3LB: Construcción de una base de datos de árboles sintáctico semánticos

Entidad financiera: MCyT (Proyecto PROFIT: FIT-150500-2002-411)

Grupos participantes: Universidad de Alicante (UA), Universidad Politécnica de Cataluña (UPC), Universidad del País Vasco (EHU), Fundación Bosch i Gimpera (FBiG), Universidad Politécnica de Valencia (UPV)

Investigadores: I. Aduriz (FBiG); A. Ageno (UPC); B. Arrieta (EHU); J.M. Arriola (EHU); E. Bisbal (UPV); N. Castell (UPC); M. Civit (FBiG); A. Díaz de Ilarraza (coordinadora EHU); B. Fernández (UA); K. Gojenola (EHU); Reda Halkoum (UPC); R. Marcos (UA); L. Marquez (UPC); M.A. Martí (coordinadora FBiG); P. Martínez-Barco (UA); A. Molina (UPV); P. Moreda (UA); L. Moreno (coordinadora UPV); B. Navarro (UA); M. Oronoz (EHU); L. Padró (UPC); M. Palomar (coordinador UA); F. Pla (UPV); H. Rodríguez (coordinador UPC); M. Saiz-Noeda (UA); E. Sanchís (UPV); K. Sarasola (EHU); A. Suárez (UA); M. Taulé (FBiG)

1 Introducción

Este proyecto, de título "Construcción de una base de datos de árboles sintáctico semánticos", fue solicitado en la convocatoria 2002 como proyecto plurianual (2002-2003) y concedida ayuda para el año 2002 con referencia FIT-150500-2002-411.

El proyecto tiene como objetivo construir un corpus anotados sintáctica, semántica y referencialmente (*treebanks*) para los idiomas español, catalán y euskera.

A pesar de que la construcción de un *treebank* es una tarea costosa, creemos que es una labor imprescindible para el desarrollo de aplicaciones reales en el área del Procesamiento del Lenguaje Natural (PLN) y como tal para el desarrollo de la sociedad de la información. En estas aplicaciones resulta imprescindible la obtención de gramáticas computacionales a partir de corpus que son un primer paso hacia procesos posteriores que requieren más elaboración. Entre estos procesos se halla la delimitación de las entidades discursivas, lo que, junto con la identificación de los elementos anafóricos y correferentes mejora sustancialmente la calidad de los sistemas de Traducción Automática (TA), Extracción de Información (EI), Recuperación de Información (RI), Resumen Automático (RA) y sistemas de Pregunta-Respuesta (PR). Otras tareas lingüísticas que pueden abordarse si se dispone de un treebank son el aprendizaje de restricciones de selección o el de los patrones de subcategorización de los verbos.

A nivel puramente lingüístico, el *treebank* es una base de datos imprescindible para el estudio de la lengua ya que proporciona ejemplos analizados/anotados de lenguaje real. El estudio lingüístico revierte directamente en la mejora de la calidad de los recursos mencionados, dotándolos de una mayor robustez.

2 Objetivos y recursos

Como se ha comentado, el objetivo global de este proyecto es construir tres corpus anotados sintácticamente (*treebanks*) para el español, catalán y euskera. Además de la anotación sintáctica, se realizará una anotación semántica mediante los *synsets* de los diferentes *wordnets*[1] elaborados en cada lengua, así como una anotación de los elementos anafóricos y elípticos y la correferencia. Para el español y el catalán el volumen del corpus será de 100.000 palabras cada uno, en el caso del euskera 50.000 por razones de mayor complejidad notacional y menor cobertura del wordnet de que se dispone (35.000 entradas frente a las 100.000 existentes para el castellano o las 65.000 para el catalán).

El corpus CLiC-TALP[2] para el español consta actualmente de 100.000 palabras anotadas manualmente a nivel morfosintáctico. El resto del corpus, hasta 5,5 millones de

[1] Véase http://www.cogsci.princeton.edu/~wn

[2] Véase http://clic.fil.ub.es/recursos/corpus.shtml

palabras está anotado a nivel morfosintáctico de forma automática, con una tasa de error del orden de un 3%.

El corpus para el euskera del que disponemos para este proyecto consta de 40.000 palabras anotadas manualmente a nivel morfosintáctico. En este proyecto se trataría de etiquetar este corpus sintáctica y semánticamente según la propuesta y ampliarlo hasta 50.000 palabras con anotación morfológica, sintáctica y semántica.

3 Avances y resultados

Los principales avances realizados se resumen en los siguientes puntos según el plan de trabajo:

- Coordinación del proyecto: En un total de 5 reuniones (Valladolid, Barcelona, San Sebastian, Alicante y Valencia) se han ido definiendo y concretando las tareas relativas a objetivos técnicos y científicos por un lado y lingüísticos por otro. Reuniones paralelas de grupos de trabajo técnico, administrativo y lingüístico han perfilado las propuestas que posteriormente se han puesto en común y han permitido entrelazar el trabajo de investigadores de los distintos centros.
- Integración de herramientas y recursos para la elaboración de los etiquetados: se han propuesto y desarrollado las herramientas de ayuda para el etiquetado de los corpus. Se ha adaptado el editor de árboles TreeTrans[3] como herramienta de etiquetado sintáctico (constituyentes y funciones) de los corpus. Para el caso de la anotación semántica se ha desarrollado una herramienta específica (3LB-SAT) para agilizar el proceso de etiquetado semántico. Se encuentran en desarrollo las herramientas de ayuda al etiquetado referencial: anafórico, de elipsis y de cadenas de correferencia.
- Anotación y supervisión de los corpus: se ha definido y diseñado el esquema de anotación sintáctica, semántica y anafórica con una sólida base lingüística y metodológica para cada una de la propuestas de anotación de cada uno de los idiomas. El etiquetado morfosintáctico abarca más de 1000 frases lo que supone un 25% del total del proyecto. La fase de etiquetado semántico con la nueva herramienta se encuentra en proceso en estos momentos mientras que el etiquetado correferencial comenzará tan pronto como la herramienta esté disponible.
- Evaluación y diseminación de los resultados: desde el comienzo del proyecto está en marcha una página web[4] que describe el proyecto desde la que es posible enlazar al servidor de información[5] cuyo objetivo es el de poner en contacto las ideas y resultados de cada grupo de trabajo creando un repositorio de información único, además de servir como plataforma divulgativa de los resultados y publicaciones derivadas del proyecto.

[3] Véase http://agtk.sourceforge.net/

[4] Véase http://www.dlsi.ua.es/projectes/3lb/

[5] Véase http://gplsi.dlsi.ua.es:9998/

Bibliografía

Aduriz I., Aldezabal I., Aranzabe M., Arrieta B., Arriola J.M, Atutxa A., Díaz de Ilarraza A., Gojenola K., Oronoz M., Sarasola K. (2002) Construcción de un corpus etiquetado sintácticamente para el euskera. Procesamiento del Lenguaje Natural 29.

Aduriz I., et al. *Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing in Corpus Linguistics Around the World*. Language and Computers. Rodopi. Netherlands. (en prensa).

Bisbal E., A. Molina, L. Moreno, F. Pla, M. Saiz-Noeda y E. Sanchís (2003). 3LB-SAT: Una herramienta de anotación semántica. XIX Congreso de la SEPLN. (en prensa).

Civit M. y M.A. Martí (2002). Design Principles for a Spanish Treebank. 1st Workshop on Treebanks and Linguistic Theories (TLT02), Sozopol, Bulgaria.

Civit M., M.A. Martí, B. Navarro, N. Bufí, B. Ferrández y R. Marcos (2003). Issues on the Syntactic Annotation of Cast3LB. Proceedings of the LINC'03. EACL Workshop. Budapest, Hungary.

Navarro B., M. Civit, M. A. Martí, R. Marcos, y B. Fernández (2003) "Syntactic, Semantic and Pragmatic Annotation in Cast3LB". Proceedings of the SproLaC'03. UCREL, Lancaster University.