XXX aniversario
novática
UPGRADE Revista de la Asociación de Técnicos de Informática

Nº 175, mayo-junio 2005, año XXXI

# sumario



**Monografía del próximo número:**
***"Estandardización y normalización en Seguridad"***

***Novática***, revista fundada en 1975 y decana de la prensa informática española, es el órgano oficial de expresión y formación continua de **ATI** (Asociación de Técnicos de Informática). ***Novática*** edita también ***UPGRADE***, revista digital de **CEPIS** (*Council of European Professional Informatics Societies*), en lengua inglesa, y es miembro fundador de **UPENET** (***UP****GRADE* ***E****uropean* ***NET****work*)

**<http://www.ati.es/novatica/>**
**<http://www.upgrade-cepis.org/>**

**ATI** es miembro fundador de **CEPIS** (*Council of European Professional Informatics Societies*) y es representante de España en **IFIP** (*International Federation for Information Processing*); tiene un acuerdo de colaboración con **ACM** (*Association for Computing Machinery*), así como acuerdos de vinculación o colaboración con **AdaSpain**, **AI2** y **ASTIC.**



**Coordinación Editorial, Redacción Central y Redacción ATI Madrid**
Padilla 66, 3º, dcha., 28006 Madrid
Tlfn.914029391; fax.913093685 <novatica@ati.es>
**Composición, Edición y Redacción ATI Valencia**
Av. del Reino de Valencia 23, 46005 Valencia
Tlfn./fax 963330392 <secreval@ati.es>
**Administración y Redacción ATI Cataluña**
Ciudad de Granada 131, 08018 Barcelona
Tlfn.934125235; fax 934127713 <secregen@ati.es>
**Redacción ATI Andalucía**
Isaac Newton, s/n, Ed. Sadiel,
Isla Cartuja 41092 Sevilla, Tlfn./fax 954460779 <secreand@ati.es>
**Redacción ATI Aragón**
Lagasca 9, 3-B, 50006 Zaragoza.
Tlfn./fax 976235181 <secreara@ati.es>
**Redacción ATI Asturias-Cantabria** <gp-astucant@ati.es>
**Redacción ATI Castilla-La Mancha** <gp-clmancha@ati.es>
**Redacción ATI Galicia**
Recinto Ferial s/n, 36540 Silleda (Pontevedra)
Tlfn.986581413; fax 986580162 <secregal@ati.es>
**Suscripción y Ventas**
<http://www.ati.es/novatica/interes.html>, o en ATI Cataluña o ATI Madrid
**Publicidad**
Padilla 66, 3º, dcha., 28006 Madrid
Tlnf.914029391; fax.913093685 <novatica.publicidad@ati.es>
**Imprenta**
Derra S.A., Juan de Austria 66, 08005 Barcelona.
**Depósito legal:** B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC

**Portada:** Antonio Crespo Foix / © ATI 2005
**Diseño:** Fernando Agresta / © ATI 2005

# Procesamiento y aplicaciones de los corpus paralelos

Xavier Gómez Guinovart
*Seminario de Lingüística Informática, Universidad de Vigo*

**<sli@uvigo.es>**

***Resumen:*** *un corpus paralelo es una colección de bitextos, siendo un bitexto el texto constituido por un texto y su traducción. En Lingüística Computacional, los corpus paralelos suelen anotarse con información sobre las equivalencias de traducción entre los segmentos de cada una de las versiones traducidas, empleando para ello los estándares XCES* (XML Corpus Encoding Standard) *o TMX* (Translation Memory eXchange) *basados en XML* (eXtensible Marrkup Language)*. El procesamiento de estos corpus paralelos enriquecidos permite elaborar aplicaciones en los campos de las concordancias bilingües, la extracción léxica y la traducción automática.*

***Palabras clave:*** *corpus paralelos, TMX, traducción automática, XCES, XML.*

## 1. Introducción

El término **corpus** se acostumbra a utilizar en el ámbito de la Lingüística Computacional para referirse a una colección electrónica de textos escritos u orales recopilada con el fin de elaborar alguna aplicación en el ámbito de las industrias de la lengua. Estos corpus textuales pueden mostrar características muy diversas, según sea su contenido textual y las aplicaciones para las que haya sido diseñado. Por ejemplo, muchas aplicaciones de las tecnologías del habla (como las agendas por voz, los sistemas de dictado o las utilidades de control oral del teléfono móvil) basan sus algoritmos de identificación de palabras en **corpus orales** de aprendizaje constituidos por enunciados grabados en condiciones acústicas controladas. Muy distintos de estos corpus orales son los corpus compuestos por textos escritos orientados, por ejemplo, a la elaboración de algoritmos para la clasificación de documentos en aplicaciones tales como el filtrado del correo electrónico basura o el direccionamiento selectivo de textos (*text routing*). Dentro de esta categoría de **corpus escritos** es donde hay que situar los corpus escritos plurilingües, constituidos por textos escritos en dos o más idiomas, y creados para satisfacer diversas necesidades relacionadas con el procesamiento informático del multingüismo. Los **corpus paralelos** representan una especialización de los corpus plurilingües, en la que los textos recopilados son traducciones los unos de los otros. Técnicamente hablando, un corpus paralelo es una colección de *bitextos*, siendo un bitexto el texto constituido por un texto y su traducción [1][2].

Sean del tipo que sean, los **corpus crudos**, es decir, los corpus que contienen exclusivamente las palabras y signos de puntuación de los textos originales, tienen una utilidad bastante limitada para la Lingüística Computacional. Para que los corpus empiecen a ser verdaderamente útiles en el desarrollo de aplicaciones, se debe enriquecer el texto de los documentos originales con diversos tipos de información lingüística, a ser posible, de manera automática o semiautomática. Estos corpus enriquecidos con información lingüística reciben la denominación de **corpus etiquetados**, ya que la información añadida al texto original se incorpora contenida dentro de etiquetas, en general, empleando alguna especificación del lenguaje XML (*eXtensible Marrkup Language*).

En la actualidad, la anotación morfosintáctica (o sea, la incorporación de una indicación de su categoría gramatical a cada palabra de un texto) puede efectuarse con un alto grado de automatización, por lo que resulta el tipo de anotación que se encuentra con más frecuencia en los corpus etiquetados. Los programas informáticos de etiquetación gramatical automática (o *taggers*) analizan las palabras en el texto y las devuelven acompañadas de su etiqueta con la categoría morfosintáctica más probable en su contexto, con una tasa de acierto muy elevada cercana al ciento por ciento [3].

## 2. Corpus paralelos y estándares de codificación

Aunque también es posible incorporar información morfosintáctica en los corpus paralelos, resulta mucho más útil enriquecer este tipo de corpus con información sobre las equivalencias de traducción entre los segmentos (palabras, frases u otro tipo de unidades textuales) de cada una de las versiones traducidas. Este proceso suele denominarse **alineamiento**, y el corpus paralelo así enriquecido suele recibir el nombre de **corpus alineado**. Los estándares XML más importantes utilizados para codificar estos alineamientos son el XCES (*XML Corpus Encoding Standard*) y el TMX (*Translation Memory eXchange*).

El XCES [4] es una adaptación a XML en fase beta de las directrices europeas de EAGLES (*Expert Advisory Group on Language Engineering Standards*) sobre anotación de corpus en SGML (*Standard Generalized Markup Language*) conocidas como CES (*Corpus Encoding Standard*) [5]. De acuerdo con este estándar, la información sobre los alineamientos entre los segmentos de dos textos se codifica en un documento independiente que contiene únicamente los enlaces entre los documentos que se han alineado. En la **figura 1**, se muestra un ejemplo de alineamiento francés-inglés a nivel de frase entre los documentos DOC1 (en francés) y DOC2 (en inglés), con la información sobre las equivalencias de traducción (qué frase o frases de DOC2 son traducción de cada frase de DOC1) codificadas en el documento ALIGN_DOC.

Un excelente ejemplo de corpus paralelo accesible a través de la web codificado en formato XCES es el corpus OPUS [6], disponible para consulta en la dirección web <http://logos.uio.no/opus/>.

El formato TMX (*Translation Memory eXchange*) [7] es un estándar para la codificación en XML de memorias de traducción independientemente de la aplicación utilizada.

El concepto de **memoria de traducción** está relacionado con la traducción asistida por ordenador y, más concretamente, con los programas denominados entornos de traducción (*translation environments*), como Trados (<http:// www.trados.com>), DéjàVu (<http://www.atril.com>), SDLX (<http://www. sdlintl.com/sdlx>), Transit (<http://www.star-group.net>), o Passolo (<http://www.passolo.com>), este último orientado a la localización de software. Los entornos de traducción integran en un único producto informático un procesador de textos especialmente diseñado para traducir, un conjunto de diccionarios bilingües, herramientas para la gestión terminológica (creación y mantenimiento de glosarios, consulta automática de glosarios durante la traducción, extracción automática de terminología...), y una utilidad de memoria de traducción. La memoria de traducción es una base de datos donde se almacenan la versión original y traducida de cada una de las frases que se traducen en el marco de la aplicación. Cuando se está traduciendo una frase, el programa detecta automáticamente

> Los corpus lingüísticos útiles para el desarrollo de aplicaciones son los **corpus etiquetados**

```
DOC1: <cesDoc version=»3.24">
          <cesHeader version="2.3">
               ...
          </cesHeader>
          <text>
             <body id="b1">
                <div type=sample id="d1">
          <p id="d1p1">
            <s id="d1p1s1">J'ai donc dû choisir un autre métier
            et j'ai appris à piloter des avions.</s>
            <s id="d1p1s2">J'ai volé un peu partout dans le monde.</s>
            <s id="d1p1s3">Et la géographie, c'est exact, m'a beaucoup servi.</s>
            <s id="d1p1s4">Je savais reconnaître, du premier coup d'oeil, la Chine
            de l'Arizona.</s>
            <s id="d1p1s5">C'est très utile, si l'on est égaré pendant la nuit.</s>
          </p>
               </div>
             </body>
          </text>
          </cesDoc>

    DOC2: <cesDoc version="3.24">
          <cesHeader version="2.3">
             ...
          </cesHeader>
          <text>
             <body id="b1">
                <div type=sample id="d1">
          <p id="d1p1">
            <s id="d1p1s1">So then I chose another profession, and learned to
            pilot aeroplanes.</s>
            <s id="d1p1s2">I have flown a little over all parts of the world;
            and it is true that geography has been very useful to me.</s>
            <s id="d1p1s3">At a glance I can distinguish China from Arizona.</s>
            <s id="d1p1s4">If one gets lost in the night, such knowledge is
            valuable.</s>
          </p>
               </div>
             </body>
          </text>
          </cesDoc>

    ALIGN_DOC:
          <cesAlign type=sent version=1.6>

          <cesHeader version="2.3">
             ...
            <translations>
              <translation trans.loc="text-f.sgml" lang=fr wsd="ISO8859-1" n=1>
              <translation trans.loc="text-e.sgml" lang=en wsd="ISO8859-1" n=2>
            </translations>
          </cesHeader>

          <linkList>

            <!— sentence alignments —>
            <linkGrp domains="d1 d1" targType="s">
              <link xtargets="d1p1s1 ; d1p1s1">
              <link xtargets="d1p1s2 d1p1s3 ; d1p1s2">
              <link xtargets="d1p1s4 ; d1p1s3">
              <link xtargets="d1p1s5 ; d1p1s4">
            </linkGrp>

          </linkList>

          </cesAlign>
```

**Figura 1.** Ejemplo de alineamiento en CES.

si esa misma frase u otra similar ya fue traducida con anterioridad, con el objeto de que se pueda reutilizar la traducción sin necesidad de reescribirla completamente, haciendo las modificaciones que se consideren más oportunas. En 1997 la industria creó e impulsó el estándar TMX para permitir el intercambio de memorias de traducción entre los distintos entornos de traducción. Con ciertas salvedades, un corpus paralelo alineado equivale a una memoria de traducción y, en la práctica, existe un buen número de corpus paralelos alineados codificados en TMX, con la ventaja adicional de que los corpus así etiquetados pueden ser ser empleados como memorias de traducción para alimentar los entornos de traducción para satisfacción de sus usuarios.

A modo de ilustración de las características generales de este formato, en la **figura 2** se muestra el alineamiento inglés-gallego en TMX simplificado de las tres primeras frases de *La Perla* de Steinbeck en el original inglés y en su traducción al gallego.

A diferencia del formato (X)CES, en TMX tanto el original, como su traducción, como la información sobre los alineamientos, se incluyen en un único archivo, en el cual los segmentos originales y traducidos discurren literalmente en paralelo, rodeados por etiquetas que explicitan su adscripción lingüística y sus equivalencias.

Un ejemplo ilustrativo de corpus paralelo codificado en TMX disponible en la red es el Corpus CLUVI (Corpus Lingüístico de la Universidad de Vigo) [8], elaborado por el grupo de investigación del Seminario de Lingüística Informática de nuestra Universidad, accesible en la dirección web <http://sli.uvigo.es/CLUVI>.

Dependiendo del tipo de textos alineados, el alineamiento a nivel de frase u oración, en el que se establecen las equivalencias de traducción entre las frases de los textos originales y las frases de sus respectivas traducciones, puede llevarse a cabo de manera automática con bastante fiabilidad.

Cuanto más literales sean las traducciones que se pretende alinear, y más cercanas sean las lenguas del original y de la traducción, mejores serán los resultados. La fiabilidad de alineamiento automático desciende cuando las lenguas no están emparentadas y cae en picado tratándose de traducción literaria.

## 4. Aplicaciones de los corpus paralelos

Las principales aplicaciones de los corpus paralelos son las concordancias bilingües, la extracción léxica y la traducción automática. Las aplicaciones de concordancia bilingüe son sistemas que permiten realizar búsquedas de traducciones en los textos alineados. Las búsquedas deben poderse hacer tanto en el texto original como en el texto de la traducción. Así, si un corpus paralelo alineado contiene textos traducidos entre la lengua A y la lengua B, el programa de concordancia debe permitirnos buscar una expresión en lengua A y ver las distintas maneras en que fue traducida en la lengua B en cada contexto. Por otra parte, el programa de concordancia también nos debe permitir buscar una expresión en lengua B y ofrecernos todos los contextos en que aparece B junto a sus contextos originales en la lengua A.

Finalmente, un buen programa de concordancia bilingüe debe permitirnos hacer búsquedas realmente bilingües, es decir, búsquedas del tipo: "*deseo consultar las frases del original y su traducción en aquellos ca-*

```
<?xml version=»1.0" ?>
<!DOCTYPE tmx SYSTEM "tmx11simp.dtd">
<tmx version="1.1">
<header creationtool="TRANS Suite 2000" creationtoolversion="1.4.2" segtype="sentence"
o-tmf="CTMTS2000" adminlang="gl" srclang="en" datatype="empty">
</header>
<body>
<tu>
<tuv 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 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 lang="en">
<seg>They tell of Kino, the fisherman, and of his wife, Juana, and of the baby,
Coyotito.</seg>
</tuv>
<tuv lang="gl">
<seg>Fálase de Kino, o pescador e da súa muller, Juana, e do neno, Coyotito.</seg>
</tuv>
</tu>
<tu>
<tuv lang="en">
<seg>And because the story has been told so often, it has taken root in every man's
mind.</seg>
</tuv>
<tuv 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>
```

**Figura 2.** Ejemplo de alineamiento en TMX.

| | | |
|---|---|---|
| CAR (1202) | She was listening to a **clatter** in the hall. | Estaba a escoitar un **rebumbio** no vestíbulo. |
| TER (973) | It was nearly five o'clock when he reached Cooler's flat, which was over an ice-cream parlour in the American zone: the bar below was full of G.I.s with their girls, and the **clatter** of the long spoons and the curious free unformed laughter followed him up the stairs. | Eran case as cinco cando chegou ao piso de Cooler, sito no portal dunha xeladería, na zona americana. O local estaba ateigado de soldados coas súas parellas e o **tintín** das longas culleres e os curiosos risos sen control, acompañárono escaleiras arriba. |
| RET (1396) | A dusk like that of the outer world obscured his mind as he heard the mare's hoofs **clattering** along the tramtrack on the Rock Road and the great can swaying and rattling behind him. | Un solpor, semellante ao do mundo exterior, escurecía a súa mente mentres escoitaba os cascos da besta **resoando** ao longo da liña do tranvía de Rock Road, ao tempo que o gran cántaro abalaba e tintinaba detrás súa. |
| GAL (429) | A typically Galician sound is the continual **clatter** of wooden sabots on granite and asphalt; they are worn by the peasants even in the finest weather. | Un son típico galego é o continuo **troupelear** dos zocos de madeira nas rúas ou no asfalto; os paisanos lévanos mesmo no bo tempo. |
| ESP (1258) | She went up in the mornings to Madame Lebrun's room, braving the **clatter** of the old sewing-machine. | Polas mañás subía ó cuarto de Madame Lebrun desafiando o **tracatá** da vella máquina de coser. |
| ESP (1457) | The crash and **clatter** were what she wanted to hear. | Desexaba escoita-lo estalido e o **estrondo**. |
| LEN (142) | He came **clattering** up to the school door with an invitation to Ichabod to attend a merry-making or "quilting frolic," to be held that evening at Mynheer Van Tassel's; | Chegou **trotando** ata a escola cunha invitación dirixida a Ichabod para asistir a unha festa, ou "charanga", que se ía celebrar aquela mesma noite na casa do honorable Van Tassel. |
| LEN (199) | Not a limb, not a fibre about him was idle; and to have seen his loosely hung frame in full motion, and **clattering** about the room, you would have thought Saint Vitus himself, that blessed patron of the dance, was figuring before you in person. | Non había un só membro do seu corpo que quedase inmóbil, e calquera que vise a súa fraca figura en movemento **pateando** por toda a sala habería pensar que tiña ó mesmo San Vito, bendito patrón do baile, diante dos seus ollos. |
| TEM (1899) | I heard your voices and the **clatter** of plates. | Escoitei as súas voces e o **ruído** dos pratos. |

**Figura 3.** Ejemplo de concordancia bilingüe.

*sos en que en el original aparezca X y en la traducción aparezca Y*".

Los programas de concordancia bilingüe son extremadamente útiles como herramienta de consulta durante una traducción, superando en funcionalidad, realismo y eficacia a los clásicos diccionarios bilingües de palabras. Los programas de concordancia no sólo permiten buscar la traducción de palabras y expresiones de una lista cerrada, como suelen permitir los diccionarios, sino que permiten buscar en los textos la traducción de cualquier tipo de fragmento textual que se pueda expresar en forma de expresiones regulares, al tiempo que facilitan no sólo su traducción, sino su contexto de uso, en traducciones reales y documentadas. Un corpus paralelo es, en este sentido, un diccionario bilingüe de frases de sentido equivalente que ilustra las posibles traducciones de una palabra en cada uno de los contextos documentados. Por ejemplo, en la **figura 3** se recoge el resultado parcial de la consulta en el Corpus CLUVI de las traducciones inglés-gallego de la palabra inglesa *clatter*.

Por motivos evidentes, las concordancias bilingües son también extremadamente interesantes para la enseñanza de segundas lenguas y en la didáctica de la traducción.

Otra aplicación destacable de los corpus paralelos es en la elaboración de repertorios léxicográficos bilingües basados en corpus. La utilidad de los corpus paralelos en este campo es doble: por un lado, es posible identificar en un corpus paralelo, mediante procedimientos informáticos, cuáles son las palabras que aparecen en los textos en lengua original y cuáles son los emparejamientos más frecuentes entre las palabras del original y las palabras de la traducción. Para realizar esta tarea se emplean programas de alineamiento automático a nivel de palabra, que implementan algoritmos capaces de identificar en un corpus paralelo alineado a nivel de frase las equivalencias más probables entre las palabras del original y las de la traducción. El resultado de la alineación de un corpus paralelo a nivel de palabras es un

clatter *intransitive verb*

*** troupelear**

EN Peasants leading ponies came into market, making a tremendous clatter on the granite-paved streets.
GL Algúns campesiños que traían poldros entraron no mercado causando un tremendo troupelear nas rúas lousadas.
*- Fonte: GAL (486)*

*noun*

*** ruído**

EN I heard your voices and the clatter of plates.
GL Escoitei as súas voces e o ruído dos pratos.
*- Fonte: TEM (1899)*

*** rebumbio**

EN She was listening to a clatter in the hall.
GL Estaba a escoitar un rebumbio no vestíbulo.
*- Fonte: CAR (1202)*

**Figura 4.** Ejemplo de diccionario basado en corpus.

| | | |
|---|---|---|
| a. | John Smith will arrive on April 4th | John Smith llegará el 4 de abril |
| b. | Dr F. Jones will arrive on 28.5.2004 | El Dr. F. Jones llegará el 28/5/2004 |
| c. | <nombre> will arrive on <fecha> | <nombre> llegará <fecha> |

**Figura 5:** Traducción automática basada en corpus.

borrador de diccionario bilingüe formado por parejas de palabras supuestamente equivalentes en las dos lenguas del corpus.
Por otro lado, los corpus paralelos también se utilizan para ilustrar las distintas traducciones para una palabra recogidas en el diccionario con ejemplos reales y extraídos de traducciones documentadas. La **figura 4** es un ejemplo de la entrada bilingüe de *clatter* en el Diccionario CLUVI Inglés-Galego, <http://sli.uvigo.es/CLIG>, que se está elaborando a partir de las equivalencias

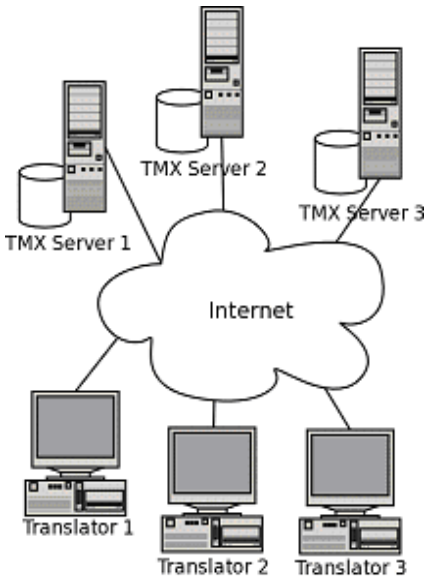


**Figura 6.** Traducción asistida con memorias de traducción.

léxicas bilingües identificadas en el Corpus CLUVI y con ejemplos seleccionados de este mismo corpus [9].

El tercer conjunto de aplicaciones importantes de los corpus paralelos se halla en el ámbito de la traducción por ordenador. Un campo actual de investigación es la traducción automática basada en corpus, en la que los patrones o reglas de traducción que aplicará el programa se extraen (semi)automáticamente por generalización a partir de los ejemplos recogidos en los corpus paralelos analizados [10] o por métodos estadísticos completamente 'ciegos' [11].

Por ejemplo, en la **figura 5** se muestra cómo a partir de un conjunto de frases traducidas semejantes a las de (a) y (b) recogidas en el corpus, es posible deducir una regla de traducción como la que se representa en (c) que sirva para traducir nuevas frases de estructura semejante. También es destacable el uso de los corpus paralelos como memorias de traducción en aplicaciones de traducción asistida por ordenador, especialmente cuando se trata de corpus de consulta libre adaptados para su uso como memorias de traducción distribuidas a través de Internet [12]. La **figura 6** muestra un diagrama de este tipo de aplicaciones distribuidas, en la que los traductores acceden a servidores de memorias de traducción codificadas en TMX para consultar posibles traducciones almacenadas que les sirvan para llevar a cabo su propia traducción en curso.

## 5. Conclusiones

Los corpus paralelos constituyen un recurso de gran importancia para las tecnologías de la lengua. Para su codificación se emplean estándares emparentados con el lenguaje XML, lo cual garantiza sus posibilidades de explotación presente y futura. Su adecuado procesamiento permite desarrollar aplicaciones de la Lingüística Computacional en ámbitos tan diversos como la enseñanza de segundas lenguas, la didáctica de la traducción, la lexicografía, la terminología y la traducción. Sin embargo, el procesamiento de corpus paralelos tiene una historia bastante corta, y la investigación en este campo sigue abierta a nuevos progresos. Sin duda alguna, sus mejores logros están aún por llegar.

## Agradecimientos

Los resultados de este trabajo son fruto de la investigación del SLI y están parcialmente financiados por el Ministerio de Ciencia y Tecnología (MCYT) y el Fondo Europeo de Desarrollo Regional (FEDER), dentro del proyecto "Procesamiento lingüístico-computacional del Corpus Lingüístico de la Universidad de Vigo (CLUVI)" (ref. BFF2002-01385), proyecto cofinanciado por la Dirección Xeral de I+D de la Xunta de Galicia y por la Universidade de Vigo; y por el Ministerio de Industria, Turismo y Comercio, dentro del proyecto "Traducción automática de código abierto para las lenguas del estado español" (ref. FIT-340101-2004-3). Más información en <http://webs.uvigo.es/sli>.

## Referencias

**[1] J. Véronis** (ed.). *Parallel Text Processing: Alignement and Use of Translation Corpora*. Kluwer, Dordrecht, 2000.
**[2] S. Laviosa.** *Corpus-based Translations Studies: Theory, Findings, Applications*. Rodopi, Amsterdam, 2002.
**[3] H. van Halteren** (ed.). *Syntactic Wordclass Tagging*. Kluwer, Dordrecht, 1999.
**[4] N. Ide, K. Suderman.** *XML Corpus Encoding Standard Document XCES 0.2.*, <http://www.cs.vassar.edu/XCES/>, 2002.
**[5] N. Ide, J. Véronis** (2000). *Corpus Encoding Standard*. <http://www.cs.vassar.edu/CES/>, 2000.
**[6] J. Tiedemann, L. Nygaard.** The OPUS corpus - parallel and free. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, Lisboa, 2004.
**[7] Y. Savourel** (ed). *TMX 1.4b Specification*. Localisation Industry Standards Association. <http://www. lisa.org/tmx/tmx.htm>, 2004
**[8] X. Gómez Guinovart, E. Sacau Fontenla.** "Parallel corpora for the Galician language: building and processing of the CLUVI" (Linguistic Corpus of the University of Vigo). *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, Lisboa, 2004.
**[9] X. Gómez Guinovart, E. Sacau Fontenla.** *Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos*. Procesamiento del Lenguaje Natural, 33, pp. 133-140, 2004.
**[10] Harold Somers.** "Example-Based Machine Translation". En **Dale, R. et al.** (eds.), *Handbook of Natural Language Processing*, pp. 611-627. Marcel Dekker, Nueva York, 2000.
**[11] Kevin Knight.** *Statistical Machine Translation Tutorial Workbook*. <http://www.clsp.jhu.edu/ws99/projects/mt/wkbk. rtf>, 1999.
**[12] Alberto Simões, Alberto, Xavier Gómez Guinovart, José João Almeida.** "Distributed Translation Memories implementation using WebServices". *Procesamiento del Lenguaje Natural*, 33, pp. 89-94, 2004.