Compilación y anotación de un corpus paralelo neerlandés - español

Organismo Financiador: Universidad de Nijmegen y ELRA (?)

Grupo Participante en el Proyecto: TOSCA Research Group for Corpus Linguistics

Persona de contacto: Jos Hallebeek. Universiteit van Nijmegen. E-mail: J.Hallebeek@let.kun.nl

Resumen

En el área de la lingüística contrastiva actualmente se siente cada vez más la necesidad de disponer de una base de datos amplia y objetiva para estudiar las diferencias reales en el uso actual de las lenguas que se contrasten. El corpus paralelo que contiene los mismos textos en diferentes lenguas resulta ser el medio más indicado para llegar a crear una base de datos de ese tipo.

Como todavía falta un corpus paralelo neerlandés - español de textos no técnicos, el Grupo de Investigación para Lingüística de Corpus, TOSCA, de la universidad de Nijmegen ha decidido pasar a la compilación de semejante corpus aprovechando los conocimientos y los recursos técnicos que tenemos.

El corpus paralelo neerlandés - español constará de un millón de palabras procedentes de novelas de autores contemporáneos españoles y holandeses que han sido traducidas. Es decir que se trata de textos originales en español y en neerlandés que tienen una versión impresa en la otra lengua. Se incluirán fragmentos de 20.000 palabras cada uno. De hecho, el corpus paralelo comprenderá cuatro subcorpora:

- . uno de textos originales en español;
- . uno de textos originales en neerlandés;
- . uno de textos traducidos al español;
- . uno de textos traducidos al holandés.

Las actividades que se irán desarrollando en el curso del proyecto serán:

- a. recopilar y escanear los textos;
- b. agregar a los textos el *markup SGML*;
- c. etiquetar los textos añadiendo información morfosintáctica (EAGLES);
- d. alinear los textos de ambas lenguas;
- e. almacenar los textos en un sistema de base de datos (COSMASII)

La base de datos obtenida como resultado de la investigación se usará para la composición de una gramática pedagógica del español para neerlandeses y también para desarrollar de material de enseñanza asistida por ordenador para estudiantes de español como segunda lengua.

Acabamos de empezar el proyecto. Hemos hecho la selección de los textos y actualmente estamos pasando los textos por el escáner.

Nombre del director del proyecto, a quien puede dirigirse para más información:

Jos Hallebeek
Vakgroep Spaans, Universiteit van Nijmegen
Erasmusplein, 1
6500 HD Nijmegen (Países Bajos)
E-mail: J.Hallebeek@let.kun.nl