

# Construcción de Analizadores Híbridos de Lenguajes Naturales Definidos sobre un Dominio Semántico Restringido.

**Organismo Financiador:** CICYT

## **Grupos Participantes en el Proyecto:**

- Grupo de Procesamiento de Lenguaje Natural (Universitat Politècnica de València y Universitat d'Alacant).
- Reconocimiento de Formas e Inteligencia Artificial (Universitat Politècnica de València).

**Persona de contacto:** Lidia Moreno Boronat. Universitat Politècnica de València. Email: lmoreno@dsic.upv.es

**Dirección de contacto:** <http://www.dsic.upv.es/users/lina/home.html>

## **Resumen**

El desarrollo de sistemas de procesamiento del Lenguaje Natural presenta dificultades específicas cuando se desea tratar con un conjunto amplio del lenguaje. Junto a fenómenos lingüísticos bien caracterizados (como es el caso de la anáfora, elipsis,...) deben considerarse otros efectos propios del lenguaje espontáneo, que pueden conducir a frases sintácticamente incorrectas. Estos problemas pueden ser abordados desde un punto de vista deductivo, diseñando un modelo para el lenguaje a partir del conocimiento lingüístico que se tiene del mismo, o bien aplicando técnicas de aprendizaje automático.

El objetivo general del proyecto que presentamos es el desarrollo de sistemas de comprensión del lenguaje escrito, de forma que dada una frase en lenguaje natural el sistema proporcione una representación de su significado. Dadas las características de este objetivo se abordarán aplicaciones definidas en dominios semánticos restringidos. La sintaxis será flexible, no se impondrán restricciones sobre las construcciones sintácticas que constituyen las frases, y se estudiarán los problemas derivados de la espontaneidad de las lenguas.

Dentro de este marco de actuación se plantea la construcción de analizadores que aseguren una adecuada cobertura del lenguaje. Para ello se utilizará información estadística sobre distintos niveles de conocimiento extraída de corpora previamente etiquetados. Además, incluso en aquellas situaciones de

análisis en las que no se obtenga una derivación completa, los análisis parciales obtenidos se pueden utilizar para completar el significado de la frase, y por tanto corregir presuntos errores, o para establecer sencillos mecanismos de diálogo que permitan descubrir el significado de la frase analizada.

Para el logro de este objetivo planteamos las siguientes líneas de actuación:

- a) Desarrollo de modelos regulares estadísticos para la descripción del lenguaje objeto de estudio.
- b) Desarrollo de gramáticas incontextuales para el análisis sintáctico-semántico de dicho lenguaje.
- c) Desarrollo de analizadores híbridos basados en gramáticas incontextuales y modelos regulares estadísticos de forma que se garantice la cobertura del lenguaje, incluidos los fenómenos de habla espontánea y las construcciones gramaticalmente incorrectas.
- d) Desarrollo de métodos de análisis parciales para la recuperación de errores producidos por construcciones sintácticas incorrectas o derivados del mismo proceso de interpretación.

El espacio temporal previsto para llevar a cabo los desarrollos presentados en estas líneas de actuación es de tres años, durante los cuales el trabajo a realizar se ha planificado en los siguientes módulos:

### *Módulo 1 : Definición de la tarea.*

T1 : Estudio de las características del lenguaje a tratar

T2 : Definición de la tarea

*Módulo 2: Modelos estadísticos y etiquetado de textos.*

T1: Definición de categorías.

T2: Construcción del diccionario.

T3: Estudio y desarrollo de técnicas de etiquetado automático de textos.

T4: Evaluación del proceso de etiquetado.

T5: Obtención automática de modelos regulares estadísticos a partir de datos.

*Módulo 3: Análisis sintáctico-semántico.*

T1: Definición de una gramática incontextual en base al conocimiento lingüístico.

T2: Estudio y desarrollo de estrategias para tratar distintos fenómenos lingüísticos (anáfora, elipsis, extraposición)

T3: Desarrollo de mecanismos de transformación de las reglas gramaticales en analizadores sintácticos-semánticos

T4: Construcción de analizadores híbridos .

T5 : Obtención de análisis parciales y estudio de su aplicación para recuperación de información.