

# *Proyectos*



## Interlex Project. MLIS-103

**Desarrollo de bases de datos generales y multilingües que se explotarán en Internet, a partir de diccionarios de traducción en formato electrónico.**

### Coordinator:

- UNIVERSIDAD ALFONSO X EL SABIO (SPAIN). (Department of Applied Languages and the Centre for Data Processing). Avda Universidad 1, Villanueva de la Cañada 28691 (Madrid) Spain. Tel.: +34-1-8109737 Fax: +34-1-8109101. Email: diez@uax.es.

### Partners:

- LA MAISON DU DICTIONNAIRE(FRANCE). 98 Bld du Montparnasse 75014 Paris France. Tel.: +33-1-43221293. Fax: +33-1-43220177. Email: feutry@aol.com.
- EDITORIAL EVEREST, S.A. (SPAIN). Carretera León-La Coruña km 5, 24080 León, Spain. Tel.: +34-87-802020 Fax: +34-87-801251 Email: publicaciones@everest.es.
- EVEREST EDITORA (PORTUGAL). Parque Industrial Meramar II Armazén nº 1, Portugal. Tel.: +351-1-9152483 Fax: +351-1-9152525 Email: publicaciones@everest.es.
- TECNOLINGUA S.A. (SPAIN). Morillo, 11,7º B Alcalá de Henares 28805 (Madrid) Spain. Tel.: +34-1-8835806 Fax: +34-1-8835806 Email: jsg38746@teleline.es

**Persona de contacto:** Pedro Luis Díez Orzas. Universidad Alfonso X El Sabio. Avenida de la Universidad, 1, E - 28691 VILLANUEVA DE LA CANADA. Tel.: +34-1-8109737 Fax: +34-1-8109101 Email: diez@uax.es

**Dirección de contacto:** <http://www.uax.es/imasd/interlex/>

### Resumen

Como proyecto MLIS, el objetivo de Interlex es convertir diccionarios bilingües y multilingües (generales y terminológicos), actualmente disponibles en formato papel o en CD-ROM, en recursos electrónicos, a los que se pueda acceder a partir de Internet.

El rápido crecimiento de Internet y la demanda, cada vez mayor, de productos electrónicos, como el CD-ROM, ha hecho que las editoriales dedicadas a la publicación de diccionarios conviertan su recursos léxicos en formato electrónico. Sin embargo, este proceso es muy lento, debido, en primer lugar, al alto coste que requiere la inversión, al miedo a la piratería de productos electrónicos y, finalmente, a la confianza en las formas tradicionales de publicación.

En este contexto, Interlex aparece como un proyecto puntero, no sólo porque es una de los primeros de este tipo en este ámbito, también porque reúne, por primera vez, a un grupo de editoriales, de reconocido prestigio, dedicadas a la publicación de diccionarios. El proyecto hará posible el acceso conjunto a recursos

comunes, la inversión que supone es subvencionada por la Unión Europea.

El socio académico y el industrial aportarán el conocimiento técnico y experto necesario para el desarrollo en las diferentes áreas que incluye el proyecto. La Facultad de Lenguas Aplicadas y el Centro de Proceso de Datos de la Universidad Alfonso X el Sabio de Madrid contribuirán con su experiencia académica, técnica y administrativa. La primera incluye, entre otras aspectos, experiencia en el desarrollo de herramientas léxicas y semánticas, análisis de corpus, diseño de interfaces HTML y explotación de bases de datos. Tecnolingua, el socio industrial, una empresa española puntera en el desarrollo de programas de tratamiento automático de la lengua, desarrollará los programas que requiere el desarrollo del producto y servicio que supone Interlex. El proyecto incluye, finalmente, a una serie de editoriales de reconocido prestigio, como son -la editorial Everest (España); Everest-Editora (Portugal) y la Maison du Dictionnaire (Francia). Estas editoriales pondrán sus diccionarios generales y terminológicos,

bilingües y multilingües a disposición de la explotación conjunta que supone Interlex.

Los objetivos principales del proyecto se podrían resumir en los siguientes puntos:

- Crear una fuente de referencia inestimable para profesionales, estudiantes y, en general, todos aquellos interesados en la traducción y las equivalencias entre lenguas.
- Poner a disposición del usuario recursos léxicos compartidos en inglés, español, alemán, francés, italiano y portugués
- Demostrar que es posible publicar recursos léxico en formato electrónico y, además, ponerlos a disposición del usuario en Internet.
- Introducir a todos los involucrados en la publicación y el uso de recursos léxicos en esta innovadora y útil forma de explotación.

Además de las ventajas que se acaban de señalar, el proyecto supondrá avances en el desarrollo tecnológico y metodológico en el área del tratamiento automático de la lengua que se pueden recoger en los siguientes puntos:

Interlex se propone la conversión de diccionarios bilingües y multilingües en bases de datos léxicos con capacidad de indización y búsqueda, que serán explotados en la Red. Con este objetivo, se utilizará GENETER; un formato estándar de codificación basado en SGML y con una estructuración en varios niveles, para los recursos terminológicos, a partir de este formato, se desarrollará una extensión del mismo con el objetivo de dar cabida a los diccionarios generales.

- En primer lugar, se desarrollará un formato común de trabajo al que se convertirán las diferentes codificaciones empleadas en editorial y en cada diccionario.
- A continuación, se desarrollará una base de datos dinámica, la herramientas y la

metodología que será necesario aplicar a cada conjunto de datos.

- En este momento, se llevará a cabo un proceso de comprobación y coherencia entre direcciones de lenguas, tanto en diccionarios bilingües como multilingües.
- En el caso de los recursos multilingües y, siguiendo los estándares propuestos por el proyecto Eurowordnet, se generará un índice oculto que se utilizará para los propósitos señalados en el punto anterior.
- Todos estos productos se integrarán en una interfaz que pondrá a disposición del usuario de Internet los recursos léxicos. Se crearán, asimismo, los motores de búsqueda e indización necesarios y se hará operativa una interfaz de demostración.

Los valores añadidos de este proyecto se pueden concretar en los siguientes puntos:

- En primer lugar, contribución a un mejor entendimiento sobre el funcionamiento de los mecanismos que permiten etiquetar y mejorar los diccionarios bilingües y multilingües, en una base de datos con rápido motor de acceso.
- En segundo lugar, establecimiento de una metodología, que adapta recursos compartidos a un formato común, para explotar diccionarios bilingües y multilingües.
- Finalmente, búsqueda y estudio de fórmulas comerciales que permitan explotar recursos léxicos en Internet; especialmente en lo que se refiere a la relación entre usuario/cliente y desarrollador/proveedor.

El proyecto supondrá una nueva forma de acceder a diccionarios bilingües y multilingües en Internet, en la que el usuario podrá beneficiarse de la continua actualización de estos recursos.

# Interfaces Multimodales para Comunicación Hombre-Máquina.

**Organismo Financiador:** Comisión Interministerial de Ciencia y Tecnología (CICYT)

**Grupos Participantes en el Proyecto:**

- Grupo de Investigación en Procesamiento de Señales y Comunicaciones (Universidad de Granada)
- Grupo de Teoría de Señales (Universidad de Vigo)
- Grupo de Aplicaciones de Procesado de Señal (Universidad Politécnica de Madrid)

**Personas de contacto:**

- Antonio José Rubio Ayuso (E-mail: [rubio@hal.ugr.es](mailto:rubio@hal.ugr.es), <http://ceres.ugr.es>)
- Carmen García Mateo (E-mail: [carmen@tsc.uvigo.es](mailto:carmen@tsc.uvigo.es), <http://www.tsc.uvigo.es/GTS/>)
- Luis A. Hdez. Gómez (E-mail: [luis@gaps.ssr.upm.es](mailto:luis@gaps.ssr.upm.es), <http://www.gaps.ssr.upm.es>)

**Direcciones de contacto:**

- Dpto. de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, Campus Universitario de Fuentenueva, 18071 Granada
- Dpto de Tecnologías de las Comunicaciones, Universidad de Vigo
- E.T.S.I. Telecomunicación (despacho C-303), Ciudad Universitaria s/n, 28040 Madrid.

**Resumen**

El desarrollo tecnológico está permitiendo la realización de tareas poco creativas por parte de máquinas más o menos inteligentes. La creciente complejidad de dichas tareas exige la búsqueda de nuevos métodos de comunicación en la interfaz hombre-máquina. Una de las formas más naturales de comunicación es la oral.

El presente proyecto pretende la creación de nuevas técnicas más naturales para la mencionada comunicación, creando un sistema

capaz de adquirir la voz humana, procesarla de forma que el ordenador sea capaz de entender el mensaje y actuar en consecuencia, generando a continuación respuestas de tipo oral.

El proyecto incluye, pues, problemas de captación de sonido, reconocimiento de voz continua (robustecimiento a nivel acústico y de lenguaje), modelado de diálogos, problemas de síntesis y compresión del sonido, así como los relacionados con la reproducción del sonido en las condiciones adecuadas.

# 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

# ITEM: Recuperación de Información Textual en un Entorno Multilíngüe con Técnicas de Lenguaje Natural.

**Organismo Financiador:** Comisión Interministerial de Ciencia y Tecnología

## Grupos Participantes en el Proyecto:

- Grupo UNED en Procesamiento de Lenguaje Natural
- Grupo UPC en Procesamiento de Lenguaje Natural
- Grupo UPV/EHU en Procesamiento de Lenguaje Natural
- Grupo EB en Procesamiento de Lenguaje Natural

## Personas de contacto:

- Dra. M<sup>a</sup> Felisa Verdejo. Coordinadora del Proyecto. Dept. IECC. ETSI Industriales UNED Madrid. Email: felisa@ieec.uned.es.
- Dr. Horacio Rodríguez. Investigador responsable grupo UPC. Departamento de Lenguajes y Sistemas Informáticos. Universidad Politécnica de Cataluña. Email: horacio@lsi.upc.es
- Dra. Arantza Díaz de Ilarraza. Investigadora responsable grupo UPV/EHU. Departamento de Lenguajes y Sistemas Informáticos. Facultad de informática. Universidad del País Vasco. Email: jipdisaa@si.ehu.es

**Dirección de contacto:** <http://sensei.ieec.uned.es/item>

## Resumen

El objetivo principal del proyecto es explorar y evaluar en qué medida el uso de técnicas de procesamiento del lenguaje natural puede mejorar los procesos de recuperación y de extracción de información en sistemas de datos textuales o multimediales.

El proyecto explora las técnicas de procesamiento de lenguaje natural aplicadas a la recuperación de información en dos vertientes: Por un lado, en la incorporación de tecnologías lingüísticas a los sistemas clásicos de recuperación de información (básicamente estadísticos), para mejorar tanto la indicación de textos como su posterior consulta. Por otro lado, en aumentar el número de usuarios potenciales, facilitando el acceso mediante un sistema de consulta multilíngüe a las bases de datos documentales.

Para ello, un primer paso es la implantación e integración de diversas herramientas lingüísticas ya disponibles, en una plataforma común estándar. A continuación, el proyecto plantea el desarrollo de técnicas de parsing robusto -según las necesidades de consulta e indicación- y la creación de una base de conocimiento conceptual común para todas las lenguas contempladas.

El proyecto está organizado en 7 tareas que se describen a continuación:

### T0-Coordinación

Duración: todo el proyecto  
Estado: activa

### T1-Implantación e integración de herramientas en una plataforma estándar.

Duración: 24 meses  
Estado: activa

### T2-Investigación en técnicas robustas de análisis de texto libre en lenguaje natural.

Duración: 24 meses  
Estado: activa

### T3-Bases de conocimiento ontológicas

Duración: 24 meses  
Estado: activa

### T4-Base documental textual y prototipo de interfaz

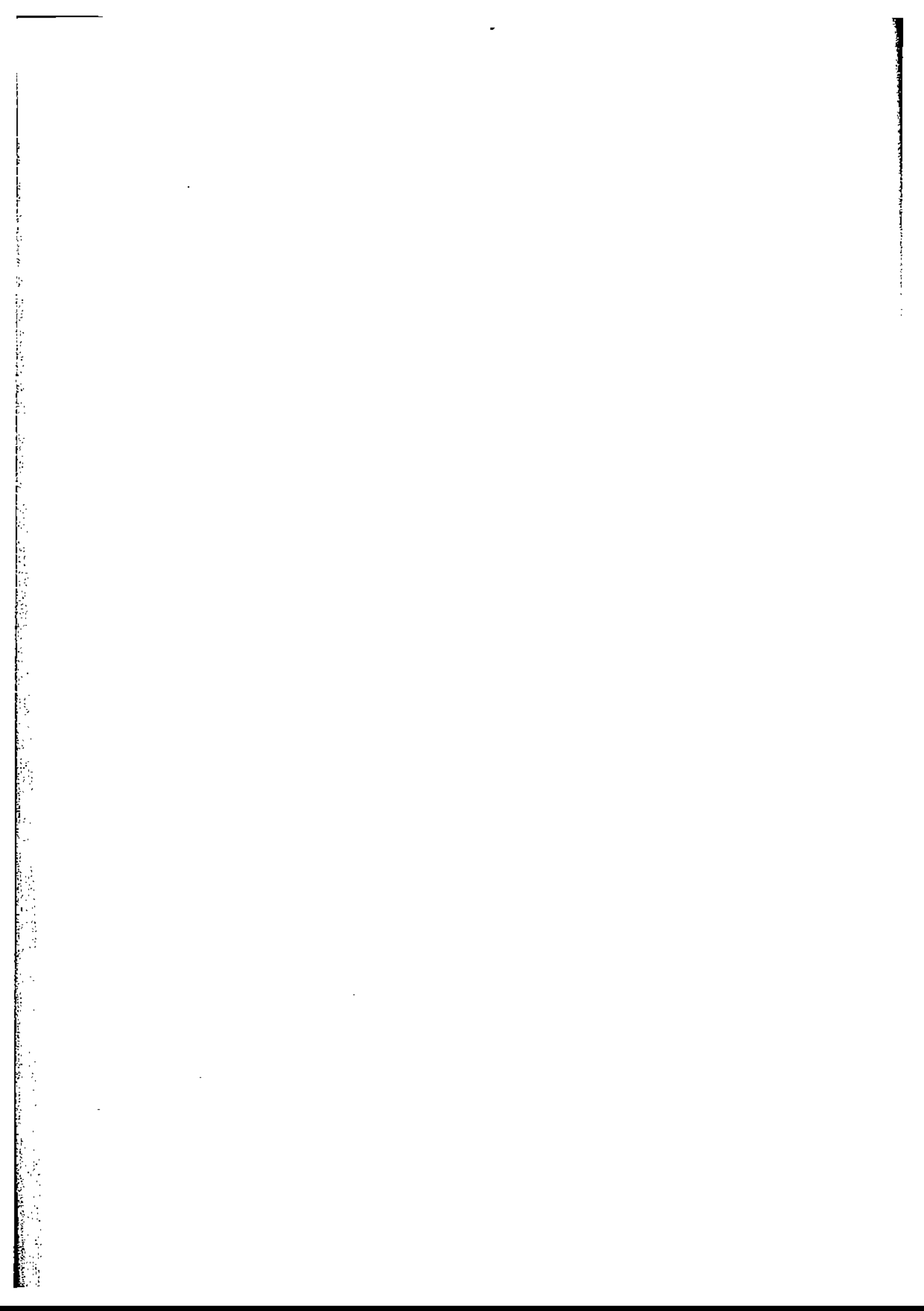
Duración: 12 meses  
Estado: activa

### T5-Extracción de información

Duración: 18 meses  
Estado: activa

### T6-Implantación, prueba y evaluación

Duración: 12 meses  
Estado: sin inicializar.



# LETRAC (Language Engineering for Translators Curricula)

Organismo Financiador: Comisión Europea, DG XIII

## Grupos participantes en el proyecto:

- IAI- Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V, Saarbrücken, Alemania  
Coordinador del proyecto
- Universität des Saarlands, Saarbrücken, Alemania
- Universität Mainz, Mainz, Alemania
- Universitat Pompeu Fabra, Barcelona, España
- Universidade do Porto, Porto, Portugal
- Ionian University, Corfu, Grecia
- Aarhus Business School, Aarhus, Dinamarca
- Conférence Internationale Permanente d'Instituts Universitaires de Traducteurs et Interprètes
- EC-Servicio de Traducción, Bruselas y Luxemburgo

## Personas de contacto:

- Toni Badia. Email: toni.badia@trad.upf.es. Facultat de Traducció i Interpretació. Universitat Pompeu Fabra. Rambla, 30-32, 08002 Barcelona.
- Carme Colominas. Email: carme.colominas@trad.upf.es. Facultat de Traducció i Interpretació. Universitat Pompeu Fabra. Rambla, 30-32, 08002 Barcelona.

## Resumen

### Introducción

El proceso cada vez mayor de tecnificación de nuestra sociedad ha tenido una doble repercusión en el campo de la traducción. Por un lado ha contribuido a aumentar de forma considerable el volumen de textos que tienen que ser traducidos en tanto que ha generado y genera un volumen cada vez mayor de textos técnicos. En este sentido, por tanto, más traducción equivale a más traducción técnica (recordemos p.e. que actualmente la traducción de software (localización) ocupa uno de los primeros puestos en la demanda de traducción).

Por otro lado ha incidido directamente en los instrumentos de trabajo del traductor y por tanto, en su labor cotidiana. La progresiva introducción de la tecnología en las humanidades y concretamente en el campo de la traducción, está dotando a los traductores de toda una serie de instrumentos y recursos informáticos que agiliza enormemente su tarea, pero que evidentemente comporta modificaciones en los requisitos para la competencia profesional.

La especialización, así como un mínimo conocimiento sobre el funcionamiento de un

ordenador y de las diferentes aplicaciones informáticas en el campo de la traducción, han pasado a ser condiciones indispensables de los traductores para su incorporación al mercado laboral.

El objetivo de LETRAC, en este sentido, es establecer una base común para la elaboración e inclusión de elementos curriculares relacionados con las tecnologías del language en los estudios de traducción. De esta manera, la formación de los nuevos traductores se adecuaría realmente a las nuevas necesidades profesionales. LETRAC se propone la confección de una "lista de contenidos ideal", con el fin de que cada centro pueda adaptarlo a sus necesidades, objetivos y otros parámetros de tipo cultural.

### 1. Recopilación de datos

La descripción de nuevos elementos curriculares se ha elaborado en base al estado actual de la cuestión, es decir, teniendo en cuenta:

- (a) los elementos curriculares actuales de los centros de formación de traductores y las posibilidades reales de introducir nuevos elementos

(b) las demandas y necesidades del mercado laboral en que se tendrán que introducir los futuros traductores.

Con este fin se redactaron 6 tipos de formularios dirigidos a los colectivos siguientes:

- (a) profesores de traducción  
estudiantes de traducción  
responsables de facultades o centros de traducción
- (b) empresas que requieren servicio de traducción  
agencias de traducción  
traductores autónomos

Los resultados obtenidos a través de estos cuestionarios aportan una visión amplia a la vez que detallada de las expectativas, necesidades, actitudes etc. de las distintas partes implicadas y corroboran la necesidad de introducir modificaciones en los currículums. De los cuestionarios dirigidos al mundo laboral se desprende que un traductor hoy debe:

- ser un usuario eficiente de PC
- tener experiencia en el uso de programas de traducción asistida y otras aplicaciones relacionadas con la traducción
- conocer los sistemas de gestión terminológica
- tener experiencia en tecnologías de la información

De los cuestionarios de los centros educativos se desprende que en general hay relativamente poca formación en estos aspectos. Parece evidente, por lo tanto, que éstos se deben adaptar a las necesidades del mundo laboral actual.

## 2. Descripción de nuevos elementos curriculares

Este trabajo se está desarrollando actualmente. Los criterios básicos que se tienen en cuenta para la inclusión de determinados contenidos son los siguientes:

- que sean relevantes para el futuro profesional del traductor, en la medida que se adecuan a las necesidades del mercado
- que contribuyan a una visión más amplia y flexible de la actividad del traductor (p.e. si los traductores del futuro tienen un conocimiento básico de las funciones y las posibilidades de los sistemas de traducción automática, podrán ellos mismos contribuir al desarrollo de nuevos proyectos, que en el pasado estaban en manos mayoritariamente de informáticos)
- que contribuyan a entender cómo funciona un ordenador y a adquirir una soltura en su uso que permita la familiarización rápida con posibles nuevas aplicaciones
- que contribuyan a ejercitar el razonamiento formal, que de forma más o menos directa representa una ayuda para la traducción de textos científicos y técnicos

Los elementos propuestos se agruparán aproximadamente bajo los siguientes módulos:

- Rudimentos de informática  
El objetivo principal de los ítems incluidos en este módulo sería la presentación de los componentes básicos de un ordenador, y los distintos tipos de software y hardware según si se trata de introducción, almacenamiento, manipulación o extracción de datos.
- Elementos prácticos de informática  
Los estudiantes deberían adquirir agilidad en el uso de sistemas de edición y en más de un sistema operativo; deberían adquirir conocimientos básicos de programación, sobre tecnologías de la información (e-mail, ftp, telnet), sobre las posibilidades de Internet etc...
- Tratamiento lingüístico  
Los ítems incluidos en este apartado tendrían un carácter teórico-práctico. Se contemplarán desde aspectos generales (procesadores de texto, gestión de documentos, sistemas de terminología, tratamiento de corpus etc.) hasta aspectos específicos de la traducción (memorias de traducción, traducción automática, lenguajes controlados etc.).

# 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.

# Integración de IRIS en ATOS (Sistema de Operador Telefónico Automático).

## Grupos Participantes en el Proyecto:

- Grupo de Investigación en Lingüística Computacional JULIETTA. Universidad de Sevilla
- Telefónica Investigación y Desarrollo. División de Tecnología del Habla.

**Persona de contacto:** Teresa López Soto. Universidad de Sevilla. Email: [teresa@fing.us.es](mailto:teresa@fing.us.es)

**Dirección de contacto:** <http://www.fing.us.es/>

## Resumen

### 1. Introducción

ATOS (Automatic Telephone Operator Service) [Alvarez et al. 1996] es un sistema interactivo dirigido por medio de la voz que permite al usuario activar y configurar todas las opciones que permiten las PABX actuales [Lopez-Soto et al. 1997]. Básicamente consta de 3 módulos: reconocedor de voz, módulo de PLN, y gestor de diálogo (GD). El módulo de RV es capaz de reconocer un vocabulario de unas 20.000 palabras y emplea un modelo estadístico del lenguaje. El sistema es independiente del usuario. La tasa media de error de palabra medida en condiciones de laboratorio es de alrededor de un 95%. Los errores más frecuentes son los debidos a falta de concordancia así como inserciones y omisiones de palabras. El reconocedor emplea unidades dependientes del contexto que se modelan mediante modelos ocultos de Markov semicontinuos.

Las unidades empleadas son trifenemas [Amores et al. 1994]. El módulo de PLN se denomina IRIS. IRIS recibe como cadena de entrada la secuencia generada por un reconocedor de voz y pasa la salida al GD. A continuación nos centraremos en los mecanismos de parsing que han servido para analizar las muestras de lenguaje hablado que genera el reconocedor de voz.

Asimismo presentamos el protocolo CTAC, que asegura una interfaz fluida entre el sistema de PLN y el GD.

### 2. IRIS: Mecanismos de parsing

Los mecanismos de parsing en IRIS se han diseñado para hacer frente a las características propias del lenguaje hablado así como a los

errores producidos en la fase de reconocimiento.

IRIS constituye una versión modificada de Episteme [Amores & Quesada 1997] para esta aplicación. Las incorporaciones más novedosas que se han introducido en la fase de análisis son las siguientes:

#### 2.1 Categoría VOID: Palabras sin información semántica relevante.

Sólo los términos léxicos que aportan información del dominio pasan al parser. El resto caen bajo la categoría VOID y son ignorados. El sistema puede analizar o no estas entradas según cambie el dominio mediante la activación o desactivación del comando correspondiente (VoidWordsIgnore).

#### 2.2 Palabras no incluidas en el léxico.

Mediante el comando NotFoundWordIgnore se previene la entrada al parser de aquellas palabras que no pertenecen al léxico. El comando NotFoundWordDefault permite al sistema aumentar progresivamente el vocabulario asignando una categoría por defecto a la palabra desconocida y comprobando su validez en el análisis.

#### 2.3 Análisis de subcadenas

El módulo de parsing permite el análisis de fragmentos del total de la cadena de entrada con el comando ConfParserSolution-PartialStrings. Cuando una unidad de habla no ha sido correctamente reconocida, o cuando sólo es posible analizar parte de la información contenida en la secuencia de entrada, IRIS puede extraer las subcadenas gramaticales de la cadena completa. Así se consigue que el proceso de análisis no quede interrumpido, y llegue información parcial al GD.

## 2.4 Resolución de la ambigüedad en el análisis

Se ha implementado un mecanismo de control de prioridad entre reglas gramaticales asociadas a un mismo símbolo en la parte izquierda de la producción. Los niveles de prioridad se establecen en función de la plausibilidad semántica de los nodos analizados cuando se emparejan a una producción gramatical. El sistema selecciona el análisis más completo desde el punto de vista de la información que aporta en relación al dominio. En segundo lugar, se han definido unos algoritmos para seleccionar el análisis más completo según los elementos consumidos en el análisis. Hemos considerado los siguientes criterios: posición, longitud, y el denominado "criterio G" o criterio de globalidad.

## 2.5 CTAC: Protocolo de comunicación entre IRIS y el gestor de diálogo.

CTAC es el protocolo que permite la comunicación fluida entre IRIS y el GD. Se define como una especificación tipificada de núcleos semánticos. Este protocolo permite recuperar información aunque las oraciones no se hayan analizado completamente. Esto se consigue gracias a la estructura de rasgos (CTAC) asociada a cada elemento de análisis, desde los nodos no terminales a los raíces, y que se mantiene durante todo el proceso de análisis.

## 3. Resultados actuales

Con los mecanismos descritos arriba se ha conseguido un análisis correcto en el 98.2% de las oraciones del corpus (4.028). El promedio de velocidad de análisis es de 70.240 palabras por segundo.

### References

- [Alvarez et al. 1996] J. Álvarez, C. Caminero, C. Crespo & D. Tapias. 1996. The Natural Language Processing Module for a Voice Assisted Operator at Telefónica I+D. ICSTP-96. Philadelphia
- [Amores & Quesada 1997] J. G. Amores & J. F. Quesada. 1997. Episteme En Procesamiento del Lenguaje natural,21:1-16.
- [Amores et al. 1994] J. G. Amores, J. F. Quesada & D. Tapias. 1994. Traducción

automática basada en el formalismo LFG con entrada y salida por voz. In Comunicaciones de Telefónica I+D,21:1-16

[Lopez-Soto et al. 1997] M. T. Lopez-Soto, J. F. Quesada & J. Álvarez-Cercadillo. 1997. Aplicación de LEKTA al entorno ATOS en Procesamiento del Lenguaje Natural. 21:49-68.

[Quesada 1997] J. F. Quesada 1997. El algoritmo SCP de análisis sintáctico mediante propagación de restricciones Tesis doctoral. Universidad de Sevilla. Abril de 1997.