Esta tesis doctoral contiene un índice que enlaza a cada uno de los capítulos de la misma.
Existen asimismo botones de retorno al índice al principio y final de cada uno de los capítulos.

Ir directamente al índice

Para una correcta visualización del texto es necesaria la versión de Adobe Acrobat Reader 7.0 o posteriores

Aquesta tesi doctoral conté un índex que enllaça a cadascun dels capítols. Existeixen així mateix botons de retorn a l'index al principi i final de cadascun deis capítols .

Anar directament a l'index

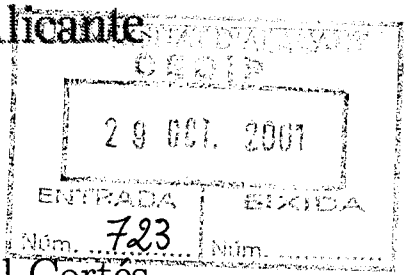
Per a una correcta visualització del text és necessària la versió d' Adobe Acrobat Reader 7.0 o posteriors.

Resolución y generación de la anáfora pronominal en español e inglés en un sistema interlingua de Traducción Automática

Tesis Doctoral

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante



Presentada por Jesús Peral Cortés

Dirigida por Dr. Antonio Ferrández Rodríguez

Alicante, 29 de octubre de 2001

Agradecimientos

Tras finalizar la realización de este trabajo quisiera agradecer su apoyo a todos aquéllos que contribuyeron a la finalización del mismo. Espero no omitir a nadie, pero en el caso de que esto sucediera se podría explicar por el elevado número de personas que han colaborado (en mayor o menor medida) en la confección de este trabajo. Espero, por tanto, que nadie se sienta ofendido.

En primer lugar, quiero destacar la labor realizada por Antonio Ferrández como director de esta Tesis. Su inestimable ayuda desde el inicio de mi etapa investigadora contribuyó decisivamente a que enfocara mi Tesis sobre un tema relacionado con el Procesamiento del Lenguaje Natural (al cual me "enganché"). Sus ideas, consejos y total disponibilidad (incluyendo fines de semana y consultas telefónicas a cualquier hora) me facilitaron enormemente mi trayectoria investigadora y la realización de este trabajo. Del mismo modo, quiero agradecerle que me permitiera usar el Sistema de Procesamiento del Lenguaje Natural orientado a la resolución de la anáfora en español desarrollado por él; éste se ha adaptado convenientemente para el inglés y los módulos para ambos idiomas se han integrado en el sistema interlingua global. Además de su labor como director ha sido un excelente compañero y amigo que me ha alentado y ayudado en todo momento.

También estoy particularmente agradecido a Manuel Palomar, director de nuestro Grupo de Investigación, por sus consejos e ideas desde el inicio de mi etapa investigadora. Él y Antonio Ferrández confiaron en mí desde el principio, me apoyaron e impulsaron notablemente mi actividad investigadora.

Quiero agradecer al Dr. Ruslan Mitkov, director del Grupo de Lingüística Computacional de la Universidad de Wolverhampton,

y a todos los miembros de su Grupo de Investigación las sugerencias y comentarios que han realizado durante la realización de este trabajo. Del mismo modo, darles las gracias ya que nos proporcionaron una serie de corpus formados por manuales técnicos en inglés que estaban etiquetados anafóricamente.

A todos los compañeros del Grupo de Procesamiento del Lenguaje y Sistemas de Información: Maxi Saiz, Patricio Martínez, Rafa Muñoz, Juan Carlos Trujillo, José Luis Vicedo, Armando Suárez, Andrés Montoyo, Jaime Gómez, Rafa Romero, Fernando Llopis, Sergio Luján, Paloma Moreda, Borja Navarro y Cristina Cachero. La mayoría han intervenido directamente en tareas de investigación. Todos ellos me han apoyado y me han ayudado siempre ante cualquier problema.

Al Departamento de Lenguajes y Sistemas Informáticos por su apoyo en mi actividad docente e investigadora.

A los compañeros del Grupo de Procesamiento del Lenguaje Natural de la Universidad Politécnica de Valencia: Lidia Moreno, Antonio Molina, Emilio Sanchís, Natividad Prieto, Encarna Segarra y Ferrán Pla. Con ellos hemos intercambiado frecuentemente impresiones, siendo éstas muy útiles, acerca de mi trabajo así como de distintos proyectos y trabajos que han complementado mi formación en otras áreas de investigación.

Por último, pero no por ello menos importante, quiero hacer un agradecimiento especial a mi familia y amigos. En primer lugar, a mi esposa María José por su comprensión y apoyo durante varios años que han culminado con la realización de esta Tesis. A mis padres, mis suegros, mis cuñados, mi hermana y mi sobrino que me han transmitido el tesón y la fuerza necesaria que me ayudaron en esta tarea. Y en general, quiero agradecer a toda mi familia y amigos que me han animado a finalizar este trabajo. Ellos siempre han sabido comprender que la infinidad de horas dedicadas a la Tesis han impedido, en ocasiones, que pudiera atenderles como se merecían.

Alicante, septiembre 2001 *Jesús Peral Cortés*

Índice General

Índice de Tablas

Índice de Figuras

1. Introducción

El lenguaje es el sistema de comunicación y expresión verbal propio de un pueblo o nación (o común a varios). Tanto en su forma escrita (sirve para transmitir el conocimiento de una generación a la siguiente durante largo tiempo) como en su forma hablada (sirve como vehículo de comunicación principal en el comportamiento cotidiano con los demás) es una de las características fundamentales del comportamiento humano.

Tal y como se define en el trabajo de Moreno *et al.* (1999), el Procesamiento del Lenguaje Natural (PLN) es una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre/máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales.

El estudio del lenguaje se lleva a cabo desde diversas disciplinas como la lingüística, la psicolingüística, la filosofía y la lingüística computacional. Nuestro interés se centra en ésta última, una disciplina que intenta reproducir la transmisión natural de la información mediante el modelado de la producción del hablante y la interpretación del oyente en un sistema de ordenador (Hausser, 1999). En esta disciplina se pueden distinguir dos tipos de aplicaciones: las basadas en el tratamiento masivo de información textual y las basadas en diálogos (Moreno *et al.*, 1999).

Las aplicaciones basadas en el tratamiento masivo de información textual llevan a cabo el procesamiento de texto escrito (libros, periódicos, revistas electrónicas, diccionarios, informes, mensajes, etc.). Entre éstas podemos destacar los trabajos en las áreas siguientes:

1. *Traducción Automática.* Los sistemas de Traducción Automática son sistemas capaces de traducir textos de un idioma origen a un idioma destino de una manera automática.
   Las características de la sociedad actual, en la que las relaciones internacionales entre los distintos países del mundo han crecido exponencialmente, tiene una necesidad creciente de producir cada vez más traducciones, lo más rápido posible y a más bajo coste. Esta necesidad ha impulsado notablemente la investigación en el campo de la traducción de textos entre dos o más idiomas de forma automática.
2. *Extracción de Información.* Los sistemas de Extracción de Información trabajan sobre textos que contienen información de forma no estructurada. La información relevante de los mismos se extrae de forma estructurada mediante el relleno de unas plantillas definidas previamente. La información que no se ha definido como relevante es ignorada por el sistema.
3. *Recuperación de Información.* Los sistemas de Recuperación de Información pueden considerarse como un paso previo a la Extracción de Información. Son sistemas que trabajan sobre una colección de textos con formato heterogéneo y proporcionan aquellos textos que se consideran relevantes según algún criterio definido previamente.
   En la actualidad, con el auge de Internet principalmente y debido a la gran cantidad de documentos disponibles, se hace cada vez más necesario el diseño de sistemas que permitan (lo más rápido posible) la extracción y recuperación de información a un requerimiento realizado por el usuario mediante la comprensión o interpretación semántica tanto del requerimiento como de los distintos textos.
4. *Búsqueda de Respuestas.* Los sistemas de Búsqueda de Respuestas son herramientas capaces de obtener respuestas concretas a necesidades de información muy precisas a partir del análisis de documentos escritos en lenguaje natural. Estos sistemas localizan y extraen las respuestas de aquellas zonas de los documentos de cuyo contenido es posible inferir la información requerida.

5. *Realización automática de resúmenes.* Sistemas capaces de realizar de una forma automática resúmenes que contienen la información relevante de los textos procesados.
6. *Producción automática de textos.* Estos sistemas permiten el mantenimiento automático de textos como páginas Web, manuales técnicos, etc. Normalmente extraen la información de bases de datos y con ésta generan el texto en lenguaje natural.
7. *Corrección automática de textos.* Los sistemas para la corrección automática de los textos se utilizan principalmente en los procesadores de texto. Inicialmente eran muy simples, pero éstos han ido evolucionando de modo que ofrecen cada vez más prestaciones, incluyendo la detección de: errores ortográficos, repetición de palabras, problemas de concordancia, construcciones sintácticamente incorrectas, etc.

Las aplicaciones basadas en diálogos tratan la comunicación hombre-máquina tanto oral como escrita. Se pueden clasificar en tres grandes áreas de trabajo:

1. *Sistemas de acceso a Bases de Datos.* Sistemas de pregunta/respuesta locales en los que el usuario hace una pregunta, expresada en lenguaje natural, solicitando información que se encuentra almacenada en una base de datos. El objetivo de estos sistemas es la compilación de las expresiones producidas en lenguaje natural a una forma directamente interpretable por un sistema de gestión de base de datos a través de su lenguaje de consulta propio.
   Estos sistemas pretenden superar las barreras que pueden suponer la realización de una consulta a una base de datos por parte de un usuario que no conoce el lenguaje formal y artificial necesario para realizar las preguntas a la base de datos.
2. *Sistemas de acceso a otros dominios.* Estos sistemas actúan sobre dominios de información heterogéneos entre los que destacan los interfaces para sistemas expertos y los accesos a sistemas operativos.
3. *Sistemas inteligentes de diálogo.* Estos sistemas examinan el aspecto multimodal de la comunicación que se establece entre el hombre y la máquina, en toda su globalidad. En ellos, el

usuario conversa con el sistema utilizando el lenguaje natural, y el sistema genera respuestas en lenguaje natural que se construyen tras la formalización de aspectos tales como las intenciones y deseos del usuario, el conocimiento y las creencias acerca de ese conocimiento, y la relación entre el conocimiento y la acción.

En esta Tesis nos centraremos en las aplicaciones basadas en el tratamiento masivo de información textual, particularmente en la Traducción Automática. Dentro de este área trataremos el problema de la resolución y generación de la anáfora pronominal en el idioma origen y destino respectivamente a partir de textos no restringidos sobre cualquier dominio.

1.1 El problema de la Traducción Automática

La Traducción Automática (TA) ha sido un tema de estudio y discusión desde hace varios siglos. Ya en el siglo XVII, Descartes y Leibniz se plantearon la posibilidad de crear diccionarios mecánicos para facilitar la traducción entre idiomas distintos. Con la aparición de los ordenadores (primera mitad del siglo XX) nacieron los primeros intentos reales para automatizar el proceso de la traducción. Desde entonces, ha sido tema de amplio debate y ha atraído la atención de especialistas en campos diversos como Lingüística, Filosofía, Informática, Matemáticas, etc.

No hay duda que la traducción (de un idioma a otro) es una de las proezas más grandes del ser humano. Se puede comparar a la creación de una obra literaria original. Para capturarla en una máquina sería necesario capturar alguna parte especial del espíritu humano, es decir, debe ser capaz de comprender sus intimidades. Pero precisamente debido a que hay demasiado de "espíritu humano" en la traducción, en algunas ocasiones podríamos rechazar las traducciones que nos ofrece la máquina porque no expresan el sentido real del texto original. No hay nada que una persona pueda saber, sentir o soñar que no sea crucial para obtener una buena traducción de un texto a otro. Para ser un traductor, por

lo tanto, uno no puede tener sólo algunas partes de humanidad, uno debe ser un ser humano completo.

Las motivaciones para la investigación en TA han sido muy diferentes. Cada vez más, las relaciones comerciales internacionales (entre los distintos países del mundo) han ido aumentando, lo que implica que las cartas, contratos, manuales, etc. se deben producir en idiomas diferentes. La tecnología moderna evoluciona cada vez más rápido, por lo tanto, los textos que la acompañan deben ser reemplazados más pronto. Más aún, la tecnología avanzada (coches, ordenadores, etc.) debe estar al alcance de todo el mundo, por esta razón los manuales para operar y mantener estos productos no pueden ser producidos exclusivamente en un único idioma. Los países de la Unión Europea, países bilingües (Canadá), países en desarrollo, etc. necesitan que los documentos oficiales se traduzcan en varios idiomas. En definitiva, hay una necesidad creciente de producir más traducciones, de producirlas más rápidas y de producirlas a más bajo coste.

En los últimos años, estas necesidades de obtener más traducciones de las que son capaces de realizar los traductores humanos han llevado a un gran incremento en la actividad en este campo. En términos reales, la TA se aplica sobre dos clases de materiales. En la primera clase se incluyen los textos que cubren un tema muy específico para los cuales se puede realizar una traducción que requieren una parte pequeña de comprensión. Como ejemplo de esta clase de textos se pueden citar los manuales de instrucciones de las máquinas. La segunda clase está formada por material que será leído por personas que buscan una idea aproximada de lo que se está diciendo en el texto, por lo tanto, simplemente con una traducción aproximada será suficiente. Un ejemplo de esta clase puede ser la búsqueda en un texto de los clientes de una compañía competidora a una compañía dada.

Sin embargo, la TA, incluso cuando se aplica sobre textos de temas específicos, todavía no produce (generalmente) resultados aceptables para un traductor humano bajo ninguna circunstancia. El futuro de la TA radica en la incorporación de nuevos métodos o estrategias a los sistemas actuales que permitan mejorar el ren-

dimiento de éstos para obtener traducciones lo más parecidas a las realizadas por el ser humano.

1.2 El problema de la anáfora y su relación con la TA

De todos los problemas que hay que tratar en el PLN, la anáfora se puede considerar como uno de los problemas más difíciles de resolver (Moreno *et al.*, 1999). La etimología de la palabra "anáfora" se remonta a la Grecia Clásica con la palabra compuesta "$\alpha\nu\alpha\varphi o\rho\alpha$" formada por las palabras $\alpha\nu\alpha$ (hacia atrás) y $\varphi o\rho\alpha$ (el acto de llevar) y significa el acto de llevar hacia atrás. En la literatura se han realizado varias definiciones del término anáfora, aunque en todas ellas subyace el mismo contenido. Halliday & Hassan (1976) definieron la anáfora como "*la cohesión que apunta a una entidad*[1] *previa*". Una definición más formal la realiza Hirst que la define como "*el mecanismo que permite hacer en un discurso una referencia abreviada a alguna entidad o entidades, con la confianza de que el receptor del discurso sea capaz de interpretar la referencia y por consiguiente determinar la entidad a la que se alude*" (Hirst, 1981). De esta definición se extraen los componentes básicos del proceso anafórico: la referencia abreviada a la que se denomina *expresión* o *elemento anafórico* y la entidad referenciada que se denomina *referente* o *antecedente*.

(1) *[María]*$_i$ fue al cine el miércoles. La película no *le*$_i$ gustó nada.

En el ejemplo 1 aparece una anáfora en la que se pueden distinguir estos dos componentes: la expresión anafórica (el pronombre *le*) y el antecedente (el sintagma nominal *María*)[2].

[1] Por *entidad* se entiende cualquier objeto o persona que aparece implícita o explícitamente en el proceso comunicativo.

[2] En esta Tesis, en los ejemplos de expresiones anafóricas, se le asociará a la anáfora un subíndice, al igual que a su antecedente (irá marcado entre corchetes "[" y "]"). La coincidencia de ambos subíndices indica que están relacionados (en caso contrario no hay relación). La anáfora y su antecedente se escribirán en letra cursiva.

Siguiendo con la presentación de Moreno *et al.* (1999), el problema de la anáfora debe tratarse como dos procesos distintos: la *resolución* y la *generación* de la anáfora. La resolución busca la entidad a la que hace referencia la anáfora, mientras que la generación crea referencias sobre una entidad del discurso. En esta Tesis estudiaremos ambos procesos en el área de la Traducción Automática.

Normalmente las estrategias para la resolución de la anáfora se han clasificado en dos grandes grupos: sistemas *integrados* y sistemas *alternativos* (Mitkov & Schmidt, 1998; Ferrández, 1998).

Los sistemas *integrados* se basan en el conocimiento, es decir, hacen uso de una serie de conocimientos (fuentes de información) que se suponen necesarios para resolver la anáfora. Los sistemas *alternativos*, a diferencia de los anteriores, no están basados en el conocimiento y utilizan una serie de técnicas y recursos distintos a los tradicionales (técnicas estadísticas, redes neurales, estudio de corpus etiquetados, etc.).

En esta Tesis se ha utilizado una estrategia integrada desarrollada mediante un mecanismo de restricciones y preferencias.

Las distintas fuentes de información utilizadas generalmente en la resolución de la anáfora se pueden clasificar en:

- Fuentes de información léxica. Como información léxica se podría incluir aquélla relativa al comportamiento de ciertas palabras o grupos de palabras en situaciones concretas. Por ejemplo, se pueden establecer un conjunto de preferencias para una serie de verbos concretos (Mitkov & Stys, 1997).
- Fuentes de información morfológica. La información morfológica hace referencia a la necesaria concordancia en número, género y persona entre la expresión anafórica y su antecedente.
- Fuentes de información sintáctica. La información sintáctica se refiere a la información que subyace en la estructura sintáctica del texto. A partir de esta información se pueden establecer una serie de reglas que permiten aceptar o rechazar antecedentes de ciertas expresiones anafóricas. Por ejemplo, el paralelismo sintáctico permite expresar la compatibilidad entre anáforas y antecedentes.

- Fuentes de información semántica. La información semántica pretende captar las distintas relaciones semánticas que existen entre los constituyentes de una oración. Por ejemplo, los papeles temáticos (agente, experimentante, benefactivo, etc.) expresan las relaciones entre los complementos (argumentos del verbo) y el verbo de la oración.
  Este tipo de información obliga a que exista una cierta compatibilidad entre los rasgos semánticos de la expresión anafórica y su antecedente.
- Otras fuentes de información. Entre estas fuentes de información destacamos la información pragmática que utiliza conocimiento del mundo y la posibilidad de inferir nuevo conocimiento para aceptar o descartar antecedentes bajo ciertas circunstancias.

En el contexto de la TA, la resolución de las expresiones anafóricas, es decir, el proceso en el que se busca y determina los antecedentes de las anáforas, es de una importancia vital para realizar una correcta traducción. Analicemos las tres oraciones presentadas en el siguiente ejemplo (Hutchins & Somers, 1992):

(2) **a** *[The monkey]*$_i$ ate the banana because *it*$_i$ was hungry.

b The monkey ate *[the banana]*$_i$ because *it*$_i$ was ripe.

c The monkey ate the banana because *it* was tea-time.

En cada una de las tres oraciones del ejemplo 2 el pronombre *it* se refiere a cosas distintas: en (a) *the monkey* –*el mono*–, en (b) *the banana* –*la banana*– y en (c) la noción abstracta de tiempo. Si traducimos estas oraciones a español o alemán (idiomas que marcan el género de los pronombres), la resolución de la anáfora es inevitable ya que el pronombre toma el género de su antecedente. Así, en español se traduciría por los siguientes pronombres: (a) *éste* (masculino, antecedente *el mono*), (b) *ésta* (femenino, antecedente *la banana*) y (c) pronombre omitido (oración impersonal). Para el caso del alemán, se generarían los siguientes pronombres:

(a) *er* (antecedente masculino), (b) *sie* (antecedente femenino) y (c) *es* (neutral).

Además de estas diferencias entre idiomas provocadas por las discrepancias de género, existen otras que influyen en el proceso de la traducción de las anáforas. Por ejemplo, las discrepancias de número ocurren con aquellas palabras que son referidas por un pronombre singular en un idioma y por un pronombre plural en otro. Por ejemplo, la palabra *policía* es plural en inglés, mientras que en alemán es singular. En la traducción inglés-alemán se traducirá un pronombre singular por un pronombre plural.

Por otra parte, aunque en la mayoría de las traducciones entre dos idiomas las expresiones anafóricas pronominales de un idioma se traducen y se generan pronominalmente en el otro idioma, hay una serie de excepciones. En algunos idiomas, los pronombres se traducen directamente por su antecedente. Por ejemplo, en la traducción inglés-malayo, hay una tendencia a reemplazar el pronombre *it* por su antecedente, por lo que el traductor debe identificar el antecedente previamente.

Otra excepción ocurre cuando los pronombres se omiten simplemente en el idioma hacia el que se está realizando la traducción. Por ejemplo, en la traducción inglés-español, aunque los pronombres ingleses tienen sus pronombres españoles correspondientes, normalmente se omiten en español cuando realizan la función de sujeto de la oración.

Algunos idiomas permiten que los elementos anafóricos se puedan traducir por distintas expresiones dependiendo de la información sintáctica y semántica del antecedente. Por ejemplo, en la traducción inglés-coreano, el pronombre inglés se puede omitir, traducir por un sintagma nominal definido, por su antecedente o por un pronombre cuando se traduce al coreano.

Todos estos problemas de la traducción de las expresiones anafóricas de un idioma a otro indican que es necesario realizar un tratamiento detallado de estas expresiones (incluyendo su resolución y determinando su antecedente) para su correcta generación en el idioma hacia el que se realiza la traducción.

Normalmente, los sistemas de TA no abordan el proceso de resolución de las anáforas en el idioma origen (como se verá en

el capítulo 3). En esta Tesis, planteamos una forma novedosa de afrontar este problema en el contexto de la TA cuyo tratamiento es necesario para la correcta generación de las expresiones anafóricas en el idioma destino.

1.3 Organización de la Tesis

El presente trabajo se ha estructurado en ocho capítulos:

En este primer capítulo se ha presentado una introducción al problema del Procesamiento del Lenguaje Natural (PLN) y, concretamente, al problema del tratamiento de las expresiones anafóricas y su relación con la Traducción Automática (TA).

En el segundo capítulo se realiza un recorrido por la historia y evolución de la TA desde el siglo XVII hasta nuestros días. Además se presenta una clasificación de las estrategias básicas que pueden seguir los sistemas de TA según tres criterios distintos (el número de idiomas, las fuentes de información y la existencia de representaciones intermedias). Por último, se presenta la arquitectura de los sistemas de TA describiendo los distintos módulos implicados en el proceso de la TA: módulo de análisis (en el que se detallan todos los tipos de ambigüedades que se tienen que abordar), módulo de transferencia (en el que se especifican los distintos niveles de transferencia que puede utilizar un sistema de TA) y módulo de generación (en el que se trata la generación en el idioma destino según la estrategia utilizada).

En el capítulo tercero se describen los sistemas de TA más relevantes que siguen las tres estrategias básicas: sistemas directos, sistemas de transferencia o sistemas interlingua. El capítulo finaliza con un resumen de las características de los mismos con el que se demuestra las deficiencias que tienen los sistemas de TA para el tratamiento y generación de las anáforas pronominales en el idioma destino.

En el cuarto capítulo se lleva a cabo un estudio profundo del fenómeno lingüístico de la anáfora. Se realiza una clasificación de las expresiones anafóricas en función de varios criterios y se presenta una revisión bibliográfica de las distintas estrategias utiliza-

das para resolver el problema de las relaciones anafóricas. Tras estudiar el problema de la anáfora en general, se aborda el problema de las expresiones anafóricas en el contexto de la TA. Por último, se presenta una revisión de diferentes estrategias específicas que tratan el problema de la anáfora en sistemas de TA.

El capítulo quinto presenta la arquitectura general del sistema interlingua AGIR que realiza la resolución y generación de la anáfora pronominal en el idioma destino. Se describen de forma independiente los módulos de análisis y generación. En el primero de ellos se realiza una descripción exhaustiva de la representación interlingua del texto. En el segundo, se explican detalladamente las discrepancias sintácticas y morfológicas ocasionadas por las diferencias entre español e inglés en el tratamiento de los pronombres. El tratamiento adecuado de estas discrepancias permitirá la correcta generación de los mismos en el idioma destino.

En el capítulo sexto se presenta la implementación realizada de los distintos módulos del sistema AGIR. En él, se explicarán con detalle las distintas etapas que se llevan a cabo en los módulos de análisis y generación del sistema.

El capítulo séptimo incluye la evaluación global del sistema AGIR. Para ello se muestra una evaluación independiente de las distintas tareas llevadas a cabo en el sistema y que conducen a la tarea final: la generación de la anáfora pronominal en el idioma destino. Las tareas evaluadas fueron las siguientes: detección de los pronombres no referenciales, detección de los cero pronombres, resolución de la anáfora pronominal, resolución de los cero pronombres y, por último, generación de la anáfora pronominal. Para cada una de ellas se presenta la fase de entrenamiento y la fase de evaluación, independientes entre sí.

En el capítulo octavo se recogen las conclusiones obtenidas con el desarrollo de este trabajo, así como diferentes líneas de investigación para el futuro. El capítulo finaliza con la presentación de todas las publicaciones científicas realizadas desde el inicio de la actividad investigadora.

Por último, se muestran las referencias bibliográficas utilizadas en la realización de este trabajo.

2. La Traducción Automática (TA). Arquitectura de un sistema de TA

En este capítulo se introducirán los conceptos de traducción y Traducción Automática (TA). Tras definir la TA, se realizará un recorrido por la historia de la misma y se realizarán distintas clasificaciones de las estrategias básicas que siguen los sistemas de TA en función de varios criterios. A continuación se presentará la arquitectura de los sistemas de TA describiendo los módulos implicados en el proceso de la TA: módulo de análisis, módulo de transferencia y módulo de generación.

2.1 Traducción y Traducción Automática (TA)

Según el Diccionario de la Real Academia Española[1], la traducción es la "*acción y efecto de traducir*", es decir, "*acción y efecto de expresar en una lengua lo que está escrito o se ha expresado antes en otra*". Por lo tanto, el término traducción es, en sí mismo, ambiguo ya que denota tanto la acción de traducir como el resultado (efecto) de este proceso.

Según Jakobson (1985) se distinguen 3 maneras de interpretar un signo verbal: (1) traducirlo a otros signos de la misma lengua, (2) a otra lengua, o (3) a cualquier otro sistema no verbal de símbolos. Estos tres tipos de traducción pueden designarse de modo diferente:

1. La traducción intralingüística o reformulación (*rewording*) es una interpretación de los signos verbales mediante otros signos de la misma lengua.

[1] Edición electrónica, Espasa Calpe, S.A., 1995.

2. La traducción interlingüística o traducción propiamente dicha (*translation proper*) es una interpretación de los signos verbales mediante cualquier otra lengua.
3. La traducción intersemiótica o transmutación (*transmutation*) es una interpretación de los signos verbales mediante los signos de un sistema no verbal.

En la traducción intralingüística de una palabra se emplea otra palabra más o menos sinónima o se recurre al circunloquio[2]. Sin embargo, por regla general, el sinónimo no suele dar una equivalencia completa: por ejemplo, "todo célibe es soltero, pero no todo soltero es célibe". Una palabra o una expresión idiomática, en suma, sólo se puede interpretar plenamente mediante un mensaje (combinación de expresiones) que se refiera a esta expresión: "todo soltero es una persona que no ha contraído matrimonio y toda persona que no ha contraído matrimonio es soltera".

De igual modo, a nivel de la traducción interlingüística no hay normalmente una equivalencia entre las expresiones, aunque los mensajes puedan servir de interpretaciones correctas de mensajes pertenecientes a otras lenguas. Sin embargo, lo más frecuente es que en la traducción de una lengua a otra se sustituyan mensajes, no por expresiones por separado sino por mensajes enteros, a su vez, en la otra lengua. Tal traducción equivale a un estilo indirecto; el traductor recodifica y transmite un mensaje recibido de otra fuente. Una traducción semejante requiere dos mensajes equivalentes en dos códigos diferentes.

La falta de algún recurso gramatical en la lengua a la cual se traduce no imposibilita la traducción literal de la totalidad de la información contenida en el original. Si en un determinado lenguaje falta alguna categoría gramatical, su significado puede traducirse a este lenguaje por medios léxicos.

Para traducir correctamente la frase inglesa *I hired a worker* (*Yo contraté un trabajador*), un ruso necesita además que se le diga si esta acción se completó o no y si el *trabajador* (*worker*)

[2] El circunloquio se define según el Diccionario de la Real Academia Española como "*rodeo de palabras para dar a entender algo que hubiera podido expresarse más brevemente*".

era hombre o mujer, porque debe elegir entre un verbo de aspecto completivo o uno no completivo –*nanjal* o *naminal*– y entre un sustantivo masculino y uno femenino –*rabotnika* o *rabotnitsu*–. Si le preguntamos al hablante inglés si el trabajador era hombre o mujer, la pregunta podrá parecer impertinente o indiscreta, mientras que en la versión rusa de esta frase la respuesta a este interrogante es obligatoria. El hecho de que los artistas alemanes pintaran al pecado en forma de mujer sorprendió al pintor ruso Repin porque desconocía que *pecado* es femenino en alemán (*die Sünde*), pero masculino en ruso (*grjech*). El título de un libro de poemas de Boris Pasternak, *Mi hermana la vida*, no presenta ninguna dificultad en ruso, lengua en que la palabra vida es femenina (*žizn'*), pero presentó muchas dificultades al poeta checo Josef Hora, en su intento de traducir el libro, ya que en checo *vida* es masculino (*život*).

Con estos ejemplos queda patente que la traducción interlingüística o traducción propiamente dicha no es una tarea sencilla y requiere un amplio dominio de las dos lenguas entre las que se quiere realizar la traducción. Si esta labor es, a veces, difícil incluso para un humano, podemos imaginar lo que supone realizar esta tarea de un modo automático. Esta idea de realizar el proceso de la traducción interlingüística entre dos lenguas de un modo automático dio lugar al nacimiento del concepto Traducción Automática.

La Traducción Automática (TA) se puede definir[3] como "*el proceso (o el resultado) de traducir un texto en un idioma origen a un idioma destino de una manera automática*". En esta definición hay que matizar dos conceptos: *texto* y *automática*. *Texto* se refiere a un texto informatizado (fichero de ordenador que contiene un texto codificado en un formato determinado) y *automática* se refiere al uso de un programa de ordenador.

[3] En esta Tesis se utilizarán los términos: *idioma*, *lengua* y *lenguaje* como palabras sinónimas, según aparecen en el Diccionario de la Real Academia Española con el significado "*sistema de comunicación y expresión verbal propio de un pueblo o nación, o común a varios*". Del mismo modo, se utilizará el término *idioma origen* (también llamado *idioma fuente* en la terminología de la TA) para denominar el idioma en el que están escritos los textos que van a ser traducidos e *idioma destino* (también llamado *idioma meta* o *idioma objetivo*) para denominar el idioma al que se va a traducir el texto original.

Otra definición de TA es la que aparece en el trabajo de Arnold (1994): "*intento de automatizar todo, o parte del proceso de traducir de un idioma humano a otro*". En esta definición hay que destacar dos aspectos. En primer lugar, la TA implica el tratamiento de sistemas de comunicación usados por los *humanos*. Por otra parte, en la definición se indica que el proceso de la TA puede ser *total* o *parcialmente automático*. Normalmente se utiliza el término *traducción automática* para aquella traducción que no necesita intervención humana, es decir, es totalmente automática. Se utiliza el término *traducción asistida por ordenador* (o *traducción semi-automática*) para aquellas traducciones que requieren la intervención humana.

Tal y como se presenta en Hutchins & Somers (1992) la mecanización de la traducción ha sido uno de los sueños más viejos de la humanidad. En el siglo XX este sueño se hizo realidad en forma de programas de ordenadores capaces de traducir una amplia variedad de textos de un idioma a otro. Sin embargo, como siempre, la realidad no es perfecta. No existen "máquinas traductoras" que tomen un texto en un idioma y produzcan una traducción perfecta en otro idioma sin la intervención o asistencia humana. Éste es un ideal para el futuro, que si bien es factible en principio, requerirá un esfuerzo considerable para llegar a alcanzarlo.

Por el momento lo que se ha obtenido es el desarrollo de programas que pueden producir traducciones "sin refinar" de textos de dominio bien definido. Estas traducciones pueden ser revisadas posteriormente por especialistas para proporcionar una mejor calidad a los textos traducidos. En algunos casos, controlando el lenguaje de los textos de entrada, se pueden obtener automáticamente traducciones de alta calidad sin la necesidad de revisiones posteriores.

Sin embargo, estos sólidos logros alcanzados por la TA son a menudo mal comprendidos. La percepción pública de la TA es distorsionada por dos extremos. Por una parte, aquéllos que están convencidos de la facilidad que supone analizar el lenguaje, ya que incluso los niños pequeños son capaces de aprender idiomas fácilmente. Además, están convencidos de que cualquiera que conozca un idioma extranjero debe ser capaz de traducir con facili-

dad. Estas personas son incapaces de apreciar las dificultades que conllevan un proceso de TA y de los logros que se han obtenido. Por otra parte, están aquéllos que creen que la traducción automática de autores literarios como Cervantes, Shakespeare, etc. no es factible y por lo tanto cualquier sistema de traducción basado en un ordenador no tiene sentido. Son incapaces de evaluar la contribución que estas traducciones "imperfectas" podrían hacer en su propio trabajo o en la mejora general de la comunicación internacional.

2.2 Objetivos de la TA

La mayoría de las traducciones que se realizan rutinariamente corresponden a textos que no tienen un alto nivel literario o cultural. La gran mayoría de traductores profesionales trabajan para satisfacer la demanda, cada vez más creciente, de traducciones de documentos científicos y técnicos, transacciones comerciales y de negocios, documentación legal, manuales de instrucciones, libros de texto de medicina y agricultura, patentes industriales, reportajes de periódicos, etc. La demanda para este tipo de traducciones está creciendo de un modo que supera la capacidad de los profesionales de la traducción. Por esta razón, la ayuda de un ordenador presenta una serie de atractivos interesantes.

La utilidad práctica de un sistema de TA está determinada por la calidad de su salida. Aunque actualmente es prácticamente imposible que un ordenador realice una traducción perfecta a partir de un texto de entrada, el ideal de igualar la mejor traducción humana permanece. La TA es parte de una amplia esfera de investigación "pura" en Lingüística Computacional e Inteligencia Artificial que exploran los mecanismos básicos del idioma y la mente mediante su modelización y simulación en programas de ordenador.

Sin embargo, los mayores obstáculos para la traducción, usando un ordenador, no son computacionales sino lingüísticos. Estos problemas lingüísticos son la ambigüedad léxica, la ambigüedad contextual, la complejidad sintáctica, las construcciones elípticas,

las referencias, etc. En definitiva, el problema de la traducción se reduce a la extracción del "significado" de las oraciones y textos a partir del análisis de signos escritos y producir oraciones y textos en otro conjunto de símbolos lingüísticos con un significado equivalente.

2.3 Historia de la TA

Los orígenes de la TA y su evolución hasta nuestros días se presentan en varios trabajos (Hutchins, 1986; Slocum, 1988; Hutchins & Somers, 1992; Arnold *et al.*, 1994; Mitkov, 2001). A continuación se realiza una pequeña revisión de la historia de la TA.

2.3.1 (1600–1930) Aproximaciones de TA basadas en el "idioma universal"

Se puede considerar que Descartes y Leibniz, en el siglo XVII, fueron los primeros autores en abordar el problema de la TA. Ellos sugirieron el uso de diccionarios mecánicos para superar las barreras del idioma. Especularon con la idea de crear diccionarios basados en códigos numéricos internacionales. A mitad de siglo, Cave Beck, Athanasius Kircher y Johann Becher pusieron en práctica dichas ideas. La inspiración fue el movimiento del "idioma universal", la idea de crear un lenguaje basado en principios lógicos y en símbolos icónicos en el cual toda la humanidad se pudiera comunicar sin problemas. El más conocido fue el interlingua elaborado por John Wilkins en 1668 presentado en su libro "Essay towards a Real Character and a Philosophical Language". Durante los siglos posteriores se desarrollaron muchas propuestas para idiomas universales, siendo el Esperanto el más conocido.

2.3.2 (1931–1963) Primeros sistemas reales de TA

Los primeros intentos para mecanizar la traducción aparecieron en la mitad del siglo XX. En 1933 aparecieron independientemente dos patentes, una en Francia y otra en Rusia:

1. El francés George Artsouni diseñó, en 1937, un prototipo en el que se usaba una cinta de papel como dispositivo de almacenamiento que permitía encontrar el equivalente de cualquier palabra en otro idioma.
2. En su aproximación, el ruso Petr Smirnov-Troyanskii concibió tres etapas en la traducción mecánica:
   a) Un editor que conocía sólo el idioma origen llevaba a cabo el análisis "lógico" de las palabras en sus formas base y funciones sintácticas.
   b) Una máquina transformaba la secuencia de formas base y funciones en las secuencias equivalentes en el idioma destino.
   c) Otro editor que conocía el idioma destino convertía esta salida en las formas normales de ese idioma.

Aunque la patente de Troyanskii sólo se refería a la máquina que llevaba a cabo la segunda etapa, él creyó que "el proceso del análisis lógico podría mecanizarse". Troyanskii fue un avanzado de su época, aunque sus inventos no fueron conocidos fuera de Rusia.

Pocos años más tarde, Warren Weaver de la Fundación Rockefeller y Andrew D. Booth, un especialista en cristalografía, fueron los primeros en plantearse la posibilidad de usar los ordenadores para la traducción. Booth exploró la mecanización de un diccionario bilingüe y colaboró con Richard H. Richens que había usado tarjetas perforadas para producir traducciones palabra a palabra de abstracts (resúmenes) científicos. Sin embargo, fue Weaver en 1949 cuando presentó en un trabajo la idea de la TA y sugirió varios métodos: el uso de las técnicas de criptografía usadas en época de guerra, análisis estadístico, exploración de las características universales del lenguaje, etc. (Locke & Booth, 1955).

Años más tarde, la investigación había empezado en numerosos centros norteamericanos, y en 1951 fue nombrado el primer investigador de TA con dedicación exclusiva: Yehoshua Bar-Hillel en el Instituto de Tecnología de Massachussets (Massachussets Institute of Technology, MIT). Un año más tarde él convocó el primer congreso de TA en el que se establecieron las líneas maes-

tras de la investigación futura. Se presentaron propuestas para tratar la sintaxis, sugerencias de escribir los textos en lenguajes controlados, argumentos para la construcción de sistemas basados en sublenguajes y el reconocimiento de la necesidad de la ayuda humana hasta que se alcance la traducción automática completa. Sin embargo, se necesitaba demostrar la viabilidad de un sistema real de TA.

En enero de 1954 se realizó la primera demostración pública de un sistema de TA. Este sistema fue desarrollado en un proyecto llevado a cabo entre la Universidad de Georgetown e IBM. En la demostración, se tradujeron 49 oraciones rusas al inglés, usando un vocabulario muy restringido de 250 palabras y sólo 6 reglas gramaticales. Aunque el experimento tenía un pequeño valor científico, sirvió para estimular la investigación en Estados Unidos y para inspirar el comienzo de nuevos proyectos de TA en todo el mundo, especialmente en la Unión Soviética.

Durante la siguiente década (1954–1964) hubieron muchos grupos activos dedicados a la investigación en TA. Básicamente se adoptaron dos aproximaciones:

1. Aproximaciones empíricas. Usaban estrategias empíricas de *prueba* y *error* y generalmente estaban basadas en técnicas estadísticas. Normalmente se les denominaba métodos basados en la "fuerza bruta".
   Ejemplos de estas aproximaciones son: la aproximación lexicográfica de la Universidad de Washington (Seattle), posteriormente continuada por IBM en un sistema ruso-inglés usado por el Ejército del Aire norteamericano, la aproximación estadística de RAND Corporation y los métodos adoptados en el Instituto de la Mecánica en la Unión Soviética y el Laboratorio Físico Nacional de Gran Bretaña. El más grande de todos ellos fue el grupo de la Universidad de Georgetown, cuyo sistema ruso-inglés es conocido como el típico de la "primera generación" en investigación en TA –Systran (Toma, 1977)–.
2. Aproximaciones teóricas. Suponían investigaciones lingüísticas y buscaban el desarrollo de un modelo genérico de traducción. Comúnmente se les denominaba métodos "perfeccionistas".

Entre los centros de las aproximaciones teóricas estaban: MIT, la Universidad de Harvard, la Universidad de Texas, la Universidad de California en Berkeley, el Instituto de Lingüística en Moscú y la Universidad de Leningrado, la Unidad de Investigación del Lenguaje de Cambridge (Cambridge Language Research Unit, CLRU) y las Universidades de Milán y Grenoble. A diferencia de las aproximaciones empíricas donde la "traducción directa" era la norma, algunas de las aproximaciones teóricas experimentaron con las primeras versiones de sistemas interlingua y de transferencia (por ejemplo, CLRU y MIT respectivamente).

Muchas de las investigaciones desarrolladas durante este período fueron importantes no sólo para la TA sino también para la Lingüística Computacional e Inteligencia Artificial, en particular, el desarrollo de diccionarios automáticos y de técnicas de análisis sintáctico. Sin embargo, el objetivo básico de construir sistemas capaces de producir buenas traducciones no se alcanzó. El optimismo había sido elevado, existían predicciones de avances inminentes, pero el pesimismo iba creciendo a medida que la complejidad de los problemas lingüísticos se hacía cada vez más aparente. En 1960, en una revisión del progreso de la TA, Bar-Hillel criticó la suposición de que el objetivo de la investigación en TA debía ser la creación de sistemas de traducción automáticos de alta calidad que produjeran resultados similares a los de los traductores humanos. Él proponía que las "barreras semánticas" de la TA sólo podrían ser superadas, en principio, con la inclusión de una gran cantidad de conocimiento enciclopédico acerca del "mundo real". Su recomendación fue que la TA debía de adoptar metas menos ambiciosas, es decir, la construcción de sistemas que rentabilicen la interacción hombre-máquina.

2.3.3 (1964–1979) Repercusiones del informe ALPAC. TA en Estados Unidos y Europa

En 1964 los patrocinadores gubernamentales de la TA en Estados Unidos formaron el Comité Consultivo del Procesamiento del Lenguaje Automático (Automatic Language Processing Advisory

Committee, ALPAC) para examinar las perspectivas de la TA. En su informe de 1966 (ALPAC, 1966) concluyó que la TA era más lenta, menos precisa y dos veces más cara que la traducción humana y afirmó que "no hay perspectiva inmediata o predecible de una TA útil". Determinó que no hay necesidad de continuar investigando en TA, en vez de esto, recomendó el desarrollo de ayudas para los traductores, tales como diccionarios automáticos, etc. La influencia del informe ALPAC fue profunda provocando el fin virtual de la investigación en TA en Estados Unidos durante una década y dañando la percepción pública de la TA durante muchos años después.

Durante la siguiente década la investigación en TA se desarrolló fuera de los Estados Unidos, en Canadá y en Europa occidental, y fue virtualmente ignorada por la comunidad científica. En América la actividad se había concentrado en las traducciones a inglés de textos rusos científicos y técnicos y en algunos proyectos subvencionados por la Iglesia mormona para la traducción de la Biblia[4]. En Canadá y Europa las necesidades eran bastante diferentes: la política bicultural canadiense creó una demanda para las traducciones inglés-francés-inglés que superaba la capacidad del mercado, mientras que la Comunidad Económica Europea (CEE) solicitaba traducciones de documentación científica, técnica, administrativa y legal desde y para todos los idiomas de la Comunidad.

En 1976, a un grupo de investigación de Montreal se le asignó la tarea de la creación del sistema Météo (Chandioux, 1976) basado en un "sublenguaje". Este sistema traducía los partes meteorológicos diarios de la previsión del tiempo.

En el mismo año, la Comisión de las Comunidades Europeas decidió instalar un sistema inglés-francés llamado Systran (Toma, 1977) que había sido desarrollado previamente por Peter Toma (antiguo miembro del equipo de Georgetown) para la traducción ruso-inglés y que fue utilizado por el Ejército del Aire norteamericano. Este sistema había estado funcionando desde 1970. Durante

[4] La investigación principal se llevó a cabo en la Universidad de Brigham Young (Provo, Utah), cuyos trabajos condujeron al desarrollo de los primeros sistemas comerciales Weidner y ALPSystems.

los años siguientes, la Comisión desarrolló sistemas para francés-inglés, inglés-italiano, inglés-alemán, etc. A finales de los 70, decidió subvencionar un ambicioso proyecto de investigación para desarrollar un sistema multilingüe para todos los idiomas de la Comunidad, basado en los últimos avances en TA y Lingüística Computacional. Éste es el proyecto Eurotra y en él trabajan grupos de investigación de todos los estados miembro (Allegranza *et al.*, 1991).

Para su diseño básico, Eurotra debe mucho a la investigación realizada en Grenoble y en Saarbrücken. Durante los 60, el grupo francés construyó un sistema "interlingua" para la traducción ruso-francés. Sin embargo, debido a la complejidad de obtener una representación interlingua "real", los resultados no fueron satisfactorios y en los 70 comenzaron a desarrollar el sistema de "transferencia" Ariane (Boitet & Nédobejkine, 1981). Desde finales de la década de los 60, el grupo de Saarbrücken también había estado construyendo un sistema multilingüe de transferencia, el sistema SUSY (Maas, 1977). Es por esta época, cuando hay consenso entre la comunidad investigadora en TA y se opina que las mejores perspectivas para los avances significativos en TA radican en el desarrollo de sistemas de transferencia. Los investigadores del Centro de Investigación de Lingüística (Linguistics Research Center, LRC) de Austin, Texas habían llegado a conclusiones similares después de experimentar con un sistema interlingua y estaban desarrollando ahora su sistema de transferencia METAL (Bennet & Slocum, 1985). También en Japón el trabajo había empezado en la Universidad de Kyoto con el sistema de transferencia Mu para la traducción japonés-inglés. El grupo Eurotra adoptó la misma aproximación básica.

2.3.4 (1980–2001) Aparición de los sistemas interlingua y sistemas comerciales

En la década de los 80, el diseño de los sistemas de transferencia se unió mediante nuevas aproximaciones a la idea interlingua. La más importante es la investigación sobre sistemas basados en el conocimiento, especialmente en la Universidad de Carne-

gie Mellon –Carnegie Mellon University (CMU, Pittsburg)–, que están basados en sistemas de comprensión del lenguaje natural enmarcados dentro del área de la Inteligencia Artificial. La idea central es que la TA debe ir más allá de la información puramente lingüística (sintaxis y semántica); la traducción supone la "comprensión" del contenido de los textos y se debe referir al conocimiento del "mundo real". Tal aproximación implica realizar la traducción utilizando representaciones intermedias basadas en elementos "universales" (extra-lingüísticos). Así surgió el sistema KANT (Goodman, 1989; Mitamura *et al.*, 1991). Otro centro que ha impulsado la investigación sobre sistemas de TA basados en conocimiento es el Laboratorio de Investigación (Computing Research Laboratory, CRL) de la Universidad de Nuevo México (Nuevo México, Las Cruces). En él se ha desarrollado el sistema de TA basado en el conocimiento denominado MikroKosmos (Mahesh & Nirenburg, 1995b) además de otros proyectos relacionados con la TA –ULTRA (Farwell & Wilks, 1991), proyecto CREST (Farwell & Helmreich, 2000), etc.– .

Otras aproximaciones interlingua que no están basadas en sistemas de comprensión del lenguaje han aparecido en dos proyectos holandeses: el sistema DLT basado en una modificación del Esperanto (Schubert, 1988), desarrollado en Utrecht, y el sistema Rosetta que está experimentando con la semántica de Montague como base para el interlingua (Appelo & Landsbergen, 1986), desarrollado por Phillips en Eindhoven.

A finales de los 80 surgieron otras alternativas para la TA. Los logros alcanzados en el reconocimiento y la generación automática del habla permitieron el desarrollo de proyectos de TA para la traducción de diálogos: UMIST-BT (Jones & Tsujii, 1990), UMIST-ATR (Somers *et al.*, 1990), VERBMOBIL (Alexandersson *et al.*, 1995), etc. Por otra parte, los avances en la tecnología de los ordenadores (accesos muy rápidos a gran cantidad de información) impulsaron al desarrollo de unos sistemas de TA basados en corpus bilingües alineados. Estos métodos, denominados *basados en ejemplos* o *memorias de traducción* (Sadler, 1989; Sato & Nagao, 1990; Sumita *et al.*, 1990) utilizan como base de conocimientos los corpus en el idioma origen y destino. Por último, la sofistica-

ción de las técnicas estadísticas desarrolladas para los proyectos de reconocimiento y generación automática del habla reavivaron el interés por la aplicación de tales métodos en los sistemas de TA. Uno de los grupos principales en este campo es el laboratorio de IBM en Yorktown Heights, NY, donde se ha desarrollado el sistema Candide (Berger *et al.*, 1994) que usa exclusivamente técnicas estadísticas para realizar el análisis y la generación del texto.

Sin embargo, los avances más significativos de la década de los 80 fue la aparición de los sistemas comerciales de TA. A los productos americanos SPANAM y ENGSPAN –para la traducción español-inglés e inglés-español, creados por la Organización de la Salud Panamericana (Pan American Health Organization, PAHO)–, ALPSystems, Weidner y Logos se les unieron muchos sistemas japoneses de compañías de ordenadores (Fujitsu, Hitachi, Mitsubishi, NEC, Oki, Sanyo, Sharp, Toshiba). Al final de los 80 aparecieron Globalink, PC-Translator, Tovna y el sistema METAL desarrollado por Siemens. La mayoría de estos sistemas son bastante toscos en la calidad lingüística de su salida y dependen en gran medida de una post-edición para producir traducciones aceptables. Al igual que la post-edición, la pre-edición también está extendida. En algunos sistemas, por ejemplo, cuando los operadores introducen el texto deben marcar la separación entre palabras, cláusulas o frases.

El resurgimiento de las investigaciones sobre TA en la década de los 80 y la salida al mercado de numerosos sistemas de TA han resaltado la importancia de las herramientas de traducción. Aún pueden existir interpretaciones erróneas de lo que se ha alcanzado y lo que puede ser el futuro de la TA, pero el buen estado de la TA se refleja en el gran número de sistemas e investigaciones[5] que

[5] Simplemente mencionar la gran cantidad de proyectos subvencionados actualmente por la Unión Europea para el desarrollo de sistemas y herramientas de TA entre los idiomas de la Unión: NL-Translex –TA entre holandés, inglés, francés y alemán, (Cucchiarini *et al.*, 2000)–, TRADUAU-PT (TA entre portugués, inglés, francés y alemán), EC-MT-GPS (TA entre griego, inglés, francés y alemán), MIS (servicio de información multilingüe para agencias de viajes), EuroTerm (enriquecimiento de EuroWordNet con terminología de un dominio específico –dominio del sector público, en concreto, terminología medioambiental– para el holandés, griego y español), etc.

se están explorando. Los avances futuros en la tecnología de los ordenadores, Inteligencia Artificial y Lingüística teórica sugieren posibles líneas futuras de investigación. Pero los problemas fundamentales de la TA no están relacionados con la tecnología sino con el lenguaje, significado, comprensión y las diferencias sociales y culturales de la comunicación humana.

2.4 Clasificación de las estrategias en TA

La clasificación de las estrategias que pueden seguir los sistemas de TA se puede realizar teniendo en cuenta varios criterios que afectan a los idiomas involucrados y a la forma de realizar la traducción. A continuación presentaremos distintas clasificaciones según los criterios de: el número de idiomas participantes, las fuentes de información utilizadas y la existencia de representaciones intermedias.

2.4.1 Conforme al número de idiomas

Según el número de idiomas que intervienen en el proceso de traducción se pueden distinguir los siguientes tipos:

Sistemas bilingües. Los sistemas bilingües son aquéllos que traducen entre dos idiomas. A su vez, en los sistemas bilingües podemos distinguir entre unidireccionales o bidireccionales, según sean capaces de traducir de un idioma a otro en una única dirección o traducir entre dos idiomas en cualquier dirección. Ejemplos de sistemas bilingües son los sistemas Météo (Chandioux, 1976) y Candide (Berger *et al.*, 1994) para inglés-francés, Ntran (Whitelock *et al.*, 1986) para inglés-japonés, etc.

Sistemas multilingües. Los sistemas multilingües se diseñan para traducir entre más de dos idiomas. En un sistema multilingüe "auténtico" los módulos de análisis y generación para un idioma concreto son independientes del resto de idiomas del sistema. Por ejemplo, en un sistema multilingüe para español, inglés y francés, el proceso de análisis para el español debe ser independiente si la traducción es a inglés o francés. Del mismo modo, el proceso de

generación a inglés debe ser el mismo si el idioma origen es español o francés. Ejemplos de sistemas multilingües son el proyecto Eurotra (Allegranza *et al.*, 1991), Systran (Toma, 1977), el sistema Ariane (Boitet & Nédobejkine, 1981), Mikrokosmos (Mahesh & Nirenburg, 1995b), KANT (Mitamura *et al.*, 1991), etc.

2.4.2 Conforme a las fuentes de información

Según las fuentes de información que intervienen en el proceso de traducción podemos clasificar los sistemas de TA en los siguientes tipos:

Sistemas basados en información lingüística. Estas aproximaciones se caracterizan por el uso exclusivo de fuentes de información lingüísticas para realizar la traducción. Generalmente utilizan información léxica, morfológica, sintáctica, semántica y contextual. En estos sistemas, las representaciones intermedias del texto utilizadas durante el proceso de la traducción están basadas en este tipo de información y son dependientes del idioma.

Ejemplos de estos sistemas son Systran (Toma, 1977), Météo (Chandioux, 1976), sistema Ariane (Boitet & Nédobejkine, 1981), proyecto Eurotra (Allegranza *et al.*, 1991), etc.

Sistemas basados en el conocimiento. La justificación básica de los sistemas basados en el conocimiento se fundamenta en el hecho de que la traducción supone la transmisión del significado de un texto de un idioma a otro; por ello, un sistema de TA debe ser capaz de "comprender" el significado de los textos. Comprender un texto supone relacionar lo que se dice en el texto (contenido lingüístico) con fenómenos (entidades, acciones, eventos) fuera del texto ("realidad" no lingüística). Sin comprensión, ningún sistema es capaz de decidir entre varias expresiones en el idioma destino cuál es la que expresa el significado del texto original.

Para comprender el significado de los textos se utilizan unas bases de conocimientos almacenadas previamente (conocimiento del mundo) y se aplican una serie de mecanismos de inferencia. Como resultado, se obtienen unas representaciones *conceptuales* (no lingüísticas) independientes del idioma que expresan el significado de los textos. La complejidad de obtener estas representa-

ciones implica que estas aproximaciones se aplican sobre textos de dominio restringido y con unas aplicaciones muy concretas.

El sistema KANT (Mitamura *et al.*, 1991) de la Universidad Carnegie Mellon y el sistema Mikrokosmos (Mahesh & Nirenburg, 1995b) de la Universidad de Nuevo México están basados en esta aproximación.

Sistemas basados en corpus. Los avances en la tecnología de los ordenadores que proporcionan memorias y almacenamientos de datos cada vez más grandes y más rápidos está fomentando la investigación de métodos basados en el acceso a grandes corpus de textos que se utilizan como principal fuente de información.

Los *sistemas basados en ejemplos*, también denominados *memorias de traducción*, siguen esta estrategia. Se basan en el hecho de que la traducción consiste, a menudo, en encontrar ejemplos análogos, es decir, descubrir o recordar cómo una expresión en el idioma origen ha sido traducida previamente al idioma destino (Samuelson-Brown, 1996).

Las bases de datos que contienen los *ejemplos* se obtienen a partir de un análisis estructural de un gran corpus de textos en el idioma origen y sus correspondientes traducciones al idioma destino. Estos corpus bilingües son alineados (automáticamente o manualmente) de modo que se identifican fragmentos de texto en cada idioma que son equivalentes.

Para realizar la traducción de un texto nuevo se recuperan fragmentos de texto de las bases de datos similares a los del texto que se desea traducir. Si éstos son idénticos y cada uno tiene una traducción distinta[6], se escoge como traducción aquél que tenga mayor frecuencia en el conjunto de frases *ejemplos* para el mismo contexto. Si por el contrario no son iguales, la identificación de la "similitud" de los fragmentos se basa en una medida de distancia del significado. Esta medida se puede basar en una clasificación de las unidades léxicas[7] en jerarquías semánticas –sistema DLT

[6] Si todos los fragmentos seleccionados tienen la misma traducción, ésta se escoge como traducción del fragmento original.

[7] Por unidad léxica se entiende aquella palabra o conjunto de palabras que tienen significado por sí mismas. Así, el conjunto de palabras *por lo tanto* se considera una unidad léxica con significado único.

con almacenamiento bilingüe de conocimiento (Sadler, 1989)–, en cálculos estadísticos, etc.

Las memorias de traducción no están restringidas a ningún corpus en particular. Aunque inicialmente fueron aplicadas sobre dominios específicos, los ejemplos extraídos de un corpus pueden ser igual de eficaces para traducir textos en otra área.

Aunque generalmente se acepta que los sistemas basados en ejemplos funcionan mejor con conjuntos estructurados de textos bilingües, se ha desarrollado algunos experimentos –por ejemplo, el sistema Candide (Berger *et al.*, 1994) desarrollado por IBM– que establecen las correspondencias de las unidades léxicas entre los textos origen y destino exclusivamente con cálculos estadísticos.

2.4.3 Conforme a la existencia de representaciones intermedias

Dependiendo de que se utilicen o no representaciones intermedias durante el proceso de traducción, podemos distinguir dos tipos:

Sistemas directos. Los sistemas directos (también denominados sistemas de "primera generación") pertenecen a la primera estrategia realizada para TA. Se denominan así porque la traducción del texto se produce directamente, sin que se genere una representación intermedia del mismo.

Esta aproximación fue adoptada por la mayoría de los sistemas de TA de la década de los 50. Si se desea conocer cómo operaban estos sistemas, hay que tener en cuenta que los ordenadores disponibles al final de los 50 y principios de los 60 eran muy primitivos en relación con los potentes ordenadores de hoy en día. En aquella época no existían los lenguajes de alto nivel y la mayoría de la programación se realizaba en lenguaje ensamblador.

A grandes rasgos, estos sistemas de primera generación empezaban con lo que se podría denominar fase de análisis morfológico, donde se identificaban las terminaciones de las palabras y se reducían las palabras a sus formas básicas (únicamente contienen las raíces de las palabras). Los resultados de esta fase serían la

entrada para un módulo de búsqueda de palabras en un gran diccionario bilingüe. Hay que destacar que no se realizaba ni análisis sintáctico ni se establecían las relaciones semánticas, etc. Tras la búsqueda de las palabras en el diccionario se aplicaban unas reglas de reordenación local (adjetivos, verbos compuestos, etc.) para producir un resultado aceptable en el idioma destino. La última fase consistía en la generación total del texto en el idioma destino.

En la figura 2.1 se resume la aproximación directa.

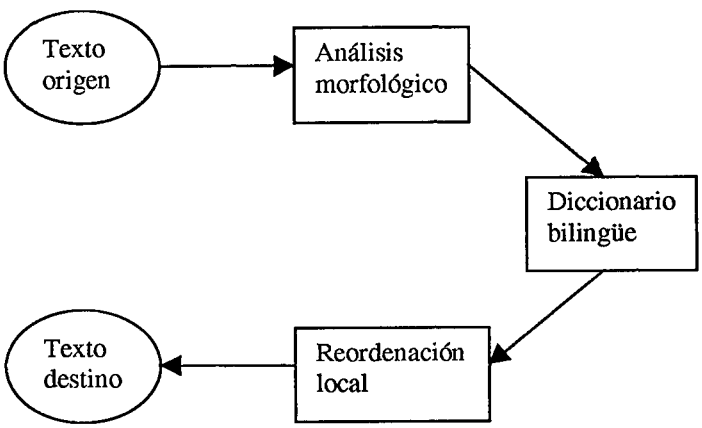


Figura 2.1. Sistemas directos

Esta aproximación se puede decir que es una traducción "palabra a palabra" con alguna reordenación local de palabras. Las grandes limitaciones de esta aproximación son obvias. Se producirán frecuentes errores en la traducción de palabras a nivel léxico y una gran cantidad de estructuras sintácticas incorrectas ya que se basarán en el idioma origen.

Rápidamente se reconoció la ingenuidad de esta aproximación desde el punto de vista lingüístico y computacional. Desde el punto de vista lingüístico lo que faltaba era realizar un análisis de la estructura interna del texto, particularmente las relaciones gramaticales entre las principales partes de las oraciones. Desde el

punto de vista computacional, las limitaciones venían impuestas por los ordenadores rudimentarios y las características de los lenguajes de programación de aquella época.

Ejemplos de sistemas que siguen la aproximación directa son Systran (Toma, 1977) y Météo (Chandioux, 1976).

Sistemas indirectos. Los fallos de los sistemas de la primera generación llevaron al desarrollo de modelos lingüísticos más sofisticados para la traducción. En concreto, se incrementó el interés por realizar un análisis de los textos en el idioma origen y representarlos mediante una estructura intermedia (representar el significado del texto de algún modo). A partir de esta representación intermedia se realizaría la generación en el idioma destino. Ésta es la esencia de los métodos indirectos (también denominados sistemas de "segunda generación"), que tienen dos variantes principales: los sistemas interlingua y los sistemas de transferencia.

- **Sistemas interlingua.** En los sistemas interlingua (históricamente la primera estrategia indirecta) se asume la posibilidad de convertir los textos a una representación independiente del idioma que contiene el "significado" de los mismos. A partir de esta representación se puede traducir el texto a cualquier idioma. De este modo, la traducción se realiza en dos fases: en la primera se traduce el texto del idioma origen (texto origen) a la representación interlingua, y en la segunda, desde el interlingua al idioma destino. Los módulos para el análisis son independientes de los módulos para la generación.

La representación intermedia incluye toda la información necesaria para la generación del texto destino sin tener que volver "atrás" a examinar de nuevo el texto original. Por tanto, esta representación es una proyección del texto origen y a su vez es la base para la generación del texto destino. El método es interlingua en el sentido de que esta representación es neutral entre dos o más idiomas. En el pasado, se pretendía la construcción de una representación interlingua que fuera completamente "universal" y pudiera servir de intermediaria entre cualquier idioma. Actualmente, los sistemas interlingua son menos ambiciosos.

La aproximación interlingua es más atractiva para sistemas multilingües. Cada módulo de análisis puede ser independiente

del resto de módulos de análisis y de los módulos de generación. En la figura 2.2 se observa un modelo interlingua para dos idiomas.

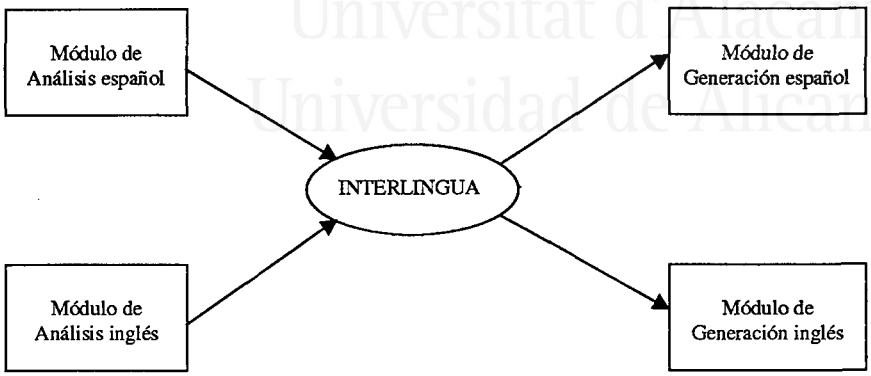


Figura 2.2. Modelo interlingua con dos idiomas

La ventaja de esta aproximación consiste en que la adición de un nuevo idioma al sistema supone únicamente la creación de dos nuevos módulos: un módulo de análisis y un módulo de generación para el nuevo idioma. En la figura 2.3 se muestra esta nueva configuración.

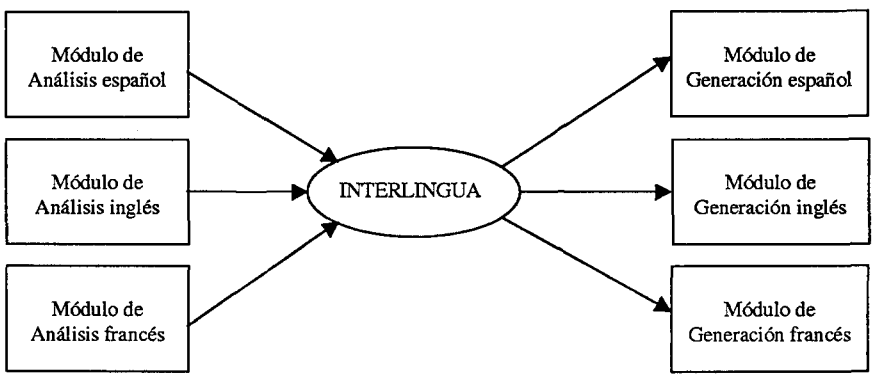


Figura 2.3. Modelo interlingua con tres idiomas

Como se puede observar, la adición de dos nuevos módulos ha incrementado el número de traducciones de dos a seis. En general, para n idiomas se obtendrán $n(n-1)$ traducciones distintas.

Hay que destacar, que esta configuración permite la traducción de un idioma a sí mismo, por ejemplo, la conversión de un texto origen español a la representación interlingua y a partir de esta representación volver a generar el texto destino en español. Esta capacidad de los sistemas interlingua es útil durante la fase de desarrollo del sistema para comprobar el correcto funcionamiento de los módulos de análisis y generación. Podría ocurrir que el texto generado no sea idéntico al texto origen, pero sí fuera equivalente en significado.

Los sistemas DLT (Schubert, 1988), Rosetta (Appelo & Landsbergen, 1986), proyecto CREST (Farwell & Helmreich, 2000), KANT (Mitamura *et al.*, 1991), Mikrokosmos (Mahesh & Nirenburg, 1995b), etc. son ejemplos de sistemas interlingua.

- **Sistemas de transferencia.** La segunda variante de la aproximación indirecta son los sistemas de transferencia (figura 2.4).

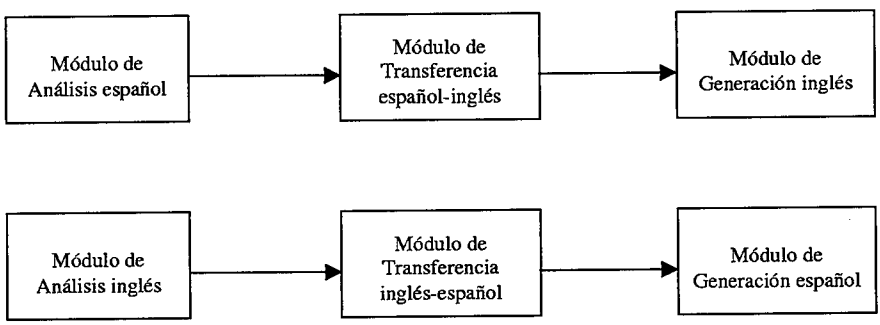


Figura 2.4. Modelo de transferencia con dos idiomas

A diferencia de los sistemas interlingua en los que la traducción se realizaba en dos fases, los sistemas de transferencia realizan la traducción en tres fases: la primera fase convierte el texto origen en una representación intermedia sin ambigüedades; en la segunda fase, esta representación se convierte a la representación equivalente en el idioma destino; en la tercera fase, el texto se genera

en el idioma destino (texto destino). Los módulos de análisis y generación son específicos para cada idioma y son independientes entre sí. Estos módulos están relacionados por medio de los módulos bilingües de transferencia intermedios.

En estos sistemas, las representaciones intermedias son dependientes del idioma: el resultado del análisis es una representación abstracta del texto en el idioma origen y la entrada para la generación es una representación abstracta del texto en el idioma destino. La función de los módulos de transferencia bilingües es convertir las representaciones intermedias en el idioma origen a las representaciones intermedias en el idioma destino.

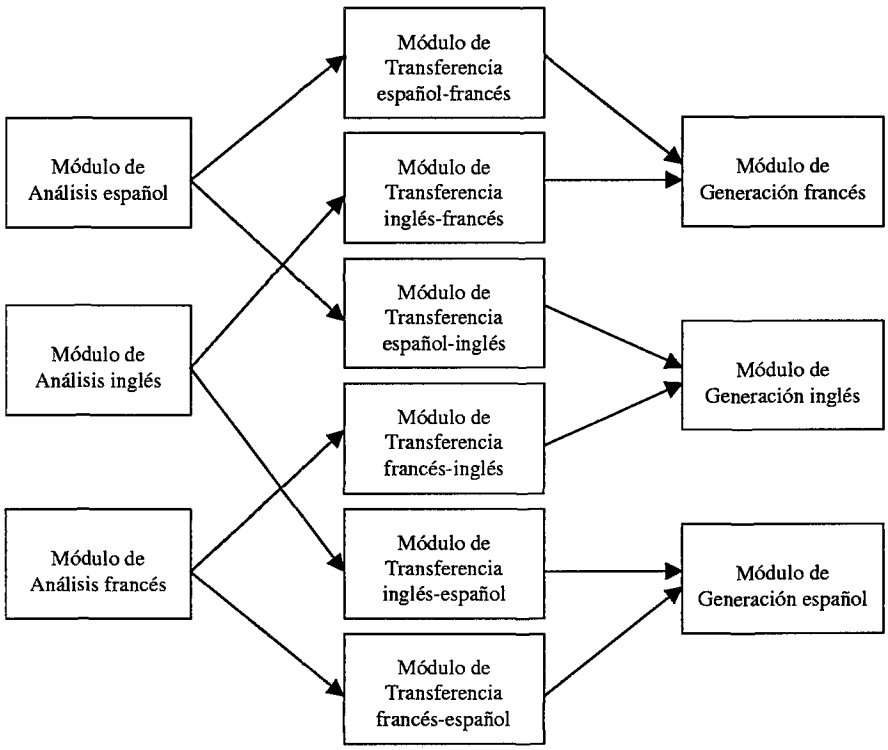


Figura 2.5. Modelo de transferencia con tres idiomas

En comparación con los sistemas interlingua hay unas claras desventajas de los sistemas de transferencia. La adición de un nuevo idioma supone, no sólo la creación de dos nuevos módulos para análisis y generación sino también la adición de nuevos módulos de transferencia, el número de los cuales varía dependiendo del número de idiomas existentes en el sistema. Por ejemplo, en el caso de un sistema con dos idiomas, la adición de un tercer idioma requeriría cuatro nuevos módulos de transferencia. En la figura 2.5 se puede observar esta nueva estructura.

La adición de un cuarto idioma supondría el desarrollo de seis nuevos módulos de transferencia. En general, el número de módulos de transferencia en un sistema de transferencia multilingüe, para todas las combinaciones de n idiomas, es $n(n-1)$, es decir, muy cercano a n^2. En la tabla 2.1 se muestra el número de módulos necesarios (análisis, generación y transferencia) en un sistema de transferencia y en un sistema interlingua.

Número de Idiomas	Módulos de Análisis	Módulos de Generación	Módulos de Transferencia	Módulos totales Sist. Transfer.	Módulos totales Sist. Interlingua
2	2	2	2	6	4
3	3	3	6	12	6
4	4	4	12	20	8
5	5	5	20	30	10
...					
9	9	9	72	90	18
n	n	n	$n(n-1)$	$n(n+1)$	$2n$

Tabla 2.1. Módulos requeridos en un sistema de transferencia y en un sistema interlingua multilingües

Como se observa en la tabla 2.1 el número de módulos requerido en un sistema multilingüe de transferencia es mucho mayor (dependiendo del número de idiomas) que en un sistema interlingua. Sin embargo, muchos sistemas prefieren el método de transferencia en vez del método interlingua. Existen dos razones que justifican esta decisión. La primera de ellas radica en la dificul-

tad de desarrollar representaciones interlingua independientes del idioma. La segunda es la complejidad de las gramáticas de análisis y generación en los sistemas interlingua. La relativa complejidad de los módulos de análisis y generación usando un método de transferencia, es mucho más reducida que en un sistema interlingua ya que las representaciones intermedias son representaciones abstractas dependientes del idioma.

Ejemplos de sistemas de transferencia son los siguientes: SUSY (Maas, 1977), sistema Ariane (Boitet & Nédobejkine, 1981), proyecto Eurotra (Allegranza *et al.*, 1991), METAL (Bennet & Slocum, 1985), etc.

2.5 Organización de los datos en los sistemas de TA

Los datos lingüísticos requeridos en los sistemas de TA se pueden dividir en dos grandes grupos: datos gramaticales y datos léxicos. Estos datos, aunque están relacionados, normalmente se almacenan de forma independiente.

- Los datos gramaticales se expresan generalmente mediante las gramáticas de análisis y generación usadas por el sistema de TA. En ellas se establecen las combinaciones aceptables de las categorías gramaticales incluyendo, además, restricciones sobre sus características (género, número, etc.).
- Los datos léxicos contienen información específica de cada unidad léxica del vocabulario del idioma implicado. Esta información es mucho más específica que la que aparece en un diccionario tradicional, ya que debe especificar: categoría gramatical, características morfológicas, características semánticas, restricciones de selección, características de sus posibles argumentos, etc. Por esta razón, generalmente se le denomina *lexicón* en lugar de diccionario.
  Según la estrategia usada en el sistema de TA[8], la organización de los datos léxicos es diferente. Así en los sistemas directos exis-

[8] En esta Tesis utilizaremos normalmente la clasificación de los sistemas de TA basada en la existencia o ausencia de representaciones intermedias. De este modo,

te básicamente un *lexicón bilingüe* que contiene datos de las unidades léxicas en el idioma origen y sus equivalentes en el idioma destino. Típicamente contendría una entrada para cada unidad léxica del idioma origen incluyendo información morfológica y semántica, su(s) traducción(es) al idioma destino (que incluye su información gramatical), información necesaria para seleccionar una de entre varias palabras destino alternativas y, por último, información que permitiera cambiar estructuras sintácticas por las apropiadas en el idioma destino. El resultado será un lexicón de gran tamaño y complejidad.

Por otra parte, los sistemas indirectos tienen módulos de análisis y generación independientes por lo que utilizan un lexicón monolingüe para los idiomas origen y destino respectivamente y un lexicón bilingüe de transferencia.

- El *lexicón monolingüe para el análisis* del idioma origen tiene la información necesaria para el análisis estructural y la desambiguación del significado de la palabra. Contiene información morfológica, categoría gramatical, características semánticas y restricciones de selección.
- El *lexicón bilingüe* para convertir las unidades léxicas del idioma origen al idioma destino (sistema de transferencia) o convertir las unidades léxicas hacia/desde la representación interlingua (sistema interlingua) es muy simple. Contiene las correspondencias léxicas de las palabras según su significado y la información gramatical mínima de los dos idiomas.
- El *lexicón para la generación* es generalmente más sencillo que el utilizado para el análisis ya que no necesita información para llevar a cabo la desambiguación.

  Hay que mencionar que algunos sistemas indirectos incorporan la mayor parte de la información para seleccionar las formas de las palabras destino en sus lexicones bilingües, por lo que los lexicones para la generación contienen únicamente información morfológica.

En la práctica, los sistemas de TA suelen dividir los lexicones en diccionarios especiales para "palabras de frecuencia alta",

los sistemas seguirán una de las tres estrategias básicas: sistemas directos, de transferencia e interlingua.

frases idiomáticas, formas irregulares, etc. Incluso suelen utilizar lexicones de dominios específicos con el objetivo de reducir los problemas de ambigüedad de las palabras.

2.6 Módulo de Análisis de un sistema de TA. Tratamiento de la ambigüedad

Como se ha visto en el diseño de los sistemas de TA en función de la utilización de representaciones intermedias (sistemas directos, de transferencia e interlingua) podemos distinguir tres componentes básicos en su estructura: un módulo monolingüe de análisis para el idioma origen, un módulo para la generación del texto destino y un interfaz entre estos dos módulos monolingües. Las diferencias básicas en las tres estrategias de los sistemas de TA radican en los tamaños de estos tres componentes. El método directo está en un extremo, el método interlingua está en el otro extremo y los sistemas de transferencia están en el punto medio. La conocida "pirámide" (Vauquois, 1968) de la figura 2.6 muestra las tres estrategias.

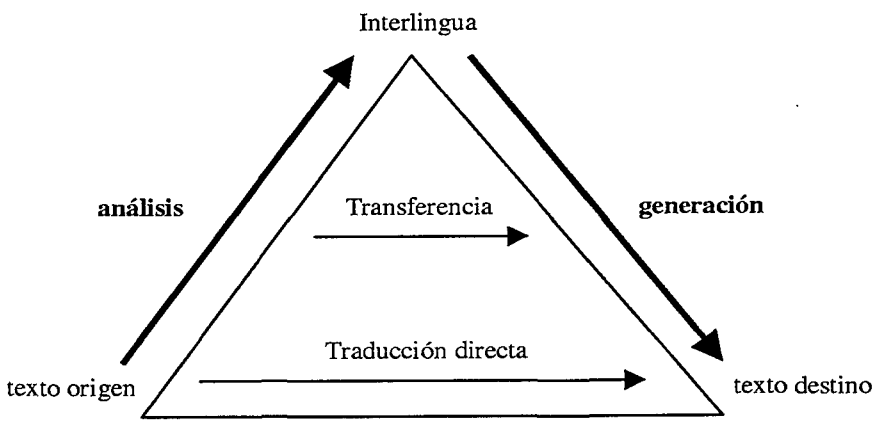


Figura 2.6. Pirámide de la transferencia en los sistemas de TA

En la figura 2.6 aparece el sistema interlingua en el ápice de la pirámide. La representación interlingua se obtiene mediante un análisis monolingüe y permite su uso directo para la generación. Sin embargo, el camino hasta la representación interlingua es largo y, como se observa en el gráfico, este análisis monolingüe se puede interrumpir en algún punto y entrar en una fase de transferencia bilingüe que evitaría las dificultades de un análisis completo y exhaustivo. Como el gráfico muestra, a medida que el texto es analizado con más detalle más fácil será la transferencia. En el otro extremo (la base de la pirámide) se encuentran los sistemas directos en los que prácticamente no hay análisis monolingüe y casi todo el trabajo se realiza en la transferencia.

En esta sección y en las dos siguientes analizaremos en detalle cada uno de los módulos implicados en el proceso de la TA: módulo de análisis, módulo de transferencia y módulo de generación.

En el módulo de análisis de un sistema de TA los principales problemas que se tienen que abordar se producen por la ambigüedad inherente del lenguaje humano.

La ambigüedad ocurre cuando una palabra se puede analizar potencialmente de dos o más modos distintos. Según se presenta en Moreno *et al.* (1999), el tratamiento de la ambigüedad contempla dos subproblemas:

- La representación del problema: Cómo las diversas interpretaciones se representan en un sistema.
- La interpretación del problema: Qué estrategias se siguen cuando aparece una ambigüedad para seleccionar una interpretación.

La ambigüedad se puede producir a nivel léxico, sintáctico, semántico o contextual por lo que su resolución es de crucial importancia en el proceso de la TA, ya que la traducción al idioma destino será diferente (aunque no siempre) dependiendo de la solución escogida. En esta sección presentaremos los distintos tipos de ambigüedad y los mecanismos utilizados para su resolución.

- Ambigüedad léxica. La ambigüedad léxica ocurre cuando una palabra tiene más de un significado, es decir, se puede interpretar de varias maneras posibles. Desde el punto de vista de

la Lingüística se puede hacer una distinción entre *homografía* y *polisemia*:

- Dos (o más) palabras son *homógrafas* si sus significados son totalmente diferentes.
- La *polisemia* ocurre cuando una palabra puede tener un conjunto de significados, pero todos ellos están relacionados de alguna manera.

La ambigüedad léxica se puede clasificar en *ambigüedad léxica pura* y *ambigüedad léxica categorial* (Moreno *et al.*, 1999).

1. La *ambigüedad léxica pura* ocurre cuando dos palabras se escriben igual, tienen la misma categoría sintáctica y significan dos cosas distintas. Por lo tanto, la ambigüedad léxica pura sólo afecta al nivel semántico.
2. La *ambigüedad léxica categorial* se presenta cuando una palabra, además de tener distintos significados, éstos desempeñan funciones sintácticas diferentes en la oración.

Para elaborar un mecanismo que determine la correcta interpretación de una palabra que posee una ambigüedad léxica, será necesario tener la representación de sus diversos significados. En el caso de la ambigüedad categorial, además se necesitará conocer sus funciones sintácticas, las cuales nos permitirán posteriormente tomar una decisión acerca de la función sintáctica y significado más apropiado en relación a los elementos que forman la oración.

Sin embargo, hay algunos casos (palabras homógrafas con la misma categoría sintáctica) en los que la información almacenada –significado de la palabra y función sintáctica– es insuficiente para determinar el significado correcto de una palabra. Por esta razón, se utiliza la información semántica. Una técnica muy común consiste en asignar *características semánticas* a cada una de las posibles interpretaciones de una palabra y determinar mediante *restricciones de selección* las características que son compatibles entre sí. Por último, si con la incorporación de información semántica no se puede averiguar la correcta interpretación de una palabra será necesario recopilar toda la

información lingüística, contextual, e incluso conocimiento del mundo real para poder descubrir la interpretación de la misma.

- Ambigüedad estructural. La ambigüedad estructural aparece a nivel de oraciones o sintagmas (a diferencia de la ambigüedad léxica que sucede con palabras individuales), siendo ésta esencialmente sintáctica. Se dice que una oración es ambigua estructuralmente cuando posee más de un árbol sintáctico de derivación, es decir, tiene diferentes estructuras sintácticas.
  Para resolver correctamente este tipo de ambigüedad es necesario aplicar distintas fuentes de información (información semántica, estadística, contextual o conocimiento del mundo real) que nos permita tomar una decisión coherente acerca de la estructura sintáctica de la oración más adecuada.
- Ambigüedad de ámbito de cuantificación. Esta clase de ambigüedad se presenta cuando aparecen en una misma oración un cuantificador existencial y un cuantificador universal. Para resolver este tipo de ambigüedad se suelen adoptar unos criterios basados en una serie de reglas de precedencia de los cuantificadores.
- Ambigüedad de función contextual. La ambigüedad de función contextual aparece a nivel oracional (al igual que la ambigüedad estructural y la ambigüedad de ámbito de cuantificación). Se define como las diversas interpretaciones o sentidos que puede tener una oración dependiendo del contexto.
  Una solución a un nivel oracional consistiría en establecer una prioridad en los contenidos semánticos que puede tener una palabra en el universo de discurso, eligiendo el significado más prioritario cuando se presenta una ambigüedad que no puede resolverse por falta de información.
  Se podría adoptar una solución a nivel pragmático que estaría basada en la búsqueda de referencias de las mismas palabras en oraciones anteriores en el contexto. Cuando aparece una palabra con ambigüedad se determinaría el contenido semántico más correcto que se debería asumir.
- Ambigüedad referencial. La ambigüedad referencial afecta exclusivamente al nivel contextual o pragmático.

La ambigüedad referencial surge cuando para un pronombre o sintagma nominal definido aparece más de un antecedente válido. En el ejemplo 3 aparece una frase con ambigüedad referencial. La expresión referencial pronominal –*él*– tiene dos antecedentes igualmente válidos –*Pedro* y *carpintero*–.

(3) El tío de Pedro es carpintero. Él arregló la mesa sin problemas.

Esta frase, por lo tanto, tendría dos interpretaciones válidas:

– Pedro arregló la mesa.
– El carpintero arregló la mesa.

Una posible solución para resolver esta ambigüedad referencial pronominal consiste en escoger el antecedente más cercano a la expresión referencial (Allen, 1995). Existen otros muchos métodos para resolver este tipo de ambigüedad que serán vistos con detalle en el capítulo 4.

Por último, hay que destacar que el establecimiento de los antecedentes de una expresión referencial es fundamental para una correcta traducción en TA. Cuando traducimos a un lenguaje que marca el género de los pronombres (expresiones referenciales pronominales) es esencial resolver correctamente estas expresiones. Por ejemplo, si traducimos de inglés a español, el pronombre *they* (válido para género masculino y femenino en inglés) debe ser traducido a su correspondiente en español: *ellos* (género masculino) o *ellas* (género femenino). La información para una u otra traducción debe ser extraída del antecedente.

Esta Tesis está centrada en la resolución de esta ambigüedad referencial pronominal (en inglés y español) y en el planteamiento de un sistema interlingua que permite generar correctamente las referencias pronominales en idiomas que marcan el género del pronombre, como el español. Además de los problemas planteados por el género de los pronombres, se estudiarán los problemas planteados por el número gramatical de los mismos, las construcciones elípticas, etc.

2.7 Módulo de Transferencia de un sistema de TA

Una vez analizadas las dificultades monolingües que causan problemas durante el análisis del texto en el idioma origen, en esta sección se presentarán, en primer lugar, las diferencias léxicas y estructurales entre idiomas que hay que tratar en el módulo de transferencia de un sistema de TA. Posteriormente se describirán, en orden creciente según la complejidad del análisis realizado, los niveles de transferencia que se pueden llevar a cabo en los distintos sistemas de TA. Para ello, se comenzará por el nivel de transferencia que ocupa la base de la pirámide de transferencia (figura 2.6), los sistemas de transferencia morfológica o directos, y se concluirá con el nivel que ocupa el ápice de la pirámide, los sistemas sin transferencia o interlingua. En cada uno de estos niveles de transferencia se explicará el tipo de análisis que es necesario para realizar la transferencia.

2.7.1 Diferencias léxicas

Las diferencias léxicas ocurren cuando una única palabra en el idioma origen se puede traducir potencialmente de diferentes formas en el idioma destino, no porque la palabra en el idioma origen sea ambigua sino porque es ambigua desde la perspectiva del idioma destino. Podemos distinguir tres grandes grupos de diferencias léxicas: ambigüedades estilísticas, gramaticales y conceptuales.

1. Ambigüedades estilísticas: ocurren cuando la elección de la palabra destino depende del tipo de texto que vaya a ser generado. Por ejemplo la palabra francesa *domicile* podría traducirse a inglés como *home* o *domicile* dependiendo del tipo de documento que se vaya a generar.
2. Ambigüedades gramaticales: ocurren cuando la elección de la palabra destino está condicionada por el contexto gramatical. Por ejemplo la palabra inglesa *know* se puede traducir al francés por *connaître* o *savoir*, dependiendo de que el objeto directo sea un sintagma nominal (*connaître*) o una oración subordinada o un infinitivo (*savoir*).

3. Ambigüedades conceptuales: ocurren cuando un único "concepto" representado por una palabra en el idioma origen corresponde a un número de conceptos (palabras) en el idioma destino. Por ejemplo, en la traducción inglés-español la palabra inglesa *leg* se puede traducir como las palabras españolas: *pierna* (de una persona), *pata* (de un animal, mesa o silla), *etapa* (de un viaje). Hay que destacar que estas ambigüedades son la causa de los problemas más importantes en el proceso de la traducción.

2.7.2 Diferencias estructurales

Las diferencias estructurales ocurren cuando la estructura de los componentes de una oración en el idioma origen es diferente a la estructura en el idioma destino. Las diferencias estructurales se pueden clasificar en: diferencias sintácticas, diferencias debidas a palabras particulares, diferencias por campos semánticos y frases idiomáticas.

1. Diferencias sintácticas: son las diferencias en la estructura de la oración producidas por las diferentes sintaxis de los idiomas. Por ejemplo, en español y francés la mayoría de los adjetivos se colocan después del nombre mientras que en inglés preceden al nombre. Otras diferencias se producen por la colocación del verbo principal de la oración: en japonés y latín, por ejemplo, se coloca al final de la oración mientras que en otros idiomas, como inglés y alemán, el verbo aparece después del primer sintagma nominal. Sin embargo, y a diferencia de los ejemplos anteriores, hay algunos casos en los que el proceso de igualar una estructura de un idioma a otro es mucho más complejo; como ejemplo se puede mencionar el tratamiento de la voz pasiva entre varios idiomas: inglés, francés, español, alemán o japonés.
2. Diferencias debidas a palabras particulares: son las diferencias estructurales producidas por algunas palabras particulares. Por ejemplo, cuando el verbo español *soler* va acompañado de un complemento en infinitivo se traduce al inglés por el adverbio *usually*.

3. Diferencias por campos semánticos: ocurren cuando una expresión se representa estructuralmente diferente según el idioma. Como ejemplo se pueden citar las expresiones de movimiento: en inglés el modo de transporte se expresa mediante el verbo (*walk*, *fly*, *drive*) y la dirección mediante preposiciones (*across*, *into*, *from*). En francés, la dirección se expresa con el verbo (*traverser*, *entrer*, *partir*) y el modo de transporte con adjuntos adverbiales (*à pied*, *en voiture*, *par avion*).
4. Frases idiomáticas: las frases o expresiones idiomáticas normalmente producen cambios en la estructura de la oración según el idioma. Podemos distinguir dos tipos. Por una parte aquéllas que se pueden traducir directamente de un idioma a otro por composición de palabras: la expresión inglesa *to ask for the moon* y su correspondiente traducción al español *pedir la luna*. Estas expresiones idiomáticas no son tratadas como tales ya que no suponen cambios en la estructura del idioma destino.
   Por otra parte están aquellas expresiones idiomáticas que suponen un cambio en la estructura: la expresión francesa *donner un coup de poing* (dar un golpe con el puño) y su correspondiente traducción al inglés *punch* (dar un puñetazo). Hay que destacar que los textos ricos en estos tipos de expresiones no son los más idóneos para ser procesados con un sistema de TA a pesar de los numerosos esfuerzos que se han realizado para resolver este problema.

2.7.3 Sistemas de transferencia morfológica (sistemas directos)

El análisis más superficial que se podría realizar sobre un texto consistiría en un análisis morfológico "palabra a palabra". Este análisis identifica las categorías gramaticales de las palabras, el número gramatical (singular o plural) de los nombres y el tiempo de los verbos, pero no identifica relaciones entre palabras (determinantes y nombres) o entre grupos de palabras (sintagmas nominales). En un sistema rudimentario "palabra a palabra" la transferencia consistiría simplemente en la sustitución de las pala-

bras origen (con sus rasgos gramaticales) por las palabras destino (con sus rasgos correspondientes).

Como se puede pensar las traducciones resultantes son de escasa calidad, incluso para oraciones donde lo único que se necesita es una pequeña reordenación del texto. En la práctica, la mayoría de los sistemas de traducción directa incluyen una mínima identificación del contexto local. Así por ejemplo la secuencia inglesa *adjetivo + nombre* se reemplazará por la secuencia española *nombre + adjetivo*. Las excepciones a este tipo de reglas se indicarán en el diccionario.

Estos pequeños cambios estructurales se suelen llevar a cabo en una fase posterior a la sustitución léxica de las palabras origen por las palabras destino. Por esta razón, la "reordenación local" normalmente se considera como el principio de la "generación" en estos sistemas.

El principal problema que tiene que resolver la aproximación directa consiste en la resolución de la ambigüedad léxica de las palabras. Normalmente la resuelve buscando en las palabras adyacentes para obtener alguna pista. En el caso de que no la pueda resolver (porque fuera necesario un análisis estructural completo del texto), normalmente el sistema proporciona traducciones alternativas para que en un proceso posterior de post-edición se eligiera la alternativa correcta.

2.7.4 Sistemas de transferencia sintáctica

El siguiente nivel de análisis posterior al análisis morfológico consiste en el análisis sintáctico. Tras realizar el análisis sintáctico se obtiene la representación de la estructura superficial del texto de entrada (árbol sintáctico) en el idioma origen.

Aunque esta representación proporciona algo de información estructural del texto original no tiene información acerca de las relaciones funcionales entre los elementos. Por ejemplo, si utilizamos este árbol sintáctico como base para una transferencia inglés-español, podríamos obtener internamente constituyentes coherentes en español, pero no se sabría la estructura de la oración como

un "todo", ya que no se ha realizado un análisis de las funciones sintácticas o semánticas de los elementos.

Por lo tanto, generalmente se acepta que la transferencia se debe basar en un análisis más "profundo" que incluya las relaciones funcionales entre los elementos. Tras realizar este análisis se obtiene un nuevo árbol sintáctico que contiene: información del verbo principal de la oración; relaciones funcionales como sujeto, complementos, modificadores y cuantificadores; el núcleo de cada constituyente y la información gramatical de todos los constituyentes (número, género, persona, etc.) que ha sido convenientemente transmitida a los nodos superiores del árbol.

En el proceso posterior para realizar la transferencia de esta nueva estructura del idioma origen al idioma destino habría que tratar la transferencia léxica y la transferencia estructural.

Transferencia léxica. La transferencia léxica consiste en la sustitución de un elemento léxico (palabra) en el idioma origen por un elemento léxico en el idioma destino. Se pueden distinguir tres casos:

1. Si sólo hay una unidad léxica destino equivalente –traducciones uno-a-uno– (*library* en inglés, *biblioteca* en español) entonces no hay problema en la transferencia.
   Sin embargo, estos tipos de transferencias léxicas normalmente sólo se producen cuando se realizan traducciones de textos técnicos (*screen – pantalla*; *printer – impresora*, etc.)
2. Las transferencias léxicas de las traducciones muchos-a-uno tampoco presentan problemas. Por ejemplo, las palabras españolas *rincón* y *esquina* se traducen al inglés por *corner*.
3. Las transferencias léxicas problemáticas surgen con las traducciones uno-a-muchos (sección 2.7.1). Para resolver estos casos y escoger una única palabra destino se utilizan varias estrategias. Podemos mencionar:
   a) La inspección de las palabras que rodean a la palabra problemática. Por ejemplo, la palabra inglesa *know* se traduce al francés o alemán de modo diferente dependiendo de que su complemento sea una cláusula o un sintagma nominal.

b) Características semánticas de los constituyentes. Por ejemplo, el pronombre personal español *él* se puede traducir al inglés por *he* o *it* dependiendo de que se refiera a un humano (*he*) o a un animal o cosa (*it*).

c) Conocimiento del mundo. Existen muchas traducciones que implican realizar distinciones que son obvias para los humanos pero que son difíciles de expresar computacionalmente. En estos casos sería necesario tener un conocimiento del mundo para realizarlas correctamente.

Transferencia estructural. La transferencia estructural es necesaria cuando la estructura heredada del idioma origen no es apropiada para el idioma destino. En teoría, cuanto más profundo sea el análisis en el idioma origen más pequeñas serán las diferencias estructurales, ya que el objetivo de profundizar en el análisis es la neutralización de las diferencias entre los idiomas.

Como ya se ha comentado en la sección 2.7.2, las diferencias estructurales entre los idiomas se han clasificado en varios tipos. Hay algunos de estos problemas que son más fáciles de resolver que otros, pero todos ellos en general son resueltos, normalmente, con el uso de "reglas de transferencia estructurales". El objetivo de estas reglas de transferencia estructurales es transformar la estructura superficial[9] del texto origen en la estructura superficial del texto destino.

Consideremos el ejemplo 4 en el que se muestra una frase en inglés y su traducción al español con las correspondientes diferencias estructurales.

(4) **I** Jones *likes* the film.

E La película le *gusta* a Jones.

En el ejemplo 4 el verbo *like* se ha traducido por *gustar*. Sin embargo, la estructura de la frase 4.i y la estructura de la frase 4.e

[9] Hay que destacar la distinción entre *estructura superficial* y *estructura profunda* de un texto. La estructura profunda contiene la representación del significado del texto, mientras que la estructura superficial contiene la secuencia de palabras correspondientes. Por ejemplo, a partir de la forma lógica que representa el significado de la oración –estructura profunda– se podría generar la secuencia de palabras correspondientes a esa oración –estructura superficial–.

no coinciden[10]. Como se puede observar en el ejemplo, el sujeto de *like* se ha convertido en el objeto indirecto de *gustar* y el objeto directo de *like* se ha convertido en el sujeto de *gustar*. Además en la traducción a español se ha introducido un nuevo pronombre, *le*, que hace referencia al objeto indirecto.

En la figura 2.7 se muestra la regla de transferencia estructural para la traducción de inglés a español del verbo *like*. En esta figura aparece la estructura sintáctica en inglés y su correspondiente en español. Como se puede observar la regla de transferencia incluye: información del verbo principal de la oración (*trans* –transitivo–, *intrans* –intransitivo–); constituyentes de la oración (*SN* –sintagma nominal–, *pron* –pronombre–, *verbo*, *SP* –sintagma preposicional–) e información de los constituyentes (*Func* –función sintáctica en la oración: *Sujeto*, *Objeto Directo*, *Objeto Indirecto*–, *Núm* –número gramatical–, *Palab* –palabra o unidad léxica–). Aunque se han omitido del gráfico, existirían otras reglas de transferencia para los elementos etiquetados con el mismo nombre (SN1 → SN1', etc.).

Reglas de transferencia similares a las de la figura 2.7 se definirán para resolver todos los tipos de diferencias estructurales entre dos idiomas planteados en la sección 2.7.2.

Una vez aplicado el conjunto de reglas de transferencia estructurales sobre la estructura en el idioma origen se obtiene una estructura en el idioma destino que servirá como entrada para el siguiente módulo de generación del idioma destino.

La mayoría de los sistemas basados en la transferencia utilizan estas representaciones intermedias o de transferencia como paso previo a la generación en el idioma destino. Hay que destacar que éstas son específicas para cada idioma: contienen las palabras en el idioma origen y en el idioma destino, y reflejan la estructura, bien superficialmente o en profundidad, de los respectivos idiomas origen y destino. Por el contrario, los sistemas interlingua intentan proporcionar representaciones que son independientes del idioma tanto en las palabras (unidades léxicas) como en la estructura.

[10] A partir de ahora en esta Tesis las frases en español e inglés referidas a un mismo ejemplo se referenciarán como n°ejemplo.e y n°ejemplo.i respectivamente.

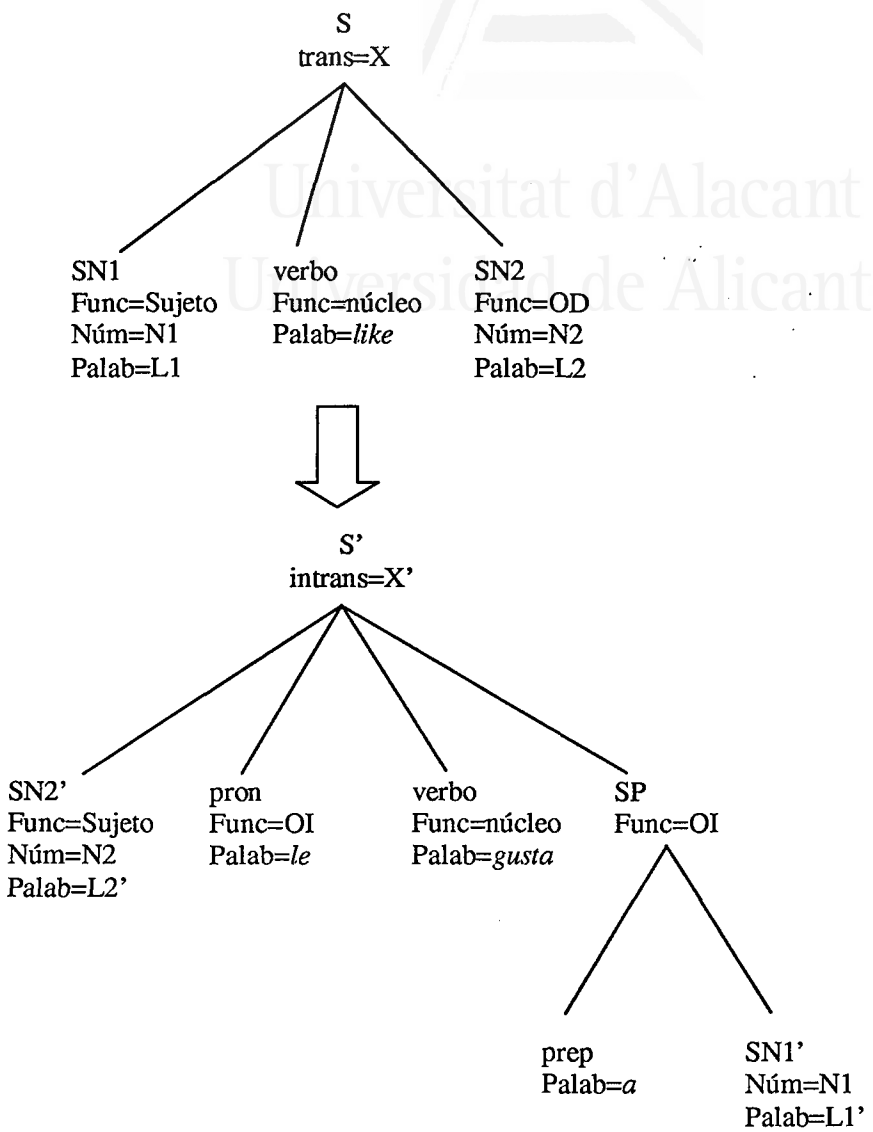


Figura 2.7. Regla de transferencia inglés-español para el verbo *like*

2.7.5 Sistemas de transferencia semántica

Los problemas de la transferencia estructural son patentes cuando se obtienen representaciones que no neutralizan suficientemente las particularidades entre las estructuras superficiales, e incluso entre las profundas, de los idiomas respectivos.

Los sistemas de TA basados en la transferencia sintáctica funcionan bien para oraciones sencillas pero necesitan realizar un análisis más profundo, debido a que las relaciones sintácticas (sujeto, objeto, etc.) y las relaciones semánticas (AGENTE, TEMA, etc.) de los constituyentes de una oración pueden variar de un idioma a otro (ejemplo 4).

Para solventar estos problemas, los sistemas de transferencia semántica utilizan representaciones intermedias más profundas. El análisis y la generación son más complejos, pero la transferencia se simplifica. Un tipo de representación usada generalmente por estos sistemas son las representaciones interlingua basadas en papeles temáticos.

Los papeles temáticos (también conocidos como "casos profundos", "papeles semánticos", "papeles caso" o "papeles theta") expresan las relaciones semánticas entre los complementos (argumentos del verbo) y el verbo. La naturaleza de la relación entre un nombre o sintagma nominal (argumento) y su verbo gobernante[11] (predicado) en una oración se puede expresar en términos de papeles tales como AGENTE, TEMA, INSTRUMENTO, etc. Estas funciones semánticas permanecen constantes entre idiomas, mientras que las funciones sintácticas (sujeto, objeto y otras) pueden cambiar.

Los papeles temáticos señalados de forma más generalizada son los que proceden de las tipologías de Gruber (1965) y Jackendoff (1972). En base a estas tipologías, Haegeman (1991) realizó la siguiente clasificación:

[11] Las relaciones de dependencia en una oración se pueden expresar mediante un *árbol de dependencias* donde los dependientes de cada gobernante se expresan como hijos de éste. El gobernante de toda la oración es el verbo principal.

- AGENTE, es el que voluntariamente causa y realiza la acción expresada por el predicado. Normalmente, el sujeto de la oración se corresponde con el AGENTE[12].

(5) Juan empujó la caja.

- TEMA, es la entidad afectada por la acción expresada por el predicado o la entidad que se mueve (con un verbo de movimiento) o la entidad cuya locación se define (con un verbo que indique locación). Generalmente, cuando el verbo es transitivo se corresponde con el objeto de la oración. Cuando el verbo es intransitivo, el sujeto de la oración es el TEMA y no desempeñará el papel de AGENTE.

(6) María colocó las carpetas en la estantería.

- EXPERIMENTANTE, es la entidad que experimenta algún estado (psicológico) expresado por el predicado o el receptor de un cierto estado psicológico.

(7) Me gustan los zumos de frutas tropicales.

- BENEFACTIVO, es la entidad que se beneficia de la acción expresada por el predicado.

(8) Le conviene aprobar.

- META, es la entidad hacia la que se dirige la actividad expresada por el predicado.

(9) Entregaron el paquete a Jesús.

- FUENTE, es la entidad de donde algo se mueve como un resultado de la actividad expresada por el predicado.

(10) Recibí un paquete de mi primo.

[12] En cada uno de los ejemplos, la expresión subrayada es la que recibe el papel temático definido previamente.

- LOCACIÓN, es el lugar en el que se sitúa la acción o estado expresado por el predicado.

(11) Andrés vive en Méjico.

- INSTRUMENTO, es el medio (material o herramienta) utilizado para alcanzar la acción.

(12) Rompió la ventana con el martillo.

- EVENTO, es la acción, proceso o estado denotado por el predicado en sí mismo.

(13) El viejo marinero caminaba torcido.

La lista anterior se puede ampliar incluyendo nuevos papeles temáticos (POSESOR, TIEMPO, etc.) o dividiendo algunos de los existentes. Por ejemplo, el papel AGENTE se puede dividir en AGENTE/CAUSA, restringiendo la etiqueta AGENTE a las entidades dotadas del rasgo semántico humano (las únicas que pueden ser sujetos voluntarios de acciones), mientras que al resto se le asignaría la etiqueta CAUSA (por ejemplo, *El terremoto destruyó numerosos edificios*). En cuanto al TEMA, éste se puede entender de una forma más restrictiva como la entidad (persona o cosa) movida por la acción expresada por el predicado, mientras que la entidad (persona o cosa) que padece o sufre la acción expresada por el predicado es el PACIENTE.

Los papeles temáticos están unidos a las características semánticas, en el sentido de que ciertos papeles son normalmente llevados a cabo por elementos que cumplen unas características semánticas determinadas. Esto puede ser general (por ejemplo los AGENTES son normalmente animados, las LOCACIONES expresan lugares físicos) o específico para los verbos o clases de verbos (por ejemplo *comer* requiere que el TEMA sea comestible, verbos que expresan *dar* requieren que la META sea una entidad animada). Estas características son importantes ya que describen la estructura sintáctica y semántica de las palabras en una oración,

y permiten hacer elecciones entre interpretaciones alternativas de palabras individuales.

Si estos papeles temáticos se usan en la representación interlingua, algunos de los problemas de la transferencia estructural desaparecen. Si consideramos las representaciones basadas en papeles temáticos para el ejemplo 4 en inglés y español (figura 2.8) se observa que son idénticas para ambos idiomas excepto en las palabras o entradas léxicas[13]. Para este ejemplo, la transferencia de un idioma a otro es prácticamente inmediata.

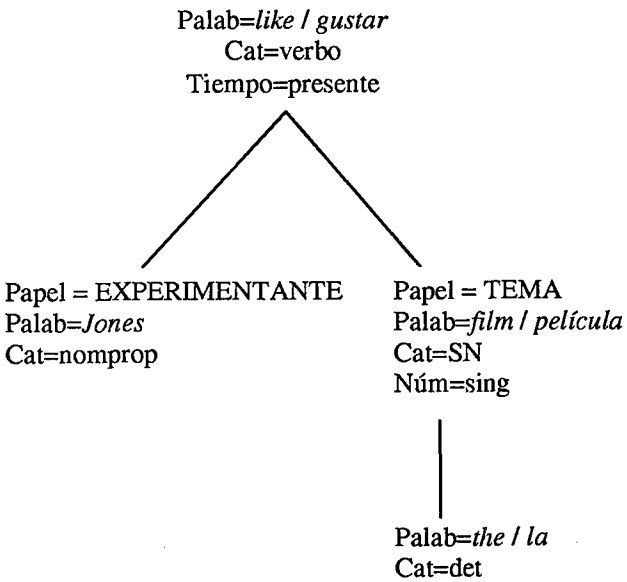


Figura 2.8. Representación basada en papeles temáticos

Como se puede apreciar en esta representación interlingua, se han eliminado las relaciones funcionales de los constituyentes (*Sujeto, Objeto Directo, Objeto Indirecto*) y ahora aparecen los papeles temáticos con sus categorías: *verbo*, *nomprop* –nombre propio–, *SN* –sintagma nominal–, *det* –determinante–, etc.

[13] Destacar que cuando se realiza el análisis con papeles temáticos se utiliza, preferentemente, un *árbol de dependencias* en vez de un árbol sintáctico.

La representación interlingua en el idioma destino obtenida tras la transferencia servirá de entrada al módulo de generación del idioma destino correspondiente que derivará la estructura superficial apropiada. En el caso de una traducción inglés-español, se seleccionará el TEMA (*la película*) como el sujeto superficial, el EXPERIMENTANTE (*Jones*) como objeto indirecto y se aseguraría la concordancia entre todos los elementos de la oración en español.

Normalmente, estas representaciones basadas en papeles adoptan una representación denominada *estructura de rasgos*. Los rasgos se pueden representar como *atributos* con sus correspondientes *valores*, normalmente denominados *pares atributo-valor*. El valor de cada atributo puede ser atómico u otra estructura de rasgos y está restringido de tal modo que un determinado atributo sólo puede tomar una serie de valores. Así, por ejemplo, podríamos citar los siguientes atributos:

- *Categoría gramatical* con los siguientes valores: nombre, adjetivo, verbo, etc.
- *Información morfológica* como *Género*, *Número*, *Tiempo*, *Persona*, etc. con sus correspondientes valores.
- *Información semántica* con valores como: humano, no humano, etc.

Si representamos de nuevo el ejemplo 4 utilizando una representación basada en una *estructura de rasgos* con papeles temáticos obtendríamos la estructura mostrada en la figura 2.9.

En general, estas estructuras interlingua usadas en la transferencia son más complejas y representan un análisis más profundo de la oración. En ellas aparecen resueltas las referencias, los verbos compuestos, el ámbito de los cuantificadores, etc. Además, en estas estructuras normalmente se eliminan (ver figura 2.9) las categorías sintácticas de los constituyentes, las unidades léxicas de las categorías inferiores (determinantes, pronombres) y el tiempo gramatical de la oración; todos ellos serán deducidos a partir de la información semántica que se encuentra almacenada en la propia estructura.

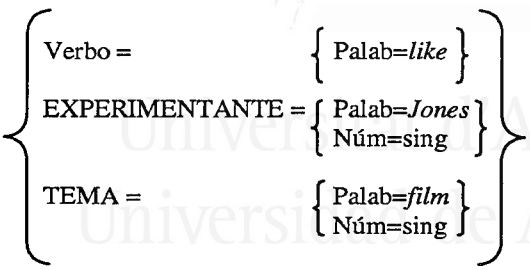


Figura 2.9. Representación basada en estructuras de rasgos con papeles temáticos

A partir de la representación interlingua basada en la estructura de rasgos en el idioma origen, y tras aplicar la transferencia léxica, se obtendrá la representación interlingua en el idioma destino. Esta nueva estructura será la entrada del módulo de generación. Podemos concluir que estas estructuras son muy cercanas a la representación interlingua "ideal", pero sólo en lo que se refiere a sus características estructurales, ya que sigue siendo necesario el proceso de la transferencia léxica bilingüe (con el tratamiento de todas sus ambigüedades).

2.7.6 Sistemas sin transferencia (interlingua)

En los sistemas interlingua, el análisis del texto origen es tan profundo que la generación se puede realizar directamente a partir del mismo sin realizar ningún tipo de transferencia. El resultado del análisis es una representación del texto independiente del idioma (lingüísticamente neutral) que servirá de base para la generación del texto en el idioma destino.

Las ventajas de los sistemas interlingua multilingües ya se han descrito en la sección 2.4.3. Los inconvenientes radican en que los módulos de análisis y generación tienen que ser totalmente independientes, es decir, el análisis no debe estar orientado hacia un idioma destino particular y la generación no debe buscar ninguna información en el texto original. Por todo ello, la representación interlingua debe incluir toda la información que podría ser requerida durante la generación del texto en cualquier idioma destino.

Este alto grado de independencia del idioma y neutralidad significa que los sistemas interlingua pretenden obtener la representación del *significado* del texto, sin tener en cuenta la estructura ni las palabras o unidades léxicas del texto.

Todas las diferencias estructurales tratadas en la sección 2.7.2: diferencias sintácticas (orden de las palabras, voz pasiva), diferencias debidas a palabras particulares o diferencias por campos semánticos (expresiones de movimiento) son resueltas cuando se encuentra una verdadera representación interlingua que contiene el significado de la oración como un todo.

Una opción frecuente consiste en adoptar la representación que es común a la mayoría de los idiomas del sistema. Por ejemplo, una posible representación interlingua neutral que resuelve las diferencias por campos semánticos entre inglés y francés para las expresiones de movimiento consiste en separar los elementos de *modo de transporte* y *dirección*.

En el ejemplo 14 aparece una expresión de movimiento en inglés y su traducción al francés (Hutchins & Somers, 1992).

(14) **I** He walked across the road.
F Il traversa la rue à pied.

En la figura 2.10 aparece la representación interlingua que se podría utilizar para las frases del ejemplo 14.

En esta figura se ha utilizado una representación interlingua formada por una estructura de rasgos basada en papeles temáticos. Para cada papel se ha definido un nuevo atributo *Pred* –Predicado– que contiene las unidades léxicas interlingua independientes del idioma (indicadas entre los símbolos "<" y ">"). Como se puede apreciar, la diferencia fundamental entre esta representación y la utilizada para los sistemas de transferencia semántica 2.7.5 radica en la representación de las unidades léxicas.

- **Representación de las unidades léxicas.** Las dificultades para encontrar representaciones neutrales para las unidades léxicas quizás exceden a las dificultades encontradas para obtener representaciones neutrales para las estructuras sintácticas. Al

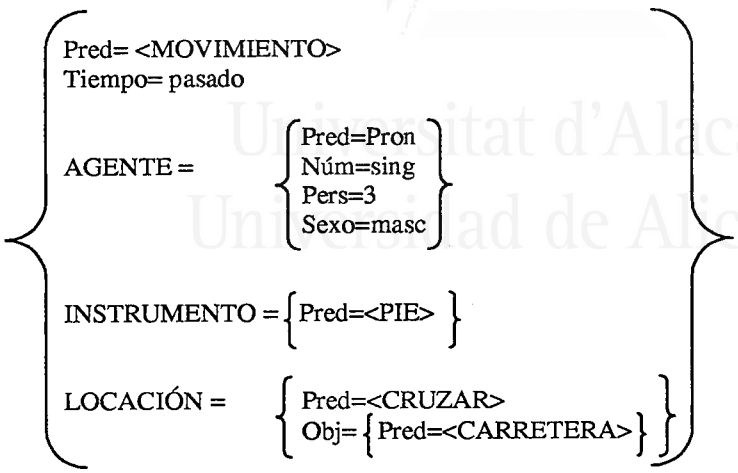


Figura 2.10. Representación interlingua para las expresiones de movimiento

igual que en las diferencias estructurales, el problema para la aproximación interlingua es doble: por una parte hay que decidir la representación neutral más apropiada y, por otra parte, hay que descubrir el método para extraer la información necesaria a partir del texto original.

Como hemos visto, en los sistemas de transferencia no hay problemas en las traducciones uno-a-uno ni en las traducciones muchos-a-uno; los problemas ocurrían sólo en las traducciones uno-a-muchos. Sin embargo, si trabajamos con un sistema interlingua multilingüe hay problemas cuando una palabra en un idioma del sistema tiene dos o más formas posibles en algún otro idioma del sistema. Para que un sistema interlingua sea completamente independiente del idioma, no debe representar las palabras en un idioma sino que debe representar unidades léxicas independientes del idioma, es decir, debe representar *conceptos*. Cualquier distinción que sea representada léxicamente en los idiomas del sistema debe ser representada explícitamente en la representación interlingua. Estas distinciones pueden reflejar las diferencias estilísticas, gramaticales y conceptuales presentadas en la sección 2.7.1.

Hay que destacar que la obtención de esta representación interlingua basada en *conceptos* es una tarea compleja incluso para idiomas muy relacionados. Por ejemplo, si el español es uno de los idiomas del sistema interlingua, según el ejemplo de ambigüedad conceptual presentado en la sección 2.7.1, se deben hacer las distinciones para *pierna* (de una persona) y *pata* (de un animal, silla o mesa), incluso cuando se traduce al inglés o alemán donde estas distinciones son irrelevantes (*leg*: *Bein*).

En estas representaciones basadas en conceptos normalmente se definen una serie de conceptos ontológicos de modo que todas las entidades del texto se definen como instancias de éstos. Las representaciones así obtenidas requieren, pues, un nivel de análisis más profundo que el utilizado en los sistemas de transferencia semántica.

Sin embargo, en los sistemas interlingua actuales se utilizan unas representaciones estructurales basadas en conceptos en las que siguen apareciendo las unidades léxicas del idioma origen. Por lo tanto, siguen reflejando las características superficiales de los idiomas particulares.

Cuando se presentan traducciones alternativas, estos sistemas extraen información del contexto o conocimiento del mundo real con el objetivo de escoger una única opción. En definitiva, utilizan representaciones más profundas que las utilizadas en la transferencia semántica y posteriormente realizan un proceso de transferencia léxica.

Se puede encontrar alguna excepción en aquellos sistemas interlingua basados en un idioma existente. Para ser un candidato, el idioma elegido debe ser no ambiguo, consistente y regular en su diccionario. En el sistema DLT (Schubert, 1988) se ha investigado el uso de una versión modificada del Esperanto, que cumple los requisitos anteriores, como idioma interlingua. Sin embargo, es dudoso que un sistema que use el Esperanto como representación interlingua pueda ser denominado estrictamente interlingua, ya que la obtención de la estructura intermedia en Esperanto a partir de otro idioma (español, inglés, francés, etc.) es, en sí misma, un sistema de TA.

2.8 Módulo de Generación de un sistema de TA

En esta sección presentaremos el último módulo que interviene en todo proceso de TA: el módulo de generación. Para presentarlo, se analizarán las tres estrategias que siguen las aproximaciones básicas: generación en sistemas directos, generación en sistemas de transferencia y generación en sistemas interlingua.

2.8.1 Generación en los sistemas directos

Si recordamos la arquitectura de los sistemas directos y el modo de realizar la transferencia de estos sistemas (sección 2.7.3), podemos afirmar que no se produce una generación en el idioma destino propiamente dicha. Si comparamos los sistemas directos con los sistemas de transferencia y los sistemas interlingua es muy difícil establecer dónde finaliza el análisis del texto origen y dónde comienza la generación del texto destino. En los sistemas directos hay una fase influenciada por el idioma destino (sustitución de las palabras origen por las palabras destino con el diccionario bilingüe) y otra fase influenciada por el idioma origen (la "reordenación local").

La fase de la "reordenación local" se puede considerar como una mezcla de transferencia y generación. Es un proceso de transferencia ya que las estructuras del texto en el idioma origen son transformadas en estructuras de texto en el idioma destino. También es un proceso de generación, ya que la salida de esta fase es una secuencia de unidades léxicas en el idioma destino. Como ya vimos en la sección 2.7.3 un ejemplo de "reordenación local" consistía en reemplazar la secuencia inglesa *adjetivo + nombre* por la secuencia española *nombre + adjetivo*.

La fase final de "generación" del texto en el idioma destino consiste básicamente en asegurar la correcta formación de las palabras en el idioma destino:

- Terminación de las palabras (ejemplo 15):

(15) establecer + presente + 1ª persona + singular → establezco

- Concordancia morfológica entre las palabras (ejemplo 16):

(16) barco + rojo + masculino + plural → barcos rojos

Esta fase es la única que se puede considerar auténtica generación, ya que trabaja con información del idioma destino y no tiene en cuenta el texto en el idioma origen.

La estrecha unión que existe, pues, entre la transferencia y la generación en los sistemas directos es una de las principales desventajas de los sistemas directos. La falta de modularidad en estos sistemas implican que, una vez que están en funcionamiento, sea muy difícil introducir nuevas mejoras al sistema global. Por esta razón, los sistemas directos más modernos han incrementado la modularidad de sus componentes y se puede distinguir alguna separación entre los módulos de transferencia y generación. Como ejemplos de estos sistemas podemos citar el sistema Systran (Toma, 1977) y el sistema Météo (Chandioux, 1976).

2.8.2 Generación en los sistemas de transferencia

En los sistemas de transferencia la fase de generación se divide normalmente en dos módulos: "generación sintáctica" y "generación morfológica".

1. En la fase de *generación sintáctica* se toma como entrada la representación intermedia obtenida tras el análisis y la transferencia. Esta representación es muy cercana a la estructura profunda del texto origen[14] y en ella aparecen las palabras expresadas en el idioma destino. Esta estructura se transforma mediante "reglas transformacionales" en una nueva que representa la estructura superficial del texto en el idioma destino, y que contiene las funciones gramaticales y características de

[14] Esta estructura normalmente en TA tiene la forma de un árbol de análisis sintáctico.

las palabras en el idioma destino. La tarea fundamental de la generación sintáctica consiste en ordenar los constituyentes en la secuencia correcta en el idioma destino.

2. La estructura superficial obtenida tras la fase de *generación sintáctica* es, a su vez, la entrada de la *generación morfológica.* En esta fase se interpretan las unidades léxicas y toda su información asociada para generar el texto destino. Normalmente, esto se realiza mediante *reglas morfológicas.* En el ejemplo 17 se observa la regla morfológica para el verbo irregular en inglés *eat*:

   (17) eat + pasado → ate

   En el ejemplo 18 se muestra la regla morfológica para la primera persona singular del pretérito imperfecto del verbo español *conducir*:

   (18) conducir + pretérito imp. + 1ª persona + singular → conduje

2.8.3 Generación en los sistemas interlingua

El proceso de generación en un sistema interlingua consta, al igual que en los sistemas de transferencia, de dos fases: generación sintáctica y generación morfológica. La principal diferencia entre ambos sistemas radica en el punto de partida para la generación. Mientras que los sistemas de transferencia toman como entrada para la generación una representación sintáctica de la estructura profunda del texto en el idioma destino, los sistemas interlingua realizan la generación a partir de una estructura que contiene el significado del texto. Normalmente, un sistema interlingua utiliza una representación basada en *estructuras de rasgos* (sección 2.7.5).

Para poder aplicar las fases de la generación sintáctica y morfológica, se debe obtener previamente la estructura profunda del texto a partir de la representación interlingua. Esto se consi-

gue mediante una fase conocida normalmente como *generación semántica.*

A continuación se muestra el proceso completo de generación en un sistema interlingua con el ejemplo 19 de una frase en inglés para estudiar cómo se realiza la correspondiente generación en español[15].

(19) Jones likes the film.

En la figura 2.11 se muestra la representación interlingua del ejemplo 19. Se han utilizado como unidades léxicas interlingua las palabras en inglés para simplificar la representación.

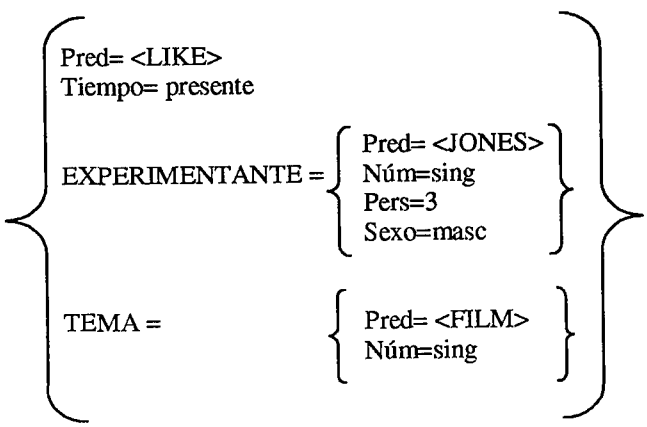


Figura 2.11. Representación interlingua de *Jones likes the film*

A partir de la representación interlingua de la figura 2.11 se aplica la fase de la generación semántica para obtener la estructura profunda (figura 2.12). Para ello, se debe seleccionar un verbo en el idioma destino (español, en este caso) como núcleo de la oración. En el ejemplo, el núcleo de la oración es el verbo *gustar* (*like*). A la vez, el tiempo del verbo (*presente*) se añade como rasgo al nodo principal de la estructura profunda de la oración.

[15] Este ejemplo se corresponde con la frase 4.i ya presentada anteriormente.

Por último, las estructuras de rasgos correspondientes a EXPERIMENTANTE (*Jones*) y TEMA (*the film*) se convierten en los correspondientes sintagmas nominales (SN) con las funciones de Sujeto (*Jones*) y OD (*la película*) respectivamente, conteniendo las palabras en español.

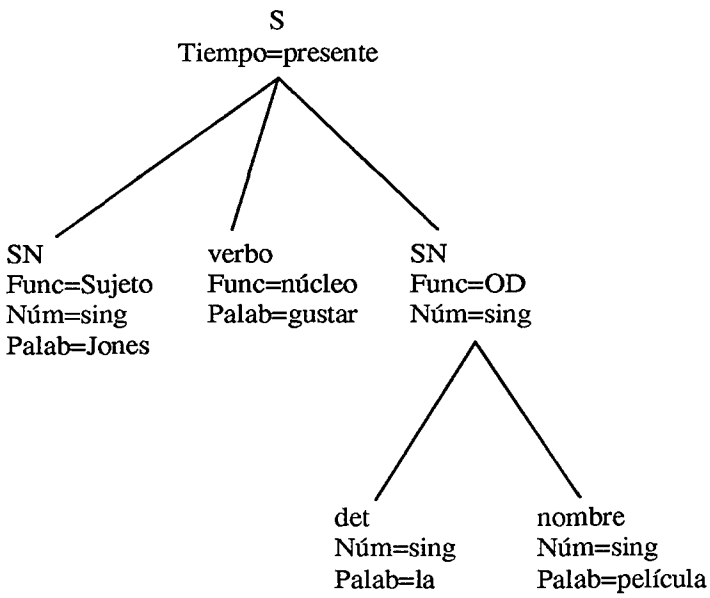


Figura 2.12. Estructura profunda de *Jones likes the film*

En la siguiente fase de generación sintáctica, se derivará la estructura superficial apropiada para el español. Primero se examina el nodo principal para seleccionar una forma verbal apropiada en español. En este caso es *presente*[16] y la estructura verbal no cambia. Sin embargo, el verbo *gustar* es intransitivo en español, por lo que el OD (*la película*) se selecciona como sujeto superficial y el Sujeto (*Jones*) como objeto indirecto. Este objeto indirecto

[16] Si fuera, por ejemplo, un verbo en voz pasiva habría que añadir el/los correspondiente/s verbo/s auxiliar/es.

en español lleva asociado un pronombre (*le*) y debe concordar en número.

Además en esta fase, se debe distribuir la información de género y número a los nodos terminales correspondientes. Por ejemplo, los determinantes y adjetivos deben concordar en género y número con su nombre gobernante en un SN; el verbo debe concordar en número con el SN Sujeto, etc.

Tras aplicar todos estos pasos intermedios se obtiene la estructura superficial de la frase en español. En la figura 2.13 se observa esta representación.

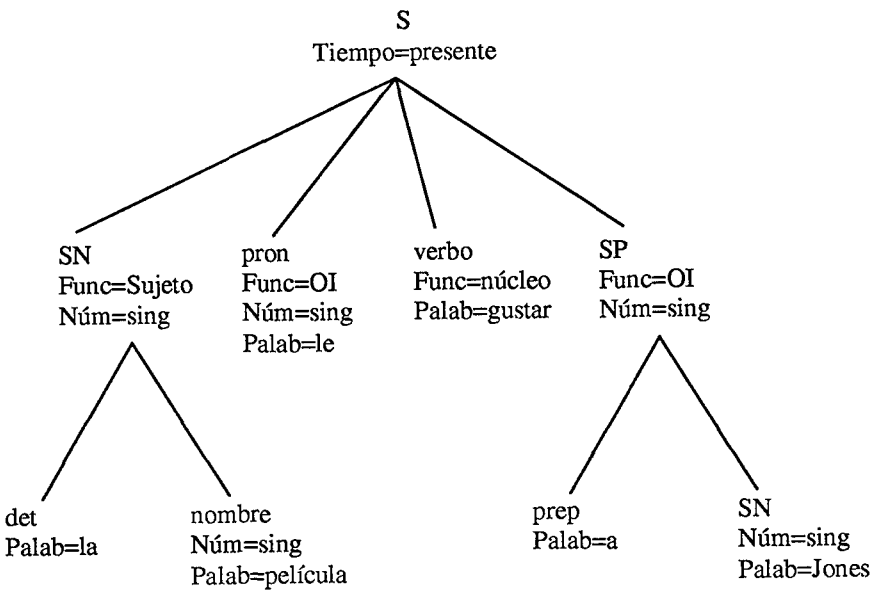


Figura 2.13. Estructura superficial de *La película le gusta a Jones*

Por último, en la fase de generación morfológica se aplica la regla morfológica correspondiente (ejemplo 20) para generar correctamente la forma verbal.

(20) gustar + presente + 3ª persona + singular
→ gusta

Como resultado global del proceso de generación se obtendría la siguiente frase en español: *La película le gusta a Jones.*

3. Sistemas de TA

Una vez que se han presentado en los capítulos anteriores los conceptos fundamentales de la TA, explicando las estrategias de los sistemas de TA y sus componentes esenciales, en este capítulo presentaremos los sistemas de TA actuales más relevantes. Para ello, mostraremos las características principales de sistemas que siguen las tres estrategias básicas: sistemas directos, sistemas de transferencia y sistemas interlingua.

3.1 Sistemas directos

3.1.1 Systran

Aunque el sistema Systran (Toma, 1977; van Slype & Pigott, 1979; Whitelock & Kilby, 1983; Wheeler, 1987) está funcionando desde hace más de treinta años, se puede considerar como uno de los mejores ejemplos de los sistemas directos de TA.

Los orígenes del Systran se pueden encontrar en las primeras aproximaciones para TA a finales de los años 50. Su diseñador, Peter Toma (un lingüista investigador), empezó su trabajo en 1957 en el Instituto de Tecnología de California. Pocos años más tarde, Toma era el principal programador del sistema GAT (traducción ruso-inglés) de la Universidad de Georgetown. Posteriormente, en 1968, Toma fundó una empresa en La Jolla, California, con un producto denominado Systran (System Translation). La empresa fue contratada para desarrollar un sistema ruso-inglés de TA para el Ejército del Aire norteamericano. El primer sistema Systran fue instalado y evaluado en 1969 en la base aérea Wright-Patterson (Dayton, Ohio). Desde 1970, el sistema ha seguido trabajando para el Ejército del Aire.

Systran fue utilizado por la NASA en el conocido proyecto conjunto (americano-ruso) Apollo-Soyouz durante 1974 y 1975. Este hecho permitió que Toma, en 1975, realizara una demostración de un prototipo de Systran para la traducción inglés-francés a representantes de la Comisión de las Comunidades Europeas. Como resultado, se formalizó un contrato para desarrollar versiones que permitieran la traducción entre los idiomas de la Comunidad Europea.

El gran auge que tuvo Systran permitió que numerosas compañías quisieran promocionar y desarrollar versiones para nuevos idiomas. El número de versiones que traducen pares de idiomas crecen cada año. Como ejemplos podríamos citar los siguientes: inglés-francés, inglés-alemán, inglés-japonés, inglés-ruso, inglés-español, inglés-italiano, inglés-portugués, alemán-francés, alemán-italiano, alemán-español, etc.

En 1996, Systran firmó un contrato con el Ejército del Aire norteamericano para desarrollar versiones que tradujeran idiomas de Europa oriental. Este contrato incluía el primer sistema de TA entre el serbo-croata y el inglés.

La tecnología de Systran también se ha usado en aplicaciones comerciales. En 1996, Systran firmó otro contrato con SEIKO mediante el cual proporcionaría datos lingüísticos y software a los productos de traducción electrónicos de SEIKO.

Systran hizo historia en 1997 con la ayuda del servicio de traducción de Altavista (BABELFISH[1]) impulsado por Systran. BABELFISH introdujo la TA en la comunidad de Internet ofreciendo traducciones gratis en tiempo real a cualquier usuario de Internet.

El diseño de Systran. La facilidad del Systran de crear versiones para nuevos pares de idiomas es debida al alto grado de modularidad que hay en el diseño del sistema.

El proceso de traducción se divide en cuatro fases: preprocesamiento, análisis, transferencia y generación. Cada una de estas fases usa un conjunto de diccionarios bilingües que contienen información léxica, gramatical y semántica. Estos diccionarios cons-

[1] http://www.babelfish.altavista.com (página visitada el 14/02/01).

tituyen el principal componente del sistema, el resto de componentes lo constituye un conjunto de programas.

- **Los diccionarios.** Las bases de datos léxicas para Systran están constituidas por el diccionario principal de raíces y un conjunto de diccionarios contextuales.

1. *Diccionario principal de raíces.*
   Es un diccionario bilingüe de entradas formadas por palabras simples. En este diccionario aparece cada palabra en el idioma origen en su forma base (raíz) con información morfológica, sintáctica y semántica: categoría gramatical, concordancia, transitividad, marcadores semánticos, etc. y una traducción a la forma base equivalente en el idioma destino que irá acompañada por la información gramatical necesaria para su generación.
   Las palabras homógrafas con distintas categorías sintácticas tienen diferentes entradas individuales, mientras que las palabras homógrafas con la misma categoría serán tratadas por los diccionarios contextuales.
   Por último, destacar que en este diccionario a cada palabra en el idioma origen le corresponde una única palabra en el idioma destino. Ésta es la traducción por defecto que permanecerá intacta mientras no sea modificada por el resto de diccionarios contextuales.
2. *Diccionario de expresiones.*
   Trata con las expresiones invariantes del idioma origen.
3. *Diccionario de semántica limitada.*
   Define el ámbito de las relaciones sintácticas dentro de los sintagmas nominales e identifica unidades léxicas formadas por varias palabras (nombres compuestos).
4. *Diccionario de homógrafos.*
   Incluye la información contextual necesaria para la resolución de determinados homógrafos.
5. *Diccionario analítico.*
   Contiene las acepciones a las reglas sintácticas generales que hay que aplicar a palabras particulares.
6. *Diccionario de semántica condicional.*

Interviene en las fases de transferencia y generación y hace la selección de la palabra en el idioma destino.

- Los programas. Hay dos tipos de programas en Systran, los de sistema y los de traducción. Los *programas de sistema*, escritos en ensamblador, son independientes del idioma e incluyen los programas para el preprocesamiento. Los *programas de traducción*, escritos en un lenguaje de alto nivel, están divididos en análisis, transferencia y generación y son diferentes según el idioma a tratar. Systran usa una estructura de datos lineal formada por una secuencia de registros para cada palabra de una oración. Ésta contiene información gramatical, la traducción y las relaciones gobernante-dependiente (ver sección 2.7.5) de cada palabra.

Las distintas etapas que llevan a cabo cada uno de estos programas en el proceso de traducción son los siguientes:

1. Entrada (input). Un programa carga el texto e identifica información del formato (títulos, párrafos, etc.)
2. Búsqueda en el diccionario de expresiones. Identificación de las expresiones invariantes.
3. Búsqueda en el diccionario principal de raíces. Las palabras restantes del texto se buscan en el diccionario principal de raíces y se almacena su información asociada.
4. Análisis morfológico. Lleva a cabo la identificación de combinaciones potenciales de raíces y terminaciones. También se aplica a aquellas palabras que no se han encontrado en el diccionario principal con el objetivo de inferir información gramatical y de su categoría.
5. Acceso al diccionario de semántica limitada. Identificación de los nombres compuestos.
6. Resolución de homógrafos. Se lleva a cabo examinando las categorías gramaticales de las palabras adyacentes. Sólo se realiza una pasada para cada oración.
7. Separación de las oraciones en principal y subordinadas. Se realiza buscando signos de puntuación, conjunciones, pronombres relativos, etc.

8. Identificación de las relaciones sintácticas primarias. Análisis parcial que identifica determinados constituyentes y relaciones entre ellos: nombre y modificadores, verbo y objetos, etc.
9. Identificación de estructuras coordinadas.
10. Identificación de sujeto y predicado de las oraciones. Se hace en un proceso relativamente sencillo: tras identificar los verbos, se definen como sujetos potenciales los nombres o pronombres que no se han identificado como objetos. Además las oraciones se marcan como declarativas, interrogativas o imperativas.
11. Identificación de las *relaciones profundas*. Por ejemplo, identificación de los sujetos gramaticales en las oraciones pasivas.
12. Transferencia léxica de "expresiones condicionales". En esta fase se traducen aquellas palabras que tienen traducciones alternativas bajo ciertas condiciones. Estas condiciones suponen análisis estructurales de fases anteriores y están definidas en el diccionario de semántica condicional.
13. Traducción de las preposiciones que no han sido tratadas en la fase anterior.
14. Transferencia estructural usando "rutinas léxicas".
15. Asignación de la traducción por defecto (en el diccionario principal de raíces) para cada palabra que no haya sido modificada durante la transferencia por los diccionarios contextuales.
16. Generación morfológica. Se toma como base la información estructural de caso, género, número, etc. de etapas anteriores e información acerca de flexiones y restricciones de dependencia.
17. Generación de las palabras en el idioma destino en su orden correspondiente.

Las cinco primeras etapas corresponden a la fase de preprocesamiento, de la etapa seis hasta la once corresponden a la fase de análisis, de la etapa doce hasta la catorce corresponden a la fase de transferencia, y las tres últimas etapas corresponden a la fase de generación.

En cuanto a la resolución y generación de la anáfora pronominal, Systran realiza un tratamiento de los pronombres omitidos en español –cero pronombres con función de sujeto (ver sección 4.2.1)– para su posterior generación en inglés. También trata la

anáfora pronominal intersentencial (ocurre cuando la expresión anafórica y su antecedente no aparecen en la misma oración) eligiendo como antecedente el sintagma nominal de la oración anterior (normalmente el más cercano). Posteriormente, ésta será generada en el idioma destino. Por último, mencionar que aunque el sistema trata estas referencias pronominales, los resultados obtenidos no son muy satisfactorios.

Características de la estrategia de traducción. Aunque Systran es un sistema directo, es evidente (a partir de las etapas del proceso de traducción) que podría ser clasificado como un sistema de transferencia. Sin embargo, hay una serie de razones que impiden que sea caracterizado como un sistema de transferencia propiamente dicho:

- No hay separación de los datos lingüísticos en bases de datos monolingües para análisis y generación. Systran, por el contrario, conserva la característica de los sistemas directos de grandes diccionarios bilingües a los que se accede en varias etapas de análisis y generación.
- No hay una clara separación entre la transferencia y la generación. Por ejemplo, en la generación encontramos procesos de transferencia léxica (etapa 15) y transferencia estructural (etapa 17). El único módulo monolingüe en todo el proceso de TA es la generación morfológica (etapa 16), característica típica de los primeros sistemas directos.
- No se realiza un análisis completo de las oraciones. Systran sólo identifica ciertas relaciones entre los constituyentes de las oraciones (nombre + adjetivo + artículo, adverbio + verbo, sujeto del verbo, etc.). El análisis estructural es parcial y selectivo, por lo que no se establece una representación de las relaciones de dependencia de todos los constituyentes de la oración. En definitiva, no existe ningún modelo lingüístico sobre el que se fundamente el análisis del texto.

Podemos concluir que aunque se han realizado numerosos esfuerzos para introducir un mayor grado de homogeneidad y modularidad, Systran todavía conserva las características básicas de un sistema directo de primera generación.

3.1.2 Météo

En 1965 se creó en la Universidad de Montreal el grupo de investigación CETADOL cuya investigación se basaba en el tratamiento automático de los datos lingüísticos. Por estas fechas, el gobierno canadiense introdujo su política bilingüe, según la cual todos los documentos oficiales debían aparecer en inglés y francés. Las demandas de servicios de traducción crecieron considerablemente, por lo que se comenzaron a subvencionar proyectos de TA.

El mencionado grupo (CETADOL) comenzó a investigar en TA y cambió su nombre por TAUM (Traducción Automática en la Universidad de Montreal). Aunque inicialmente el grupo TAUM había adoptado una aproximación basada en la transferencia había estudiado las ventajas de una aproximación basada en un "sublenguaje" para resolver las dificultades semánticas de la traducción. En 1975, firmó un contrato para desarrollar un sistema que tradujera los partes metereológicos de inglés a francés. En 1976 se desarrolló un prototipo, el sistema Météo (Chandioux, 1976; Chevalier *et al.*, 1978; Chandioux, 1989), y desde 1977 está en funcionamiento proporcionando los partes metereológicos diarios para prensa y televisión. Posteriormente, en 1989 se puso en funcionamiento la versión que permitía realizar la traducción inversa (francés-inglés).

El diseño de Météo. El sistema Météo adoptó un diseño esencialmente directo. No había necesidad de un módulo de transferencia debido, en parte, a que el estilo de los informes en inglés y francés eran estructuralmente muy parecidos. Tampoco se incluía un módulo de análisis morfológico ya que la variación morfológica de las palabras también estaba restringida. Por último, el análisis sintáctico también se simplificaba debido a las restricciones: no hay referencias pronominales, ni oraciones de relativo ni pasivas.

El proceso de traducción (inglés-francés) de los partes metereológicos en el sistema Météo consta de cinco fases: preprocesamiento, búsqueda en los diccionarios, análisis sintáctico en inglés, generación sintáctica en francés y, por último, generación morfológica en francés. Al finalizar todas las fases, la intervención humana sólo es necesaria si se ha producido algún error (provoca-

do principalmente por palabras desconocidas o errores en el texto de entrada). A continuación se muestran estas fases:

1. Preprocesamiento.
   Los informes metereológicos son recibidos en inglés en un formato estándar: un código de cabecera, el origen del informe, una lista de regiones sobre las que se hace la predicción, la previsión propiamente dicha y un código de terminación del informe. El vocabulario de los informes es fijo y está restringido a un conjunto de frases de la cabecera, nombres de lugares y descripciones de las condiciones metereológicas. Por todo ello, es muy fácil identificar las palabras desconocidas (normalmente son producidas por errores en el texto original).
   A partir de este informe se identifican las unidades de traducción y se transforman en el formato "Q-system" (Colmerauer, 1970). Para ello se separan las palabras, oraciones, etc. por una serie de códigos y se identifican las abreviaturas y los acrónimos.
2. Búsqueda en los diccionarios.
   Los datos lingüísticos de Météo lo forman tres diccionarios bilingües para expresiones, nombres de lugares y vocabulario en general (incluye terminología metereológica).
   La traducción comienza con la extracción de los datos léxicos. El *diccionario de expresiones* se usa para traducir expresiones inglesas formadas por más de una palabra a las que les corresponde una única palabra francesa.
   El *diccionario de lugares* contiene aquellos nombres propios que difieren en inglés y francés. Los nombres que no cambian de un idioma a otro no se incluyen, ya que el programa los tratará como palabras desconocidas y, al ser nombres propios, no se traducirán.
   El *diccionario principal* contiene todas las formas morfológicas. Cada entrada para una palabra inglesa contiene la palabra francesa equivalente, la categoría gramatical, características semánticas e información morfológica francesa.
3. Análisis sintáctico en inglés.

De un estudio realizado se dedujo que en las frases de los informes metereológicos sólo aparecen cinco tipos de árboles sintácticos. Por lo tanto, el análisis sintáctico del sistema Météo consiste en descubrir el árbol de análisis particular para una frase determinada.
Durante el análisis se van creando todas las posibles soluciones parciales. Éstas se van podando en fases posteriores hasta que sólo queda un análisis. El análisis sintáctico usa una técnica de análisis ascendente y se realiza en tres etapas:

a) En la primera etapa se reconocen las fechas, horas y grados de temperatura.
b) La siguiente etapa identifica los sintagmas nominales restantes, es decir, aquéllos que expresan condiciones metereológicas. En esta etapa se consulta tanto las categorías sintácticas como las características semánticas de las palabras.
c) En la última etapa se reconocen estructuras complejas.

El resultado de la fase de análisis sintáctico es un único árbol que tiene palabras inglesas en los nodos terminales.

4. Generación sintáctica en francés.
   La tarea de la generación sintáctica consiste en la derivación del orden de las palabras en francés a partir de la representación estructural y de la información que contiene cada palabra.
5. Generación morfológica en francés.
   En esta fase se asegura la correcta terminación de las palabras en francés, en particular de los adjetivos.

Aunque aparentemente Météo se puede considerar como un sistema de segunda generación, en el proceso de la traducción no existe una separación clara de las operaciones de análisis, transferencia y generación. Por otra parte, Météo conserva la característica principal de los sistemas directos: la transferencia léxica se realiza antes de la etapa de análisis sintáctico.

En cuanto al tratamiento de las referencias pronominales, hay que mencionar que el estilo de los partes metereológicos era muy sencillo y no incluía expresiones anafóricas pronominales, consecuentemente éstas no eran tratadas por el sistema.

3.2 Sistemas de transferencia

3.2.1 SUSY

A mediados de los 60 la investigación sobre TA en la Universidad de Saarlandes (Saarbrücken, Alemania) había comenzado con el desarrollo de un analizador para alemán. En 1967, comenzó un proyecto más importante, en el que se exploraba la posibilidad de adaptar el sistema Systran para la traducción ruso-alemán. Este intento falló[2] y el grupo comenzó a trabajar en su propio prototipo. Este prototipo ruso-alemán fue el comienzo de un sistema multilingüe denominado SUSY (Maas, 1977; Luckhardt, 1982; Hutchins, 1986; Maas, 1987), en el que se trabajaba con los siguientes idiomas: alemán (idioma principal), ruso, inglés, francés y esperanto.

El diseño de SUSY. Debido a las múltiples modificaciones sufridas por el sistema SUSY provocadas por los avances en TA, presentaremos el diseño del sistema SUSY-I.

SUSY data de mediados de 1980 y fue programado en Fortran. Es básicamente un sistema de transferencia con fases de análisis y generación monolingües y una fase bilingüe en la que se realiza la transferencia léxica y estructural. Las estructuras usadas por el sistema se basan en árboles de dependencias en las que se identifican las relaciones gobernante-dependiente para cada palabra. La entrada del sistema puede ser opcionalmente pre-editada y no hay un proceso interactivo de post-edición.

El diseño del sistema es totalmente modular. Los distintos módulos se aplican secuencialmente. En general, los módulos de análisis y generación son específicos para cada idioma, mientras que los módulos de transferencia están diseñados para pares de idiomas específicos. Los componentes del sistema son, pues, los distintos módulos y el conjunto de diccionarios.

---

[2] Debido principalmente a la falta de modularidad del sistema original ruso-inglés. En la versión original de Systran, el análisis ruso estaba muy dirigido hacia la posterior generación en inglés.

- Los diccionarios.

1. *Diccionarios monolingües.*
   Existen diccionarios monolingües para los idiomas origen y destino que contienen esencialmente información morfológica y sintáctica de las unidades léxicas. Cada idioma tiene diccionarios separados para análisis y generación. Existen tres diccionarios monolingües en SUSY:
   a) Diccionario con las palabras de mayor frecuencia.
   b) Diccionario de raíces.
   c) Diccionario de expresiones.
2. *Diccionarios semánticos monolingües.*
   Existen diccionarios monolingües para los idiomas origen y destino que contienen características semánticas y condiciones sintácticas y semánticas para tratar la polisemia (en particular las preposiciones). Estos diccionarios se usan en las etapas de análisis y generación, aunque a diferencia de los diccionarios anteriores, el mismo diccionario monolingüe es usado en ambas etapas.
3. *Diccionarios bilingües de transferencia.*
   Estos diccionarios contienen las equivalencias léxicas con alguna información del contexto estructural.

- Los módulos. Las fases de análisis, transferencia y generación del sistema SUSY son llevadas a cabo por una serie de módulos que pasamos a detallar:

1. Pre-edición. SUSY permite la posibilidad de la pre-edición de los textos origen para insertar códigos especiales que ayuden al análisis. Si se realiza, este proceso permite la mejora del rendimiento del sistema. De cualquier modo, el módulo de análisis funciona igualmente con los códigos especiales o sin ellos.
2. Entrada de texto. Este módulo es una etapa de preprocesamiento no lingüístico. En él se lee el texto en el idioma origen, se separa en palabras y oraciones y se genera la estructura (árbol de dependencia) de cada palabra. La información del proceso de pre-edición, si existió, también se incluirá en la estructura generada.

3. Búsqueda en los diccionarios y análisis morfológico. En este módulo se buscan en los diccionarios monolingües las palabras del texto origen, excepto aquéllas marcadas como nombres propios en la pre-edición.
   También se incluye un módulo de análisis morfológico que intenta descomponer las palabras en raíz y afijos. La salida de este módulo proporciona para cada palabra sus posibles raíces (será desambiguada en un módulo posterior). Si no se ha podido encontrar la raíz para una palabra determinada, ésta será tratada como "palabra desconocida". Por último, tras este proceso se realiza la identificación de expresiones invariantes.
4. Desambiguación de palabras homógrafas. La salida del módulo anterior es una secuencia de palabras que han sido buscadas en el diccionario, muchas de las cuales son ambiguas en categoría. En este módulo se intentan resolver estas ambigüedades.
   Cuando una palabra es ambigua en su categoría, en una primera etapa se eliminan las combinaciones imposibles de secuencias de categorías. Las categorías restantes se ordenan según una tabla de "probabilidades" y "compatibilidades". La primera mide la probabilidad de que dos categorías ocurran en una misma secuencia. La segunda es una medida similar, pero tiene en cuenta las probabilidades mutuas, es decir, cuando hay dos palabras ambiguas juntas, la probabilidad de la solución para una depende de su compatibilidad con la solución para la otra. Los valores de estas dos medidas fueron inicialmente estimados por lingüistas y han sido ajustados paulatinamente durante un período de tiempo en respuesta a traducciones incorrectas del sistema.
5. Segmentación de la oración. El objetivo principal de este módulo es separar la oración en oración principal y resto de cláusulas (oraciones subordinadas, aposiciones, etc.). Para ello, se identifican los signos de puntuación, las conjunciones subordinadas y los pronombres relativos.
   Tras la salida del segmentador, se realizan dos procesos para identificar sintagmas nominales, sintagmas preposicionales y núcleos verbales (verbos simples, compuestos, auxiliares modales, etc.).

6. Análisis estructural. El objetivo de este módulo es determinar la estructura de valencias[3] de cada cláusula. Para ello, intenta averiguar cuáles de los sintagmas nominales identificados previamente se corresponden con los complementos de la palabra principal en la cláusula, por ejemplo, sujeto y objeto para un verbo transitivo con valencia 2. El resto de elementos son analizados como adverbiales o aposiciones. Además, el módulo intenta reconstruir estructuras elípticas y averiguar los sujetos profundos de las oraciones.
7. Desambiguación semántica. Los procedimientos de este módulo están basados en las características semánticas asignadas a los nombres y a algunos pronombres en los diccionarios semánticos. Estas características son de dos clases: universales (humano, abstracto, animado, etc.) y específicas (lugar, profesión, animal, planta, etc.).
   El propósito de este módulo es asignar una interpretación a las estructuras sintácticas que son semánticamente ambiguas. Esto se realiza aplicando una serie de reglas que expresan preferencias semánticas. Otro tipo de desambiguación semántica de este módulo consiste en la distinción de homógrafos de la misma categoría.
8. Transferencia. La transferencia se realiza mediante una secuencia de procesos, usando los diccionarios bilingües para sustituir las formas léxicas en el idioma origen por sus equivalentes en el idioma destino. Normalmente, la estructura de dependencias se conserva, aunque excepcionalmente puede ser alterada. Este módulo también contiene un subproceso que intenta traducir las palabras identificadas como desconocidas.
9. Generación semántica. Se generan las expresiones y las preposiciones correctas para verbos y nombres.
10. Generación sintáctica. Genera una secuencia de elementos léxicos (raíces de las palabras) que incluyen la información morfológica necesaria para la generación morfológica. Parte

[3] La *valencia* es el número de complementos que tiene un verbo determinado. Además, proporciona información acerca de la naturaleza de los complementos del verbo. Por ejemplo, los verbos monovalentes son intransitivos y sólo tienen un argumento: el sujeto. Los verbos divalentes tienen dos argumentos y pueden ser transitivos con sujeto y objeto o sujeto y sintagma preposicional.

de la tarea de este módulo es determinar la estructura superficial en el idioma destino. Además, resuelve las terminaciones de los nombres, tratamiento de las palabras compuestas y la generación de los pronombres y adjetivos posesivos intrasentenciales[4].

11. Generación morfológica. El objetivo de este módulo es convertir la raíz de las palabras y su información morfológica asociada en cadenas de texto. Éste es un proceso relativamente sencillo, aunque hay algunos idiomas (como el alemán) en los que las formas morfológicas de las palabras dependen del contexto que les rodea. Por último, se tratan las mayúsculas y los signos de puntuación.

Los siete primeros módulos corresponden a la fase de análisis, el octavo módulo corresponde a la fase de transferencia y los tres últimos módulos corresponden a la fase de generación.

Respecto a la resolución de las anáforas pronominales, SUSY plantea un mecanismo (en la etapa de la generación sintáctica) para tratar la anáfora intersentencial usando un sistema de ponderación, según el cual cada antecedente tiene un peso dependiendo de su posición y función en las oraciones anteriores. Una vez resueltas las anáforas pronominales, éstas se generarán convenientemente en el idioma destino.

3.2.2 Ariane

El grupo de investigación de la Universidad de Grenoble liderado por Bernard Vauquois –quizás una de las personas más relevantes en la historia de la TA– comenzó a trabajar en la TA al principio de los 60.

Inicialmente se concentró en el desarrollo de formalismos para expresar la información lingüística y los algoritmos para su posterior implementación. La traducción se realizaría mediante una sucesión de "representaciones lingüísticas". El sistema original, desarrollado entre 1960 y 1970, usaba una estrategia inter-

[4] Tanto los pronombres como los adjetivos posesivos deben concordar en género con sus antecedentes.

lingua para tres pares de idiomas: ruso-francés, alemán-francés y japonés-francés. De todos ellos, destacaba el sistema ruso-francés.

Posteriormente, en 1971 la introducción de las nuevas tecnologías en los ordenadores, entre otras razones, llevó al grupo de Grenoble a rediseñar su sistema interlingua original y desarrollar un sistema de trasferencia, oficialmente conocido como Ariane (Boitet & Nédobejkine, 1981; Vauquois & Boitet, 1985; Boitet, 1987; Boitet, 1989), aunque a menudo es denominado como sistema GETA (el nombre del grupo francés).

La investigación principal se centró en la traducción ruso-francés aunque paulatinamente se fueron añadiendo nuevos idiomas: portugués, malayo, chino, etc. Además de ser un buen ejemplo de un sistema de "segunda generación", Ariane animó al establecimiento de proyectos de TA por todo el mundo (particularmente en Asia). Los primeros trabajos del proyecto Eurotra (Allegranza *et al.*, 1991) fueron influenciados en sus inicios por el grupo de Grenoble, y posteriormente por los trabajos del grupo de Saarbrücken –sistema SUSY (Maas, 1977)–.

El sistema ha sufrido grandes modificaciones desde su primer prototipo dando lugar a diferentes versiones. A continuación describiremos la versión Ariane-78.

El diseño de Ariane. Ariane es un sistema de transferencia con tres módulos principales: análisis, transferencia y generación. Los módulos de análisis y generación a su vez se dividen en módulos morfológicos y sintácticos. El módulo de transferencia se divide en transferencia léxica y estructural. Ariane utiliza un formalismo de reglas de propósito específico[5] para cada uno de los módulos y una estricta separación del conocimiento lingüístico y algorítmico en cada etapa.

El sistema es multilingüe en el sentido que los formalismos y los algoritmos que los implementan son independientes de los idiomas a ser tratados. Además, desde un punto de vista lingüístico, los programas de análisis y generación para un par de idiomas pueden ser reutilizados por otros idiomas origen o destino indistintamente.

[5] Estas reglas son parecidas a las reglas de producción o reglas de reescritura de una Gramática Independiente del Contexto (Context Free Grammar, CFG).

Los niveles de representación usados en Ariane son interesantes desde el punto de vista de la Lingüística, ya que usan *estructuras multinivel* que combinan las relaciones de dependencia y las estructuras de los constituyentes, conteniendo información lingüística profunda y superficial.

Las *estructuras multinivel* combinan información de distintos niveles: morfológico, sintáctico y "lógico-semántico". En el nivel morfológico, se almacenan la información morfológica y léxica de las palabras. En el nivel sintáctico, se almacenan la información de los constituyentes –sintagma nominal, núcleo verbal– y las relaciones sintácticas superficiales. Por último, en el nivel "lógico-semántico" se almacenan las relaciones de dependencia de una forma lógica, por ejemplo, predicados y argumentos con sus papeles temáticos (agente, locación, etc.).

El flujo del proceso de traducción sigue una estratificación lingüística estándar: al texto origen se le aplica un módulo de análisis morfológico obteniendo un árbol etiquetado. El análisis multinivel proporciona una estructura intermedia en el idioma origen, que sufre dos etapas de transferencia: en la transferencia léxica, las unidades léxicas origen se reemplazan por las correspondientes unidades léxicas en el idioma destino, proporcionando una estructura en el idioma origen con unidades léxicas en el idioma destino. La transferencia estructural produce una estructura intermedia en el idioma destino que sirve de entrada para la generación sintáctica y morfológica.

La arquitectura del sistema Ariane se basa en los distintos procesos lingüísticos (análisis, transferencia y generación) y los programas que implementan los formalismos.

- **Los programas.** Ariane tiene cuatro formalismos y sus implementaciones asociadas. Estos programas corresponden a los cuatro tipos diferentes de procesamiento de datos lingüísticos. Ésta es una característica importante de Ariane ya que cada uno de los formalismos está específicamente diseñado para facilitar un determinado tipo de procesamiento. Los cuatro programas son los siguientes:

1. ATEF está diseñado para el análisis morfológico y convierte las cadenas de caracteres en grupos de características (*estructuras de rasgos*) que se organizan en un árbol etiquetado.
2. ROBRA es una herramienta muy potente que manipula estructuras complejas de árboles y proporciona unas estructuras de árboles nuevas.
3. TRANSF es un formalismo menos potente (usado para la transferencia léxica) que sólo altera las etiquetas de las hojas de los árboles.
4. SYGMOR (usado en la generación morfológica) tiene la función de convertir los árboles etiquetados en cadenas de caracteres.

- Los procesos lingüísticos.

1. Análisis morfológico.
   En primer lugar, el texto original se somete a una fase de análisis morfológico, usando las reglas escritas en el formalismo ATEF. Como resultado se obtiene las etiquetas de cada nodo terminal, obtenidas tras la segmentación del texto y la búsqueda de las palabras en el diccionario. A menudo se obtienen resultados ambiguos.
   La salida de esta fase es un conjunto de árboles etiquetados, donde cada árbol tendrá un padre que domina al resto de nodos (todas las palabras de la frase) y no hay constituyentes intermedios. Si hay palabras ambiguas, se representarán todas las posibles combinaciones en árboles diferentes.
2. Análisis multinivel.
   En esta fase se utiliza el formalismo ROBRA. Se realiza una combinación de dos procesos: "análisis" para construir una estructura de árbol sintáctico inicial con constituyentes intermedios y una etapa "transformacional" en la que las estructuras de árbol sintácticas superficiales se transforman en representaciones más profundas (por ejemplo, se unen las unidades léxicas que forman expresiones y las funciones sintácticas como sujeto y objeto se reemplazan por las relaciones lógico-semánticas).
3. Transferencia.

La salida del análisis multinivel es un árbol que muestra las relaciones de dependencia, los constituyentes, la información lógico-semántica, así como las funciones sintácticas superficiales. La siguiente etapa del proceso de traducción es la transferencia léxica en la que las unidades léxicas en el idioma origen son sustituidas por las correspondientes en el lenguaje destino. A menudo la elección de las unidades léxicas depende del contexto de las palabras.

La estructura obtenida se pasa a la transferencia estructural que realiza las alteraciones estructurales necesarias. La entrada son estructuras en el idioma origen (con unidades léxicas en el idioma destino ya insertadas), mientras que la salida serán las correspondientes estructuras profundas en el idioma destino. En esta etapa se utiliza de nuevo el formalismo ROBRA para realizar los cambios estructurales.

4. Generación.

   Durante la fase de generación, la siguiente etapa es la generación sintáctica en la que se asignan las etiquetas de la estructura superficial, es decir, la elección de sujetos y objetos superficiales, la selección de los verbos auxiliares apropiados, la elección del orden correcto de las palabras y el establecimiento de los valores morfológicos de las palabras, es decir, concordancia de género y número entre los distintos componentes. De nuevo ROBRA se usa durante esta fase.

   Por último, en la generación morfológica, el formalismo SYGMOR tiene la función de convertir los árboles etiquetados generados en la fase de generación sintáctica en cadenas de caracteres, incluyendo los signos de puntuación.

   El tratamiento de las referencias pronominales se lleva a cabo en esta etapa y permitirá la correcta generación de las mismas en el idioma destino.

Proyectos relacionados. Tras el prototipo Ariane-78 se realizaron una serie de mejoras dando lugar a las versiones Ariane-85 y Ariane-G5 (Boitet, 1997). Basándose en esta última versión, se han desarrollado una serie de proyectos colaterales:

- El proyecto LIDIA es un sistema interactivo para usuarios monolingües que permite la traducción del texto en el idioma origen (francés) al texto en el idioma destino (alemán o ruso) sin tener conocimientos del idioma destino (Boitet & Blanchon, 1994).
- El proyecto UNL se centra en la comunicación interpersonal multilingüe a través de Internet (Boitet, 1997).
- El proyecto C-STAR es un sistema de TA de diálogos (Blanchon *et al.*, 1999).

3.2.3 Eurotra

Eurotra (Varile & Lau, 1988; Allegranza *et al.*, 1991; Bech *et al.*, 1991) se puede considerar como uno de los proyectos más grandes de TA por dos razones: el número de personas involucradas en él y la amplia distribución geográfica de los grupos de investigación implicados. El proyecto Eurotra estableció los principios básicos y los requerimientos lingüísticos y computacionales del diseño de un sistema de TA multilingüe complejo, los cuales servirán para el desarrollo de futuros sistemas de TA.

En la década de los 70, la política multilingüe de la Comunidad Europea incrementó el número de peticiones de traducciones entre los distintos idiomas de la Comunidad. Por ello, en 1976 se utilizó el sistema Systran inicialmente para inglés-francés. Sin embargo, desde el principio se reconocieron las limitaciones de Systran como un sistema multilingüe. En 1978, se acordó llevar a cabo un proyecto para la creación de un sistema de TA de diseño avanzado (Eurotra) capaz de tratar con todos los idiomas oficiales de la Comunidad. Estos idiomas eran danés, holandés, inglés, francés, alemán e italiano. Griego, portugués y español serían añadidos posteriormente. Por lo tanto, el proyecto Eurotra realizaría las traducciones entre nueve idiomas, es decir, 72 combinaciones de pares de idiomas.

En 1978 se tomó una decisión crucial: Eurotra sería un sistema de transferencia. Aunque algunos proponían que con 9 lenguajes (72 pares de traducciones) sería conveniente elegir una aproximación interlingua, sin embargo, hay que recordar que por estas

fechas se opinaba que las mejores perspectivas para los avances en TA radicaban en el desarrollo de sistemas de transferencia.

El proyecto, tras su aprobación por el Consejo de Ministros en 1982, comenzó a desarrollarse en tres fases. En la primera de ellas (1982-84), se estableció la organización global formada por grupos de investigación de cada estado miembro y se definieron las bases lingüísticas y las especificaciones de los programas.

La segunda fase (1984-88), se centró en la investigación lingüística necesaria para el desarrollo de un pequeño prototipo diseñado para un corpus particular.

La tercera fase (1988-90), consistía en la ampliación del sistema para que fuera capaz de tratar con textos más complejos. El resultado final no consistía en crear un sistema completo, sino un prototipo que pudiera ser la base de futuros sistemas.

Tras estas tres fases y varios informes de la Comisión Europea en los que se concluía que Eurotra constituía un prototipo de definición científica en vez de un prototipo pre-industrial, se pasó a una nueva etapa –conocida como *programa de transición*–.

Esta nueva etapa se centra en el establecimiento de las bases para el desarrollo de un prototipo industrial que mejore el rendimiento lingüístico y computacional del sistema anterior. Además, promueve la investigación científica sobre las bases teóricas de TA en una estructura multilingüe.

Para presentar el sistema, no describiremos el prototipo sino las especificaciones de un sistema potencial parcialmente implementado.

Organización y diseño del sistema Eurotra. La traducción en el sistema Eurotra se realiza mediante la transformación del texto origen al texto destino en varias fases correspondientes a los módulos del sistema. En las tres fases básicas, *análisis* es la transformación del texto origen en una representación que refleja el texto origen, *transferencia* es la transformación de la representación anterior en una representación que refleja el texto destino y *generación* es la transformación de ésta en el texto destino.

Con 72 combinaciones en el sistema global, es obvio que la transferencia ha de ser lo más sencilla posible. Como consecuencia, los módulos de análisis y generación son más complejos que en los

sistemas de transferencia originales. Debido a esta complejidad, el análisis y la generación se realizan en varios niveles. Estos niveles intermedios de representación son los siguientes:

- Estructura de Texto Eurotra (*ETS*, Eurotra Text Structure).
  Contiene el texto de entrada original, incluyendo códigos, datos no textuales y diagramas.
- Texto Normalizado Eurotra (*ENT*, Eurotra Normalised Text).
  Contiene el texto de entrada del que se han eliminado los datos no textuales y códigos, obteniendo un fichero ASCII.
- Estructura Morfológica Eurotra (*EMS*, Eurotra Morphological Structure).
  Es una representación de las palabras y morfemas en una secuencia de árboles etiquetados. Contiene información léxica y morfológica.
  En esta etapa no se han reconocido constituyentes intermedios en los árboles etiquetados. Las ambigüedades reconocidas durante el análisis morfológico se resolverán en fases posteriores.
- Estructura de Constituyentes Eurotra (*ECS*, Eurotra Constituent Structure).
  Es una representación de la estructura sintáctica superficial basada en las relaciones de las categorías sintácticas.
  En esta etapa se realiza la transmisión de las características sintácticas de las hojas a los nodos dominantes apropiados, por ejemplo, información de género, número, etc.
- Estructura Relacional Eurotra (*ERS*, Eurotra Relational Structure).
  Es una representación de las relaciones gramaticales superficiales (sujeto, objeto, etc.) junto con las categorías sintácticas. Además refleja las relaciones de dependencia de gobernante, dependiente, etc.
- Estructura Interfaz (*IS*, Interface Structure).
  Es una representación basada en la dependencia semántica, incorporando *estructuras de rasgos* y características semánticas (animado, humano, etc.).
  Estas representaciones son la entrada y la salida del proceso de transferencia. La transferencia de un idioma a otro consiste,

básicamente, en la sustitución de las unidades léxicas del idioma origen al idioma destino.
Hay que destacar que la definición de estas representaciones en un sistema multilingüe tan complejo es crucial. Por esta razón, se estableció un conjunto de principios generales sobre el contenido y el nivel de abstracción de la información lingüística que debían contener.

En todos estos niveles intermedios de representación se utilizan los distintos componentes del sistema. Éstos son los siguientes:

- **Formalismo gramatical.** Aunque no se puede considerar como un componente del sistema, es el mecanismo en el que se basan los distintos procesos del sistema. El formalismo gramatical usado por Eurotra se denomina E-estructura –E-framework (Bech & Nygaard, 1988)– y es un formalismo gramatical basado en la unificación (Shieber, 1986). A diferencia de otros, no ha sido diseñado para tratar con análisis sintáctico superficiales; E-estructura es un mecanismo más general que trata los objetos lingüísticos basados en *estructuras de rasgos* y características.

- **Objetos y estructuras.** En cada nivel intermedio, las representaciones se componen de *objetos* primitivos y *estructuras* construidas a partir de estos objetos.

Los objetos son estructuras de rasgos, donde las características o rasgos son pares atributo-valor. Los posibles valores de las características y las posibles combinaciones de los atributos son definidos por una "teoría de rasgos" para cada nivel de representación.

Las estructuras se definen mediante reglas que expresan las dos propiedades estructurales de *dominio* (relación padre-hijo) y de *precedencia* (ordenación entre los hermanos).

- **Traductores y generadores.** La transformación de los objetos y las estructuras de un nivel de representación a otro es llevada a cabo por una serie de procesos denominados *traductores* y *generadores* que tienen un conjunto de reglas basadas en el formalismo E-estructura.

Respecto al tratamiento de fenómenos lingüísticos, Eurotra puede manejar una gran variedad de tipos de oraciones: oraciones

subordinadas, aposiciones, construcciones con participio, etc. Sin embargo, otros problemas tales como la elipsis, negación, ámbito de los cuantificadores o resolución de la anáfora pronominal no se han abordado. En el futuro se prevé la investigación sobre estos temas y su incorporación al prototipo final.

Proyectos relacionados. Las investigaciones llevadas a cabo en el proyecto Eurotra crearon las bases para el desarrollo de nuevos sistemas experimentales de TA:

- En el sistema CAT2 (Sharp, 1988) se implementaba un nuevo formalismo gramatical para un sistema de TA multilingüe.
- El sistema multilingüe Mimo (Arnold & Sadler, 1990; van Noord *et al.*, 1990) surge como un sistema de transferencia basado en las gramáticas de unificación. Utiliza una estructura interlingua basada en formas lógicas para realizar la transferencia –sistema de transferencia con una estructura interlingua (ver sección 2.7.5)–. En la gramática de unificación de este sistema se definen las relaciones de transferencia entre formas lógicas de dos idiomas distintos.
- El proyecto MULTILINT (Reuther, 1998) se centra en el tratamiento y la generación de documentos multilingües en el dominio de la industria del automóvil.

3.2.4 METAL

Los orígenes del sistema METAL (Bennet & Slocum, 1985; Slocum, 1987; Thurmair, 1990; Alonso, 1990) hay que buscarlos en el establecimiento del Centro de Investigación de Lingüística (Linguistics Research Center, LRC) en la Universidad de Texas (Austin, Texas). Las investigaciones desarrolladas en inglés y alemán impulsaron la propuesta de desarrollar un sistema de transferencia bidireccional. Esta fase de investigación inicial finalizó en 1968.

Una segunda fase de la investigación en este centro, desde 1970 hasta 1975, se centró en la exploración de una aproximación interlingua para alemán e inglés. El diseño interlingua obtenido era interlingua sólo en el sentido de las representaciones sintácticas: la traducción de las unidades léxicas se llevaba a cabo mediante reglas de transferencia léxica.

En 1978 comenzó una nueva fase financiada por la compañía Siemens. La motivación de esta compañía para la creación de un sistema de TA era doble: por una parte la necesidad de incrementar la productividad de su propio servicio de traducción y, por otra parte, el deseo de producir un sistema de traducción para otros clientes potenciales.

A partir de este momento, el sistema METAL cambió de un diseño interlingua a una aproximación de transferencia, y no fue pensado para funcionar de un modo totalmente automático sino para aumentar las facilidades de edición de textos y el acceso a grandes bases de datos con terminología específica (en particular la propia de Siemens). En 1989, el primer sistema METAL apareció en el mercado para la traducción alemán-inglés. Posteriormente, a la versión alemán-inglés le sucedieron las versiones inglés-alemán, holandés-francés, francés-holandés, alemán-español, alemán-francés y alemán-danés.

A continuación describiremos la arquitectura del sistema de la versión inicial alemán-inglés.

El diseño de METAL. El proceso de traducción en el sistema METAL se lleva a cabo en las siguientes etapas:

1. Adquisición del texto. El texto en el idioma origen se proporciona al sistema.
2. Formateado del texto. Del texto original se elimina toda la información no textual como diagramas, gráficos, etc. y se insertan las marcas correspondientes para su posterior reinserción en el idioma destino.
   Esta etapa incluye la identificación de las unidades léxicas y de las oraciones completas.
3. Pre-análisis. Búsqueda de las palabras en los diccionarios obteniendo tres listas:
   a) Lista de palabras desconocidas.
   b) Lista de nombres compuestos desconocidos, que incluye las traducciones posibles al inglés.
   c) Lista de las palabras técnicas conocidas, que el usuario posteriormente comprobará manualmente.

4. Diccionarios. Existen diccionarios monolingües para los idiomas origen y destino, conteniendo información morfológica, sintáctica y semántica. Estos diccionarios están diseñados para ser neutrales e independientes y pueden ser empleados durante el análisis y la generación por cualquier idioma involucrado en la traducción.
   Los diccionarios de transferencia bilingües están diseñados para la traducción en un sentido entre un par de idiomas específicos.
5. Programas de traducción.
   a) Análisis. Incluye análisis léxico y morfológico: extracción de raíces potenciales y afijos para cada una de las palabras de la oración.
      Posteriormente, se realiza un análisis sintáctico (teniendo como base la gramática del sistema) que produce distintas estructuras alternativas ordenadas según una puntuación obtenida dependiendo de sus constituyentes. En esta etapa no se usan características semánticas.
      En la mayoría de los casos la ambigüedad de las palabras homógrafas se resuelve por preferencias léxicas.
      Como resultado, se selecciona el árbol sintáctico con mayor puntuación. Si no se puede obtener un análisis completo de una oración, la transferencia se realizaría con análisis parciales de fragmentos de la misma.
   b) Transferencia. La entrada a este módulo son representaciones sintácticas que incluyen asignaciones de papeles temáticos y algunas características semánticas.
      Se aplican las reglas de transferencia léxica (de los diccionarios bilingües) y las reglas de transferencia estructural (reglas gramaticales). Ambos tipos de reglas interactúan entre ellas.
      La salida del módulo son representaciones superficiales con la especificación completa del orden de las palabras y los constituyentes morfológicos.
   c) Generación. Como consecuencia de la fase de transferencia la etapa final de la generación se limita a producir palabras en el idioma destino que son correctas morfológicamente.

6. Reformateado del texto. Mezcla de la información textual con los datos no lingüísticos.
7. Post-edición. Revisión de la traducción realizada. Se puede llevar a cabo oración por oración o con segmentos más grandes, estando el texto original presente o no (en una parte de la pantalla).
8. Salida. Composición final del texto en el idioma destino.

La gramática usada en el sistema METAL consiste en un conjunto de reglas de producción aumentadas por condiciones y especificaciones de las estructuras de salida. Las reglas incluyen morfología flexiva así como estructuras sintácticas. Además, combinan operaciones para ser realizadas durante el análisis con operaciones que se realizarán durante la transferencia.

La resolución de las anáforas intrasentenciales e intersentenciales se lleva a cabo en METAL durante una fase denominada Integración. Esta cuarta fase se lleva a cabo entre el Análisis y la Transferencia y tiene como objetivo la resolución de estos problemas lingüísticos. Ya que METAL incorpora una fase más a las típicas utilizadas en una estrategia de transferencia, normalmente se define como un sistema de "transferencia modificado".

3.2.5 Ntran

En el Instituto de Ciencia y Tecnología de la Universidad de Manchester (University of Manchester Institute of Science and Technology, UMIST) se realizaron los primeros estudios sobre sistemas de TA para usuarios monolingües. El primer prototipo fue el sistema Ntran (Whitelock *et al.*, 1986; Wood & Chandler, 1988) para inglés-japonés. En este sistema, las ambigüedades de la transferencia léxica al japonés eran resueltas por un usuario que no tenía conocimientos de japonés. Para ello, las posibles ambigüedades del idioma destino (japonés) que podían ocurrir durante la transferencia léxica, se presentaban en un modo interactivo al usuario inglés con un texto en el idioma origen (inglés). Con este mecanismo, un usuario inglés que no tuviera conocimientos de japonés podría resolver todas las ambigüedades producidas durante la transfe-

rencia. Finalmente, la generación en japonés era totalmente automática.

Proyectos relacionados. Basados en el éxito logrado con Ntran, los investigadores de la Universidad de Manchester trabajaron en otros proyectos relacionados con la construcción de sistemas de TA para usuarios monolingües:

- Proyecto UMIST-British Telecom (Jones & Tsujii, 1990). El objetivo de este proyecto era el desarrollo de un sistema de TA que permitiera la composición de cartas de negocios en un idioma desconocido por el usuario mediante una serie de menús con distintas opciones.
  El sistema está basado en la idea de textos "pro-forma" para ciertos tipos de cartas de negocios (propuesta, solicitud de información, queja, etc.). Los textos pro-forma son plantillas que contienen diferentes campos para introducir nombres, fechas, direcciones, etc. El usuario introduce estos datos en su propio idioma. Una vez que el pro-forma se ha rellenado, el sistema generará el texto en el idioma destino comparando la información introducida por el usuario con una base de datos de fragmentos de texto pre-traducidos. La gran ventaja de este sistema consiste en la buena calidad del texto generado en el idioma destino para un dominio determinado. Esta característica se debe, fundamentalmente, a que la base de datos está formada por textos escritos por traductores humanos.
- Proyecto UMIST-ATR (Somers *et al.*, 1990). Es un proyecto llevado a cabo entre UMIST y el laboratorio japonés ATR (Advanced Telecommunications Research). El dominio del proyecto es una conversación on-line (normalmente por teléfono) entre una oficina en Japón y un usuario inglés que solicita información. El objetivo del proyecto es que el sistema actúe como intermediario entre los dos interlocutores y traduzca el diálogo entre inglés y japonés. Hay que destacar que la traducción del diálogo es una tarea compleja debido a la gran cantidad de construcciones elípticas, referencias anafóricas, frases incompletas, etc. que contiene.

Al igual que ocurre en el proyecto UMIST-British Telecom, el sistema tiene una base de datos (*modelo de diálogo* bilingüe) e interactúa con el usuario para intentar emparejar su pregunta con el rango de posibilidades existentes en el modelo. Ya que estos sistemas contienen una base de datos con fragmentos de textos pre-traducidos pertenecen al género de los sistemas basados en ejemplos.
Entre otros proyectos que trabajan en la traducción de diálogos podemos citar experimentos desarrollados en British Telecom Research Laboratories para la traducción de diálogos telefónicos de negocios entre inglés, alemán, francés y español, el sistema VERMOBIL (Maier & Glashan, 1994; Alexandersson *et al.*, 1995) que permite la traducción de diálogos de negocios entre dos personas en inglés, alemán y japonés, etc.

3.2.6 Candide

Desde 1989 se están llevando a cabo unas investigaciones en el centro de investigación Thomas J. Watson de IBM (Laboratorios de Investigación de Yorktown Heights, NY) para el desarrollo de un sistema de traducción automática basado casi exclusivamente en técnicas estadísticas. Así, se ha desarrollado el sistema Candide (Berger *et al.*, 1994) que usa las técnicas estadísticas (Brown *et al.*, 1990) como única herramienta para análisis y generación.

La investigación de IBM se basa en un gran corpus (Canadian Hansard) que contiene debates parlamentarios en inglés y francés. El corpus para el experimento contenía 40.000 oraciones en cada idioma. La base del método consiste en la alineación automática de las oraciones en los dos idiomas. Una vez alineadas, se calculan las probabilidades de que una palabra en una oración de un idioma corresponda a dos, una o ninguna palabras en la oración traducida en el idioma destino. Las probabilidades se estimaban observando los bigramas (dos palabras consecutivas) de cada oración inglesa y su oración francesa correspondiente.

Se calcularon dos tipos de probabilidades. Por una parte, se calculó para cada palabra inglesa las probabilidades de sus posibles palabras francesas correspondientes. Por ejemplo, la palabra

inglesa *the* corresponde a la palabra francesa *le* con una probabilidad de 0,610, a la palabra francesa *la* con una probabilidad de 0,178, etc. Por otra parte se calcularon las probabilidades de que dos, una o ninguna palabras francesas correspondan a una única palabra inglesa. Por ejemplo, *the* corresponde a dos palabras francesas con una probabilidad de 0,004, a una con 0,871 y a ninguna con 0,124.

Tras una evaluación del primer prototipo para la traducción de inglés a francés se obtuvo un 48% de éxito[6]. Se podría mejorar este porcentaje con el uso de un corpus más grande, aplicando segmentación probabilística de las oraciones, usando trigramas y bigramas e incluyendo los datos sobre la flexión de un grupo de palabras de una manera conjunta. En este prototipo no se trataban las expresiones anafóricas pronominales.

Independientemente de los resultados obtenidos por el sistema, la importancia de esta investigación se basa en la posibilidad de desarrollar un sistema de TA bilingüe unidireccional sin utilizar análisis lingüístico.

3.2.7 InterNOSTRUM

En la Universidad de Alicante se está llevando a cabo un proyecto (vigente desde noviembre de 1998) para desarrollar un sistema de TA del castellano a las variantes estándar del catalán y el inverso correspondiente. En este proyecto –denominado interNOSTRUM (Canals-Marote *et al.*, 2001; Canals *et al.*, 2000; Garrido *et al.*, 1999)– están participando investigadores del departamento de Lenguajes y Sistemas Informáticos y del Instituto Interuniversitario de Filología Valenciana de la mencionada universidad[7].

El diseño de interNOSTRUM. El proceso de traducción en interNOSTRUM se realiza mediante una serie de módulos o subprogramas basados en transductores de estados finitos que llevan a cabo las distintas etapas de la traducción del texto origen.

[6] En la evaluación del prototipo del sistema realizada en 1993 (Berger *et al.*, 1994), estos resultados fueron mejorados y se alcanzaba un 62% de éxito.

[7] El proyecto está financiado por la Caja de Ahorros del Mediterráneo (CAM) y la Universidad de Alicante.

- Los subprogramas. Todos los subprogramas utilizados en interNOSTRUM se generan automáticamente a partir de los datos lingüísticos correspondientes, usando una serie de programas compiladores (Canals *et al.*, 2000; Garrido *et al.*, 1999). Ésta es una de las principales ventajas del sistema, ya que se puede aplicar fácilmente a otros idiomas. Los subprogramas usados son los siguientes:

1. Subprograma de análisis morfológico.
   El subprograma de análisis morfológico se genera automáticamente a partir de un diccionario morfológico del idioma origen que contiene las raíces de las palabras, los paradigmas de flexión y las conexiones entre ellos.
2. Subprograma de desambiguación léxica categorial.
   Realiza la desambiguación de las palabras con ambigüedad categorial usando un modelo basado en trigramas (secuencias de tres categorías léxicas). Este modelo está basado en las frecuencias observadas para estos trigramas en un corpus de entrenamiento. El subprograma asigna una probabilidad a cada posible desambiguación de la palabra ambigua (según los trigramas obtenidos) y se elige la mayor de todas.
3. Subprograma de consulta del diccionario bilingüe.
   Este subprograma se genera automáticamente a partir de un fichero que contiene las correspondencias bilingües. Es invocado por el subprograma de tratamiento de patrones.
4. Subprograma de transferencia (tratamiento de patrones).
   Este subprograma reconoce patrones sintácticos sobre el texto (en la representación proporcionada por el analizador morfológico). Se genera automáticamente a partir de un fichero de reglas que especifican los patrones y las acciones asociadas. Traduce el fragmento de texto reconocido por el patrón, utilizando el subprograma de consulta del diccionario bilingüe, y opera sobre él según indique la regla correspondiente.
5. Subprograma de generación morfológica.
   Obtiene un texto legible en el idioma destino a partir de la representación abstracta del texto proporcionada por el sub-

programa de transferencia, es decir, genera la forma superficial de la palabra a partir de la forma léxica destino.

6. Subprograma de post-generación (post-generador).
   Los apóstrofes y guiones de las palabras en el idioma destino se tratan en este subprograma que normalmente se encuentra inactivo. El post-generador se genera a partir de reglas sencillas para el tratamiento de apóstrofes, guiones y pronombres.

- El proceso de traducción. InterNOSTRUM es un sistema de transferencia morfológica avanzada[8] que realiza la traducción en seis etapas:

1. Análisis morfológico. Se realiza el análisis morfológico de las palabras del texto origen utilizando el diccionario morfológico del idioma origen.
   La entrada de esta etapa está constituida por las formas superficiales del texto origen y la salida por las formas léxicas consistentes en raíz, categoría léxica e información de la flexión.
2. Desambiguación categorial de los homógrafos. Se utiliza la información estadística sobre la aparición conjunta de categorías léxicas en textos del idioma origen (bigramas y trigramas) para desambiguar los homógrafos.
3. Consulta del diccionario bilingüe. Se utiliza un diccionario de correspondencia de raíces en el idioma origen y raíces en el idioma destino. La consulta se lleva a cabo utilizando la raíz y la categoría gramatical elegida de la palabra.
4. Tratamiento de patrones. Se detectan y tratan las secuencias de palabras que constituyen ámbitos de concordancia o que se deben de reordenar (mediante la aplicación de las reglas correspondientes).
5. Generación morfológica. Generación de las formas superficiales del texto destino. Se usa información sobre el idioma destino análoga a la que usa el análisis morfológico sobre el idioma origen.

[8] Es definido así por sus autores ya que realiza un análisis morfológico y algunas operaciones de análisis sintáctico parcial para resolver problemas de la transferencia (ambigüedad léxica categorial, concordancia, orden de las palabras, etc.).

6. Post-generación. Tratamiento de los apóstrofes y guiones de las formas superficiales.

Las dos primeras etapas constituyen la fase de análisis, las dos siguientes constituyen la fase de transferencia y el resto corresponden a la fase de generación.

Por último, mencionar que InterNOSTRUM no aborda el problema de la resolución de las expresiones anafóricas pronominales.

3.2.8 Sistema de TA multilingüe (Grupo IXA)

En el Grupo de Investigación IXA[9] de la Universidad del País Vasco se está trabajando en la actualidad con un prototipo de TA multilingüe para el euskara, castellano e inglés basado en transferencia –Díaz de Ilarraza *et al.* (2001; 2000b; 2000a)–. Este prototipo traduce sintagmas nominales y sintagmas preposicionales de textos reales integrando diversas herramientas y recursos de amplia cobertura.

En el trabajo de Díaz de Ilarraza *et al.* (2000a) se presentó el primer prototipo para el par inglés-euskara. Este primer prototipo incluía: el módulo de análisis del inglés, los módulos de transferencia, el lexicón inglés-euskara y módulos de generación sintáctica y morfológica del euskara.

Actualmente se está trabajando con el par castellano-euskara que implica el desarrollo de los trabajos siguientes: el módulo de análisis del castellano, los módulos de transferencia y el lexicón castellano-euskara[10].

El diseño del sistema. El diseño del sistema sigue el esquema típico de un sistema de transferencia y consta de tres fases: análisis, transferencia y generación. En estas fases se hace uso de una serie de datos lingüísticos separados claramente de los programas o procesos de traducción.

[9] http://ixa.si.ehu.es (página visitada el 05/03/01).

[10] Los módulos de generación del prototipo inglés-euskara son totalmente portables a este sistema.

- Bases de datos lingüísticas. Las bases de datos lingüísticas están constituidas por los diccionarios monolingües, la base de datos léxica para el vasco, los lexicones bilingües y las reglas de la gramática.

1. *Diccionarios monolingües.*
   Los diccionarios monolingües para español e inglés se usan en los módulos de análisis.
2. *Base de datos léxica para el vasco.*
   La base de datos léxica ha sido desarrollada por el Grupo IXA y contiene 75.000 entradas correspondientes a lemas y afijos. Cada entrada tiene asociadas sus características lingüísticas: categoría, subcategoría, caso, número, etc.
3. *Lexicones bilingües.*
   Los lexicones bilingües contienen información acerca de la palabra en el idioma origen y su equivalente en euskara. En cada entrada del lexicón se distinguen cinco rasgos:
   a) Entrada en el idioma origen.
   b) Categoría en el idioma origen.
   c) Equivalente en euskara: Representa su forma superficial. Esta información será utilizada cuando no sea necesaria la generación morfológica.
   d) Categoría euskara. Aporta la información requerida para la fase de generación sintáctica. Contiene la categoría y eventualmente la subcategoría del equivalente en euskara.
   e) Segmentación morfológica: Aporta la información necesaria para la generación morfológica.
4. *Reglas de la gramática.*
   Las reglas gramaticales usan información de la categoría, subcategoría y lema de las palabras para decidir el orden correcto de las mismas en sintagmas nominales y preposicionales. Estas reglas se ejecutan durante la etapa de la generación sintáctica.

- Los procesos de traducción. Las distintas etapas que se llevan a cabo en el proceso de traducción son las siguientes:

1. Análisis de los sintagmas nominales y preposicionales. Para el inglés se utiliza el analizador morfológico ENGCG[11]. Para el castellano, se han utilizado los analizadores y etiquetadores MACO, TACAT y RELAX (Atserias *et al.*, 1998).
2. Conversión a una estructura de árbol. La representación intermedia será una estructura tipo árbol que contiene la siguiente información: valor léxico, información morfológica y sintáctica y los índices de sus nodos padre e hijos.
3. Transferencia léxica. En esta etapa se hace uso del lexicón bilingüe, recogiendo en cada nodo de la representación intermedia la información asociada a su equivalente en euskara.
   El sistema es parametrizable y permite decidir la traducción correcta de una palabra cuando hay más de una alternativa en el diccionario. Se puede llevar a cabo de una forma interactiva en la que se pregunta al usuario para que decida la palabra adecuada. De otro modo, se selecciona la primera traducción que aparece en el diccionario (la más usada).
4. Transferencia estructural. En esta etapa los nodos sin valor léxico desaparecen y la información sobre declinación y postposición se transfiere a los nodos hijos.
5. Generación sintáctica. Usando las reglas gramaticales el sistema establece el orden correcto de las palabras en los sintagmas nominales y preposicionales.
   Las reglas se aplican recursivamente con el objetivo de agrupar varios nodos del árbol en un nuevo nodo que contiene las palabras en el orden correcto. Este proceso se repite hasta que el árbol se reduce a un único nodo que contiene todas las palabras ordenadas correctamente. Por último, la información acerca de la declinación se transfiere a la última palabra del sintagma.
6. Generación morfológica. En la última etapa, el sistema llama al analizador/generador del euskara con la información de la segmentación morfológica para generar la forma de las palabras que tengan información sobre declinación, ya que el resto de las palabras aparecerán con el lema.

[11] Este analizador se encuentra disponible en la web en la dirección http://www.lingsoft.fi/cgi-pub/engcg (página visitada el 05/03/01).

Las dos primeras etapas corresponden a la fase de análisis, la etapa tres y cuatro corresponden a la fase de transferencia y las dos últimas etapas corresponden a la fase de generación.

Respecto a la resolución de la anáfora pronominal (intrasentencial e intersentencial) hay que mencionar que el primer prototipo del sistema no resuelve fenómenos lingüísticos de este tipo. Actualmente, la investigación se centra en la traducción de sintagmas nominales y preposicionales y no trata –por el momento– la traducción de oraciones completas.

3.2.9 Episteme

En el Grupo de Investigación Julietta de la Universidad de Sevilla se están abordando desde hace varios años una serie de proyectos relacionados con TA. Los sistemas JULIETTA (Amores, 1992) y Lekta (Amores *et al.*, 1994) se pueden considerar las bases de Episteme (Quesada & Amores, 2000; Amores & Quesada, 1997).

Episteme se puede describir como el núcleo de un entorno para el desarrollo de sistemas de TA. Episteme no es un traductor automático, sino un entorno o "shell" que permite el desarrollo de traductores automáticos. El sistema cuenta con un conjunto de lenguajes de especificación mediante los que se introduce el conocimiento lingüístico necesario para la tarea de la TA y un conjunto de comandos de control mediante los que se configura el funcionamiento del sistema.

El diseño del sistema. Desde un punto de vista teórico, Episteme está basado en gramáticas de unificación, y el proceso de traducción sigue el modelo de transferencia. Aunque en el diseño de Episteme se pretendía la mayor independencia posible respecto a teorías y formalismos gramaticales, el sistema se puede describir como una herramienta inspirada en las Gramáticas Léxico Funcionales –Lexical Functional Grammar, LFG (Bresnan, 1982)–.

El enfoque de transferencia utilizado en Episteme encaja perfectamente con los dos niveles de representación propuestos en LFG. Por una parte, la estructura de constituyentes (c-estructura) incluye información específica del idioma y es, consecuentemente, descartada durante la transferencia. Por otra parte, las rela-

ciones gramaticales que se encuentran en la estructura funcional (f-estructura) proporcionan una información más abstracta, ideal para la transferencia.

Episteme genera c-estructuras y f-estructuras para cada oración en el idioma origen. La transferencia se efectúa desde la f-estructura del idioma origen y produce una f-estructura en el idioma destino, desde la cual se genera su correspondiente c-estructura.

La construcción principal de Episteme es la de "lenguaje". El sistema permite la definición de uno o más lenguajes. Para cada uno de ellos se podrá definir opcionalmente una gramática de análisis, un léxico de análisis, un conjunto de módulos de transferencia a otros lenguajes, una gramática de generación y un léxico de generación. El sistema se puede configurar de modo que se puedan determinar cuáles son los idiomas origen y destino de la traducción.

Las distintas etapas que se llevan a cabo en el proceso de traducción son las siguientes:

1. Análisis léxico-morfológico. La etapa de análisis léxico recibe una cadena de palabras como entrada. El objetivo de esta etapa es la generación de una lista de categorías sintácticas con las correspondientes estructuras funcionales asociadas.
   El léxico se construye siguiendo la sintaxis de Mph que define un sofisticado lenguaje para la especificación de léxicos para gramáticas de unificación. La salida de éste se puede enlazar con el sistema Vtree diseñado para el almacenamiento y recuperación eficientes de grandes bases de conocimiento basadas en estructuras de rasgos. La utilización conjunta de ambos sistemas constituye de forma simultánea un entorno de especificación cercano al lingüista y un módulo de análisis léxico-morfológico muy eficiente.
2. Análisis sintáctico. Episteme utiliza como analizador sintáctico un chart[12] ascendente que incorpora predicciones descenden-

[12] Básicamente un chart se puede definir como un conjunto de árboles. Cada uno de los árboles es un grafo acíclico dirigido en el que las hojas son las palabras de la cadena de entrada y cada uno de los nodos internos se corresponde con una estructura sintáctica de la gramática. Una descripción del proceso de análisis

tes con el objetivo de mejorar la eficiencia del analizador. Para ello, se han implementado un conjunto de relaciones matemáticas entre los símbolos de la gramática que permiten decidir si un evento (arco) en determinadas situaciones no va a ser útil en las primeras fases de análisis; si esto sucede, el evento será eliminado.

3. Unificación. El módulo de unificación se divide en dos componentes funcionalmente diferentes:
   a) El núcleo del unificador (independiente del formalismo gramatical). Utiliza una estrategia de unificación reversible basada en desunificación con post-copia: una vez finalizado el proceso de unificación, si éste ha fallado, se aplica el algoritmo de desunificación recuperando íntegramente las estructuras iniciales de datos. Si ha tenido éxito, se copia (post-copia) el resultado y se aplica la desunificación para recuperar las estructuras de entrada al unificador.
   b) La capa que incorpora las estrategias del formalismo LFG. En lo que concierne a la capa de adaptación a LFG, el unificador permite además de la unificación ecuacional básica otra serie de instrucciones: asignación de estructuras, instrucciones de evaluación y ejecución condicional, etc.
4. Transferencia. Durante la fase de análisis, el parser y el unificador habrán obtenido una c-estructura y una f-estructura. Esta fase consiste en la transferencia de la f-estructura obtenida en el idioma origen hasta la f-estructura equivalente en el idioma destino.
   El proceso de transferencia se divide en dos etapas. En la primera (transferencia léxica) se modifican los valores de los atributos de tipo atómico. En la segunda (transferencia estructural) se pueden modificar los nombres de los atributos, lo que permite realizar modificaciones en la estructura misma.
5. Generación. El proceso de generación utiliza como entrada la f-estructura y obtiene como salida una cadena de palabras en el idioma destino. Durante el proceso utiliza una gramática del idioma destino así como un léxico para este mismo idioma.

distingue entre nodos y arcos. En este caso, los arcos se corresponden con la aplicación de una regla de la gramática sobre un conjunto de nodos.

Las tres primeras etapas corresponden a la fase de análisis, la etapa cuatro corresponde a la fase de transferencia y la última etapa corresponde a la fase de generación.

En cuanto a la resolución de problemas de traducción, Episteme resuelve problemas complejos como las construcciones resultativas y los verbos de dirección de movimiento. Sin embargo, el problema de la resolución de las anáforas pronominales no es tratado en el sistema.

3.3 Sistemas interlingua

3.3.1 KANT

La investigación del Procesamiento del Lenguaje Natural dentro del contexto de la Inteligencia Artificial (IA) se consideró como origen de nuevas técnicas para mejorar la calidad de la TA. El desarrollo de investigaciones en IA en proyectos de TA comenzó a principios de los años 70. Hay que destacar los trabajos de Yorick Wilks en la Universidad de Stanford y la investigación de Roger Schank y sus colaboradores en la Universidad de Yale. Todos estos trabajos fueron posteriores al informe ALPAC (ALPAC, 1966) que señalaba las debilidades de las aproximaciones de TA de aquella época.

Las aproximaciones basadas en IA se fundamentaban en el argumento de que la traducción suponía la transmisión del significado de un texto de un idioma a otro; por esta razón, un sistema de TA debía ser capaz de "comprender" el significado de los textos. Las características de las aproximaciones basadas en la IA son: la adopción de análisis sintáctico orientado a la semántica, la interpretación de textos según unas bases de conocimiento almacenadas y mediante mecanismos de inferencia, y representaciones independientes del idioma que expresan el significado de los textos.

Uno de los mayores centros de investigación de estas aproximaciones basadas en IA fue la Universidad de Carnegie-Mellon (Carnegie-Mellon University, CMU) de Pittsburgh. Desde 1983

llevan trabajando en sistemas experimentales basados en una metodología descrita como *TA orientada al significado en un paradigma interlingua*. Se prestó especial atención a la creación de herramientas para la adquisición del conocimiento, dando el nombre a su investigación: *TA basada en el conocimiento* (Knowledge-based MT, KBMT). Los sistemas desarrollados se consideraban como aproximaciones graduales de un sistema de TA interlingua ideal.

El primer prototipo de un sistema de TA basado en el conocimiento creado por la Universidad de Carnegie-Mellon (Goodman, 1989; Nirenburg, 1989) fue el precursor del sistema KANT (Mitamura *et al.*, 1991). El prototipo se diseñó para la traducción inglés-japonés (en ambas direcciones) de manuales de ordenadores personales[13]. Tenía un pequeño "dominio de modelo" de 1500 conceptos y diccionarios de análisis y generación para ambos lenguajes (contenían aproximadamente 900 entradas). El sistema se basaba en el formalismo gramatical de las Gramáticas Léxico Funcionales (LFG). A continuación se muestra el diseño de este primer prototipo:

El diseño de KANT.

- Los módulos. Los módulos básicos del sistema KANT son los siguientes:

1. Un analizador sintáctico con restricciones semánticas. Este analizador usa una gramática LFG.
2. Un intérprete semántico. Utiliza unas reglas que convierten la salida del módulo anterior en una representación interlingua.
3. Un incrementador (Augmentor) interactivo para las ambigüedades restantes.
4. Un generador semántico que produce estructuras sintácticas con selección léxica.
5. Un generador sintáctico que produce el texto destino.

Las bases de datos son específicas para cada idioma y consisten en gramáticas de análisis y generación y diccionarios de análisis y

[13] En versiones posteriores del sistema KANT se realizaba la traducción a otros idiomas: francés, alemán y español (Mitamura *et al.*, 1991; Nyberg *et al.*, 1998; Czuba *et al.*, 1998).

generación que proporcionan información sintáctica. El *diccionario de conceptos*[14] y la información semántica de los diccionarios de análisis y generación (definen algunas restricciones semánticas) son independientes del idioma pero específicos para el dominio.

- Proceso de traducción. El proceso de traducción consta de cinco etapas:

1. El analizador recibe el texto de entrada y produce un conjunto de f-estructuras[15] (típicas de las gramáticas LFG).
2. El intérprete semántico comprueba las ambigüedades de las f-estructuras y construye, mediante la aplicación de reglas, la representación interlingua formada por papeles temáticos: agente, tema, etc.
   El núcleo central del sistema es la representación interlingua. Representa los eventos actuales del texto origen en forma de redes de proposiciones, es decir, eventos o estados con sus argumentos y enlaces (causales, temporales, espaciales, etc.) a otros eventos o estados. Las representaciones se producen como instanciaciones de conceptos (eventos, individuales, etc.) del *diccionario de conceptos.*
3. La tarea del incrementador (Augmentor) es producir un texto interlingua sin ambigüedades que será la entrada de la etapa siguiente (generador).
   Ya que la salida de las etapas anteriores todavía refleja características del idioma origen, deberá ser transformada por el Augmentor en una forma independiente del idioma. Además, el Augmentor desambigua los candidatos interlingua posibles mediante el conocimiento almacenado en el *diccionario de conceptos.* La mayoría de las operaciones de desambiguación del Augmentor se realizan de un modo interactivo con los usuarios.
4. El generador semántico produce una f-estructura en el idioma destino mediante la selección de las unidades léxicas y la

[14] El diccionario de conceptos es una base de datos del conocimiento acerca de los eventos y entidades de un dominio específico.

[15] La f-estructura (f-structure) es una representación de la estructura de dependencias basada en rasgos.

aplicación de reglas similares a las usadas por el intérprete semántico.

5. El generador sintáctico produce una estructura superficial y el texto de salida.
   Mientras que el resultado del análisis es una salida múltiple de posibles interpretaciones, la generación se detiene en el momento en que se genera una cadena válida en el idioma destino. Las gramáticas usadas para la generación y el análisis son las mismas, pero sólo en el caso del japonés se pueden aplicar las reglas de una manera reversible.

Respecto a la resolución de las anáforas pronominales, hay que destacar que esta tarea es llevada a cabo por el Augmentor. La mayoría de las tareas del Augmentor se llevan a cabo de un modo interactivo con el usuario. Sin embargo, la resolución de las referencias pronominales para su posterior generación en el idioma destino se lleva a cabo de un modo totalmente automático.

3.3.2 DLT

El proyecto DLT, sistema de traducción distribuido –Distributed Language Translation, DLT (Witkam, 1983; Schubert, 1986; Schubert, 1988)–, es un proyecto de TA llevado a cabo por la empresa de software de Utrecht llamada BSO. En 1985 firmó un contrato de seis años de duración con el gobierno holandés con el objetivo inicial de crear un prototipo inglés-francés en 1987 y una versión comercial en 1993. La finalidad del proyecto consistía en la construcción de un sistema para la traducción entre idiomas de la Comunidad Europea (francés, alemán, inglés, italiano) con posibles extensiones a otros idiomas.

DLT fue concebido como un sistema interactivo multilingüe, no como una herramienta para los traductores sino, principalmente, como una herramienta para los usuarios monolingües que manejaban textos informativos (informes, manuales, etc.) o mensajes comerciales.

La traducción se denomina "distribuida" debido a que los procesos de análisis y generación se realizan en diferentes terminales. Para realizar el proceso de traducción, un usuario monolingüe

introduce el texto en un idioma (por ejemplo en inglés) en un terminal. Éste se traduce inmediatamente a un interlingua (idioma intermedio) que está basado en esperanto. El análisis y la traducción se realizan en tiempo real. Los problemas que no puede resolver el sistema se transmiten al usuario monolingüe (no necesita tener conocimientos del idioma destino) en un diálogo interactivo. En algunos casos, las interacciones monolingües provocan que el texto sea expresado de otra forma con el objetivo de simplificar y eliminar los problemas de la traducción.

El texto obtenido en el interlingua se transmite posteriormente a otro terminal donde otro usuario inicia la traducción del interlingua al idioma destino.

DLT es un sistema modular ya que los módulos de análisis (hacia la representación interlingua) y los módulos de generación (desde la representación interlingua) para los idiomas origen y destino se pueden crear, en principio, sin afectar a los módulos existentes en el sistema. Además, es considerado como un sistema que puede ser fácilmente extendido a otros idiomas independientemente de sus estructuras.

El diseño de DLT.

- El interlingua. En la mayoría de los sistemas de TA interlingua, las representaciones intermedias no son auténticamente interlingua. En muchos casos, la transferencia léxica está basada en diccionarios bilingües, es decir, no hay unidades léxicas interlingua (independientes del idioma).

En DLT, por el contrario, el interlingua esperanto es un "idioma natural" que tiene sus propias estructuras independientes y sus unidades léxicas. Según la opinión de sus investigadores, un interlingua sólo puede ser tan explícito y expresivo como un idioma humano si tiene el carácter de "idioma humano". Por esta razón, el interlingua usado en DLT es una versión modificada del esperanto. El esperanto combina la expresividad de los idiomas naturales con la regularidad y consistencia.

Entre las ventajas del esperanto podemos citar las siguientes:

- Como idioma (semi) natural tiene una expresividad, riqueza y flexibilidad que superan las representaciones lógicas interlingua.

- Proporciona un vocabulario estándar basado en las raíces de las palabras indo-europeas.
- Es regular y consistente.
- Es autónomo e independiente de otros idiomas.
- Puede ser aprendido y comprendido como cualquier otro idioma humano.

Sin embargo, el esperanto también tiene algunos inconvenientes como representación interlingua: ha adquirido con el paso de los años palabras homógrafas, imprecisiones estructurales y ambigüedades léxicas.

- **El proceso de traducción.** El análisis de los textos origen y la generación de los textos fuente se realizan en dos sistemas de traducción: del origen al esperanto y del esperanto al destino. Como consecuencia, el sistema DLT se puede considerar como una red de sistemas de TA bilingües cuyo centro es el esperanto modificado.

A continuación presentaremos las etapas del proceso de traducción en el prototipo original inglés-francés:

1. Análisis en el idioma origen. El analizador reconoce las palabras inglesas, sus características morfológicas y sintácticas, identifica relaciones de dependencia (sujeto, objeto, etc.) y produce un árbol de dependencias, generando las alternativas posibles cuando hay ambigüedad sintáctica. En esta fase la semántica no está involucrada y se consulta el diccionario del idioma origen.
2. Transformaciones de los árboles monolingües en el idioma origen. En esta etapa las contracciones se separan y los verbos auxiliares se reducen a verbos simples con sus correspondientes características (tiempo, voz, etc.).
3. Transformaciones de los árboles bilingües. Ésta es la tarea del Metataxor (Schubert, 1987; Maxwell & Schubert, 1989), que sustituye las palabras inglesas por las correspondientes en esperanto y sustituye las etiquetas de dependencia sintáctica por las equivalentes en esperanto.
   El sistema de reglas que usa es específico para un par de idiomas y en una única dirección. Por lo tanto, para el prototipo

inglés-francés se han desarrollado dos sistemas de reglas: de inglés a esperanto y de esperanto a francés.
En esta etapa se producen reestructuraciones en el árbol y la inserción de las funciones de las palabras.
Debido a que normalmente existen traducciones alternativas para cada unidad léxica en inglés (provocadas por las ambigüedades léxicas del inglés y las ambigüedades en la transferencia léxica del inglés al esperanto) se generarán un número de árboles alternativos en esperanto. Al igual que en la primera etapa no se utiliza la información semántica ni la información pragmática.

4. Elección semántica-pragmática. De los distintos árboles interlingua presentados, se selecciona el más probable para el contexto particular involucrado. Esta selección se realiza según los patrones de palabras en esperanto codificados en el Banco de Conocimiento Léxico (Lexical Knowledge Bank, LKB[16]).
Esta etapa la realiza el componente semántico SWESIL – Semantic Word Expert System for the Intermediate Language (Papegaaij *et al.*, 1986)–. Como resultado se obtiene una ordenación de todas las interpretaciones alternativas.
5. Desambiguación. Si en la fase anterior SWESIL no pudo elegir una única interpretación, los problemas de interpretación se presentan al operador en un diálogo interactivo con el ordenador. En el diálogo se expresan los fragmentos de las representaciones interlingua en el idioma origen que requieren desambiguación. Después de esta etapa se obtendrá una única representación interlingua de la oración de entrada.
6. Transformaciones del árbol interlingua. En esta etapa se determinan las características morfológicas para las representaciones interlingua correctas, incluyendo indicadores de las relaciones de gobierno y de concordancia.
7. Generación del texto interlingua. Se transforma el árbol interlingua en una forma lineal de esperanto. Implica determinar el orden correcto de las palabras y la eliminación de etiquetas

[16] El Banco de Conocimiento Léxico de textos en esperanto está formado por pares de palabras unidas por una función o relación. Éstas han sido extraídas tras un análisis de dependencia de un corpus de 500.000 palabras de textos en esperanto.

y características. El resultado es un texto plano en esperanto que puede ser leído directamente por las personas.

8. Corrección. Para seguridad, un analizador sintáctico analiza cada oración para comprobar que está bien formada. Las oraciones rechazadas vuelven a la etapa 6.
9. Codificación y decodificación. Se transmite el texto en esperanto a otro terminal.
10. Analizador de esperanto. El texto codificado se transforma en un árbol de dependencias.
11. Transformación del árbol interlingua. Tratamiento de contracciones y verbos en esperanto (igual que la etapa 2, en este proceso para esperanto).
12. Transformaciones del árbol bilingüe. El Metataxor genera del único árbol esperanto un conjunto de árboles alternativos de dependencias en francés, debido a que en la transferencia léxica una palabra en esperanto puede tener más de una palabra equivalente en francés.
13. Elección semántica-pragmática. En esta etapa se escogen las palabras correctas francesas seleccionadas a partir de la información de patrones almacenadas en el LKB[17].
    Ya que en esta etapa no se produce interacción con los usuarios la selección se realiza automáticamente.
14. Transformaciones de los árboles en el idioma destino. Se realizan adaptaciones en los árboles de dependencias franceses incorrectos y se insertan las relaciones de gobierno y concordancia.
15. Transformación del árbol en el idioma destino. El árbol francés se transforma en una forma lineal, donde aparecen las contracciones.

[17] En esta ocasión, el LKB está formado por un diccionario bilingüe esperanto-francés. En este diccionario los pares de unidades léxicas de esperanto y francés van acompañados por información contextual indicando las circunstancias bajo las cuales se debe realizar la transferencia léxica.

3.3.3 DLT con BKB

En una evaluación realizada en 1988, se observaron las deficiencias en la transferencia léxica del componente semántico SWESIL del sistema DLT. Por ello, en 1989, se decidió adoptar un nuevo modelo de traducción para la versión comercial de DLT. Este nuevo modelo estaba basado en el concepto de Banco de Conocimiento Bilingüe (Bilingual knowledge Bank, BKB).

En la evaluación realizada, se concluyó que las grandes bases de datos requeridas por el sistema no se construirían de la misma forma que en el prototipo inicial. La información léxica no se basaría en el esfuerzo humano para la construcción de grandes diccionarios, sino que se derivaría de los textos. En general, el procesamiento basado en las reglas se tendría que reducir y sustituir por el procesamiento basado en ejemplos usando datos de corpus de textos paralelos bilingües. El BKB representa la etapa final de un movimiento gradual en DLT de la aproximación basada en reglas a una aproximación basada en ejemplos (memorias de traducción).

El objetivo del BKB (Sadler, 1989) es servir como fuente fundamental de conocimiento lingüístico para todos los módulos en el proceso de traducción. Consiste en un corpus de textos equivalentes en dos idiomas que se han analizado estructuralmente (por el mismo tipo de analizador sintáctico) en "unidades de traducción" y que se han alineado[18]. Los investigadores del DLT usaron un corpus bilingüe estructurado para la resolución automática de las ambigüedades en el idioma origen, problemas de transferencia léxica y estructural y dificultades en la selección en el idioma destino.

El diseño del sistema. Las etapas de transferencia del idioma origen al destino permanecen como en el modelo inicial del DLT. La diferencia radica en el tipo de información que se aplica en cada etapa. Los módulos trabajan sobre una estructura de datos común, en vez de pasar secuencialmente estructuras de árbol. Las

[18] Las unidades de traducción son fragmentos de texto en los dos idiomas que un traductor consideraría equivalentes y mutuamente sustituibles.

bases de datos (diccionarios y otras bases de conocimiento) se integran dentro del BKB.

- **El proceso de traducción.** En el proceso de traducción hay cuatro mecanismos que operan interactivamente: analizador, Experto de Textos (Text Expert), Metataxor y Examinador (Examiner).

1. El analizador y el Metataxor se corresponden básicamente con los mecanismos descritos para el análisis sintáctico y la transformación del sistema DLT, excepto que ninguno está restringido a un análisis basado en reglas. Ellos pueden usar información del BKB para conocer qué patrones estructurales son los más probables. Además, el Metataxor puede valorar estructuras en el idioma destino potenciales según los subárboles bilingües alineados.
2. El Experto de Textos es un nuevo módulo propuesto para tratar las relaciones referenciales intersentenciales.
3. El Examinador (que sustituye al componente semántico SWESIL del DLT original) selecciona el mejor análisis del analizador. Para ello, escoge las mejores transformaciones del Metataxor y resuelve las incertidumbres gramaticales del Experto de Textos. Valora la corrección semántica y pragmática de las interpretaciones y propone traducciones consultando el BKB[19].

3.3.4 Rosetta

El proyecto Rosetta (Appelo & Landsbergen, 1986; Landsbergen, 1987a; Landsbergen, 1987b) desarrollado en los laboratorios Philips de Eindhoven (Holanda) es uno de los sistemas experimentales más innovadores. Su característica fundamental es el intento de desarrollar representaciones interlingua basadas en los principios de las gramáticas de Montague (Dowty *et al.*, 1981). Esta teoría une directamente la sintaxis y la semántica. En este proyecto se exploraron temas como la reversibilidad de las gramáticas,

[19] Para resolver un problema particular, el Examiner buscará ejemplos en un contexto similar en los idiomas origen y destino.

la composicionalidad de significados y el isomorfismo potencial de las gramáticas.

El proyecto tiene sus raíces en las investigaciones de Philips sobre un sistema de búsqueda de respuestas (question-answering system) llamado PHLIQA. Su objetivo era convertir una pregunta formulada en inglés en una representación lógica que pudiera entender la base de datos. Se llevaba a cabo mediante un analizador basado en una gramática independiente del contexto donde a cada regla gramatical le correspondía una regla de traducción en el lenguaje lógico. Sin embargo, la traducción no era directa y las representaciones independientes del contexto se transformaban en representaciones intermedias antes de que se obtuviera la representación lógica. Los resultados de esta aproximación híbrida no fueron satisfactorios y condujeron al diseño de una nueva gramática que fuera totalmente composicional y cuyas reglas fueran más potentes que las reglas de las gramáticas independientes del contexto. Se concluyó que las gramáticas definidas por Richard Montague ofrecían un modelo atractivo para esta aproximación.

Las posibilidades de las gramáticas de Montague se empezaron a explorar en el proyecto Rosetta que empezó en 1980. Inicialmente se diseñaron dos pequeños sistemas experimentales: Rosetta1 y Rosetta2. Un proyecto más grande que constaba de dos fases comenzó en 1985. La primera fase se concentró en la base lingüística y computacional y en el desarrollo de un sistema (Rosetta3) para la traducción de frases cortas holandés-inglés, holandés-español, inglés-holandés y español-holandés. Utilizaba unos pequeños diccionarios y el sistema generaba todas las traducciones posibles. La segunda fase, que empezó en 1989 se concentró en el desarrollo de una versión más robusta de Rosetta3 y en la construcción de un prototipo para una aplicación real (Rosetta4). El objetivo consistía en el desarrollo de un sistema para usuarios que no conocieran el idioma destino, incluyendo desambiguación interactiva durante el análisis y produciendo resultados que no requerían post-edición.

Características de Rosetta. El sistema Rosetta está basado en cuatro principios:

1. Composicionalidad.
   La principal característica de las gramáticas de Montague es la unión de las interpretaciones semánticas con las relaciones estructurales. Las gramáticas de Montague cumplen el principio de composicionalidad, según el cual el significado de una expresión es una función del significado de sus partes. Ya que las partes están definidas por la sintaxis, hay una estrecha relación entre sintaxis y semántica.
   Una gramática de Montague especifica un conjunto de "expresiones básicas" y un conjunto de reglas sintácticas. Las expresiones básicas son las unidades más pequeñas con significado propio y las reglas describen cómo se construyen expresiones más grandes a partir de las expresiones básicas. Estas reglas se aplican en sentido ascendente (bottom-up). Si consideramos que la expresión básica más grande es una oración, el proceso de su derivación se puede realizar de una forma explícita en un *árbol sintáctico de derivación.*
   Por ejemplo, si consideramos dos expresiones básicas: *hombre* y *venir*, y dos reglas: la formación de un sintagma nominal con un nombre y un artículo indefinido y la formación de una oración con un sintagma nominal y un verbo, se podría generar la siguiente oración:

   (21) Un hombre vino

   El componente semántico de una gramática de Montague asigna una representación semántica a una expresión asignándole un dominio semántico. Así, cada expresión básica se asocia directamente con un objeto del dominio y cada regla se asocia con una operación sobre objetos del dominio. De este modo se obtienen "significados básicos" correspondientes a las expresiones básicas y "operaciones con significado" correspondientes a las reglas sintácticas.
   El proceso de derivar el significado es paralelo al proceso de derivar una representación sintáctica. Se puede representar en un *árbol semántico de derivación* que tiene la misma geometría que su correspondiente árbol sintáctico de derivación pero eti-

quetado con los nombres de los significados básicos y las operaciones con significado (también denominadas reglas con significado).

En el proyecto Rosetta se han usado los árboles semánticos de derivación como representaciones interlingua. Hay que destacar que estos árboles conservan información acerca de la estructura superficial de las oraciones, a diferencia de las formas lógicas que se podrían derivar de estos árboles que carecen de esta información. El árbol semántico de derivación contiene exactamente la información relevante necesaria durante el proceso de traducción.

2. Explicitud.
   Además de que todas las gramáticas en los idiomas origen y destino se definen independientemente, todos los procesos lingüísticos y de traducción están expresados exactamente en las gramáticas.
3. Reversibilidad.
   La misma gramática se usa para el análisis y la generación de las oraciones, es decir, las gramáticas son reversibles.
   Por ejemplo, si consideramos la oración del ejemplo 21 y aplicamos las reglas de modo inverso: la obtención de un nombre a partir de un sintagma nominal y la obtención de un sintagma nominal y un verbo a partir de una oración, se podrían obtener las dos expresiones básicas (un nombre *–hombre–* y un verbo *–venir–*).
4. Isomorfismo.
   Una oración se considera traducción de otra oración si tienen el mismo árbol semántico de derivación y, por lo tanto, el mismo árbol sintáctico de derivación. Como consecuencia, las gramáticas de los dos idiomas tienen que estar emparejadas, es decir, cada expresión básica en una gramática tiene al menos una expresión básica correspondiente en la otra con el mismo significado, y cada regla en una gramática tiene al menos una regla correspondiente en la otra.

El proceso de traducción. Las gramáticas de Rosetta llamadas M-gramáticas (mostrando su afinidad a las gramáticas de Mon-

tague) tienen tres componentes básicos, cada uno de los cuales es reversible: un componente morfológico, un componente sintáctico y un componente semántico. El componente sintáctico se divide en una parte que trata con las estructuras sintácticas superficiales y una parte intermedia entre estas estructuras y los árboles sintácticos de derivación. El componente semántico trata con las representaciones intermedias entre los árboles sintácticos de derivación y los árboles semánticos de derivación.

La traducción se realiza en 8 etapas:

1. Análisis morfológico. Las palabras del texto de entrada se descomponen en raíz y afijos produciendo una secuencia de árboles léxicos formados por unidades léxicas (una o varias palabras[20]). Estos árboles constituyen las expresiones básicas.
2. S-análisis. Es la primera parte del análisis sintáctico en la que el S-analizador produce un conjunto de análisis alternativos. En esta etapa se resuelve únicamente la ambigüedad léxica categorial.
3. M-análisis. En la segunda parte del análisis sintáctico, el M-analizador selecciona las estructuras de árboles sintácticamente correctas y produce un conjunto de árboles sintácticos de derivación.
4. Transferencia analítica. Los árboles sintácticos de derivación se usan para obtener la representación del significado (árboles semánticos de derivación). Se transforma cada regla en su correspondiente "operación con significado" y se sustituyen las expresiones básicas interlingua por las expresiones básicas en el idioma destino.
5. Transferencia generativa. Mediante los árboles semánticos de derivación se obtienen los correspondientes árboles sintácticos de derivación en el idioma destino (siguiendo el principio de isomorfismo tienen el mismo significado que los árboles en el idioma origen).

[20] Las expresiones idiomáticas serán tratadas como unidades léxicas formadas por varias palabras.

6. M-generador. En esta etapa el M-generador (al igual que el M-analizador) verifica los árboles sintácticos de derivación, selecciona los correctos y los convierte en estructuras superficiales.
7. Generación de los árboles léxicos. Se generan los árboles léxicos escogiendo las palabras de la estructura superficial.
8. Generación morfológica. Por último, se generan las formas morfológicas correctas de las palabras.

Proyectos relacionados. Tras las investigaciones realizadas en el proyecto Rosetta sobre las gramáticas reversibles surgieron otros sistemas de TA basados en el mismo principio:

- ULTRA (Farwell & Wilks, 1991). Sistema multilingüe desarrollado en la Universidad de Nuevo México (Nuevo México, Las Cruces). Incluye todas las traducciones entre inglés, alemán, español, japonés y chino.
- Proyecto XTRA (Huang, 1988). Sistema bilingüe para la traducción inglés-chino.
- SWETRA (Sigurd, 1988). Desarrollado en la Universidad de Lund para la traducción inglés-sueco. Posteriormente se realizaron experimentos para la traducción de francés, polaco, ruso, georgiano e irlandés.
- CRITTER (Isabelle *et al.*, 1988). Sistema basado en un sublenguaje desarrollado en Canadá. Fue diseñado para realizar la traducción de informes del mercado agrícola entre inglés y francés.

Todos estos sistemas tienen en común el uso de las gramáticas de cláusulas definidas –Definite Clause Grammars, DCG (Pereira & Warren, 1980)– en las fases de análisis y generación del texto.

3.3.5 Proyecto CREST

En el Laboratorio de Investigación (Computing Research Laboratory, CRL) de la Universidad de Nuevo México (Nuevo México, Las Cruces) se está llevando a cabo desde 1999 un proyecto de investigación (subvencionado por DARPA –Defense Advanced Research Projects Agency–) denominado proyecto CREST (Cross-lingual Retrieval, Extraction, Summarization and Translation). El objetivo de este proyecto consiste en el desarrollo de un sistema

que construya automáticamente la base de conocimientos a partir del texto. Además, el proyecto incluye investigación sobre TA –sistema Mikrokosmos (Mahesh & Nirenburg, 1995b) presentado en la sección 3.3.6–, y sobre tareas de Procesamiento del Lenguaje Natural.

En el trabajo de Farwell & Helmreich (2000) se presenta una aproximación interlingua que forma parte de las investigaciones llevadas a cabo en el proyecto CREST. Esta aproximación interlingua, que no ha sido implementada aún, está orientada principalmente a la resolución de las referencias del texto.

El diseño del interlingua. Para la representación interlingua se ha adoptado una representación del significado del texto basada en la ontología ONTOS (Mahesh & Nirenburg, 1995a) como base de conocimiento general. La obtención de la representación interlingua a partir del texto original se realiza en dos etapas:

1. Análisis sintáctico. En la primera etapa se realiza un análisis sintáctico del texto original generando una f-estructura. Como parte del proceso de la generación de la f-estructura, se resolverán algunas relaciones anafóricas intrasentenciales.
2. Obtención de la representación interlingua. En la segunda etapa, se genera una representación del significado del texto (representación interlingua) a partir de la f-estructura. Esta representación se construye a partir de los conceptos ontológicos que se asocian con las unidades léxicas de la f-estructura[21]. La representación incluye: instanciaciones de los objetos, relaciones entre ellos, etc.

   Por ejemplo, al verbo español *comprar* se le puede asociar el concepto ontológico PURCHASE que es una estructura genérica correspondiente a las acciones de compra. La estructura se muestra en la figura 3.1.

   En la figura 3.1 aparecen los conceptos ontológicos de TIEMPO, LUGAR, AGENTE, TEMA, persona, organización, objeto, etc. Posteriormente cuando se utilice el verbo *comprar* a lo largo del texto, la f-estructura PURCHASE se instanciará,

[21] Una f-estructura se irá "rellenando" paulatinamente con las f-estructuras que le rodean.

TIEMPO: T [fecha]
LUGAR: L [lugar]
PURCHASE
AGENTE: P [persona/organización]
TEMA: O [objeto]
FUENTE: F [persona/organización]
CANTIDAD: C [moneda]
....

Figura 3.1. Estructura del concepto ontológico PURCHASE

es decir, se indexará y rellenará con otros objetos instanciados derivados de otros conceptos ontológicos asociados con las unidades léxicas del contexto de la f-estructura.
Por ejemplo, si consideramos el texto *Roche compra Docteur Andreu* (Farwell & Helmreich, 2000), en la estructura PURCHASE aparecerá un nuevo concepto ontológico: COMPAÑÍA (uno para cada una de las dos compañías: *Roche* y *Docteur Andreu*). La estructura final del concepto ontológico PURCHASE se muestra en la figura 3.2.

COMPAÑÍA-1
NOMBRE: *Roche*
SEDE: S [oficinas]
....

COMPAÑÍA-2
NOMBRE: *Docteur Andreu*
SEDE: S [oficinas]
....

TIEMPO: T [fecha]
LUGAR: L [lugar]
PURCHASE-1
AGENTE: *COMPAÑÍA-1* [persona/organización]
TEMA: *COMPAÑÍA-2* [objeto]
FUENTE: F [persona/organización]
CANTIDAD: C [moneda]
....

Figura 3.2. Estructura final del concepto ontológico PURCHASE

En la generación de la representación interlingua, se resuelve la referencia de los distintos elementos f-estructuras. En el proceso de resolución, los objetos se clasificarán en dos tipos posibles: *correferencial* y *referencia inicial*. Un objeto será *correferencial* si se puede inferir a partir de la información ontológica una conexión entre el objeto y un objeto interlingua ya existente. En caso contrario, será *referencia inicial*. Las acciones que se llevan a cabo para los dos tipos son las siguientes:

a) Correferencial: se añade la información del objeto a objetos ya existentes en la representación interlingua.
b) Referencia inicial: se crea un nuevo objeto en la representación interlingua y se añade al dominio del discurso.

- Los algoritmos. Debido a que no está disponible el mecanismo interlingua para su procesamiento, se han implementado una serie de algoritmos para determinar la correferencia de las expresiones dependiendo de su categoría sintáctica (nombre propio, pronombre, sintagma nominal indefinido y sintagma nominal definido). Los cuatro algoritmos son los siguientes:

1. Nombre propio. Se intenta emparejar en su totalidad o en parte con cualquier nombre propio previo. Si el emparejamiento es satisfactorio se establece un enlace de correferencia. De otro modo, el nombre propio es una referencia inicial.
2. Pronombre. Se utiliza un algoritmo basado en la cercanía del pronombre a sus posibles correferentes. Este algoritmo utiliza restricciones morfosintácticas (género, número,etc.) y, si es posible, semánticas para filtrar los posibles candidatos. Cuando un candidato satisface las restricciones, se establece la correferencia.
3. Sintagma nominal indefinido (sin artículo). Se asume que es una referencia inicial.
4. Sintagma nominal definido. Se intenta emparejar el núcleo del sintagma nominal con los núcleos de los anteriores sintagmas nominales. Si coinciden y además concuerdan sus complementos, se establece la correferencia entre los dos sintagmas nominales.

3.3.6 Mikrokosmos

Mikrokosmos –μK (Mahesh & Nirenburg, 1995a; Mahesh & Nirenburg, 1995b; Beale *et al.*, 1997)– es un sistema de TA basado en el conocimiento que está siendo desarrollado en el CRL de la Universidad de Nuevo México. El objetivo principal del proyecto Mikrokosmos es el desarrollo de un sistema que produzca una representación del significado de un texto de entrada (Text Meaning Representation, TMR) para un conjunto de idiomas origen. Inicialmente se empezó la investigación con el español para el dominio de las fusiones y adquisiciones de empresas; en el futuro está previsto la construcción de un sistema de TA multilingüe (árabe, japonés, ruso y tailandés) para un dominio más amplio. Como consecuencia de la investigación llevada a cabo para obtener una representación interlingua del significado del texto, se ha desarrollado una ontología para facilitar la interpretación y generación del lenguaje natural.

El diseño de Mikrokosmos. Un estudio exhaustivo del tratamiento computacional de los textos es un esfuerzo que requiere cubrir un amplio rango de fenómenos lingüísticos y pragmáticos. Debido a que las diferentes facetas del conocimiento son complejas en sí mismas, un estudio de los fenómenos individuales conduce a un aislamiento con respecto al estudio de otros fenómenos relacionados. Sin embargo, en un sistema de TA basado en el conocimiento, se requiere un conocimiento acerca de los distintos fenómenos lingüísticos interrelacionados. Una manera de combinar las distintas facetas del conocimiento en un sistema para tratarlas como un "todo" consiste en que los distintos fenómenos sean tratados por distintas *microteorías*[22] lingüísticas computacionales.

En el proyecto Mikrokosmos, se llevó a cabo un profundo estudio de las microteorías con el objetivo de definir una metodología para la representación en un formato interlingua (independiente del idioma) del significado del texto (TMR). El TMR representa el resultado del análisis de un texto de entrada y sirve como en-

[22] El nombre Mikrokosmos se refiere a un conjunto de microteorías que actúan como especialistas del significado. Cada una de ellas contribuye a la construcción de la representación del significado de un texto de entrada (TMR).

trada al generador en el idioma destino. El significado del texto de entrada se representa en el TMR como elementos instanciados de un modelo del mundo (ontología). El enlace entre la ontología y el TMR es proporcionado por el diccionario, donde los significados de las unidades léxicas se definen en términos de sus correspondencias con los conceptos ontológicos y sus contribuciones resultantes a la estructura TMR. La ontología y el diccionario son las dos fuentes de conocimiento principales del sistema Mikrokosmos.

- **La ontología Mikrokosmos.** Todas las entidades en la ontología Mikrokosmos están clasificadas en: *objetos*, *eventos* y *propiedades* (niveles más altos en la jerarquía de la ontología). Éstos constituyen los *conceptos* en la ontología y son representados como *estructuras*. Cada estructura es una colección de huecos. El conjunto de huecos define el concepto especificando el modo en el que el concepto se relaciona con otros conceptos en la ontología (a través de *relaciones*) o con constantes alfabéticas o numéricas (a través de *atributos*).

Cada concepto se representa por una estructura que tiene un nombre y los siguientes huecos: una definición, un atributo de tiempo, enlaces (*es-un* y *subclases* para conceptos; *instancias* e *instancia-de* para instancias) y otros huecos (otras propiedades).

- **El diccionario.** Los principales tipos de las entradas del diccionario corresponden a las categorías ontológicas de *objetos*, *eventos* y *propiedades*.

1. *Objetos.*
   Los objetos son normalmente representados por nombres, aunque la correspondencia puede ser más compleja. En general, pueden ser simples (correspondencia uno-a-uno entre concepto y palabra) o complejos (contienen algunas propiedades).
2. *Eventos.*
   Los eventos son representados por nombres y verbos o verbos nominalizados. Las entradas del diccionario para los eventos determinan la estructura de la cláusula por la correspondencia de los papeles temáticos esperados con los elementos del verbo. También se disponen de reglas de transformación para pasivas y oraciones de relativo.

3. *Propiedades.*
   Las propiedades son el tipo de entradas del diccionario más interesantes debido a su flexibilidad. Se pueden representar por adjetivos, oraciones de relativo, sintagmas nominales complejos y oraciones completas. A menudo, una propiedad se incluye en la definición de otro objeto o evento.

- El proceso de traducción. El sistema Mikrokosmos realiza el proceso de traducción en tres fases fundamentales:

1. Análisis morfológico y sintáctico.
   La primera fase en el procesamiento de la oración de entrada es el reconocimiento de sus distintas palabras y el análisis morfológico y sintáctico.
   La salida del análisis es la estructura sintáctica de la oración.
2. Obtención del TMR.
   Para producir la representación del significado a partir de la estructura sintáctica, Mikrokosmos usa tanto el conocimiento semántico representado en el diccionario como el conocimiento del mundo representado en la ontología independiente del idioma.
   La representación del significado del texto es el resultado de la instanciación de los conceptos de la ontología que corresponden a los sentidos elegidos de las palabras en el texto y de las distintas relaciones que se establecen entre ellos. El TMR básico obtenido es posteriormente mejorado por las distintas microteorías (expertos especializados) de distintos tipos de conocimientos del idioma, tales como microteorías de espacio, tiempo, aspecto, actitudes del hablante, etc.
   Existe un componente, el analizador semántico, que realiza la tarea de la selección del sentido correcto de las palabras ambiguas. Para ello, combina la información de sus fuentes lingüísticas y de conocimiento del mundo y establece un conjunto de restricciones sobre los huecos de las estructuras de cada uno de los conceptos.
3. Generación del texto destino.
   El sistema Mikrokosmos utiliza una arquitectura de control basada en restricciones, denominada "Hunter-Gatherer (HG)"

(Beale *et al.*, 1996). Se utiliza tanto en el análisis semántico como en la generación.
HG usa un conocimiento de restricciones (semánticas, sintácticas, etc.) para dividir el problema de entrada en problemas más pequeños, relativamente independientes y que se pueden procesar de un modo aislado. Posteriormente, se usan unas técnicas de síntesis para combinar los conjuntos de soluciones optimizadas de los subproblemas y obtener la solución global.

3.4 Resumen de los sistemas de TA

En la tabla 3.1 se muestra un resumen de los rasgos principales de los sistemas de TA presentados con detalle en este capítulo.

Para cada sistema estudiado se presentan una serie de características, destacando aquéllas relacionadas con la resolución de las referencias pronominales y generación de los pronombres omitidos. Con este estudio se pretende demostrar las deficiencias que tienen los sistemas de TA para el tratamiento y generación de las referencias pronominales en el idioma destino (tema central de esta Tesis). Las diferentes características estudiadas son las siguientes[23]:

1. Estrategia bilingüe *versus* multilingüe. Aparece en la primera columna de la tabla junto con el nombre del sistema e indica si el sistema traduce entre dos idiomas (bilingüe) o entre más de dos (multilingüe).
2. Estrategia de traducción. Indica la estrategia de TA utilizada por el sistema: sistema directo, transferencia, o interlingua.
3. Dominio restringido. Expresa si los textos del idioma origen tienen un dominio restringido. En el caso de que sí lo tengan se especifica cuál es.
4. Análisis sintáctico parcial. Se utiliza para indicar si el sistema de TA realiza un análisis sintáctico parcial del texto origen, identificando únicamente algunos constituyentes (sintagmas

[23] Las características aparecen ordenadas por columnas tal y como aparecen en la tabla 3.1.

nominales, sintagmas verbales, etc.) y relaciones entre ellos, o por el contrario se realiza un análisis sintáctico completo.

5. Información semántica. Con esta característica se especifica si el sistema utiliza información semántica para el análisis o generación del texto. Esta información también se utiliza normalmente para la resolución de los problemas lingüísticos: referencias pronominales, elipsis, etc.
   La información semántica se suele almacenar en los diccionarios utilizados por el sistema y expresan las relaciones semánticas existentes entre las palabras del texto.
6. Resolución de la anáfora intersentencial. Especifica si el sistema de TA lleva a cabo la resolución de la anáfora pronominal intersentencial.
   Éste es el problema crucial de la generación de los pronombres, ya que si no se realiza una resolución de la anáfora pronominal intersentencial es imposible generar correctamente los pronombres en el idioma destino cuando su antecedente se encuentre en una oración previa. Más aún, si no se resuelven estas anáforas no se podrá generar correctamente los pronombres omitidos que hagan referencia a oraciones anteriores.
7. Resolución de las cadenas de correferencia[24]. Esta característica indica si se identifican las cadenas de correferencia del texto, tras resolver las anáforas intersentenciales.
8. Generación de los pronombres omitidos. Expresa si el sistema trata el problema de resolver los pronombres omitidos en el idioma origen y que son obligatorios en el idioma destino.

Tras examinar la tabla 3.1 se puede observar que no hay ningún sistema real de TA que resuelva las referencias pronominales intersentenciales, las cadenas de correferencias y la generación de los pronombres omitidos en el idioma destino (si éstos son obligatorios) realizando un análisis sintáctico parcial del texto origen. En esta Tesis presentamos la aproximación de un sistema interlingua que trata todos estos problemas tras realizar un análisis sintáctico parcial del texto origen.

[24] Una cadena de correferencia está formada por una serie de constituyentes (referencias pronominales, sintagmas nominales, etc.) que hacen referencia a la misma entidad a lo largo de un texto.

	Estrategia traducción	Dominio restringido	Análisis sintáct. parcial	Informac. semántica	Resolución anáfora intersentenc.	Resolución cadenas correferenc.	Generación Pronombres omitidos
Systran [multilingüe]	Sistema directo	No	Sí	Sí	Sí	No	Sí
Météo [bilingüe]	Sistema directo	Sí. Sublenguaje partes meteorológicos	No	Sí	No hay referencias pronominales	No hay	No hay
SUSY [multilingüe]	Sistema transferen.	No	No	Sí	Sí	No	No
Ariane [multilingüe]	Sistema transferen.	No	No	Sí	Sí	No	Sí
Eurotra [multilingüe]	Sistema transferen.	No	No	Sí	No	No	No
METAL [multilingüe]	Sistema transferen.	No	Sí	Sí	Sí	No	Sí
Candide [bilingüe]	Sistema transferen.	Sí. Debates parlamentarios	No	No	No	No	No
Inter-Nostrum [multilingüe]	Sistema transferen.	Sí. Transacciones bancarias	Sí	No	No	No	No
IXA [multilingüe]	Sistema transferen.	No	Sí	No	No	No	No
Epísteme [multilingüe]	Sistema transferen.	No	No	Sí	No	No	No
KANT [multilingüe]	Sistema interlingua	Sí. Manuales de ordenadores personales	No	Sí	Sí	No	Sí
DLT [multilingüe]	Sistema interlingua	No	No	Sí	No	No	No
DLT (BKB) [multilingüe]	Sistema interlingua	No	No	Sí	Sí	No	No
Rosetta [multilingüe]	Sistema interlingua	No	No	Sí	No	No	No
CREST [multilingüe]	Sistema interlingua	Sí. Fusiones y adq. de compañías	No	Sí	Sí	Sí	Sí
μkosmos [multilingüe]	Sistema interlingua	Sí. Fusiones y adq. de compañías	No	Sí	No	No	No

Tabla 3.1. Características de los principales sistemas de TA

4. La anáfora en la TA: clasificación y resolución

En este capítulo se realiza un estudio profundo del fenómeno lingüístico de la anáfora. Comenzará con una revisión del concepto de anáfora y de todos los elementos que intervienen en el proceso de su resolución. Posteriormente, se realizará una clasificación exhaustiva de las anáforas en función de varios criterios y se presentará una revisión bibliográfica de las distintas estrategias utilizadas para resolver el problema de las relaciones anafóricas. Tras estudiar el problema de la anáfora en general, se aborda el problema de las expresiones anafóricas en el contexto de la Traducción Automática y se analizan los problemas asociados a la traducción de las expresiones referenciales. Por último, el capítulo finaliza con una revisión de diferentes estrategias específicas que tratan el problema de la anáfora en sistemas de TA.

4.1 El fenómeno lingüístico de la anáfora

Según la definición realizada por Hirst (1981), la anáfora se puede definir como "*el mecanismo que permite hacer en un discurso una referencia abreviada a alguna entidad o entidades, con la confianza de que el receptor del discurso sea capaz de interpretar la referencia y por consiguiente determinar la entidad a la que se alude*". Los componentes básicos del proceso anafórico son, por lo tanto, dos: la referencia abreviada a la que se denomina *expresión* o *elemento anafórico* y la entidad referenciada que se denomina *referente* o *antecedente*.

Los conceptos de *referente* y *antecedente* son normalmente utilizados de una forma equivalente. Sin embargo, en el trabajo de Brown & Yule (1983) se hace una diferenciación importante en-

tre estos conceptos. Se considera como referente la representación mental de los objetos que han sido evocados en el texto, mientras que el antecedente constituye la representación lingüística del referente. Veamos estos conceptos con un ejemplo:

(22) Juan compró *[el periódico]*$_i$ y *lo*$_i$ perdió en el restaurante.

En el ejemplo 22 se muestra una oración en la que existe una anáfora generada por el pronombre *lo*. En este caso, el referente sería la representación mental del periódico, mientras que el antecedente sería la representación lingüística, *el periódico*.

Otros conceptos relacionados con la anáfora y que normalmente son confusos son la *referencia* y *correferencia*. En el trabajo de Martínez-Barco (2001) se aclaran estos conceptos. De acuerdo con el *Linguistic Glossary*[1] la referencia, en su acepción más general, se define como la relación simbólica que una expresión lingüística tiene con el objeto concreto o abstracto al que representa. De esta definición se deriva el concepto de correferencia que se define como la referencia en una expresión al mismo referente en otra expresión.

La referencia, a su vez puede clasificarse en *endófora* y *exófora*. La endófora se puede definir como la correferencia de una expresión lingüística con otra que se encuentra antes o después de ella pero siempre dentro del discurso lingüístico. Una de las expresiones contendrá la información necesaria para interpretar a la otra. Cuando la expresión que contiene la información necesaria para interpretar a la otra se encuentra antes de ésta se habla de *anáfora* propiamente dicha (ejemplo 22). Sin embargo, cuando la expresión que lleva la carga semántica se encuentra después de la otra, entonces se habla de *catáfora*. En el ejemplo 23 (Covington, 1994) aparece una catáfora:

[1] http://www.sil.org/linguistics/glossary (página visitada el 30/01/01).

(23) Cerca de *él*$_i$, *[Pedro]*$_i$ vio una serpiente.

La exófora se define como la referencia directa a un objeto extralingüístico. Las exóforas más comunes son la *deixis* y la *homófora*. La deixis se puede definir (Lyons, 1977) como "*la localización e identificación de personas, objetos, eventos, procesos y actividades de las que se habla, o a las que se refiere, en relación con el contexto espacio-temporal creado y sostenido por el acto del enunciado y la participación en él, típicamente, de un hablante simple y al menos un direccionamiento (alguna forma de apuntar a algo)*". Por ejemplo, si imaginamos un situación concreta en la que estamos ante un conjunto de objetos y decimos la frase del ejemplo 24 señalando con el dedo hacia uno de ellos, el antecedente de la expresión anafórica (*éste*) será el objeto señalado que no ha sido nombrado explícitamente.

(24) Quiero *éste*.

La *homófora*, por su parte, se define como una referencia que depende del conocimiento cultural o de otro conocimiento general más que de características específicas de un contexto particular. En el ejemplo 25, la expresión anafórica (*El presidente*) hace referencia al presidente del Gobierno que no ha sido mencionado previamente.

(25) *El presidente* asistió al acto de clausura de la Cumbre Iberoamericana.

En el problema de la anáfora se pueden distinguir dos procesos distintos: la *resolución* y la *generación* de la anáfora (Moreno *et al.*, 1999). La resolución de la anáfora busca la entidad a la que hace referencia, mientras que la generación crea referencias sobre una entidad del discurso. En esta Tesis nos centramos en ambos procesos en el contexto de la TA. Por una parte se detectan y resuelven las anáforas de un texto escritos en un idioma. Con la información obtenida tras la resolución y, aplicando una serie de mecanismos, se generan las anáforas en el idioma al que se desea traducir el texto original.

Normalmente las estrategias para la resolución de la anáfora se clasifican en dos grandes grupos (Mitkov & Schmidt, 1998; Ferrández, 1998):

- Sistemas *integrados*. Se basan en el conocimiento, es decir, utilizan una serie de conocimientos (fuentes de información) que se suponen necesarios para resolver la anáfora. Dentro de este grupo se puede hacer una clasificación según el modo de combinar las distintas fuentes de información:
  - Sistemas *democráticos basados en restricciones y preferencias.*
  - Sistemas *consultivos.*

  Los sistemas democráticos se caracterizan porque dan igual protagonismo a cada una de las fuentes de información, mientras que en los sistemas consultivos aparece una fuente de información que propone candidatos a antecedente y las restantes fuentes se limitan a confirmar o rechazar estos candidatos.
- Sistemas *alternativos*. Se basan en el uso de técnicas y recursos distintos a los tradicionales, es decir, que no están basados en el conocimiento. Podemos destacar las aproximaciones basadas en: técnicas estadísticas, redes neurales, técnicas de aprendizaje, razonamiento, corpus etiquetados, etc.

Por lo tanto, y como ha quedado patente en las estrategias utilizadas para la resolución de la anáfora, es necesaria la combinación de distintas fuentes de información para realizar un correcto tratamiento de la anáfora. Estas fuentes de información se pueden clasificar en:

- Fuentes de información léxica. Como información léxica se podría incluir aquélla relativa al comportamiento de ciertas palabras o grupos de palabras en situaciones concretas. Por ejemplo, en el tratamiento de la anáfora una información léxica que se ha utilizado consiste en establecer una serie de preferencias para una serie de verbos concretos (Mitkov & Stys, 1997).
  Este tipo de información se suele almacenar en el lexicón o diccionario del sistema. El analizador-etiquetador léxico extrae la información necesaria del diccionario y se la asocia a cada palabra del texto.

- Fuentes de información morfológica. La información morfológica necesaria para la resolución de la anáfora consiste en la concordancia en número, género y persona entre la expresión anafórica y su antecedente. Esta regla general presenta algunas excepciones. Por ejemplo, en la oración del ejemplo 26 aparece una anáfora en plural que tiene como antecedente un grupo de antecedentes singulares. Para resolver estos casos (y otros similares) de discrepancia entre el género, número o persona de la anáfora y su antecedente se utilizan dos estrategias: estudiar y tratar las excepciones por separado (almacenándolas en el diccionario o en el sistema de resolución de la anáfora) o utilizar este tipo de información como preferencia relativa en lugar de utilizarla como restricción absoluta.

  (26) *[Juan]*$_i$ le dijo a *[Ana]*$_j$ que *ellos*$_{ij}$ tenían que salir inmediatamente.

  Normalmente la información morfológica se obtiene a partir de un analizador-etiquetador morfológico que la incluye como una etiqueta a cada una de las palabras del texto. Generalmente, las fuentes de información léxicas y morfológicas se generan conjuntamente en la misma herramienta proporcionando una única etiqueta a cada palabra.
- Fuentes de información sintáctica. La información sintáctica se refiere a la información que subyace en la estructura sintáctica del texto. Esta información es obtenida a partir del analizador sintáctico. Éste puede realizar un análisis sintáctico completo del texto (se obtienen todos los constituyentes) o un análisis sintáctico parcial (se extraen sólo unos constituyentes concretos).
  Una vez que se ha extraído la información sintáctica del texto se pueden formular una serie de reglas que permiten aceptar o rechazar antecedentes de ciertas expresiones anafóricas. Por ejemplo, en la resolución de la anáfora pronominal se utilizan las restricciones *c-dominio* o *c-command* (basadas en información sintáctica) para determinar casos de correferencialidad y no correferencialidad entre la anáfora y sus posibles antecedentes

(Reinhart, 1983). Otro tipo de información sintáctica utilizada es el paralelismo sintáctico entre constituyentes que permite expresar la compatibilidad entre anáforas y antecedentes.

- Fuentes de información semántica. La información semántica pretende captar las distintas relaciones semánticas que existen entre los constituyentes de una oración. Por ejemplo, los papeles temáticos expresan las relaciones entre los complementos (argumentos del verbo) y el verbo de la oración. Otro tipo de información semántica utilizada son las características semánticas que deben tener el sujeto, el objeto directo o el objeto indirecto de un determinado verbo.
  Esta información se puede obtener de herramientas tales como EuroWordNet (una red semántica multilingüe).
- Otras fuentes de información. La información pragmática, la información obtenida a partir de análisis de corpus o la información sobre la expresión anafórica constituyen otras fuentes de información que son utilizadas en el proceso de resolución de la anáfora.
  En la primera de ellas, se utiliza conocimiento del mundo y la posibilidad de inferir nuevo conocimiento para aceptar o descartar antecedentes bajo ciertas circunstancias.
  Con el análisis de corpus se pretende obtener un conjunto de reglas para cada tipo de expresión anafórica y específicas para cada tipo de texto. Estas reglas se pueden obtener con la información probabilística extraída tras el estudio del corpus.
  Por último, el tipo de expresión anafórica puede determinar qué información utilizar para establecer la correferencia y el número de oraciones a considerar en la búsqueda de posibles antecedentes.

4.2 Clasificación de las relaciones anafóricas

La clasificación de las relaciones anafóricas se puede realizar teniendo en cuenta varios criterios que afectan a la expresión anafórica, al antecedente y a la relación entre ambos. A continuación presentaremos las distintas clasificaciones según los criterios

de: la categoría gramatical de la expresión anafórica, el marco en el que ocurre la expresión anafórica, la naturaleza del antecedente, el tipo de referencia y la accesibilidad del antecedente.

4.2.1 Conforme a la categoría gramatical de la expresión anafórica

Según la categoría gramatical de la expresión anafórica podemos distinguir los siguientes tipos de relaciones anafóricas: anáforas pronominales, cero pronombres, descripciones definidas, anáforas adjetivas, anáforas verbales y referencias temporales o locales.

Anáfora pronominal. Una de las anáforas más habituales en el lenguaje natural son las originadas por los pronombres personales, demostrativos y posesivos. Dependiendo del tipo de pronombre las anáforas pronominales se podrían clasificar en:

- Originadas por los pronombres personales. Los pronombres personales son aquéllos que directamente representan personas, animales o cosas. Tienen formas diferentes según las funciones que desempeñan. De este modo, podríamos distinguir los siguientes tipos:
  - Anáfora pronominal de sujeto. Ocurrencias de los pronombres *yo, tú, usted, él, ella, ello, nosotros, nosotras, vosotros, vosotras, ustedes, ellos, ellas* con función sintáctica de sujeto.

    (27) *[Isidro]*$_i$ asistió al juicio. *Él*$_i$ era el hermano del acusado.

  - Anáfora pronominal de objeto. Ocurrencias de los pronombres *me, mí, conmigo, te, ti, contigo, lo, él, la, ella, le, se*[2], *consigo, nos, nosotros, nosotras, os, vosotros, vosotras, los, las, ellos, ellas, les* con función sintáctica de objeto (directo o indirecto).

    (28) A *[Julia]*$_i$ no *se*$_i$ lo he dicho.

    Este tipo de pronombres pueden aparecer pospuestos al verbo dando lugar a los *pronombres enclíticos* (ejemplo 29).

[2] El pronombre *se* del objeto indirecto es la forma que adoptan los pronombres *le* o *les* cuando deben ir delante del objeto directo *lo, la, los, las.*

(29) Si vas a ver a *[Julia]$_i$*, díle$_i$ que no irás.

– Anáfora pronominal reflexiva. Ocurrencias de los pronombres personales *me, te, se, nos, os, se* (y *sí*) cuando sustituyen en la oración a un sintagma nominal que es igual al sintagma nominal que funciona en la oración como sujeto.

(30) *[Laura]$_i$* se$_i$ lava la cara en el río.

– Anáfora pronominal recíproca. Ocurrencias de los pronombres personales *nos, os, se* cuando sustituyen a dos o más sintagmas nominales que, siendo sujetos, intercambian sus acciones y las reciben como objetos directos o indirectos. Actúan, pues, recíprocamente.

(31) *[El perro de Juan]$_i$* y *[el gato de María]$_j$* se$_{ij}$ odian a muerte.

• Originadas por los pronombres demostrativos. Los pronombres demostrativos son aquéllos con los que material o intelectualmente se demuestran o señalan personas, animales o cosas. Dan lugar al siguiente tipo:

– Anáfora pronominal demostrativa. Ocurrencias de los pronombres *éste, ésta, éstos, éstas, ése, ésa, ésos, ésas, aquél, aquélla, aquéllos, aquéllas, esto, eso, aquello.*

(32) *[El aparato]$_i$* es muy sencillo. *Éste$_i$* está construido con plástico y madera.

• Originadas por los pronombres posesivos. Los pronombres posesivos son los pronombres que denotan posesión o pertenencia. Se clasifican en los siguientes tipos:

– Anáfora pronominal posesiva pura. Ocurrencias de los pronombres *mío, mía, míos, mías, tuyo, tuya, tuyos, tuyas, suyo, suya, suyos, suyas, nuestro, nuestra, nuestros, nuestras, vuestro, vuestra, vuestros, vuestras.*

(33) *[El coche de Andrés]$_i$* está estropeado. Mi coche no está disponible y el *suyo$_i$* tampoco.

– Anáfora pronominal posesiva como determinante. Ocurrencias de los pronombres posesivos *mi, tu, su, mis, tus, sus, nuestro, nuestra, nuestros, nuestras, vuestro, vuestra, vuestros, vuestras* que actúan como determinante y hacen referencia al poseedor de la entidad a la que modifican.

(34) *[Antonio]$_i$* tiene 7 hermanos. *Su$_i$* madre se casó muy joven.

Por último, hay que destacar que existen pronombres que no tienen antecedente, es decir, no son anafóricos. Hirst los denominó *pronombres no referenciales* (Hirst, 1981). Este fenómeno es muy habitual en inglés, principalmente con el pronombre *it* cuando aparece como sujeto de oraciones impersonales (hay que recordar que en inglés es obligatorio la existencia del sujeto en la oración). Un ejemplo podría ser el siguiente: *It is necessary to carry out the project.* En español, este tipo de pronombres suelen aparecer con los pronombres *se* impersonales. Por ejemplo: *Se rumorea que ha quebrado la compañía.*

Cero pronombres. Los *cero pronombres* o *cero anáforas* (*zero pronouns*) se pueden considerar como un caso particular de la anáfora pronominal ya que éstos se originan cuando los pronombres se omiten en las oraciones.

En español, los cero pronombres aparecen frecuentemente cuando se omite el pronombre de sujeto de la oración (omisión del sujeto pronominal), recayendo la carga semántica sobre el verbo. Mientras que en español los cero pronombres sólo aparecen en la posición de sujeto, en otros idiomas pueden aparecer en la posición de sujeto u objeto de la oración (como el japonés).

Aunque los cero pronombres con función de sujeto se pueden considerar como un caso de elipsis, normalmente se tratan como anáforas puesto que la información morfológica sobre el pronombre omitido que contiene el verbo actúa como expresión anafórica

que se relaciona con el antecedente. Conociendo la información morfológica se puede reconstruir fácilmente el pronombre de sujeto. Tras su reconstrucción, la tarea de resolución de la anáfora se convierte en un caso más de anáfora pronominal de sujeto.

(35) *[Pedro]$_i$* suspendió el examen. *Ø$_i$* Estaba muy disgustado.

En el ejemplo 35, el símbolo *Ø* indica la ausencia del sujeto pronominal de la oración (el pronombre *él*), es decir, la existencia de una anáfora pronominal de sujeto cuyo antecedente es *Pedro*.

Descripciones definidas. Las descripciones definidas son expresiones anafóricas formadas por sintagmas nominales introducidos por un artículo determinado o por un demostrativo. Estas expresiones anafóricas hacen referencia a otro sintagma nominal enunciado previamente. El enlace entre el antecedente y la anáfora se hace a través de repeticiones léxicas, repeticiones con adición de modificadores o mediante relaciones semánticas como la hiponimia e hiperonimia (relación es-un), meronimia (relación parte-todo), sinonimia y antonimia, etc.

(36) Mi amigo tiene *[un apartamento]$_i$* en San Juan. *El apartamento$_i$* tiene vistas a la playa.

Anáfora adjetiva. La anáfora adjetiva es aquella expresión anafórica formada por un sintagma nominal en el que ha sido omitido el núcleo nominal, función que es realizada por el adjetivo.

(37) *[Las manzanas verdes]$_i$* están buenas, pero prefiero *las rojas$_i$*.

En ocasiones, la anáfora adjetiva puede aparecer como un sintagma preposicional o como una cláusula de relativo desempeñando una función de modificador adjetivo.

(38) Cada hora sale *[un tren]$_i$*. *El de las dos$_i$* ya ha salido.

Dentro de la anáfora adjetiva destaca la formada por las expresiones anafóricas *el mismo, la misma, los mismos, las mismas.*

(39) Después de tirar *[el vidrio]$_i$*, se puede utilizar *el mismo$_i$* para su reciclaje.

Normalmente, la anáfora adjetiva en español equivale a la *one-anaphora* en inglés originada por el pronombre *one*[3]. La diferencia consiste en que la *one-anaphora* inglesa incluye el pronombre *one* para sustituir al núcleo del sintagma nominal (cuando es un nombre contable) que queda omitido.

(40) John chose *[the black shirt]$_i$* because he didn't like *the white one$_i$*.

Un problema que pueden plantear la anáfora adjetiva y la *one-anaphora* es que permiten tanto referencias parciales como completas (se verá en detalle en la sección 4.2.4).

(41) Compró *[una manzana verde]$_i$* y *[otra roja]$_j$*. Ella se comió *la verde$_i$*.

En el ejemplo 41, la anáfora adjetiva *otra roja* realiza una referencia parcial (introduce un nuevo objeto en el discurso) mientras que la anáfora *la verde* realiza una referencia completa (es correferente con *una pera verde*).

Un tipo de anáfora similar a la anterior de tipo adjetivo es la *anáfora superficial numérica.* En este tipo de anáforas, las expresiones anafóricas presentan la siguiente estructura: determinante + adjetivo + modificadores. En este caso, el componente adjetivo se especializa (en ello se distingue de la anáfora adjetiva) en un numeral de tipo ordinal, cardinal o, en general, una expresión

[3] Hay algunos casos, por ejemplo la oración inglesa: *Sport gives one health*, en los que el pronombre *one* aparece sólo (sin ningún modificador) y con un significado indefinido, es decir, sin referirse a ningún antecedente ya que no se trata de una anáfora.

que represente algún tipo de orden como ocurre en *el primero*, *el segundo*, *el último*, etc.

(42) Compró *[fresas]*$_i$, manzanas y peras. *Las primeras*$_i$ estaban muy ácidas.

Para resolver la anáfora superficial numérica es necesario mantener la estructura superficial del texto donde aparece el antecedente con el orden de los constituyentes que lo acompañan. En el ejemplo 42, se puede observar cómo la resolución del antecedente no sería posible sin mantener en la memoria del oyente el orden en el que aparecieron los tres elementos.

Este tipo de anáfora en inglés tiene una estructura similar a la *one-anáfora*, ya que también incluye como núcleo el pronombre *one* (*the second one, the last one*, etc.) sin embargo, es conveniente distinguirla ya que realmente introduce un nuevo tipo de información (la relación de orden) indispensable para su correcta resolución.

Anáfora verbal. La anáfora verbal se produce por la sustitución de un verbo, o incluso un sintagma verbal completo, por un verbo auxiliar para evitar la repetición del mismo. En español suele ir acompañada por un pronombre de objeto que es quien realiza realmente la función referencial, por lo que puede ser tratada generalmente como una anáfora pronominal de objeto.

(43) Ayer, mi hermana *[cocinó el pescado]*$_i$, pero *lo hizo*$_i$ muy mal.

En inglés, este tipo de anáfora se produce con la utilización de verbos auxiliares (*do*, *have*, etc.). Normalmente, se utiliza para responder de forma abreviada a alguna pregunta y evitar la repetición del verbo principal.

(44) *[Do you smoke]*$_i$? Yes, I *do*$_i$.

La anáfora verbal puede ser confundida en ocasiones con la elipsis verbal. La principal diferencia consiste en que en la elipsis verbal el verbo se elimina y no se sustituye por ningún otro com-

ponente, mientras que en la anáfora verbal existe una expresión anafórica que la sustituye.

Referencias temporales o locales. Las referencias temporales o locales (Hirst, 1981) están originadas por la ocurrencia de adverbios o complementos circunstanciales que hacen referencia a lugares, modos, tiempos, etc.

(45) *[La casa]*$_i$ está situada en la cima de la colina. Es muy difícil llegar *allí*$_i$.

Hirst denominó de este modo a este tipo de expresiones anafóricas para indicar que sus antecedentes consisten siempre en la localización temporal (o local) más reciente en el texto, tal y como aparece en el ejemplo 46. Este tipo de expresiones necesitan obtener el antecedente para poder ser resueltas.

(46) El partido comenzó a *[las 5 de la tarde]*$_i$. *Dos horas después*$_i$, aún no había terminado.

4.2.2 Conforme al marco en el que ocurre

En función del marco en el que ocurre la anáfora podemos hablar de los siguientes tipos:

La anáfora intrasentencial. La anáfora intrasentencial (también denominada intraoracional) ocurre dentro de la oración, es decir, tanto el antecedente como la expresión anafórica se encuentran en la misma oración.

(47) Ayer estuve con *[Pedro]*$_i$ y *le*$_i$ pregunté si jugaría el partido con nosotros.

La anáfora intersentencial. En la anáfora intersentencial (también denominada anáfora interoracional o discursiva) la expresión anafórica y su antecedente no aparecen en la misma oración. Este tipo de anáfora actúa como un mecanismo de cohesión textual, es decir, la anáfora actúa como un instrumento lingüístico que

contribuye a mantener el discurso como una unidad de sentido completa gracias a la creación de vínculos de unión entre diferentes partes del texto. De este modo contribuye a la conservación de un foco de interés previamente establecido mediante las entidades discursivas, las cuales crean unos puntos de atención en el discurso que son confirmados y consolidados gracias a las expresiones anafóricas.

(48) *[Mi hermana y su amiga]$_i$* fueron al cine. A *ellas$_i$* les encantan las películas de amor.

4.2.3 Conforme a la naturaleza del antecedente

Según la naturaleza del antecedente, Asher (Asher, 1993) realiza la siguiente clasificación:

Anáfora individual. La anáfora individual es aquel tipo de anáfora cuyo antecedente es un objeto concreto (o individual). Esta característica hace que el antecedente corresponda sintácticamente a un sintagma nominal.

(49) El vecino tenía *[un canario]$_i$* que cantaba muy bien. *Lo$_i$* alimentaba con alpiste y miel.

Anáfora abstracta. La anáfora abstracta, por el contrario, tiene como referente una entidad abstracta (eventos, hechos o proposiciones). Sintácticamente, el referente no es por tanto un sintagma nominal sino un sintagma verbal, una oración, o incluso una cadena de oraciones.

(50) José dijo que *[llegaría una hora tarde]$_i$*. Yo creo que *eso$_i$* no está bien.

4.2.4 Conforme al tipo de referencia

En función del tipo de relación que se establece entre la expresión anafórica y su antecedente, Allen (Allen, 1995) hace la siguiente distinción:

Anáfora profunda. El concepto de anáfora profunda engloba la referencia completa a un objeto que ha aparecido previamente en el discurso. De este modo, se establece una relación de correferencia entre la expresión anafórica y su antecedente, es decir, ambos tienen el mismo referente.

(51) *[La camisa amarilla]$_i$* no tiene botones. *La$_i$* compré en rebajas.

Anáfora superficial. La anáfora superficial es aquélla que realiza una referencia parcial a parte de un antecedente. En este tipo de anáforas, la expresión anafórica introduce un nuevo objeto que no aparece explícitamente en el texto y que se relaciona con otro objeto anteriormente mencionado, es decir, no se establece una relación de correferencia entre la expresión anafórica y su antecedente.

(52) Yo compré *[una camisa amarilla]$_i$*. Juan también se compró *otra$_j$*.

En la literatura estos dos tipos de anáfora han sido denominados de manera diferente. Por ejemplo, Woods (Woods, 1977) las denominó anáfora completa y anáfora parcial respectivamente. Del mismo modo, Hirst (Hirst, 1981) utilizó los nombres de anáfora con identidad de referencia (Identity of Reference Anaphora, IRA) y anáfora con identidad de sentido (Identity of Sense Anaphora, ISA) para referirse a la anáfora profunda y superficial respectivamente.

4.2.5 Conforme a la accesibilidad del antecedente

Según la accesibilidad del antecedente, es decir, la facilidad con la que se puede acceder al antecedente de una expresión anafórica concreta, Rico (Rico, 1994) las clasifica en[4]:

[4] Los grupos aparecen ordenados de mayor a menor accesibilidad, de modo que la anáfora morfosintáctica es la de mayor accesibilidad y la anáfora pragmática la de menor.

Anáfora morfosintáctica. En la anáfora morfosintáctica las relaciones anafóricas se explican mediante criterios morfológicos y sintácticos, y son las que mayor accesibilidad presentan. Los antecedentes de este tipo de anáforas pueden ser sintagmas nominales, verbos o frases verbales y oraciones.

(53) *[Tu padre]$_i$* me dijo que llegaría pronto. Yo no *le$_i$* creo.

Anáfora semántica. En la anáfora semántica las relaciones anafóricas sólo pueden explicarse acudiendo a criterios semánticos. Normalmente en este tipo de anáforas, los antecedentes no están explícitos en el discurso, y por lo tanto, son antecedentes muy poco accesibles. Entre los antecedentes y las expresiones anafóricas existen unas relaciones semánticas tales como sinonimia, hiperonimia, meronimia, etc.

(54) Se compró *[el coche]$_i$* hace un mes y ya le ha tenido que cambiar *las ruedas$_i$*.

Anáfora pragmática. La anáfora pragmática supone una relación entre la expresión anafórica y su antecedente que no depende de factores contextuales. En este tipo de anáforas el oyente es capaz de identificar el antecedente independientemente del contexto. Para ello, es necesario tener un conocimiento pragmático (o conocimiento del mundo) presupuesto por el discurso.

(55) Ayer me compré *La isla del tesoro*.

Para identificar el antecedente de la anáfora del ejemplo 55 es necesario que el oyente tenga el conocimiento del mundo necesario que ayuda a interpretar *La isla del tesoro* como el título de un libro.

En esta Tesis nos centraremos en la resolución (en español e inglés) de las anáforas pronominales y cero pronombres (con función de sujeto) para su posterior generación en el idioma destino. Según los distintos criterios de clasificación comentados, las

anáforas a tratar serán de los siguientes tipos: intersentenciales (interoracionales o discursivas), individuales, profundas y morfosintácticas.

4.3 Estrategias de resolución de las anáforas

Como ya se ha presentado en la sección 4.1, las estrategias para la resolución de la anáfora se pueden clasificar en dos grandes grupos: sistemas integrados y sistemas alternativos.

A continuación presentaremos las aproximaciones más características de los distintos tipos de sistemas integrados y de los sistemas alternativos. Previamente se realizará una revisión de las primeras aproximaciones desarrolladas para el tratamiento de la anáfora.

4.3.1 Primeras aproximaciones al tratamiento de la anáfora

Desde hace varias décadas se han planteado una serie de aproximaciones computacionales que intentan resolver el problema de la anáfora en el contexto del lenguaje natural. Estas primeras aproximaciones tenían la característica común de centrarse prácticamente en una única fuente de información, no tratando de manera conjunta cada una de las distintas fuentes de información (presentadas en la sección 4.1) y basándose principalmente en la utilización de reglas heurísticas para la resolución de ciertos casos de anáfora. A continuación presentaremos estos sistemas pioneros en la resolución computacional de la anáfora:

- Una de estas primeras aproximaciones fue el sistema STUDENT desarrollado en 1964 por Bobrow (Bobrow, 1969). Este sistema resolvía problemas de álgebra e incorporaba una serie de procedimientos heurísticos para resolver algunos tipos de anáfora y repeticiones incompletas. En este sistema no se realizaba un análisis sintáctico del texto por lo que los procedimientos planteados no eran excesivamente fiables.

- El sistema ELIZA (Weizenbaum, 1966) fue uno de los primeros sistemas de simulación del comportamiento humano. Representaba el comportamiento de un psiquiatra dialogando en inglés con su paciente. ELIZA, que no pretendía ser un sistema experto, únicamente era capaz de construir nuevos enunciados enlazados a partir de los enunciados que emitía el usuario mediante un mecanismo de reconocimiento de patrones (*pattern-matching*). Respecto al tratamiento de la anáfora, el sistema sólo realizaba un tratamiento de la anáfora pronominal personal de primera y segunda persona. En concreto sólo trataba el intercambio entre pronombres: cuando el usuario utilizaba el pronombre *yo*, ELIZA al contestar lo intercambiaba por el pronombre *tú*.
- Otro sistema de diálogos fue el sistema interactivo SHRDLU (Winograd, 1972; Winograd, 1986). Permitía al usuario entablar un diálogo con el sistema obteniendo información sobre el mundo de las figuras geométricas, las posiciones, formas y colores que éstas pueden adoptar. Además, SHRDLU era capaz de aceptar instrucciones para el movimiento de un brazo robot. El sistema podía tratar algunos pronombres personales, sintagmas nominales definidos y determinados casos de la *one-anaphora*. SHRDLU estaba formado básicamente por un analizador sintáctico, un analizador semántico y el módulo de razonamiento para la resolución de las cuestiones. Para la resolución de la anáfora, el sistema usaba una lista con todos los sintagmas nominales que aparecían en el diálogo. Cuando aparecía una anáfora, se aplicaban restricciones de concordancia morfológica en género, número y persona entre la expresión anafórica y los candidatos a antecedente (cada uno de los elementos de la lista mencionada anteriormente). El candidato más próximo a la anáfora que cumplía dichas restricciones era elegido como solución. El propio diálogo era usado por el sistema para que el usuario confirmara la elección del antecedente.

  El sistema planteaba reglas sencillas para determinar la correferencia entre pronombres de la misma oración u oraciones adya-

centes[5]. El principal inconveniente de este método consistía en que no utilizaba restricciones sintácticas por lo que situaciones de no correferencialidad dentro de una misma oración, como en el ejemplo 56, no se podrían evitar.

(56) Él colocó *[la pirámide]*$_i$ sobre *[la mesa]*$_j$. Como *ésta*$_j$ no estaba nivelada, *ésta*$_i$ se cayó.

- El sistema LSNLIS (*Lunar Sciences Natural Language Information System*), también conocido como LUNAR (Woods, 1977), era un sistema de diálogo creado para actuar como interfaz en lenguaje natural sobre una base de datos de minerales extraídos de las piedras lunares que se adquirieron en la expedición del Apolo-XI a la luna. El sistema constaba de tres componentes básicos: un analizador sintáctico, un módulo de interpretación semántica y un módulo gestor de la base de datos.
  Durante el diálogo, el sistema almacenaba la información sintáctica y semántica de cada entidad aparecida (de modo similar al sistema SHRDLU). Con esta información, LUNAR resolvía ciertos casos de anáforas profundas y superficiales (ver sección 4.2.4), denominadas por Woods completas y parciales respectivamente. Para la resolución de ambos tipos de anáforas, la estrategia a aplicar dependía del tipo de pronombre encontrado. Si el pronombre era demostrativo, el sistema buscaba un sintagma nominal cuyo núcleo fuera el mismo que acompañaba al pronombre demostrativo. En caso de ser un pronombre personal, se usaría la información semántica para establecer una concordancia semántica.
  Uno de las principales limitaciones que tenía el sistema consistía en que dependía del paralelismo estructural entre la expresión anafórica y su antecedente para resolver una anáfora, por lo que era incapaz de hacer correferir entidades con estructuras diferentes. Además, no podía resolver la anáfora intraoracional ya que los sintagmas nominales de la oración que se estaba anali-

[5] Por ejemplo, se usaban reglas del tipo "En el caso que aparezcan dos veces los pronombres *it* o *they* en la misma oración o en dos oraciones adyacentes, se considerará que estos pronombres son correferenciales".

zando no estarían disponibles hasta que la oración se analizara completamente.

- Otro sistema que trataba de resolver los problemas de la ambigüedad anafórica es el descrito por Charniak (Charniak, 1972). El dominio de este sistema eran las historias infantiles y su objetivo principal consistía en estudiar el tipo de inferencias sobre el mundo real que son necesarias para hacer comprender al ordenador pequeñas historias, contestar a preguntas sobre estas historias y resolver problemas de ambigüedad anafórica. Respecto a la anáfora, realizaba un tratamiento de sintagmas nominales definidos y pronombres. Para realizar inferencias y razonamientos el sistema empleaba un método esencialmente heurístico.

  Para la resolución de las anáforas el sistema trabajaba con una lista de posibles antecedentes de la expresión anafórica. Si existía más de un candidato, se aplicaban una serie de procedimientos que sustituían la anáfora por cada uno de los posibles candidatos y realizaban una interpretación de las expresiones obtenidas. Aquel candidato que generaba una interpretación válida del texto sería seleccionado como solución de la expresión anafórica. El inconveniente de este sistema se presentaba en la selección del modo de aplicar los distintos procedimientos para resolver una relación anafórica, ya que siempre tenía preferencia aquél que se hubiera empleado antes, lo cual no producía habitualmente resultados correctos. Además, aunque el dominio era restringido, eran necesarios desarrollar una gran cantidad de procedimientos para realizar un correcto tratamiento de toda la información.
- En Webber (1978) se presentaba una aproximación computacional a la resolución de la anáfora en la que se consideraba el discurso como una colección de diferentes tipos de entidades (individuales, conjuntos, acontecimientos o acciones). Respecto a la resolución de la anáfora trataba la *one-anaphora* y pronombres. En el modelo de Webber aparecían lo que él denominaba *descripciones invocadoras* (*invoking descriptions*, ID) de cada entidad del discurso (refiriéndose a la primera mención de cada

una de las entidades del discurso). Las expresiones anafóricas tendrían como antecedentes estas ID.

4.3.2 Sistemas democráticos basados en restricciones y preferencias

En los trabajos de Carbonell & Brown (1988) y Rich & Luperfoy (1988) se reconoce la necesidad de tratar la anáfora como un fenómeno complejo en el que intervienen una gran variedad de fuentes de información. Según ellos, estas fuentes de información pueden ser aplicadas como *restricciones* o *preferencias*.

Las *restricciones* contienen una serie de reglas que tienen que ser cumplidas por cualquier candidato a antecedente de una anáfora. El objetivo de las restricciones es eliminar candidatos, de modo que un candidato que incumpla alguna de ellas sea automáticamente rechazado por el sistema. Las fuentes de información siguientes se tratan normalmente como restricciones: morfológicas, semánticas[6] y restricciones impuestas por la acción que se realiza en la oración.

Por otra parte, las *preferencias* facilitan la selección del antecedente correcto indicando qué candidatos se consideran mejores que otros. Están formadas por una serie de reglas cuyo cumplimiento no es necesario, pero que en caso de cumplirse destacan a un candidato determinado respecto a aquéllos que no las cumplen. Como criterios de preferencia se suelen utilizar los siguientes: paralelismo sintáctico (se escoge aquel candidato que realiza la misma función sintáctica que la anáfora), paralelismo semántico (se escoge aquel candidato cuya categoría semántica coincide con la categoría semántica de la expresión anafórica) y paralelismo pragmático (obliga a seleccionar como antecedente aquel candidato que ayude a mantener la cohesión discursiva).

Tratando de un modo conjunto las restricciones y las preferencias, las restricciones se usan para descartar candidatos imposibles y las preferencias para ordenar los candidatos restantes. El can-

[6] Se eliminan aquellos candidatos cuyos rasgos semánticos no concuerden con las restricciones semánticas exigidas por las expresiones anafóricas.

didato que alcance mejor posición en esa ordenación es elegido como antecedente.

A continuación presentaremos una serie de sistemas que siguen esta estrategia para la resolución computacional de la anáfora:

- Hobbs definió una aproximación para la resolución de la anáfora pronominal que trabajaba únicamente con información sintáctica y morfológica (Hobbs, 1978). En su algoritmo, Hobbs trabaja con los árboles obtenidos tras realizar el análisis sintáctico de las oraciones. Cuando aparece un pronombre, realiza una búsqueda por el árbol de análisis sintáctico hasta que encuentra un candidato que cumpla una serie de restricciones morfológicas (basadas en la concordancia entre expresión anafórica y antecedente) y sintácticas (basadas en las relaciones de dominio[7]).
  La búsqueda de los sintagmas nominales candidatos a solución de la expresión anafórica se realiza de izquierda a derecha (primero en anchura, *breadth-first*) comenzando por la oración donde se encuentra la anáfora. Hobbs evaluó[8] su algoritmo sobre 100 frases que contenían ocurrencias de los pronombres *he, she, it* y *they*, obteniendo un porcentaje de éxito del 81,8%.
  Se puede considerar que este algoritmo es computacionalmente eficiente, por lo que se tomó como referencia para el desarrollo de otras nuevas aproximaciones basadas en un sistema de restricciones a las que se le incorporan nuevas fuentes de información.
- El sistema de Rico es una variante de los sistemas basados en restricciones y preferencias (Rico, 1994). En su sistema se utilizan una serie de fuentes de información en las que se han eliminado las restricciones, de modo que todos los tipos de información son utilizados como preferencias.
  El sistema se fundamenta en los siguientes procesos:

[7] La relación de dominio es una relación gramatical que se puede establecer entre los distintos constituyentes de un árbol de análisis sintáctico. Así se define que *un nodo A domina a un nodo B, si A se encuentra por encima de B en el árbol de análisis sintáctico.*

[8] Hay que destacar que el módulo de análisis sintáctico no estaba implementado, por lo que la evaluación fue manual basándose en un análisis sintáctico completo, correcto y sin ambigüedades.

- Tratamiento simultáneo de toda la información lingüística (sintáctica, morfológica, semántica y pragmática).
- Asignación de relevancia a cada fuente de información en función del contexto lingüístico.
- Comparación en términos de igualdad de cada antecedente posible y su expresión anafórica. Ningún candidato es eliminado a priori ya que todas las fuentes de información se consideran como preferencias.

La estrategia planteada por Rico parte de la idea de codificar de forma numérica cada fuente de información representándola en forma de vector. De este modo cada expresión anafórica y sus posibles antecedentes tendrán asignados un vector numérico que simboliza los valores que tienen para cada fuente de información. La comparación entre dos vectores de este tipo se realizará mediante el producto escalar de los mismos.

La codificación de las distintas fuentes de información supone asignar un valor numérico a cada una de estas fuentes. Rico elige los siguientes atributos anafóricos: información morfológica (género, número y persona), información sintáctica (sujeto, objeto directo e indirecto, complemento circunstancial y preposicional), información semántica (rasgos semánticos de carácter general que aseguran la consistencia entre antecedente y anáfora: humano/no humano, animado/no animado, etc.) e información pragmática (prominencia[9], distancia oracional y distancia clausal).

- En el trabajo de Lappin & Leass (1994) se describe un algoritmo para la resolución de anáforas pronominales, reflexivas y recíprocas que utiliza exclusivamente información morfológica (concordancia de género, número y persona entre el pronombre anafórico y el sintagma nominal candidato a antecedente) y sintáctica (después de realizar el análisis sintáctico completo se establecen una serie de condiciones de correferencialidad y no correferencialidad entre pronombres y sintagmas nominales que aparecen en la misma oración).

[9] La prominencia mide el factor de importancia que tiene un candidato dependiendo de su función sintáctica en la oración o de las repeticiones que éste tenga a lo largo del texto.

Tras aplicar las restricciones morfológicas y sintácticas y descartar los candidatos incompatibles con la expresión anafórica, se calculan los valores de una serie de *factores de relevancia* para los candidatos restantes cuyo objetivo es medir la relevancia discursiva de unos candidatos frente a otros. Algunos de los factores de relevancia empleados en el algoritmo son los siguientes: oración actual (se otorga la valoración máxima a los candidatos que se encuentren en la misma oración que la anáfora), énfasis de sujeto (los candidatos con función de sujeto tienen mayor relevancia que cualquier otra función gramatical), etc. Tras calcular los factores de relevancia, el candidato con mayor valor será elegido por el algoritmo como solución de la expresión anafórica. Para la evaluación del algoritmo, Lappin & Leass utilizaron una serie de manuales de informática obteniendo un 85% de análisis correctos.

- Kennedy & Boguraev se basaron en la aproximación de Lappin & Leass y propusieron un algoritmo para la resolución de la anáfora pronominal (Kennedy & Boguraev, 1996). A diferencia del algoritmo de Lappin & Leass que utiliza un análisis sintáctico completo del texto, su algoritmo trabaja sobre la salida de un etiquetador (*part-of-speech tagger*, *POS tagger*).
  Tras ser aplicado el etiquetador sobre un texto, se obtienen las características léxicas, morfológicas, gramaticales y sintácticas de cada una de las palabras que hay en él. A la salida obtenida por el etiquetador, el algoritmo propuesto añade la información correspondiente a la posición numérica de cada palabra en el texto para determinar relaciones de precedencia entre las palabras. La identificación de los sintagmas nominales del texto se obtiene mediante una serie de reglas gramaticales que definen la composición de un sintagma nominal.
  Para la resolución de la anáfora pronominal el algoritmo utiliza una serie de restricciones morfológicas (concordancia de género, número y persona) y sintácticas (c-dominio[10]) sobre los posibles antecedentes de la expresión anafórica.

[10] Ya que no se dispone de una estructura sintáctica completa del texto, para definir las restricciones c-dominio el algoritmo realiza inferencias de las funciones gramaticales según la precedencia entre constituyentes.

Sobre los candidatos restantes, se aplican una serie de preferencias para obtener una única solución. Estas preferencias tienen asignadas un valor numérico y están basadas en información contextual, gramatical y sintáctica. Como ejemplos podemos citar las siguientes: si el candidato se encuentra en la misma oración que la anáfora hay que sumarle un valor de 100, si la función del candidato es sujeto hay que sumarle un valor de 80, etc. La suma total de estos valores numéricos dará el valor total según el cual se ordenarán todos los candidatos para la correferencia, eligiéndose como solución el que tenga mayor valor (si existen dos candidatos con el mismo valor total se escoge como solución el más cercano a la anáfora).

En la evaluación del algoritmo, Kennedy & Boguraev obtuvieron un porcentaje de éxito del 75%. Aunque este porcentaje es menor que el obtenido por la aproximación de Lappin & Leass (obtenían un 85%) se debe principalmente a que los textos tratados eran más variados y menos formales que los utilizados por estos últimos (manuales de informática).

- En los trabajos de Mitkov se presentan dos aproximaciones basadas en restricciones y preferencias para la resolución de la anáfora:
  - La primera aproximación está basada en restricciones y preferencias (Mitkov, 1994). Utiliza información morfológica, sintáctica, semántica y pragmática. Como restricciones escoge la concordancia en género y número, las restricciones sintácticas c-dominio y la consistencia semántica. Las preferencias que aplica son las siguientes: la teoría del foco del discurso (se verá en detalle en la sección 4.3.3), el paralelismo sintáctico, el paralelismo semántico y la distancia entre la expresión anafórica y su antecedente. Las preferencias son aplicadas en el orden especificado de modo que los primeros criterios de preferencia serán los que más peso tengan en la elección final.
  - La segunda aproximación está basada únicamente en preferencias (Mitkov, 1995a). Utiliza técnicas de *razonamiento con incertidumbre* (*uncertainty reasoning*) para aceptar o rechazar los candidatos posibles de la expresión anafórica. El uso de estas técnicas viene justificado en la base de que el trata-

miento de un texto en lenguaje natural incluye información incompleta, información que no se ha comprendido completamente, etc.

Esta aproximación trabaja con los mismos *factores*[11] que la aproximación anterior de Mitkov pero sin distinguir entre restricciones y preferencias. Para ello, se le asigna a cada factor un valor numérico que indica su aportación a la identificación del antecedente. A este valor numérico le llamará *factor de certeza* (*certainty factor*, CF). Cada posible antecedente de la anáfora tendrá un conjunto de factores de certeza que especificarán un único valor numérico. El candidato más cercano a la anáfora cuyo valor numérico sobrepase un determinado umbral será elegido como solución.

Ambas aproximaciones se compararon sobre textos de informática, obteniendo un 83% de precisión para la basada en restricciones y preferencias y un 82% para la basada en la técnica de razonamiento con incertidumbre. En el trabajo de Mitkov (1995b) se sugiere un nuevo método que está basado en la combinación de ambas aproximaciones. En él, cada candidato se evalúa simultáneamente por ambos métodos, deteniéndose el proceso cuando la respuesta de ambos coincida. De este modo se reduce el proceso de búsqueda en comparación con el uso independiente de cada uno de los métodos.

- En el trabajo de Stuckardt (1996) se propone un algoritmo para la resolución de la anáfora pronominal y descripciones definidas. El algoritmo está basado principalmente en las restricciones sintácticas derivadas del trabajo de Chomsky en su *teoría de rección y ligamiento* –*Government and Binding Theory*, (Chomsky, 1981)–, por lo que necesita un análisis sintáctico completo del texto.

  En la resolución de la anáfora aplica una serie de restricciones sobre los posibles antecedentes de la anáfora: morfológicas (concordancia de número, género y persona) y sintácticas (basadas en las restricciones sintácticas planteadas por Chomsky acerca de las relaciones anafóricas intraoracionales). Sobre los candi-

[11] Las fuentes de información son denominadas *factores* por Mitkov.

datos restantes se aplican una serie de preferencias (paralelismo sintáctico, proximidad, preferencia por el sujeto, etc.) que asignan un valor numérico a cada uno de ellos. Los candidatos se ordenarán en función del valor obtenido de mayor a menor. Las anáforas también se ordenarán descendentemente en función del valor numérico de su mejor antecedente. Siguiendo este orden determinado por las anáforas, se irán asignando sucesivamente los mejores antecedentes de cada expresión anafórica siempre que no aparezcan problemas de interdependencia entre ellos[12].
Tras la evaluación realizada, Stuckardt estimó que su sistema resolvía aproximadamente un 90% de los pronombres aparecidos en textos sobre biografías de arquitectos. Posteriormente, Stuckardt presentó un trabajo (Stuckardt, 1997) en el que se realizaba un análisis sintáctico parcial del texto, obteniendo una precisión de 82% en la resolución de la anáfora sobre textos en alemán.

- CogNIAC (Baldwin, 1997) es el sistema de resolución de la anáfora pronominal presentado por Baldwin. La principal característica de este sistema consiste en que únicamente resuelve los pronombres que no presentan ninguna ambigüedad. Para ello, se exigirá que después de aplicar el algoritmo de resolución de la anáfora sólo quede un candidato, en caso contrario el pronombre no se resuelve indicando que era un pronombre ambiguo (tiene, como mínimo, dos candidatos igualmente probables).
  El algoritmo se basa en la salida de un etiquetador (*POS tagger*) para identificar los sintagmas nominales simples. Posteriormente identifica manualmente las cláusulas de cada oración, utilizando expresiones regulares para identificar sujeto, objetos y verbo.
  En la resolución de la anáfora aplica una serie de restricciones morfológicas y c-dominio que reducirán el número de antecedentes. Sobre los candidatos restantes se aplican una serie de preferencias heurísticas (si sólo hay un candidato se escoge éste como solución, si el pronombre es reflexivo se escoge el candidato más cercano de la oración actual, etc.) en riguroso orden,

[12] La interdependencia se evita asegurando que un mismo candidato no pueda ser solución de dos anáforas distintas en la misma oración.

de modo que si se cumple una de ellas no se continúa probando con las demás. Si no se cumple ninguna de ellas, el pronombre no se resuelve indicando que es un pronombre ambiguo.
En la evaluación del sistema CogNIAC, Baldwin obtuvo un 90% de precisión.

- Mitkov & Stys desarrollaron otra aproximación para la resolución de la anáfora pronominal con unas fuentes de información muy limitadas (Mitkov & Stys, 1997).
  El algoritmo trabaja sobre la salida de un etiquetador que proporciona información léxica y morfológica de las palabras. Se aplicarán una serie de reglas gramaticales para la identificación de los sintagmas nominales.
  En la resolución de la anáfora se utilizaba la concordancia en número, género y persona como restricción y una serie de *indicadores de antecedente* (*antecedent indicators*) a modo de preferencias[13]. Ejemplos de estos indicadores serían los siguientes: los sintagmas nominales definidos se prefieren sobre los indefinidos, se prefieren los sintagmas nominales que representan términos del dominio del texto, paralelismo sintáctico, reiteración léxica, etc. Cada uno de estos indicadores asignará valores numéricos a los antecedentes, escogiéndose finalmente como antecedente de la expresión anafórica el que tenga mayor suma de estos valores. En caso de que dos candidatos tengan el mismo valor total se escogerá aquél que tenga mayor valor para el indicador de reiteración léxica y si continúa existiendo más de uno, se escogerá el más cercano.
  La evaluación de la aproximación fue realizada sobre manuales técnicos en inglés y polaco, obteniendo una precisión de 95,8% y 92,1% respectivamente.
- En Ferrández (1998) se presenta una aproximación computacional para el tratamiento de la anáfora pronominal, adjetiva, superficial numérica y *one-anaphora* basada en un sistema de restricciones y preferencias. A diferencia de otras aproximaciones, el objetivo de las preferencias no es la ordenación de candidatos sino descartar aquellos candidatos que no las cumplan, es

[13] Estos indicadores constituyen una serie de reglas heurísticas que se han obtenido tras el estudio de un corpus de entrenamiento.

decir, las preferencias funcionan de modo similar a las restricciones. La excepción se produce cuando ningún candidato cumple una determinada preferencia, en este caso no se descarta ningún candidato y el proceso continúa con la preferencia siguiente.
Esta aproximación utiliza las restricciones siguientes: morfológicas (concordancia de género, número y persona), restricciones sintácticas c-dominio y consistencia semántica (cuando se trabaja con textos de dominio restringido).
Si tras aplicar las restricciones queda más de un candidato se aplican las preferencias. Para ello, se designan diversos niveles de preferencia (según el tipo de anáfora) que se irán aplicando secuencialmente comenzando por las de nivel 1. El proceso finaliza cuando, tras aplicar un nivel de preferencia, sólo queda un candidato. Si se han aplicado todas las preferencias y aún queda más de un candidato, el más cercano a la anáfora se elige como solución. Las preferencias que usa esta aproximación dependen del tipo de expresión anafórica a resolver. Podemos mencionar las siguientes para la anáfora pronominal: preferencia por los candidatos que se encuentren en la misma oración que la anáfora o en la oración anterior, preferencia por los nombres propios o sintagmas nominales indefinidos, paralelismo sintáctico, etc.
La aproximación de Ferrández fue evaluada sobre textos en español de dominio restringido y no restringido obteniendo un porcentaje de éxito de 84% y 82,2% respectivamente en la resolución de la anáfora pronominal.

4.3.3 Sistemas consultivos basados en la teoría del foco del discurso

Los sistemas que siguen el enfoque consultivo se caracterizan por la selección de antecedentes mediante la información obtenida por una única fuente de información. Cuando un antecedente es propuesto, se consultarán otros tipos de conocimiento con el fin de obtener la información suficiente para confirmar o rechazar dicho antecedente.

La fuente de información que normalmente se usa en este tipo de sistemas consultivos para la selección de los antecedentes

es la correspondiente a la estructura del discurso. Por ello, generalmente estos sistemas se agrupan bajo el título de *sistemas basados en la teoría del foco del discurso*. El objetivo que persiguen es la modelización de la estructura del discurso para que fenómenos lingüísticos discursivos, como la anáfora, sean resueltos utilizando dicha estructura.

A continuación describiremos las aproximaciones basadas en la teoría del foco del discurso más relevantes:

- El método del *centering* (Grosz *et al.*, 1983; Grosz *et al.*, 1995) ha sido usado por una gran variedad de trabajos que usan la estructura del discurso para la resolución de la anáfora. Estos trabajos han utilizado el centering como marco para modelar la coherencia local en el discurso.
  El marco conceptual del centering es un modelo de focalización que explica las relaciones existentes entre un foco local (la entidad que tiene mayor interés en el enunciado actual) y las expresiones referenciales.
  El modelo establece unos mecanismos de focalización que limitan la búsqueda a las entidades relacionadas con elementos focalizados, no considerando aquellas entidades menos accesibles. Esta teoría se aplica fundamentalmente a la resolución de referencias pronominales.
  La regla básica de interpretación anafórica está basada en el proceso de identificación de los focos locales. Para ello se crean las siguientes estructuras:
  - Dado un enunciado U_i, $Cf(U_i)$ es la lista de *focos locales que miran hacia delante* (*forward-looking centers*). Esta lista proporciona un conjunto de entidades a las que pueden hacer referencia los siguientes enunciados. Está parcialmente ordenada e incluye todas las entidades del discurso del enunciado U_i. Su primer elemento es el *center* preferido, $Cp(U_i)$, y será el candidato que se espera encontrar en $Cb(U_{i+1})$.
    Un posible criterio para ordenar los elementos de la lista C_f está basado en los roles gramaticales (Grosz *et al.*, 1995), de modo que las entidades con rol de sujeto se prefieren a aquéllas

que tienen rol de objeto. A su vez, los objetos se prefieren al resto (complementos circunstanciales, etc.).

– $Cb(U_{i+1})$ es el *foco local que mira hacia atrás* (*backward-looking center*). Une el enunciado al discurso precedente y supone la entidad focalizada en la oración. Es el elemento más alto de $Cf(U_i)$ que será referido por otro elemento en el siguiente enunciado U_{i+1}.

El proceso de identificación de antecedentes se basa en los siguientes principios:

– Dado un enunciado U_i, el modelo predice qué entidad del discurso será el foco de U_{i+1}.
– Cuando el foco local se mantiene entre enunciados, el modelo predice que se expresará mediante un pronombre.
– Cuando se encuentra un pronombre, el modelo proporciona un orden de preferencia sobre los antecedentes posibles del enunciado anterior.

El *centering* define un orden de preferencia basado en técnicas para efectuar un cambio de tópico (tabla 4.1).

	$Cb(U_i) = Cb(U_{i-1})$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	continuación	desplazamiento
$Cb(U_i) \neq Cp(U_i)$	retención	desplazamiento

Tabla 4.1. Tipos de transiciones en el centering (Grosz *et al.*, 1983)

Una vez definidos todos los conceptos necesarios, el núcleo de la teoría se basa en dos reglas de *centering*:

– Si cualquier miembro de $Cf(U_i)$ es referido por un pronombre en U_{i+1}, entonces $Cb(U_{i+1})$ debe ser un pronombre.
– Las secuencias de *continuaciones* se prefieren a las secuencias de *retenciones*, y las secuencias de *retenciones* se prefieren sobre las secuencias de *desplazamientos*.

• Uno de los numerosos estudios sobre la teoría del centering lo constituye el trabajo de Brennan *et al.* (1987) para el tratamien-

to de la anáfora pronominal. En él, se realiza un refinamiento de las relaciones de transición entre enunciados presentadas en el trabajo de Grosz *et al.* (1983). Los nuevos tipos de transición se muestran en la tabla 4.2.

	$Cb(U_i) = Cb(U_{i-1})$ o indefinido $Cb(U_{i-1})$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	continuación	desplazamiento débil
$Cb(U_i) \neq Cp(U_i)$	retención	desplazamiento fuerte

Tabla 4.2. Refinamiento de los tipos de transiciones en el centering (Brennan et al., 1987)

El algoritmo propuesto se divide en las siguientes fases:

- Localización de las expresiones anafóricas en el texto. Tras encontrar un pronombre en U_i, los elementos en $Cf(U_{i-1})$ se prueban generando todas las asignaciones posibles de $Cb(U_i)$ a elementos en $Cf(U_{i-1})$ teniendo en cuenta que el antecedente del pronombre será $Cb(U_i)$. La lista Cf se ordenará usando los roles gramaticales de modo que las entidades se preferirán en este orden: sujeto, objeto directo, objeto indirecto, otros y complementos circunstanciales.
- Se aplican una serie de restricciones (morfosintácticas, semánticas, etc.) que filtran las asignaciones del pronombre a las entidades en Cf.
- Por último, la lista de las asignaciones obtenidas se ordena en función del tipo de transición. De esta forma se prefiere la *continuación* a la *retención*, ésta al *desplazamiento-débil* y ésta al *desplazamiento-fuerte*. La asignación de mayor orden se considerará la solución de la anáfora.

La evaluación del algoritmo propuesto obtuvo una precisión del 72,9%.

- En los trabajos de Strube (Strube, 1998; Strube & Hahn, 1999) se presenta un nuevo criterio para la ordenación de los elemen-

tos de la lista Cf usada en el método del *centering*. Este criterio pretende superar las dificultades planteadas por la ordenación de Cf basada en roles gramaticales (Grosz *et al.*, 1995) para idiomas como alemán o español que tienen un orden libre de palabras. Estos idiomas presentan una gran dificultad para realizar una definición correcta de los roles gramaticales, por ello se propone un nuevo criterio de ordenación basado en información sobre la familiaridad de las entidades del discurso. Estas entidades se clasifican en dos conjuntos disjuntos:

- En las *entidades del discurso conocidas para el oyente* (*Old*) se incluyen aquellas entidades que se han evocado previamente, es decir, expresiones correferentes que ya han sido resueltas (pronombres, descripciones definidas, nombres propios ya mencionados, etc.). Además, se incluyen las entidades que son nombres propios y títulos que no han sido usadas previamente.
- En las *entidades del discurso nuevas para el oyente* (*New*) se incluyen el resto de entidades que no estén en el conjunto *Old*.

Según la nueva propuesta, para la resolución de la anáfora pronominal la lista Cf –será denominada lista de entidades de discurso relevantes (*S-list*)– contendrá aquellas entidades que han sido referidas en el enunciado actual y en el previo. Quedará ordenada de modo que las entidades *Old* se prefieren sobre las entidades *New*. Si ambas son del mismo tipo, se preferirán las entidades del enunciado actual (y dentro de éstas se ordenarán según su posición).

El algoritmo de resolución de la anáfora pronominal consta de las siguientes fases:

- Localización de las anáforas pronominales. Comprobar los elementos de *S-list* siguiendo el orden establecido hasta que alguno sea válido para la referencia pronominal concreta.
- Actualización de *S-list* al finalizar el enunciado U, eliminado de *S-list* aquellas entidades que no han sido referidas en U.

El algoritmo fue evaluado obteniendo una precisión del 85,4%.

- En el trabajo de Mitkov (1996) se presenta una aproximación que combina los métodos lingüísticos clásicos con la teoría del

foco del discurso. En ella, se utiliza la información clásica (morfológica, sintáctica, semántica) junto con la información proporcionada por el discurso.

En su aproximación, Mitkov utiliza un método probabilístico para la selección del foco del discurso en la primera frase del texto. Se lleva a cabo en las siguientes fases:

- Selección de un conjunto de reglas heurísticas obtenidas a partir de un estudio previo sobre textos del mismo dominio. Estas reglas indican los modelos seguidos por el foco del discurso a lo largo del texto analizado.
- A partir de las reglas heurísticas anteriores se calcula la probabilidad de que un determinado segmento del texto sea el foco en la primera frase.

Cuando exista ambigüedad en la selección del antecedente de la anáfora usando exclusivamente información tradicional, se utilizará la información proporcionada por el foco del discurso (aplicada como preferencia) para seleccionar un único candidato.

4.3.4 Sistemas alternativos

Los sistemas alternativos se caracterizan por el uso de técnicas distintas a las tradicionales. La mayoría de estos sistemas utilizan información estadística que se ha obtenido tras el estudio de un corpus para la resolución de la anáfora.

Las aproximaciones más relevantes que siguen esta estrategia son las siguientes:

- En el trabajo de Dagan & Itai (1990) se desarrolló un método *estadístico basado en corpus* para la resolución de las anáforas pronominales (en concreto para el pronombre *it*).
  Su método se fundamentaba en un análisis previo de un fragmento de corpus (elegido aleatoriamente) que contenía oraciones con ocurrencias de pronombres *it*. Con este análisis se obtenía una serie de patrones que asociaban la ocurrencia de la anáfora con la posición sintáctica (sujeto u objeto) de su antecedente y el verbo con el que aparecía. Después del análisis previo, en la

resolución de una nueva expresión anafórica se aplican los patrones obtenidos que tienen el mismo verbo y posición sintáctica que la anáfora, seleccionando de todos los candidatos aquél que tiene mayor número de ocurrencias.
En la evaluación del método se obtuvo un 87% de éxito en la resolución de los *it* anafóricos.

- Nasukawa presentó una aproximación (Nasukawa, 1994) que utilizaba una serie de reglas heurísticas (extraídas de un corpus de entrenamiento) para mejorar el rendimiento de su sistema de resolución de la anáfora pronominal. Esta aproximación, conocida como *independiente del conocimiento*, se caracteriza porque no realiza análisis morfológico, sintáctico o semántico del texto y tampoco necesita conocimiento del mundo exterior.
  Su método consiste en la extracción del corpus de entrenamiento de una serie de plantillas que especifican las relaciones entre los constituyentes de las distintas oraciones. Utilizando las plantillas obtenidas y aplicando una serie de preferencias (repeticiones, paralelismo sintáctico, etc.) es capaz de resolver la anáfora pronominal.
  Evaluando su aproximación sobre manuales de informática obtuvo una precisión del 93,8%.
- El método presentado en Connolly *et al.* (1994) está basado en el *aprendizaje computacional.* Se fundamenta en el uso de un corpus de entrenamiento en el que se anotan las expresiones anafóricas y sus correspondientes antecedentes con una serie de características de clasificación. Del corpus se extraen un conjunto de reglas que muestran las preferencias de las distintas anáforas por determinadas clases de antecedentes.
  El proceso de resolución de la anáfora se limita a un problema de elección (cuando hay más de dos candidatos posibles) entre distintas clases (cada uno de los antecedentes). Inicialmente se seleccionan dos candidatos y se elige uno de ellos en función de las preferencias *aprendidas.* A continuación se compara el candidato elegido con otro, repitiendo este proceso hasta que quede un único candidato.
- El trabajo de Rocha (1999) presenta la *teoría de la probabilidad de los antecedentes.* Se basa en la construcción de un modelo

probabilístico para la resolución de las anáforas en diálogos en inglés y portugués.

El modelo se basa en la información probabilística obtenida de un corpus de entrenamiento anotado previamente. Esta información permitirá especificar para una anáfora determinada, la probabilidad que tiene una entidad de ser su antecedente.

Se realizará una anotación del corpus con el objetivo de marcar las entidades relevantes que participan en el discurso: tópico del discurso (local y global), expresiones anafóricas y sus antecedentes respectivos. Cada expresión anafórica será anotada con las cuatro propiedades siguientes:

- *Tipo de anáfora.* Clasificación de la anáfora según la categoría gramatical de la misma.
- *Tipo de antecedente.* El antecedente es clasificado en explícito o implícito (no aparece en el texto). Además se añade información que indica si es referencial o no referencial (casos de pronombres no anafóricos, por ejemplo, algunas ocurrencias del pronombre *it*).
- *Rol tópico del antecedente.* Indica si el antecedente pertenece a alguno de los elementos de la topicalidad del discurso: tópico global del discurso, local del segmento o local del subsegmento.
- *Estrategia de procesamiento.* Es una propiedad que indica la estrategia que se puede tomar para resolver una nueva expresión anafórica. Está basada en el análisis del corpus de entrenamiento realizado previamente. Las estrategias se clasificarán en cuatro grupos: procesos léxicos (conocimiento del mundo, reiteración léxica, etc.), procesos sintácticos (búsqueda del primer candidato, paralelismo sintáctico, etc.), procesos de discurso (conocimiento del discurso, deixis, etc.) y coloquialismos.

Tras la anotación del corpus del modo mencionado, se obtiene un modelo estadístico que combina cuatro variables entre sí mediante un árbol de probabilidad, cuya raíz será el tipo de expresión anafórica. La aplicación de la información estadística recopilada consistirá en la obtención de esquemas genéricos de

resolución de expresiones anafóricas según los criterios que hayan alcanzado una mayor probabilidad.

4.4 Anáfora y TA

Tal y como se ha presentado en la sección 3.4, uno de los principales problemas de los sistemas de TA experimentales y comerciales es que no realizan un tratamiento correcto de las expresiones anafóricas.

Según se menciona en el trabajo de Mitkov & Schmidt (1998), la resolución de las expresiones anafóricas y el establecimiento de sus antecedentes es un proceso clave que se ha de llevar a cabo para una correcta traducción en el idioma destino. Por ejemplo, cuando traducimos a un idioma que distingue el género de sus pronombres (como el español) es esencial resolver las anáforas para su correcta generación en el idioma destino. Por otra parte, cuando se traduce un fragmento de texto, es decir, no se trata de una oración aislada, es fundamental resolver las relaciones anafóricas entre las distintas entidades del discurso. Desafortunadamente, la mayoría de los sistemas de TA no abordan el problema de las expresiones anafóricas, y si lo hacen, se limitan a la resolución de la anáfora intraoracional.

La traducción de las expresiones anafóricas no es una tarea sencilla ya que se pueden plantear una serie de problemas debido a diferencias entre el idioma origen y el idioma destino. A continuación presentamos algunos de los problemas que se pueden producir en la traducción de las anáforas pronominales:

- Las anáforas pronominales del idioma origen se omiten en el idioma destino. Por ejemplo, en la traducción inglés-español los pronombres ingleses con función de sujeto son normalmente omitidos en español (omisión del sujeto pronominal). Estas construcciones denominadas *cero anáforas* con función de sujeto (*zero-subject* constructions) son típicas de idiomas como italiano, tailandés o chino. En otros idiomas, las *cero anáforas* pueden aparecer en la posición de sujeto u objeto de la oración, como el japonés.

- Los pronombres del idioma origen carecen de género, mientras que en el idioma destino se hace distinción entre masculino y femenino. Por ejemplo, en la traducción inglés-español, los pronombres *we, you* o *they* no tienen género, mientras que sus correspondientes en español distinguen entre las formas masculina y femenina (*nosotros, nosotras, vosotros, vosotras, ellos* y *ellas*).
- Existen determinadas palabras que son referidas por un pronombre singular en el idioma origen y por un pronombre plural en el idioma destino o viceversa. Por ejemplo, en la traducción español-inglés la palabra *ganado* es singular, mientras que su traducción a inglés (*cattle*) es plural.
- Los pronombres del idioma origen se traducen directamente por su antecedente en el idioma destino. Por ejemplo, en la traducción inglés-malayo hay una tendencia a sustituir el pronombre *it* por su antecedente.
- Un caso particular es el coreano. Los pronombres del idioma origen que se traducen a coreano se pueden omitir, traducir por un sintagma nominal definido, por su antecedente o por un pronombre coreano dependiendo del tipo de información sintáctica y semántica que tiene el antecedente de la anáfora.

Todos estos problemas ponen de manifiesto que el proceso de la traducción automática de las expresiones anafóricas es complejo y requiere un amplio estudio para que se pueda superar con unos resultados satisfactorios.

Por último, un importante aspecto en la traducción automática de la anáfora pronominal consiste en la aplicación de dos posibles técnicas (Mitkov & Schmidt, 1998). En la primera técnica, los pronombres en el idioma origen son traducidos directamente al idioma destino sin estudiar sus posibles relaciones con otras palabras del texto. La segunda técnica considera que los pronombres no son autónomos en su significado o función, sino que dependen de otras unidades del texto. Por lo tanto, un modo más eficiente de tratar los pronombres en TA sería el siguiente:

- El análisis del texto origen tiene que determinar su estructura de referencias, es decir, se tienen que resolver todas las expresiones anafóricas especificando sus antecedentes.
- Ésta es la única información que debe ser transmitida al módulo de generación del idioma destino.
- El módulo de generación, crea la estructura superficial apropiada en el idioma destino teniendo en cuenta los antecedentes de las expresiones anafóricas ya resueltas y de acuerdo a las reglas de este idioma.

En esta Tesis utilizaremos la segunda técnica, basada más en una reconstrucción de las expresiones anafóricas que en una traducción directa. Para ello, plantearemos un sistema interlingua inglés-español-inglés que trata las diferencias (a las que hemos denominado *discrepancias*) entre ambos idiomas y que permite la correcta generación de las expresiones anafóricas (incluyendo las anáforas intersentenciales) en el idioma destino.

4.4.1 Aproximaciones para la resolución de las expresiones anafóricas en TA

Debido a que la mayoría de los sistemas de TA realizan un tratamiento del texto origen *oración-por-oración*, normalmente no abordan el problema de la resolución de la anáfora. Sin embargo, hay una serie de trabajos específicos desarrollados para resolver el problema de las expresiones anafóricas en sistemas de TA. A continuación describiremos los trabajos más destacados en este campo:

- En el trabajo de Wilks (1975) se describe el modo en el que una teoría de semántica de preferencias se usa en un sistema de TA (inglés-francés) para abordar el problema de las expresiones anafóricas clasificadas como semánticas y pragmáticas.
  Esta teoría utiliza patrones semánticos para interpretar las palabras en un contexto determinado. De este modo, el texto se concibe como una unidad semántica (formada por un conjunto de bloques semánticos) que no necesita la sintaxis para su análisis. A su vez, cada bloque semántico está compuesto por

plantillas, las cuales se unen mediante patrones y reglas de inferencia de sentido común. El vínculo de unión entre estos elementos lo constituyen las fórmulas.

Para la resolución de la anáfora pronominal el sistema utiliza cuatro niveles de resolución dependiendo del tipo de pronombre y del mecanismo necesario para resolverlo. Estos niveles se irán aplicando sucesivamente mientras exista más de un candidato para un determinado pronombre:

- En el primer nivel, se aplican las plantillas y patrones existentes utilizando únicamente conocimiento de las palabras que aparecen en el texto.
- En el segundo nivel, se construyen nuevas plantillas semánticas *implícitas* en las plantillas que ya existen. Será necesario utilizar métodos de inferencia para obtener conocimiento del mundo real.
- El tercer nivel intenta encontrar el foco o tópico de la oración que se considera como antecedente. Supone el empleo de reglas de inferencia sobre el mundo real que van más allá de las definiciones y significados codificados en las fórmulas.
- En el cuarto nivel se selecciona el antecedente por defecto, suponiendo que el foco se ha mantenido.

- Wada (Wada, 1990) presenta un mecanismo para la resolución de la anáfora pronominal en un sistema de TA inglés-japonés. Dicho mecanismo consta de los siguientes módulos:
  - Identificación de los distintos segmentos del discurso.
  - Ordenación de los candidatos del segmento del discurso actual. Para ello aplica tres filtros (función gramatical, uso de pronombres y construcción sintáctica) sobre los sintagmas nominales del segmento actual.
  - Búsqueda del antecedente de una anáfora determinada. En primer lugar, se realiza la búsqueda de candidatos en el segmento actual, aplicando restricciones morfológicas (género y número) y sintácticas. Si la búsqueda falla, se continúa buscando en los segmentos precedentes. Los resultados de la búsqueda determinan tres clases de pronombres:
    1. El antecedente se encuentra en el segmento actual.

2. El antecedente no se encuentra en el segmento actual, pero es controlado por el foco del discurso.
3. No se encuentra ningún antecedente válido. Se escoge la traducción por defecto *palabra-a-palabra.*

- Chen (Chen, 1992) describe un algoritmo para resolver los pronombres personales y reflexivos en un sistema de TA inglés-chino-inglés. Su algoritmo está basado en restricciones sintácticas c-dominio y restricciones semánticas. Además, propone otro algoritmo para la resolución de las cero anáforas en chino. Por último, investiga la distribución estadística de las anáforas y sus antecedentes en ambos idiomas para su posible aplicación en TA.
- Nakaiwa presenta en varios trabajos (Nakaiwa & Ikehara, 1992; Nakaiwa & Shirai, 1996) un algoritmo para la resolución de las cero anáforas en japonés para su aplicación en un sistema de TA japonés-inglés. Su algoritmo usa restricciones semánticas y pragmáticas (atributos semánticos de los verbos, expresiones modales, etc.) para la resolución de las cero anáforas intraoracionales.
- Un algoritmo para la resolución de la anáfora pronominal en un sistema de TA portugués-inglés que traduce abstracts científicos se presenta en Saggion & Carvalho (1994). El algoritmo usa concordancia sintáctica y restricciones sintácticas c-dominio para resolver la anáfora intraoracional. Para la resolución de la anáfora intersentencial utiliza el análisis sintáctico de la oración previa y una lista de los antecedentes anteriores.
- El sistema experimental KIT-FAST para la Traducción Automática inglés-alemán nace como una investigación complementaria del proyecto Eurotra (Allegranza *et al.*, 1991) e incluye un módulo para la resolución de las expresiones anafóricas (Preuss *et al.*, 1994).

  KIT-FAST usa dos niveles para la representación del texto. La representación estructural proporciona información acerca de la estructura del texto, incluyendo las relaciones entre verbos y nombres, relaciones entre adjetivos y sus complementos, papeles temáticos, características semánticas, etc. La representación referencial expresa aspectos del contenido del texto.

Para la resolución de la anáfora utiliza una serie de factores:

- *Proximidad.* Los pronombres personales tienen con más probabilidad sus antecedentes en la oración previa, mientras que los pronombres posesivos los suelen tener en la misma oración.
- *Ligadura* (*binding*). Las anáforas pronominales no pueden tener como antecedentes entidades hermanas en la representación estructural del texto.
- *Énfasis* (*themehood*). Se prefieren las entidades que son sujeto o tópico de una oración.
- *Paralelismo.* Concordancia e identidad de roles.
- *Consistencia conceptual.* Compatibilidad entre antecedente y anáfora.

- En el trabajo de Mitkov *et al.* (1994) se describe una extensión del sistema de TA inglés-coreano MATES que permite la resolución de las anáforas pronominales. El método usado es una versión simplificada del algoritmo desarrollado por Mitkov basado en restricciones y preferencias (Mitkov, 1994). Básicamente usa restricciones morfológicas, sintácticas c-dominio y la consistencia semántica. En cuanto a las preferencias, utiliza el paralelismo sintáctico, el paralelismo semántico y la distancia entre la expresión anafórica y su antecedente.
- En Mitkov *et al.* (Mitkov *et al.*, 1995) se describe una extensión del sistema CAT2 (Sharp, 1988) para el tratamiento de la anáfora pronominal intersentencial. Este sistema, que realiza la traducción inglés-alemán, también es una investigación complementaria del proyecto Eurotra.

  El módulo de resolución de la anáfora está basado en un conjunto de restricciones y preferencias. Se usan las siguientes restricciones: morfológicas (género, número y persona), sintácticas c-dominio y semánticas (consistencia semántica). Como preferencias se utilizan las siguientes: paralelismo sintáctico, topicalidad y paralelismo semántico.

  Para resolver la anáfora intersentencial, CAT2 presenta el mismo problema que la mayoría de los sistemas de TA: procesan el texto de entrada *oración-por-oración.* El algoritmo propuesto simula la intersentencialidad uniendo las oraciones tal y como se muestra en el ejemplo 57 (Mitkov & Schmidt, 1998):

(57) **I** *[The decision]*$_i$ was adopted by *[the council]*$_j$; *it*$_j$ published *it*$_i$.

E *[El consejo]*$_j$ tomó *[la decisión]*$_i$; *éste*$_j$ *la*$_i$ publicó.

Para la resolución de la anáfora intersentencial, en cada sintagma nominal se almacena la siguiente información: género, número y persona; características semánticas y tipo (indica si es un pronombre o no lo es). Todos los sintagmas nominales de una oración se agrupan en una estructura. Si en la oración siguiente aparece un pronombre, sus características son comparadas con todos los candidatos (aplicando técnicas de *backtracking*) hasta que concuerden. Dicho candidato será elegido como solución del pronombre.

Como se puede observar en todas estas aproximaciones presentadas para la resolución de las expresiones anafóricas en sistemas de TA se sigue una estrategia basada en restricciones y preferencias utilizando distintos tipos de información (morfológica, sintáctica, semántica o pragmática).

5. Generación de la anáfora con el sistema interlingua AGIR

En este capítulo presentamos el sistema interlingua AGIR (*Anaphora Generation with an Interlingua Representation*) que permite la generación de las expresiones anafóricas en el idioma destino a partir de la representación interlingua del texto origen.

El sistema utiliza una serie de fuentes de información lingüísticas: léxicas, morfológicas, sintácticas y semánticas utilizadas para resolver los problemas lingüísticos del texto origen y, posteriormente, generar una representación interlingua global del texto que permita la correcta generación de las expresiones anafóricas.

El sistema AGIR en la actualidad es capaz de resolver y generar en inglés y español la anáfora pronominal originada por los pronombres personales de tercera persona y los cero pronombres. Ya que el sistema es capaz de resolver otros tipos de anáforas (adjetivas, *one-anáforas*, descripciones definidas, etc.), éste se podría enriquecer para la generación de este tipo de anáforas en español e inglés.

5.1 Arquitectura general del sistema AGIR

La arquitectura del sistema AGIR (*Anaphora Generation with an Interlingua Representation*) se corresponde con la arquitectura general de un sistema de TA que sigue una estrategia interlingua. En AGIR se ha utilizado esta estrategia por dos razones fundamentales. En primer lugar, el número de módulos requerido en un sistema interlingua multilingüe –$2n$, siendo n el número de idiomas– es mucho menor (dependiendo del número de idiomas) que los módulos requeridos en un sistema de transferencia multilingüe –$n(n+1)$–. Tal y como menciona Trujillo (Trujillo, 2000),

"*la traducción automática por transferencia es un modo extremadamente despilfarrador y muy caro en la traducción entre más de dos lenguas, puesto que es necesario un módulo específico para cada par de lenguas*". En segundo lugar, el diseño de estos módulos de transferencia para dos idiomas determinados implica la creación de reglas de transferencia estructurales que transforman el árbol de análisis del texto origen en el correspondiente del texto destino; estas reglas son muy específicas y deben tener en cuenta todas las posibles divergencias y excepciones que se puedan producir entre estos idiomas, por lo que su diseño puede ser muy complejo.

En la figura 5.1 se muestra la arquitectura general del sistema AGIR. La generación del texto destino se lleva a cabo en dos etapas: en primer lugar, se realiza el análisis del texto en el idioma origen y se obtiene la representación interlingua de éste; en la segunda etapa, se genera el texto en el idioma destino a partir de la representación interlingua. Como se puede observar en la estrategia interlingua adoptada, los módulos de análisis y generación son independientes.

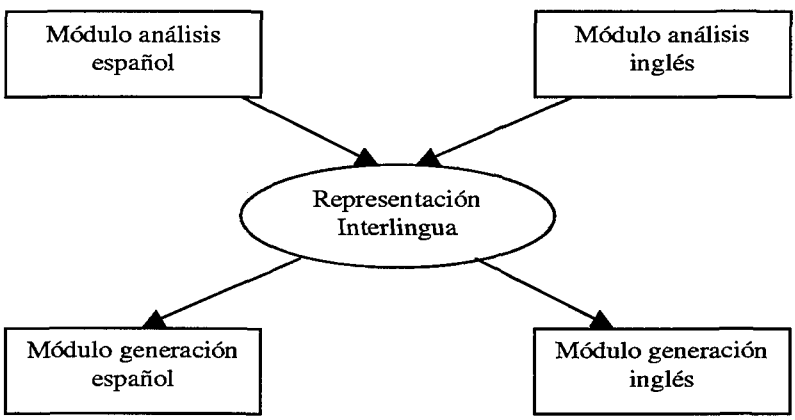


Figura 5.1. Arquitectura general del sistema AGIR

Hay que destacar, que en la representación interlingua obtenida se especifican las expresiones anafóricas de un modo par-

ticular, ya que cada anáfora contiene un enlace a su antecedente respectivo. Esta representación también permitirá identificar las cadenas de correferencia del texto, ya que éstas estarán formadas por aquellas expresiones anafóricas que tengan un enlace al mismo antecedente. Toda esta información junto con un estudio de las diferencias entre los idiomas origen y destino permitirá una generación correcta de las expresiones anafóricas (incluyendo las anáforas intersentenciales) en el idioma destino.

Aunque en esta Tesis nos hemos centrado en la generación de las expresiones anafóricas en español e inglés, nuestra aproximación se puede extender fácilmente a otros idiomas, es decir, es un sistema multilingüe en el que pueden existir múltiples módulos de análisis y generación en idiomas distintos, unidos por la representación interlingua.

El diseño de los módulos de análisis y generación de AGIR sigue un esquema de funcionamiento general independiente del idioma. En las dos siguientes secciones, presentaremos los módulos genéricos de análisis y generación de nuestro sistema.

5.2 Módulo de análisis del sistema AGIR

El objetivo principal del módulo de análisis consiste en analizar el texto del idioma origen y obtener una representación interlingua del mismo. El módulo consta de una serie de etapas que se llevan a cabo secuencialmente de modo que la salida de una etapa constituye la entrada de la etapa posterior.

Las etapas son las siguientes:

1. Análisis léxico y morfológico. Esta etapa recibe como entrada el texto en el idioma origen que se desea traducir al idioma destino. Para cada una de las unidades léxicas del texto se obtienen la categoría gramatical, información morfológica y características semánticas. Normalmente, esta información se extrae de los lexicones monolingües o de los etiquetadores léxico-morfológicos.
2. Análisis sintáctico. En la siguiente etapa se realiza el análisis sintáctico de cada una de las oraciones del texto en el que

se incluye la información léxica, morfológica y semántica para cada unidad léxica obtenida en la etapa anterior. Como resultado se obtienen los distintos constituyentes (sintagmas nominales, sintagmas preposicionales, núcleos verbales, etc.) de cada oración.

Además, hay que destacar, que en este módulo se genera una estructura que almacena la información de todos los posibles antecedentes que han ido apareciendo en las distintas oraciones con el objetivo de resolver las expresiones anafóricas pronominales intersentenciales.

3. Desambiguación del significado de las palabras. Esta etapa recibe como entrada la estructura sintáctica del texto enriquecida con información léxico, morfológica y semántica obtenida de etapas anteriores. A partir de esta estructura, se utiliza un módulo que proporciona un único significado o sentido para los nombres y verbos que aparecen en ella.
4. Resolución de problemas lingüísticos. En la cuarta etapa se resuelven los problemas lingüísticos (anáforas, elipsis, problemas de ambigüedad estructural, etc.) del texto[1].

   Este módulo recibe como entrada la estructura sintáctica enriquecida con la información que se ha obtenido en etapas anteriores. La salida estará constituida por otra estructura de información donde se habrán resuelto todos los problemas lingüísticos.

   Para el caso particular de resolución de la anáfora, se aplicará un algoritmo basado en restricciones y preferencias. Tras aplicar el algoritmo a las expresiones anafóricas del texto, se almacenará para cada anáfora el antecedente escogido junto con toda su información (morfológica, sintáctica, semántica, etc.).
5. Obtención de la representación interlingua. En esta etapa se genera la representación interlingua del texto a partir de la estructura obtenida en la etapa anterior.

[1] En esta Tesis nos centraremos exclusivamente en la resolución de las expresiones anafóricas pronominales y en el estudio de su posterior generación en el idioma destino.

Para ello, las oraciones se dividen en cláusulas y se genera para cada una de ellas una estructura de rasgos basada en papeles temáticos (AGENTE, TEMA, etc.). En cada cláusula se identificarán los distintos papeles temáticos que aparecen en ella y a cada uno de ellos se le asociará una entidad[2] del texto. En el caso concreto de un papel temático desempeñado por una expresión anafórica, se le asociará la entidad que representa su antecedente. Las cadenas de correferencia se identificarán cuando dos o más papeles temáticos distintos tengan su enlace a la misma entidad.

Una vez presentadas a grandes rasgos las distintas etapas del módulo de análisis del sistema AGIR, a continuación presentaremos en detalle la etapa en la que se obtiene la representación interlingua del texto origen. Con este estudio detallado se pretende explicar una de las principales aportaciones de este trabajo, la obtención de la representación interlingua, que consiste en obtener una representación independiente del idioma que permita generar (a partir de ella) el texto en el idioma destino.

El resto de etapas (análisis léxico, morfológico, sintáctico, etc.) se explicarán en el capítulo en el que se presenta la implementación del sistema AGIR (capítulo 6), en el que se incluye un estudio detallado de las distintas herramientas que se han utilizado en el mismo.

5.2.1 Representación interlingua en el sistema AGIR

La principal innovación que presenta el sistema AGIR consiste en la obtención de una representación interlingua del texto completo que se desea traducir. A diferencia del resto de sistemas de TA que realizan un tratamiento del texto oración por oración –Météo (Chandioux, 1976), Candide (Berger *et al.*, 1994), Eurotra (Allegranza *et al.*, 1991), KANT (Mitamura *et al.*, 1991), DLT (Schubert, 1988), Rosetta (Appelo & Landsbergen, 1986), CREST (Farwell & Helmreich, 2000), etc.– en nuestro sistema se obtiene

[2] Por *entidad* entendemos cualquier objeto o persona que aparece explícitamente en el proceso comunicativo, tal y como se definió en el capítulo 1.

una representación global del texto origen. Esta representación permitirá generar correctamente en el idioma destino la anáfora intrasentencial e intersentencial. Además, el sistema permite la identificación de las cadenas de correferencia y su posterior generación en el idioma destino. La representación interlingua del sistema AGIR ha sido presentada con detalle en diversas publicaciones (Peral & Ferrández, 2000a), (Peral *et al.*, 1999b).

Para obtener la representación interlingua (etapa 5 del módulo de análisis), previamente las distintas oraciones del texto se han analizado sintácticamente (etapa 2), se ha realizado la desambiguación del sentido de las palabras (etapa 3) y se han resuelto los problemas lingüísticos (etapa 4). La etapa en la que se obtiene la representación interlingua se puede dividir en varias fases:

1. Identificación de las cláusulas del texto. Debido a que una oración puede estar formada por más de una *cláusula*[3], la representación interlingua de nuestro sistema utiliza la cláusula (en vez de la oración) como unidad básica de representación. De este modo, un texto estará formado por un conjunto de cláusulas. Esta característica distingue, de nuevo, nuestro sistema respecto a los sistemas de TA tradicionales.

Para realizar la división del texto en cláusulas, previamente el sistema identifica las *conjunciones libres* que aparecen en él. Definimos conjunción libre como *aquella conjunción que se utiliza para coordinar las distintas proposiciones (unidades lingüísticas de estructura oracional constituidas por sujeto y predicado) de una oración compuesta*[4]. Si en una oración no aparece ninguna conjunción libre, esta oración tendrá una única cláusula[5]. En caso contrario, si se ha identificado alguna conjunción libre, las distintas proposiciones que ella enlaza serán definidas como cláusulas de la oración compuesta.

[3] Una cláusula se puede definir como *la parte mínima de una oración que contiene un verbo y un grupo de palabras que modifican al mismo.*

[4] Hay que destacar que aquellas conjunciones que coordinan sintagmas nominales o preposicionales no se identifican como conjunciones libres, ya que estos constituyentes no pueden formar, por sí solos, una proposición dentro de una oración.

[5] En inglés existen excepciones ya que estas conjunciones se pueden omitir (*The jury said it did find ...*) por lo que será necesario detectar la ausencia de estas conjunciones omitidas para separar las cláusulas correspondientes.

En el ejemplo 58 se muestra una oración con dos cláusulas:

(58) *[Juan y Pedro llegaron tarde]* porque *[ellos se quedaron dormidos]*.

En este ejemplo aparecen dos conjunciones: *y*, *porque*. La primera de ellas (*y*) coordina dos sintagmas nominales y no se identificará como conjunción libre. Por el contrario, la conjunción *porque* será identificada como conjunción libre que coordina dos cláusulas: (1) *Juan y Pedro llegaron tarde* y (2) *ellos se quedaron dormidos*.

2. Identificación de los papeles temáticos. El elemento principal de una cláusula es el verbo. Un verbo, en sí mismo, contiene información sobre los papeles temáticos de los constituyentes que lo acompañan. Para identificar estos papeles temáticos[6], se utiliza un lexicón monolingüe en el que se incluye para cada unidad léxica, además de otras características, información sintáctica, semántica y la relación entre ambas.

En la figura 5.2 aparece de un modo simplificado la entrada léxica correspondiente al verbo español *asistir*. Este verbo tiene 3 sentidos distintos (*ir*, *atender* y *socorrer*). Para cada uno de estos sentidos, se muestran un ejemplo con una frase, su estructura sintáctica y su estructura semántica con los papeles temáticos. La estructura sintáctica contiene las funciones sintácticas de los distintos constituyentes y asocia un identificador de entidad a cada uno de éstos. La estructura semántica asigna a cada uno de los constituyentes de la estructura sintáctica un papel temático. Estos papeles temáticos, que incluyen restricciones semánticas (aparecen entre paréntesis), se relacionan con la estructura sintáctica a través del identificador de la entidad correspondiente. Por ejemplo, para el sentido 1 de *asistir*, la entidad que es el sujeto constituye el *AGENTE* de la cláusula y la entidad que aparece en el objeto indirecto es la *LOCACIÓN* de la misma.

[6] Se han utilizado los papeles temáticos definidos por Haegeman (1991) presentados en la sección 2.7: AGENTE, TEMA, EXPERIMENTANTE, BENEFACTIVO, META, FUENTE, LOCACIÓN e INSTRUMENTO.

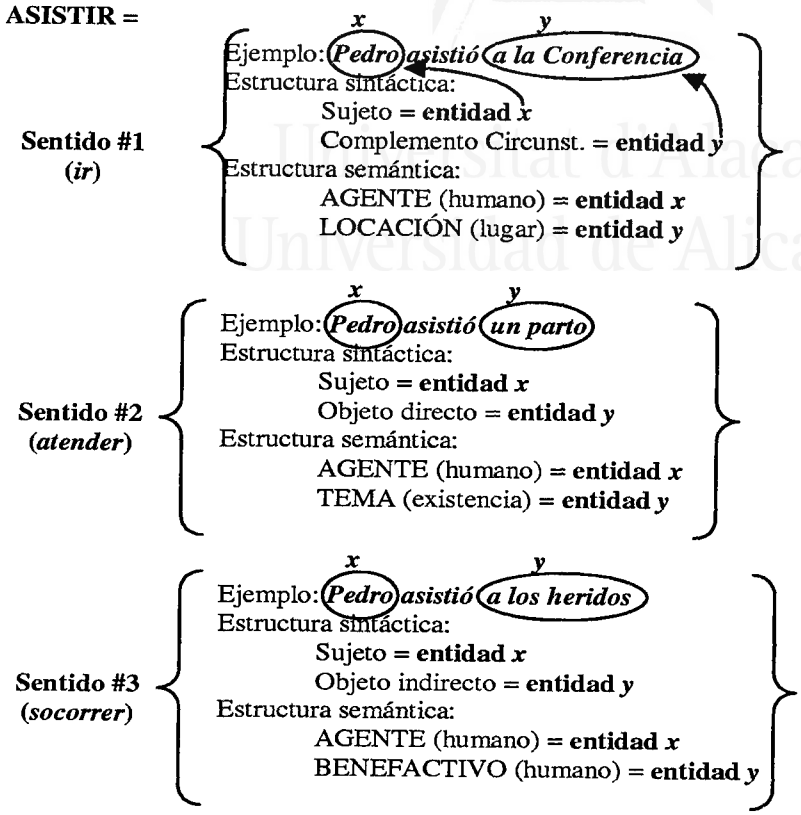


Figura 5.2. Representación en el lexicón de la unidad léxica *asistir*

3. Representación interlingua de las cláusulas. Tras identificar las distintas cláusulas del texto y los papeles temáticos que en ellas aparecen, la siguiente fase consiste en generar la representación interlingua para cada una de las cláusulas del texto. Para este fin, hemos usado una *estructura de rasgos* compleja con papeles temáticos para cada cláusula. Si realizamos una generalización de la estructura interlingua de una cláusula, en ella únicamente pueden aparecer 3 tipos distintos de estructuras de rasgos que se utilizan para representar:

1. Un verbo. Se representa con la estructura de rasgos *ACCIÓN* y contiene toda la información correspondiente al verbo prin-

cipal de la cláusula que permite su correcta generación en el idioma destino.

2. Un sintagma nominal. Representa una entidad del texto y contiene información del núcleo de la misma y de sus modificadores. Los papeles temáticos *AGENTE*, *TEMA*, *EXPERIMENTANTE* y *BENEFACTIVO* se representarán con esta estructura.
3. Un sintagma preposicional. Representa una entidad del texto que viene introducida por una preposición. Contiene información de la preposición y de la entidad que introduce. Los papeles temáticos *META*, *FUENTE*, *LOCACIÓN* e *INSTRUMENTO* se representarán con esta estructura.

Veamos la representación interlingua de una cláusula con el ejemplo 59:

(59) Los chicos de las montañas estaban en el jardín porque ellos estaban cogiendo flores.

En este ejemplo, aparece una oración que contiene dos cláusulas separadas por la conjunción libre *porque*. En la figura 5.3 se muestra la representación interlingua[7] para la primera cláusula (*Los chicos de las montañas estaban en el jardín*).

Como se observa en la figura 5.3, la representación interlingua de una cláusula es una estructura de rasgos formada por una serie de atributos que identifican la cláusula y por las distintas estructuras de rasgos de los papeles temáticos de sus constituyentes.

Los atributos de identificación de la cláusula son los siguientes:

- *ID_CLÁUSULA*, contiene el identificador de la cláusula dentro de la oración en la que se encuentra.
- *ID_ORACIÓN*, contiene el identificador de la oración.
- *CONJUNCIÓN*, almacena la conjunción que une las distintas cláusulas del texto.

[7] En esta figura sólo aparecen los atributos relevantes de cada papel temático de un modo simplificado. Se podrán añadir otros atributos adicionales a la representación interlingua final para disponer de la información necesaria para su posterior generación en el idioma destino.

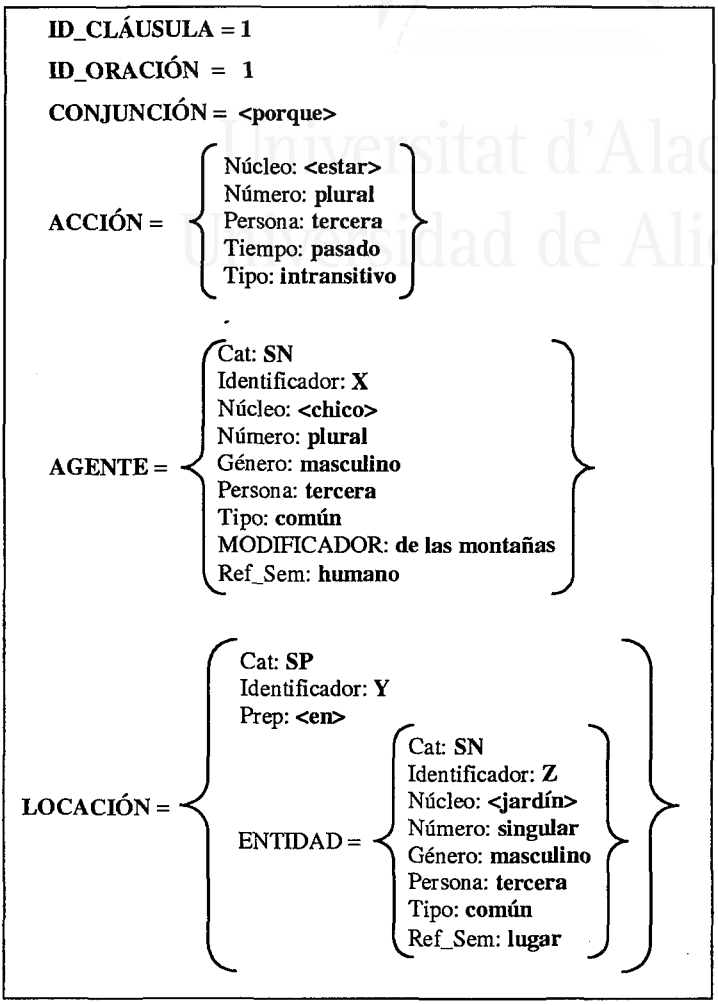


Figura 5.3. Representación interlingua de la cláusula *Los chicos de las montañas estaban en el jardín*

Por otra parte, en la cláusula de la figura 5.3 aparecen tres estructuras de rasgos correspondientes al verbo y a los papeles temáticos de sus constituyentes:

- *ACCIÓN*, formado por el verbo *estaban.*
- *AGENTE*, formado por el sintagma nominal *los chicos de las montañas.*

- *LOCACIÓN*, formado por el sintagma preposicional *en el jardín*.

A continuación, analizaremos cada una de estas estructuras de rasgos. La estructura de rasgos *ACCIÓN* tiene los siguientes atributos:

- *Núcleo*, contiene la unidad léxica interlingua independiente del idioma que representa al verbo principal de la cláusula. En esta Tesis se representarán las unidades léxicas interlingua entre los símbolos "<" y ">" para destacar que no se está especificando una palabra, sino que se está representando un *concepto*.
- *Número*, *Persona* y *Tiempo*, contienen las características gramaticales del verbo de la cláusula.
- *Tipo*, indica el tipo de verbo: transitivo, intransitivo, impersonal, etc.

El papel temático *AGENTE* tiene los siguientes atributos:

- *Cat*, contiene la categoría sintáctica del constituyente: sintagma nominal, sintagma preposicional, etc.
- *Identificador*, contiene el identificador de la entidad.
- *Núcleo*, contiene la unidad léxica interlingua que constituye el núcleo de la entidad.
- *Número*, *Género* y *Persona*, contienen las características gramaticales del núcleo del constituyente.
- *Tipo*, indica el tipo de la entidad: común, propio, pronombre, etc.
- *MODIFICADOR*, contiene toda la información de los modificadores (adjetivos, sintagmas preposicionales, etc.) del constituyente.
- *Ref_Sem*, contiene información semántica del núcleo del constituyente.

El papel temático *LOCACIÓN* tiene los siguientes atributos:

- *Cat*, contiene la categoría sintáctica del constituyente.
- *Identificador*, contiene el identificador del constituyente.
- *Prep*, contiene la preposición que introduce la ENTIDAD del sintagma preposicional.

- *ENTIDAD*, contiene todos los atributos propios de un sintagma nominal: *Cat*, *Identificador*, *Núcleo*, *Número*, *Género*, *Persona*, *Tipo*, *MODIFICADOR* y *Ref_Sem*.

4. Representación interlingua del texto completo. Como ya hemos mencionado, el sistema AGIR se caracteriza por la generación de una representación interlingua (basada en conceptos) del texto completo que se desea traducir, a diferencia del resto de sistemas de TA que se basan en la oración como unidad básica de tratamiento. Esta representación se fundamenta en la cláusula como unidad básica, y en ella se muestran las distintas entidades del texto y las relaciones que existen entre ellas. Para conseguir este objetivo, el esquema de la figura 5.3 se extiende para representar todo el discurso utilizando las cláusulas como unidades principales. En la figura 5.4 aparece la representación interlingua del texto completo del ejemplo 59.

En la parte izquierda de la figura 5.4 se representan las nuevas entidades u objetos del discurso. Estos objetos se denominan *ENTIDADES* y contienen los atributos propios de un sintagma nominal ya mencionados anteriormente. Las *ENTIDADES* pueden representar cualquier papel temático que aparezca en una cláusula.

La representación de las distintas entidades que aparecen en el texto, así como de las relaciones existentes entre las mismas, constituye otra característica innovadora de nuestro sistema. Esta representación permite la resolución de la anáfora intrasentencial e intersentencial, así como de la resolución y generación de las cadenas de correferencia del texto. Del mismo modo, permite la resolución de problemas lingüísticos variados (elipsis, omisión de sujeto, etc.).

En la parte derecha del gráfico se muestran, de un modo simplificado, las distintas cláusulas del texto. Contienen los atributos de identificación (*ID_CLÁUSULA*, *ID_ORACIÓN* y *CONJUNCIÓN*), el verbo y los papeles temáticos que hayan aparecido en la misma. Estos papeles temáticos están relacionados (mediante la variable que representa el identificador de la entidad) con la *ENTIDAD* a la que ellos se refieren.

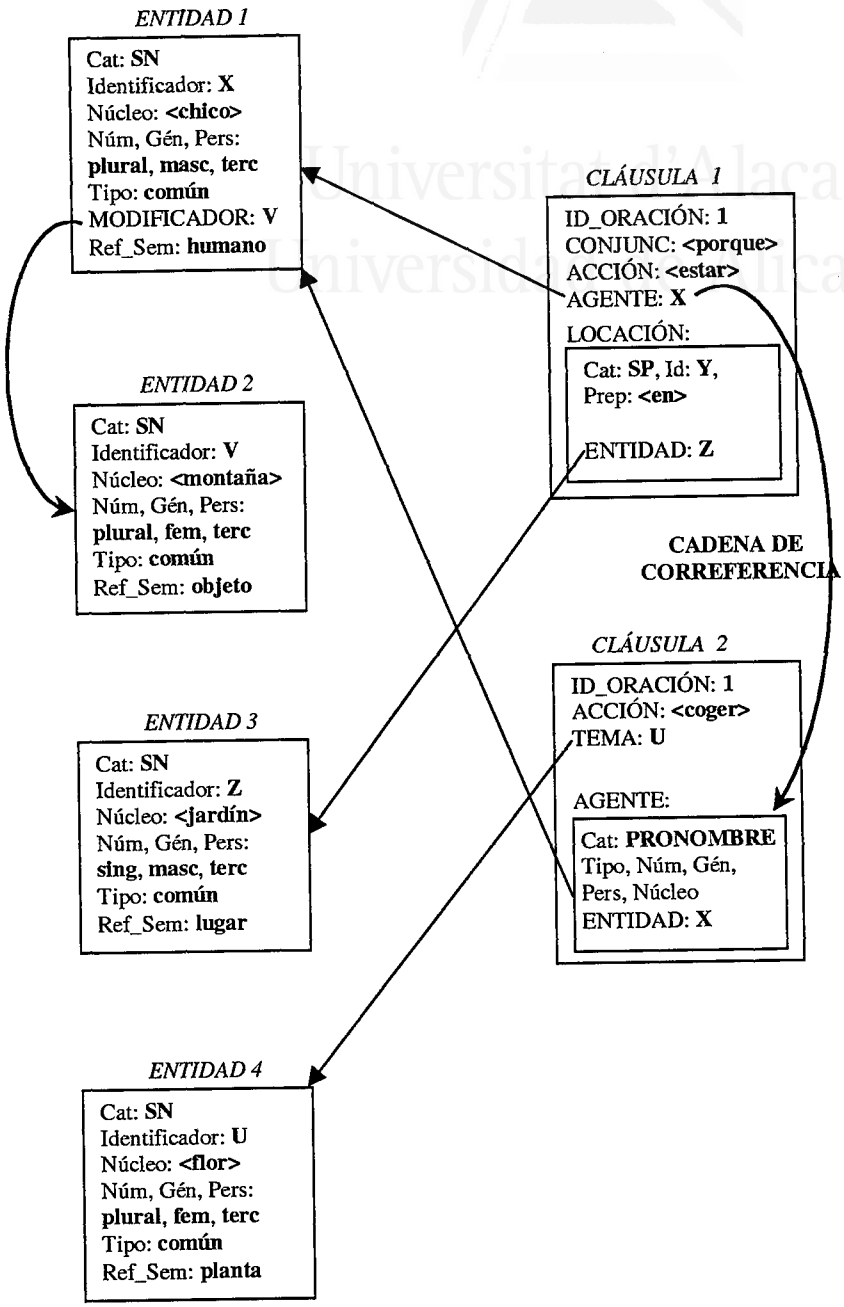


Figura 5.4. Representación interlingua de un fragmento de texto

En la figura 5.4 podemos distinguir cuatro *entidades* y dos *cláusulas*. Las *entidades* son las siguientes: *ENTIDAD* 1 ("chico"), *ENTIDAD* 2 ("montaña"), *ENTIDAD* 3 ("jardín") y *ENTIDAD* 4 ("flor"). Además, se observa una relación entre dos *entidades* (número 1 y número 2) debido a que la *ENTIDAD* 1 (un SN) contiene un *MODIFICADOR* (un SP).

La *CLÁUSULA* 1 contiene: *ID_ORACIÓN* ("1"), *CONJUNCIÓN* ("porque"), *ACCIÓN* ("estar"), *AGENTE* ("X", el enlace a la *ENTIDAD* 1) y *LOCACIÓN* (un SP que contiene un enlace a la *ENTIDAD* 3). La *CLÁUSULA* 2 contiene: *ID_ORACIÓN* ("1"), *ACCIÓN* ("coger"), *AGENTE* (un *PRONOMBRE*, que contiene el enlace a la *ENTIDAD* 1) y *TEMA* ("U", el enlace a la *ENTIDAD* 4).

La cadena de correferencia se puede identificar si distintos papeles temáticos tienen enlaces a la misma entidad. En este ejemplo, los *AGENTES* de la *CLÁUSULA* 1 y la *CLÁUSULA* 2 ("los chicos" y "ellos") tienen sus enlaces a la misma *ENTIDAD* (la número 1). Por lo tanto, se establece una cadena de correferencia entre ambos constituyentes. Hay que destacar que estos enlaces pueden ocurrir entre constituyentes de diferentes cláusulas u oraciones. De este modo, el sistema global es capaz de generar las anáforas intersentenciales y de identificar las cadenas de correferencia del texto.

5.3 Módulo de generación del sistema AGIR

El módulo de generación recibe como entrada la representación interlingua del texto origen. La salida de este módulo es la traducción de este texto al idioma destino, es decir, expresar el significado del texto origen con palabras del idioma destino. En esta Tesis nos hemos centrado exclusivamente en la generación en el idioma destino de las anáforas pronominales originadas por los pronombres personales de tercera persona y los cero pronombres españoles.

Hay que destacar que la omisión del sujeto pronominal (*cero anáforas* o *cero pronombres* en la posición de sujeto) es un

fenómeno muy habitual en español, ya que sólo se menciona dicho sujeto si se quiere enfatizar el mismo o se quiere distinguir entre varios posibles candidatos a sujeto. En el resto de los casos, el sujeto se podría omitir.

En el ejemplo 60 el pronombre *ella* se podría omitir de la segunda oración. La mención explícita del pronombre en este ejemplo se utiliza para distinguir de los dos posibles candidatos, cuál es el sujeto de la segunda oración.

(60) Julián se encontró con $[Ana]_i$ en la estación.
$Ella_i$ estaba esperando a su hijo.

En esta Tesis estudiamos cuál es el pronombre correspondiente en el idioma destino al pronombre que aparece en el texto origen. Este pronombre se tratará posteriormente para generarlo de un modo adecuado en el idioma destino.

El módulo de generación consta de las siguientes etapas: generación semántica, generación sintáctica y generación morfológica. A continuación presentaremos cada una de las 3 etapas. Aunque la aproximación presentada es multilingüe, en esta Tesis nos hemos centrado en la generación de las anáforas pronominales en inglés y español.

5.3.1 Generación semántica

La etapa de la generación semántica consiste en generar la estructura profunda del texto en el idioma destino a partir de la representación interlingua. En la figura 5.5 se muestra la estructura profunda de la primera cláusula del ejemplo 59 (cuando se traduce al inglés) tras la etapa de la generación semántica.

Básicamente, en este módulo se identifica el verbo de la cláusula y se asignan a cada uno de los papeles temáticos de la misma una función sintáctica. Además, la información que se encuentra almacenada en la estructura interlingua (gramatical, sintáctica, semántica, información sobre el antecedente de las expresiones anafóricas y las cadenas de correferencias, etc.) se transmite al constituyente correspondiente de la nueva estructura sintáctica.

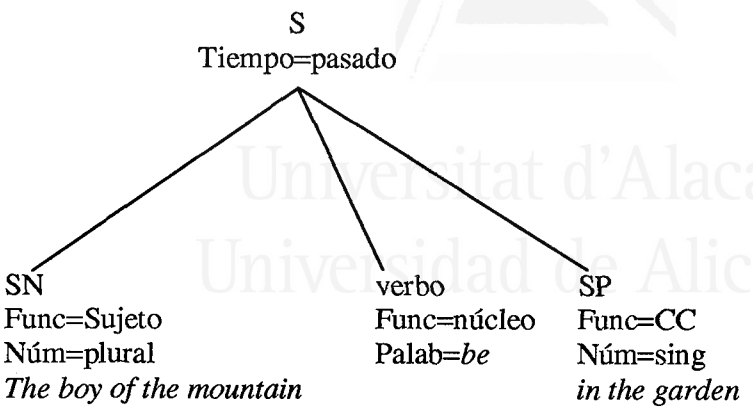


Figura 5.5. Estructura de la cláusula *Los chicos de las montañas estaban en el jardín* obtenida tras la generación semántica

Así, en la figura 5.5 se observa que se ha seleccionado como núcleo de la cláusula el verbo estar (*be*). Al papel temático *AGENTE* (tiene un enlace a la *ENTIDAD* 1 –chico–, que contiene un *MODIFICADOR*, el enlace a la *ENTIDAD* 2 –montaña–) se le asigna la función de Sujeto (*the boy of the mountain*). Por último, al papel temático *LOCACIÓN* (tiene un enlace a la *ENTIDAD* 3 –jardín–) se le ha asignado la función de Complemento Circunstancial (*in the garden*)[8].

Como se puede observar en la figura 5.5, en la estructura sintáctica obtenida las palabras en el idioma destino se presentan en su forma base. Esta estructura se utilizará como entrada de las etapas posteriores, generación sintáctica y generación morfológica, tras las cuales se obtendrá el texto en el idioma destino en su forma correcta.

5.3.2 Generación sintáctica

En la etapa de la generación sintáctica, la estructura sintáctica obtenida tras la etapa de la generación semántica se transforma mediante "reglas transformacionales" en una nueva estructura que

[8] Hay que destacar que para obtener esta representación, previamente se ha realizado la correspondiente selección de las unidades léxicas en el idioma destino.

representa la estructura superficial del texto en el idioma destino, y que contiene las funciones gramaticales y características de las palabras en el idioma destino. La tarea principal de la generación sintáctica es la ordenación de los constituyentes en la secuencia correcta en el idioma destino.

El objetivo fundamental de esta Tesis consiste en el desarrollo de un sistema interlingua que permita la correcta generación de la anáfora pronominal en el idioma destino (español o inglés), es decir, no pretendemos realizar la traducción de un texto completo de un idioma a otro, sino que nuestro estudio se centra exclusivamente en la generación de los pronombres. Por esta razón, en las etapas de generación sintáctica y generación morfológica únicamente presentaremos las diferencias relativas al tratamiento de los pronombres en español e inglés. El estudio realizado sobre las diferencias (a las que hemos denominado *discrepancias*) entre ambos idiomas permitirá la correcta generación de las expresiones anafóricas pronominales en el idioma destino. Estas discrepancias se han presentado exhaustivamente en diversas publicaciones (Peral *et al.*, 1999b), (Peral, 1999), (Peral *et al.*, 1999a), (Peral & Ferrández, 2000a).

En la etapa de generación sintáctica, podemos encontrar las siguientes discrepancias entre español e inglés: *cero pronombres* con función de sujeto y pronombres pleonásticos.

Cero pronombres (*zero pronouns*) con función de sujeto. Como ya hemos comentado a lo largo de la Tesis, la gramática española permite la omisión de los pronombres con función de sujeto en las oraciones. Estos pronombres omitidos se denominan *cero anáforas* o *cero pronombres* (*zero pronouns*). Mientras que en otros idiomas los *cero pronombres* pueden aparecer en la posición de sujeto u objeto de la oración (por ejemplo, el japonés), en español los *cero pronombres* sólo pueden aparecer en la función de sujeto de la oración.

En inglés, el uso de los pronombres con función de sujeto es obligatorio, aunque podemos encontrar algunos ejemplos en los que se omite el sujeto pronominal[9]. En el ejemplo 61, el pronombre

[9] Tal y como se define en el *Linguistic Glossary* –http://www.sil.org/linguistics/glossary, visitada el 16/06/01–, "*el sujeto es una relación gramatical que*

she se ha omitido de la segunda cláusula coordinada (se representa por el símbolo $\emptyset$, que indica la posición del pronombre omitido).

(61) *[Ross]* carefully folded his trousers and $\emptyset_i$ climbed into bed.

Otros idiomas en los que aparecen construcciones elípticas originadas por *cero pronombres* son el italiano, tailandés, chino o japonés.

En los trabajos de Peral y Ferrández (Peral & Ferrández, 2000b; Ferrández & Peral, 2000) se presentan el tratamiento de los *cero pronombres* llevado a cabo en el sistema AGIR.

Para realizar una traducción correcta de español a inglés, los *cero pronombres* con función de sujeto se deben detectar en español para que puedan ser generados convenientemente en el idioma destino.

En la detección de la omisión del sujeto pronominal, hay que destacar una serie de excepciones. Si el verbo de la oración en español es imperativo o impersonal, la oración no tiene sujeto, por lo que no hay que detectarlo. En su traducción al inglés, estas oraciones no tendrán sujeto o tendrán un sujeto *dummy* (definido en el *Linguistic Glossary* como "*una unidad gramatical que no tiene significado pero completa una oración para hacerla gramatical*"). En los ejemplos 62 y 63 se muestran una oración en español con un verbo imperativo y con un verbo impersonal respectivamente y sus traducciones al inglés en las que no aparece ningún sujeto.

(62) **E** ¡$\emptyset$ Coge los papeles del suelo!
I $\emptyset$ Pick up the papers on the floor!

(63) **E** $\emptyset$ Hubo un accidente de tráfico.

se caracteriza por ciertas propiedades sintácticas independientes. Una de ellas formula que puede ser obligatoriamente u opcionalmente eliminado de determinadas construcciones gramaticales, tales como las cláusulas adverbiales, las de complemento y las coordinadas".

I Ø There was a road accident.

Si el verbo es impersonal pero se utiliza en una serie de expresiones (temporales, relacionadas con las condiciones climáticas, construcciones de la voz pasiva, etc.)[10], en español no tiene sujeto pero en la traducción al inglés se utiliza el pronombre *it* como sujeto *dummy*. En el ejemplo 64 se muestra un ejemplo con una expresión temporal.

(64) **E** Ø Son las tres en punto.

I *It* is three o'clock.

En el resto de los casos (si el verbo no es imperativo o impersonal) la omisión del sujeto de una oración se detecta cuando, tras haber realizado el análisis sintáctico de la misma, no se ha identificado ningún constituyente que realice dicha función.

Tras haber detectado el *cero pronombre* con función de sujeto, la información gramatical del mismo (persona y número) se extrae del verbo de la oración, ya que el sujeto de la oración debe concordar en persona y número con el verbo de la misma. Esta información se utilizará posteriormente en la etapa de resolución de problemas lingüísticos (en concreto, en la resolución de la anáfora pronominal) para identificar el antecedente del *cero pronombre*.

(65) *[Juan]*$_i$ era un boxeador profesional. Ø$_i$ Perdió únicamente dos combates.

En el ejemplo 65, el sujeto pronominal de la segunda oración ha sido omitido. Del verbo de la oración (*perdió*) se extrae la información de persona y número (tercera persona y singular) del sujeto de la misma. Esta información se utilizará en la resolución de la anáfora que identificará el sintagma nominal *Juan* como antecedente de la misma.

Una vez que se ha identificado el antecedente del *cero pronombre*, se obtiene la representación interlingua de la oración, que

[10] Estas expresiones serán explicadas con detalle en la siguiente sección, en la que se presentan los pronombres pleonásticos.

contiene toda la información necesaria para la correcta generación en inglés del pronombre omitido.

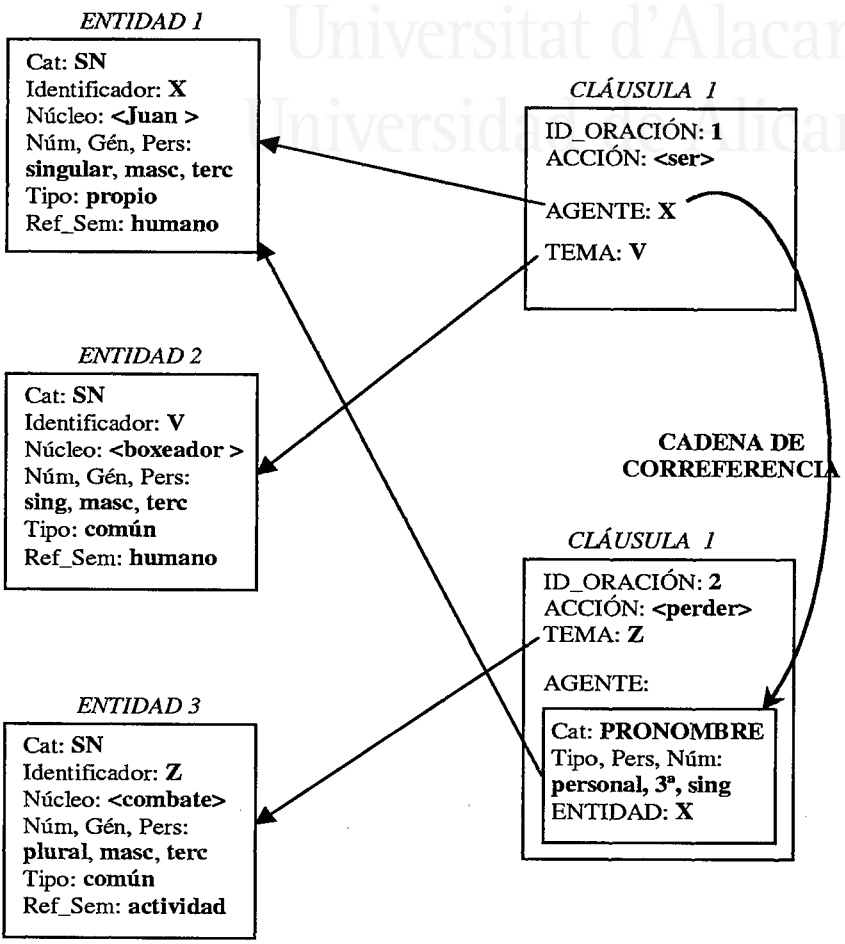


Figura 5.6. Representación interlingua con *cero pronombre* con función de sujeto

En la figura 5.6 aparece la representación interlingua del ejemplo 65. En la cláusula 1 de la oración 2 se observa que el *AGENTE* de la misma es un pronombre (el *cero pronombre* con función de sujeto) que tiene un enlace a la *ENTIDAD* 1 (el sintagma nominal *Juan*). En la representación interlingua aparece toda la información necesaria para la generación de este pronombre en inglés, ya

que la información de persona y número se extraen del *PRONOMBRE* (tercera persona y singular) y la información del género se extrae del *Núcleo* de su antecedente (masculino). Con esta representación interlingua se generaría en inglés el texto del ejemplo 66, donde aparece el *cero pronombre* generado correctamente (el pronombre inglés *he*).

(66) *[Juan]$_i$* was a professional boxer. *He$_i$* only lost two fights.

Como se puede observar en este ejemplo, es fundamental resolver la anáfora pronominal para su correcta generación en inglés, ya que la información del género del *cero pronombre* se extrae del núcleo de su antecedente, y ello nos permitirá decidir entre el pronombre inglés masculino *he* y el pronombre inglés femenino *she*.

En algunas ocasiones la información del género del *cero pronombre* se puede obtener directamente de la oración sin tener que resolver la anáfora pronominal. Por ejemplo, si el verbo de la oración es copulativo, la información del género del *cero pronombre* se puede extraer del objeto.

(67) **E** Pedro vio a *[Ana]$_i$* en el parque. *Ø$_i$* Estaba muy guapa.

I Pedro saw *[Ana]$_i$* in the park. *She$_i$* was very beautiful.

En el ejemplo 67, el verbo de la oración (*estaba*) es copulativo y su objeto (*guapa*) concuerda en género y número (*femenino singular*) con el sujeto de la oración[11]. Por esta razón, podemos concluir que el sujeto pronominal omitido de la oración es un pronombre femenino, singular y en tercera persona, es decir, el

[11] Esta afirmación será cierta siempre que el objeto de la oración pueda tener formas lingüísticas distintas para el género masculino y femenino como sucede en el ejemplo 67 (guapa: femenino, guapo: masculino). Si el objeto no presenta estas dos formas, no se podrá extraer la información del género, como sucede en las oraciones *Pedro es un genio* y *Ana es un genio.*

pronombre *ella*, que generará el correspondiente pronombre inglés *she*.

Por último, hay que mencionar que en español es posible que el sujeto de la oración se encuentre situado a la derecha del verbo. Esto es debido a que el español no tiene un orden sintáctico estricto y los distintos constituyentes de una oración pueden aparecer, prácticamente, en cualquier orden. Por ejemplo, la oración del ejemplo 68 tiene el sujeto de la misma a la derecha del verbo.

(68) Abrió Antonio la puerta y todos se quedaron sorprendidos.

Aunque en esta oración aparentemente existe un *cero pronombre* con función de sujeto (ya que no aparece ningún sintagma nominal antes que el verbo), esto no es cierto ya que el sujeto es el sintagma nominal *Antonio*. Si se realiza una correcta identificación de los papeles temáticos, este sintagma nominal se identificaría como *AGENTE*, con lo que se obtendría la representación interlingua adecuada que permitiría la correcta generación en el idioma destino.

Pronombres pleonásticos. Según el Diccionario de la Real Academia Española, un pleonasmo[12] es una "*figura de construcción, que consiste en emplear en la oración uno o más vocablos innecesarios para el recto y cabal sentido de ella, pero con los cuales se da gracia o vigor a la expresión*". En inglés, existe un tipo de pronombre de uso frecuente en los textos que en ciertas construcciones tiene un uso pleonástico: el pronombre *it*.

El pronombre inglés *it* tiene una gran variedad de usos. Normalmente son anafóricos y sus antecedentes aparecen previamente en el texto.

(69) Look at *[that dog]*$_i$. *It*$_i$ can´t run.

En otras ocasiones, el pronombre *it* tiene su antecedente en el texto, pero se establece una relación catafórica entre ambos.

[12] Pleonástico es el adjetivo que indica que algo pertenece al pleonasmo.

(70) When it_i fell, *[the bottle]*$_i$ broke.

Por último, existen los pronombres *it* pleonásticos. Estos pronombres no se refieren a nada, es decir, no son anafóricos, son semánticamente vacíos, pero aparecen debido a una serie de reglas de la gramática inglesa, es decir, actúan como una palabra *dummy*.

(71) **I** *It* is raining.
E Ø Está lloviendo.

En Traducción Automática es muy importante la detección de estos pronombres pleonásticos, ya que como no son referenciales, normalmente no se traducen al idioma destino. En el ejemplo 71 se observa que el pronombre pleonástico *it* no se traduce al español.

En nuestro sistema, estos pronombres se detectan antes de la etapa de la resolución de problemas lingüísticos, por lo que no se incluyen, a diferencia del resto de pronombres anafóricos, en el proceso de búsqueda de su antecedente. Estos pronombres aparecerán marcados como pleonásticos en la representación interlingua, no tendrán antecedente y no se generarán en español. Además, estos pronombres no anafóricos nunca podrán formar parte de una cadena de correferencia.

Para detectar los pronombres *it* pleonásticos se han construido una serie de reglas que permiten la identificación automática de esta clase de pronombres. Estas reglas se basan en el estudio realizado por otros autores (Paice & Husk, 1987; Lappin & Leass, 1994; Denber, 1998) que abordan el problema de modo similar, es decir, proponen unos métodos que se fundamentan en el reconocimiento de patrones para identificar el uso pleonástico del pronombre *it*.

Las distintas reglas usadas en el sistema AGIR son las presentadas en el trabajo de Paice & Husk (1987). En este trabajo se realiza una descripción detallada de un número de construcciones que identifican el uso pleonástico del pronombre *it* y que abordan

prácticamente todas las construcciones de este tipo. Las distintas reglas planteadas son las siguientes[13]:

1. *it . . . adjetivo_de_estado . . . to + verbo_infinitivo*
   donde *adjetivo_de_estado* representa un adjetivo del tipo: important, impossible, necessary, sufficient, useful, unusual, etc.

   (72) *it* is necessary to limit the speed . . .

2. *it . . . conjunción_that*

   (73) *it* is inevitable that I repeat . . .

3. *it . . . adjetivo_de_conocimiento + whether/if/what/how/why/when/where*
   donde *adjetivo_de_conocimiento* representa un adjetivo del tipo: known, certain, clear, doubtful, unknown, questionable, wondered, etc.

   (74) *it* is not known whether such an exchange was performed . . .

4. *it . . . which/who*

   (75) *it* is rather the teacher or parent who is moved by them . . .

5. *it . . . verbo_be . . . palabra_metereológica*
   donde *verbo_be* representa cualquier tiempo del verbo be y *palabra_metereológica* indica una palabra del tipo: rain, drizzle, snow, warm, stormy, foggy, spring, summer, etc.

   (76) *it* was drizzling . . .

6. *it . . . verbo_be . . . construcción_temporal*

[13] Los puntos suspensivos de cada una de las reglas indican que en esa posición pueden aparecer cualquier palabra o conjunto de palabras.

donde *construcción_temporal* representa una construcción que indica una expresión temporal: five o'clock, tea time, high time, etc.

(77) *it* is high time he found a job . . .

7. *it . . . worth . . . verbo_gerundio*

(78) *it* is now worth considering . . .

8. *, it . . . (3 palabras o menos) . . . ,*

(79) Each chromosome, *it* is now clear, has a characteristic . . .

9. *expresiones idiomáticas*

(80) Stick *it* out . . .
as *it* were . . .

Como se puede observar en las distintas construcciones presentadas, el uso pleonástico del pronombre *it* es bastante común en inglés. Por esta razón, su detección se tiene que abordar en todo sistema de Traducción Automática para permitir un tratamiento adecuado en el idioma destino.

Respecto al español, también existen ciertos pronombres que tienen un uso no referencial y, por lo tanto, no se deben generar en inglés. Por ejemplo, en algunas construcciones el pronombre *se* sirve únicamente para intensificar el significado del verbo y no tiene un significado anafórico (ejemplo 81).

(81) **E** Pascual *se* fue pronto.
I Pascual left early.

En otras construcciones, los pronombres españoles no se traducen al inglés cuando en la oración aparece también el sustantivo al que hace referencia (ejemplos 82 y 83).

(82) **E** *Se* lo di a tu hermano.
I I gave it to your brother.

(83) **E** A Rubén *le* vi ayer.
I I saw Rubén yesterday.

Por último, existen construcciones impersonales en las que el pronombre tiene un uso impersonal y no se traduce al inglés (ejemplo 84).

(84) **E** *Se* habla inglés.
I English spoken.

Todas estas construcciones en las que los pronombres no son referenciales se deben tratar en español para que se puedan generar correctamente en el idioma destino.

5.3.3 Generación morfológica

En la etapa de generación morfológica se tratarán y resolverán las discrepancias de número y género (Peral *et al.*, 1999b; Peral, 1999; Peral *et al.*, 1999a; Peral & Ferrández, 2000a) ocasionadas por las diferencias entre español e inglés en el tratamiento de los pronombres.

Discrepancias de número. El problema de las discrepancias de número se produce por las diferencias del número gramatical entre palabras en distintos idiomas que expresan el mismo concepto.

Estas palabras en el idioma inglés son clasificadas por Alcaraz & Moody (1997) dentro de la categoría de *nombres colectivos* y las define como "*aquellos nombres que requieren un verbo en plural, aunque en su forma externa no aparece morfema de plural, y denotan un conjunto de personas*". Por el contrario, en español estos nombres colectivos tienen número singular y van acompañados de un verbo en singular. Por esta razón, en la traducción inglés-español-inglés estos nombres serán referidos por un pronombre en plural en el idioma origen y por un pronombre en singular en el idioma destino o viceversa.

Por ejemplo, en inglés la palabra *público* es plural y requiere un verbo en plural, mientras que en español es singular y el verbo de la oración en la que aparece es singular[14].

(85) **E** *[El público]$_i$* es muy exigente. *Éste$_i$* siempre tiene la razón.

I *[Public]$_i$* are very demanding. *They$_i$* are always right.

En el ejemplo 85.e se puede observar que la palabra *público* es singular y ha sido reemplazada en la segunda oración por un pronombre en singular. Sin embargo, en la traducción a inglés (ejemplo 85.i) se utiliza un pronombre en plural para referirnos a la misma palabra. En ambos ejemplos, el verbo y el sujeto concuerdan en el número gramatical.

Las discrepancias de número no sólo existen en la traducción entre español e inglés. Hay otras traducciones, por ejemplo entre alemán e inglés, en las que este tipo de problema también se produce.

Para tener en cuenta las discrepancias de número en la generación del pronombre en el idioma destino, en nuestro sistema se ha construido un conjunto de reglas morfológicas (de número) que se deben de tener en cuenta en la etapa de la generación morfológica. Dependiendo de la función del pronombre se ha hecho la siguiente clasificación: pronombres con función de sujeto y pronombres con función de complemento.

- **Pronombres con función de sujeto.** Esta clase de pronombres se pueden identificar en la representación interlingua porque tienen el papel temático de *AGENTE* en una *CLÁUSULA*. Además, tienen los enlaces a las *ENTIDADES* que son sus antecedentes correspondientes.

[14] Cuando indicamos que estas palabras requieren un verbo en singular o en plural se cumple siempre que desempeñen la función de sujeto en la oración en la que aparecen; en caso contrario, el verbo debe concordar con el sujeto de la misma. Por ejemplo, en la oración *Los anuncios previos a la proyección de una película no interesan al público*, aunque aparece la palabra *público*, el verbo está en plural porque concuerda con el sujeto.

En la figura 5.7 se muestran algunos ejemplos de las reglas morfológicas que se han construido para la traducción de español a inglés de pronombres con función de sujeto con discrepancias de número.

AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <público>)
→ **they** (PRONOMBRE + 3ª persona + plural)

AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <gente>)
→ **they** (PRONOMBRE + 3ª persona + plural)

AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <policía>)
→ **they** (PRONOMBRE + 3ª persona + plural)

AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <juventud>)
→ **they** (PRONOMBRE + 3ª persona + plural)

AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <ganado>)
→ **they** (PRONOMBRE + 3ª persona + plural)

Figura 5.7. Discrepancias de número. Traducción español-inglés de pronombres con función de sujeto

La parte izquierda de las reglas morfológicas de la figura 5.7 contienen la representación interlingua del pronombre del idioma origen y la parte derecha contienen el pronombre en el idioma destino.

Por ejemplo, para la generación en inglés del pronombre *éste* del ejemplo 85.e, la representación interlingua tendrá un *AGENTE* constituido por un *PRONOMBRE* (*éste*, tercera persona y singular) que tendrá un enlace a su antecedente, la *ENTIDAD* cuyo núcleo es el nombre *público*. Después de consultar las reglas morfológicas, se genera el correspondiente pronombre inglés (*they*, tercera persona y plural).

Del mismo modo, se construyen un conjunto de reglas morfológicas para la generación de los pronombres en español. En la figura 5.8 se muestran algunos ejemplos de estas reglas.

AGENTE (PRONOMBRE + 3ª persona + plural + antecedente <public>)
→ **éste** (PRONOMBRE + 3ª persona + masculino + singular)

AGENTE (PRONOMBRE + 3ª persona + plural + antecedente <police>)
→ **ésta** (PRONOMBRE + 3ª persona + femenino + singular)

AGENTE (PRONOMBRE + 3ª persona + plural + antecedente <cattle>)
→ **éste** (PRONOMBRE + 3ª persona + masculino + singular)

Figura 5.8. Discrepancias de número. Traducción inglés-español de pronombres con función de sujeto

- **Pronombres con función de complemento.** Los pronombres con función de complemento se identifican en la representación interlingua porque ellos tienen un papel temático distinto al de *AGENTE*. Pueden desempeñar los papeles temáticos de *EXPERIMENTANTE*, *BENEFACTIVO*, *TEMA*, etc.

En el ejemplo 86 aparece un texto en inglés y español con un pronombre de complemento con discrepancias de número.

(86) **E** Pedro cuidaba mucho *[su ganado]*$_i$. Él *lo*$_i$ alimentaba con heno.

I Pedro took great care over *[his cattle]*$_i$. He fed *them*$_i$ on hay.

En el ejemplo 86.e se utiliza un pronombre de complemento en singular (*lo*) para referirnos al *ganado*. En su traducción al inglés (ejemplo 86.i) se utiliza un pronombre de complemento en plural (*them*) para referirnos a la misma palabra.

NO_AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <policía>)
→ **them** (PRONOMBRE + 3ª persona + plural)

NO_AGENTE (PRONOMBRE + 3ª persona + singular + antecedente <ganado>)
→ **them** (PRONOMBRE + 3ª persona + plural)

Figura 5.9. Discrepancias de número. Traducción español-inglés de pronombres con función de complemento

En las figuras 5.9 y 5.10 se muestran unos ejemplos de las reglas morfológicas que se han construido para resolver estas discrepancias generadas por los pronombres de complemento en la traducción español-inglés e inglés-español respectivamente.

TEMA (PRONOMBRE + 3ª persona + plural + antecedente <police>)
→ **la** (PRONOMBRE + 3ª persona + femenino + singular)

TEMA (PRONOMBRE + 3ª persona + plural + antecedente <cattle>)
→ **lo** (PRONOMBRE + 3ª persona + masculino + singular)

NO_TEMA_y_NO_AGENTE (PRONOMBRE + 3ª persona + plural +
+ antecedente <police>) → **ésta** (PRONOMBRE + femenino + singular)

NO_TEMA_y_NO_AGENTE (PRONOMBRE + 3ª persona + plural +
+ antecedente <cattle>) → **éste** (PRONOMBRE + masculino + singular)

Figura 5.10. Discrepancias de número. Traducción inglés-español de pronombres con función de complemento

Discrepancias de género. Las discrepancias de género se originan por las diferencias morfológicas existentes entre distintos idiomas. Para el tratamiento de los idiomas español e inglés, hay que tener en cuenta que el primero es un idioma con más riqueza morfológica que el segundo. En español las marcas de concordancia de las distintas palabras de una oración están más explícitas que en inglés.

(87) **E** Las chicas altas y guapas trabajaron hasta medianoche.

I The tall pretty girls worked until midnight.

El ejemplo 87.e tiene marcas de concordancia de género (femenino, *-a*) y de número (plural, *-s*) para el nombre y sus modificadores: nombre (*chic<u>as</u>*), artículo (*<u>las</u>*), adjetivos (*alt<u>as</u>* y *guap<u>as</u>*). Además, el verbo tiene una marca de número (plural, *-on*): verbo (*trabajar<u>on</u>*). En su traducción al inglés (ejemplo 87.i) se observa que no existen marcas de concordancia para las distintas palabras

de la oración (artículo, adjetivos y verbo) y únicamente podemos afirmar que el nombre es plural (*girl<u>s</u>*) porque tiene una marca de plural, *-s*.

En estos idiomas en que las marcas de concordancia no son manifiestas, el orden sintáctico necesariamente es más estricto que en aquéllos, como el español, en que aparecen marcados los morfemas de género, número, etc.

En cuanto al tratamiento de los pronombres personales, también existen unas diferencias en las marcas de género entre español e inglés, principalmente en los pronombres en plural. Así, por ejemplo, el pronombre inglés *we* puede ser traducido al español por *nosotr<u>o</u>s* (masculino) o *nosotr<u>a</u>s* (femenino), el pronombre *you* por *vosotr<u>o</u>s* o *vosotr<u>a</u>s* y el pronombre *they* por *ell<u>o</u>s* o *ell<u>a</u>s*. Para realizar una correcta traducción de estos pronombres al español, es necesario identificar el antecedente de los mismos y averiguar su género.

(88) **I** *[Women]$_i$* were in the shop. *They$_i$* were buying gifts for their husbands.

E *[Las mujeres]$_i$* estaban en la tienda. *Ellas$_i$* estaban comprando regalos para sus maridos.

En el ejemplo 88.i el pronombre *they* es válido tanto para masculino como para femenino. En su traducción al español (ejemplo 88.e) debemos escoger entre dos posibilidades: *ell<u>o</u>s* y *ell<u>a</u>s*. Para realizar esta elección, se debe identificar el antecedente del pronombre (el sintagma nominal *women*) y extraer su género (femenino). Por lo tanto, el pronombre español correspondiente será *ellas*.

Estas discrepancias no significan que todas las palabras en español tengan más información morfológica que sus homólogas en inglés. Por ejemplo, los adjetivos posesivos en español no tienen información de género (*<u>su</u> casa*) mientras que en inglés se distingue si el poseedor es masculino (*<u>his</u> house*) o femenino (*<u>her</u> house*).

Las discrepancias de género no se producen exclusivamente en la traducción entre inglés y español. Este problema también se

presenta en la traducción entre otros pares de idiomas, como en la traducción entre francés y alemán.

Como se ha podido observar, debido a la riqueza morfológica del español, es muy importante realizar una correcta resolución de la anáfora pronominal (identificación de su antecedente y su género gramatical) en el idioma origen (inglés) para generarla correctamente en español. En la generación de los pronombres ingleses también aparecen otros problemas ocasionados por el género del antecedente. Para tratar todas estas discrepancias de género en la generación de la anáfora pronominal en el idioma destino, en nuestro sistema se ha construido un conjunto de reglas morfológicas (de género) que se deben de tener en cuenta en la etapa de la generación morfológica. Dependiendo de la función del pronombre se ha hecho la siguiente clasificación: pronombres con función de sujeto y pronombres con función de complemento.

- **Pronombres con función de sujeto.** Estos pronombres se identifican en la representación interlingua porque tienen el papel temático de *AGENTE* en una *CLÁUSULA*.

En la traducción español-inglés de las anáforas pronominales originadas por los pronombres personales, el principal problema se plantea en la generación del pronombre *it*. Si en la representación interlingua aparece un pronombre con la siguiente información: tercera persona, masculino y singular, éste puede ser generado en los pronombres ingleses *he* o *it*[15]. Si el antecedente del pronombre se refiere a una persona entonces generaremos el pronombre *he*. De otro modo, si el antecedente se refiere a un animal o a un objeto, generaremos el pronombre *it*. Esta característica del antecedente se puede obtener de la información semántica almacenada en su atributo *Ref_Sem* en la representación interlingua.

(89) **E** *[El león]*$_i$ bebía leche porque *él*$_i$ estaba muy hambriento.

[15] Además de los pronombres personales mencionados anteriormente, también existen dificultades para la generación de los *cero pronombres*.

l *[The lion]$_i$* was drinking milk because *it$_i$* was very hungry.

En el ejemplo 89.i se ha generado el pronombre *it* debido a que el núcleo del antecedente del pronombre (*león*) tiene información semántica que indica que es de tipo animal.

Una estrategia similar se sigue para generar los pronombres ingleses *she* o *it*. En cuanto a los pronombres españoles en plural (*ellos* y *ellas*) se traducirán indistintamente por el pronombre inglés *they*. En la figura 5.11 se muestra el conjunto de reglas morfológicas que tratan las discrepancias de género para la generación de los pronombres ingleses con función de sujeto.

AGENTE (PRONOMBRE + 3ª persona + masculino + singular + + antecedente (persona)) → **he**

AGENTE (PRONOMBRE + 3ª persona + masculino + singular + + antecedente (animal u objeto)) → **it**

AGENTE (PRONOMBRE + 3ª persona + femenino + singular + + antecedente (persona)) → **she**

AGENTE (PRONOMBRE + 3ª persona + femenino + singular + + antecedente (animal u objeto)) → **it**

AGENTE (PRONOMBRE + 3ª persona + masculino/femenino + plural) → **they**

Figura 5.11. Discrepancias de género. Traducción español-inglés de pronombres con función de sujeto

En la traducción inglés-español, el principal problema se plantea, de nuevo, por la generación del pronombre *it* en español, ya que se puede traducir por cuatro pronombres españoles distintos (*él*, *ella*, *éste* y *ésta*). Estos pronombres españoles se pueden referir tanto a animales como a objetos, pero normalmente *él/ella* se refiere a un animal y *éste/ésta* se refiere a un objeto. Por ello, en nuestro sistema automático, cuando el antecedente del pronombre *it* es de tipo animal se traducirá por *él/ella* y cuando es de tipo

objeto se traducirá por *éste/ésta*, ya que es el uso más común en español.

(90) **I** The monkey ate *[the banana]*$_i$ because *it*$_i$ was ripe.

E El mono se comió *[la banana]*$_i$ porque *ésta*$_i$ estaba madura.

En el ejemplo 90.e el pronombre *it* se ha traducido por *ésta* (femenino y singular) ya que su antecedente (el sintagma nominal *the banana*) es femenino, singular y de tipo objeto.

En la figura 5.12 se muestra el conjunto de reglas morfológicas que tratan las discrepancias de género para la generación de los pronombres españoles con función de sujeto.

AGENTE (PRONOMBRE + 3ª persona + masculino + singular +
+ antecedente (persona)) → **él**

AGENTE (PRONOMBRE + 3ª persona + singular +
+ antecedente (animal con género masculino)) → **él**

AGENTE (PRONOMBRE + 3ª persona + singular +
+ antecedente (objeto con género masculino)) → **éste**

AGENTE (PRONOMBRE + 3ª persona + femenino + singular +
+ antecedente (persona)) → **ella**

AGENTE (PRONOMBRE + 3ª persona + singular +
+ antecedente (animal con género femenino)) → **ella**

AGENTE (PRONOMBRE + 3ª persona + singular +
+ antecedente (objeto con género femenino)) → **ésta**

AGENTE (PRONOMBRE + 3ª persona + plural +
+ antecedente (con género masculino)) → **ellos**

AGENTE (PRONOMBRE + 3ª persona + plural +
+ antecedente (con género femenino)) → **ellas**

Figura 5.12. Discrepancias de género. Traducción inglés-español de pronombres con función de sujeto

- Pronombres con función de complemento. Estos pronombres se identifican en la representación interlingua porque tienen un papel temático distinto al de *AGENTE* en una *CLÁUSULA*.

En el ejemplo 91 se muestra una oración en español con un pronombre personal de complemento y su traducción a inglés.

(91) **E** *[El jarrón]*$_i$ era muy caro. El niño *lo*$_i$ rompió.
I *[The vase]*$_i$ was very expensive. The child broke *it*$_i$.

En el ejemplo 91.i se ha generado el pronombre *it* debido a que el antecedente del pronombre es un objeto (*el jarrón*) y está en singular. Si el antecedente es una persona en singular se generará el pronombre *him/her* dependiendo del género gramatical del mismo. Por último, si el antecedente está en plural se generará el pronombre *them*. En la figura 5.13 se muestran las reglas morfológicas que tratan las discrepancias de género para la generación de los pronombres ingleses con función de complemento.

NO_AGENTE (PRONOMBRE + 3ª persona + singular) +
+ antecedente (persona con género masculino)) → **him**

NO_AGENTE (PRONOMBRE + 3ª persona + singular) +
+ antecedente (persona con género femenino)) → **her**

NO_AGENTE (PRONOMBRE + 3ª persona + singular) +
+ antecedente (animal u objeto)) → **it**

NO_AGENTE (PRONOMBRE + 3ª persona + plural)
→ **them**

Figura 5.13. Discrepancias de género. Traducción español-inglés de pronombres con función de complemento

En la traducción inglés-español hay que hacer la distinción dependiendo de que el pronombre de complemento tenga el papel temático *TEMA* u otro papel distinto al de *AGENTE* o *TEMA*.

(92) **I** I like *[that car]*$_i$. I will buy *it*$_i$ when I save.

E Me gusta *[ese coche]*$_i$. *Lo*$_i$ compraré cuando ahorre.

En el ejemplo 92 el pronombre de complemento *it* con papel temático *TEMA* genera el pronombre español *lo* porque el antecedente del mismo es masculino singular y es de tipo objeto.

(93) **I** *[Pedro]*$_i$ is an absolute bore. We are sick and tired of *him*$_i$.

E *[Pedro]*$_i$ es muy pesado. Estamos hartos y cansados de *él*$_i$.

En el ejemplo 93 el pronombre de complemento *him* tiene el papel temático *META* y genera el pronombre español *él* porque el antecedente del mismo tiene género masculino.

En la figura 5.14 se muestran las reglas morfológicas que tratan las discrepancias de género para la generación de los pronombres españoles con función de complemento.

- Excepciones en las discrepancias de género. Tal y como presentan Alcaraz & Moody (1997), hay una serie de excepciones en la traducción de los pronombres personales entre español e inglés que afectarían a las reglas presentadas para tratar las discrepancias de género. Entre estas excepciones podemos citar las siguientes:

- Cuando se habla de animales domésticos, los pronombres españoles *él/éste*, *ella/ésta* se traducen al inglés por *he/she* en lugar del pronombre *it*.

(94) **E** Tuve que llevar *[mi perra]*$_i$ al veterinario porque *ella*$_i$ tenía una infección.

I I had to take *[my dog]*$_i$ to the vet because *she*$_i$ had an infection.

- En poesía al referirse a *Time*, *Love*, *Death*, etc. de carácter abstracto, y a palabras concretas como *sun*, *mountain*, *river*, etc., siempre que se personifiquen se utiliza el pronombre *he* en lugar del pronombre *it*.

TEMA (PRONOMBRE + 3ª persona + singular) +
+ antecedente (persona)) → **le**

TEMA (PRONOMBRE + 3ª persona + plural) +
+ antecedente (persona)) → **les**

TEMA (PRONOMBRE + 3ª persona + singular) +
+ antecedente (animal u objeto con género masculino)) → **lo**

TEMA (PRONOMBRE + 3ª persona + singular) +
+ antecedente (animal u objeto con género masculino)) → **la**

TEMA (PRONOMBRE + 3ª persona + plural) +
+ antecedente (animal u objeto con género masculino)) → **los**

TEMA (PRONOMBRE + 3ª persona + plural) +
+ antecedente (animal u objeto con género femenino)) → **las**

NO_TEMA_y_NO_AGENTE (PRONOMBRE + 3ª persona + singular) +
+ antecedente (género masculino)) → **él**

NO_TEMA_y_NO_AGENTE (PRONOMBRE + 3ª persona + singular) +
+ antecedente (género femenino)) → **ella**

NO_TEMA_y_NO_AGENTE (PRONOMBRE + 3ª persona + plural) +
+ antecedente (género masculino)) → **ellos**

NO_TEMA_y_NO_AGENTE (PRONOMBRE + 3ª persona + plural) +
+ antecedente (género femenino)) → **ellas**

Figura 5.14. Discrepancias de género. Traducción inglés-español de pronombres con función de complemento

(95) **E** Mira al *[sol]*$_i$. *Él*$_i$ realmente es el padre del mundo.

I Look at *[the sun]*$_i$. *He*$_i$ is really the father of the world.

- Cuando nos referimos a coches, barcos, motores, al hablar de países, ciudades, *the moon*, *the earth*, *the sea*, *the Church*, y al

Volver al índice/Tornar a l'índex

personificar conceptos como *Beauty*, *Peace*, *Liberty*, *Virtue*, etc., se utiliza el pronombre *she* en lugar del pronombre *it*.

(96) **E** Cuando se construyó el *[Queen Mary]*$_i$, $\emptyset_i$ era el buque mayor que existía.

I When the *[Queen Mary]*$_i$ was built, *she*$_i$ was the biggest vessel in existence.

- El pronombre *it* se emplea al referirse a recién nacidos en lugar de utilizar los pronombres *he/she*.

(97) **E** Mira a *[ese bebé]*$_i$; $\emptyset_i$ es realmente muy hermoso.

I Look at *[that baby]*$_i$; *it*$_i$ is really very beautiful.

Estas excepciones, debido al carácter tan específico y al tipo de texto en las que pueden aparecer (por ejemplo en poesía), no se han tenido en cuenta en nuestro sistema. Para incluirlas, habría que diseñar un conjunto de reglas específicas para cada una de ellas dependiendo del tipo de texto que se esté traduciendo.

6. Implementación del sistema AGIR

En este capítulo se presenta la implementación que se ha realizado de los distintos módulos del sistema AGIR cuyo funcionamiento básico se ha explicado previamente en el capítulo 5. Para ello se explicarán en detalle las distintas etapas que se llevan a cabo en los módulos de análisis y generación del sistema.

6.1 Implementación del módulo de análisis del sistema AGIR

El módulo de análisis del sistema AGIR analiza el texto del idioma origen y obtiene la representación interlingua del mismo. Este módulo se basa en el sistema computacional SUPAR –*Slot Unification Parser for Anaphora Resolution* (Ferrández *et al.*, 1999a)– que está orientado principalmente a la resolución de problemas lingüísticos, en particular a la resolución de la anáfora. Tanto SUPAR como la mayoría de los módulos del sistema AGIR han sido implementados en PROLOG, en concreto en LPA Win-Prolog[1].

A partir de la estructura de datos generada por SUPAR (tras el análisis del texto y la resolución de problemas lingüísticos) se obtiene la representación interlingua del texto. Esta representación se utilizará como entrada del módulo de generación del sistema AGIR.

Los detalles sobre la implementación de las distintas etapas del módulo de análisis: análisis léxico y morfológico, análisis sintáctico, desambiguación del sentido de las palabras y resolución de problemas lingüísticos, así como la etapa de la obtención

[1] Ciertos procesos de la etapa de análisis léxico-morfológico han sido implementados en el lenguaje de programación AWK y C++ para Linux.

de la representación interlingua, se explicarán en las siguientes secciones. Los ejemplos que aparecerán para ilustrar algunos procesos se mostrarán indistintamente en español o inglés ya que el sistema tiene el mismo esquema de funcionamiento para ambos idiomas.

6.1.1 Análisis léxico y morfológico

El módulo de análisis léxico y morfológico recibe como entrada el texto en el idioma origen que se desea traducir al idioma destino y una herramienta que proporciona toda la información léxica y morfológica de cada una de las unidades léxicas del mismo. Esta herramienta puede ser un lexicón monolingüe o un etiquetador léxico-morfológico (part-of-speech-tagger, POS tagger). Como salida, este módulo devuelve una lista con las unidades léxicas del texto en las que se almacena toda la información necesaria para los restantes módulos.

Etiquetado léxico-morfológico. Aunque el sistema AGIR puede trabajar con un lexicón monolingüe como fuente léxico-morfológica[2], debido a la dificultad de crear un lexicón de propósito general para trabajar con cualquier tipo de textos (dominio no restringido), AGIR trabaja sobre la salida de un etiquetador léxico-morfológico. Éste asigna a cada unidad léxica del texto una etiqueta que contiene la categoría gramatical e información morfológica de la misma. Tanto para el inglés como para el español se han utilizado distintos etiquetadores que usaban sus propios conjuntos de etiquetas.

Para inglés se han utilizado corpus etiquetados por el Xerox POS Tagger (Cutting *et al.*, 1992), por el etiquetador estocástico de Brill (Brill, 1992) y por el etiquetador TreeTagger (Schmid, 1994). Estos etiquetadores utilizan un conjunto de etiquetas basadas en las etiquetas que fueron definidas para anotar el Brown Corpus (Francis, 1964; Francis & Kucera, 1982).

[2] Se han realizado experimentos sobre corpus de dominio restringido en los que se utilizaba un pequeño lexicón, que contenía todas las unidades léxicas del texto, como fuente de información léxico-morfológica.

Para español se han utilizado corpus que utilizan el conjunto de etiquetas del Xerox POS Tagger adaptado al español en el Proyecto CRATER (Proyecto CRATER, 1994-1995) y corpus que utilizan el conjunto de etiquetas PAROLE (Martí *et al.*, 1998) definido en el proyecto ITEM (Proyecto ITEM, 1996-1999). En la actualidad se están incorporando al sistema corpus procesados con dos nuevos etiquetadores para español: el etiquetador léxico-morfológico desarrollado por el Grup de Processament del Llenguatge Natural de la Universitat Politècnica de València (Pla, 2000; Pla *et al.*, 2000) que utiliza las etiquetas PAROLE y el etiquetador léxico-sintáctico Conexor desarrollado en Finlandia – SpaCG-2, (Samuelsson & Voutilainen, 1997)– que usa un conjunto de etiquetas propio.

El etiquetador léxico-morfológico recibe como entrada un texto plano y obtiene como salida la forma base o lema y la etiqueta (incluye la categoría gramatical e información morfológica) de cada una de las palabras del texto. El formato es el siguiente: $PALABRA_1$ $LEMA_1$ $ETIQUETA_1$ $PALABRA_2$ $LEMA_2$ $ETIQUETA_2$...

En la figura 6.1 se muestra un ejemplo de un fragmento de texto en inglés y la salida producida por el etiquetador Xerox POS Tagger.

Texto de entrada:

```
All Plmn configurations can be deducted from this basic configu
ration.
```

Salida tras el etiquetador léxico-morfológico:

```
All all DB
Plmn plmn NN1
configurations configuration NN2
can can VM
be be VBI
deducted deduct VVN
from from II
this this DD1
basic basic JJ
configuration configuration NN1
. . .
```

Figura 6.1. Salida del etiquetador Xerox POS Tagger

Segmentación del texto. Puesto que SUPAR trabaja con el texto dividido en oraciones[3], el siguiente proceso al etiquetado léxico-morfológico será la segmentación del texto en oraciones.

En la figura 6.2 se muestra un fragmento de texto en español y la salida producida tras el segmentador de oraciones.

El segmentador de oraciones reconoce los caracteres de fin de oración(., ?, !, etc.) y segmenta el texto de entrada (etiquetado previamente con el formato $PALABRA_1$ $LEMA_1$ $ETIQUETA_1$ $PALABRA_2$ $LEMA_2$ $ETIQUETA_2$...) en oraciones, introduciendo cada una de ellas en un predicado Prolog denominado *o()*. Este predicado tiene tres argumentos con la siguiente información:

- *Número.* Es un número único que identifica cada oración del texto.
- *Descripción.* Breve descripción del tipo de texto (corpus) a tratar.
- *Oración.* Lista de predicados *w()* de las palabras que forman la oración[4]. Cada uno de estos predicados contiene a su vez tres argumentos correspondientes a la palabra, la forma base (lema) y la etiqueta. Si se dispone de información semántica de las palabras, ésta se añadiría como un argumento adicional a cada una de las palabras.

Interfaz entre el etiquetador y la gramática. El último proceso que se lleva a cabo en esta etapa consiste en la transformación de las etiquetas proporcionadas por el etiquetador para convertirlas en las etiquetas apropiadas de la gramática (símbolos terminales) que se utilizará en la etapa posterior de análisis sintáctico. Esta transformación se realiza mediante un interfaz apropiado. Actualmente el sistema AGIR dispone de interfaces capaces de convertir tres juegos de etiquetas: el juego de etiquetas definido para anotar el Brown Corpus inglés (Francis, 1964; Francis &

[3] Hay que destacar que aunque SUPAR trabaja oración por oración almacena información del discurso que será usada en módulos posteriores. Esta información es la lista de antecedentes de oraciones anteriores usada para la resolución de la anáfora.

[4] Los símbolos *[...]* se utilizarán para representar una lista Prolog. Por ejemplo, *[s, t, u]* representa la lista formada por tres elementos: *s*, *t* y *u*, donde *s* es su cabeza y *u* es su cola.

Texto de entrada:

```
Las descripciones indicadas a continuación se limitan a la apli
cación móvil. La parte de aplicación móvil  será sustentada por
las capacidades de transacción.
```

Salida tras el etiquetador léxico-morfológico:

```
Las el ARTDFP
descripciones descripción NCFP
indicadas indicado ADJGFP
a a PREP
continuación continuación NCFS
se se SE
limitan limitar VLPI3P
a a PREP
la el ARTDFS
aplicación aplicación NCFS
móvil móvil ADJGFS
. . FS
La el ARTDFS
parte parte NCFS
de de PREP
aplicación aplicación NCFS
móvil móvil ADJGFS
será ser VSFI3S
sustentada sustentar VLPXFS
por por PREP
las el ARTDFP
capacidades capacidad NCFP
de de PREP
transacción transacción NCFS
. . FS
```

Salida tras el segmentador de oraciones:

```
o(1,'frase_BBE',[w('Las','el','ARTDFP'),w('descripciones','desc
ripción','NCFP'),w('indicadas','indicado','ADJGFP'),w('a','a','
PREP'),w('continuación','continuación','NCFS'),w('se','se','SE'
),w('limitan','limitar','VLPI3P'),w('a','a','PREP'),w('la','el'
,'ARTDFS'),w('aplicación','aplicación','NCFS'),w('móvil','móvil
','ADJGFS'),w('.','.','FS')]).

o(2,'frase_BBE',[w('La','el','ARTDFS'),w('parte','parte','NCFS'
),w('de','de','PREP'),w('aplicación','aplicación','NCFS'),w('mó
vil','móvil','ADJGFS'),w('será','ser','VSFI3S'),w('sustentada',
'sustentar','VLPXFS'),w('por','por','PREP'),w('las','el','ARTDF
P'),w('capacidades','capacidad','NCFP'),w('de','de','PREP'),w('
transacción','transacción','NCFS'),w('.','.','FS')]).
```

Figura 6.2. Salida del etiquetador Xerox POS Tagger adaptado al español

Kucera, 1982), el juego de etiquetas del Xerox POS Tagger adaptado al español en el Proyecto CRATER (Proyecto CRATER, 1994-1995) y el juego de etiquetas PAROLE (Martí *et al.*, 1998) definido en el proyecto ITEM (Proyecto ITEM, 1996-1999).

En la figura 6.3 se puede observar el proceso de transformación de etiquetas para la oración 2 de la figura 6.2. En ella aparece la oración tras el segmentador de oraciones, el interfaz etiquetador Xerox–gramática SUG[5] y la salida (*lista SUG*). En la *lista SUG* cada palabra es una estructura con dos argumentos: la forma base (lema) y una lista con las distintas categorías gramaticales que pueda tener la palabra y que se corresponden con los símbolos terminales SUG. En el sistema AGIR, esta lista de símbolos terminales SUG sólo contiene un elemento ya que el etiquetador proporciona una única etiqueta (categoría gramatical) a cada palabra.

6.1.2 Análisis sintáctico

El módulo de análisis sintáctico toma como entrada la lista de palabras que contiene las etiquetas apropiadas de la gramática (*lista SUG*) y la información sintáctica representada por medio del formalismo gramatical SUG –*Slot Unification Grammar* (Ferrández *et al.*, 1997a)–. La salida de este módulo está formada por las *estructuras de huecos* (*slot structures*) que almacenan toda la información necesaria para el resto de módulos. Además, devuelve información del discurso mediante una lista de antecedentes de oraciones anteriores que será usada en el módulo de resolución de la anáfora.

Slot Unification Grammar (SUG). Las SUG fueron desarrolladas como una extensión de las DCG –*Definite Clause Grammars* (Pereira & Warren, 1980)– con el objetivo de ampliar las capacidades de éstas para facilitar la resolución de diversos problemas lingüísticos de manera modular. Las SUG se denominan así debido a las *estructuras de huecos* (*slot structures*) generadas

[5] El sistema SUPAR utiliza el formalismo gramatical SUG para realizar el análisis sintáctico del texto. Este formalismo será definido con detalle en la siguiente sección.

Entrada:

```
. . .
o(2,'frase_BBE',[w('La','el','ARTDFS'),w('parte','parte','NCFS'
),w('de','de','PREP'),w('aplicación','aplicación','NCFS'),w('mó
vil','móvil','ADJGFS'),w('será','ser','VSFI3S'),w('sustentada',
'sustentar','VLPXFS'),w('por','por','PREP'),w('las','el','ARTDF
P'),w('capacidades','capacidad','NCFP'),w('de','de','PREP'),w('
transacción','transacción','NCFS'),w('.','.','FS')]).
```

Interfaz etiquetador Xerox-gramática SUG:

```
. . .
interfaz('ARTDFS',art(sing,fem,det)).
interfaz('NCFS',sust(sing,fem,comun)).
interfaz('PREP',prepSimple)).
interfaz('ADJGFS',adjSimple(sing,fem,cal)).
interfaz('VSFI3S',verbo(sing,terc,presente,copul)).
interfaz('VLPXFS',verbo(sing,_,participio,_)).
interfaz('ARTDFP',art(pl,fem,det)).
interfaz('NCFP',sust(pl,fem,comun)).
. . .
```

Salida: lista SUG utilizada en la etapa de análisis sintáctico

```
. . .
o(2,'frase_BBE',[w(el,[art(sing,fem,det)]),w(parte,[sust(sing,f
em,comun)]),w(de,[prepSimple]),w(aplicación,[sust(sing,fem,comu
n)]),w(móvil,[adjSimple(sing,fem,cal)]),w(ser,[verbo(sing,terc,
presente,copul)]),w(sustentar,[verbo(sing,_,participio,_)]),w(p
or,[prepSimple]),w(el,[art(sing,fem,det)]),w(capacidad,[sust(pl
,fem,comun)]),w(de,[prepSimple]),w(transacción,[sust(sing,fem,c
omun)]),w(.,[conj]).
```

Figura 6.3. Salida del interfaz entre el etiquetador Xerox y la gramática SUG

automáticamente por el analizador donde se incluye de forma automática toda la información morfológica, sintáctica y semántica necesaria para resolver problemas lingüísticos variados.

Las SUG se definen como una quíntupla *(NT,T,H,P,S)*, donde NT es un conjunto finito de símbolos no terminales. T es un conjunto finito de símbolos terminales disjunto con NT. H son hechos SUG, pudiendo ser *coordination, juxtaposition, fusion, basicWord* o *isWord*. P son las reglas de producción de la gramática, un conjunto finito de reglas de la forma: $\alpha ++ > \beta$ ó β_1 ,donde $\alpha \in NT$, $\beta \in (T \cup NT \cup \{\text{llamadas a procedimientos}\})^*$ y $\beta_1 \in H$. S es el conjunto de símbolos iniciales de la gramática.

Ya que las SUG son una extensión de las DCG, heredan muchas de sus características. La principal diferencia entre ambas consiste en que en las reglas de producción SUG de la forma $\alpha ++ > \beta$, cada subconstituyente de β puede omitirse en la oración si se escribe entre el operador opcional: << *constituyente* >>. Este operador opcional tiene la posibilidad de recordar si el constituyente opcional ha sido analizado en la oración o no. Esta información es muy útil para la resolución de problemas lingüísticos tales como la elipsis o extraposición. Para comprobar la presencia o ausencia de un constituyente opcional se le añade una etiqueta, por ejemplo << *SP : sp* >> para el caso de un sintagma preposicional. Esta etiqueta será una variable Prolog sin instanciar si no existe el constituyente *sp* (el predicado Prolog *var(SP)* tendrá éxito).

Estructuras de huecos (slot structures). El sistema SUPAR lleva incorporado un traductor que transforma las reglas de la gramática SUG en cláusulas Prolog. Este traductor genera un programa Prolog que realizará el análisis sintáctico del texto. La salida del módulo de análisis es una serie de estructuras denominadas *estructuras de huecos* (*slot structures*). Estas estructuras almacenan la información léxica, morfológica, sintáctica y semántica de cada uno de los constituyentes de la gramática.

En la figura 6.4 se muestra un ejemplo simplificado de una estructura de huecos de un *sn* extraído de la oración 2 de la figura 6.2.

Como se observa en la figura 6.4 la estructura de huecos es un estructura Prolog cuyo nombre (funtor) se corresponde con el constituyente que representa (*sn*, *sp*, etc.) y que tiene los siguientes argumentos:

- *Conc.* Incluye la información de concordancia del constituyente (información morfológica de número, género y persona).
- *X.* Variable usada para identificar el constituyente (identificador de discurso o de entidad). Hay que destacar que esta variable se utilizará para establecer el enlace correspondiente entre la anáfora y su antecedente.
- El resto de los argumentos contiene las estructuras de huecos de sus subconstituyentes.

Entrada: texto a analizar

```
. . . La parte de aplicación móvil será sustentada . . .
```

Fragmento de la gramática SUG:

```
sn ++> <<determinante>>, <<adj>>, sust, <<adj>>, <<sp>>.
sp ++> preposicion, sn.
. . .
```

Estructura de huecos del *sn* tras el análisis sintáctico:

```
sn( conc(singular, femenino), S,
    determinante( conc(singular, femenino), T, la ),
    _,
    sust( conc(singular, femenino), U, parte ),
    _,
    sp( conc(prep_de), V,
        preposicion ( conc(prep_de), W, de),
        sn( conc(singular, femenino), X,
            _,
            _,
            sust( conc(singular, femenino), Y, aplicación ),
            adj( conc(singular, femenino), Z, móvil ),
            _
        )
    )
)
```

Figura 6.4. Ejemplo de una estructura de huecos

Las variables sin instanciar que aparecen en la estructura de huecos de un constituyente (representadas por el símbolo "_") representan los huecos correspondientes a los constituyentes opcionales que no han aparecido en el mismo.

Analizador sintáctico parcial. Una de las principales ventajas del sistema SUPAR es que permite realizar un análisis sintáctico completo o un análisis sintáctico parcial del texto.

En el sistema AGIR se ha optado por realizar un análisis sintáctico parcial del texto motivado principalmente por el uso de textos (corpus) de dominio no restringido. Sobre estos textos se considera que es más adecuado realizar un análisis parcial en lugar de un análisis completo debido a los posibles errores e inevitable incompletitud de la gramática (Palomar *et al.*, 1999; Martínez-Barco *et al.*, 1998). La estrategia que se ha utilizado es la presentada en el trabajo de Ferrández *et al.* (1998a) para el sistema SUPAR. En este trabajo se muestra que la resolución de

la anáfora se puede llevar a cabo realizando un análisis sintáctico parcial que identifique únicamente los siguientes constituyentes: sintagmas nominales y preposicionales que pueden estar coordinados o no, pronombres, conjunciones y verbos. Ésta es la única información necesaria para resolver correctamente las expresiones anafóricas. Las palabras no analizadas se corresponden con constituyentes que no están cubiertos por la gramática (por ejemplo los adverbios) o palabras incorrectamente etiquetadas o erróneas. En la figura 6.5 se muestra de un modo simplificado un fragmento de la gramática SUG para el inglés.

La gramática SUG de la figura 6.5, además de usar el operador opcional `<< a >>`, utiliza el operador `<# a, b #>` que permite analizar el constituyente *a*, y si no se encuentra en el texto, entonces intenta analizar *b*, en caso de no encontrarse fracasará. El operador `<## a, b ##>` es similar al anterior salvo que tendrá éxito aunque no se encuentren ni *a* ni *b*.

El símbolo inicial de la gramática es *oracTag* que comenzará el proceso del análisis parcial aplicando las reglas de la gramática anterior. Con este símbolo inicial se buscarán los sintagmas nominales (*sn*) y preposicionales (*sp*) que estén coordinados o no, las conjunciones (*conj*) y los verbos (*nucleoverbal*) en cualquier orden que aparezcan en el texto. Como salida, el analizador sintáctico parcial generará una estructura de datos que mostrará los constituyentes analizados, las palabras no analizadas y el orden en que los encontró.

Hay que destacar la gran flexibilidad de este analizador parcial ya que se puede fácilmente añadir o eliminar constituyentes a analizar, simplemente modificando la regla gramatical *oracTag*. Del mismo modo, hay que resaltar la modularidad que presenta el sistema global, ya que una misma gramática SUG puede ser utilizada por distintos etiquetadores o diccionarios, simplemente hay que definir el interfaz apropiado entre éstos y la gramática.

Por último, mencionar que tras la etapa del análisis sintáctico parcial también se devuelve una lista que contiene todos los posibles antecedentes de las oraciones anteriores con el objetivo de resolver la anáfora intersentencial. Esta lista contiene las estruc-

```
oracTag ++>
  << V:nucleoVerbal >>, << C:conj >>, << SP:sp >>, << SN:sn >>,
  <# ['.'],
      oracTagSuelto(V,C,SP,SN)
  #> .

oracTagSuelto(V,C,SP,SN) ++>
  <## ( {( var(V), var(C), var(SP), var(SN))}, [W]),
        ( _                                , _  )
  ##>,
  oracTag.
%----------- Reglas gramaticales para cada constituyente
coordination(sp,simpleSP).
simpleSP ++> preposicion, sn.
coordination(sn,simpleSN(_) ).
simpleSN(tipoSUST) ++>
  << det >>, << preModificadorSN >>, nucleoSN,
  << genitivoSajon >>, << postModificadorSN >>, << sp >>.
simpleSN(tipoADJ) ++>
  << det >>, << preModificadorSN >>, sadjetival, << sp >>.
simpleSN(tipoPRON) ++> pronEnglish.
det ++> art.
det ++> adjSimple.
preModificadorSN ++> verbo.
preModificadorSN ++> sadjetival.
nucleoSN ++> sustantivo.
nucleoSN ++> verbo.
genitivoSajon ++>
  genitive, << preModificadorSN >>, sustantivo.
postModificadorSN ++> oracRelat.
postModificadorSN ++> adjSimple.
postModificadorSN ++> aposicion.
sadjetival ++> << adv >>, adjetivo, << adjSimple >>.
nucleoVerbal ++>
  nucleoVerbalAux, << particle >>, << perifrasis >>.
perifrasis ++> verbo(_,_,gerundio,_).
perifrasis ++> [to], verbo(_,_,infinitivo,_).
nucleoVerbalAux ++>
  <# verbCompHave,
     verbCompBe,
     verbCompDo,
     verbCompWill,
     verboSolo
  #> .
```

Figura 6.5. Fragmento de una gramática SUG para inglés

turas de huecos de los sintagmas nominales que se han analizado en oraciones previas.

En la figura 6.6 se muestra un ejemplo de la aplicación del análisis parcial sobre un fragmento de texto en inglés.

6.1.3 Desambiguación del sentido de las palabras

En esta etapa se resuelve la ambigüedad léxica de las palabras del texto origen, es decir, se realiza la desambiguación del sentido de las palabras (*Word Sense Disambiguation*, *WSD*). Este módulo recibe como entrada la estructura sintáctica enriquecida con la información que se ha obtenido en etapas anteriores y proporciona un único significado a los nombres y verbos que aparecen en ella.

En AGIR, en concreto, se han utilizado textos en los que las palabras están etiquetadas con su sentido correcto tras un proceso de desambiguación. En este proceso se ha utilizado el recurso léxico EuroWordNet (Vossen, 1998) como herramienta que proporciona los distintos sentidos que puede tener una palabra, así como, un conjunto de relaciones semánticas que existen entre las palabras del texto. Toda la información semántica obtenida de EuroWordNet se utilizará por el sistema en etapas posteriores.

EuroWordNet. El EuroWordNet es un recurso léxico que está formado por varios WordNets de diferentes lenguas europeas (inglés, holandés, español, italiano, alemán, checoslovaco, estonio y francés). Cada uno de estos WordNets es una base de datos léxica en la que los nombres, verbos, adjetivos y adverbios están organizados en conjuntos de sinónimos (*synsets*) que representan un mismo concepto léxico. Entre estos *synsets* se establecen distintas relaciones semánticas como hiponimia, hiperonimia, meronimia, etc. Los distintos WordNets se relacionan entre sí a través de un módulo interlingua denominado ILI (*Inter-Lingual Index*) que establece las relaciones de equivalencia correspondientes entre *synsets* de distintos idiomas.

EuroWordNet proporciona una ontología con 63 categorías (conceptos ontológicos) diferentes para describir todas las palabras almacenadas en la base de conocimientos. En el nivel principal se distinguen los siguientes conceptos ontológicos denomi-

Texto de entrada:

```
Rockwell International Corp. s Tulsa unit said it signed a tent
ative agreement extending its contract . . .
```

Salida tras el segmentador de oraciones:

```
o(1,'frase_Brown',[w('Rockwell','rockwell','NNP'),w('Internatio
nal','international','NNP'),w('Corp.','corp.','NNP'),w('s','s',
'POS'),w('Tulsa','tulsa','NNP'),w('unit','unit','NN'),w('said',
'say','VBD'),w('it','it','PPH1'),w('signed','sign','VBD'),w('a'
,'a','DT'),w('tentative','tentative','JJ'),w('agreement','agree
ment','NN'),w('extending','extend','VBG'),w('its','its','PRP$')
,w('contract','contract','NN') . . .
```

Salida tras el analizador parcial:

```
** ORACION ANALIZADA PARCIALMENTE:
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** NUCLEOSN:
               ** SUSTANTIVO: rockwell
                  ** SUSTANTIVO: international
                     ** SUSTANTIVO: corp.
         ** GENITIVO SAJÓN:
            ** GENITIVO: s
               ** SUSTANTIVO: tulsa
                  ** SUSTANTIVO: unit
   ** NUCLEO VERBAL:
      ** VERBO: said
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** PRONOMBRE: it
   ** NUCLEO VERBAL:
      ** VERBO: signed
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** DETERMINANTE 1:
            ** ARTICULO: a
         ** PREMODIFICADOR:
            ** SINT.ADJETIVAL:
               ** ADJETIVO COORDINADO:
                     ** ADJETIVO SIMPLE: tentative
         ** NUCLEOSN:
               ** SUSTANTIVO: agreement
   ** NUCLEO VERBAL:
      ** VERBO: extending
   ** PALABRA: its
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** NUCLEOSN:
               ** SUSTANTIVO: contract
. . .
```

Figura 6.6. Salida del analizador parcial de un texto en inglés

nados *conceptos principales* (*Top Concepts*): *entidades de primer orden* (1stOrderEntities), *entidades de segundo orden* (2ndOrderEntities) y *entidades de tercer orden* (3rdOrderEntities). Éstos a su vez agrupan el resto de conceptos denominados *conceptos base* (*Base Concepts*). En la figura 6.7 se muestra de un modo esquemático la ontología de EuroWordNet.

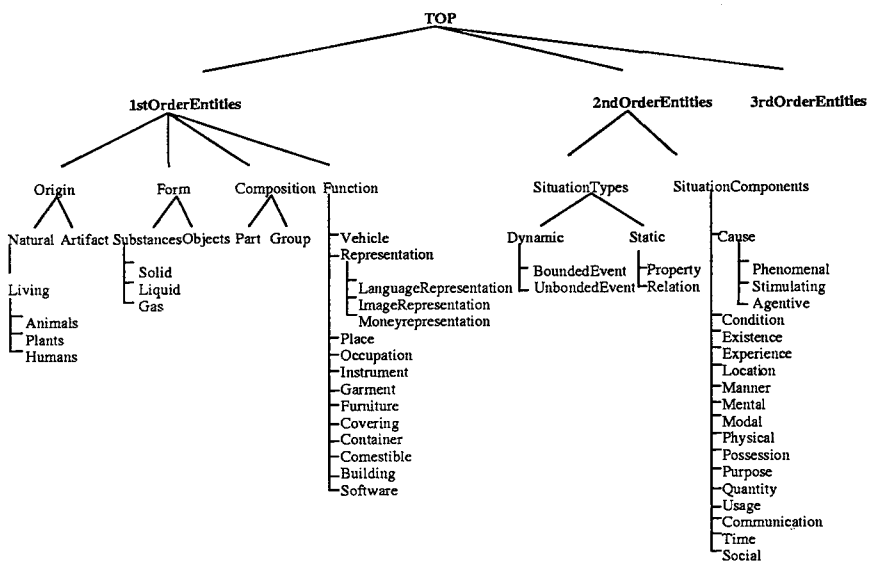


Figura 6.7. Ontología de EuroWordNet

Si las palabras de un texto en un idioma están etiquetadas con su sentido en el WordNet correspondiente, se puede identificar para cada una de ellas: (1) las relaciones semánticas existentes con otras palabras (sinonimia, antonimia, hiponimia, etc.), (2) su concepto ontológico y (3) el *synset* equivalente en otro idioma a través del ILI. Toda esta información semántica que se obtiene mediante EuroWordNet se aplicará en las etapas posteriores del sistema (resolución de problemas lingüísticos, obtención de la representación interlingua, etc.).

Por el momento, toda esta información se obtiene partiendo de textos etiquetados con su sentido correcto en WordNet y reali-

zando los accesos correspondientes a EuroWordNet. En el futuro se prevé la integración en AGIR de una herramienta para resolver la ambigüedad léxica de las palabras que proporcione un único sentido o significado a las palabras del texto automáticamente. En concreto, existen varios miembros de nuestro Grupo de Investigación que están trabajando en este campo utilizando WordNet, destacando los trabajos de Montoyo & Palomar (2000a; 2000b) y el trabajo de Saiz-Noeda *et al.* (2001) en el que se presenta una propuesta para integrar un método de desambiguación léxica en sistemas de resolución de la anáfora.

6.1.4 Resolución de problemas lingüísticos

En esta etapa se tratan los distintos problemas lingüísticos (anáforas, elipsis, problemas de ambigüedad estructural, etc.) del texto. En esta Tesis resolveremos exclusivamente las anáforas pronominales originadas por los pronombres personales de tercera persona y los cero pronombres para tratar posteriormente su generación en el idioma destino.

Módulo de resolución anafórica en AGIR. El sistema AGIR utiliza dos módulos de resolución anafórica distintos para inglés y español. En esta Tesis hemos desarrollado el módulo de resolución anafórica para el inglés, adaptando convenientemente el módulo desarrollado previamente para español y que resolvía distintos tipos de anáfora –pronominal, adjetiva, etc.– (Ferrández *et al.*, 1998b; Ferrández *et al.*, 1997b).

- El algoritmo de resolución de la anáfora. Como el sistema SUPAR trabaja oración por oración, el algoritmo de resolución de la anáfora pronominal recibe como entrada la estructura de huecos de la oración obtenida tras el análisis sintáctico y una lista con todos los posibles antecedentes de oraciones anteriores (contiene las estructuras de huecos de todos los sintagmas nominales de oraciones previas). La salida del algoritmo será una nueva estructura de huecos para cada oración en la que cada anáfora llevará incorporada la información relativa a su antecedente elegido de entre todos los posibles candidatos. Este proceso se repite para todas las oraciones del texto.

En la figura 6.8 aparece un texto en inglés que contiene una anáfora pronominal. En ella se muestra de un modo simplificado el modo de almacenar en la estructura de huecos de la anáfora la información relativa al antecedente.

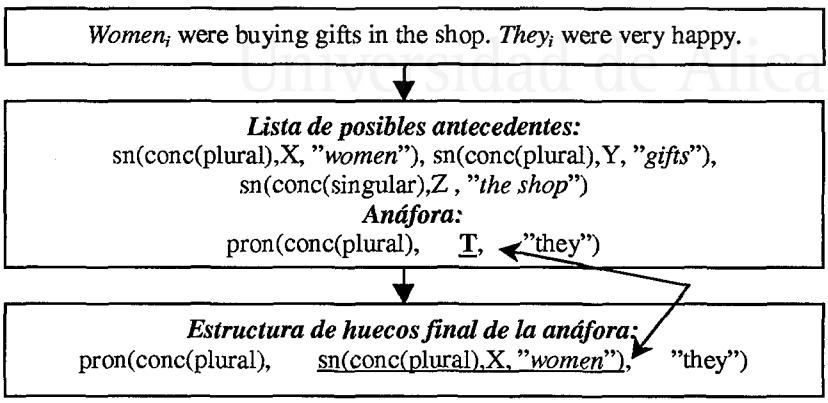


Figura 6.8. Almacenamiento del antecedente en la estructura de huecos de la anáfora

En esta figura se observa que la información del antecedente se almacena en la estructura del pronombre en la posición del identificador de discurso del mismo, ya que esta información no será útil en etapas posteriores. La cadena de correferencia se establecerá entre aquellos constituyentes que tengan el mismo antecedente (representado por su identificador de discurso). Esta información sobre las cadenas de correferencia se almacena en una estructura, denominada *paral*, que contiene información adicional del análisis sintáctico para cada candidato de la lista de antecedentes[6].

[6] La estructura *paral* almacena información estructural del constituyente. Entre sus distintos argumentos destacamos los siguientes: *oración* (número único de oración donde aparece el constituyente), *cláusula* (número de cláusula dentro de la oración donde aparece le constituyente), *posición respecto al núcleo verbal* (toma dos valores, *av* si el constituyente se encuentra antes del núcleo verbal en la cláusula, o *dv* si se encuentra detrás), *frecuencia* (número de veces que el núcleo del constituyente aparece en el texto), *cadena de correferencia* (contiene los identificadores de discurso de los constituyentes que forman la cadena de correferencia donde aparece el constituyente) e *información semántica* (almacena información semántica del constituyente si está disponible, por ejemplo el sentido del núcleo del constituyente en WordNet).

El algoritmo para la resolución de la anáfora pronominal se divide en dos fases: en primer lugar, detecta las distintas anáforas de la oración (recorriendo la estructura de huecos de la oración de izquierda a derecha). En segundo lugar, tras la detección de la anáfora, activa el mecanismo de resolución de anáforas asociado al tipo identificado. A continuación explicaremos ambas fases:

- La detección de la anáfora pronominal se realiza mediante la información almacenada en su estructura de huecos, es decir, su funtor y aridad. La anáfora pronominal originada por un pronombre personal tendrá una estructura de huecos cuyo funtor es *sn* e incluye un constituyente con nombre *pron* y cuyo tipo es *personal.*
- Tras la detección de la anáfora se aplica el correspondiente método de resolución basado en un sistema de restricciones y preferencias –Ferrández (1998), Ferrández *et al.* (1998a; 1999a)–. En este sistema cada tipo de anáfora tiene su propio conjunto de restricciones y preferencias aunque todas ellas siguen el mismo algoritmo: primero se aplican las restricciones y después las preferencias.

  El objetivo de las restricciones es descartar candidatos (por ejemplo aquellos antecedentes que no concuerden en género y número con la anáfora se descartarán). En cuanto al tratamiento de las preferencias existen dos tendencias (Ferrández, 1998): (1) *tratamiento ordenado*, se aplican secuencialmente distintos niveles de preferencia que van descartando candidatos, es decir, no se ordenan los candidatos sino que se eliminan aquéllos que no cumplan la preferencia; (2) *tratamiento ponderado*, se asigna un peso a cada una de las preferencias de forma que cada candidato tendrá asignado un valor correspondiente al sumatorio de los pesos de las preferencias que cumple.

  En el algoritmo que se ha utilizado se aplican las preferencias con un *tratamiento ordenado* de modo que son consideradas como restricciones, excepto cuando ningún candidato satisface una preferencia en cuyo caso no se descarta ninguno de ellos. Este tipo de tratamiento es computacionalmente más eficiente que el *tratamiento ponderado.*

El número de oraciones previas consideradas para resolver la anáfora viene determinado por el tipo de anáfora. Para el caso de las anáforas pronominales se consideran los antecedentes de la misma oración en la que aparece la anáfora y los que aparecen en las cuatro oraciones anteriores[7]. Esta lista de candidatos posibles se ordena por proximidad a la anáfora. Sobre esta lista se aplica el conjunto de restricciones y preferencias.

El funcionamiento del algoritmo es muy sencillo. En primer lugar se aplican las restricciones a todos los candidatos. Si queda más de un candidato se aplican las preferencias. La secuencia de preferencias se para cuando tras aplicar una de ellas sólo queda un candidato. Si después de aplicar todas las preferencias queda más de un candidato, se extraen los candidatos más repetidos en el texto. Después de esto, si aún queda más de uno, se seleccionan los candidatos que hayan aparecido más veces con el verbo de la anáfora. Por último, si tras todo el proceso aún queda más de un candidato, se escoge como solución el más cercano a la anáfora.

El conjunto de restricciones y preferencias que se han utilizado para resolver la anáfora pronominal en español e inglés está basado en la combinación de distintas fuentes de información. En concreto, las restricciones utilizadas se basan en información morfológica –concordancia de género, número y persona entre la anáfora y su antecedente–, información sintáctica –restricciones *c-dominio* (Reinhart, 1983) adaptadas para análisis sintáctico parcial; básicamente se establecen las condiciones de no correferencia entre un pronombre y un sintagma nominal (Palomar *et al.*, 2001)– y, por último, información semántica si está disponible –compatibilidad semántica entre la anáfora y el antecedente–.

Por otra parte, las preferencias se basan en información lingüística y definen una serie de condiciones lingüísticas que hacen más probable a un candidato que las cumple respecto a otro que no las cumple. Ejemplos de preferencias serían los siguientes: preferencia por los candidatos que estén en la misma oración

[7] Esta determinación se ha tomado tras realizar un profundo estudio del comportamiento de las anáforas pronominales en distintos tipos de textos.

que la anáfora, preferencia por los candidatos que tengan la misma posición sintáctica que la anáfora, preferencia por los candidatos que sean nombres propios, etc.
Tras establecer el conjunto de preferencias para el sistema AGIR, se realizaron los experimentos de entrenamiento oportunos con el objetivo de obtener la ordenación de preferencias de modo que los resultados sean óptimos. Los resultados se muestran en el capítulo 7, en el que se presenta la evaluación del sistema.

6.1.5 Obtención de la representación interlingua

La etapa en la que se obtiene la representación interlingua es la última etapa del módulo de análisis del sistema AGIR. Esta etapa recibe como entrada la estructura de huecos de la oración en la que cada anáfora pronominal lleva asociada la información relativa de su antecedente. La salida es una representación interlingua basada en la cláusula como unidad principal y en la que aparecen las distintas entidades de la oración y las relaciones existentes entre ellas. Este proceso se repite para todas las oraciones tras lo cual se obtiene la representación interlingua del texto completo. Detalles sobre la obtención de la representación interlingua en AGIR se han presentado en los trabajos de Peral *et al.* (Peral & Ferrández, 2000a; Peral *et al.*, 1999b).

A continuación presentaremos la implementación realizada de este módulo en el sistema AGIR según las distintas fases presentadas en la sección 5.2.1.

Identificación de las cláusulas del texto. Para identificar las distintas cláusulas de una oración cuando se ha realizado un análisis sintáctico parcial se aplica la siguiente heurística (H_1):

H_1 Asumimos que se encuentra el principio de una nueva cláusula cuando se ha analizado un núcleo verbal que no está en infinitivo o gerundio y posteriormente aparece una *conjunción libre* tras la que se encuentra otro núcleo verbal.

En la oración en inglés del ejemplo 98 se ha analizado la siguiente secuencia de constituyentes: *sn(Juan and Pedro),verbo(were), sp(for work),conj(because),pron(they),verbo(slept in)*

(98) Juan and Pedro were late for work *because* they slept in.

Aplicando la heurística H_1 tras analizar un verbo (*were*) aparece una conjunción libre (*because*) que marca el comienzo de una nueva cláusula cuyo verbo es *slept in.*

Identificación de los papeles temáticos. La siguiente fase consiste en la identificación de los distintos papeles temáticos de los constituyentes que han aparecido en la cláusula.

El verbo de una cláusula es el elemento principal de la misma ya que contiene, intrínsecamente, información sobre los papeles temáticos de los constituyentes que le rodean. Para identificar estos papeles temáticos normalmente se utiliza como herramienta adicional un lexicón monolingüe que proporciona información léxico-morfológica, sintáctica y semántica para las unidades léxicas y en particular para los verbos. El sistema AGIR trabaja con el etiquetador (*POS tagger*) como herramienta léxico-morfológica por lo que no dispone de esta información.

En ausencia de un lexicón, AGIR reduce el número de posibles papeles temáticos a los siguientes: *ACCIÓN*, *AGENTE*, *TEMA* y *MODIFICADOR*, que se corresponden con el verbo, sujeto, objeto directo y sintagmas preposicionales de la cláusula respectivamente. Para identificar estos papeles temáticos en una cláusula cuando se ha realizado un análisis parcial y no se dispone de información semántica se establece la siguiente heurística (H_2):

H_2 Asumimos que el sintagma nominal analizado antes del verbo es el *AGENTE* de la cláusula. Del mismo modo, el sintagma nominal analizado después del verbo es el *TEMA* de la cláusula. Por último, todos los sintagmas preposicionales encontrados en la cláusula son sus *MODIFICADORES.*

Hay que destacar que si el verbo de la cláusula está en pasiva la heurística H_2 determina que el sintagma nominal analizado antes del verbo es el *TEMA* de la cláusula y el sintagma preposicional que hay después encabezado por la preposición "*por*" introduce el *AGENTE.*

Representación interlingua de las cláusulas. Tras identificar las distintas cláusulas del texto y los papeles temáticos de las mismas, la siguiente fase consiste en obtener la representación interlingua de cada cláusula. Para tal fin se ha utilizado una *estructura de rasgos* compleja para cada cláusula. Consideremos el ejemplo 99 en el que aparece un fragmento de texto en español (se corresponde con el ejemplo 59 del capítulo 5; aquí se ha repetido por claridad).

(99) Los chicos de las montañas estaban en el jardín porque ellos estaban cogiendo flores.

En la figura 6.9 se muestra la representación interlingua tal y como se ha implementado en el sistema AGIR de la primera cláusula (*Los chicos de las montañas estaban en el jardín*[8]).

En la figura 6.9 se han encontrado los siguientes elementos: *ACCIÓN*= "estaban", *AGENTE*= "los chicos de las montañas", *TEMA*= "Ø" (ya que no se ha encontrado ningún sintagma nominal después del verbo) y *MODIFICADOR*= "en el jardín". Los distintos atributos que contienen estas estructuras de rasgos ya se han explicado con detalle en la sección 5.2.1.

Como se puede apreciar en la implementación realizada en AGIR, en la representación interlingua sólo aparecen tres estructuras de rasgos distintas que se utilizan para representar: (1) un verbo (papel temático *ACCIÓN*), (2) un sintagma nominal (papeles temáticos *AGENTE* y *TEMA*) y (3) un sintagma preposicional (papel temático *MODIFICADOR*).

Hay que destacar que el atributo *Núcleo* de la *estructura de rasgos* contiene la unidad léxica interlingua. En AGIR, ésta ha sido representada por la palabra y su sentido correcto en WordNet. Realizando el acceso correspondiente al módulo interlingua ILI de EuroWordNet nos permitirá la generación correcta de esta unidad léxica en el idioma destino. Una aproximación similar aparece en el trabajo de Dorr *et al.* (1997) en el que se presenta

[8] La representación interlingua de esta cláusula se ha presentado previamente en la figura 5.3 en la que se presentaba la propuesta general de representación interlingua.

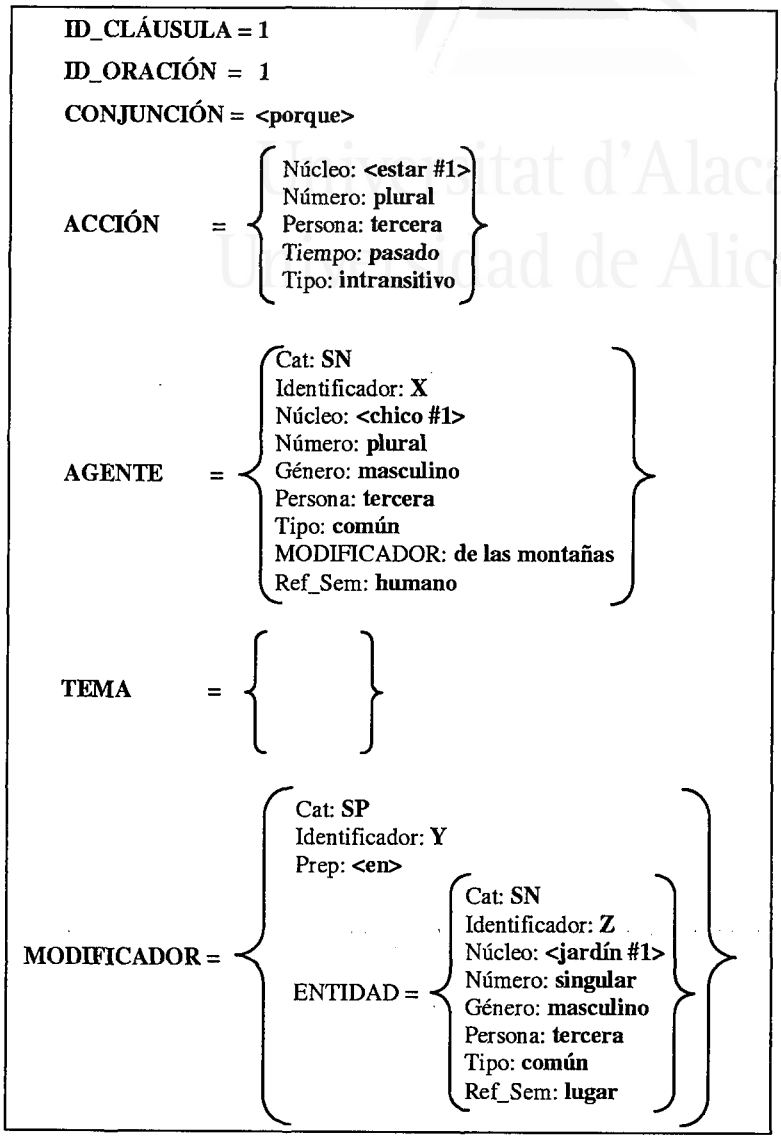


Figura 6.9. Representación interlingua en AGIR de la cláusula *Los chicos de las montañas estaban en el jardín*

una estructura interlingua que incluye enlaces bilingües (mediante EuroWordNet) entre las palabras del idioma origen y destino respectivas. En él se muestra la construcción manual de una base de datos que contiene únicamente verbos en inglés (con sus *synsets* correspondientes) y los enlaces a los respectivos verbos en español. Esta estructura interlingua es independiente del idioma y permitirá la correcta generación de los verbos en el idioma destino. Tomando como base este trabajo, en AGIR se extiende el uso de los enlaces bilingües entre verbos al uso de enlaces multilingües (utilizando EuroWordNet) entre todas las palabras del texto incluyendo: nombres, verbos, adjetivos, etc.

Representación interlingua del texto completo. En la última fase, se obtiene la representación interlingua de todo el texto. Para ello se representan todas las cláusulas del texto y las distintas entidades que aparecen en el mismo incluyendo las relaciones entre ellas. La representación interlingua obtenida ya se ha explicado exhaustivamente en la sección 5.2.1. Como salida de esta fase se obtienen dos listas: en la primera de ellas se almacenan las distintas entidades del texto; la segunda contiene las cláusulas en las que se incluyen los papeles temáticos con los enlaces correspondientes a las entidades. Estas listas contienen la información necesaria para tratar en el idioma destino las anáforas intersentenciales y las cadenas de correferencia.

6.2 Implementación del módulo de generación del sistema AGIR

El módulo de generación del sistema AGIR recibe como entrada la representación interlingua del texto en el idioma origen (lista de entidades y lista de cláusulas) y produce como salida el texto en el idioma destino.

Como en esta Tesis no se pretende realizar la traducción de un texto completo de un idioma a otro sino que el objetivo principal es la generación de la anáfora pronominal en el idioma destino, mostraremos exclusivamente las etapas de generación sintáctica

y morfológica centrándonos en las *discrepancias* entre español e inglés en el tratamiento de los pronombres.

En esta sección presentaremos el algoritmo utilizado en el sistema AGIR para la generación de la anáfora pronominal en el idioma destino y, a continuación, analizaremos la implementación de las discrepancias sintácticas (pronombres pleonásticos y cero pronombres) y las discrepancias morfológicas (discrepancias de número y género).

6.2.1 Algoritmo para la generación de la anáfora pronominal del sistema AGIR

El algoritmo para la generación de la anáfora en AGIR va precedido del correspondiente algoritmo de análisis. En la figura 6.10 se muestran ambos algoritmos en pseudocódigo.

Estos algoritmos son válidos tanto para el inglés como el español. El algoritmo de análisis de AGIR se ha presentado porque las discrepancias sintácticas entre español e inglés (*it* pleonásticos y cero pronombres) son identificadas antes de obtener la representación interlingua del texto. En concreto, la detección de los pronombres *it* pleonásticos se realiza después del análisis léxico-morfológico del texto. Por otra parte, la detección de los cero pronombres se realiza después del análisis parcial.

A partir de la representación interlingua se lleva a cabo la última etapa de la generación (generación morfológica) en la que se tratan las discrepancias de número y género.

A continuación presentamos la implementación de las discrepancias sintácticas y las discrepancias morfológicas.

6.2.2 Identificación y tratamiento de los pronombres pleonásticos

En el sistema AGIR se han implementado las reglas basadas en patrones para la detección de los pronombres *it* pleonásticos presentadas en la sección 5.3.2. El algoritmo para la detección de los pronombres *it* pleonásticos se ha implementado con el lenguaje de programación AWK para Linux.

```
Procedimiento ANÁLISIS (T)
Sea T = texto en el idioma origen a traducir

Análisis léxico y morfológico de T
Detección de pronombres no referenciales y referenciales de T
Para cada oración de T
    Análisis sintáctico parcial (creación lista "antecedentes"
    oraciones previas)
    Detección cero pronombres
    Para cada anáfora pronominal
        Resolución anáfora
    FinPara
FinPara
Obtención de la representación interlingua de T
Devolver E = lista de "entidades" y C = lista de "cláusulas"
FinProcedimiento


Procedimiento GENERACIóN (E,C)
Sea E = lista de "entidades" y C = lista de "cláusulas"

Para cada cláusula
    Para cada anáfora pronominal
       Si pertenece a discrepancias de número
          Entonces generación aplicando reglas morfológicas
                   número
          Sino (pertenece a discrepancias de género) generación
               aplicando reglas morfológicas género
       FinSi
    FinPara
FinPara
FinProcedimiento
```

Figura 6.10. Algoritmo para la generación de la anáfora pronominal en AGIR

El algoritmo escrito en pseudocódigo en el que aparecen las reglas para la detección de los pronombres pleonásticos numeradas del mismo modo en el que aparecieron en la sección 5.3.2 se muestra en la figura 6.11.

Como se observa en la figura 6.11, el algoritmo comienza con la separación del texto en fragmentos finalizados con un signo de puntuación (“.” “;” “:” “?” “!” “)”). Esta fragmentación se realiza ya que la búsqueda de un patrón que contiene un pronombre *it* no debe sobrepasar estos signos de puntuación.

Una vez separado el texto en fragmentos se busca en cada uno de ellos alguna ocurrencia del pronombre *it*. Para cada una de éstas se comprueba si es referencial (anafórico) o pleonástico. Si

```
Procedimiento DETECCIÓN_IT_PLEONÁSTICOS (T)
Sea T = texto en el idioma origen a traducir

Separar T en fragmentos de texto (F) según signos de puntuación
Para cada fragmento de texto F
   Para cada pronombre "it" de F
       Si "it" aparece precedido de una preposición
          Entonces marcar como Referencial Tipo 1
          Sino Si pertenece al patrón de la regla 1
                  Entonces marcar como Pleonástico Tipo 1
                  Sino Si pertenece al patrón de la regla 2
                          Entonces marcar Pleonástico Tipo 2
                          .
                          .
                          .
                       Si pertenece al patrón de la regla 9
                          Entonces marcar Pleonástico Tipo 9
                          Sino marcar como Referencial Tipo 2

       FinSi
   FinPara
FinPara
FinProcedimiento
```

Figura 6.11. Algoritmo para la detección de los pronombres *it* pleonásticos en AGIR

va precedido de una preposición se determina que es referencial y se etiqueta como Referencial[9] Tipo 1. En caso contrario, se comprueba si pertenece al patrón de alguna de las reglas que lo identifica como pleonástico. Por último, si ha superado todas las reglas y no cumple ninguna, se determina que es referencial y se etiqueta como Referencial Tipo 2.

Como se puede comprobar, este algoritmo marca los pronombres *it* pleonásticos y referenciales según la regla que cumplan. Esta información es útil ya que permite realizar un estudio estadístico de las distintas apariciones de cada uno de los patrones que permite modificar, añadir o eliminar algunas de las reglas aplicadas.

[9] El marcado de los pronombres *it* se realiza añadiendo información a su etiqueta léxico-morfológica del tipo correspondiente. Por ejemplo, si la etiqueta del pronombre es *PPH1* y es Referencial Tipo 1, la etiqueta se modifica resultando la etiqueta *PPH1R1*. Si es Pleonástico Tipo 1 la etiqueta resultante será *PPH1P1*, etc.

Por último, destacar que aquellos pronombres *it* que hayan sido marcados como referenciales serán resueltos como tales en el módulo de resolución anafórica y serán generados en el idioma destino. Aquéllos que hayan sido marcados como pleonásticos no serán tratados en el mencionado módulo y no se generarán en el idioma destino.

En la figura 6.12 se muestra, a modo de ejemplo, un fragmento del código en AWK de las reglas para Referencial Tipo 1 y Pleonástico Tipo 1, Tipo 2 y Tipo 3.

6.2.3 Identificación y tratamiento de los cero pronombres

Aunque el tratamiento computacional de los cero pronombres ya se ha estudiado en otros idiomas, como por ejemplo el japonés (Nakaiwa & Ikehara, 1992; Nakaiwa & Shirai, 1996), en español aún no se ha tratado este problema. En esta Tesis presentamos la implementación realizada en el sistema AGIR para tratar este fenómeno.

Para generar correctamente en inglés los cero pronombres españoles con función de sujeto, éstos se deben detectar previamente (fase de detección) y después se deben resolver (fase de resolución anafórica). En la fase de detección se debe de obtener información del cero pronombre (persona, género y número) del verbo de la cláusula en la que aparece; esta información se usará posteriormente para identificar el antecedente del cero pronombre (fase de resolución).

La identificación y tratamiento de los cero pronombres se llevan a cabo en AGIR tras haber realizado un análisis sintáctico parcial del texto. En Peral & Ferrández (2000b) y Ferrández & Peral (2000) se presenta la implementación realizada para resolver los cero pronombres cuando se ha realizado este tipo de análisis. A continuación se muestran las dos fases:

Fase de detección de los cero pronombres. Tras realizar el análisis sintáctico parcial de una oración se realiza la identificación de cláusulas de la misma aplicando la heurística H_1. Una vez que

```
function itReferencial_1(linea)
# PREPOSICION + IT --> it referencial
# Ej: after it
{
   gsub("IN [A-Za-z]* [A-Za-z]*\nit PPH1", "&R1", linea);
   gsub("TO [A-Za-z]* [A-Za-z]*\nit PPH1", "&R1", linea);
   gsub("OF [A-Za-z]* [A-Za-z]*\nit PPH1", "&R1", linea);
   return linea;
}

function itConstruccion_1(linea)
# IT ... TASK_STATUS_WORD ... TO + VERBO_INFINITIVO -->
# it pleonastico. Ej: it is necessary to limit
{
   sub("PPH1 .*[(abnormal)|(advantageous)|(advisable)|
   (appropiate)|(bad)|(beneficial)|(best)|(better)|(common)|
   (correct)|(customary)|(dangerous)|(decided)|(difficult)|
   (easier)|(easiest)|(easy)|(essential)|(faster)|(fastest)|
   (feasible)|(fitting)|(foolish)|(good)| . . . |(surprising)]
    .* TO [A-Za-z]* [A-Za-z]*\n[A-Za-z]* VB","LL&",linea);
   sub("LLPPH1","PPH1P1",linea);
   return linea;
}

function itConstruccion_2(linea)
# IT ... SUBORDINATING_CONJUNCTION_THAT --> it pleonastico
# Ej: it is inevitable that
{
   sub("PPH1 .*that IN ","LL&",linea);
   sub("LLPPH1","PPH1P2",linea);
   return linea;
}

function itConstruccion_3(linea)
# IT ... STATE_OF_KNOWLEDGE_WORD ... (WHETHER|IF...WHERE) -->
# it pleonastico. Ej: it is not known whether
{
   sub("PPH1 .*[(certain)|(clear)|(debatable)|(doubted)|
   (doubtful)|(dubious)|(known)|(questionable)|(questioned)|
   (uncertain)|(unclear)|(understood)|(unknown)|(wondered)|
   (could)|(should)|(would)] .*whether","LL&",linea);
#  . . . (if, what, how, why, when, where)
   sub("LLPPH1","PPH1P3",linea);
   return linea;
}
```

Figura 6.12. Código en AWK para la detección de los pronombres *it* pleonásticos

la oración se ha dividido en cláusulas, la siguiente tarea a realizar es la detección de la omisión del sujeto de cada cláusula.

Si se han aplicado técnicas de análisis parcial podemos establecer la heurística H_3 para detectar la omisión del sujeto de cada cláusula:

H_3 Después de que la oración haya sido dividida en cláusulas, se busca un sintagma nominal o un pronombre para cada cláusula entre todos los constituyentes situados a la izquierda del verbo (siempre que éste no sea imperativo o impersonal). Si se encuentra, éste debe concordar en número y persona con el verbo de la cláusula.

En el ejemplo 100 (traducción a español del ejemplo 98) se observa que la primera cláusula con el verbo *llegaron* no tiene su sujeto omitido ya que aparece el sintagma nominal *Juan y Pedro* antes del verbo. Sin embargo, hay un cero pronombre (*ellos*) para la segunda cláusula con verbo *durmieron* ya que éste no tiene ningún sintagma nominal o pronombre a su izquierda.

(100) *[Juan y Pedro]*$_i$ llegaron tarde al trabajo porque $\emptyset_i$ se durmieron.

Con respecto al número de constituyentes que se buscan antes del verbo para encontrar un sintagma nominal o un pronombre, se establece una ventana de búsqueda que contiene los cuatro constituyentes anteriores al verbo. La razón por la que se determina este tamaño viene justificada por la posible aparición de adverbios y pronombres reflexivos o de complemento antes del verbo. En el ejemplo 101 aparece una oración en la que aparecen tres constituyentes (marcados con subíndices numéricos) entre el sujeto y el verbo de la misma.

(101) *Eva* no$_1$ se$_2$ lo$_3$ *ha comprado* todavía.

Fase de resolución de los cero pronombres. Tras la detección del cero pronombre, nuestro sistema computacional inserta el pronombre en la posición en la cual se había omitido. Este pronombre se resolverá en el siguiente módulo de resolución anafórica.

La información sobre el número y la persona del cero pronombre se extrae del verbo de la cláusula. La información del género del mismo no estará normalmente disponible en el módulo de resolución anafórica. Sin embargo, algunas veces la información del género del cero pronombre se puede obtener cuando el verbo de la cláusula sea copulativo.

(102) *[Carlos]$_i$* fue con Isabel aunque $\emptyset_i$ estaba disgustado con su actitud.

En el ejemplo 102 la segunda cláusula tiene su sujeto omitido y el verbo de la misma es copulativo. En estos casos, el sujeto de la cláusula debe concordar en género y número con su objeto siempre que éste tenga formas lingüísticas distintas para el género masculino y femenino. En este ejemplo el objeto es *disgustado* y tiene las dos formas (disgustado: masculino, disgustada: femenino). Por lo tanto, la información de número y persona se extrae del verbo (*singular* y *tercera*) y la información del género se extrae del objeto (*masculino*). El pronombre omitido será, pues, el pronombre español *él*.

Con respecto a la obtención de la información del objeto del verbo cuando se ha realizado análisis parcial, simplemente hay que buscar un sintagma nominal a la derecha del verbo copulativo. Ya que AGIR trabaja sobre la salida del etiquetador (*POS tagger*), éste no proporciona la información relativa a la posibilidad de que el objeto pueda tener formas lingüísticas distintas para masculino y femenino. Por esta razón, nuestro sistema computacional siempre añadirá la información del género al cero pronombre aunque ésta se tratará como una preferencia en el módulo de resolución anafórica. Sin embargo, la información de número y persona se considerará como una restricción[10].

Una vez que se ha resuelto el cero pronombre (identificando su antecedente) se procede a la obtención de la representación in-

[10] Esta característica es una diferencia importante entre la resolución de anáforas pronominales y cero pronombres. Para las primeras, se utiliza la información de género, número y persona como restricción. Por su parte, los segundos utilizan la información de número y persona como restricción mientras que la información de género se utiliza como preferencia.

terlingua en la que se almacena toda la información necesaria del cero pronombre (incluyendo la de su antecedente) para su correcta generación en el idioma destino. Posteriormente, se tratará la generación morfológica en la que se obtendrá la forma definitiva del cero pronombre en el idioma destino.

6.2.4 Tratamiento de las discrepancias de número

Las discrepancias de número y género entre español e inglés en AGIR (Peral *et al.*, 1999b; Peral, 1999; Peral *et al.*, 1999a; Peral & Ferrández, 2000a) se tratan en la última etapa de la generación (generación morfológica).

Para el tratamiento de las discrepancias de número se han utilizado las reglas morfológicas (de número) presentadas en la sección 5.3.3 adaptándolas a la representación interlingua que se ha obtenido tras haber realizado un análisis sintáctico parcial del texto. Recordemos que en esta representación interlingua sólo existen cuatro papeles temáticos: *AGENTE*, *ACCIÓN*, *TEMA* y *MODIFICADOR*. Con estos papeles temáticos el tratamiento de las discrepancias de número sería el siguiente:

Pronombres con función de sujeto. En la implementación realizada de AGIR estos pronombres se identifican cuando tienen el papel temático de *AGENTE*, por lo tanto, las reglas morfológicas para el tratamiento de las discrepancias de número con estos pronombres son las ya presentadas en las figuras 5.7 y 5.8.

Pronombres con función de complemento. Los pronombres con función de complemento se identifican en la implementación realizada de AGIR porque tienen los papeles temáticos de *TEMA* o *MODIFICADOR*.

En la figura 6.13 se presentan unos ejemplos de las reglas morfológicas construidas en AGIR para tratar las discrepancias de número en la traducción español-inglés de estos pronombres.

En la figura 6.14 se presentan las reglas morfológicas para la traducción inglés-español de estos pronombres.

TEMA_o_MODIFICADOR (PRONOMBRE + 3ª persona + singular + antecedente <policía>) → **them** (PRONOMBRE + 3ª persona + plural)

TEMA_o_MODIFICADOR (PRONOMBRE + 3ª persona + singular + antecedente <ganado>) → **them** (PRONOMBRE + 3ª persona + plural)

Figura 6.13. Discrepancias de número en AGIR. Traducción español-inglés de pronombres con función de complemento

TEMA (PRONOMBRE + 3ª persona + plural + antecedente <police>) → **la** (PRONOMBRE + 3ª persona + femenino + singular)

TEMA (PRONOMBRE + 3ª persona + plural + antecedente <cattle>) → **lo** (PRONOMBRE + 3ª persona + masculino + singular)

MODIFICADOR (PRONOMBRE + 3ª persona + plural + + antecedente <police>) → **ésta** (PRONOMBRE + femenino + singular)

MODIFICADOR (PRONOMBRE + 3ª persona + plural + + antecedente <cattle>) → **éste** (PRONOMBRE + masculino + singular)

Figura 6.14. Discrepancias de número en AGIR. Traducción inglés-español de pronombres con función de complemento

6.2.5 Tratamiento de las discrepancias de género

Para el tratamiento de las discrepancias de género se han adaptado las reglas morfológicas (de género) presentadas en la sección 5.3.3 para la representación interlingua con análisis parcial. El tratamiento de las discrepancias de género es el siguiente:

Pronombres con función de sujeto. Estos pronombres se identifican porque tienen el papel temático de *AGENTE*. Las reglas de las figuras 5.11 y 5.12 son válidas para la implementación realizada en AGIR.

En las reglas mostradas en estas figuras se utiliza información semántica acerca del antecedente para clasificarlo como *persona*, *animal* u *objeto*. Esta información es de vital importancia para la generación correcta en el idioma destino. Si el corpus de trabajo no dispone de esta información (porque las palabras no se han etiquetado con información semántica de WordNet) se utilizan una

serie de heurísticas para la generación de estos pronombres en el idioma destino basadas en información obtenida tras el análisis sintáctico y en el tipo de corpus de trabajo. Por ejemplo, el análisis proporciona información del tipo de sintagma nominal –fecha, dirección cantidad, persona (si el sintagma nominal comienza con un tratamiento de persona: Sr., Sra., etc.), hora, etc.– que se puede utilizar para su generación en el idioma destino. Por otra parte, el corpus puede restringir, en cierto modo, el tipo de pronombres que aparecen en él –por ejemplo, si el corpus de trabajo es un manual técnico aparecerán muy pocos pronombres (o ninguno) que se refieran a una *persona* o *animal*–.

Pronombres con función de complemento. Estos pronombres se identifican porque tienen el papel temático de *TEMA* o *MODIFICADOR* en la representación interlingua de AGIR.

En la traducción español-inglés de este tipo de pronombres se utilizarán las reglas morfológicas presentadas en la figura 6.15.

TEMA_o_MODIFICADOR (PRONOMBRE + 3ª persona + singular) +
+ antecedente (persona con género masculino)) → **him**

TEMA_o_MODIFICADOR (PRONOMBRE + 3ª persona + singular) +
+ antecedente (persona con género femenino)) → **her**

TEMA_o_MODIFICADOR (PRONOMBRE + 3ª persona + singular) +
+ antecedente (animal u objeto)) → **it**

TEMA_o_MODIFICADOR (PRONOMBRE + 3ª persona + plural)
→ **them**

Figura 6.15. Discrepancias de género en AGIR. Traducción español-inglés de pronombres con función de complemento

Si no se dispone de información semántica para obtener el tipo de antecedente (*persona*, *animal* u *objeto*), se utilizarán las heurísticas mencionadas anteriormente.

Por último, en la traducción inglés-español se emplearán las reglas de la figura 6.16.

TEMA (PRONOMBRE + 3ª persona + singular) +
+ antecedente (persona)) → **le**

TEMA (PRONOMBRE + 3ª persona + plural) +
+ antecedente (persona)) → **les**

TEMA (PRONOMBRE + 3ª persona + singular) +
+ antecedente (animal u objeto con género masculino)) → **lo**

TEMA (PRONOMBRE + 3ª persona + singular) +
+ antecedente (animal u objeto con género masculino)) → **la**

TEMA (PRONOMBRE + 3ª persona + plural) +
+ antecedente (animal u objeto con género masculino)) → **los**

TEMA (PRONOMBRE + 3ª persona + plural) +
+ antecedente (animal u objeto con género femenino)) → **las**

MODIFICADOR (PRONOMBRE + 3ª persona + singular) +
+ antecedente (género masculino)) → **él**

MODIFICADOR (PRONOMBRE + 3ª persona + singular) +
+ antecedente (género femenino)) → **ella**

MODIFICADOR (PRONOMBRE + 3ª persona + plural) +
+ antecedente (género masculino)) → **ellos**

MODIFICADOR (PRONOMBRE + 3ª persona + plural) +
+ antecedente (género femenino)) → **ellas**

Figura 6.16. Discrepancias de género en AGIR. Traducción inglés-español de pronombres con función de complemento

7. Evaluación del sistema AGIR

En este capítulo se presenta la evaluación del sistema AGIR. Para ello se mostrará una evaluación independiente de las distintas tareas llevadas a cabo en el sistema global y que conducen a la tarea final: la generación de la anáfora pronominal en el idioma destino.

Para la evaluación de las distintas tareas se ha utilizado la misma metodología: en primer lugar se han seleccionado una serie de corpus sobre los que se van a realizar los distintos experimentos. Posteriormente, se han elegido aleatoriamente fragmentos de estos corpus para realizar la fase de entrenamiento (corpus de entrenamiento) y se ha reservado el resto para la evaluación (corpus de evaluación). En la fase de entrenamiento se realiza una evaluación sobre el corpus de entrenamiento y, tras analizar los resultados obtenidos, se realizan las modificaciones oportunas con el objetivo de mejorar los resultados. Este proceso se repite con la realización de distintos experimentos hasta que se obtiene la configuración óptima. Por último, cuando el sistema se ha ajustado con la mejor configuración, se realiza la evaluación final del sistema sobre el corpus de evaluación. Para garantizar la fiabilidad de los resultados se han seleccionado corpus de distintos tipos: textos técnicos (incluyen manuales sobre ordenadores, aplicaciones informáticas, conceptos de telecomunicaciones, etc.) y textos narrativos (incluyen textos de diferentes estilos que tratan temas diversos –deportes, religión, leyes, fragmentos de novelas, etc.– y que han sido desarrollados por autores diversos). Esta metodología asegura que todos los experimentos realizados en la fase de entrenamiento son independientes de la fase de evaluación y que los resultados obtenidos son fiables.

Las distintas tareas que se han evaluado son las siguientes: detección de pronombres pleonásticos en inglés, detección y resolución de cero pronombres y, por último, resolución y generación de la anáfora pronominal. Para cada una de ellas se presentará la fase de entrenamiento y la fase de evaluación.

El capítulo comenzará con una breve descripción de los corpus que se han utilizado en la evaluación global del sistema, posteriormente se presentará la metodología de evaluación y, por último, se mostrará la evaluación de las diferentes tareas.

7.1 Corpus

Como se ha comentado previamente, para garantizar la fiabilidad de los resultados obtenidos en la evaluación de las distintas tareas se han seleccionado corpus de dominio no restringido de diferentes tipos tanto para español como para inglés. Un fragmento de cada uno de ellos se ha utilizado en la fase de entrenamiento y el resto del corpus en la fase de evaluación.

A continuación presentamos las características principales de los diferentes corpus utilizados:

- **LEXESP.**
  Corpus en español que contiene 5 millones de palabras etiquetadas léxica y morfológicamente perteneciente al proyecto del mismo nombre llevado a cabo en el Departamento de Psicología de la Universidad de Oviedo y desarrollado por el Grupo de Lingüística Computacional de la Universidad de Barcelona con la colaboración del Grupo de Procesamiento del Lenguaje de la Universidad Politécnica de Cataluña. Para el etiquetado léxico-morfológico del texto se ha utilizado el conjunto de etiquetas PAROLE (Martí *et al.*, 1998).
  El corpus consta de una serie de textos de tipos diferentes que han sido escritos por autores diversos. En su mayoría son textos periodísticos y artículos de opinión sobre temas variados: política, deporte, asuntos sociales, etc. Otro grupo de textos son breves narraciones literarias sobre un tópico concreto. Por último, hay textos que son fragmentos de novelas literarias.

En la evaluación del sistema AGIR se ha utilizado un fragmento de corpus formado por 31 textos que contienen 38.999 palabras. En la tabla 7.1 se muestra el número de oraciones de cada texto, las palabras que contiene y el número medio de palabras por oración.

- **BLUE BOOK (BB).**
  El corpus Blue Book es un texto técnico que contiene el manual de telecomunicaciones de CCITT (International Telecommunications Union CCITT) y que ha sido publicado en español, inglés y francés. En el etiquetado léxico-morfológico para español se ha utilizado el conjunto de etiquetas del Xerox POS Tagger adaptado al español en el Proyecto CRATER (Proyecto CRATER, 1994-1995). En inglés, el corpus ha sido etiquetado por el Xerox POS Tagger (Cutting *et al.*, 1992).
  Para la evaluación de AGIR se ha utilizado un fragmento del corpus en español que contiene 15.571 palabras (509 oraciones) con una media de 30,6 palabras por oración.
- **Federal Register (FR).**
  Corpus en inglés que forma parte de las colecciones de datos utilizadas en la Conferencia de Recuperación de Información TREC[1] (Text REtrieval Conference) patrocinada por el Instituto Nacional de Estándares y Tecnología (National Institute of Standards and Technology, NIST) y la Agencia de Investigación de Defensa (Defense Advanced Research Projects Agency, DARPA) de los Estados Unidos. En la evaluación realizada este texto se ha etiquetado previamente con el etiquetador TreeTagger (Schmid, 1994) que utiliza un conjunto de etiquetas basadas en las que fueron definidas para anotar el Brown Corpus (Francis, 1964; Francis & Kucera, 1982).
  El corpus contiene una recopilación de noticias internacionales sobre temas diversos separadas en distintos documentos. En la evaluación del sistema se ha utilizado un fragmento que contiene 313 documentos con 156.831 palabras (6.325 oraciones) con una media de 24,7 palabras por oración.

[1] http://trec.nist.gov (página visitada el 25/08/00).

LEXESP	Número oraciones	Número palabras	Palabras por oración
texto 1	38	965	25,395
texto 2	55	988	17,964
texto 3	34	921	27,088
texto 4	34	944	27,765
texto 5	59	1006	17,051
texto 6	40	977	24,425
texto 7	36	943	26,194
texto 8	32	974	30,438
texto 9	34	1015	29,853
texto 10	31	976	31,484
texto 11	35	1153	32,943
texto 12	47	712	15,149
texto 13	54	1058	19,593
texto 14	44	1039	23,614
texto 15	39	810	20,769
texto 16	28	593	21,179
texto 17	29	840	28,966
texto 18	41	740	18,049
texto 19	21	846	40,286
texto 20	44	1085	24,659
texto 21	16	343	21,438
texto 22	43	782	18,186
texto 23	38	940	24,737
texto 24	97	1895	19,536
texto 25	115	2998	26,070
texto 26	171	4972	29,076
texto 27	41	1039	25,341
texto 28	33	988	29,939
texto 29	102	2323	22,775
texto 30	90	2630	29,222
texto 31	67	1504	22,448
TOTAL	**1588**	**38999**	**24,559**

Tabla 7.1. Corpus LEXESP. Número de oraciones y palabras

- **SEMCOR.**
  Corpus en inglés que se incluye como un paquete adicional de WordNet, presentado en Landes *et al.* (1998). El etiquetado léxico-morfológico ha sido realizado por el etiquetador estocástico de Brill (Brill, 1992). La principal característica de este corpus consiste en que además del etiquetado léxico-morfológico, cada una de las palabras ha sido etiquetada semánticamente con su sentido en WordNet. A diferencia del resto de corpus es el único que proporciona información semántica de las palabras del mismo.
  Contiene textos de estilo narrativo sobre temas diferentes: leyes, deporte, religión, naturaleza, etc. Se ha utilizado un fragmento formado por 11 textos con 23.788 palabras. En la tabla 7.2 se muestra el número de oraciones de cada texto, las palabras que contiene y el número medio de palabras por oración.

SEMCOR	Número oraciones	Número palabras	Palabras por oración
texto a01	90	2074	23,044
texto a02	88	2018	22,932
texto a11	79	2097	26,544
texto a12	100	2265	22,650
texto a13	102	2113	20,716
texto a14	123	2145	17,439
texto a15	101	2248	22,257
texto d01	82	2134	26,024
texto d02	61	2240	36,721
texto d03	75	2207	29,427
texto d04	77	2247	29,182
TOTAL	**978**	**23788**	**24,323**

Tabla 7.2. Corpus SEMCOR. Número de oraciones y palabras

- **Manuales Técnicos de Informática (MTI).**
  Corpus en inglés formado por un conjunto de textos técnicos proporcionados por el Grupo de Lingüística Computacional de la Escuela de Humanidades, Idiomas y Ciencias Sociales de la

Universidad de Wolverhampton (Reino Unido). Este grupo de investigación etiquetó anafóricamente el corpus, marcando la anáfora y su antecedente correcto con un formato propio. En la evaluación de AGIR previamente el corpus se ha etiquetado léxicamente con el etiquetador TreeTagger.
Contiene manuales de Informática sobre temas diversos: aplicaciones informáticas, procesadores de textos, funcionamiento de dispositivos, etc. Se ha utilizado un fragmento de 7 textos con 101.843 palabras. En la tabla 7.3 se muestra el número de oraciones de cada texto, las palabras que contiene y el número medio de palabras por oración.

MTI	Número oraciones	Número palabras	Palabras por oración
BEOWULF	396	7175	18,119
CDROM	650	10033	15,435
MAC	1044	17189	16,465
PSW	537	7432	13,840
WINDOWS	170	3258	19,165
SCANWORX	3181	45482	14,298
GIMP	609	11274	18,512
TOTAL	**6587**	**101843**	**15,461**

Tabla 7.3. Corpus MTI. Número de oraciones y palabras

7.2 Metodología de evaluación

Para comprobar el rendimiento de AGIR se ha utilizado un módulo de evaluación automática de resultados. Este módulo compara la solución correcta de cada una de las anáforas del corpus (previamente se han etiquetado manualmente y se han almacenado electrónicamente) con la salida proporcionada por el sistema AGIR.

El etiquetado anafórico del corpus se realiza definiendo un predicado PROLOG *oSol* para cada anáfora que aparece en una nueva oración, el cual contiene dos parámetros:

1. *Oración.* Contiene el número de oración donde aparece la anáfora.
2. *Lista de anáforas.* Contiene una lista PROLOG en la que aparece un predicado llamado *anaf* por cada una de las anáforas que hay en esa oración. Contiene 7 argumentos:
   a) *Anáfora.* Lista que contiene las palabras que forman la expresión anafórica.
   b) *Antecedente.* Lista que contiene la solución (el antecedente) de la expresión anafórica formada por el predicado *sol* que contiene 4 argumentos: una lista con las palabras que constituyen el antecedente, el número de oración donde éste aparece, el número de palabra de la oración donde éste comienza y el número de palabra donde acaba.
   c) *Tipo.* Indica el tipo de relación que existe entre la expresión anafórica y su antecedente. En AGIR se han utilizado los siguientes tipos para español: *compl* (complemento), *complRefl* (reflexiva), *pronSuj* (pronombre no incluido en un sintagma preposicional), *pronSP* (dentro de un sintagma preposicional) y *pronOmitido* (cero pronombres). Para inglés se han usado: *persIt* (pronombre *it*), *persNoRefl* (pronombres personales de sujeto y complemento no reflexivos excepto el pronombre *it*).
   d) *Dirección.* Puede adoptar 3 valores: "<" indica que el antecedente se encuentra antes de la anáfora, ">" indica que el antecedente se encuentra después de la anáfora (catáfora) y "!" indica que el antecedente no aparece en el texto o está formado por varios sintagmas nominales que han aparecido de una forma separada en el texto (*split antecedents*). En esta evaluación sólo se tratarán las anáforas del primer tipo.
   e) *Palabra Inicio.* Indica el número de la palabra dentro de la oración donde empieza la expresión anafórica.
   f) *Palabra Fin.* Indica el número de la palabra dentro de la oración donde finaliza la anáfora.
   g) *Generación.* Indica la generación correcta de la anáfora en el idioma destino.

En la figura 7.1 se presenta un ejemplo de anotación correferencial en español e inglés de fragmentos de corpus utilizados en la evaluación de AGIR.

```
oSol(2572,[anaf([w('gritaron','gritar','VMIS3PO')],[sol([w('sus
','su','DP3CP00'),w('compañeros','compañero','NCMP000')],2572,0
,2)],pronOmitido,<,7,8,they)]).

oSol(56,[anaf([w('It','it','PPH1')],[sol([w('the','the','DT'),w
('error','error','NN'),w('message','message','NN'),w('appendix'
,'appendix','NN')],55,3,7)],persIt,<,0,1,éste),anaf([w('them','
them','PPHO2')],[sol([w('error','error','NN'),w('conditions','c
ondition','NNS')],56,5,7)],persNoRefl,<,13,14,ellas)]).
```

Figura 7.1. Ejemplo de anotación correferencial en español e inglés

En el ejemplo de anotación en español aparece un cero pronombre (*tipo pronOmitido*) que empieza en la palabra *7* y finaliza en la palabra *8* de la oración *2572*. En este caso, la expresión anafórica (como es un cero pronombre) está formada por el verbo cuyo sujeto está omitido (*gritaron*). La solución de esta anáfora es el sintagma nominal *sus compañeros* de la oración *2572* que está entre las posiciones *0* y *2* de la oración. El pronombre inglés correspondiente a este cero pronombre es el pronombre *they*.

En el ejemplo en inglés se observa que la oración *56* tiene *2* anáforas. La primera de ellas es el pronombre *it* (*tipo persIt*) cuya solución es *the error message appendix* y el correspondiente pronombre español es *éste*. La segunda anáfora es el pronombre *them* (*tipo persNoRefl*) que tiene como solución *error conditions* y el pronombre español es *ellas*.

Para medir la capacidad que tiene el sistema de resolver correctamente las anáforas se han utilizado dos medidas[2]:

- *Precisión*. Se calcula como el cociente entre el número de anáforas correctamente resueltas por el sistema (N) y el número total de anáforas resueltas (T), $Precisión = N/T$.

[2] Para calcular estas medidas hay que tener en cuenta que el sistema detecta e intenta resolver todas las anáforas que han sido etiquetadas como pronombres personales por los etiquetadores léxico-morfológicos utilizados.

- *Cobertura.* Se calcula como el cociente entre el número de anáforas correctamente resueltas por el sistema (N) y el número de anáforas existentes (E), $Cobertura = N/E$.

7.3 Detección de los pronombres no referenciales

En la implementación del sistema AGIR se realiza la detección de los pronombres no referenciales en inglés. En concreto, ésta se centra en la identificación de los pronombres *it* pleonásticos.

Este módulo de identificación de los *it* pleonásticos ha sido presentado en detalle en la sección 6.2.2. El algoritmo utilizado clasifica todas las ocurrencias de pronombres *it* y los marca como pleonásticos o referenciales identificando la regla que cumplen. Aquéllos que hayan sido marcados como pleonásticos no serán tratados en el módulo de resolución anafórica y no se generarán en el idioma destino.

7.3.1 Fase de entrenamiento

En la fase de entrenamiento se pretendía comprobar el funcionamiento de las distintas reglas planteadas en la sección 6.2.2 para la detección de los pronombres *it* pleonásticos y su adecuación al corpus utilizado. En concreto, para esta tarea se utilizó el corpus Federal Register (FR) usando 94 documentos (50.243 palabras) elegidos aleatoriamente como corpus de entrenamiento y 219 (106.588 palabras) como corpus de evaluación. Se realizaron 2 experimentos hasta obtener la precisión óptima.

Experimento 1: Aplicación de 9 reglas. En el primer experimento se utilizaron las 9 reglas presentadas en el trabajo de Paice & Husk (1987) para la detección de los pronombres *it* pleonásticos. No se utilizó el POS tagger para etiquetar previamente el texto e identificar las categorías gramaticales de las palabras, consecuentemente el algoritmo trabajaba sobre texto plano. Las reglas utilizadas se presentaron en la sección 5.3.2. Básicamente eran las siguientes:

1. *it . . . adjetivo_de_estado . . . to + verbo_infinitivo*
2. *it . . . conjunción_that*
3. *it . . . adjetivo_de_conocimiento + whether/if/what/how/why/when/where*
4. *it . . . which/who*
5. *it . . . verbo_be . . . palabra_metereológica*
6. *it . . . verbo_be . . . construcción_temporal*
7. *it . . . worth . . . verbo_gerundio*
8. *, it . . . (3 palabras o menos) . . . ,*
9. *expresiones idiomáticas*

Tras aplicar el algoritmo cada uno de los pronombres *it* del corpus será clasificado como Referencial –Tipo 1 si va precedido de una preposición (R1) y Tipo 2 si no cumple ninguna regla de pleonástico (R2)– o Pleonástico –Tipo 1 (P1), Tipo 2 (P2), etc. hasta Tipo 9 (P9) si cumple la regla 1, la regla 2, etc–.

La evaluación fue totalmente automática tras realizar el etiquetado anafórico de los pronombres *it* del texto. Si se encontraba un pronombre *it* clasificado como pleonástico y no existía solución para ese pronombre (es decir no es anafórico) se contabilizaba como acierto en caso contrario como fallo. Los resultados obtenidos en este experimento se muestran en la tabla 7.4.

Para esta tarea la precisión se entiende como el cociente entre el número de ocurrencias del pronombre *it* correctamente clasificadas para un tipo determinado y el número de ocurrencias del pronombre para ese tipo. Por ejemplo, para el Pleonástico Tipo 1 (P1) se han clasificado correctamente 50 ocurrencias de un total de 86 tratadas como Tipo 1. Los 36 fallos se producen cuando los pronombres han sido clasificados como pleonásticos y son referenciales.

Discusión. Tras analizar los resultados se han extraído una serie de conclusiones:

- Los tipos P1, P3 y P4 tienen una precisión relativamente baja (58,14%, 60% y 40% respectivamente) y debe ser estudiada. Para el tipo P1 se observa que la mayoría de los fallos se producen por la detección incorrecta de la partícula *to* que debe

Corpus	Tipo pronombre *it*	Aciertos	Total	Precisión (%)
Federal Register Corpus	P1	50	86	58,14
	P2	35	41	85,37
	P3	9	15	60
	P4	2	5	40
	P5	0	0	0
	P6	0	0	0
	P7	0	0	0
	P8	4	5	80
	P9	0	0	0
	R1	63	68	92,65
	R2	108	110	98,18
TOTAL		**271**	**330**	**82,12**

Tabla 7.4. Detección de los pronombres *it* pleonásticos. Fase de entrenamiento: experimento 1

introducir un verbo en infinitivo. El resto de fallos del tipo P1 y los fallos de los tipos P3 y P4 ocurren en oraciones que tienen varias cláusulas en las que se identifica erróneamente un patrón que abarca varias cláusulas.

- Tal y como mencionan Paice & Husk se pueden mejorar los resultados con el uso de un POS tagger. Por ejemplo, los resultados obtenidos para el tipo P1 se podrían mejorar con las palabras etiquetadas (léxica y morfológicamente) ya que permitiría detectar correctamente la partícula *to* que introduce un verbo en infinitivo.
- En el corpus de entrenamiento no hay ninguna ocurrencia de los tipos P5, P6, P7 y P9. Se podría pensar en la eliminación de estas reglas para trabajar con este corpus concreto pero como el objetivo consiste en plantear un sistema general que permita la correcta identificación de los pronombres pleonásticos en cualquier corpus se decide conservarlas.

Tras analizar los resultados del experimento 1 se observa que se obtiene una precisión global para la correcta detección de los pronombres *it* como pleonásticos o referenciales de un 82,12%. Esta precisión es muy inferior a la obtenida por Paice & Husk

(92,2%) y se decide realizar el segundo experimento introduciendo una serie de cambios con el objetivo de mejorar la precisión global.

Experimento 2: Uso del POS tagger. La principal innovación del segundo experimento y una diferencia relevante respecto al trabajo original de Paice & Husk consiste en la utilización de un POS tagger. Para ello, se redefinen las reglas de detección de pleonásticos utilizando la información del etiquetado léxico y morfológico. Por ejemplo, para la regla 1 se especifica que la partícula *to* tiene que ir seguida de una palabra cuya etiqueta sea un verbo en infinitivo. Las reglas que se modifican sustancialmente son la regla 1 (verbo en infinitivo), la regla 2 (conjunción *that*) y la regla 7 (verbo en gerundio).

Tras estas modificaciones se obtienen los resultados de la tabla 7.5.

Corpus	Tipo pronombre *it*	Aciertos	Total	Precisión (%)
Federal Register Corpus	P1	84	86	97,67
	P2	40	41	97,56
	P3	9	15	60
	P4	2	5	40
	P5	0	0	0
	P6	0	0	0
	P7	0	0	0
	P8	4	5	80
	P9	0	0	0
	R1	63	68	92,65
	R2	108	110	98,18
TOTAL		**310**	**330**	**93,94**

Tabla 7.5. Detección de los pronombres *it* pleonásticos. Fase de entrenamiento: experimento 2

Discusión. Los resultados obtenidos en este experimento son sustancialmente mejores que los obtenidos en el experimento anterior. Se pueden observar las mejoras obtenidas en la identificación de los tipos P1 y P2 con precisiones de 97,67% y 97,56% respectivamente. Finalmente, se obtiene una precisión global de 93,94%. Con

este experimento se da por concluida la fase de entrenamiento. A continuación se realiza la fase de evaluación con la configuración obtenida.

7.3.2 Fase de evaluación

Para la fase de evaluación se utilizaron los 219 documentos restantes del corpus Federal Register que no fueron usados en la fase de entrenamiento.

Los resultados obtenidos se muestran en la tabla 7.6.

Corpus	Tipo pronombre *it*	Aciertos	Total	Precisión (%)
Federal Register Corpus	P1	105	146	71,92
	P2	59	68	86,76
	P3	26	35	74,29
	P4	7	7	100
	P5	0	0	0
	P6	0	0	0
	P7	1	1	100
	P8	0	0	0
	P9	0	0	0
	R1	138	148	93,24
	R2	232	235	98,72
TOTAL		**568**	**640**	**88,75**

Tabla 7.6. Detección de los pronombres *it* pleonásticos. Fase de evaluación

Discusión. Como se observa en la tabla se ha obtenido una precisión global para el sistema de 88,75% (568/640). Los fallos en la identificación de los pronombres pleonásticos (tipos P1...P9) se producen principalmente en oraciones muy largas (con varias cláusulas) en las que se identifica erróneamente un patrón que abarca varias cláusulas (87,5% de los fallos). El resto de fallos son originados por algunas excepciones no tenidas en cuenta (9,7%) y por errores en el etiquetado léxico-morfológico de las palabras (2,8%).

Como conclusión se puede considerar que este resultado es realmente bueno. Si queremos comparar nuestro resultado con los obtenidos por otros autores, hay que tener en cuenta siempre el tipo de corpus sobre el que se ha realizado la evaluación y las condiciones en que ésta se ha realizado.

Los trabajos de Paice & Husk (1987) y Evans (2000) realizan una identificación de los pronombres *it* pleonásticos y referenciales mediante la definición de una serie de patrones. Al igual que en nuestra propuesta, para los patrones definen una serie de propiedades: pueden incluir palabras intermedias que no se han especificado previamente (se expresan con puntos suspensivos en la definición de las reglas), tienen una serie de signos de puntuación permitidos y, por último, se establece que las ocurrencias de pronombres *it* precedidas por una preposición son referenciales. En sus aproximaciones Paice & Husk y Evans obtuvieron precisiones ligeramente superiores a la nuestra, 92,2% y 93,9% respectivamente. Esto se justifica principalmente por el tipo de corpus que ellos han utilizado, ya que éstos son textos técnicos donde las construcciones gramaticales son más estrictas. En nuestro caso, se ha utilizado textos de dominio no restringido de estilo narrativo con oraciones muy largas (con una media de 24,7 palabras por oración) que contienen noticias internacionales escritas por diferentes autores y que tienen un mayor grado de libertad gramatical.

Otros investigadores han propuesto métodos basados en el reconocimiento de patrones como Lappin & Leass (1994) y Denber (1998). La diferencia fundamental entre estos métodos y el nuestro es que ellos no realizan una definición formal de las propiedades que deben cumplir los patrones de modo que éstos están formados por una serie de expresiones muy concretas formadas por un número fijo de palabras. El método de Lappin & Leass no fue formalmente evaluado mientras que la propuesta de Denber no se ha implementado.

Por último, destacar el porcentaje elevado de pronombres *it* pleonásticos –32,97% (211/640)– existentes en el corpus de evaluación. Este porcentaje refuerza la importancia que tiene la correcta detección de estos pronombres para que sean tratados convenientemente en un sistema de Traducción Automática.

7.4 Detección de los cero pronombres

Como ya se ha explicado en esta Tesis (sección 6.2.3), en el sistema AGIR se realiza la detección de los cero pronombres con función de sujeto en español. Tras la detección del cero pronombre el sistema computacional inserta el pronombre en la posición en la cual se había omitido. Este pronombre se resolverá en el módulo posterior de resolución anafórica. Tras su resolución, se procederá a la generación de este pronombre en el idioma destino.

En esta sección presentaremos la evaluación de la fase de detección de los cero pronombres. En una sección posterior se mostrará la evaluación de la fase de resolución de estos pronombres.

7.4.1 Fase de entrenamiento

El objetivo de la fase de entrenamiento consistía en comprobar el funcionamiento del algoritmo para la detección de los cero pronombres que se ha explicado previamente en la sección 6.2.3. Básicamente éste se fundamenta en la aplicación de la heurística H_3:

H_3 Después de que la oración haya sido dividida en cláusulas, se busca un sintagma nominal o un pronombre para cada cláusula entre todos los constituyentes situados a la izquierda del verbo (siempre que éste no sea imperativo o impersonal). Si se encuentra, éste debe concordar en número y persona con el verbo de la cláusula.

En la evaluación de esta tarea se han utilizado dos corpus: el corpus Blue Book (BB) en español y el corpus LEXESP. En concreto, se ha utilizado un fragmento de Blue Book (4.723 palabras) elegido aleatoriamente y los 12 primeros textos de LEXESP (11.574 palabras) como corpus de entrenamiento. El resto (10.848 palabras de Blue Book y los 19 textos de LEXESP restantes – 27.425 palabras–) se ha utilizado como corpus de evaluación. En la fase de entrenamiento se realizó un experimento que sirvió para realizar refinamientos en la gramática que permitieron obtener los resultados presentados.

Experimento 1: Aplicación de la heurística H_3. En este experimento se comprobó el algoritmo para la detección de los cero pronombres aplicando la heurística H_3.

En primer lugar, el algoritmo debe detectar todos los verbos del texto. Esta fase es relativamente sencilla ya que los corpus se han etiquetado previamente con el POS tagger por lo que todos los verbos son detectados (asumiendo que el etiquetado es correcto). En segundo lugar, el algoritmo debe clasificar los verbos en dos categorías: (a) verbos cuyos sujetos se han omitido, y (b) verbos cuyos sujetos no se han omitido.

La evaluación de esta tarea fue totalmente automática tras realizar el etiquetado anafórico de los cero pronombres del texto. Tal y como se presentó en la figura 7.1, los antecedentes de los cero pronombres (*pronOmitido*) se asocian al verbo de la oración cuyo sujeto está omitido. Si el sistema encuentra un verbo con sujeto omitido y existe solución para ese verbo se contabiliza como acierto, en caso contrario como fallo. Los resultados obtenidos en este experimento para los corpus LEXESP y BB se muestran en las tablas 7.7 y 7.8 respectivamente.

Cada una de las tablas aparece dividida en dos partes: (a) *verbos con sujeto omitido*, y (b) *verbos con sujeto no omitido*. Para cada una de estas partes aparece el número de pronombres de primera, segunda y tercera persona (1^a *p.*, 2^a *p.*, 3^a *p.*) que hay en el corpus junto con su precisión (P). En este caso la precisión se entiende como el cociente entre el número de ocurrencias de verbos correctamente detectadas para una categoría determinada y el número de ocurrencias de verbos para esa categoría. Por ejemplo, en el texto 1 del corpus LEXESP hay 14 verbos de primera persona con su sujeto omitido y la precisión obtenida ha sido del 100%, es decir, se han detectado todos correctamente.

Discusión. De los resultados se pueden extraer las siguientes conclusiones:

- Los resultados obtenidos con este experimento son los siguientes: en el corpus LEXESP se ha obtenido una precisión de 95,18% y 85,68% para los verbos con sujeto omitido y no omitido respectivamente; en el corpus BB las precisiones obtenidas

	Verbos con sujeto omitido						Verbos con sujeto no omitido					
	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)
txt 1	14	100	2	100	38	92,11	1	100	0	0	29	89,66
txt 2	4	100	8	100	65	95,38	1	0	0	0	27	100
txt 3	9	100	0	0	31	90,32	0	0	0	0	29	79,31
txt 4	11	100	3	100	37	100	1	0	0	0	34	91,18
txt 5	17	94,12	5	100	67	97,01	10	100	1	100	14	92,86
txt 6	7	100	2	100	32	87,50	3	100	3	100	34	94,12
txt 7	22	90,91	0	0	45	100	8	87,50	0	0	20	80
txt 8	8	87,50	11	100	22	90,91	1	0	0	0	28	89,29
txt 9	13	100	8	100	18	72,22	2	0	1	100	29	58,62
txt 10	3	100	1	100	36	91,67	0	0	0	0	28	71,43
txt 11	1	100	0	0	41	100	0	0	0	0	30	86,67
txt 12	9	100	2	100	30	96,67	0	0	0	0	36	97,22
TOTAL	**118**	**96,61**	**42**	**100**	**462**	**94,37**	**27**	**77,78**	**5**	**100**	**338**	**86,09**
	PRECISIÓN GLOBAL = 95,18%						**PRECISIÓN GLOBAL = 85,68%**					

Tabla 7.7. Corpus LEXESP: Detección de los cero pronombres. Fase de entrenamiento: experimento 1

	Verbos con sujeto omitido						Verbos con sujeto no omitido					
	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)
txt 1	0	0	0	0	59	98,31	0	0	0	0	162	83,33
TOTAL	**0**	**0**	**0**	**0**	**59**	**98,31**	**0**	**0**	**0**	**0**	**162**	**83,33**
	PRECISIÓN GLOBAL = 98,31%						**PRECISIÓN GLOBAL = 83,33%**					

Tabla 7.8. Corpus Blue Book: Detección de los cero pronombres. Fase de entrenamiento: experimento 1

han sido de 98,31% y 83,33% respectivamente. Con estos resultados no hay diferencias significativas entre ambos corpus.

- El corpus BB no tiene ningún verbo en primera ni en segunda persona. Se puede explicar por el estilo usado en este manual técnico que normalmente consta de una serie de definiciones aisladas distribuidas en distintos párrafos y que no están relacionadas entre sí.

- La precisión para detectar los verbos con sujeto no omitido es peor que la relativa a los verbos con sujeto omitido (entre un 10% y un 15% dependiendo del corpus). Se puede justificar por varias razones:
  - El POS tagger no identifica los verbos impersonales. Este problema se ha resuelto parcialmente con el uso de heurísticas para identificar verbos impersonales (por ejemplo, *llover*) pero que han fallado en ocasiones con el uso impersonal de algunos verbos (por ejemplo, el verbo *ser* normalmente no es impersonal aunque en la oración *Es hora de desayunar* tiene un uso impersonal).
  - La inevitable incompletitud de la gramática utilizada y los distintos problemas de ambigüedad que tienen que ser resueltos en el módulo de análisis afectan al proceso de división de cláusulas y, por lo tanto, al proceso de la detección de los posibles sujetos situados a la izquierda del verbo. Por esta razón, la fase de entrenamiento se utilizó para realizar mejoras y refinamientos en la gramática.

Los resultados de este experimento son satisfactorios y se decide pasar a la fase de evaluación con la configuración obtenida.

7.4.2 Fase de evaluación

En esta fase se utilizaron los fragmentos restantes del corpus LEXESP y BB no usados en la fase de entrenamiento. Los resultados obtenidos para cada corpus se muestran en las tablas 7.9 y 7.10 respectivamente.

La evaluación total para los dos corpus aparece en la tabla 7.11

Discusión. Como se observa en la tabla resumen para los dos corpus (tabla 7.11), se han obtenido los siguientes resultados: en el corpus LEXESP una precisión de 98,24% y 80,08% para los verbos con sujeto omitido y no omitido respectivamente; en el corpus BB las precisiones obtenidas han sido de 97,52% y 82,05% respectivamente. La precisión global del sistema para esta tarea ha sido de 89,20% (2394/2684).

Los resultados confirman las conclusiones extraídas del entrenamiento:

	Verbos con sujeto omitido						Verbos con sujeto no omitido					
	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)
txt 13	6	83,33	1	100	36	97,72	1	100	0	0	53	83,02
txt 14	14	100	0	0	31	96,77	0	0	0	0	43	72,09
txt 15	18	100	0	0	29	100	0	0	0	0	30	80
txt 16	2	100	0	0	25	96,00	0	0	0	0	27	66,67
txt 17	4	75	0	0	24	100	0	0	0	0	33	84,85
txt 18	7	100	0	0	27	96,30	0	0	0	0	35	91,43
txt 19	2	100	1	100	19	100	0	0	0	0	24	62,50
txt 20	4	50	0	0	39	100	0	0	0	0	43	79,07
txt 21	6	100	0	0	14	100	0	0	0	0	12	91,67
txt 22	4	100	0	0	21	100	0	0	3	100	40	87,50
txt 23	11	100	0	0	54	100	1	0	1	0	15	93,33
txt 24	17	100	2	100	63	95,24	1	0	0	0	63	87,30
txt 25	13	100	1	100	125	100	0	0	0	0	93	83,87
txt 26	1	100	0	0	178	99,44	0	0	0	0	164	82,32
txt 27	8	100	0	0	42	97,62	0	0	5	100	42	80,95
txt 28	5	100	6	85,71	38	100	1	0	3	100	30	83,33
txt 29	4	100	0	0	107	99,07	1	100	2	100	103	74,76
txt 30	3	100	6	100	95	96,84	0	0	0	0	85	71,76
txt 31	16	87,50	1	100	62	98,39	2	50	0	0	63	77,78
TOTAL	**145**	**95,86**	**18**	**94,74**	**1029**	**98,64**	**7**	**42,86**	**14**	**92,86**	**998**	**80,16**
	PRECISIÓN GLOBAL = 98,24%						**PRECISIÓN GLOBAL = 80,08%**					

Tabla 7.9. Corpus LEXESP: Detección de los cero pronombres. Fase de evaluación

	Verbos con sujeto omitido						Verbos con sujeto no omitido					
	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)	1ª p.	P (%)	2ª p.	P (%)	3ª p.	P (%)
txt 2	0	0	0	0	121	97,52	0	0	0	0	351	82,05
TOTAL	**0**	**0**	**0**	**0**	**121**	**97,52**	**0**	**0**	**0**	**0**	**351**	**82,05**
	PRECISIÓN GLOBAL = 97,52%						**PRECISIÓN GLOBAL = 82,05%**					

Tabla 7.10. Corpus Blue Book: Detección de los cero pronombres. Fase de evaluación

	Verbos con sujeto omitido Precisión (%)	Verbos con sujeto no omitido Precisión (%)
LEXESP	98,24% (1172/1193)	80,08% (816/1019)
BB	97,52% (118/121)	82,05% (288/351)
	PRECISIÓN GLOBAL = 98,17%	**PRECISIÓN GLOBAL = 80,58%**
PRECISIÓN SISTEMA = 89,20% (2394/2684)		

Tabla 7.11. Detección de los cero pronombres: resultados globales de la evaluación

- En el corpus BB no ha aparecido ningún verbo en primera ni en segunda persona.
- La precisión para detectar los verbos con sujeto no omitido ha sido peor (aproximadamente un 18%) que la relativa a los verbos con sujeto omitido. Esto es provocado principalmente por la incorrecta detección de verbos impersonales y por la incorrecta división de las cláusulas.
  Un ejemplo de un error provocado por una incorrecta división de las cláusulas se muestra en la siguiente oración extraída del corpus LEXESP *La humedad de los muros roía sus propios huesos y su respiración se vaciaba en el gélido tesón de la piedra.* En este caso la conjunción *y* coordina las dos cláusulas de la oración cuyos sujetos no están omitidos. Sin embargo, el sistema analiza como un único sintagma nominal coordinado la expresión *sus propios huesos y su respiración*, por lo que clasifica el verbo de la segunda cláusula como *verbo con su sujeto omitido.*

No hemos comparado nuestros resultados con los obtenidos por otros autores, ya que (como hemos explicado a lo largo de esta Tesis) nuestro trabajo es el primer estudio que se ha hecho específicamente para textos en español y el diseño de la tarea de detección de los cero pronombres depende principalmente de la estructura del idioma en cuestión. Cualquier comparación que se hubiera podido hacer con otras idiomas hubiera sido realmente irrelevante.

Por último, destacar la importancia de este fenómeno en español. En concreto, en los dos corpus estudiados el 48,96% (1314/2684) de los verbos tienen su sujeto omitido. Incluso, pode-

mos concluir que este fenómeno es más habitual en textos narrativos, el 53,93% (1193/2212) en el LEXESP, que en textos técnicos, el 25,47% (121/475) en el BB. Estos porcentajes justifican la importancia de realizar una correcta detección de estos pronombres en un sistema de Traducción Automática para que sean generados correctamente en el idioma destino.

7.5 Resolución de la anáfora pronominal

En la etapa de resolución de problemas lingüísticos del sistema AGIR nos hemos centrado en la resolución de la anáfora pronominal originada por los pronombres personales de tercera persona y los cero pronombres. En esta sección presentaremos la evaluación de la anáfora pronominal originada por pronombres personales. La evaluación de la resolución de los cero pronombres se tratará en la siguiente sección.

El algoritmo utilizado para la resolución de la anáfora pronominal está basado en un sistema de restricciones y preferencias y se ha presentado previamente en la sección 6.1.4.

Según comenta Martínez-Barco (2001), un sistema para la resolución de la anáfora basado en restricciones y preferencias requiere una definición del sistema que permita obtener la mejor configuración posible para sus metas. Esta configuración incluye tres objetivos diferentes: (a) la definición de un *espacio de accesibilidad anafórica* óptimo[3], (b) la definición de un conjunto de restricciones válido, y (c) la definición de un conjunto de preferencias junto con una valoración de la importancia de cada una de ellas.

En AGIR, la definición de un espacio de accesibilidad anafórico óptimo y de un conjunto de restricciones válido se ha efectuado tras un estudio de las características propias de cada idioma. Sin embargo, la definición de un conjunto de preferencias adecuado así como la importancia de cada una de ellas necesita un proceso de entrenamiento en el que se van ajustando los distintos

[3] El *espacio de accesibilidad anafórica* se puede definir como el espacio –normalmente expresado en oraciones– donde hay que buscar para encontrar el antecedente de una anáfora determinada.

parámetros que regulan dicha importancia hasta obtener la configuración óptima.

A continuación presentaremos de un modo separado la evaluación de este algoritmo para español e inglés ya que el sistema AGIR trabaja con ambos idiomas.

7.5.1 Resolución de la anáfora pronominal en inglés

El algoritmo planteado para la resolución de la anáfora pronominal en inglés está basado en el algoritmo correspondiente para español. Utiliza un sistema de restricciones y preferencias que ha sido adaptado convenientemente para el inglés. El objetivo de las restricciones es descartar candidatos mientras que las preferencias se utilizan para ordenar los candidatos restantes.

Fase de entrenamiento. El objetivo de la fase de entrenamiento consiste en definir el espacio de accesibilidad anafórica óptimo y un conjunto de restricciones y preferencias válido.

Para la evaluación de esta tarea se utilizaron los corpus SEMCOR y los Manuales Técnicos de Informática (MTI). Como corpus de entrenamiento se eligieron aleatoriamente 3 textos de SEMCOR (6.473 palabras) y 2 textos de MTI (24.264 palabras). El resto, 8 textos de SEMCOR (17.315 palabras) y 5 textos de MTI (77.479 palabras), se utilizó como corpus de evaluación.

Se realizaron dos experimentos hasta obtener la configuración óptima.

Experimento 1: Uso de información léxica, morfológica y sintáctica. En este primer experimento se utilizó como base el algoritmo para español que utilizaba exclusivamente información léxica, morfológica y sintáctica[4] y fue adaptado para inglés.

El sistema completo para este experimento se define a continuación:

[4] Para textos de dominio restringido también utilizaba información semántica que se añadía a las unidades léxicas del texto. La compatibilidad semántica entre dos unidades léxicas se comprobaba por medio de la integración del módulo IRSAS (Moreno *et al.*, 1991) en el sistema.

Espacio de accesibilidad anafórica. El espacio considerado para encontrar el antecedente de una anáfora pronominal está basado en una ventana de accesibilidad fija según un número de oraciones. En este caso se ha considerado el uso de una ventana de 4 oraciones más la actual. Por lo tanto, se tomarán como candidatos todos los sintagmas nominales encontrados en este espacio. Esta determinación se ha tomado tras realizar un profundo estudio del comportamiento de las anáforas pronominales en inglés en distintos tipos de textos.

La lista obtenida con todos los candidatos posibles se ordena por proximidad a la anáfora y sobre ella se aplica el conjunto de restricciones y preferencias.

Restricciones.

- *Restricciones morfológicas*: concordancia de género, número y persona entre la anáfora y su antecedente.
- *Restricciones sintácticas*: restricciones *c-dominio* propuestas por Reinhart (1983) y condiciones de no correferencia (Lappin & Leass, 1994) adaptadas para análisis sintáctico parcial. Estas restricciones *c-dominio* se utilizan normalmente para eliminar antecedentes que no van a correferir con el pronombre en cuestión (condiciones de no correferencia).

Preferencias. Las preferencias utilizadas en este experimento para todos los tipos de pronombres personales aparecen a continuación:

1. Candidatos en la misma oración que la anáfora.
2. Candidatos de la oración anterior a la anáfora.
3. Candidatos que no estén en complementos circunstanciales.
4. Candidatos que han aparecido con el mismo verbo de la anáfora más de una vez.
5. Candidatos que tienen la misma posición respecto al verbo de la anáfora (antes o después).
6. Candidatos que son SN propios.
7. Candidatos que se repiten más en el texto.
8. Candidatos que aparecen con mayor frecuencia con el mismo verbo de la anáfora.
9. El candidato más cercano a la anáfora.

Con esta ordenación de preferencias (ordenadas de mayor a menor importancia) se realizó una prueba inicial obteniendo los resultados que aparecen en la tabla 7.12. La evaluación fue totalmente automática tras realizar el etiquetado anafórico del texto.

Corpus		He	She	It	They	Him	Her	Them	Ac.	Total	P (%)
SEMCOR	a01	7	0	12	3	1	0	2	20	25	80
	a12	41	0	6	7	4	0	2	40	60	66,67
	d01	5	1	25	7	0	0	5	26	43	60,47
	TOTAL SEMCOR	**53**	**1**	**43**	**17**	**5**	**0**	**9**	**86**	**128**	**67,19**
MTI	BEOWULF	3	0	29	9	3	0	3	34	47	72,34
	MAC	0	0	100	3	0	0	10	85	113	75,22
	TOTAL MTI	**3**	**0**	**129**	**12**	**3**	**0**	**13**	**119**	**160**	**74,38**
	TOTAL	**56**	**1**	**172**	**29**	**8**	**0**	**22**	**205**	**288**	**71,18**

Tabla 7.12. Resolución de la anáfora pronominal en inglés. Ordenación inicial de las preferencias

En la tabla aparecen el número de ocurrencias de los pronombres (separados por tipos) para cada documento. Las 3 últimas columnas representan para cada documento el número de aciertos, el total de pronombres y la precisión obtenida respectivamente. En este caso, la precisión es el cociente entre el número de anáforas correctamente resueltas y el número total de anáforas tratadas. Por ejemplo, para el documento a01 del corpus SEMCOR se ha obtenido una precisión de 80% (20/25).

Con el objetivo de mejorar los resultados obtenidos se realizaron distintas pruebas cambiando el orden de aplicación de las preferencias hasta obtener la configuración óptima. Ésta se obtuvo tras realizar los siguientes cambios:

- La preferencia número 3 (preferencia por los candidatos que no estén en complementos circunstanciales) se intercambió con la preferencia número 6 (preferencia por los SN propios) para todos los pronombres personales excepto para el pronombre *it*.

- Para el pronombre *it* la preferencia número 3 se cambió de lugar colocándose tras la preferencia número 6.

Los resultados finales obtenidos en el experimento 1 con esta ordenación de preferencias aparecen en la tabla 7.13.

Corpus		Aciertos	Total	P (%)
SEMCOR	a01	20	25	80
	a12	42	60	70
	d01	30	43	69,77
	TOTAL SEMCOR	**92**	**128**	**71,88**
MTI	BEOWULF	38	47	80,85
	MAC	92	113	81,42
	TOTAL MTI	**130**	**160**	**81,25**
	TOTAL	**222**	**288**	**77,08**

Tabla 7.13. Resolución de la anáfora pronominal en inglés. Fase de entrenamiento: experimento 1

Discusión. Tras analizar los resultados obtenidos con este experimento se extraen las siguientes conclusiones:

- Los tipos de pronombres varían notablemente según el corpus que se estudie. Así, en el corpus SEMCOR un 33,59% de los pronombres son ocurrencias del pronombre *it* mientras que el 66,41% son ocurrencias del resto de pronombres personales. Sin embargo, en el corpus MTI el 80,63% de los pronombres son *it* y el 19,37% son ocurrencias del resto de pronombres. Esto viene justificado principalmente por el tipo de dominio de cada corpus. SEMCOR es un corpus de estilo narrativo formado por documentos en los que aparece una gran cantidad de entidades del tipo *persona*[5] que son referenciadas en el texto mediante pronombres. Por otra parte, MTI es un corpus formado por

[5] Si se utiliza una jerarquía de rasgos semánticos muy básica, en un primer nivel las palabras se pueden clasificar como pertenecientes a 3 tipos: *persona*, *animal* u *objeto*.

documentos técnicos donde prácticamente no aparece ninguna *persona* y se hacen referencias a entidades del tipo *objeto*.

- En el corpus SEMCOR hay una serie de fallos originados por la falta de información semántica (33,33%). Así, hay ciertas ocurrencias del pronombre *it* (concretamente 5) para las cuales el sistema devuelve un antecedente de tipo *persona*. Por otra parte, 7 ocurrencias de pronombres *he/she/him/her* se resuelven proporcionando un antecedente que no es del tipo *persona*.
  Otros fallos en este corpus (61,11%) se originan por excepciones en la aplicación del conjunto de preferencias debido a la existencia de muchos candidatos compatibles con la anáfora[6]. Por ejemplo, para los pronombres *he/she/him/her* existen, en ocasiones, muchos candidatos del tipo *persona* y el sistema escoge como solución uno incorrecto. Del mismo modo, se producen fallos en la resolución de los pronombres *they/them* por la existencia de varios candidatos con las mismas características (por ejemplo, equipos de fútbol: Dodgers, Pirates, Bears, etc.).
  El resto de fallos (5,56%) se originan por errores en el etiquetado léxico-morfológico de las palabras.
- En el corpus MTI los fallos se producen principalmente en la resolución de los pronombres *it* (73,33% de los fallos). Los fallos en la resolución de este tipo de pronombres se originan por la ausencia de género gramatical del mismo (es válido para masculino y femenino)[7] lo que provoca un elevado número de candidatos entre los que hay que decidir para escoger el correcto[8]. Esta circunstancia hace que se produzcan errores en la aplicación de las preferencias. El resto de fallos se producen por falta de información semántica.

[6] Tal y como se mostró en la tabla 7.2, las oraciones del SEMCOR son muy largas (con una media de 24,3 palabras por oración) lo que implica el elevado número de antecedentes. En particular, este corpus tiene una media de 15,2 candidatos para cada anáfora tras aplicar las restricciones.

[7] Aunque generalmente el etiquetador no proporciona información de género de los sustantivos, en ocasiones está disponible: por ejemplo la palabra *man* tiene género masculino y la palabra *woman* femenino.

[8] En este corpus las oraciones no son muy largas y tienen una media de 15,5 palabras por oración (tabla 7.3). Sin embargo, el número medio de candidatos posibles para una anáfora determinada tras aplicar las restricciones es elevado (13).

Tras analizar los resultados del experimento 1, se puede observar que se ha obtenido una precisión de 71,88% y 81,25% para los corpus SEMCOR y MTI respectivamente. El porcentaje inferior (aproximadamente en un 10%) obtenido con SEMCOR es originado fundamentalmente por la falta de información semántica. En global, para los dos corpus se obtiene una precisión de 77,08%.

Con el objetivo de mejorar estos resultados se decide realizar el experimento 2 que supone la inclusión de información semántica.

Experimento 2: Adición de información semántica. Para realizar este experimento se parte del trabajo realizado con otros compañeros del Grupo de Investigación GPLSI (M. Saiz-Noeda y A. Suárez) en el que se presenta una serie de técnicas de compatibilidad semántica para aplicarlas en el módulo de resolución de la anáfora (Saiz-Noeda *et al.*, 2000; Saiz-Noeda *et al.*, 1999).

En esta Tesis, se han utilizado las herramientas proporcionadas por Saiz-Noeda que permiten acceder desde un programa PROLOG al recurso léxico WordNet 1.6 y obtener para una palabra determinada (etiquetada semánticamente con su sentido en WordNet) su *concepto ontológico principal* (*Top concept*).

Los cambios introducidos en el segundo experimento son los siguientes:

- A las restricciones morfológicas y sintácticas del primer experimento se le añaden 2 *restricciones semánticas* (Saiz-Noeda *et al.*, 2000):
  1. Los pronombres *he/she/him/her* deben tener antecedentes cuyo tipo semántico sea *persona.*
  2. Los pronombres *it* deben tener antecedentes con un tipo semántico que no sea *persona.*

  Para ello se realiza un agrupamiento de los 25 *Top concepts* existentes en el WordNet para los nombres y se reducen a 3 tipos semánticos: *persona*, *animal* y *objeto*. Con el núcleo de cada uno de los candidatos de una anáfora determinada se accede a WordNet, se determina su tipo semántico y se aplican las restricciones semánticas anteriores.

- Sólo se realiza este experimento con el corpus SEMCOR ya que es el único que está etiquetado semánticamente con el sentido de las palabras en WordNet.

Tras estos cambios en el sistema se obtienen los resultados de la tabla 7.14.

Corpus		Aciertos	Total	P (%)
SEMCOR	a01	21	25	84
	a12	50	60	83,33
	d01	33	43	76,74
	TOTAL	**104**	**128**	**81,25**

Tabla 7.14. Resolución de la anáfora pronominal en inglés. Fase de entrenamiento: experimento 2

Discusión. Los resultados obtenidos con este experimento mejoran aproximadamente un 10% los resultados anteriores para el corpus SEMCOR. Así, se obtiene una precisión de 81,25%. Con la configuración obtenida se decide dar paso a la fase de evaluación.

Fase de evaluación. Para la fase de evaluación se utilizaron el resto de documentos del corpus SEMCOR y MTI que no se usaron en la fase de entrenamiento. En concreto, para el corpus SEMCOR se utilizó la configuración obtenida en el experimento 2 ya que se disponía de información semántica; para el corpus MTI se utilizó la configuración del experimento 1.

Los resultados obtenidos se muestran por separado para el corpus SEMCOR (tabla 7.15) y MTI (tabla 7.16) ya que se aplican distintas restricciones.

Discusión. Como se puede observar en las tablas que presentan la evaluación del sistema para cada corpus se han obtenido unas precisiones de 86,61% (220/254) y 76,81% (361/470) para los corpus SEMCOR y MTI respectivamente. Para calcular la cobertura

Corpus		He	She	It	They	Him	Her	Them	Ac.	Total	P (%)
SEMCOR	a02	13	0	10	2	0	0	2	25	27	92,59
	a11	6	2	2	2	3	0	0	14	15	93,33
	a13	11	2	0	4	5	0	0	18	22	81,82
	a14	36	0	6	1	7	0	1	49	51	96,08
	a15	17	1	0	14	7	0	2	33	41	80,49
	d02	2	0	8	8	1	0	0	18	19	94,74
	d03	1	3	3	6	0	0	1	11	14	78,57
	d04	30	2	9	13	11	0	0	52	65	80
	TOTAL	**116**	**10**	**38**	**50**	**34**	**0**	**6**	**220**	**254**	**86,61**

Tabla 7.15. Corpus SEMCOR: Resolución de la anáfora pronominal en inglés. Fase de evaluación

Corpus		He	She	It	They	Him	Her	Them	Ac.	Total	P (%)
MTI	CDROM	1	0	44	10	0	0	14	49	69	71,01
	PSW	0	0	56	4	0	0	2	53	62	85,48
	WINDOWS	0	0	31	2	0	0	4	30	37	81,08
	SCANWORX	0	0	150	16	0	0	27	146	193	75,65
	GIMP	0	0	66	24	0	0	19	83	109	76,15
	TOTAL	**1**	**0**	**347**	**56**	**0**	**0**	**66**	**361**	**470**	**76,81**

Tabla 7.16. Corpus MTI: Resolución de la anáfora pronominal en inglés. Fase de evaluación

se calculó el número total de anáforas existentes, obteniendo los siguientes porcentajes 82,09% (220/268) y 72,93 % (361/495) para SEMCOR y MTI respectivamente. Tras analizar los resultados se extraen las siguientes conclusiones:

- La precisión obtenida en los manuales técnicos es inferior (aproximadamente un 10%) a la del corpus SEMCOR. Se justifica, en primer lugar, por la falta de información semántica del corpus MTI. En segundo lugar, tal y como se dedujo del entrenamiento, la falta de información de género gramatical del pronombre *it*, así como el elevado número de candidatos (una media de 13,6 candidatos por anáfora tras aplicar las restricciones) perjudican notablemente la elección del antecedente correcto.

- La inclusión de dos sencillas restricciones semánticas ha mejorado considerablemente la precisión para el corpus SEMCOR. Explotando más este tipo de información semántica, utilizando preferencias que establezcan la compatibilidad semántica entre un candidato y una anáfora[9], los resultados deben ser mejores. El mayor número de fallos en este corpus son originados, al igual que en la fase de entrenamiento, por la existencia de varios candidatos con las mismas características semánticas, de entre los cuales el sistema elige como solución uno incorrecto tras aplicar las preferencias.

Para comparar nuestros resultados con los obtenidos por otros autores hay que tener en cuenta el tipo de corpus sobre el que se ha hecho la evaluación y el modo de realizarla. Hobbs obtuvo un porcentaje de éxito de 81,8% sobre 100 frases con distintas ocurrencias de pronombres *he*, *she*, *it* y *they*. El algoritmo fue evaluado manualmente basándose en un análisis sintáctico completo y sin ambigüedades. Lappin & Leass obtuvieron un porcentaje de éxito de 85% sobre manuales de informática utilizando análisis sintáctico completo. Kennedy & Boguraev propusieron un algoritmo basado en de Lappin & Leass, obteniendo un 75% sobre textos más variados y menos formales que los anteriores. Mitkov & Stys obtuvieron una precisión de 95,8% y 92,1% sobre manuales técnicos para inglés y polaco respectivamente. Su sistema estaba basado en restricciones y una serie de preferencias muy específicas para el corpus sobre el que se realizó la evaluación. Por último, en los trabajos de Strube en los que se presenta el *centering* funcional, se obtiene una precisión de 85,4% en la resolución de la anáfora pronominal.

En general, nuestro sistema ha obtenido unos resultados muy satisfactorios (precisiones de 86,61% y 76,81% para los corpus SEMCOR y MTI respectivamente) teniendo en cuenta los tipos de corpus que hemos utilizado y que el análisis sintáctico utilizado ha sido parcial. Para realizar una comparación real, se han implementado una serie de algoritmos significativos (adaptándolos para

[9] Por ejemplo, en la línea de investigación establecida por Saiz-Noeda *et al.* (2000) en la que se determinan patrones semánticos *nombre-verbo* para determinar la compatibilidad semántica entre anáfora y antecedente.

análisis sintáctico parcial) sobre los que se ha realizado la evaluación de los corpus SEMCOR y MTI. Los resultados se muestran en la tabla 7.17 en la que aparecen las precisiones obtenidas.

	Cercanía	Hobbs	Lappin	Strube	AGIR
SEMCOR	37,01	57,09	59,45	59,45	**86,61**
MTI	54,89	65,96	75,11	63,19	**76,81**

Tabla 7.17. Resolución de la anáfora pronominal en inglés. Comparación con otros autores

En la tabla aparece, en primer lugar, la aproximación denominada *cercanía* que escoge como solución de la anáfora el candidato más cercano a la misma tras haber aplicado las restricciones. Este método se puede utilizar como base (*baseline*) para determinar las mejoras obtenidas con el resto de métodos. En las siguientes columnas de la tabla se muestran los resultados de los métodos de Hobbs, Lappin & Leass, el método de *centering* funcional de Strube y el sistema AGIR[10]. Como se observa, la precisión de AGIR es mejor que la obtenida con el resto de aproximaciones implementadas.

7.5.2 Resolución de la anáfora pronominal en español

El algoritmo para la resolución de la anáfora pronominal en español en el sistema AGIR se ha presentado en varios trabajos (Palomar *et al.*, 2001; Ferrández *et al.*, 1999a; Ferrández *et al.*, 1998a; Ferrández, 1998).

Los resultados obtenidos en la evaluación de esta tarea difieren ligeramente de los obtenidos en el último trabajo publicado en el que se presenta el algoritmo para la resolución de la anáfora en español (Palomar *et al.*, 2001). Esto se justifica principalmente por

[10] En el trabajo de Palomar *et al.* (2001) se describen con detalle el modo en el que se ha realizado la implementación de las distintas aproximaciones.

dos motivos: en primer lugar, los fragmentos del corpus LEXESP utilizados para entrenamiento y evaluación son distintos en cada trabajo; por otra parte, aquí presentamos una modificación de la aplicación del conjunto de preferencias del trabajo de Palomar *et al.* (2001).

Fase de entrenamiento. El objetivo de la fase de entrenamiento consiste en definir el espacio de accesibilidad anafórica óptimo y un conjunto de restricciones y preferencias válido.

Para la evaluación de esta tarea se utilizó el corpus LEXESP. Como corpus de entrenamiento se eligieron los 3 últimos textos del corpus (6.457 palabras), usando los 28 textos restantes (32.542 palabras) como corpus de evaluación.

Se realizó un único experimento para obtener la configuración óptima.

Experimento 1: Uso de información léxica, morfológica y sintáctica. El sistema completo para este experimento se define a continuación:

Espacio de accesibilidad anafórica. Para la anáfora pronominal reflexiva se toma como espacio de accesibilidad anafórica una ventana fija de 1 oración (la oración actual), es decir, se tomarán como candidatos todos los sintagmas nominales que aparezcan en la misma oración que la anáfora. Para el resto de anáforas pronominales se toma como espacio de accesibilidad anafórica una ventana fija de 4 oraciones más la actual.

Sobre la lista obtenida con todos los candidatos posibles (ordenada por proximidad a la anáfora) se aplica el conjunto de restricciones y preferencias.

Restricciones.

- *Restricciones morfológicas*: concordancia de género, número y persona entre la anáfora y su antecedente.
- *Restricciones sintácticas*: restricciones *c-dominio* propuestas por Reinhart (1983) y condiciones de no correferencia (Lappin & Leass, 1994) adaptadas para análisis sintáctico parcial (Palomar *et al.*, 2001).

Preferencias. Las preferencias utilizadas dependen del tipo de pronombre (Palomar *et al.*, 2001). A continuación aparecen las que se han utilizado para cada tipo:

Pronombres personales de complemento:

1. Candidatos que no son del tipo *tiempo*, *dirección*, *cantidad* o tipo *abstracto*[11].
2. Candidatos en la misma oración que la anáfora.
3. Candidatos de la oración anterior a la anáfora.
4. Candidatos que no están incluidos en otro sintagma nominal.
5. Candidatos que no están incluidos en un sintagma preposicional o están incluidos en un sintagma preposicional introducidos por la preposición *a* o *de*.
6. Candidatos que han aparecido con el verbo de la anáfora en más de una ocasión.
7. Candidatos que se repiten más en el texto.
8. Candidatos que aparecen con mayor frecuencia con el mismo verbo de la anáfora.
9. El candidato más cercano a la anáfora.

Pronombres personales que no están incluidos en un sintagma preposicional; pronombres reflexivos:

1. Candidatos que no son del tipo *tiempo*, *dirección*, *cantidad* o tipo *abstracto*.
2. Candidatos en la misma oración que la anáfora.
3. Candidatos de la oración anterior a la anáfora.
4. Candidatos que no están incluidos en otro sintagma nominal.
5. Candidatos que no están incluidos en un sintagma preposicional o están incluidos en un sintagma preposicional introducidos por la preposición *a* o *de*.
6. Para el caso de pronombres personales no incluidos en un sintagma preposicional, candidatos que no estén incluidos en un sintagma preposicional con la preposición *en*.
7. Candidatos que aparecen antes del verbo de la oración en la que ellos aparecen.

[11] Esta información se extrae durante el análisis sintáctico mediante una serie de reglas particulares. Por ejemplo, un *número* seguido de la palabra *ptas.* expresa una entidad (sintagma nominal) del tipo *cantidad*.

8. Candidatos que se repiten más en el texto.
9. Candidatos que aparecen con mayor frecuencia con el mismo verbo de la anáfora.
10. El candidato más cercano a la anáfora.

Pronombres personales que están incluidos en un sintagma preposicional:

1. Candidatos que no son del tipo *tiempo*, *dirección*, *cantidad* o tipo *abstracto*.
2. Candidatos en la misma oración que la anáfora.
3. Candidatos de la oración anterior a la anáfora.
4. Candidatos que no están incluidos en otro sintagma nominal.
5. Candidatos que han aparecido con el verbo de la anáfora en más de una ocasión.
6. Candidatos que están incluidos en un sintagma preposicional.
7. Candidatos que tienen la misma posición respecto al verbo de la anáfora (antes o después).
8. Candidatos que se repiten más en el texto.
9. Candidatos que aparecen con mayor frecuencia con el mismo verbo de la anáfora.
10. El candidato más cercano a la anáfora.

Con esta ordenación de preferencias (ordenadas de mayor a menor importancia) se llevó a cabo una prueba inicial[12] obteniendo los resultados que aparecen en la tabla 7.18. La evaluación realizada fue totalmente automática.

En la tabla aparecen el número de ocurrencias de pronombres personales para cada documento. Los distintos tipos estudiados son: *Compl* (personales de complemento), *Refl* (reflexivos), *PPSuj* (personales no incluidos en un sintagma preposicional) y *PPSP* (personales incluidos en un sintagma preposicional). Junto a cada tipo aparece la precisión obtenida. Las dos últimas columnas representan el total de pronombres personales por documento y la precisión obtenida en el mismo. Por ejemplo, en el texto 29 hay

[12] Esta ordenación de preferencias es la que se estableció en el algoritmo para la resolución de la anáfora pronominal en español presentado en Palomar *et al.* (2001).

	Compl	P (%)	Refl	P (%)	PP Suj	P (%)	PP SP	P (%)	Total	P (%) Total
txt 29	19	52,63	11	100	2	50	8	75	40	70
txt 30	11	63,64	6	100	9	66,67	11	72,73	37	72,97
txt 31	7	85,71	3	100	8	62,50	1	100	19	78,95
TOTAL	**37**	**62,16**	**20**	**100**	**19**	**63,16**	**20**	**75**	**96**	**72,92**

Tabla 7.18. Corpus LEXESP: Resolución de la anáfora pronominal en español. Ordenación inicial de las preferencias

40 ocurrencias de pronombres personales y se ha obtenido una precisión de 70%.

Para mejorar los resultados obtenidos se realizaron distintas pruebas cambiando el orden de las preferencias hasta llegar a la configuración óptima. Ésta se alcanzó tras efectuar los siguientes cambios:

- Para los pronombres personales de complemento, la preferencia número 5 (candidatos no incluidos en un sintagma preposicional) se intercambió con la preferencia número 6 (candidatos que han aparecido con el verbo de la anáfora en más de una ocasión).
- Para los pronombres personales que no están incluidos en un sintagma preposicional, la preferencia número 6 (candidatos no incluidos en un sintagma preposicional con la preposición *en*) se intercambió con la preferencia número 7 (candidatos que aparecen antes del verbo).
- Para los pronombres personales que están incluidos en un sintagma preposicional, la preferencia número 6 (candidatos incluidos en un sintagma preposicional) se intercambió con la preferencia número 7 (candidatos que tienen la misma posición respecto al verbo de la anáfora).

Los resultados finales obtenidos en el experimento 1 con la configuración óptima aparecen en la tabla 7.19.

Discusión. De los resultados obtenidos se extraen las siguientes conclusiones:

- Las precisiones obtenidas para cada tipo de pronombre oscilan entre 70% y 80% excepto para las pronombres reflexivos que

	Compl	P (%)	Refl	P (%)	PP Suj	P (%)	PP SP	P (%)	Total	P (%) Total
txt 29	19	68,42	11	100	2	50	8	75	40	77,50
txt 30	11	81,82	6	100	9	77,78	11	81,82	37	83,78
txt 31	7	71,43	3	100	8	62,50	1	100	19	73,68
TOTAL	**37**	**72,97**	**20**	**100**	**19**	**68,42**	**20**	**80**	**96**	**79,17**

Tabla 7.19. Corpus LEXESP: Resolución de la anáfora pronominal en español. Fase de entrenamiento: experimento 1

alcanzan una precisión del 100%. El elevado porcentaje obtenido con estos pronombres viene justificado porque normalmente el antecedente de este tipo de pronombres es el sintagma nominal más cercano al mismo y se encuentra en su misma oración por lo que tras aplicar las preferencias no se ha producido ningún fallo.

- Para analizar los fallos producidos en el resto de pronombres hay que tener en cuenta, en primer lugar, la complejidad del corpus LEXESP en sí mismo. Son textos narrativos (en ocasiones complejos) con oraciones muy largas (con una media de 24,6 palabras por oración), lo que implica el elevado número de candidatos para cada anáfora tras aplicar las restricciones (una media de 16,6 candidatos).
  El mayor porcentaje de fallos (65%) se produce por excepciones en la aplicación del conjunto de preferencias. La falta de información semántica provoca otro elevado porcentaje de fallos[13] (30%). El resto de fallos se deben a errores producidos por el etiquetador léxico-morfológico (5%).

Tras los resultados obtenidos en este experimento en el que se alcanza una precisión de 79,17% para el fragmento del corpus LEXESP utilizado como entrenamiento se decide pasar a la fase de evaluación.

[13] Por ejemplo, al resolver la anáfora pronominal en la siguiente oración extraída del corpus LEXESP *El Salvador de la Patria y* él *cabalgaban juntos...* el sistema propone como solución de la misma el sintagma nominal *fuego del hogar*. Con información semántica, este candidato se podría descartar ya que es incompatible semánticamente con el verbo en el que aparece la anáfora (*cabalgaban*).

Fase de evaluación. En la fase de evaluación se utilizaron los 28 primeros textos del corpus LEXESP que no fueron usados en la fase de entrenamiento. Se obtuvieron los resultados que se presentan en la tabla 7.20.

	Compl	P (%)	Refl	P (%)	PP Suj	P (%)	PP SP	P (%)	Total	P (%) Total
txt 1	3	66,67	2	100	4	100	1	100	10	90
txt 2	7	85,71	5	100	3	66,67	1	100	16	87,50
txt 3	4	75	7	100	2	50	1	100	14	85,71
txt 4	4	75	4	100	2	100	1	0	11	81,82
txt 5	4	75	1	100	2	0	1	100	8	62,50
txt 6	1	100	1	100	5	80	0	0	7	85,71
txt 7	3	100	3	0	6	83,33	4	100	16	75
txt 8	0	0	2	100	2	50	0	0	4	75
txt 9	3	66,67	0	0	6	83,33	1	100	10	80
txt 10	2	50	2	100	0	0	1	100	5	80
txt 11	0	0	2	100	2	100	1	0	5	80
txt 12	4	75	4	100	5	60	0	0	13	76,92
txt 13	0	0	5	100	0	0	0	0	5	100
txt 14	0	0	6	100	2	100	0	0	8	100
txt 15	3	100	4	100	1	0	5	100	13	92,31
txt 16	1	100	0	0	0	0	2	100	3	100
txt 17	1	100	4	100	1	100	0	0	6	100
txt 18	4	75	2	100	2	50	0	0	8	75
txt 19	0	0	5	100	1	100	1	100	7	100
txt 20	3	100	0	0	1	0	1	0	5	60
txt 21	2	100	1	100	0	0	0	0	3	100
txt 22	1	0	3	100	2	50	2	100	8	75
txt 23	4	75	2	100	1	100	2	50	9	77,78
txt 24	6	83,33	2	100	2	0	2	100	12	75
txt 25	11	100	11	90,91	8	87,50	5	100	35	94,29
txt 26	16	81,25	16	75	6	50	11	36,36	49	65,31
txt 27	6	83,33	6	100	3	66,67	1	100	16	87,50
txt 28	5	80	5	100	2	100	2	100	14	92,86
TOTAL	**98**	**82,65**	**105**	**92,38**	**71**	**70,42**	**46**	**76,09**	**320**	**82,19**

Tabla 7.20. Corpus LEXESP: Resolución de la anáfora pronominal en español. Fase de evaluación

Discusión. Como se puede observar en la tabla que muestra los resultados de la evaluación, se ha obtenido una precisión de 82,19% (263/320) para el corpus LEXESP. La cobertura alcanzada fue de 78,98% (263/333). Analizando los resultados se extrajeron las siguientes conclusiones:

- Hay algunos documentos (como los textos 5, 20, 26 etc.) en los que la precisión obtenida ha sido realmente baja (aproximadamente un 60%). En estos textos las oraciones, normalmente, son muy largas y se usa un estilo de narración muy complejo donde aparecen muchas conjunciones, oraciones subordinadas, aposiciones, etc. Del mismo modo, algunos de ellos son fragmentos de novelas donde aparecen muchos personajes. Todas estas circunstancias provocan que el número de candidatos para una anáfora determinada sea muy elevado lo que influye negativamente en la elección del antecedente correcto.
- En general, los fallos producidos en esta fase son originados por las mismas causas que en la fase de entrenamiento: excepciones en la aplicación de las preferencias (66,67%), falta de información semántica (29,82%) y errores de etiquetado (3,51%).

Si queremos comparar nuestros resultados con los de otros autores, son válidas las aproximaciones presentadas en la fase de evaluación de la resolución de la anáfora pronominal en inglés de la sección 7.5.1. Hay que tener en cuenta que estas aproximaciones fueron propuestas y evaluadas por sus autores originalmente para textos en inglés.

Si nos fijamos en las diferencias entre español e inglés en cuanto a la resolución de la anáfora podemos hacer las siguientes observaciones:

- Las palabras en español tienen, en general, más información morfológica que en inglés. Como consecuencia de esto, en el proceso de resolución de la anáfora en español las restricciones morfológicas descartan más candidatos que las correspondientes en inglés.
- El español es un idioma que presenta un orden "casi" libre de palabras en el que los distintos papeles sintácticos (sujeto, objeto, etc.) pueden aparecer, prácticamente, en cualquier lugar en

una oración. Por esta razón, el paralelismo sintáctico juega un papel más importante en la resolución de la anáfora en inglés que en español.

- Las frases en español son, generalmente, de una longitud media superior a las correspondientes en inglés, por lo que existe una mayor cantidad de candidatos para los pronombres españoles que para los ingleses.

Todas estas observaciones ponen de manifiesto que el proceso de resolución de la anáfora en español es más complejo que el respectivo para inglés.

Si comparamos los resultados obtenidos en AGIR para la anáfora en español con el resto de aproximaciones para inglés, los nuestros son ligeramente inferiores que los obtenidos en otras aproximaciones. Para realizar una comparación real se implementaron las aproximaciones más significativas adaptándolas para análisis sintáctico parcial y para aplicarlas sobre textos en español[14]. Estas aproximaciones fueron evaluadas sobre el mismo fragmento de corpus LEXESP utilizado en la evaluación de AGIR. Las precisiones obtenidas se muestran en la tabla 7.21.

	Cercanía	Hobbs	Lappin	Strube	**AGIR**
LEXESP	52,48	65,35	73,27	68,32	**82,19**

Tabla 7.21. Resolución de la anáfora pronominal en español. Comparación con otros autores

Como se observa en la tabla, la precisión obtenida en AGIR supera notablemente a la alcanzada por el resto de aproximaciones.

[14] Las distintas aproximaciones implementadas se corresponden con las presentadas anteriormente en la tabla 7.17.

7.6 Resolución de los cero pronombres

Tras la detección de los cero pronombres con función de sujeto en español, AGIR inserta el pronombre en la posición en la cual se había omitido. Este pronombre se resolverá en el módulo de resolución anafórica. El algoritmo para la resolución de este tipo de pronombres es el mismo que se utiliza para la resolución de la anáfora pronominal en español ya que una vez que se ha recompuesto el pronombre, la tarea de resolución del mismo se convierte en un caso de anáfora pronominal de sujeto. La diferencia fundamental en la resolución de los cero pronombres respecto al resto de anáforas pronominales consiste en la definición de un nuevo conjunto de restricciones y preferencias.

En esta sección presentamos la evaluación de la fase de resolución de los cero pronombres.

7.6.1 Fase de entrenamiento

El objetivo de la fase de entrenamiento es la definición de un espacio de accesibilidad anafórico óptimo y un conjunto de restricciones y preferencias válido para los cero pronombres.

En la evaluación de esta tarea se utilizaron los corpus utilizados para evaluar la fase de detección de los cero pronombres, es decir, el corpus LEXESP y el corpus BB. Concretamente se seleccionaron un fragmento de BB (4.723 palabras) elegido aleatoriamente y los 3 últimos textos de LEXESP (6.457 palabras) como corpus de entrenamiento. Los restantes fragmentos de corpus (10.848 palabras de BB y 28 textos de LEXESP –32.542 palabras–) se utilizaron como corpus de evaluación. En la fase de entrenamiento se realizó un único experimento.

Experimento 1: Uso de información léxica, morfológica y sintáctica. El sistema completo para este experimento se define a continuación:

Espacio de accesibilidad anafórica. Para los cero pronombres se ha considerado como espacio de accesibilidad anafórica una ventana de 4 oraciones más la actual.

Restricciones.

- *Restricciones morfológicas*: concordancia de número y persona entre la anáfora y su antecedente. La información del género de la anáfora no estará normalmente disponible por lo que ésta se usará como preferencia y no como restricción[15]. Esta característica es una diferencia fundamental entre la resolución de las anáforas pronominales y los cero pronombres.
- *Restricciones sintácticas*: restricciones *c-dominio* propuestas por Reinhart (1983) y condiciones de no correferencia (Lappin & Leass, 1994) adaptadas para análisis sintáctico parcial.

Preferencias. El conjunto de preferencias utilizado para los cero pronombres está basado en el presentado para la resolución de la anáfora pronominal en español. Se definieron las siguientes preferencias:

1. Candidatos que no son del tipo *tiempo*, *dirección*, *cantidad* o tipo *abstracto*.
2. Candidatos en la misma oración que el cero pronombre.
3. Candidatos de la oración anterior a la anáfora.
4. Candidatos que no están incluidos en otro sintagma nominal.
5. Candidatos que no están incluidos en un sintagma preposicional o están incluidos en un sintagma preposicional introducidos por la preposición *a* o *de*.
6. Candidatos que aparecen antes del verbo de la oración en la que ellos aparecen.
7. Candidatos que se repiten más en el texto.
8. Candidatos que aparecen con mayor frecuencia con el mismo verbo de la anáfora.
9. El candidato más cercano a la anáfora.

Con esta ordenación de preferencias (ordenadas de mayor a menor importancia) se realizó una prueba inicial. Tras la evaluación automática se obtuvieron los resultados que se muestran en la tabla 7.22.

[15] Tal y como se presentó en la sección 6.2.3, la información de número y persona del cero pronombre se extraen del verbo de la cláusula donde aparece. En algunas ocasiones (verbos copulativos) la información del género se puede extraer del objeto de la oración.

Corpus		Catáfora	Exófora	Anáfora		
				Aciertos	Total	P (%)
LEXESP	txt 29	31	0	64	76	84,21
	txt 30	43	0	36	52	69,23
	txt 31	39	1	15	22	68,18
	TOTAL LEXESP	**113**	**1**	**115**	**150**	**76,67**
BB	txt 1	37	4	14	18	77,78
	TOTAL BB	**37**	**4**	**14**	**18**	77,78
	TOTAL	**150**	**5**	**129**	**168**	**76,79**

Tabla 7.22. Resolución de los cero pronombres en español. Ordenación inicial de las preferencias

En la tabla aparecen los verbos en tercera persona de los fragmentos de corpus LEXESP y BB que tienen cero pronombres y que serán tratados en AGIR[16]. Estos cero pronombres se han dividido en tres categorías:

1. Catafóricos. Son aquellos cero pronombres cuyos antecedentes (sujetos de la cláusula) vienen después del verbo. Por ejemplo, en la oración *Avisó Juan que llegaría tarde*, el sujeto de la cláusula, *Juan*, aparece después del verbo.
   Estos pronombres no se resolverán en AGIR ya que se necesita información semántica que sea capaz de descartar los candidatos de oraciones anteriores y determinar el antecedente correcto de entre todos los candidatos que aparezcan después del verbo de la cláusula.
2. Exofóricos. Son aquellos cero pronombres cuyos antecedentes no aparecen, lingüísticamente, en el texto.
   Estos pronombres no se resolverán en AGIR.
3. Anafóricos. Son los cero pronombres cuyos antecedentes se encuentran antes del verbo.
   Este tipo de pronombres serán detectados y resueltos en el sistema AGIR.

[16] Los cero pronombres que tengan verbos en primera y segunda persona no se resolverán.

En la tabla 7.22 aparecen para cada texto el número de cero pronombres catafóricos, exofóricos y anafóricos, distinguiendo para estos últimos el número de aciertos tras su resolución, el total y la precisión obtenida. Por ejemplo, para el texto 31 del corpus LEXESP hay 39 cero pronombres catafóricos, 1 exofórico y 22 anafóricos, de los cuales se han resuelto correctamente 15, obteniendo una precisión de 68,18%.

Para mejorar los resultados iniciales obtenidos se realizaron distintas pruebas hasta alcanzar la configuración óptima. El cambio introducido es el siguiente:

- Se mantuvo la ordenación inicial de las preferencias y se introdujeron dos nuevas preferencias que distinguen los cero pronombres de las anáforas pronominales:
  - Candidatos en la misma oración de la anáfora y que han sido solución de otros cero pronombres.
    Con esta nueva preferencia se pretende dar más prioridad a aquellos candidatos que han sido solución de cero pronombres en la misma oración donde se encuentra la anáfora asumiendo que si en una oración hay varios cero pronombres, normalmente, tendrán el mismo antecedente. Esta preferencia se añadió después de la preferencia número 2 (candidatos en la misma oración que el cero pronombre).
  - Si el cero pronombre tiene información de género, aquellos candidatos que concuerden en género.
    Esta preferencia establece una prioridad para los cero pronombres que aparecen en una cláusula con verbo copulativo. En este caso, se favorece a aquellos candidatos cuyo género concuerde con el género del objeto gramatical de la cláusula donde aparece el cero pronombre[17]. Esta preferencia se añadió después de la preferencia número 6 (candidatos que aparecen antes del verbo de la oración en la que ellos aparecen).

[17] Esta preferencia ocupa una de las últimas posiciones porque el etiquetador no proporciona la información relativa a la posibilidad de que el objeto pueda tener formas lingüísticas distintas para masculino y femenino. Por ejemplo, en las oraciones *Pedro es un genio* y *Ana es un genio* el objeto *genio* sólo tiene una forma lingüística para masculino, consecuentemente la información del género del objeto no se podrá usar como restricción para los cero pronombres y se usará como preferencia.

Los resultados finales obtenidos en el experimento 1 con esta configuración óptima aparecen en la tabla 7.23.

Corpus		Catáfora	Exófora	Anáfora		
				Aciertos	Total	P (%)
LEXESP	txt 29	31	0	66	76	86,84
	txt 30	43	0	37	52	71,15
	txt 31	39	1	17	22	77,27
	TOTAL LEXESP	**113**	**1**	**120**	**150**	**80**
BB	txt 1	37	4	15	18	83,33
	TOTAL BB	**37**	**4**	**15**	**18**	**83,33**
	TOTAL	**150**	**5**	**135**	**168**	**80,36**

Tabla 7.23. Resolución de los cero pronombres en español. Fase de entrenamiento: experimento 1

Discusión. Con los resultados obtenidos se han extraído las siguientes conclusiones:

- No existen diferencias significativas en la resolución de los cero pronombres entre ambos corpus obteniendo una precisión de 80% aproximadamente.
- Los fallos producidos en el corpus LEXESP, al igual que en la resolución de la anáfora pronominal, están influenciados, en primer lugar, por la complejidad del corpus. El número de candidatos para cada cero pronombre tras aplicar restricciones es muy elevado: 19. Esta cantidad es mayor que la obtenida para la resolución de la anáfora pronominal en español e inglés y se justifica, principalmente, por la inexistencia de la restricción de concordancia en género entre cero pronombre y antecedente.
  Los fallos en la resolución de los cero pronombres son originados por las mismas causas que provocaban los errores en la resolución de las anáforas pronominales: excepciones en la aplicación del conjunto de preferencias (63,34%), falta de información semántica (33,33%) y errores en el etiquetado de las palabras (3,33%).

- Los fallos en el corpus BB no son significativos (sólo hay 3 fallos). Éstos son provocados por excepciones en la aplicación de las preferencias.

Tras los resultados obtenidos en este experimento en el que se ha obtenido una precisión media de 80,36% para ambos corpus se pasa a la fase de evaluación.

7.6.2 Fase de evaluación

En esta fase se utilizaron los fragmentos de corpus LEXESP y BB no usados durante el entrenamiento. Los resultados que se obtuvieron aparecen en la tabla 7.24.

Discusión. Como se observa en la tabla que presenta los resultados de la evaluación para ambos corpus, se ha obtenido una precisión global de 81,38% (485/596) en la resolución de los cero pronombres anafóricos en tercera persona. La cobertura obtenida fue de 79,12% (485/613). Analizando los resultados se hacen las siguientes observaciones:

- Considerando todos los cero pronombres en tercera persona que se han evaluado (1.348), el 53,12% son catafóricos, el 44,21% son anafóricos y el restante 2,67% son exofóricos. Estos porcentajes ponen de manifiesto que de los verbos en español en tercera persona que tienen su sujeto omitido, aproximadamente la mitad tienen cero pronombres anafóricos y la otra mitad catafóricos. Consecuentemente, la correcta detección y resolución de este tipo de pronombres tiene una gran importancia en un sistema de Traducción Automática (sobre todo si se traduce a un idioma que no omite el sujeto pronominal, como el inglés) para que se puedan generar convenientemente en el idioma destino.
- La precisión obtenida en la resolución de los cero pronombres es inferior (aproximadamente en un 1%) que la obtenida en la resolución de las anáforas pronominales en español (82,19%). Este hecho se justifica, principalmente, por la falta de información de género en los cero pronombres que implica un mayor número de candidatos y, consiguientemente, un mayor porcentaje de fallos.

Corpus		Catáfora	Exófora	Anáfora		
				Aciertos	Total	P (%)
LEXESP	txt 1	20	3	12	15	80
	txt 2	32	1	27	32	84,38
	txt 3	9	3	10	19	52,63
	txt 4	26	0	7	11	63,64
	txt 5	55	1	8	11	72,73
	txt 6	17	3	10	12	83,33
	txt 7	29	0	15	16	93,75
	txt 8	12	2	6	8	75
	txt 9	10	5	3	3	100
	txt 10	16	3	12	17	70,59
	txt 11	36	0	4	5	80
	txt 12	9	1	16	20	80
	txt 13	19	1	12	16	75
	txt 14	22	0	7	9	77,78
	txt 15	14	0	10	15	66,67
	txt 16	14	0	7	11	63,64
	txt 17	11	0	10	13	76,92
	txt 18	20	0	7	7	100
	txt 19	13	0	5	6	83,33
	txt 20	23	0	12	16	75
	txt 21	10	0	4	4	100
	txt 22	11	0	10	10	100
	txt 23	40	0	8	14	57,14
	txt 24	41	3	14	19	73,68
	txt 25	41	0	77	84	91,67
	txt 26	51	1	110	126	87,30
	txt 27	20	1	16	21	76,19
	txt 28	19	0	16	19	84,21
	TOTAL LEXESP	**640**	**28**	**455**	**559**	**81,40**
BB	txt 2	76	8	30	37	81,08
	TOTAL BB	**76**	**8**	**30**	**37**	**81,08**
	TOTAL	**716**	**36**	**485**	**596**	**81,38**

Tabla 7.24. Resolución de los cero pronombres en español. Fase de evaluación

Por otra parte, una característica que influye positivamente en la correcta resolución de los cero pronombres es su elevado número de ocurrencias en relación con las anáforas pronominales para un fragmento determinado de texto. Por ejemplo, en los 28 fragmentos de LEXESP utilizados en la evaluación hay 320 anáforas pronominales y 559 cero pronombres anafóricos. Esto quiere decir que hay mayor cantidad de cero pronombres que de anáforas por oración (aproximadamente el doble). Con la aplicación de la nueva preferencia número 3 (candidatos en la misma oración y que han sido solución de otros cero pronombres) se favorece la resolución de los cero pronombres que se encuentren en la misma oración, es decir, si en una oración existen varios cero pronombres con el mismo antecedente y se resuelve el primero correctamente, el resto tiene mayor posibilidad de resolverse correctamente.

A continuación se muestra un ejemplo de una oración extraída del corpus LEXESP en la que aparecen 5 cero pronombres –que el sistema resuelve correctamente– que hacen referencia al mismo antecedente: *$\emptyset_i$ cambió rápidamente las marchas para tratar de atenuar el estruendo y $\emptyset_i$ salió rodando suavemente por el paseo de Garay arriba, para volver por la calle de las Animas hacia la plaza de la Cruz Roja, no lejos de donde $\emptyset_i$ había estado aparcado; $\emptyset_i$ atravesó el puente sobre el río Segura, y por Princesa y Florida blanca $\emptyset_i$ enfiló la salida hacia Cartagena por la carretera nacional 301, todo ello circulando suavemente y sin sensación de prisas.*

- Los fallos producidos en ambos corpus corroboran las conclusiones extraídas en la fase de entrenamiento.

Para realizar una comparación real con otros autores se implementaron las aproximaciones más significativas para la resolución de la anáfora pronominal y se adaptaron para análisis sintáctico parcial y textos en español[18]. Aunque estas aproximaciones no se diseñaron para los cero pronombres, éstas se evaluaron sobre los cero pronombres detectados previamente por el sistema. Las pre-

[18] Las distintas aproximaciones implementadas se corresponden con las presentadas anteriormente en la tabla 7.17.

cisiones obtenidas aparecen en la tabla 7.25, en la que se observa que la precisión obtenida en AGIR es mejor que la obtenida en el resto de aproximaciones.

	Cercanía	Hobbs	Lappin	Strube	AGIR
LEXESP	54,86	60,42	65,97	59,72	**81,40**
BB	48,65	62,16	67,57	59,46	**81,08**

Tabla 7.25. Resolución de los cero pronombres en español. Comparación con otros autores

7.7 Generación de la anáfora pronominal

La última etapa para la generación de la anáfora pronominal en el idioma destino es la generación morfológica. Tal y como se presentó en el algoritmo de la sección 6.2.1, esta etapa recibe como entrada la representación interlingua del texto origen en la que aparecen las distintas entidades del mismo y las relaciones que existen entre ellas. En esta representación se han identificado los papeles temáticos de las entidades y se han tratado convenientemente las distintas anáforas pronominales (pronombres no referenciales, cero pronombres y pronombres personales).

A partir de la representación interlingua se tratarán y resolverán las discrepancias de número y género ocasionadas por las diferencias entre español e inglés en el tratamiento de los pronombres. Estas discrepancias se han presentado en detalle en las secciones 6.2.4 y 6.2.5 y un tratamiento adecuado de las mismas permitirá la correcta generación de las anáforas pronominales en el idioma destino.

A continuación presentaremos de un modo separado la evaluación de la generación de la anáfora pronominal en español y en inglés.

7.7.1 Generación de la anáfora pronominal en español

El algoritmo planteado para la generación de la anáfora pronominal en español se presentó previamente en la sección 6.2.1 y básicamente consiste en el tratamiento de las discrepancias de número y género.

Fase de entrenamiento. El objetivo de la fase de entrenamiento consistía en comprobar el funcionamiento de las distintas reglas morfológicas planteadas en las secciones 6.2.4 y 6.2.5 para tratar las discrepancias de número y género en la generación inglés-español.

En la evaluación de esta tarea se utilizaron los mismos fragmentos de los corpus SEMCOR y MTI utilizados previamente para la resolución de la anáfora pronominal en inglés (sección 7.5.1).

Se realizó un único experimento.

Experimento 1: Aplicación de reglas morfológicas de número y género. En este experimento se aplicaron sobre todos los pronombres personales anafóricos en tercera persona (excluidos los pronombres reflexivos) las reglas morfológicas de número y género para la generación inglés-español de las secciones 6.2.4 y 6.2.5.

Para la aplicación de las reglas morfológicas es necesario conocer el tipo semántico (*persona*, *animal* u *objeto*) y el género gramatical (*masculino* o *femenino*) del antecedente. En el corpus SEMCOR se ha utilizado la información semántica del sentido en WordNet de las palabras del texto para identificar el tipo semántico del antecedente. Esta información permite clasificar el antecedente de una anáfora como perteneciente a alguno de estos 3 tipos semánticos bien diferenciados: *persona*, *animal* u *objeto*. En el corpus MTI debido a la ausencia de este tipo de información, se han utilizado una serie de heurísticas que permiten identificar el tipo semántico del antecedente.

Con respecto a la información del género gramatical (*masculino*, *femenino*) del antecedente de la anáfora, hay que tener en cuenta que el etiquetador léxico-morfológico no proporciona (en la

mayoría de los casos) este tipo de información. Por esta razón, en AGIR se ha utilizado un diccionario electrónico inglés-español[19].

Con este tipo de información semántica y morfológica se han aplicado las reglas morfológicas de número y género.

La evaluación de esta tarea fue totalmente automática tras realizar el etiquetado anafórico de los pronombres en el que se incluía la información relativa a la generación del pronombre en el idioma destino. Los resultados que se obtuvieron aparecen en la tabla 7.26.

Corpus		Sujeto	Complemento		Aciertos	Total	P (%)
		AGENTE	TEMA	MODIF.			
SEMCOR	a01	21	3	1	21	25	84
	a12	49	8	3	58	60	96,67
	d01	28	6	9	31	43	72,09
	TOTAL SEMCOR	**98**	**17**	**13**	**110**	**128**	**85,94**
MTI	BEOWULF	35	10	2	40	47	85,11
	MAC	35	65	13	85	113	75,22
	TOTAL MTI	**70**	**75**	**15**	**125**	**160**	**78,13**
	TOTAL	**168**	**92**	**28**	**235**	**288**	**81,60**

Tabla 7.26. Generación de la anáfora pronominal inglés-español. Fase de entrenamiento: experimento 1

En la tabla aparecen para cada documento las ocurrencias de los pronombres clasificadas según su papel temático. Así, aparecen los papeles *AGENTE* (desempeñado por los pronombres *he*, *she*, *it* y *they*), *TEMA* y *MODIFICADOR* (desempeñados por los pronombres *him*, *her*, *it* y *them*). Las 3 últimas columnas representan para cada documento el número de aciertos, el total de pronombres y la precisión obtenida respectivamente. En esta tarea la precisión se entiende como el cociente entre el número

[19] Este diccionario se ha incorporado al sistema como una base de datos y proporciona para una palabra determinada en inglés una única traducción, su género y su número en español.

de pronombres correctamente generados en español[20] y el número de pronombres totales. Por ejemplo, el documento a01 del corpus SEMCOR tiene 21 pronombres con papel temático *AGENTE*, 3 con *TEMA* y 1 con *MODIFICADOR* y en él se ha obtenido una precisión de 84% (21/25).

Discusión. De los resultados obtenidos se extrajeron las siguientes conclusiones:

- En el corpus SEMCOR todas las ocurrencias de los pronombres *he*, *she*, *him* y *her* han sido generadas correctamente en español. Se justifica por dos razones:
  - Los papeles temáticos desempeñados por estos pronombres se han identificado correctamente en todos los casos.
  - Estos pronombres incorporan la información gramatical necesaria (*género* y *número*) para su correcta generación en español, independientemente del antecedente propuesto como solución de los mismos.

  Los fallos en la generación de los pronombres *it*, *they* y *them* han sido originados por diversas causas:
  - Fallos en la resolución de la anáfora, es decir, el antecedente propuesto por el sistema no es el correcto (44,44% de los fallos). Esto provoca una incorrecta generación en español motivada, principalmente, porque el antecedente propuesto y el correcto tienen distinto género gramatical.
  - Fallos en la identificación del papel temático de los pronombres lo que provoca la aplicación de una regla morfológica incorrecta (44,44%). Son originados, principalmente, por una incorrecta división de cláusulas.
  - Fallos producidos por el diccionario inglés-español (11,12%). Pueden ocurrir dos circunstancias: (a) la palabra no se encuen-

[20] Un pronombre se considera que está bien generado cuando la salida proporcionada por el sistema coincide con el pronombre propuesto tras realizar el etiquetado anafórico del corpus. Con esta medida se pretende evaluar la correcta aplicación de la regla morfológica correspondiente. Por ejemplo, para el pronombre *he* la regla morfológica propone únicamente el pronombre español *él*, aunque el pronombre *éste* sería igualmente válido; en este caso si el sistema hubiera propuesto como solución el pronombre *éste*, habría sido considerado como fallo en la evaluación automática.

tra en el diccionario[21], y (b) la palabra tiene una traducción distinta y con género distinto a la que aparece en el diccionario debido a que tiene diferentes acepciones.

- En el corpus MTI prácticamente sólo hay ocurrencias de los pronombres *it*, *they* y *them* (96,25% del total de pronombres). Los fallos en la generación de estos pronombres son originados por las mismas causas que en el corpus SEMCOR pero con distintos porcentajes:
  - Fallos en la resolución de la anáfora (22,86% de los fallos).
  - Fallos en la identificación del papel temático de los pronombres (62,86%). En este tipo de corpus la incorrecta división de las cláusulas se origina, principalmente, por errores del POS tagger que no identifica correctamente los verbos imperativos (carecen de sujeto).
  - Fallos producidos por el diccionario inglés-español (14,28%). En este tipo de corpus existen muchas palabras técnicas que no están en el diccionario.

Tras analizar los resultados de este experimento se observa que se ha obtenido una precisión de 85,94% y 78,13% para los corpus SEMCOR y MTI. El porcentaje inferior (aproximadamente un 7%) obtenido con MTI viene originado, principalmente, por el tipo de corpus (prácticamente sólo hay ocurrencias de los pronombres *it*, *they* y *them*) y por la falta de información semántica.

Con este experimento se concluye la fase de entrenamiento y se pasa a la fase de evaluación.

Fase de evaluación. Los resultados obtenidos en esta fase para los corpus SEMCOR y MTI se muestran en la tabla 7.27.

Discusión. En la tarea de generación de la anáfora pronominal en español se ha obtenido una precisión global de 80,39% (582/724). En particular se han obtenido unas precisiones de 90,16% y 75,11% para los corpus SEMCOR y MTI respectivamente.

[21] Si una palabra no se ha encontrado en el diccionario se selecciona, por defecto, el género *masculino*. Esta medida se ha adoptado tras realizar pruebas empíricas en las que se seleccionaba por defecto género *masculino* o *femenino*, observando que se obtuvieron mejores resultados con el primero.

Corpus		Sujeto	Complemento		Aciertos	Total	P (%)
		AGENTE	TEMA	MODIF.			
SEMCOR	a02	21	5	1	23	27	85,19
	a11	10	5	0	14	15	93,33
	a13	17	2	3	21	22	95,45
	a14	40	10	1	48	51	94,12
	a15	32	5	4	34	41	82,93
	d02	14	2	3	18	19	94,74
	d03	13	0	1	12	14	85,71
	d04	50	6	9	59	65	90,77
	TOTAL SEMCOR	**197**	**35**	**22**	**229**	**254**	**90,16**
MTI	CDROM	38	24	7	47	69	68,12
	PSW	24	36	2	52	62	83,87
	WINDOWS	16	19	2	30	37	81,08
	SCANWORX	95	87	11	142	193	73,58
	GIMP	66	33	10	82	109	75,23
	TOTAL MTI	**239**	**199**	**32**	**353**	**470**	**75,11**
	TOTAL	**436**	**234**	**54**	**582**	**724**	**80,39**

Tabla 7.27. Generación de la anáfora pronominal inglés-español. Fase de evaluación

Estos resultados reafirman las conclusiones extraídas de la fase de entrenamiento en la que se concluía que el tipo de corpus y la información semántica influía considerablemente en la resolución y posterior generación de la anáfora pronominal. Los fallos producidos en la fase evaluación son originados por las mismas causas y en unos porcentajes aproximados que los originados durante la fase de entrenamiento.

No vamos a comparar nuestros resultados con los obtenidos por otras aproximaciones por dos razones fundamentales. En primer lugar, en este trabajo se presenta el primer estudio detallado de las discrepancias entre español e inglés en el tratamiento de los pronombres que permite realizar una correcta generación de la anáfora pronominal (incluidos los cero pronombres) en el idio-

ma destino, consecuentemente, la comparación con otros pares de idiomas sería irrelevante.

En segundo lugar, con este trabajo se pretende evaluar la correcta aplicación de las reglas morfológicas que proporcionan el pronombre adecuado en el idioma destino. Una tarea posterior debe decidir si el pronombre en el idioma destino se debe generar tal y como lo propone nuestro sistema, se debe sustituir por otro tipo de pronombre (como por ejemplo un pronombre posesivo o un pronombre enclítico) o simplemente se debe eliminar (por ejemplo, los cero pronombres en español). La comparación, pues, con otras aproximaciones o sistemas reales de Traducción Automática que abordan este tipo de problemas no tendría mucho sentido.

7.7.2 Generación de la anáfora pronominal en inglés

El algoritmo para la generación de la anáfora pronominal en inglés se presentó en la sección 6.2.1 (es válido tanto para español como para inglés) y básicamente consiste en el tratamiento de las discrepancias de número y género.

Fase de entrenamiento. En esta fase se pretendía comprobar el correcto funcionamiento de las reglas morfológicas presentadas en las secciones 6.2.4 y 6.2.5 para el tratamiento de las discrepancias de número y género en la generación español-inglés.

En la evaluación de esta tarea se utilizaron los mismos fragmentos del corpus LEXESP utilizados previamente para la resolución de la anáfora pronominal en español (sección 7.5.2).

En esta fase sólo se realizó un experimento.

Experimento 1: Aplicación de reglas morfológicas de número y género. En este experimento se trataron todos los pronombres personales anafóricos en tercera persona de sujeto, complemento y cero pronombres (se excluyeron los pronombres reflexivos). Sobre ellos se aplicaron las reglas morfológicas de número y género para la generación español-inglés.

Para la aplicación de estas reglas es necesario conocer el tipo semántico (*persona*, *animal* u *objeto*) y el género gramatical (*masculino* o *femenino*) del antecedente. Ya que en el corpus LEXESP no se incluye información semántica, se han utilizado una serie de

heurísticas para identificar el tipo semántico del antecedente. Por otra parte, la información del género gramatical del antecedente es proporcionada por la etiqueta léxico-morfológica del núcleo del mismo.

Con este tipo de información semántica y morfológica se aplicaron las correspondientes reglas morfológicas.

Los resultados de la evaluación automática aparecen en la tabla 7.28.

Corpus		Sujeto	Complemento		Aciertos	Total	P (%)
		AGENTE	TEMA	MODIF.			
LEXESP	txt29	78	19	8	88	105	83,81
	txt30	61	11	11	64	83	77,11
	txt31	30	7	1	29	38	76,32
	TOTAL	**169**	**37**	**20**	**181**	**226**	**80,09**

Tabla 7.28. Generación de la anáfora pronominal español-inglés. Fase de entrenamiento: experimento 1

En la tabla aparecen para cada documento las ocurrencias de los pronombres clasificadas según su papel temático: *AGENTE* (desempeñado por los pronombres *él*, *ella*, *ellos*, *ellas* y cero pronombres), *TEMA* y *MODIFICADOR* (desempeñados por los pronombres *él*, *ella*, *lo*, *la*, *le*, *ellos*, *ellas*, *los*, *las*, *les*). En las 3 últimas columnas aparecen para cada documento el número de aciertos, el total de pronombres y la precisión obtenida.

Discusión. Tras realizar el experimento se extrajeron las siguientes conclusiones:

- Todas las ocurrencias de los pronombres en plural (*ellos*, *ellas*, *les*, *los*, *las* y cero pronombres en plural) han sido generadas correctamente en inglés. Hay dos razones que justifican esta circunstancia:
  - Los papeles temáticos de estos pronombres (*AGENTE*, *TEMA* o *MODIFICADOR*) se han identificado correctamente.

- Los pronombres plurales en inglés correspondientes (*they* y *them*) carecen de género, es decir, son válidos para masculino y femenino, por lo que el género del antecedente no influye en la generación de los mismos.

- Los fallos se han producido en la generación de los pronombres *él*, *ella*, *le*, *lo*, *la* y cero pronombres en singular y se han originado por diversas causas:
  - Fallos en la resolución de la anáfora, es decir, el antecedente propuesto por el sistema no es el correcto (82,22% de los fallos). Esta circunstancia provoca una incorrecta generación en inglés motivada, principalmente, porque el antecedente propuesto y el correcto difieren en su género gramatical. En otras ocasiones, aunque el género coincide, difieren en tipo semántico; por ejemplo, el antecedente propuesto es de tipo *objeto* y el correcto es de tipo *persona*.
  - Fallos de la heurística que se aplica para identificar el tipo semántico del antecedente (17,78%). Este hecho supone la aplicación de una regla morfológica incorrecta. Por ejemplo, si el antecedente de una anáfora es una ciudad (nombre propio) se asume que es una entidad de tipo *persona* por lo que se producirá una generación incorrecta. Del mismo modo puede ocurrir que una entidad del tipo *persona* no sea nombre propio, por lo que se asumirá que es del tipo *animal-objeto* y se realizará una generación incorrecta[22].

Tras los resultados obtenidos en este experimento, una precisión de 80,09% en la generación español-inglés, se realiza la fase de evaluación.

Fase de evaluación. Los resultados obtenidos en la fase de evaluación aparecen en la tabla 7.29.

Discusión. En la tarea de generación de la anáfora pronominal en inglés se ha obtenido una precisión de 84,77% (657/775). Los fallos

[22] En el siguiente fragmento de texto extraído del corpus LEXESP hay un cero pronombre cuyo antecedente es del tipo *persona* y no es nombre propio: *Había* [otro viejo]$_i$ *allí;* $\emptyset_i$ *Tiraba piedrecitas a los peces que acudían a la superficie...* En este caso el sistema propone el antecedente correcto pero lleva a cabo una generación incorrecta.

Corpus		Sujeto	Complemento		Aciertos	Total	P (%)
		AGENTE	TEMA	MODIF.			
LEXESP	txt1	19	3	1	21	23	91,30
	txt2	35	7	1	33	43	76,74
	txt3	21	4	1	19	26	73,08
	txt4	13	4	1	15	18	83,33
	txt5	13	4	1	14	18	77,78
	txt6	17	1	0	16	18	88,89
	txt7	22	3	4	28	29	96,55
	txt8	10	0	0	9	10	90
	txt9	9	3	1	8	13	61,54
	txt10	17	2	1	19	20	95
	txt11	7	0	1	7	8	87,5
	txt12	25	4	0	29	29	100
	txt13	16	0	0	12	16	75
	txt14	11	0	0	10	11	90,91
	txt15	16	3	5	18	24	75
	txt16	11	1	2	13	14	92,86
	txt17	14	1	0	11	15	73,33
	txt18	9	4	0	10	13	76,92
	txt19	7	0	1	7	8	87,5
	txt20	17	3	1	13	21	61,90
	txt21	4	2	0	6	6	100
	txt22	12	1	2	15	15	100
	txt23	15	4	2	19	21	90,48
	txt24	21	7	2	25	30	83,33
	txt25	92	11	5	100	108	92,59
	txt26	132	16	11	129	159	81,13
	txt27	24	6	1	27	31	87,10
	txt28	21	5	2	24	28	85,71
	TOTAL	**630**	**99**	**46**	**657**	**775**	**84,77**

Tabla 7.29. Generación de la anáfora pronominal español-inglés. Fase de evaluación

Volver al índice/Tornar a l'índex

producidos en esta fase han sido provocados por una mala resolución de la anáfora (79,66%) y por excepciones de la heurística aplicada (20,34%).

Estos porcentajes ponen de manifiesto que si se pudieran obtener mejores resultados en el proceso de resolución de la anáfora (por ejemplo, con la incorporación de información semántica) se mejoraría notablemente la precisión obtenida en el proceso de generación inglés-español.

8. Conclusiones y trabajos futuros

En este trabajo se ha presentado una aproximación de un sistema interlingua de Traducción Automática que lleva a cabo la resolución de la anáfora pronominal originada por los pronombres personales de tercera persona en español e inglés y su posterior generación en el idioma destino.

El sistema, al que hemos denominado AGIR (*Anaphora Generation with an Interlingua Representation*), utiliza una serie de fuentes de información lingüísticas (léxicas, morfológicas, sintácticas y semánticas) en los módulos de análisis y generación del mismo. Concretamente, estas fuentes de información se utilizan en las etapas de análisis sintáctico y resolución de problemas lingüísticos que conducen a la última etapa del módulo de análisis, la representación interlingua del texto origen. El módulo de generación recibe como entrada la representación interlingua (en la que se incluye la información lingüística) y, tras las etapas de generación sintáctica y morfológica, realiza la correcta generación de la anáfora pronominal en el idioma destino (español o inglés).

Gracias al uso de información no dependiente del dominio, el sistema AGIR es capaz de resolver y generar las anáforas pronominales en textos no restringidos de cualquier dominio.

Las principales aportaciones de este trabajo son las siguientes:

- La propuesta general de un sistema interlingua de Traducción Automática. La representación interlingua intermedia tiene las siguientes características:
  - Representa el texto completo utilizando la cláusula (en vez de la oración) como unidad básica de la representación. En cada cláusula se utiliza una representación basada en papeles temáticos en la que aparecen los roles o papeles de las

distintas entidades que en ellas aparecen. Consecuentemente, la representación interlingua contiene todas las entidades del texto.

- Contiene y expresa las relaciones que existen entre las distintas entidades del texto mediante los enlaces correspondientes. Así se expresan: la relación entre una entidad y su modificador, la relación entre una entidad anafórica y su antecedente y las relaciones entre distintas entidades anafóricas que tienen el mismo antecedente (cadenas de correferencia).
- Representa las unidades léxicas interlingua del texto utilizando su sentido correcto en WordNet. De este modo, tras realizar el acceso correspondiente al módulo interlingua ILI de EuroWordNet se puede obtener la unidad léxica correspondiente en el idioma destino.
- Se obtiene tras realizar un análisis sintáctico parcial del texto origen. Esta característica garantiza su aplicabilidad sobre textos no restringidos de cualquier dominio.

- La construcción de los correspondientes módulos de análisis para inglés y español en AGIR que implican la realización de las siguientes tareas:
  - Análisis léxico y morfológico. En esta etapa se utiliza un etiquetador léxico-morfológico que proporciona a cada unidad léxica del texto origen su categoría gramatical e información léxico-morfológica.
    Esta tarea implica la definición de un conjunto de herramientas interfaz que adaptan esta información para su uso en la siguiente etapa de análisis sintáctico. La existencia de estas herramientas proporcionan una gran flexibilidad al sistema ya que éste es capaz de trabajar con distintos etiquetadores así como con distintas gramáticas para realizar el análisis sintáctico (sólo es necesario definir el interfaz apropiado).
  - Análisis sintáctico parcial. En esta etapa se utiliza el analizador sintáctico parcial del sistema SUPAR (Ferrández *et al.*, 1999a) para español y la adaptación correspondiente para inglés. Se obtiene la estructura sintáctica parcial para cada una de las oraciones del texto y una lista con todos los an-

tecedentes candidatos que se usará en la etapa posterior de resolución de la anáfora.

- Resolución de problemas lingüísticos. En esta tarea nos hemos centrado en la detección de los pronombres *it* pleonásticos en inglés y en la resolución de la anáfora pronominal originada por los pronombres personales (para español e inglés) y los cero pronombres españoles. En la resolución de las expresiones anafóricas, AGIR ha utilizado un algoritmo basado en restricciones y preferencias lingüísticas que tiene las siguientes características:
  - Permite la resolución multilingüe de la anáfora. En concreto se ha utilizado para español y se ha adaptado convenientemente para el inglés aunque puede ser adaptado fácilmente para otros idiomas.
  - Requiere la definición de un espacio de accesibilidad anafórica óptimo así como la definición de un conjunto de restricciones morfológicas, sintácticas y semánticas válidas para cada tipo de anáfora. Todo ello se obtiene tras un profundo estudio del idioma en cuestión.
  - Requiere la definición de un conjunto válido de preferencias basadas en información lingüística para cada tipo de anáfora. El grado de importancia de cada una de ellas se determina tras un proceso de entrenamiento que culmina con la obtención de la configuración óptima.
  - Aplica, en primer lugar, el conjunto de restricciones como un filtro lingüístico con el objetivo de descartar candidatos. Posteriormente, sobre los candidatos restantes, se aplican secuencialmente los distintos niveles de preferencia con el objetivo de eliminar aquellos candidatos que no cumplan la preferencia. Con esta estrategia las preferencias son consideradas como restricciones, excepto cuando ningún candidato satisface una preferencia en cuyo caso no se descarta ninguno de ellos y se aplica la siguiente preferencia.
- Obtención de la representación interlingua. Esta etapa es la última del módulo de análisis y en ella se obtiene la representación interlingua global del texto. Las características principa-

les de esta representación ya se han mencionado previamente en el inicio de este capítulo.

- La construcción de los correspondientes módulos de generación para inglés y español en AGIR. En estos módulos nos hemos centrado exclusivamente en la generación de la anáfora pronominal y cero pronombres en el idioma destino. Para llevar a cabo esta tarea se ha desarrollado un estudio profundo sobre las diferencias (*discrepancias*) entre español e inglés en cuanto al tratamiento de los pronombres. Este estudio nos permitirá realizar una correcta generación de las expresiones anafóricas en el idioma destino.
  En el módulo de generación se llevan a cabo dos tareas:
  - Generación sintáctica. En esta etapa se tratan las discrepancias sintácticas entre español e inglés. Básicamente se han estudiado dos:
    - Pronombres pleonásticos. Concretamente, se ha llevado a cabo la identificación y clasificación de los pronombres *it* pleonásticos en inglés. Estos pronombres no son anafóricos y, por lo tanto, no se deben generar en español.
    - Cero pronombres. En AGIR se ha desarrollado el primer estudio para el español que permite la detección, resolución y generación en inglés de los cero pronombres españoles con función de sujeto.
  - Generación morfológica. En esta etapa se tratarán y resolverán las discrepancias de número y género existentes entre los pronombres en español y en inglés:
    - Discrepancias de número. Se producen por las diferencias del número gramatical entre palabras de distintos idiomas (en este caso español e inglés) que expresan el mismo concepto. Particularmente, estas palabras serán referidas por un pronombre en singular en el idioma origen y por un pronombre plural en el idioma destino o viceversa.
      Para resolver estas discrepancias se ha construido un conjunto de reglas morfológicas para la traducción inglés-español y español-inglés. Estas reglas utilizan la información

que se encuentra disponible en la representación interlingua para las distintas entidades del texto: papel temático, información léxico, morfológica y semántica y, por último, información de su antecedente (para el caso de las expresiones anafóricas).

- Discrepancias de género. Se originan por las diferencias morfológicas existentes entre el español y el inglés en el tratamiento de los pronombres personales. El español marca el género gramatical de los pronombres personales y distingue entre las formas masculina y femenina de los mismos mientras que el inglés, en general, no lo hace (principalmente para los pronombres es plural).
  Para resolver estas discrepancias se ha construido un conjunto de reglas morfológicas que utilizan toda la información disponible en la representación interlingua y permiten la correcta generación de los pronombres en el idioma destino.

- La evaluación global del sistema AGIR. Para ello se realizó una evaluación independiente de las distintas tareas llevadas a cabo en el sistema global y que conducen a la tarea final: la generación de la anáfora pronominal en el idioma destino.
  En la evaluación de las distintas tareas se utilizó la misma metodología. En primer lugar, se realiza la fase de entrenamiento en la que se llevan a cabo distintos experimentos con el objetivo de obtener la configuración óptima que proporcione los mejores resultados. Los corpus usados en esta fase son fragmentos de los corpus originales elegidos aleatoriamente (corpus de entrenamiento) reservando el resto de corpus para la fase de evaluación (corpus de evaluación). En segundo lugar, cuando el sistema se ha ajustado con la mejor configuración, se realiza la fase de evaluación sobre el corpus de evaluación.
  Para garantizar la fiabilidad de los resultados y su posible aplicación sobre textos no restringidos de cualquier dominio se seleccionaron cinco corpus de tipos muy variados: textos técnicos que incluyen manuales de Informática (corpus MTI en inglés) y manuales de telecomunicaciones (corpus Blue Book en español) y, por otra parte, textos narrativos que tratan temas muy di-

versos (corpus Federal Register y SEMCOR en inglés y corpus LEXESP en español).
Las distintas tareas se evaluaron automáticamente tras realizar un etiquetado anafórico de las anáforas pronominales y cero pronombres de los distintos corpus. Las tareas evaluadas fueron las siguientes:

- Detección de los pronombres *it* pleonásticos en la que se obtuvo una precisión de 88,75%. Hay que destacar la importancia de la correcta detección de este tipo de pronombres ya que el 32,97% de pronombres *it* en el corpus de evaluación eran de este tipo.
- Detección de los cero pronombres españoles en la que se obtuvieron unas precisiones de 98,17% y 80,58% en la detección de verbos con su sujeto omitido y verbos con su sujeto no omitido respectivamente. En los corpus estudiados, el 48,96% de los verbos tenían su sujeto omitido; este porcentaje pone de manifiesto la importancia de este fenómeno en español.
- Resolución de la anáfora pronominal en inglés en la que se obtuvieron precisiones de 86,61% y 76,81% para los corpus SEMCOR y MTI respectivamente. Para el primer corpus se utilizó información semántica proporcionada por WordNet que permitió aplicar una serie de restricciones semánticas.
- Resolución de la anáfora pronominal y cero pronombres en español en las que se alcanzaron precisiones de 82,19% y 81,38% respectivamente.
- Generación de la anáfora pronominal en español y en inglés en las que se obtuvieron precisiones de 80,39% y 84,77% respectivamente.

Como conclusión final indicaremos que en esta Tesis se ha presentado un sistema interlingua en el que se lleva a cabo la resolución y generación de la anáfora pronominal en el idioma destino. El sistema trabaja sobre textos no restringidos de cualquier dominio en español e inglés aunque se puede extender fácilmente a otros idiomas (sistema interlingua multilingüe).

Las precisiones obtenidas en la generación de las expresiones anafóricas son muy cercanas a las alcanzadas en la resolución de

las mismas. Tras analizar los resultados se dedujo que el proceso de resolución influye decisivamente en el proceso de generación, consiguientemente, todas las mejoras que se puedan conseguir en el primero influirán positivamente en el segundo.

8.1 Trabajos futuros

Este trabajo sirve como punto de partida para otras líneas de investigación que actualmente se están llevando a cabo o que en el futuro se convertirán en nuevas vías de investigación. Entre éstas podemos citar las siguientes:

- Incorporación de información semántica a las unidades léxicas del texto, de modo que este tipo de información se pueda utilizar en las distintas etapas de los módulos de análisis y generación del sistema interlingua: análisis sintáctico, resolución de las expresiones anafóricas, obtención de la representación interlingua, generación sintáctica y generación morfológica.
  En AGIR sólo se ha realizado un experimento en el que se incluía información semántica (proporcionada por el recurso léxico WordNet) en el proceso de resolución de la anáfora. Los resultados obtenidos demostraron que la inclusión de este tipo de información mejoró notablemente la precisión de esta tarea. Consecuentemente, si esta información se incorporara en las distintas tareas que se llevan a cabo en el sistema AGIR, la precisión obtenida en la tarea final (generación de la anáfora pronominal en el idioma destino) mejoraría considerablemente.
- Tratamiento, resolución y generación de otro tipo de referencias. Se proponen las anáforas pronominales originadas por los pronombres reflexivos y demostrativos para el inglés. Este tipo de anáforas se han resuelto previamente en español por lo que el algoritmo se podría adaptar convenientemente para el inglés. Del mismo modo, se propone la resolución de las descripciones definidas en inglés. Las descripciones definidas, a diferencia de las anáforas pronominales, contienen mayor información léxica que permite definir mejor la compatibilidad entre la referencia

y el antecedente. Previamente, es necesario clasificar las mismas como anafóricas y no anafóricas, ya que no todas tienen propiedades referenciales. En esta línea de investigación se puede continuar y adaptar para inglés los trabajos realizados en la resolución de las descripciones definidas para español por compañeros de nuestro Grupo de Investigación (Muñoz *et al.*, 2000; Muñoz, 2001).

- Realización automática de la alineación anafórica de dos corpus bilingües, es decir, establecer las correspondencias de las anáforas pronominales entre dos textos iguales expresados en distintos idiomas. A partir de los corpus bilingües alineados (a nivel de párrafo, oración y expresiones anafóricas) se podrán estudiar las discrepancias reales que existen entre los dos idiomas en cuanto a la traducción de la anáfora pronominal. Por ejemplo un pronombre personal en un idioma se puede traducir por un pronombre posesivo, enclítico o puede ser eliminado en otro idioma.
- Inclusión de otros idiomas en el sistema interlingua con el desarrollo de los correspondientes módulos de análisis y generación para los mismos.
- Análisis de la influencia de la aplicación del proceso de resolución de las expresiones anafóricas sobre diversas aplicaciones en el campo del Procesamiento del Lenguaje Natural. Concretamente, en nuestro Grupo de Investigación se ha trabajado sobre las siguientes aplicaciones:
  - Extracción de Información. Estos sistemas extraen información relevante de forma estructurada a partir de un texto. La importancia de la resolución de las descripciones definidas en español en estos sistemas ha sido tratada por Muñoz (2001).
  - Sistemas inteligentes de diálogo. En ellos se produce una comunicación entre hombre y máquina donde el usuario conversa con el sistema utilizando el lenguaje natural, y el sistema genera respuestas en lenguaje natural. La resolución de la anáfora pronominal y adjetiva en español en estos sistemas ha sido tratada por Martínez-Barco (2001).
  - Búsqueda de Respuestas. Los sistemas de Búsqueda de Respuestas son herramientas capaces de obtener respuestas con-

cretas a necesidades de información muy precisas a partir del análisis de documentos escritos en lenguaje natural. Localizan y extraen las respuestas de aquellas zonas de los documentos de cuyo contenido es posible inferir la información requerida. La importancia de la resolución de la anáfora pronominal en inglés en estos sistemas ha sido estudiada por varios compañeros (Vicedo *et al.*, 2000; Vicedo & Ferrández, 2000).

– Realización automática de resúmenes de texto. Los sistemas de generación de resúmenes trabajan sobre un texto de forma que proporcionan automáticamente una sinopsis de todo el texto procesado que contiene la información relevante del mismo.

8.2 Producción científica

Como consecuencia de la labor de investigación llevada a cabo en esta Tesis se han producido las siguientes publicaciones relacionadas directamente con la resolución y generación de la anáfora pronominal en español e inglés:

1. Revistas internacionales:
   - Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., & Muñoz, R. 2001. An Algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics*, **27**(4).
   - Ferrández, A., Moreno, L., Palomar, M., y Peral, J. 1998. Un método de resolución de la anáfora discursiva en textos no restringidos mediante la unificación. *Revista Iberoamericana de Inteligencia Artificial*, **4**(1), 20–29.
2. Revistas nacionales:
   - Saiz-Noeda, M., Suárez, A., y Peral, J. 1999. Propuesta de incorporación de información semántica desde Wordnet al análisis sintáctico parcial orientado a la resolución de la anáfora. *Procesamiento del Lenguaje Natural*, **25**, 167–173.
   - Ferrández, A., Palomar, M., Moreno, L., Martínez-Barco, P., Peral, J., Muñoz, R., y Saiz-Noeda, M. 1999. Demostración

del "Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística". *Procesamiento del Lenguaje Natural*, **25**, 217–218.

- Peral, J., Martínez-Barco, P., Ferrández, A., y Navarro, B. 1998. Sistema de adquisición automática de reglas gramaticales. *Procesamiento del Lenguaje Natural*, **23**, 110–117.

3. Actas de congresos internacionales:
   - Peral, J. 2001. El problema de la anáfora: resolución, aplicaciones y líneas futuras. *In: Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies (SLPLT-2)*. Jaén (Spain). pp. 159–163.
   - Ferrández, A., & Peral, J. 2000. A computational approach to zero-pronouns in Spanish. *In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*. Hong Kong (China). pp. 166–172.
   - Peral, J., & Ferrández, A. 2000. Generation of Spanish zero-pronouns into English. *In:* Christodoulakis, D.N. (ed), *Natural Language Processing - NLP'2000*. Lecture Notes in Artificial Intelligence, vol. 1835. Patras (Greece): Springer-Verlag. pp. 252–260.
   - Peral, J., & Ferrández, A. 2000. An application of the Interlingua System ISS for Spanish-English Pronominal Anaphora Resolution. *In: Proceedings of the Third AMTA/SIGIL Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP (ANLP/NAACL'2000)*. Seattle, Washington (U.S.A.). pp. 42–51.
   - Saiz-Noeda, M., Peral, J., & Suárez, A. 2000. Semantic compatibility Techniques for Anaphora Resolution. *In: Proceedings of the International Conference on Artificial and Computational Intelligence For Decision, Control and Automation In Engineering and Industrial Applications (ACIDCA'2000)*. Monastir (Tunizia). pp. 43–48.
   - Peral, J., Saiz-Noeda, M., Ferrández, A., & Palomar, M. 1999. Anaphora resolution and generation in a multilingual system. Interlingua mechanism. *In: Proceedings of Venezia*

per il Trattamento Automatico delle Lingue (VEXTAL'99). Venice (Italy). pp. 315–324.

- Peral, J. 1999. Proposal of an English-Spanish interlingual mechanism focused on pronominal anaphora resolution and generation in Machine Translation systems. *In: Proceedings of the Student Session of the 11th European Summer School in Logic, Language and Information (ESSLLI'99).* Utrecht (Holland). pp. 169–179.
- Palomar, M., Ferrández, A., Moreno, L., Saiz-Noeda, M., Muñoz, R., Martínez-Barco, P., Peral, J., & Navarro, B. 1999. A Robust Partial Parsing Strategy based on the Slot Unification Grammars. *In: Proceedings of the Sixth Conference on Natural Language Processing (TALN'99).* Corsica (France). pp. 263–272.
- Peral, J., Palomar, M., & Ferrández, A. 1999. Coreference-oriented Interlingual Slot Structure & Machine Translation. *In: Proceedings of the ACL Workshop Coreference and its Applications.* College Park, Maryland (U.S.A.). pp. 69–76.
- Martínez-Barco, P., Peral, J., Ferrández, A, Moreno, L., y Palomar, M. 1998. Analizador Parcial SUPP. *In: Proceedings of the VI Biennial Iberoamerican Conference on Artificial Intelligence (IBERAMIA'98).* Lisbon (Portugal). pp. 329–342.

4. Actas de congresos nacionales:
   - Ferrández, A., Moreno, L., Palomar, M., y Peral, J. 1997. Un método de resolución de la anáfora discursiva mediante la unificación. *In: Actas de la VII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'97).* Málaga (España). pp. 805–814.

Además de las anteriores, se citan a continuación algunas publicaciones de temas relacionados que corresponden a desarrollos anteriores al de la Tesis, así como a trabajos futuros.

1. Revistas nacionales:
   - Vicedo, J.L., Ferrández, A., y Peral, J. 2000. ¿Cómo influye la resolución de la anáfora pronominal en los Sistemas de

Búsqueda de Respuestas? *Procesamiento del Lenguaje Natural*, **26**, 231–237.

- Ferrández, A., Peral, J., Martínez, P., Saiz-Noeda, M., y Romero, R. 1997. Resolución de la extraposición a izquierdas con las Gramáticas de Unificación de Huecos. *Procesamiento del Lenguaje Natural*, **21**, 167–182.

2. Actas de congresos internacionales:
   - Peral, J., Martínez-Barco, P., Muñoz, R., Ferrández, A., Moreno, L., y Palomar, M. 1999. Una técnica de análisis parcial sobre textos no restringidos (SUPP) aplicada a un Sistema de Extracción de Información (EXIT). *In: Actas del VI Simposio Internacional de Comunicación Social.* Santiago de Cuba (Cuba). pp. 662–669.
3. Informes Técnicos:
   - Llopis, F., Muñoz, R., Suárez, A., Montoyo, A., Palomar, M., Ferrández, A., Peral, J., Martínez-Barco, P., Romero, R., y Saiz-Noeda, M. 1998. Sistema EXIT. *Report Interno - DLSI. Universidad de Alicante.*

Referencias

Alcaraz, E., & Moody, B. 1997. *Morfosintaxis inglesa para hispanohablantes. Teoría y práctica.* Alcoy: Editorial Marfil.

Alexandersson, J., Maier, E., & Reithinger, N. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. *Pages 188–193 of: Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95).*

Allegranza, V., Krauwer, S., & Steiner, E. 1991. Introduction. *Machine Translation (Eurotra Special Issue)*, **6**(2), 61–71.

Allen, J. 1995. *Natural Language Understanding.* The Benjamin/Cummings Publishing Company, Inc.

Alonso, J.A. 1990. Transfer InterStructure: designing an 'interlingua' for transfer-based MT systems. *Pages 189–201 of: Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'90).*

ALPAC. 1966. *Language and machines: computers in translation and linguistics.* Tech. rept. Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Research Council. National Academy of Sciences, Washington, DC.

Amores, J.G. 1992. *A Lexical Functional Grammar-based Machine Translation System for Medical Abstracts.* Ph.D. thesis, Universidad de Sevilla.

Amores, J.G., & Quesada, J.F. 1997. Episteme. *Procesamiento del Lenguaje Natural*, **21**, 1–15.

Amores, J.G., Quesada, J.F., & Tapias, D. 1994. Traducción Automática basada en el formalismo LFG con entrada y salida por voz. *Comunicaciones de Telefónica I+D*, **5**(2), 132–147.

Appelo, L., & Landsbergen, J. 1986. The machine translation project Rosetta. *Pages 34–51 of:* Gerhardt, T.C. (ed), *I. International Conference on the State of the Art in Machine Translation in America, Asia and Europe: Proceedings of IAI-MT86, IAI/EUROTRA-D.*

Arnold, D., & Sadler, L. 1990. The theoretical basis of MiMo. *Machine Translation*, **5**, 195–222.

Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S., & Sadler, L. 1994. *Machine Translation. An introductory guide.* Oxford: NCC Blackwell.

Asher, N. 1993. *Reference to Abstract Objects.* Dordrecht: Kluwer.

Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Màrquez, L., Martí, M.A., Padró, L., Placer, R., Rodríguez, H., Taulé, M., & Turmo, J. 1998. Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. *Pages 1267–1272 of: Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98).*

Baldwin, B. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. *Pages 38–45 of: Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution.*

Beale, S., Nirenburg, S., & Mahesh, K. 1996. HUNTER-GATHERER: Three search techniques integrated for natural language semantics. *Pages 1056–1061 of: Proceedings of the National Conference on AI (AAAI'96).*

Beale, S., Viegas, E., & Nirenburg, S. 1997. Breaking down barriers: the Mikrokosmos generator. *In: Proceedings of the Natural Language Pacific Rim Symposium (NLPRS'97).*

Bech, A., & Nygaard, A. 1988. The E-framework: a formalism for natural language processing. *Pages 36–39 of: Proceedings of the 12th International Conference on Computational Linguistics (COLING'88).*

Bech, A., Maegaard, B., & Nygaard, A. 1991. The Eurotra MT formalism. *Machine Translation*, **6**, 83–101.

Bennet, W.S., & Slocum, J. 1985. The LRC machine translation system. *Computational Linguistics*, 111–121. Reprinted in (Slocum, 1988), 49–84.

Berger, A., Brown, P., Pietra, S.D., Pietra, V.D., Gillett, J., Lafferty, J., Mercer, R.L., Printz, H., & Ures, L. 1994. The Candide system for Machine Translation. *Pages 157–163 of: Proceedings of the ARPA Workshop on Speech and Natural Language.*

Blanchon, H., Boitet, C., & Caelen, J. 1999. Participation francophone au Consortium C-STAR II. *La tribune des industries de la langue et du multimedia*, **31/32**, 15–23.

Bobrow, D. 1969. Natural Language Input for a Computer Problem Solving Program. *In:* Minsky, M. (ed), *Semantic Information Processing.* Cambridge, Mass: MIT Press.

Boitet, C. 1987. Research and development on MT and related techniques at Grenoble University. *Pages 133–153 of:* King, M. (ed), *Machine translation today: the state of the art.* Edinburgh Information Technology Series 2. Edinburgh University Press.

Boitet, C. 1989. GETA project. *Pages 54–65 of:* Nagao, M. (ed), *Machine Translation Summit.* Tokyo: Ohmsha.

Boitet, C. 1997. GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects. *Pages 23–57 of: Proceedings of PACLING'97.*

Boitet, C., & Blanchon, H. 1994. Multilingual dilogue-based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation*, **9**(2), 99–132.

Boitet, C., & Nédobejkine, N. 1981. Recent developments in Russian-French machine translation at Grenoble. *Linguistics*, **19**, 199–271.

Brennan, S.E., Friedman, M.W., & Pollard, C.J. 1987. A centering approach to pronouns. *Pages 155–162 of: Proceedings of the 25st Annual Meeting of the Association for Computational Linguistics (ACL'87).*

Bresnan, J. 1982. *The mental representation of grammatical relations.* Cambridge, Mass: MIT Press.

Brill, E. 1992. A Simple Rule-based Part of Speech Tagger. *Pages 152–155 of: Proceedings of the Third Conference on Applied Natural Language Processing (ANLP'92).*

Brown, G., & Yule, G. 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.

Brown, P., Cocke, J., Pietra, S.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., & Roossin, P.S. 1990. A statistical approach to machine translation. *Computational Linguistics*, **16**, 79–85.

Canals, R., Garrido, A., Guardiola, M., Iturraspe, A., Montserrat, S., Pastor, H., & Forcada, M. 2000. Herramientas para la construcción de sistemas de traducción automática: aplicación al par castellano-catalán. *In: Actas del IV Congreso de Lingüística General.*

Canals-Marote, R., Esteve-Guillén, A., Garrido-Alenda, A., Guardiola-Savall, M.I., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Pérez-Antón, P.M., & Forcada, M.L. 2001. El sistema de traducción automática castellano↔catalán interNOSTRUM. *Procesamiento del Lenguaje Natural*, **27**, 151–156.

Carbonell, J.G., & Brown, R.D. 1988. Anaphora resolution: a multi-strategy approach. *Pages 96–101 of: Proceedings of the 12th International Conference on Computational Linguistics (COLING'88).*

CEC, Commission of the European Communities (ed). 1977. *Third European Congress on Information Systems and Networks, Overcoming the language barrier.*

Chandioux, J. 1976. MÉTÉO: un système opérationnel pour la traduction automatique des bulletins météreologiques destinés au grand public. *META*, **21**, 127–133.

Chandioux, J. 1989. Météo: 100 million words later. *Pages 449–453 of:* Hammond, D.L. (ed), *American Translators Association Conference 1989: Coming of Age.* Learned Information, Medford, NJ.

Charniak, E. 1972. *Toward a model of children's story comprehension.* Cambridge, Mass: MIT Press.

Chen, H.H. 1992. The transfer of anaphors in translation. *Literary and Linguistic Computing*, **7**(4), 231–238.

Chevalier, M., Dansereau, J., & Poulin, G. 1978. *TAUM-MÉTÉO: Description du système.* Tech. rept. TAUM, Université de

Montréal.

Chomsky, N. 1981. *Lectures on Government and Binding.* Dordrecht: Foris Publications.

Colmerauer, A. 1970. *Les systèmes Q, ou un formalisme pour analyser et synthétiser des phrases sur ordinateur.* Tech. rept. 43. TAUM, Université de Montréal.

Connolly, D., Burger, J., & Day, D. 1994. A Machine learning approach to anaphoric reference. *Pages 255–261 of: Proceedings of the International Conference on New Methods in Language Processing (NEMLAP'94).*

Covington, M.A. 1994. *Natural Language Processing for Prolog Programmers.* New Jersey: Prentice-Hall.

Cucchiarini, C., van Hoorde, J., & D'Halleweyn, E. 2000. NL-Translex: machine translation for Dutch. *In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000).*

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. 1992. A Practical Part-of-Speech Tagger. *Pages 133–140 of: Proceedings of the Third Conference on Applied Natural Language Processing (ANLP'92).*

Czuba, K., Mitamura, T., & Nyberg, E. 1998. Can practical interlinguas be used for difficult analysis problems? *Pages 13–21 of: Proceedings of the Second AMTA/SIG-IL Workshop on Interlinguas.*

Dagan, I., & Itai, A. 1990. Automatic processing of large corpora for the resolution of anaphora references. *Pages 1–3 (Vol. III) of: Proceedings of the 13th International Conference on Computational Linguistics (COLING'90).*

Díaz-Ilarraza, A., Mayor, A., & Sarasola, K. 2000a. Building a Lexicon for an English-Basque Machine Translation System from Heterogeneous Wide-coverage dictionaries. *Pages 2.1–9 of: Proceedings of the Machine Translation and multilingual applications in the new millennium (MT'2000).*

Díaz-Ilarraza, A., Mayor, A., & Sarasola, K. 2000b. Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual Machine Translation System. *Pages 12.1–9*

of: Proceedings of the Machine Translation and multilingual applications in the new millennium (MT'2000).

Díaz-Ilarraza, A., Mayor, A., & Sarasola, K. 2001. Inclusión del par castellano-euskara en un prototipo de traducción automática multilingüe. *Pages 107–111 of: Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies (SLPLT-2).*

Denber, M. 1998. *Automatic Resolution of Anaphora in English.* Eastman Kodak Co., Imaging Science Division.

Dorr, B.J., Martí, M.A., & Castellón, I. 1997. Spanish EuroWordNet and LCS-Based Interlingual MT. *Pages 19–32 of: Proceedings of the First Workshop on Interlinguas in MT, MT Summit.*

Dowty, D.R., Wall, R.E., & Peters, S. 1981. *Introduction to Montague semantics.* Dordrecht: Reidel.

Evans, R. 2000. A Comparison of Rule-based and Machine Learning Methods for Identifying Non-nominal It. *Pages 233–241 of:* Christodoulakis, D.N. (ed), *Natural Language Processing - NLP'2000.* Lecture Notes in Artificial Intelligence, vol. 1835. Patras, Greece: Springer-Verlag.

Farwell, D., & Helmreich, S. 2000. An interlingual-based approach to reference resolution. *Pages 1–11 of: Proceedings of the Third AMTA/SIG-IL Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP (ANLP/NAACL'2000).*

Farwell, D., & Wilks, Y. 1991. Ultra: a multilingual machine translator. *Pages 19–24 of: Proceedings of Machine Translation Summit III.*

Ferrández, A. 1998. *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos.* Ph.D. thesis, Universidad de Alicante.

Ferrández, A., & Peral, J. 2000. A computational approach to zero-pronouns in Spanish. *Pages 166–172 of: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*

Ferrández, A., Palomar, M., & Moreno, L. 1997a. Slot Unification Grammar. *Pages 523–532 of: Proceedings of the Joint Conference on Declarative Programming. APPIA-GULP-PRODE.*

Ferrández, A., Moreno, L., Palomar, M., & Peral, J. 1997b. Un método de resolución de la anáfora discursiva mediante la unificación. *Pages 805–814 of: Actas de la Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'97).*

Ferrández, A., Palomar, M., & Moreno, L. 1998a. Anaphora resolution in unrestricted texts with partial parsing. *Pages 385–391 of: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Ferrández, A., Moreno, L., Palomar, M., & Peral, J. 1998b. Un método de resolución de la anáfora discursiva en textos no restringidos mediante la unificación. *Revista Iberoamericana de Inteligencia Artificial*, **4**(1), 20–29.

Ferrández, A., Palomar, M., & Moreno, L. 1999a. An empirical approach to Spanish anaphora resolution. *Machine Translation*, **14**(3/4), 191–216.

Ferrández, A., Palomar, M., Moreno, L., Martínez-Barco, P., Peral, J., Muñoz, R., & Saiz-Noeda, M. 1999b. Demostración del "Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística". *Procesamiento del Lenguaje Natural*, **25**, 217–218.

Francis, W.N. 1964. *A standard sample of present-day English for use with digital computers.* Tech. rept. U.S Office of Education on Cooperative Research Project No. E–007. Brown University, Providence.

Francis, W.N., & Kucera, H. 1982. *Frequency analysis of English usage. Lexicon and grammar.* Boston: Houghton Mifflin.

Garrido, A., Iturraspe, A., Montserrat, S., Pastor, H., & Forcada, M. 1999. A compiler for morphological analysers and generators based on finite-state transducers. *Procesamiento del Lenguaje Natural*, **25**, 93–98.

Goodman, K. 1989. Special Issues on Knowledge-Based Machine Translation, Parts I and II. *Machine Translation*, **4**(1/2).

Grosz, B., Joshi, A., & Weinstein, S. 1983. Providing a unified account of definite noun phrases in discourse. *Pages 44–50 of: Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL'83).*

Grosz, B., Joshi, A., & Weinstein, S. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203–225.

Gruber, J. 1965. *Studies in Lexical Relations.* Ph.D. thesis, MIT.

Haegeman, L. 1991. *Introduction to Government and Binding Theory.* Cambridge, Mass: Basil Blackwell.

Halliday, M., & Hasan, R. 1976. *Cohesion in English.* Longman English Language Series 9. London: Longman.

Hausser, R. 1999. *Foundations of Computational Linguistics.* Berlin Heidelberg New York: Springer-Verlag.

Hirst, G. 1981. *Anaphora in Natural Language Understanding.* Berlin: Springer-Verlag.

Hobbs, J. 1978. Resolving pronoun references. *Lingua*, **44**, 311–338. Reprinted in (Hobbs, 1986).

Hobbs, J. 1986. Resolving pronoun references. *In:* B.L. Webber, B. Grosz, & Spark-Jones, K. (eds), *Readings in Natural Language Processing.* Palo Alto, California: Morgan Kaufmann.

Huang, X. 1988. Semantic analysis in XTRA, an English–Chinese machine translation system. *Computers and Traslation*, **3**, 101–120.

Hutchins, W.J. 1986. *Machine translation: past, present, future.* Chichester: Ellis Horwood.

Hutchins, W.J., & Somers, H.L. 1992. *An Introduction to Machine Translation.* London: Academic Press Limited.

Isabelle, P., Dymetman, M., & Macklovitch, E. 1988. CRITTER: a translation system for agricultural market reports. *Pages 261–266 of: Proceedings of the 12th International Conference on Computational Linguistics (COLING'88).*

Jackendoff, R. 1972. *Semantic Interpretation in Generative Grammar.* Cambridge, Mass: MIT Press.

Jakobson, R. 1985. *Ensayos de lingüística general.* Obras maestras del pensamiento contemporáneo. Barcelona: Planeta-Agostini.

Jones, D., & Tsujii, J. 1990. High quality machine-driven text translation. *Pages 43–46 of: Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'90).*

Kennedy, C., & Boguraev, B. 1996. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. *Pages 113–118 of: Proceedings of the 16th International Conference on Computational Linguistics (COLING'96).*

Landes, S., Leacock, C., & Tengi, R. 1998. Building Semantic Concordances. *Pages 199–216 of:* Fellbaum, C. (ed), *WordNet: An Electronic Lexical Database.* Cambridge, Mass: MIT Press.

Landsbergen, J. 1987a. Isomorphic grammars and their use in the ROSETTA translation system. *Pages 351–372 of:* King, M. (ed), *Machine translation today: the state of the art.* Edinburgh Information Technology Series 2. Edinburgh University Press.

Landsbergen, J. 1987b. Montague grammar and machine translation. *Pages 113–147 of:* Whitelock, P.J., Wood, M.M., Somers, H.L., Johnson, R., & Bennet, P. (eds), *Linguistic theory and computer applications.* London: Academic Press.

Lappin, S., & Leass, H.J. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

Locke, W.N., & Booth, A.D. 1955. *Machine translation of languages.* Cambridge, Mass: MIT Press.

Luckhardt, H.-D. 1982. SUSY: capabilities and range of application. *Multilingua*, **1**, 213–220.

Lyons, J. 1977. *Semantics.* Vol. 2. Cambridge University Press.

Maas, H.D. 1977. The Saarbrücken automatic translation system (SUSY). *In:* (CEC, 1977).

Maas, H.D. 1987. The MT system SUSY. *Pages 209–246 of:* King, M. (ed), *Machine translation today: the state of the*

art. Edinburgh Information Technology Series 2. Edinburgh University Press.

Mahesh, K., & Nirenburg, S. 1995a. A situated ontology for practical NLP. *In: Proceedings of Workshop on basic ontological issues in knowledge sharing (IJCAI'95).*

Mahesh, K., & Nirenburg, S. 1995b. Semantic classification for practical Natural Language Processing. *Pages 79–94 of: Proceedings of the Sixth ASIS SIG/CR Classification Research Workshop: An interdisciplinary meeting.*

Maier, E., & Glashan, S.M. 1994. *Semantic and dialog processing in the VERBMOBIL spoken dialog translation system.* Tech. rept. 51. DFKI Gmbh, Germany.

Martí, M.A., Rodríguez, H., & Serrano, J. 1998. *Declaración de categorías morfosintácticas.* Proyecto ITEM. Doc. núm. 2. http://sensei.ieec.uned.es/item (página visitada el 17/04/01).

Martínez-Barco, P. 2001. *Resolución computacional de la anáfora en diálogos: estructura del discurso y conocimiento lingüístico.* Ph.D. thesis, Universidad de Alicante.

Martínez-Barco, P., Peral, J., Ferrández, A., Moreno, L., & Palomar, M. 1998. Analizador Parcial SUPP. *Pages 329–342 of: Proceedings of the VI Biennial Iberoamerican Conference on Artificial Intelligence (IBERAMIA'98).*

Maxwell, D., & Schubert, K. 1989. *Metataxis in practice: dependency syntax for multilingual machine translation.* Distributed Language Translation 6. Dordrecht: Foris.

Mitamura, T., Nyberg, E., & Carbonell, J. 1991. An efficient interlingua translation system for multi-lingual document production. *In: Proceedings of Machine Translation Summit III.*

Mitkov, R. 1994. An integrated model for anaphora resolution. *Pages 1170–1176 of: Proceedings of the 15th International Conference on Computational Linguistics (COLING'94).*

Mitkov, R. 1995a. An uncertainty reasoning approach to anaphora resolution. *Pages 149–154 of: Proceedings of the Natural Language Pacific Rim Symposium (NLPRS'95).*

Mitkov, R. 1995b. Two engines are better than one: generating more power and confidence in the search for the antecedent.

Pages 225–234 of: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'95).

Mitkov, R. 1996. Anaphor resolution: a combination of linguistic and statistical approaches. *Pages 76–85 of: Proceedings of the Discurse Anaphora and Anaphor Resolution Colloquium (DAARC'96).*

Mitkov, R. 2001. *Handbook of Computational Linguistics.* Oxford: Oxford University Press. To appear.

Mitkov, R., & Schmidt, P. 1998. On the complexity of pronominal anaphora resolution in Machine Translation. *In:* Martín-Vide, C. (ed), *Mathematical and computational analysis of natural language.* Amsterdam: John Benjamins Publishers.

Mitkov, R., & Stys, M. 1997. Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish. *Pages 74–81 of: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'97).*

Mitkov, R., Kim, H.K., Lee, H.K., & Choi, K.S. 1994. Lexical transfer and resolution of pronominal anaphors in Machine Translation: the English-to-Korean case. *Procesamiento del Lenguaje Natural,* **15**, 23–37 (Grupo 2. Traducción Automática e Interfaces).

Mitkov, R., Choi, S.K., & Sharp, R. 1995. Anaphora resolution in Machine Translation. *Pages 87–95 of: Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95).*

Montoyo, A., & Palomar, M. 2000a. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. *Pages 103–108 of: Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'2000).*

Montoyo, A., & Palomar, M. 2000b. WSD algorithm applied to a NLP system. *Pages 54–65 of: Proceedings of the 5th International Conference on Application of Natural Language to Information Systems (NLDB'2000).* Lecture Notes in Computer Science, vol. 1959. Versailles, France: Springer-Verlag.

Moreno, L., Andrés, F., & Palomar, M. 1991. Incorporar Restricciones Semánticas en el Análisis Sintáctico: IRSAS. *Procesamiento del Lenguaje Natural*, **11**, 75–88.

Moreno, L., Palomar, M., Molina, A., & Ferrández, A. 1999. *Introducción al procesamiento del lenguaje natural.* Alicante: Universidad de Alicante.

Muñoz, R. 2001. *Tratamiento y resolución de las descripciones definidas y su aplicación en sistemas de extracción de información.* Ph.D. thesis, Universidad de Alicante.

Muñoz, R., Palomar, M., & Ferrández, A. 2000. Processing of Spanish Definite Descriptions. *Pages 526–537 of:* Cairo, O., Sucar, L.E., & Cantu, F.J. (eds), *MICAI 2000: Advances in Artificial Intelligence.* Lecture Notes in Artificial Intelligence, vol. 1793. Acapulco, México: Springer-Verlag.

Nakaiwa, H., & Ikehara, S. 1992. Zero pronoun resolution in a Japanese-to-English Machine Translation system by using verbal semantic attributes. *Pages 201–208 of: Proceedings of the Third Conference on Applied Natural Language Processing (ANLP'92).*

Nakaiwa, H., & Shirai, S. 1996. Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. *Pages 812–817 of: Proceedings of the 16th International Conference on Computational Linguistics (COLING'96).*

Nasukawa, T. 1994. Robust method of pronoun resolution using full-text information. *Pages 1157–1163 of: Proceedings of the 15th International Conference on Computational Linguistics (COLING'94).*

Nirenburg, S. 1989. Knowledge-based machine translation. *Machine Translation*, **4**, 5–24.

Nyberg, E., Mitamura, T., & Kamprath, C. 1998. The KANT translation system: from R&D to large-scale deployment. *LISA Newsletter*, **2**(1), 1–7.

Paice, C.D., & Husk, G.D. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it". *Computer Speech and Language*, **2**, 109–132.

Palomar, M., Ferrández, A., Moreno, L., Saiz-Noeda, M., Muñoz, R., Martínez-Barco, P., Peral, J., & Navarro, B. 1999. A

Robust Partial Parsing Strategy based on the Slot Unification Grammars. *Pages 263–272 of: Proceedings of the Sixth Conference on Natural Language Processing (TALN'99).*

Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., & Muñoz, R. 2001. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, **27**(4), 545–567.

Papegaaij, B.C., Sadler, V., & Witkam, A.P.M. 1986. *Word expert semantics: an interlingual knowledge-based approach.* Distributed Language Translation 1. Dordrecht: Foris.

Peral, J. 1999. Proposal of an English-Spanish interlingual mechanism focused on pronominal anaphora resolution and generation in Machine Translation systems. *Pages 169–179 of: Proceedings of the Student Session of the 11th European Summer School in Logic, Language and Information (ESSLLI'99).*

Peral, J. 2001. El problema de la anáfora: resolución, aplicaciones y líneas futuras. *Pages 159–163 of: Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies (SLPLT-2).*

Peral, J., & Ferrández, A. 2000a. An application of the Interlingua System ISS for Spanish-English pronominal anaphora generation. *Pages 42–51 of: Proceedings of the Third AMTA/SIG-IL Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP (ANLP/NAACL'2000).*

Peral, J., & Ferrández, A. 2000b. Generation of Spanish zero-pronouns into English. *Pages 252–260 of:* Christodoulakis, D.N. (ed), *Natural Language Processing - NLP'2000.* Lecture Notes in Artificial Intelligence, vol. 1835. Patras, Greece: Springer-Verlag.

Peral, J., Martínez-Barco, P., Ferrández, A., & Navarro, B. 1998. Sistema de adquisición automática de reglas gramaticales. *Procesamiento del Lenguaje Natural*, **23**, 110–117.

Peral, J., Saiz-Noeda, M., Ferrández, A., & Palomar, M. 1999a. Anaphora resolution and generation in a multilingual system. Interlingua mechanism. *Pages 315–324 of: Proceedings of Venezia per il Trattamento Automatico delle Lingue (VEXTAL'99).*

Peral, J., Palomar, M., & Ferrández, A. 1999b. Coreference-oriented Interlingual Slot Structure and Machine Translation. *Pages 69–76 of: Proceedings of the ACL Workshop Coreference and its Applications.*

Pereira, F., & Warren, D. 1980. Definite Clause Grammars for Language Analysis - A survey of the Formalism and a comparison with augmented transition networks. *Artificial Intelligence*, **13**(3), 231–278.

Pla, F. 2000. *Etiquetado léxico y análisis sintáctico superficial basado en modelos estadísticos.* Ph.D. thesis, Universidad Politécnica de Valencia.

Pla, F., Molina, A., & Prieto, N. 2000. Tagging and Chunking with Bigrams. *In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000).*

Preuss, S., Schmitz, B., Hauenschild, C., & Umbach, C. 1994. Anaphora resolution in Machine Translation. *Pages 29–52 of:* Ramm, W. (ed), *Studies in Machine Translation and Natural Language Processing*, vol. 6 "Text and content in Machine Translation: Aspects of discourse representation and discourse processing". Luxembourg: Office for Official Publications of the European Community.

Proyecto CRATER. 1994-1995. *Corpus Resources And Terminology ExtRaction.* MLAP-93/20. http://www.lllf.uam.es/proyectos/crater.html (página visitada el 17/04/01).

Proyecto ITEM. 1996-1999. *Recuperación de Información Textual en un Entorno Multilingüe con Técnicas de Lenguaje Natural.* Comisión Interministerial de Ciencia y Tecnología TIC96-1243-C03. http://sensei.ieec.uned.es/item (página visitada el 17/04/01).

Quesada, J.F., & Amores, J.G. 2000. *Diseño e implementación de sistemas de Traducción Automática.* Sevilla: Universidad de Sevilla. Secretariado de publicaciones.

Reinhart, T. 1983. *Anaphora and Semantic Interpretation.* Croom Helm linguistics series. Beckenham, Kent: Croom Helm Ltd.

Reuther, U. 1998. Controlling Language in an Industrial Application. *In: Proceedings of the Second International Workshop on Controlled Language Applications (CLAW'98).*

Rich, E., & LuperFoy, S. 1988. An architecture for anaphora resolution. *Pages 18–24 of: Proceedings of the Second Conference on Applied Natural Language Processing (ANLP'88).*

Rico, C. 1994. *Aproximación estadístico-algebraica al problema de la resolución de la anáfora en el discurso.* Ph.D. thesis, Universidad de Alicante.

Rocha, M. 1999. Coreference resolution in dialogues in English and Portuguese. *Pages 53–60 of: Proceedings of the ACL Workshop Coreference and its Applications.*

Sadler, V. 1989. *Working with analogical semantics: disambiguation techniques in DLT.* Distributed Language Translation 5. Dordrecht: Foris.

Saggion, H., & Carvalho, A. 1994. Anaphora resolution in a Machine Translation system. *Pages 4.1–14 of: Proceedings of the International Conference "Machine Translation, 10 years on".*

Saiz-Noeda, M., Suárez, A., & Peral, J. 1999. Propuesta de incorporación de información semántica desde Wordnet al análisis sintáctico parcial orientado a la resolución de la anáfora. *Procesamiento del Lenguaje Natural*, **25**, 167–173.

Saiz-Noeda, M., Peral, J., & Suárez, A. 2000. Semantic compatibility techniques for anaphora resolution. *Pages 43–48 of: Proceedings of the International Conference on Artificial and Computational Intelligence For Decision, Control and Automation In Engineering and Industrial Applications (ACIDCA'2000).*

Saiz-Noeda, M., Suárez, A., & Palomar, M. 2001. Semantic pattern learning through Maximum Entropy-based WSD technique. *Pages 23–29 of: Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL'2001).*

Samuelson-Brown, G. 1996. New technology for translators. *In:* Owens, R. (ed), *The translator's handbook.* London: Aslib.

Samuelsson, C., & Voutilainen, A. 1997. Comparing a Linguistic and a Stochastic Tagger. *Pages 246–253 of: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European*

Chapter of the Association for Computational Linguistics (ACL/EACL'97).

Sato, S., & Nagao, M. 1990. Toward memory-based translation. *Pages 247–252 of: Proceedings of the 13th International Conference on Computational Linguistics (COLING'90).*

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. *Pages 44–49 of: Proceedings of the International Conference on New Methods in Language Processing (NEMLAP'94).*

Schubert, K. 1986. Linguistic and extra-linguistic knowledge. *Computers and Translation*, **1**, 125–152.

Schubert, K. 1987. *Metataxis: contrastive dependency syntax for machine translation.* Distributed Language Translation 2. Dordrecht: Foris.

Schubert, K. 1988. The architecture of DLT – interlingual or double direct? *Pages 131–144 of:* Maxwell, D., Schubert, K., & Witkam, T. (eds), *New directions in machine translation.* Distributed Language Translation 4. Dordrecht: Foris.

Sharp, R. 1988. CAT2 – Implementing a formalism for multilingual MT. *In: Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'88).*

Shieber, S.M. 1986. An introduction to unification-based approaches to grammar. CSLI-Lecture Notes, vol. 4. Stanford, CA, Center for the Study of Language and Information.

Sigurd, B. 1988. Translating to and from Swedish by SWETRA – a multilanguage translation system. *Pages 205–217 of:* Maxwell, D., Schubert, K., & Witkam, T. (eds), *New directions in machine translation.* Distributed Language Translation 4. Dordrecht: Foris.

Slocum, J. 1987. METAL: the LRC machine translation system. *Pages 319–350 of:* King, M. (ed), *Machine translation today: the state of the art.* Edinburgh Information Technology Series 2. Edinburgh University Press.

Slocum, J. 1988. *Machine translation systems.* Cambridge University Press.

Somers, H.L., Tsujii, J., & Jones, D. 1990. Machine Translation without a source text. *Pages 271–276 of: Proceedings of the 13th International Conference on Computational Linguistics (COLING'90).*

Strube, M. 1998. Never Look Back: An Alternative to Centering. *Pages 1251–1257 of: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Strube, M., & Hahn, U. 1999. Functional Centering - Grounding Referential Coherence in Information Structure. *Computational Linguistics*, **25**(5), 309–344.

Stuckardt, R. 1996. Anaphor resolution and the scope of syntactic constraints. *Pages 937–942 of: Proceedings of the 16th International Conference on Computational Linguistics (COLING'96).*

Stuckardt, R. 1997. Resolving anaphoric references on deficient syntactic descriptions. *Pages 30–37 of: Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution.*

Sumita, E., Iida, H., & Kohyama, H. 1990. Translating with examples: a new approach to machine translation. *Pages 203–212 of: Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'90).*

Thurmair, G. 1990. Complex lexical transfer in METAL. *Pages 91–107 of: Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'90).*

Toma, P. 1977. Systran as a multilingual machine translation system. *In:* (CEC, 1977).

Trujillo, A. 2000. Estrategias de traducción automática. *Quark, Ciencia, Medicina, Comunicación y Cultura*, **19**.

van Noord, G., Dorrepaal, J., van der Eijk, P., Florenza, M., & des Tombe, L. 1990. The MiMo2 research system. *Pages 213–233 of: Proceedings of the Third International Confe-*

rence on Theoretical and Methodological Issues in Machine Translation (TMI'90).

van Slype, G., & Pigott, I. 1979. Description du système de traduction automatique Systran de la Commission des Communautés Européennes. *Documentaliste*, **16**, 150–159.

Varile, G.B., & Lau, P. 1988. Eurotra: practical experience with a multilingual machine translation system under development. *Pages 160–167 of: Proceedings of the Second Conference on Applied Natural Language Processing (ANLP'88).*

Vauquois, B. 1968. A survey of formal grammars and algorithms for recognition and transformation in machine translation. *Pages 254–260 of: Proceedings of IFIP'68.*

Vauquois, B., & Boitet, C. 1985. Automated translation at Grenoble University. *Computational Linguistics*, **11**, 28–36. Reprinted in (Slocum, 1988), 85–110.

Vicedo, J.L., & Ferrández, A. 2000. Importance of pronominal anaphora resolution in Question Answering systems. *Pages 555–562 of: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*

Vicedo, J.L., Ferrández, A., & Peral, J. 2000. ¿Cómo influye la resolución de la anáfora pronominal en los Sistemas de Búsqueda de Respuestas? *Procesamiento del Lenguaje Natural*, **26**, 231–237.

Vossen, P. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, **3**(1), 7–12.

Wada, H. 1990. Discourse processing in MT: problems in pronominal translation. *Pages 73–75 of: Proceedings of the 13th International Conference on Computational Linguistics (COLING'90).*

Webber, B.L. 1978. *A formal approach to discourse anaphora.* Ph.D. thesis, Division of Applied Mathematics, Harvard University, Cambridge, MA.

Weizenbaum, J. 1966. ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, **9**, 36–45.

Wheeler, P.J. 1987. SYSTRAN. *Pages 192–208 of:* King, M. (ed), *Machine translation today: the state of the art.* Edinburgh Information Technology Series 2. Edinburgh University Press.

Whitelock, P.J., & Kilby, K.J. 1983. *Linguistic and computational techniques in machine translation system design.* Tech. rept. 84/2. CCL/UMIST, Centre for Computational Linguistics, UMIST, Manchester.

Whitelock, P.J., Wood, M.M., Chandler, B.J., Holden, N., & Horsfall, H.J. 1986. Strategies for interactive machine translation: the experience and implications of the UMIST Japanese project. *Pages 329–334 of: Proceedings of the 11th International Conference on Computational Linguistics (COLING'86).*

Wilks, Y. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, **6**(1), 53–74.

Winograd, T. 1972. *Understanding Natural Language.* New York: Academic Press.

Winograd, T. 1986. A procedural model of language understanding. *In:* B. J. Grosz, K. Sparck-Jones, B. L. Webber (ed), *Readings in Natural Language.* Los Altos, California: Morgan Kaufman.

Witkam, A.P.M. 1983. *Distributed language translation: feasibility study of multilingual facility for videotex information networks.* Utrecht: BSO.

Wood, M.M., & Chandler, B.J. 1988. Machine translation for monolinguals. *Pages 760–763 of: Proceedings of the 12th International Conference on Computational Linguistics (COLING'88).*

Woods, W. 1977. Lunar rocks in natural English: Explorations in natural language question answering. *In:* Zampolli, A. (ed), *Linguistics Structures Processing.* New York: Elsevier.

Volver al índice/Tornar a l'índex