Procesamiento del Lenguaje Natural, num. 29 (2002), pp.293-294 recibido 12-07-2002; aceptado 15-07-2002

BancTrad: un banco de corpus anotados con interfaz web

Toni Badia, Gemma Boleda, Jenny Brumme, Carme Colominas,
Mireia Garmendia, Martí Quixal
Universitat Pompeu Fabra
Rambla 30-32
08002 Barcelona
{toni.badia, gemma.boleda, jenny.brumme, carme.colominas, marti.quixal}@trad.upf.es,
mireia.garmendia01@campus.upf.es

Resumen: BancTrad es un proyecto que proporciona, vía web, acceso a corpus alineados. Lo novedoso del proyecto es la integración de herramientas de PLN, explotación de corpus e interacción cliente/servidor (CGI)
Palabras clave: corpus paralelos, interfaz web, *shallow parsing*, anotación de corpus

Abstract: BancTrad has the goal of creating a web interface to aligned corpora. The novelty of BancTrad is the integration of a few pre-existing tools: morphosyntactic parsers, corpus exploitation tools and client/server communication tools (CGIs)
Keywords: parallel corpora, web interface, shallow parsing, corpus annotation

1 Objetivos

Las herramientas desarrolladas en el campo de la ingeniería lingüística suponen un gran potencial para la traducción, tanto para la didáctica como para la investigación y la práctica profesional. El proyecto BancTrad (http://glotis.upf.es/bt/index.html, v. Badia *et al.* 2002) pretende explorar este campo y proporcionar, vía web, una interfaz de acceso a corpus paralelos.

La particularidad de BancTrad es que, al integrar herramientas de procesamiento del lenguaje natural, no sólo permite la búsqueda de palabras por forma, sino que también ofrece la posibilidad de realizar búsquedas por rasgos lingüísticos (lema, categoría morfológica y, en el caso del catalán, función sintáctica) o extralingüísticos (p. ej. , género textual, registro o tema).

2 Proceso de confección de los corpus

Las lenguas del corpus de BancTrad son las lenguas de trabajo de la Facultad de Traducción y Interpretación (FTI) de la Universidad Pompeu Fabra: catalán, castellano, inglés, francés y alemán.

Se pretende que los textos de BancTrad sean representativos en cuanto a textos traducidos se refiere, es decir, que no tengan un carácter normativo sino descriptivo. Por este motivo se recopilan documentos de fuentes muy diversas (internet, editoriales, docencia) que representan un amplio abanico de tipos de texto, temas y registros.

Una vez seleccionados, los textos se procesan para etiquetarlos (en formato SGML) con información extralingüística, mediante un formulario de MS Word con código *Visual Basic*. Ello permite agilizar el marcaje, a la vez que se evitan errores tipográficos.

En cuanto a la alineación, los textos se alinean a nivel oracional mediante la aplicación comercial *DéjàVu Database Maintenance* de Atril. Este programa realiza una prealineación automática, que se puede editar en una interfaz gráfica amigable. Una vez revisados[1], los textos se transfieren al servidor Linux con el fin de continuar el procesamiento, que a partir de este momento será completamente automático.

Los pasos restantes son el análisis lingüístico y el formateo de los corpus. No todas las lenguas siguen el mismo proceso de etiquetado: los textos en catalán se analizan mediante CATCG (Alsina et. al. 2002), un etiquetador morfosintáctico superficial basado en el formalismo de la *Constraint Grammar* y desarrollado en la UPF. En cambio, los textos

[1] El tiempo para realizar el marcaje y revisar la alineación de un documento de 400 palabras es de 5 a 10 minutos. La alineación semiautomática evita muchos errores en la identificación de oraciones que causarían errores en el procesamiento lingüístico ulterior.

ISSN 1135-5948 ©2002 Sociedad Española para el Procesamiento del Lenguaje Natural

en inglés, alemán y francés se etiquetan mediante *TreeTager*, un etiquetador morfológico de base estadística desarrollado en el IMS de la Universidad de Stuttgart (v. Schmid 1995, 1997).

A pesar de que los procesos son diferentes, todas las lenguas reciben un mismo tipo de información lingüística, lo que permite tratar de manera uniforme tanto el procesamiento de los corpus como su consulta. Este diseño modular del proceso hace que se puedan cambiar las herramientas utilizadas en cualquier punto del proceso sin necesidad de cambiar ni la interfaz ni el resto de módulos.

Una vez etiquetados los ficheros de texto, se formatean y se procesan con las herramientas del *Corpus WorkBench* (CWB; v. Christ 1994). Así, los corpus pueden consultarse a través del CQP (*Corpus Query Processor*), herramienta que ofrece una gran flexibilidad y expresividad en las búsquedas. No obstante, al tratarse de una herramienta poco amigable, se diseñó una interfaz web para los usuarios potenciales de BancTrad.

Técnicamente, la novedad de BancTrad es la integración de varias herramientas para acceder a corpus paralelos a través de Internet. La interfaz gráfica se hizo en HTML para la independencia de plataforma y sistema operativo. La EPI (*External Program Interface*) se realizó con el lenguaje Perl, usando la CGI estándar y un módulo de CQP diseñado por sus autores con el fin de efectuar la interacción con la web.

3 Resultados: la interfaz de búsqueda

La interfaz web de BancTrad permite acceder a los corpus a través de tres niveles de búsqueda, en función de los conocimientos y las necesidades del usuario (los distintos niveles permiten acceder a los corpus sin conocimientos profundos de lingüística, de expresiones regulares o de CQP):

- **básico**: búsqueda de secuencias de palabras por forma
- **intermedio**: búsqueda de secuencias de hasta cinco palabras, con posibilidad de realizar búsquedas por forma, lema, categoría y, para el catalán, función sintáctica
- **experto**: búsqueda con la sintaxis de CQP

En los tres niveles existe la opción de formular restricciones no sólo sobre los vocablos, sino sobre las características extralingüísticas de los textos. Asimismo, son posibles las búsquedas de texto entero, para obtener textos paralelos con miras a la didáctica y la práctica profesional de la traducción.

Finalmente, es importante destacar que BancTrad permite incorporar otros corpus (multilingües o monolingües) con muy poco esfuerzo técnico, lo que permitiría consultar no sólo los corpus creados expresamente para BancTrad, sino también otros (p. ej., el *British National Corpus* o el *Frankfurter Rundschau*) utilizando la misma interfaz.

Así, BancTrad permite aprovechar los resultados obtenidos en lingüística computacional, aplicándolos a un campo diferente, pero relacionado, como es la traducción. A pesar de que inicialmente fue creado como herramienta para la didáctica de la traducción, sus posibilidades abarcan otros campos como la investigación lingüística (sobre todo en lingüística comparada), la creación de recursos lingüísticos (traducción automática, diccionarios bilingües) o la traducción profesional.

Bibliografía

Alsina, A., Badia, T., Boleda, G., Bott, S., Gil, A., Quixal, M. y Valentín, O. (2002) CATCG: a general purpose parsing tool applied, en Proceedings of the Third International Conference on Language Resources and Evaluation, LREC, Las Palmas.

Badia, T., Boleda, G., Colominas, C., González, A., Garmendia, M. y Quixal, M. (2002) BancTrad: a web interface for integrated access to parallel annotated corpora, en Proceedings of the LREC'02 workshop on Language Resources for Translation Work and Research, Las Palmas, 28 mayo 2002.

Christ, Oliver (1994) "A modular and flexible architecture for an integrated corpus query system", COMPLEX'94, Budapest. (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/)

Schmid, H. (1995) Improvements in Part-of-Speech Tagging with an Application to German, in Proceedings of the ACL SIGDAT-Workshop, pp. 47-50.

Schmid, H. (1997) Probabilistic Part-of-Speech Tagging Using Decision Trees, in Daniel Jones and Harold Somers, editors, New Methods in Language Processing Studies in Computational Linguistics, UCL Press, London, pp. 154-164.