía Sunny.

```
<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></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. ANOTACIÓN MORFOSINTÁCTICA PARALELA

Para a etiquetaxe morfosintáctica dos corpus paralelos do CLUVI empregamos o estándar XML e o etiquetario morfosintáctico elaborado polo SLI (Aguirre et al., 2003) consonte as directrices de EAGLES (Leech and Wilson 1996; Monachini and Calzolari 1996). O sistema probabilístico de etiquetaxe e desambiguación empregado no CLUVI, desenvolvido conxuntamente polo SLI e Imaxin Software, utiliza un léxico computacional do galego que contén as especificacións morfosintácticas definidas no etiquetario do SLI. Os textos en inglés do CLUVI son etiquetados co programa Trigram's Tags (TnT) (Brants, 2000), como se fixo no corpus paralelo IJS-ELAN (Erjavec, 2002). O etiquetario do SLI segue as propostas de EAGLES tocante ás categorías gramaticais e aos trazos morfosintácticos que cómpre distinguir. No seu deseño, non nos limitamos a seguir as directrices xerais de EAGLES, senón que aplicamos estrictamente o esquema de atributos e valores recomendado por Leech e Wilson (1996), axeitándoo para o galego de modo análogo ao que se ten feito para outras linguas, como o italiano e o alemán (Teufel 1996).

Outro aspecto fundamental do deseño do etiquetario SLI do galego é o estabelecemento das correspondencias co etiquetario intermedio de EAGLES. O etiquetario intermedio é unha representación lingüisticamente neutral que describe os trazos lingüísticos (descritos en forma de pares atributo-valor) recollidos nun etiquetario, de xeito que se poidan facer corresponder doadamente coas marcas doutro conxunto de etiquetas (Leech e Wilson 1996). Grazas ao etiquetario intermedio, podemos etiquetar os textos galegos cun etiquetario definido segundo a terminoloxía

gramatical propia da lingua galega, e os textos ingleses cun etiquetario propio da tradición da lingüística de corpus inglesa, e finalmente converter ambos os dous ao etiquetario intermedio estándar de EAGLES. Deste xeito, pódense observar as correspondencias entre a información gramatical dos textos en galego e a dos textos en inglés na sección inglés-galego do corpus paralelo CLUVI. Estas correspondencias poden así ser aproveitadas posteriormente na extracción de información léxica bilingüe contextual e fraseolóxica. De forma máis xeral, a correspondencia do etiquetario galego co etiquetario intermedio permítenos reutilizar os textos etiquetados en aplicacións adaptadas ao estándar EAGLES.

4. PROCESAMENTO LÉXICO E TERMINOLÓXICO

Unha das liñas de investigación máis prometedoras desenvolvidas no SLI é a xeración de dicionarios bilingües baseados nos datos do CLUVI. Neste momento, a nosa investigación céntrase na extracción léxica dun dicionario bilingüe inglés-galego a partir do corpus paralelo TECTRA, utilizando o programa de aliñamento léxico NATools (Simões and Almeida 2003) baseado no aliñador Twente (Hiemstra, 1998). Estamos a levar a cabo diversos experimentos para mellorar a precisión dos resultados da extracción bilingüe. Neste sentido, estamos a explorar a posibilidade de empregar unha versión "limpa" dos corpus paralelos, pre-editada para facilitar a extracción bilingüe mediante a eliminación de certas palabras gramaticais de alto índice de frecuencia e dos segmentos de texto que están marcados no corpus como omisións ou como adicións. Paralelamente, estamos tratando de estabelecer un sistema automático de filtraxe dos resultados da extracción automática, para podermos seleccionar os resultados máis fiábeis de acordo con criterios como a frecuencia de aparición da palabra na lingua fonte ou como o índice de probabilidade do aliñamento asociado coas palabras da lingua termo (Vintar, 2001). Tamén estamos a comprobar a utilidade da etiquetaxe morfosintáctica do CLUVI para resolver problemas de ambigüidade nos pares léxicos bilingües extraídos.

Outros proxectos de investigación a curto termo baseados no Corpus CLUVI oriéntanse á extracción terminolóxica bilingüe galego-español a partir do corpus paralelo xurídico-administrativo LEGA, e á extracción léxica e terminolóxica plurilingüe con máis de dúas linguas a partir do corpus paralelo tetralingüe inglés-galego-francés-español de *O Correo da Unesco*.

5. APLICACIÓNS E SERVIZOS WEB

Desde setembro de 2003, o SLI permite pescudar nos corpus paralelos do CLUVI a través dunha interfaz web de consulta dispoñíbel no enderezo http://sli.uvigo.es/CLUVI/. Os corpus paralelos cos que traballa a aplicación web do SLI están almacenados na devandita especificación CLUVI do XML. A ferramenta de busca e visualización, deseñada en PHP polo SLI, está concibida para realizar buscas bilingües en textos etiquetados conformes coa especificación TMX, incluída a especificación usada do CLUVI. 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 posíbel tamén realizar consultas autenticamente bilingües, isto é, consultas a partir de dous termos, un de cada lingua.

Outro dos servizos que imos ofrecer a través da web do SLI é a dispoñibilización dos corpus paralelos do CLUVI como memoria de tradución en Internet de libre consulta, accesíbel na rede para ambientes informáticos de tradución asistida por ordenador baseados en memorias de tradución do estilo de Trados ou DéjàVu (Simões et al., 2004).

6. CONCLUSIÓNS

Neste artigo presentamos a metodoloxía desenvolvida no SLI (Seminario de Lingüística Informática da Universidade de Vigo) para a construción e procesamento do Corpus CLUVI, mostrando en concreto as solucións adoptadas para a anotación morfosintáctica e traductolóxica dos seus corpus paralelos. Presentamos tamén as técnicas lingüístico-computacionais empregadas para a xeración de dicionarios bilingües baseados nos datos do corpus. Finalmente, presentamos a aplicación web deseñada pelo SLI para a consulta pública dos corpus paralelos do CLUVI (dispoñíbel en http://sli.uvigo.es/CLUVI/). Con este traballo, pretendemos contribuír ao avance da investigación e desenvolvemento nas áreas da lingüística de corpus e das tecnoloxías lingüísticas da lingua galega, e fornecer dunha serie de recursos básicos de lingüística informática que consideramos imprescindíbeis para a normalización da nosa lingua.

7. REFERENCIAS BIBLIOGRÁFICAS

AGUIRRE, José Luis, Alberto ÁLVAREZ LUGRÍS & Xavier GÓMEZ GUINOVART (2003): "Aplicación do etiquetario morfosintáctico do SLI ó corpus de traduccións TECTRA". *Viceversa*. 7-8. 189-212.

BRANTS, Thorsten (2000): "TnT: A Statistical Part-of-Speech Tagger". *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*.

ERJAVEC, Tomaz (2002): "Compiling and Using the IJS-ELAN Parallel Corpus". *Informatica*. 26. 299-307.

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

LEECH, Geoffrey & Andrew WILSON (1996): *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Guidelines.

MONACHINI, Monica & Nicoletta CALZOLARI (1996). *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora*. EAGLES Guidelines.

SAVOUREL, Yves (2002): *TMX 1.4a Specification*. Technical Report. Localisation Industry Standards Association.

SIMÕES, Alberto M. & J. João ALMEIDA, (2003): "NATools: A Statistical Word Aligner Workbench". *Procesamiento del Lenguaje Natural*. 31. 217-224.

SIMÕES, Alberto M., J. João ALMEIDA & Xavier GÓMEZ GUINOVART (2004): "Memórias de tradução distribuídas". In José Ramalho, C. and Simões, A. (eds.), XATA2004 - XML, Aplicações e Tecnologias Associadas (pp. 59-68). Porto: U. of Porto.

TEUFEL, Simone (1996): *ELM-DE: EAGLES Specifications for German Morphosyntax*. EAGLES Guidelines.

VINTAR, Špela (2001): "Using parallel corpora for translation-oriented term extraction". *Babel Journal*. 47:2. 121-132.