El *corpus tècnic* del IULA: corpus textual especializado plurilingüe

Teresa Cabré y Carme Bach*

Institut Universitari de Lingüística Aplicada: *BwanaNet: Programa d'explotació del corpus tècnic de l'IULA*. <brangaene.upf.es/bwananet/index.htm>. Corpus textual especializado en cinco idiomas (catalán, español, inglés, francés y alemán), con instrucciones de ayuda e interfaz de consulta en tres idiomas (catalán, español e inglés)

1. Presentación

El Instituto Universitario de Lingüística Aplicada (IULA) es un centro de la Universidad Pompeu Fabra, de Barcelona, dedicado a la investigación y a la formación de postgrado. Fue creado en 1993 y organizado desde su creación por M.ª Teresa Cabré.[1] El IULA se organiza en grupos de investigación: Léxico, Terminología y discurso especializado (Grupo IULATERM, que acoge la Lingüística Computacional), Lexicografía (Grupo INFOLEX), Variación lingüística (Grupo UVAL), Documentación y edición digital (Grupo DIGIDOC) y tres laboratorios: OBNEO (Observatorio de Neología), LATEL (Laboratorio de Tecnologías Lingüísticas) y el Laboratorio de Lingüística Forense.

Desde 1993 hasta la actualidad, el proyecto Corpus ha sido el proyecto de investigación común en el que han participado todos los miembros del IULA. Recopila textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de especialidad de la economía, el derecho, el medio ambiente, la medicina y la informática. El corpus comprende además documentos paralelos, con el objetivo de facilitar estudios de traducción. A su vez, el corpus multilingüe del IULA cuenta con un subcorpus de lengua general, extraído de la prensa de gran difusión y constituido como corpus contrastivo.

El objetivo de este corpus es facilitar el análisis de los datos lingüísticos a fin de poder establecer las leyes que rigen el comportamiento de cada lengua en cada área. Sus destinatarios son los investigadores y todos los usuarios que requieran consultas sobre los ámbitos de especialidad tratados. De la explotación del corpus se han derivado estudios de carácter terminológico, discursivo, morfológico, sintáctico, neológico o traductológico. Para facilitar la explotación de los datos, el IULA ha desarrollado una serie de herramientas de exploración. Una muestra de estas herramientas son un extractor automático de neología, un detector automático de terminología, un alineador de textos, un alimentador de diccionarios, etc. De hecho, este corpus es el soporte principal de las actividades de investigación y docencia de nuestro instituto.

La herramienta que permite acceder a los datos del corpus a través de Internet es BwanaNet, que puede encontrarse en la página principal de la web del IULA (<www.iula.upf.edu>), en el apartado denominado «Portal de recursos del IULA».

2. Los textos

El corpus del IULA, como se ha dicho, contiene textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de especialidad de economía, derecho, medio ambiente, medicina e informática, además de documentos paralelos sobre estas materias. Cada una de las áreas fue estructurada en diferentes subáreas por un especialista, a fin de que los textos pudieran recuperarse con mayor precisión temática. Véase a continuación cómo está estructurada el área de la medicina:

Anatomía (AN)
Organismos (OR)
Enfermedades (MA)
Productos químicos y fármacos (PQ)
Técnicas y equipamientos analíticos, diagnósticos y terapéuticos (TE)
Psiquiatría y psicología (PS)
Ciencias biológicas (CB)
Ciencias físicas (CF)
Antropología, educación, sociología y fenómenos sociales (FS)
Tecnología, industria, agricultura (TI)
Humanidades (HU)
Información científica (IC)
Grupos nominales (GN)
Planificación y gestión sanitaria (GS)
Asesor: Toni Valero

3. Tratamiento de los textos

El procesamiento de los textos del corpus sigue los siguientes pasos:

*Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra de Barcelona (España).
Dirección para correspondencia: teresa.cabre@upf.edu.

a) Fase de selección de los textos

Los especialistas en cada materia seleccionan aquellos textos que consideran pertinentes y los clasifican temáticamente dentro de una estructuración del dominio previamente consensuada por especialistas de la materia.

b) Fase de anotación y registro de la información del documento

Los documentos se marcan de acuerdo con el estándar SGML y siguiendo las directrices marcadas por el *Corpus Encoding Standard* (CES) de la iniciativa EAGLES. Posteriormente se registra la información documental de los textos (autor, título, edición, páginas seleccionadas, subdominio al cual pertenece, idiomas en que ese mismo documento se encuentra en el corpus...).

c) Fase de procesamiento lingüístico

El procesamiento lingüístico de los documentos está automatizado y consta de un preproceso, a través del cual se tratan lingüísticamente aquellas entidades que admiten una detección automática previa al análisis morfológico (fechas, números, locuciones, nombres propios, abreviaturas...), un análisis morfológico, mediante el cual se lematizan todas las palabras de los documentos y se les da una o más etiquetas morfológicas, de acuerdo con los etiquetarios morfosintácticos diseñados en el IULA, y una posterior desambiguación lingüística y estadística, de forma que a cada palabra le acabe correspondiendo un solo lema y una sola etiqueta.

d) Almacenamiento en una base de datos textual

Finalmente, cuando ya cada palabra tiene el lema y la categoría gramatical que le corresponde, los textos se almacenan en una base de datos textual, que contiene toda la información que se ha generado sobre el documento.

El resultado de todo el proceso de tratamiento de los textos puede consultarse actualmente en línea en <brangaene.upf.es/bwananet/index.htm>.

4. Estado actual

El corpus del IULA contiene actualmente más de 22 millones de palabras, con la siguiente distribución por ámbito temático y lengua.

Área	Catalán	Español	Inglés	Francés	Alemán	Total
Derecho	1 463 000	2 085 000	431 000	44 000	16 000	4 039 000
Economía	1 776 000	1 091 000	274 000	78 000	27 000	3 246 000
Medio ambiente	1 506 000	1 062 000	599 000	230 000	429 000	3 826 000
Informática	655 000	1 227 000	338 000	194 000	83 000	2 497 000
Medicina	2 619 000	4 077 000	1 555 000	27 000	198 000	8 476 000
Total . . .	8 019 000	9 542 000	3 197 000	573 000	753 000	22 084 000

Cuadro 1. Número de palabras por lengua y ámbito.

El corpus de medicina incluye un subcorpus de genoma humano, elaborado por el grupo Iulaterm, que contiene 945 000 palabras en catalán, 1 447 000 en español y 1 119 000 en inglés.

Los datos en relación con el corpus paralelo de las parejas lingüísticas más significativas catalán-español, catalán-inglés, español-inglés, se presentan en el cuadro 2.

Área	Catalán español	Catalán inglés	Español inglés
Derecho	460 000	12 000	57 000
Economía	600 000	250 000	283 000
Medio ambiente	214 000	213 000	144 000
Informática	28 000	-	300 000
Medicina	118 000	40 000	640 000
Total . . .	1 420 000	515 000	1 424 000

Cuadro 2. Número de palabras en corpus paralelos por ámbito y parejas de lenguas.

Finalmente, los datos del corpus de contraste se muestran en el cuadro 3.

Área	Catalán	Español	Total
General	1 526 000	3 230 000	4 756 000

Cuadro 3. Número de palabras en el corpus de lengua general.

5. Disponibilidad del corpus

La consulta del corpus del IULA se realiza vía Internet a través de BwanaNet, una interfaz desarrollada en el IULA. El Corpus Técnico del IULA (CT-IULA) está indexado con un paquete de herramientas desarrolladas por el Institut für Maschinelle Sprachverarbeitung, de la Universidad de Stuttgart (Corpus Workbench). El IULA ha desarrollado la herramienta que permite la interrogación del CT-IULA en línea (<brangaene.upf.es/bwananet/index.htm>).

Las consultas pueden hacerse bien sobre la totalidad del corpus, bien sobre un subcorpus determinado a elegir (conjunto de documentos, documentos de un mismo subdominio...). De momento, hay que hacer consultas por cada lengua, pero en la actualidad se está desarrollando una herramienta que permita hacer búsquedas multilingües y que estará disponible a finales de este año.

6. Posibilidades actuales de búsqueda

Una de las utilidades de los corpus etiquetados es que se pueden efectuar búsquedas por atributos. En el caso de nuestro corpus, disponemos de los atributos «palabra», «lema» y «categoría morfológica», de modo que podemos hacer búsquedas a través de cada uno de ellos y con todas sus posibles combinaciones.

BwanaNet ofrece cinco posibilidades de interrogación:

1) Búsqueda por unidades fuera de contexto

Permite generar una lista de formas, lemas o categorías morfológicas del subcorpus escogido. Debe especificarse la frecuencia mínima de aparición de elementos que nos interese listar. Esta opción de búsqueda no se activa si se selecciona todo el corpus.

2) Búsqueda por frecuencias

La opción de frecuencias solo está disponible cuando la selección se hace sobre todo el corpus. Permite generar dos tipos de información: a) lista de frecuencias de formas, lemas o etiquetas sobre todo el corpus seleccionado, y b) lista de frecuencias sobre secuencias de formas, lemas o categorías morfológicas de todo el corpus escogido.

Ejemplo: queremos saber cuáles son las preposiciones más frecuentes que aparecen después del verbo 'hacer', con una frecuencia mínima de dos apariciones. El resultado sería:

858	23,64%	**hacer**	**de**
797	21,96%	**hacer**	**en**
512	14,11	**hacer**	**con**
440	12,12%	**hacer**	**a**
390	10,75%	**hacer**	**por**
299	8,24%	**hacer**	**para**
63	1,74%	**hacer**	**mediante**
47	1.30%	**hacer**	**sobre**
44	1,21%	**hacer**	**sin**
43	1,18%	**hacer**	**desde**
34	0,94%	**hacer**	**entre**
26	0,72%	**hacer**	**hasta**
12	0,33%	**hacer**	**según**
10	0,28%	**hacer**	**ante**
10	0,28%	**hacer**	**hacia**
8	0,22%	**hacer**	**bajo**
8	0,22%	**hacer**	**cerca de**
6	0,17%	**hacer**	**tras**
6	0,17%	**hacer**	**acerca de**
4	0,11%	**hacer**	**contra**
4	0,11%	**hacer**	**por medio de**
2	0,06%	**hacer**	**incluso**
2	0,06%	**hacer**	**frente a**
2	0,06%	**hacer**	**a cambio de**

3) Concordancia simple

Permite interrogar sobre un lema o forma concreta, así como escoger el contexto de aparición completo o parcial.

4) Concordancia estándar

Permite la búsqueda de hasta doce unidades diferentes. Las interrogaciones pueden hacerse sobre la forma, el lema y/o la categoría morfológica de forma combinada.

Puede escogerse el tipo de contexto que se desee para los resultados, los elementos textuales sobre los cuales se quiere hacer la búsqueda y el nivel de información que se quiere en el resultado (formas, lemas o categorías morfológicas).

Ejemplo: buscamos apariciones del lema *enfermedad* seguido de un adjetivo calificativo, en documentos de medicina. El resultado de esta búsqueda sería:

y rasgos patológicos. La	***enfermedad congénita***	indica que la alteración está
y especificidad, pudiendo ocurrir	***enfermedades degenerativas***	como la demencia, con
fenómeno se acentúa en algunas	***enfermedades pulmonares***	Los pulmones resuenan a
arteria femoral es propensa a	***enfermedades arteriales***	y el vaso es accesible
En algunos pacientes, una	***enfermedad grave***	del miocardio da lugar
renal crónica: pielonefritis. Las	***enfermedades primarias***	o secundarias del intersticio
La artritis reumatoide es una	***enfermedad evolutiva***	. En conclusión, diremos
De hipertensión arterial y de	***enfermedad coronaria***	soplo cardíaco, orgánico o

5) Concordancia compleja

Este tipo de búsqueda es la que ofrece más posibilidades de interrogación en el corpus técnico del IULA. Esta facilidad se debe a que permite utilizar buena parte de la potencialidad del lenguaje de interrogación CQP. Con esta opción se podrán hacer, además de las que ya se podían hacer en la concordancia estándar, interrogaciones sobre un número ilimitado de unidades, interrogaciones sobre todos los tipos de combinaciones de formas, lemas y/o categorías, cálculos de frecuencias sobre formas, lemas o categorías, etcétera.

Para especificarlo en la búsqueda hay que hacerlo de la manera siguiente:

Búsqueda de	Expresión
Una forma concreta	[word = "ejemplos"]
Un lema	[lemma = "ejemplo"]
Una categoría morfológica	[pos = "N.*"]
Opciones combinadas	[lemma="ser" & !(word= "soy"| word= "somos") & pos="V.*"]

Ejemplo: En un subcorpus de anatomía, buscamos todas las combinaciones de nombre común con adjetivo, ordenadas por frecuencia, con la intención de encontrar posibles adjetivos con valor especializado que coocurran con distintos nombres confiriendo valor especializado a la unidad poliléxica. El resultado de la búsqueda sería:

línea	medio	69
célula	eucariota	68
cara	anterior	63
tubo	digestivo	56
pared	abdominal	47
cara	posterior	46
	lateral	38
célula	folicular	36
pared	torácico	35
plexo	braquial	35
visión	anterior	35
membrana	plasmático	34
glándula	tiroides	34
parte	superior	34
sistema	inmunitario	33
ganglio	linfático	31
sistema	nervioso	31
miembro	superior	30

Esta es sólo una muestra parcial de los resultados que pueden obtenerse con BwanaNet, pues la búsqueda compleja permite explotar íntegramente el corpus etiquetado y lematizado del IULA. Para finales del año en curso (2004) está previsto además que sea operativo el acceso multilingüe a los datos.

Notas

[1] Han participado como responsables de áreas de trabajo: Carme Bach y Jordi Vivaldi.

Hierbas, plantas, animales..., lengua y traducción (y II)

Enrique Bernárdez
Universidad Complutense de Madrid (España)

Vimos en el último número de *Panace@* (pág. 5) el error histórico al que nos llevaba traducir el inglés *corn* como *maíz* sin pensar más que en (parte de) la equivalencia léxica. Pero no son éstos los únicos errores con los que nos encontramos y de los que, con frecuencia, ni nos damos cuenta. Sucede con los nombres de plantas y de animales, sobre todo aves y peces. Los diccionarios no suelen ser demasiado útiles, porque su función no es proporcionar información sobre el hábitat, la forma de vida y demás detalles interesantes de plantas y animales. Podemos encontrar en uno, por ejemplo, que el aleman *Eiche* puede ser tanto *encina* como *roble* (ambos son *Quercus* en la denominación científica). La única posibilidad de decidir bien es conocer suficientemente ambos árboles para identificar las diferencias, sean de hábitat (en los Alpes son más frecuentes los robles, pese a lo que se tradujo en una novela alemana) o de cualquier otra característica. El traductor tendrá que familiarizarse con el nombre de la planta, del ave o el pez, aunque a lo mejor la primera entrada del diccionario fuese la correcta; pero es imprescindible asegurarse, para no situar en el frío norte escandinavo un pajarito de nuestros campos estivales o para evitar que un pez de río aparezca bogando feliz por el océano Índico. Habrá que echar mano, por tanto, no solo de enciclopedias, sino también de guías especializadas en estos seres, de los que, en general, nunca sabemos suficiente. Muchas veces habremos de trabajar a partir de la denominación científica, que es lo único seguro a ciencia cierta. Incluso en una traducción tuve que optar por usar esos nombres científicos (del estilo de *Myrica gale* y *Espidia tormentosa*) para *traducir* nombres ingleses de hierbas norteamericanas inexistentes en español corriente; solución imposible, ciertamente, si se hubiera tratado de una novela, por ejemplo. Claro que a veces surgen problemas aun más curiosos. Nada más fácil, por ejemplo, que traducir el inglés *robin*: es un *petirrojo* (o *pechicolorado*); pero resulta que en Inglaterra y España es un simpático y huidizo pajarillo, mientras en Norteamérica tiene un tamaño mucho mayor, camina frecuentemente por el suelo sin miedo a las personas y en realidad no está emparentado con el europeo, pues es una especie de tordo o mirlo, aunque con plumas rojas en el pecho. ¿Cómo traducir, entonces? *Petirrojo* no sería opción adecuada para el pájaro norteamericano, porque nos produciría quizá una impresión completamente distinta a la realidad, y podría dar lugar a confusiones con otras referencias en el texto (en el supuesto, claro, de que sepamos reconocer en nuestros parques a un bonito pajarito como petirrojo). Como el traductor no puede saberlo siempre todo, la solución es: ¡mucho ojo y a buscar confirmación!

Reproducido con autorización de *El Trujamán,*
del Centro Virtual Cervantes (<cvc.cervantes.es/trujaman/>).