Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del Dicionario Moderno Inglés-Galego

Alberto Álvarez Lugrís
Universidade de Vigo - Grupo TALG
Xavier Gómez Guinovart
Universidade de Vigo - Grupo TALG

Corpus-based practical bilingual lexicography: planning and building the *Dicionario Moderno Inglés-Galego*

Abstract

One of the lines of research of the TALG Group (Galician Language Tecnologies and Applications) of the University of Vigo is the generation of bilingual dictionaries from lexical equivalences identified in a translation parallel corpus aligned at sentence level. In a sense, the process consists of converting a parallel corpus into a new corpus tagged and aligned down to the word level. In this paper we will describe the compilation of the Dicionario moderno inglés-galego *from the creation of the corpus to the building of the dictionary.*

Keywords

corpus-based lexicography, parallel corpora, bilingual dictionaries,English, Galician

1. Introducción

Los corpus paralelos[1], formados por compilaciones digitales de textos en su versión original y traducida, permiten observar directamente la realidad lingüística plasmada en las traducciones, facilitando el estudio empírico de muchos fenómenos traductológicos que no sería posible examinar de otro modo sin grandes dificultades.

1 Este trabajo fue financiado por el Ministerio de Economía y Competitividad, dentro del proyecto *Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)* (ref. TIN2012-38584-C06-04); y por la Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia, gracias a la convocatoria de Ayudas para la consolidación y estructuración de unidades de investigación competitivas del Sistema Universitario de Galicia, dentro de la *Rede de Lexicografía (Relex)* (ref. CN 2012/290) y de la *Rede de Tecnoloxías e análise dos datos lingüísticos* (ref. CN 2012/179).

Con todo, los corpus paralelos no sólo permiten acercarse a los estudios de traducción y de lingüística comparada con una perspectiva realista inimaginable antes de su compilación, sino que también posibilitan el desarrollo de aplicaciones lingüísticas y de tecnologías de la lengua que basan su funcionamiento en los datos suministrados por los corpus, al tiempo que ofrecen diversas posibilidades de explotación en el ámbito del procesamiento del lenguaje natural. Los resultados más importantes de la explotación de estos corpus se obtienen en las aplicaciones computacionales relacionadas con la traducción y la lexicología, en los campos de la traducción automática estadística (Koehn 2010), de las memorias de traducción y de la traducción automática basada en ejemplos (Carl y Way 2002), de la extracción léxica para recuperación de información multilingüe (Brown *et al.*, 2000), de la extracción de terminología bilingüe (Gómez 2012) y de la extracción de léxico bilingüe (Simões 2008).

En este último campo de trabajo, en el que se enmarca nuestra investigación, el objetivo es la generación de diccionarios bilingües basados en las equivalencias léxicas de traducción identificadas en un corpus paralelo, en el que previamente se han establecido a nivel de frase u oración las correspondencias de traducción. En cierto sentido, el problema central de la extracción de léxico bilingüe consiste en convertir un corpus paralelo etiquetado con los alineamientos (es decir, con las equivalencias de traducción) a nivel de oración, en un corpus etiquetado paralelo con los alineamientos a nivel de palabra. Para lograr esta tarea, se han desarrollado diversos algoritmos, gran parte de ellos basados en medidas estadísticas relacionadas con la asociación mutua o con la coaparición de los elementos léxicos en las frases u oraciones bilingües alineadas (Och / Ney 2003), y mostrando todos un margen de error nada desdeñable en los resultados (Tiedemann 2003) debido a la naturaleza intrínsecamente "no literal" de la traducción y a otras dificultades relacionadas con las características del corpus, como la distancia lingüística entre las lenguas implicadas, el tipo de textos o el estilo de la traducción.

En este artículo presentaremos de manera global el trabajo llevado a cabo para realizar el *Dicionario Moderno Inglés-Galego* (*DMIG*), desde la compilación del corpus paralelo fuente de la información lexicográfica, al proceso informatizado de extracción de la información léxica, y a la labor lexicográfica práctica de elaboración del diccionario bilingüe. El *DMIG* es un diccionario basado en la colección de textos ingleses traducidos al gallego que forma parte del Corpus CLUVI y constituye, a nuestro entender, el primer diccionario basado en corpus de la lexicografía gallega. Todas las palabras inglesas que aparecen en sus entradas están documentadas en los textos en inglés traducidos al gallego recopilados en el corpus paralelo CLUVI. Así mismo, todas las traducciones gallegas recogidas en el diccionario para esas palabras son traducciones reales identificadas en las versiones gallegas de los textos ingleses del corpus. Finalmente, para cada traducción identificada, el diccionario suministra un ejemplo real de su uso tal como está documentado en el corpus.

2. El Corpus CLUVI

El CLUVI (Corpus Lingüístico da Universidade de Vigo), elaborado por nuestro grupo de investigación de la Universidad de Vigo, es un conjunto de corpus paralelos de unos 23 millones de palabras, formado principalmente con traducciones al gallego o del gallego, accesibles para consulta en la web desde septiembre de 2003 en la dirección <http://sli.uvigo.es/CLUVI>. Ciñéndonos únicamente a las secciones del CLUVI que incluyen traducciones con el gallego, el Corpus CLUVI está formado por seis corpus paralelos principales pertenecientes a cuatro registros especializados de los ámbitos jurídico-administrativo, literario, de la informática y de divulgación científica; y a cinco combinaciones lingüísticas diferentes: bilingüe gallego-español, bilingüe inglés-gallego, bilingüe francés-gallego, tetralingüe inglés-gallego-francés-español y tetralingüe español-gallego-catalán-eusquera. Estos seis corpus, con los datos actuales sobre su extensión, son el Corpus Lega de textos jurídico-administrativos gallego-español (6 582 415 palabras), el Corpus Logaliza de localización de software inglés-gallego (3 706 242 palabras), el Corpus Unesco de divulgación científica inglés-gallego-francés-español (3 724 620 palabras), el Corpus Tectra de textos literarios inglés-gallego (2 407 539 palabras), el Corpus Fega de textos literarios francés-gallego (1 863 95 palabras), y el Corpus Consumer español-gallego-catalán-eusquera de información sobre salud y consumo (5 586 431 palabras).

El corpus jurídico Lega gallego-español contiene material textual bilingüe de dos vertientes especializadas del lenguaje jurídico muy próximas, mas bien diferenciadas: por una parte, el ámbito administrativo, representado por 30 ejemplares íntegros del *Diario Oficial de Galicia* (DOG) publicados entre 2000 y 2005; y por otra parte, el ámbito legislativo, representado por un conjunto de 76 textos publicados entre 1978 y 2008 con legislación diversa de ámbito nacional (publicada en el DOG), estatal (publicada en el *Boletín Oficial del Estado*) y supraestatal (*Constitución Europea*). Más concretamente, las leyes y legislaciones publicadas en el BOE y en el DOG pertenecen a distintos campos dentro del ámbito legislativo: al campo judicial, al ámbito de la Constitución y de los Estatutos, al campo económico, al campo social, al derecho ambiental, al derecho informático y al ámbito relacionado con sectores específicos (universidades, pesca, circulación, etc.). Los textos administrativos suman un total de 2 394 407 palabras, mientras que los textos legislativos representan un total de 4 188 008 palabras.

Por otra parte, el Corpus Logaliza de localización de software inglés-gallego contiene la localización al gallego del paquete ofimático *OpenOffice*, del sistema operativo *Windows XP* de Microsoft, del navegador *Firefox* y de los escritorios *Gnome* y *KDE* para Linux; el Corpus Unesco inglés-gallego-francés-español de divulgación científica está constituido por 32 ejemplares íntegros de la revista mensual *The Unesco Courier* publicados entre 1998 y 2001; y el Corpus Consumer español-gallego-catalán-eusquera de información sobre consumo incluye 1 036 artículos de la revista *Consumer Eroski* publicados entre 1998 y 2005. Finalmente, el corpus literario Tectra inglés-gallego recopila 50 textos literarios completos en lengua inglesa con sus traducciones para el gallego; y el corpus literario Fega francés-gallego, 29 textos literarios completos en lengua francesa con sus

traducciones para el gallego. Los corpus Tectra inglés-portugués (875 595 palabras), e inglés-español (122 251 palabras), aun en desarrollo, incluyen los textos literarios bilingües del Tectra inglés-gallego en su traducción al portugués y al español.

El CLUVI incluye también otros seis corpus paralelos del gallego en distintas fases de elaboración, en particular, el Corpus Egal de economía gallego-español (718 642 palabras), el Corpus Dega alemán-gallego (366 038 palabras), el Corpus Veiga de subtitulación inglés-gallego (294 714 palabras) (Sotelo y Gómez 2012), el Corpus Turigal de turismo español-gallego (325 389 palabras), el Corpus literario Galea gallego-español (162 795 palabras) y el Corpus literario Pega portugués-gallego (68 431 palabras).

Otros corpus paralelos fruto de proyectos realizados en el seno de nuestro grupo de investigación en el ámbito del Corpus CLUVI, que no incluyen directamente el gallego como objeto de estudio, son el Corpus Palop de literatura poscolonial portugués-español (566 590 palabras) (Malvar 2007) y el Corpus Turigal de turismo portugués-inglés (1 285 764 palabras) (Moreira 2011). Así mismo, cabe destacar que a través de la interface de consulta del Corpus CLUVI se puede acceder a la consulta del Corpus Lege-Bi eusquera-español de textos jurídico-administrativos (2 384 053 palabras), desarrollado por el grupo DELi de la Universidad de Deusto.

En la siguiente sección presentaremos las características generales de la anotación del Corpus CLUVI, así como las soluciones adoptadas en su desarrollo para la codificación de las equivalencias de traducción cuando la correspondencia entre original y traducción no es directa por causa de la omisión, adición o reordenamiento de frases en la traducción.

2.1. Criterios de anotación

El alineamiento de los textos paralelos se almacena en el CLUVI en el formato TMX (Translation Memory eXchange), el estándar para la codificación en XML de memorias de traducción independientemente de la aplicación utilizada (Savourel / Lommel 2005). Las memorias de traducción se utilizan sobre todo en los programas de traducción asistida por ordenador y, más comúnmente, en los entornos de traducción como SDL Trados, DéjàVu, Wordfast, Transit o Passolo, este último más orientado a la localización de software. Estos asistentes informáticos para la traducción integran en un mismo producto un procesador de textos especialmente diseñado para traducir, un conjunto de diccionarios bilingües, herramientas para la gestión terminológica (para la creación y mantenimiento de glosarios, la consulta automática de glosarios durante la traducción y la extracción automática de terminología) y una utilidad de gestión de memorias 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 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 rescribirla completamente, haciendo las modificaciones que se consideren más idóneas. En 1997 la industria creó e impulsó el estándar TMX para permitir el intercambio de memorias de traducción entre los

distintos programas de traducción asistida. Con ciertas matizaciones, un corpus paralelo alineado equivale a una memoria de traducción y, en la práctica, existe un número considerable de corpus paralelos alineados codificados en TMX, con la ventaja adicional de que los corpus así etiquetados pueden ser empleados como memorias de traducción para alimentar los programas de traducción asistida (tal y como se propone en Simões *et al.* 2004).

La unidad básica de segmentación para el alineamiento de los bitextos del Corpus CLUVI es la frase ortográfica del texto original. Por tanto, la correspondencia entre el texto original y la traducción va a ser siempre del tipo *1:n*. Con frecuencia, a una frase del original le corresponde una frase de la traducción (*1:1*). Sin embargo, se dan también casos en los que una frase del original no se traduce (*1:0*), o en los que a una frase del original le corresponde en la traducción media frase (*1:1/2*) o dos frases (*1:2*), o incluso en los que una frase de la traducción no se corresponde con ninguna frase del original (*0:1*). A parte de eso, la traducción implica a veces desplazamientos de frases enteras, o movimientos de fragmentos de frases del original a otras frases en la traducción. Estos movimientos se reordenan en la sección de textos traducidos de los corpus paralelos del CLUVI para cumplir el requisito del alineamiento *1:n*, que preserva la integridad y el orden de las unidades de traducción del texto original. Este criterio es crucial cuando se aplica al procesamiento de corpus plurilingües de más de dos lenguas, debido a que las frases del original son las que, actuando a modo de intermediarias, nos permiten establecer las correspondencias entre las frases equivalentes de las distintas lenguas.

La especificación TMX no tiene en cuenta la codificación de estos aspectos de las traducciones, ya que fue diseñada para el almacenamiento e intercambio de memorias de traducción, y no para la representación de segmentos equivalentes en corpus paralelos. El sistema de codificación del CLUVI utiliza una versión adaptada de algunas de las etiquetas que forman parte de la especificación TMX para representar las correspondencias que no son *1:1* y los reordenamientos codificados en el corpus paralelo. Los aspectos traductológicos codificados en el Corpus CLUVI, en concreto, las omisiones, adiciones y reordenamientos, son etiquetados mediante una versión adaptada de los elementos <*hi*> e <*ph*> que forman parte del estándar TMX.

En la omisión, hay una porción del texto de partida que no tiene correspondencia en el texto de llegada, es decir, una frase o parte de una frase no es traducida. La omisión se codifica en los corpus paralelos del CLUVI con el elemento <*hi*>. De acuerdo con la especificación TMX, el elemento (de nombre derivado del 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 / Lommel, 2005). En la especificación del CLUVI, basada en la TMX, el elemento <*hi*> marca en el texto de partida el elemento que se omite en el texto de llegada. Indicamos este uso de la etiqueta <*hi*> mediante un atributo *type* caracterizado con el valor de *"supr"*, tal como ejemplificamos en (1), donde (1a) es el bitexto inglés-gallego y (1b) su codificación en el formato de representación del CLUVI:

(1)

a. [en] 'Hello', I said.

[gl] -Ola.
b. <tu>
<tuv xml:lang="en">
<seg>'Hello',<hi type="supr">I said.</hi></seg>
</tuv> <tuv xml:lang="gl">
<seg>-Ola.</seg>
</tuv>
</tu>

La adición en la traducción implica una inserción de fragmentos en el texto de llegada que no tienen correspondencias en el texto de partida. La adición también se codifica en el CLUVI con el elemento <*hi*>, haciendo que este indique el fragmento insertado en la traducción. Este uso de la etiqueta se distingue mediante un atributo *type* caracterizado con el valor de *"incl"*. El fragmento añadido se incorpora a la unidad de traducción en la que está insertado. Cuando el nuevo fragmento es una oración (o una secuencia de oraciones), se incorpora bien a la unidad de traducción anterior, bien a la siguiente, de acuerdo con su contexto, respetando así el criterio de alineamiento *1:1*, como se ilustra en el ejemplo (2):

(2)
a. [en] 'Hello.'
[gl] -Ola - dixen.
b. <tu>
<tuv xml:lang="en">
<seg>'Hello.'</seg>
</tuv>
<tuv xml:lang="gl">
<seg>-Ola <hi type="incl">- dixen.</hi>
</tuv>
</tu>

El reordenamiento implica desplazamientos de frases enteras o movimientos de fragmentos de frases del original a otras frases en la traducción. Estos movimientos se reordenan en la sección de textos traducidos de los corpus paralelos del CLUVI para cumplir el requisito del alineamiento *1:n*, que preserva la integridad y el orden de las unidades de traducción del texto original. El reordenamiento se codifica en el CLUVI mediante una combinación de los elementos <*hi*> y <*ph*>. Anotamos el fragmento o la oración movida mediante un elemento <*hi*> que incluye un atributo *type* con valor de *"reord"* y un atributo *x* con un valor numérico que actúa de índice. Por otra parte, indicamos con un elemento <*ph*> el lugar en el texto que ocupaba originalmente el elemento desplazado. Según la especificación TMX, el elemento <*ph*> (o *placeholder*) se utiliza «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 / Lommel 2005). En la especificación del CLUVI, basada en la TMX, el elemento adaptado <*ph*> indica el punto de partida del movimiento, mientras que la relación entre el elemento desplazado y el lugar de partida es codificada en el elemento <*ph*> mediante un atributo *x* que comparte valor con el índice codificado en el elemento <*hi*> del segmento movido.

Obviamente, la etiqueta que indica el lugar de origen siempre es una etiqueta vacía. Como criterio de etiquetado en la codificación del CLUVI, y con la finalidad de evitar incoherencias entre las distintas personas que participan en la codificación del corpus, los segmentos reordenados siempre son desplazados en dirección al inicio del texto. En consecuencia, en el CLUVI no hay ninguna secuencia semejante a *<ph x="n"/> [...] <hi type="reord" x="n">Reordered element</hi>*; en vez de eso, las secuencias son siempre del tipo *<hi type="reord" x="n">Reordered element</hi> [...] <ph x="n"/>*, como se puede observar en el ejemplo (3) de codificación de un reordenamiento sencillo:

(3)

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

b.
```
<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>
```

Estas son, por tanto, las ampliaciones del estándar TMX adoptadas en el Corpus CLUVI para la codificación de las equivalencias de traducción no biunívocas. Como veremos en la siguiente sección del artículo, la anotación de estas equivalencias no biunívocas tendrá una importancia capital en el procesamiento del corpus para la extracción de las equivalencias de traducción que finalmente se codificarán en el *DMIG*.

3. Extracción de la información lexicográfica bilingüe

De acuerdo con nuestros presupuestos metodológicos, la primera fase en el trabajo de elaboración del *DMIG* fue la preparación del corpus paralelo CLUVI que constituye su fuente textual. En este corpus, de forma general, a cada unidad del texto original le corresponde otra unidad del texto traducido. Hay, no obstante, excepciones: las llamadas asimetrías de la traducción, que constituyen la principal dificultad para el alineamiento y, en consecuencia, también para el tratamiento automático de la extracción léxica bilingüe a partir del corpus paralelo. Como

vimos, las asimetrías de traducción se pueden definir como correspondencias no biunívocas y alteraciones en el orden de las unidades de traducción. Si lo habitual es que las correspondencias entre original y traducción sean biunívocas (*1:1*), abundan los casos en los que esta correspondencia puede ser *1:0*, *1:1/2*, *1:2*, *0:1*, etc. Del mismo modo, podemos encontrar en las traducciones fragmentos desplazados, incluso oraciones enteras que cambian de lugar.

La existencia de estos fenómenos motivó el uso de una serie de etiquetas especiales para identificarlos y facilitar el trabajo de extracción automática de equivalencias léxicas bilingües candidatas a ser entradas del diccionario. A este efecto, una vez identificados y marcados los casos de omisión, adición y reordenamiento, se generó una nueva copia del corpus en el que estas asimetrías se eliminaron, al igual que algunos signos de puntuación, los dígitos y las palabras gramaticales con mayor índice de frecuencia. Todos estos elementos complican el proceso de extracción léxica bien por ser repetitivos, bien por suponer casos de dislocación entre el texto inglés y el gallego.

Dependiendo de la lengua del texto, se eliminaron unas unidades u otras, aunque en general para las dos lenguas se eliminaron los signos de puntuación, excluyendo los guiones de unión de palabras compuestas; los dígitos; y los segmentos etiquetados como omisiones o inserciones, ya que indican unidades sin correspondencia de traducción. Concretamente, en la versión modificada del corpus paralelo para la extracción léxica bilingüe, eliminamos de la lengua inglesa determinantes (*the, a, an*), pronombres personales (*I, you, he, she, it, we, you, they, me, you, him, her, you*), posesivos (*my, his, her*), demostrativos (*this, that*), conjunciones (*and, but, if, or*), preposiciones (*to, of, in, at, on, with, out, around, about*), partículas negativas (*no, not*), pronombres indefinidos (*all*), verbos auxiliares (*do, does, did, is, are, was, were, has, had*) y la marca de genitivo sajón. Por otro lado, para el gallego eliminamos artículos (*o, a, os, as*), indefinidos (*un, uns, unha, unhas*), pronombres personales tónicos (*eu, ti, el, ela, nós, vós, eles, eles, me, se, nos*), posesivos (*meu, meus, seu, seus*), preposiciones (*a, con, de, en, para, por*), contracciones de preposición con artículo (*ó, ao, á, ós, aos, ás, co, coa, cos, coas, do, da, dos, das, no, na, nos, nas, polo, pola, polos, polas*), conjunciones (*que, e, se, nin, ou, pero*), verbos de tipo auxiliar (*é, era*) y partículas negativas (*non*).

Para conseguir que un corpus paralelo etiquetado quede alineado en el nivel de la palabra (es decir, para conseguir que se establezcan en el corpus las equivalencias léxicas de traducción) existen diversos algoritmos estadísticos que trabajan con índices de atracción o asociación léxica y con los datos de la coaparición de elementos léxicos en las unidades de traducción bilingües alineadas. En nuestro caso, empleamos la herramienta informática para el alineamiento léxico NATools (Simões / Almeida 2003). Los diccionarios de traducción probabilísticos (DTP) de las NATools, generados automáticamente a partir de corpus paralelos alineados a nivel oracional, indican para cada forma léxica en la lengua origen documentada en el corpus el conjunto de posibles traducciones en la lengua destino identificadas por la herramienta en el corpus. Cada entrada léxica del DTP indica el número de apariciones en el corpus del elemento léxico en la lengua fuente, y sus posibles traducciones con sus probabilidades de equivalencia calculadas a partir de los

alineamientos oracionales, como se puede comprobar en los ejemplos de entradas léxicas inglés-gallego del DTP para el Corpus Logaliza recogido en (4):

```
(4) "library" => {
      count => 434,
      trans => {
        "biblioteca" => 0.92184907,
        "bibliotecas" => 0.00992135,
        "librería" => 0.00357169,
      },
    },
    "files" => {
      count => 3622,
      trans => {
        "ficheiros" => 0.92156011,
        "ficheiro" => 0.01192421,
        "arquivos" => 0.00071648,
      },
    },
    "windows" => {
      count => 1686,
      trans => {
        "windows" => 0.52585220,
        "xanelas" => 0.22492823,
        "ventás" => 0.10177911,
        "fiestras" => 0.06585271,
      },
    }
```

El resultado obtenido de la extracción automática es un diccionario bilingüe probabilístico al que le hace falta una revisión manual para mejorar su precisión; se trata fundamentalmente de eliminar equivalencias fallidas y de añadir otras que, a pesar de estar documentadas en el Corpus CLUVI, no se recogen en el diccionario probabilístico por distintas razones de índole cuantitativo. Para identificar estas ausencias en la lista inicial de equivalencias, comprobamos la presencia en el diccionario de todas las formas léxicas presentes tres o más veces en los textos en inglés del corpus, y consultamos los lemarios y las equivalencias ofrecidas por otras listas léxicos bilingües de orientación elemental y escolar, como el *Diccionario inglés-gallego* de Xerais (Salgado 1999). Se encontrarán más detalles técnicos sobre el proceso de extracción léxica bilingüe en Gómez / Sacau (2004).

Tras esta revisión, el lemario del *DMIG* recoge todos los lemas que aparecen un mínimo de tres veces en el conjunto de los corpus de los que se partió (principalmente, los corpus Tectra y Unesco, mas también los corpus Logaliza y Veiga), junto a un conjunto de palabras que consideramos de interés a pesar de poseer una presencia menor en los textos. En total, el diccionario contiene 20 000 entradas documentadas en el corpus, acompañadas de 30 000 traducciones y 60 000

ejemplos, un caudal léxico importante que constituye la base de la presente edición impresa del *DMIG*.

4. Diseño del *DMIG*

El proceso de edición del diccionario continuó por su microestructura, es decir, por la estructura interna de las entradas o artículos lexicográficos (Gómez *et al*. 2008). De este modo, se añadieron las categorías gramaticales; notas gramaticales y de uso; un ejemplo bilingüe para cada una de las traducciones suministradas con su respectiva referencia al Corpus CLUVI; y, de ser relevante, información sobre los usos fraseológicos del lema, un apartado presente en 1 de cada 4 entradas. La información fraseológica, que acrecienta el valor de este diccionario, recoge las combinaciones lexicalizadas del lema con otra u otras palabras, acompañadas de sus traducciones documentadas en el corpus.

La fuente del *DMIG* está íntegramente editada en el formato XML, respetando unas sencillas normas de codificación definidas en su DTD (definición del tipo de documento). La DTD es la declaración formalizada de la estructura lógica de un documento, en este caso un diccionario. Está compuesta por una serie de *declaraciones de elementos* codificadas en SGML. A continuación presentamos la DTD de nuestro diccionario:

```
<!ELEMENT dicionario (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)>
```

Esta DTD se lee de la siguiente manera: un documento de tipo diccionario está formado por un conjunto de entradas; cada entrada incluye además del lema un conjunto de informaciones traductológicas agrupadas en función de las categorías gramaticales del lema. Cada uno de estos conjuntos (denominados *super_cat* en la DTD) puede contener una o más acepciones, dependiendo de la polisemia de cada lema en cada categoría gramatical. Cada acepción incluye una traducción al gallego, un ejemplo de uso y, opcionalmente, la expresión plurilexemática (fraseológica) de la que forme parte el lema. Por último, cada ejemplo consta de un fragmento textual del Corpus CLUVI en inglés, su traducción al gallego y la referencia de la obra en la que se documenta el ejemplo. De este modo, cada entrada del diccionario puede incluir una o más categorías gramaticales con una o más traducciones, siendo

codificada internamente como se ilustra en el siguiente ejemplo, en el que por claridad hemos resaltado en negrita las etiquetas XML:

<entrada>
<lema>own**</lema>**
<super_cat><categoria>transitive verb**</categoria>**
<acepcion><traducion>ser dono**</traducion>**
<exemplo><en>One relic hast thou in thy treasury, handed down from the Moslems who once @owned# Toledo --a box of sandal-wood containing a silken carpet: **</en>**
<gl>Tes unha reliquia no teu tesouro, herdada dos musulmáns que nalgún tempo @foron donos# de Toledo... unha arqueta de pao de sándalo que contén unha alcatifa de seda**.</gl><fonte>**ALH (778)**</fonte></exemplo></acepcion>**
<acepcion><traducion>posuír**</traducion>**
<exemplo><en>Kino and Juana came slowly down to the beach and to Kino's canoe, which was the one thing of value he @owned# in the world. **</en>**
<gl>Kino e Juana baixaran de vagar ata a praia e achegáranse á canoa de Kino, que era o único de valor que el @posuía# no mundo**.</gl><fonte>**PER (239)**</fonte></exemplo></acepcion>**
<acepcion><plurilex>own up**</plurilex>**
<traducion>confesar**</traducion>**
<exemplo><en>But she did not mind @owning up# to it in the least; one must admit that**.</en>**
<gl>Pero alomenos non lle importaba @confesalo#; iso había que recoñecelo.
</gl><fonte>CAR (1083)**</fonte></exemplo></acepcion></super_cat>**
<super_cat><categoria>adjective**</categoria>**
<acepcion><traducion>propio**</traducion>**
<exemplo><en>It was about this little kid that wouldn't let anybody look at his goldfish because he'd bought it with his @own# money. **</en>**
<gl>Falaba dun neno pequeno que non deixaba ve-lo seu peixe a ninguén porque o mercara co seu @propio# diñeiro**.</gl><fonte>**VIX (20)**</fonte></exemplo>**
</acepcion></super_cat>
<super_cat><categoria>pronoun**</categoria>**
<acepcion><plurilex>of one's own**</plurilex><traducion>**de seu**</traducion>**
<exemplo><en>He thought it very possible that Master Randolph's sister was a coquette; he was sure she had a spirit @of her own#; but in her bright, sweet, superficial little visage there was no mockery, no irony. **</en>**
<gl>Considerou que sería moi posible que a irmá de Master Randolph fose unha coqueta. Estaba seguro de que tiña un espírito @de seu#. Mais no seu pequeno, elegante, doce e superficial rostro non había farsa nin ironía ningunha**.</gl><fonte>**DAI (124)**</fonte></exemplo></acepcion></super_cat>**
</entrada>

Las entradas, codificadas en XML, son convertidas al formato de salida, para su visualización en la web o para su versión impresa, mediante una hoja de estilo XSL.

4.1. Microestructura

La microestructura del *DMIG* es semejante a la de otros diccionarios bilingües, pero se le da más importancia (y más espacio) a las traducciones de los lemas y a los ejemplos de contextualización. Los elementos que componen cada una de las entradas son los habituales en la mayoría de los diccionarios bilingües: lema en inglés, categoría del lema, indicaciones gramaticales y de uso, traducciones al gallego, ejemplos de cada traducción con indicación de su fuente textual y un apartado de fraseología en aquellos casos en los que el lema participa en refranes, frases hechas, etc.

Si analizamos detalladamente la estructura interna de los artículos, podemos reparar en que sus elementos constituyentes muestran cierta variabilidad, reflejo de la heterogeneidad y de las irregularidades de la lengua inglesa. Así por ejemplo, se destacan en el diccionario los casos de lemas con flexión irregular, como por ejemplo en *louse*:

louse ▪ *noun* ♀ Pl: *lice*
piollo ▸ *A louse crawled over the nape of his neck and, putting his thumb and forefinger deftly beneath his loose collar, he caught it.* Un piollo andaba a arrastrarse pola súa caluga, e metendo destramente o polgar e o furabolos no colar frouxo apañouno. [RET]

Figura 1: Entrada para *louse*

o las diferencias entre las formas típicamente británicas y estadounidenses, como se puede ver en las cabeceras de las siguientes entradas:

plead ▪ *intransitive verb* ♀ Este verbo é irregular en inglés americano, adquirindo a forma *pled* para o pasado e o participio. En inglés británico, simplemente engade *-ed*.

Figura 2: Entrada para *plead*

jumper (🇺🇸 sweater) ▪ *noun*
xersei ▸ *A kind-faced woman in a hand-knitted jumper said wistfully, "Don't you agree, Mr. Dexter, that no one, no one has written about feelings so poetically as Virginia Woolf?* Unha muller de rostro infantil que levaba un xersei de calceta dixo ansiosa: – Señor Dexter, non está de acordo comigo en que ninguén, repito, ninguén describiu os sentimentos dun modo tan poético coma Virginia Woolf? [TER]

Figura 3: Entrada para *jumper*

film (🇺🇸 movie) ▪ *noun* 💡 Para referirse aos filmes cinematográficos e televisivos, o inglés americano utiliza preferentemente o termo *movie*.

Figura 4: Entrada para *film*

lorry (🇺🇸 truck) ▪ *noun*

camión ▸ *Loaded into trucks and vans, they are exported from one country to the next, even making it to the display windows of genuine bookshops –perhaps innocently, perhaps not.* Cargados en camións ou furgonetas, expórtanse dun país a outro e, o que é peor, chegan a invadir os escaparates das librarías legais. Inxenuidade ou complicidade dos libreiros? [C07]

Figura 5: Entrada para *lorry*

Pueden acompañar a los lemas, además, diversas notas sobre sus usos en inglés, que sirven para delimitar más correctamente el sentido que se le da a la palabra en cada caso:

must ▪ *modal verb* 💡 A diferenza entre *must* e *have to* radica en que o modal designa unha obriga imposta polo suxeito, unha obriga interna, mentres que *have to* expresa obrigas externas. Ex.: I *must* finish my work today > *Debo* rematar o traballo hoxe (é un deber que me propoño); You *have to drive* on the right > *hai que conducir* pola dereita (obríganme as leis nacionais).

Figura 6: Entrada para *must*

El *DMIG* ofrece también información ortográfica, morfológica y de uso sobre las traducciones gallegas de los lemas. A pesar de que muchos de los textos del corpus CLUVI de los que se extraen los ejemplos son anteriores a la reforma más reciente de las normas ortográficas y morfológicas del gallego (Real Academia Galega 2004), el *DMIG* ha unificado y regularizado la ortografía de todos los ejemplos para hacerlos concordar con las nuevas directrices de la lengua:

hassle ▪ *noun*

conflito ▸ *By starting a hotline, some 50 Slovenian teenagers have become pros at listening, conversing and gently settling everyday hassles.* É o nome dunha liña aberta iniciada e animada por uns cincuenta adolescentes eslovenos. Cal é a súa función? Escoitar, dialogar e resolver con calma os conflitos de todos os días. [C06]

Figura 7: Entrada para *hassle*

thank ▪ *transitive verb*

dar as grazas ► *He thanked heaven she had left the neighborhood, and was equally thankful that he did not know where she had gone.* Dáballe grazas ao ceo de que tivese abandonado a veciñanza e agradecía tamén non coñecer o seu paradoiro. [ESP]

Figura 8: Entrada para *thank*

De gran ayuda creemos que pueden ser las advertencias sobre errores de lengua habituales, como algunos castellanismos muy extendidos:

syndrome ▪ *noun*

síndrome En galego este termo ten xénero femenino ► *The Stendhal syndrome is not the only experience shared by modern cultural tourists and wayfarers of the past.* A síndrome de Stendhal non é a única experiencia que comparten os turistas culturais modernos e os viaxeiros do pasado. [C12]

Figura 9: Entrada para *syndrome*

En cuanto al caudal léxico recogido en este diccionario, se ha procurado incluir, además del que podríamos llamar léxico estándar o común de la lengua inglesa (y gallega), otros grupos de palabras como:

a) palabras de creación reciente o con nuevos significados, procedentes sobre todo de los campos de las tecnologías de la información y algunas disciplinas científicas:

hacking ▪ *noun*

piratería informática ► *The treaty also expands the definition of hacking to include violations of contracts and the terms of service posting on websites.* A convención amplía tamén a definición do concepto de pirataría informática para que abranga a violación de contratos e condicións de servizo nos sitios web. [C07]

Figura 10: Entrada para *hacking*

b) nombres propios de persona, país, ciudad, ríos, mares, etc.; es decir, artículos de contenido más enciclopédico que lexicográfico:

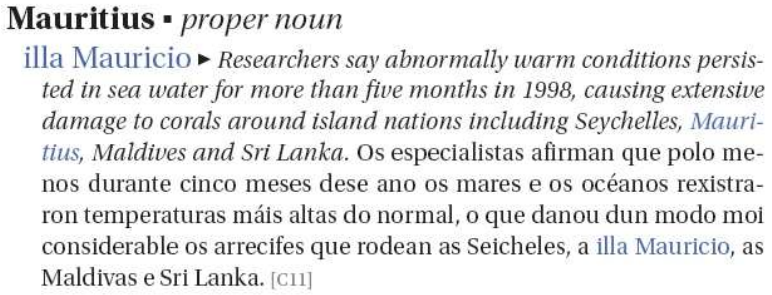

Figura 11: Entrada para *Mauritius*

c) términos especializados que llegan a la lengua común a través de textos de divulgación científica:

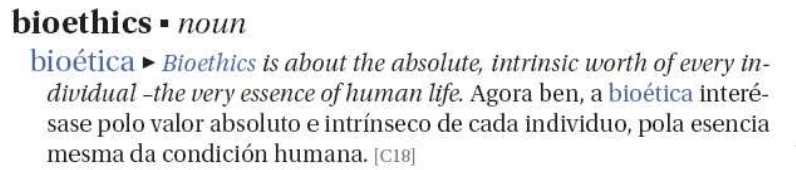

Figura 12: Entrada para *bioethics*

Por lo que respecta al contenido de las entradas, cabe destacar, por último, la sección de fraseología del diccionario, que en algunos casos, como el de los *phrasal verbs*, llega a ser muy amplio. Consúltese por ejemplo la entrada del verbo *come,* que contiene 14 unidades de este tipo; o las entradas para los verbos *go* o *break*, con 12.

5. Conclusiones

El *Dicionario moderno inglés-galego* elaborado por el Seminario de Lingüística Informática (SLI) de la Universidade de Vigo es una obra con características propias dentro de la tradición lexicográfica gallega. Se trata del primer gran diccionario inglés-gallego elaborado con una metodología moderna de orientación textual empírica. El *DMIG* está basado en un corpus representativo de textos ingleses traducidos al gallego que forman parte del Corpus CLUVI. Todas las palabras inglesas que aparecen como lemas de las entradas del *DMIG* están documentadas en los textos ingleses traducidos al gallego recopilados en el CLUVI. Además de esto, todas las traducciones gallegas que se recogen en el diccionario para esas palabras son traducciones reales identificadas en las versiones gallegas de los textos ingleses. Para cada traducción seleccionada recogemos un ejemplo real de uso documentado en el CLUVI.

Una versión actualizada del *DMIG*, en formato electrónico, puede consultarse en Internet en la dirección <http://sli.uvigo.es/dicionario>. Así mismo, puede

consultarse el Corpus CLUVI, fuente textual de este diccionario, en la dirección <http://sli.uvigo.es/CLUVI>. El trabajo de nuestro grupo de investigación en el *DMIG* se centrará, a partir de este momento, en la revisión y ampliación de las entradas y de la información que contienen, al tiempo que se mejorará y ampliará la sección de corpus paralelos inglés-gallego del Corpus CLUVI que supone la base empírica de conocimiento léxico en la que se sustenta el diccionario. Desde el Seminario de Lingüística Informática de la Universidade de Vigo deseamos que el *DMIG* sea una herramienta útil tanto para estudiantes y personas usuarias del inglés de todos los niveles, como para profesionales de la traducción y de la interpretación entre inglés y gallego, que hasta ahora tenían que recurrir a diccionarios puente inglés-portugués o inglés-castellano para poder realizar consultas lexicográficas bilingües en su trabajo.

Referencias bibliográficas

Brown, Ralf D. / Carbonell, Jaime G. / Yang, Yiming (2000): «Automatic dictionary extraction for Cross-Language Information Retrieval», en J. Véronis, ed., *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer, 275-298.

Carl, Michael / Way, Andy (eds.) (2002): *Recent Advances in Example-Based Machine Translation*. Dordrecht: Springer.

[DMIG]: Gómez Guinovart, Xavier (coord.) / Álvarez Lugrís, Alberto / Díaz, Eva (2012): *Dicionario moderno inglés-galego*. Ames (Santiago de Compostela): 2.0 Editora.

Gómez Guinovart, Xavier (2012): «A Hybrid Corpus-Based Approach to Bilingual Terminology Extraction», en Isabel Moskowich-Spiegel Fandiño / Begoña Crespo, eds., *Encoding the Past, Decoding The Future: Corpora in the 21st Century*. Newcastle upon Tyne: Cambridge Scholar Publishing, 147-175.

Gómez Guinovart, Xavier / Díaz Rodríguez, Eva / Álvarez Lugrís, Alberto (2008): «Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego», *Viceversa: Revista Galega de Traducción*, 14, 71-87.

Gómez Guinovart, Xavier / Sacau Fontenla, Elena (2004): «Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos», *Procesamiento del Lenguaje Natural*, 33, 133-140.

Koehn, Philipp (2010): *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Malvar, Paulo (2007): «Aproximação à linguística de corpus como metodologia de base empírica. Compilação e anotação do Corpus Paralelo PALOP (português-espanhol) de Narrativa Pós-colonial», *Agália*, 89-90, 9-80.

Moreira, Adonay (2011): «The Translator as Cultural Mediator: a corpus-based study of omissions and additions in translations of tourism brochures», *The Journal of Cultural Mediation*, 1, 86-95.

Och, Franz Josef / Ney, Heermann (2003): «A Systematic Comparison of Various Statistical Alignment Models», *Computational Linguistics*, 29(1), 19-51.

Real Academia Galega (2004): *Normas ortográficas e morfolóxicas do idioma galego*. Vigo: Galaxia.

Salgado, Benigno (1999): *Diccionario elemental inglés-galego galego-inglés*. Vigo: Xerais.

Savourel, Yves / Lommel, Arle (2005): *TMX 1.4b Specification*. Technical Report. Localisation Industry Standards Association, disponible en <http://www.gala-global.org/oscarStandards/tmx/tmx14b.html>, [Consulta: 28/05/13].

Simões, Alberto (2008): *Extracção de recursos de tradução com base em dicionários probabilísticos de tradução*. Tese de doutoramento. Braga: Universidade do Minho.

Simões, Alberto / Almeida, José João (2003): «NATools: a statistical word aligner workbench», *Procesamiento del Lenguaje Natural*, 31, 217-224.

Simões, Alberto / Gómez Guinovart, Xavier / Almeida, José João (2004): «Distributed translation memories implementation using WebServices», *Procesamiento del Lenguaje Natural*, 33, 89-94.

Sotelo Dios, Patricia / Gómez Guinovart, Xavier (2012): «A Multimedia Parallel Corpus of English-Galician Film Subtitling», en Alberto Simões / Ricardo Queirós / Daniela da Cruz, eds., *1st Symposium on Languages, Applications and Technologies*. OASIcs: Open Access Series in Informatics, vol. 21. Saarbrücken: Dagstuhl Publishing, 255-266.