>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus lingüístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de corpus textuales

Jorge Vivaldi Palatresi
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona
jorge.vivaldi@upf.edu

1 Introducción

Desde hace ya algunas décadas el estudio del lenguaje utilizando técnicas de lingüística de corpus ha ido incrementando su importancia. En la actualidad, la utilización de corpus se considera un recurso básico para prácticamente cualquier estudio relacionado con la lingüística. Esto es así porque los corpus son un recurso que permite obtener gran cantidad de información sobre el comportamiento real de la lengua. Por lo tanto, un corpus correctamente diseñado es fundamental para la validez y aplicabilidad de los resultados obtenidos en cualquier investigación lingüística.

Hoy en día es difícil concebir un corpus que no sea en soporte electrónico y con una codificación tal que tanto investigadores como ordenadores puedan interrogarlo de una manera ágil y eficaz.

Los corpus pueden clasificarse según diferentes puntos de vista: dominio, registro, idioma, fecha de los documentos, método de compilación, información lingüística incorporada, entre otros (Vivaldi, 2008). Algunos de estos aspectos afectan significativamente la manera en que dichos recursos se explotan. Así, por ejemplo, un corpus puede clasificarse según su cobertura (lengua general vs lenguaje de especialidad) o bien según el método empleado para su compilación (oportunístico o planificado). En ambos casos la clasificación que se aplique a un corpus determinado por sí misma no afecta prácticamente el método de explotación.

No puede decirse lo mismo en relación a la clasificación de un corpus en función de la información lingüística añadida (de texto plano, con información morfológica, sintáctica, etc.). Las características intrínsecas de los documentos del corpus en cuanto a la información lingüística añadida condiciona decisivamente tanto el método de explotación a emplear como, en muchos casos, la herramienta a utilizar.

A continuación se recogen un cierto número de recursos que se pueden utilizar para obtener información de corpus. En primer lugar mencionaremos los corpus lingüísticos existentes y disponibles para su consulta. En segundo lugar mencionaremos algunas herramientas para la creación de corpus y algunas herramientas informáticas útiles para hacer una explotación sencilla de corpus planos. Finalmente, presentaremos una serie de recursos que permiten añadir información lingüística de distinto tipo a los corpus planos.

2 Corpus existentes y disponibles para su consulta

En esta sección indicamos algunos de los proyectos de corpus más conocidos limitándonos a revisar los más utilizados en los últimos años en estudios lingüísticos y lexicográficos. Generalmente son fruto de una iniciativa académica con participación de empresas editoriales u otras iniciativas privadas. Muchos de ellos han contado también con el apoyo y la financiación de asociaciones académicas y/o de organismos oficiales.

En general, todos estos recursos son estructuras monolíticas donde el usuario se limita a consultar los datos que las respectivas instituciones ponen a su disposición. Algunos de estos recursos están sujetos a algún tipo de licencia y en consecuencia limitan el número de textos consultables y/o el número de líneas de concordancia obtenidas.

A continuación se incluye una lista no exhaustiva de los principales corpus existentes y accesibles desde Internet.

Información general	
Institución	RAE
Nombre del recurso	CREA
URL	http://corpus.rae.es/creanet.html
Tamaño	154 M palabras
Idiomas	ES
Dominio	General
Características especiales	
Información lingüística	Texto plano (el análisis morfológico existe pero no es accesible desde la web)
Textos paralelos	No
Otras	· Se pueden limitar las consultas por tema, período temporal, ámbito geográfico, etc. · Contextos reducido y amplio

Información general	
Institución	Brigham Young University
Nombre del recurso	Corpus del Español
URL	http://www.corpusdelespanol.org/x.asp
Tamaño	100 M palabras
Idiomas	ES
Dominio	General
Características especiales	
Información lingüística	Analizado morfológicamente
Textos paralelos	No
Características especiales	- Búsqueda según registros y/o períodos temporales - Búsqueda de colocaciones - Ordenación/agrupación de resultados

Información general	
Institución	IULA
Nombre del recurso	IULACT
URL	http://bwananet.iula.upf.edu
Tamaño	30 M palabras
Idiomas	ES, CA, EN
Dominio	General (CA y ES) y lenguajes de especialidad
Características especiales	
Información lingüística	Analizado morfológicamente (etiquetario IULA: demo: http://www.iula.upf.edu/recurs01ca.htm)
Textos paralelos	Sí (subcorpus)
Otras	- Permite la generación personalizada de subcorpus - Búsquedas por combinación de formas, lemas y categorías - Ordenación/agrupación de resultados - Permite búsquedas utilizando la sintaxis CQP

Información general	
Institución	BNC
Nombre del recurso	BNC
URL	http://www.natcorp.ox.ac.uk/index.xml Accesos alternativos: http://corpus.byu.edu/bnc/ (Brigham Young University) http://bncweb.info/ (Universidad de Lancaster)
Tamaño	100 M palabras
Idiomas	EN
Dominio	General
Características especiales	
Información lingüística	Analizado morfológicamente (etiquetario BNC/CLAWS, demo: http://ucrel.lancs.ac.uk/claws/trial.html)
Textos paralelos	No
Otras	- Búsquedas por combinación de formas y categorías - Utilización de expresiones regulares en las búsquedas - Los accesos alternativos pueden incluir características específicas

Información general	
Institución	ANC
Nombre del recurso	ANC/OANC
URL	ANC: (no es accesible, más información en http://www.americannationalcorpus.org/) OANC: se puede descargar y explotar localmente
Tamaño	ANC: 22 M palabras / OANC: 15 M palabras
Idiomas	EN
Dominio	General

Características especiales	
Información lingüística	Analizado morfológicamente (etiquetario Penn Treebank)
Textos paralelos	No
Otras	- Búsquedas por combinación de formas y categorías - Utilización de expresiones regulares en las búsquedas - Los accesos alternativos pueden incluir características específicas

Información general	
Institución	Syddansk University
Nombre del recurso	Corpuseye
URL	http://corp.hum.sdu.dk/
Tamaño	SP (53 M palabras), EN (189 M palabras), FR (71 M palabras), DE (99 M palabras), etc
Idiomas	SP, EN, FR, DE, etc.
Dominio	General
Características especiales	
Información lingüística	Análisis morfológico, dependencias sintácticas (CG grammar)
Textos paralelos	No
Otras	- Permite búsquedas utilizando la sintaxis CQP

Otros corpus disponibles en Internet (y su correspondiente URL) son los siguientes:

- European Corpus Initiative Multilingual Corpus I (ECI/MCI): http://www.elsnet.org/resources/eciCorpus.html
- Corpus disponibles a través de la Universidad de Leeds: http://corpus.leeds.ac.uk/list.html
- European Parliament Proceedings Parallel Corpus 1996-2006: http://www.statmt.org/europarl/
- CORGA: Corpus de Referencia do Galego Actual: http://corpus.cirp.es/corga/
- Corpus del Alemán: http://www.ids-mannheim.de/
- Corpus de alemán/inglés: http://quickie.ikw.uni-osnabrueck.de/CQPdemo/
- CUCweb (Corpus de uso del catalán en la web): http://ramsesii.upf.es/cgi-bin/cucweb/search-form.pl

Es importante recordar que esta no es una lista completa. Además, algunos de estos recursos están sujetos a un registro previo o bien a algún tipo de licencia; en consecuencia, cuando permiten su consulta, se limita el número de textos consultables y/o el número de líneas de concordancia obtenidas. En algunos es necesaria una licencia y la instalación de una herramienta específica de consulta.

También puede consultarse en este mismo número de la revista Tradumática el artículo "Catalogue of Free-Access Translation Related Corpora" donde se incluye un listado amplio y variado de corpus gratuitos.

3 Creación de corpus

Todos los recursos mencionados en la sección 2 son estructuras compactas donde el usuario se limita a consultar los datos que las respectivas instituciones ponen a su disposición. Además, están disponibles para pocas lenguas y a veces son colecciones heterogéneas construidas ad hoc restringidas en cuanto a su tamaño y/o tipos de texto. También es frecuente encontrar corpus construidos en base a noticias de prensa, estos recursos son fáciles de obtener y tienen múltiples aplicaciones especialmente en el desarrollo de herramientas de tratamiento de corpus.

Hoy en día la web es una fuente prácticamente inagotable de documentos de todo tipo y en muchos idiomas que pueden utilizarse para compilar un corpus. Hay que tener en cuenta, de todas maneras, que algunos sitios de la web sólo indexan documentos, es decir que estos sitios sólo cumplen una función de portal de acceso. El acceso al documento real requiere algún otro tipo de iniciativa.

Los corpus obtenidos a partir de la web se suelen denominar “oportunísticos” en oposición a los corpus planificados (sólo incorporan documentos previamente seleccionados) y a pesar de sus inconvenientes (falta de control, documentación, falta de representatividad, etc.) suelen utilizarse para diferentes propósitos (para complementar corpus de referencia con material actualizado y crear corpus exploratorios de lenguajes de especialidad entre otros usos).

Para compilar un corpus de este tipo es suficiente con la utilización sistemática de cualquiera de los motores de búsqueda comerciales existentes (tales como Yahoo, Google y AltaVista entre otros). Esta aproximación tiene el inconveniente de que la inversión de tiempo puede ser importante y no siempre justificable. Para solventar este inconveniente existen algunas herramientas que automatizan el proceso de búsqueda de documentos. Tanto si la exploración se hace manualmente o a través de programas especializados, la idea básica es utilizar una o más palabras semilla que permitan recuperar (combinándolas de alguna manera) algunos documentos. Estos documentos una vez explorados convenientemente, permiten obtener nuevos términos que se utilizan en combinación (o no) con los anteriores para obtener nuevos documentos y así se repite el ciclo hasta completar un corpus del tamaño deseado.

Los corpus así obtenidos son necesariamente sobre lenguajes de especialidad. Obtener un corpus general es más difícil aunque no imposible. Existen varias técnicas para explorar sistemáticamente la web o bien utilizar la iniciativa ODP (Open Directory Project, http://www.dmoz.org/) como se sugiere en Liu et al. (2006).

Existen al menos tres programas de acceso libre que tienen esta funcionalidad y otro programa comercial que combina esta función de creación de corpus con el análisis lingüístico. A continuación se incluyen los datos más relevantes de estas iniciativas.

Nombre	BootCaT
URL	http://sslmit.unibo.it/~baroni/bootcat.html
Funcionalidad	Esta es una herramienta que no hace más que aplicar el método de compilación automática de corpus. Su utilización puede ser difícil para usuarios no iniciados en informática. Más información en Baroni et al. (2004)

Nombre	Sketch Engine
URL	http://www.sketchengine.co.uk/
Funcionalidad	Es un sistema de interrogación de corpus que permite obtener de manera automática el comportamiento colocacional de una palabra en un corpus. Incorpora una serie de corpus de varias lenguas y permite la creación de corpus a partir de Internet (utilizando una versión de WaCky)

Nombre	Jaguar
URL	http://rc16.upf.es/cgi-bin/jaguar/jaguar.pl
Funcionalidad	Esta es una herramienta experimental que permite tanto la creación de un corpus como su explotación. La creación de corpus permite limitar la búsqueda de documentos a ciertos formatos (pdf, postscript, etc.) que en ciertos casos puede dar óptimos resultados. Proporciona concordancias, listados de n-gramas, asociación, distribución y similitud. Más información en Nazar et al. (2008)

Nombre	WebCorp
URL	http://www.webcorp.org.uk/
Funcionalidad	Esta herramienta usa motores de búsqueda comerciales para localizar páginas relevantes en Internet en relación a la/s palabra/s introducidas por el usuario. A continuación recupera dichas páginas y presenta el resultado en forma de concordancias superando así las limitaciones de los motores de búsqueda como generadores de concordancias. La secuencia a buscar se puede expresar en forma de comodines. Además el usuario dispone de opciones para formatear los resultados, escoger el motor de búsqueda a utilizar, limitar la búsqueda a ciertos dominios y/o ámbitos, etc. Más información en Renouf et al. (2007)

4 Explotación de corpus planos

Este tipo de corpus son los más fáciles de obtener y personalizar; en consecuencia son también los más utilizados a pesar de que no incluyen información lingüística asociada. Este hecho, justamente, hace que sea relativamente fácil crear herramientas de consulta y exploración. Además tienen la ventaja adicional de que pueden ser independientes de la lengua ya que consideran una palabra como una simple cadena de caracteres y no necesitan hacer análisis lingüístico alguno. Eventualmente es el usuario que, a través de ficheros

auxiliares, tiene la posibilidad de incorporar alguna información específica de la lengua (p. ej. lematización).

El uso principal de estos programas es generar concordancias, es decir, dada una palabra cualquiera obtener todas la ocurrencias de dicha palabra en un texto o una colección de ellos.

Otra aplicación de los corpus es en el estudio de los usos colocacionales de las palabras. Éste es un aspecto muy importante en las actividades relacionadas con la enseñanza y aprendizaje de lenguas, la traducción automática y la lexicografía, tanto monolingüe como bilingüe. Por esta razón, es muy útil contar con herramientas computacionales que ofrezcan listados de colocaciones, así como la posibilidad de ordenarlas según diferentes cálculos estadísticos. Es por estos motivos que esta es otra característica muy común en este tipo de herramientas.

Algunos añaden alguna función adicional como puede ser listados de frecuencias y n-gramas (secuencia de n palabras, siendo n un número entero positivo), ordenación de las concordancias en función de diferentes criterios, cálculos estadísticos sencillos, etc.

A continuación se enumeran algunas de estas herramientas. En cada una de las URLs es posible encontrar información detallada sobre cada una de las herramientas.

WordSmith Tools

URL	http://www.lexically.net/wordsmith/
Entorno de funcionamiento	Windows
Producto	Comercial
Características adicionales	- Texto de entrada: plano. Incorpora también conversores desde otros formatos (PDF, Word, etc.) - Identifica agrupaciones consecutivas (o no) - Posibilidad de alinear dos textos - Facilidades para construir un corpus con textos de Internet

MonoConc Pro

URL	http://www.athel.com/mono.html
Entorno de funcionamiento	Windows
Producto	Comercial
Características adicionales	-

AntConc

URL	http://www.antlab.sci.waseda.ac.jp/software.html
Entorno de funcionamiento	Windows, Linux, Macintosh
Producto	Freeware
Características adicionales	-

WConcord

URL	http://www.linglit.tu-darmstadt.de/index.php?id=linguistics
Entorno de funcionamiento	Windows
Producto	Freeware
Características adicionales	-

kfNgram

URL	http://www.kwicfinder.com/kfNgram/kfNgramHelp.html
Entorno de funcionamiento	Windows
Producto	Freeware
Características adicionales	-

5 Otros recursos relacionados con los corpus

Existen en la web múltiples recursos relacionados con los corpus. Estos van desde corpus ya construidos hasta herramientas informáticas para incorporar información lingüística a textos.

Para obtener corpus ya compilados en multitud lenguas cabe destacar estas dos organizaciones:

- ELRA: http://www.elra.info
- LDC: http://www.ldc.upenn.edu/

También existe un cierto número de herramientas informáticas de uso libre que permiten añadir información lingüística a textos (p. ej. información morfológica). Cabe señalar sin embargo que muchas de estas herramientas están diseñadas básicamente para el inglés, su aplicación a otras lenguas es posible aunque, en la mayoría de casos, con un cierto esfuerzo.

También hay que destacar que el uso de estas herramientas generalmente requiere de cierta destreza y familiaridad con el uso de herramientas informáticas.

Los recursos más interesantes y con funcionalidades muy diversas son los siguientes:

- GATE (http://gate.ac.uk/): entorno para el desarrollo de herramientas para el procesamiento del lenguaje.
- APACHE UIMA (http://incubator.apache.org/uima/): herramientas (y entorno para desarrollo de herramientas) para el PLN (procesamiento del lenguaje natural).
- JULIE NLP Toolsuite (http://www.julielab.de/Resources/Software/Tools.html): colección de herramientas para el PLN.
- NLTK (http://www.nltk.org/): herramienta concebida para la enseñanza de técnicas de PLN. Incluye también recursos (corpus, diccionarios, etc.)
- Freeling (http://garraf.epsevg.upc.es/freeling/): herramienta para el PLN de textos en varias lenguas (inglés, español, catalán, gallego e italiano). Incluye preproceso, análisis y desambiguación morfológica, análisis sintáctico, etc.
- Language Technology World (http://www.lt-world.org/). Catálogo sobre recursos, proyectos, organizaciones etc. relacionados con tecnologías de la lengua.

- The ACL NLP/CL Universe (http://tangra.si.umich.edu/clair/universe-rk/html/u/db/acl/): Catálogos sobre recursos disponibles en la web y relacionados con la lingüística computacional y la lingüística de corpus en particular.

Cabe destacar también que algunas de las herramientas mencionadas tienen la posibilidad de ser consultadas en línea.

6 Conclusiones

Es indudable que los corpus son un recurso importante en todo tipo de estudios relacionados de una u otra manera con la lengua. La existencia de gran cantidad de recursos ya compilados y/o herramientas para su creación y procesamiento confirma claramente esta afirmación. Esto es particularmente cierto para los corpus planos, es decir, sin ningún tipo de información lingüística asociada donde un gran número de organizaciones relevantes permiten el acceso libre a sus recursos. A medida que se incorpora más información lingüística la gran diferencia entre el nivel de esta información y sus formatos correspondientes hace que las posibilidades de consulta sea mucho más dificultosa y consecuentemente más escasa la oferta.

7 Referencias

Baroni M. y S. Bernardini (2004). “BootCaT: Bootstrapping corpora and terms from the web”. *Actas de LREC 2004*. Lisboa.

Liu V. y J.R. Curran (2006). “Web Text Corpus for Natural Language Processing”. *EACL* p. 233-240.

Nazar, R.; J. Vivaldi y M.T. Cabré (2008). “A Suite to Compile and Analyze an LSP Corpus”. *Actas de LREC 2008*. Marrakech.

Renouf, A., A. Kehoe y J. Banerjee (2007). “WebCorp: an integrated system for web text search”. En C. Nesselhauf, M. Hundt & C. Biewer (eds.), *Corpus Linguistics and the Web.* Ámsterdam: Rodopi.

Vivaldi, J. (2009). “Corpus and exploitation tool: IULACT and *bwanaNet”*. *Actas de CILC-09*, p. 224-239. Murcia.