Tratamiento y resolución de las descripciones definidas y su aplicación en sistemas de extracción de información

Tesis Doctoral

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Autor: **Rafael Muñoz Guillena**

Director: **Dr. Manuel Palomar Sanz**

Alicante, 23 de marzo de 2001

Agradecimientos

Quiero agradecer a todos aquellos que han contribuido a la finalización de este trabajo.

En primer lugar quiero destacar a Manuel Palomar como director de esta Tesis. Sus consejos, ideas y directrices me han facilitado mi trayectoria investigadora y la realización de esta Tesis. No sólo se ha limitado a realizar una labor de dirección sino que también ha sabido ser un amigo que me ha alentado y ayudado en todos los momentos.

También quiero agradecer a Lidia Moreno y a Antonio Ferrández los comentarios realizados y los consejos dados a lo largo de mi trayectoria investigadora, que me han servido para depurar los trabajos realizados. También quiero destacar la incalculable ayuda que ha supuesto poder usar el sistema de Procesamiento de Lenguaje Natural orientado a la resolución de la anáfora pronominal desarrollado por Antonio Ferrández, donde he integrado y probado los algoritmos que he desarrollado.

A Armando Suárez y Maximiliano Saiz por la realización de las librerías necesarias para el acceso a la información presente en el WordNet español y su integración en mi programa en PROLOG.

A mis compañeros de asignaturas, Andrés Montoyo y Jaime Gómez. En especial a Andrés, que nunca se ha quejado de las muchas veces que me ha sustituído en clase para poder asistir a los congresos.

A mis compañeros de despacho Patricio Martínez, Jesús Peral y Maximiliano Saiz que tantas horas hemos compartido y que mutuamente nos hemos animado para proseguir en esta larga y ardua tarea.

Al resto de mis compañeros del Grupo de Procesamiento del Lenguaje y Sistemas de Información Rafa Romero, Fernando Llopis, Juan Carlos Trujillo, Cristina Cachero, Jose Luis Vicedo, Borja Navarro, Sergio Luján y Paloma Moreda, por el apoyo que me han dado y en muchas ocasiones la ayuda desinteresada en la realización de diferentes tareas que me ha permitido centrarme más en la realización de este trabajo.

A todos mis compañeros del Departamento de Lenguajes y Sistemas Informáticos que me han alentado en la realización de mi labor investigadora y docente.

A mis compañeros del Grupo de Procesamiento de Lenguaje Natural de la Universidad Politécnica de Valencia Ferrán Pla, Antonio Molina, Natividad Prieto, Emilio Sanchís y Encarna Segarra con los que he compartido proyectos y trabajos que me han permitido complementar conocimientos con trabajos relacionados con otras áreas. Además, quiero destacar la ventaja que ha supuesto poder contar con el etiquetador léxico desarrollado por Ferrán Pla para etiquetar los textos utilizados como corpus en este trabajo.

Por último, pero no menos importante, quiero agradecer a mi familia y amigos. A mis padres que siempre me han apoyado y me han transmitido el tesón y fuerza necesaria para la finalización de este trabajo. A mi hermana y mi cuñado, cuyos consejos en muchas ocasiones me han servido para depurar esta Tesis. A mis sobrinas, María y Elena, por los buenos momentos. A mi tía Isabel que es como mi segunda madre, siempre se preocupa por mis tareas. Y en general, quiero agradecer a mis amigos que siempre han sabido comprender que en ocasiones no disponía de tiempo para poder atenderles y me han animado a finalizar este trabajo.

Alicante, febrero 2001 *Rafael Muñoz*

Índice General

Índice de Tablas

Índice de Figuras

1. Introducción

El uso y compresión de los lenguajes naturales de forma automática es una de las áreas en la que más esfuerzos se han invertido por parte de los investigadores en los últimos años. En esta área, denominada *Lingüística Computacional*, y también conocida como *Procesamiento del Lenguaje Natural* (PLN), se estudian los diferentes problemas que genera el lenguaje en su tratamiento automático, tanto en conversaciones habladas como escritas.

Tal y como recoge Moreno et *al.* (1999), todo sistema de Procesamiento de Lenguaje Natural intenta simular un comportamiento lingüístico humano. Para ello, debe tomar conciencia tanto de las estructuras propias del lenguaje, como del conocimiento general acerca del universo del discurso. De esta forma, una persona que participa en un diálogo sabe como combinar las palabras para formar una oración, conoce los significados de las mismas, sabe cómo éstos afectan al significado global de la oración y posee un conocimiento del mundo en general que le permite participar de la conversación. Estas informaciones son las que los sistemas deben combinar para poder comprender y usar el lenguaje natural como medio de comunicación (entrada, procesamiento y salida de datos).

Uno de los avances que más ha impulsado las investigaciones dentro del Procesamiento del Lenguaje Natural es la aparición y/o evolución de Internet, ya que el número de textos (páginas HTML) a tratar es inmanejable y son necesarias aplicaciones que hagan más efectivas las búsquedas, las recuperaciones y/o las extracciones de información; además de la posibilidad de generar eficientes resúmenes de texto y bajo el condicionante de que que nos encontramos en una sociedad multilingüe, todo ello con el

objetivo de facilitar un adecuado y fácil acceso a la denominada *sociedad de la información*. Para alcanzar este objetivo se destacan, entre otras, las siguientes aplicaciones:

- *Extracción de información* (EI). Los sistemas de extracción de información procesan textos que contienen una información relevante de forma no estructurada y proporcionan dicha información de forma estructurada a través de unas plantillas definidas previamente. La información que no es definida como relevante se desecha.

- *Recuperación de información* (RI). Los sistemas de recuperación de información procesan colecciones de textos que pueden contener o no la información relevante y proporcionan todos aquellos textos en los que se encuentra dicha información. Muchos autores consideran la extracción de información como un proceso posterior a la recuperación de información.

- *Traducción automática* (TA). Los sistemas de traducción automática traducen textos escritos en un lenguaje origen a un lenguaje destino. En una sociedad multilingüe como hoy vivimos, sistemas de estas características son totalmente necesarios.

- *Sistemas de búsquedas de respuestas (QA, question answering)*. Los sistemas de búsqueda de respuestas son una combinación entre la recuperación de información y la extracción de información, ya que ante una pregunta concreta, los sistema de respuestas proporcionan las oraciones donde se encuentra la solución a la pregunta, previo procesamiento de una colección de textos en los que puede o no estar la respuesta buscada.

- *Generación de resúmenes (summarization)*. Los sistemas de generación de resúmenes procesan un texto de forma que proporcionan una sinopsis de todo el texto procesado.

Desde el punto de vista del tratamiento automático del Lenguaje Natural (LN), estas aplicaciones tienen en común una serie de tareas, herramientas y módulos de preprocesamiento, además

de compartir diversos recursos lingüísticos que permitirán manejar, procesar y comprender los diferentes tipos de textos.

En esta memoria, se toma como marco global de actuación la *extracción de información*. Por ello, a continuación, se muestran las principales características y necesidades de un sistema de EI desde el punto de vista del PLN.

Los sistemas de extracción de información procesan textos que contienen información denominada relevante en su dominio de trabajo y la extraen mediante la utilización de técnicas de PLN. En un sistema de EI distinguimos tres fases o etapas principales: (1) fases de preproceso, (2) fases de análisis lingüístico y (3) fases dependiente del dominio.

1. *Preproceso del texto.*
   Antes de efectuar cualquier tipo de análisis sobre un texto se debe procesar para realizar las siguientes tareas:

   - Eliminación de información irrelevante. Los textos utilizados, textos electrónicos, incorporan una serie de etiquetas o marcas que pueden afectar al correcto funcionamiento de módulos posteriores. Este tipo de etiquetas son etiquetas de formato tipo HTML (*HyperText Markup Language*), XML (*eXtended Markup Language*) o simplemente marcas que usan los procesadores de textos que crearon estos textos para indicar diferentes tipos de formatos (cursiva, negrilla, cambio de línea, etc.)

   - División del texto. La mayoría de los módulos que trabajan en cualquier sistema de PLN lo hacen a nivel de oración o de palabra. Para ello, los textos se deben segmentar en oraciones y posteriormente, éstas en palabras. Estos procesos se denominan *segmentación en oraciones* y *tokenización*, respectivamente.

2. *Fases de análisis lingüístico.*
   Distinguimos, entre otros, cuatro tipos de análisis diferentes: análisis léxico-morfológico, análisis sintáctico, análisis semántico

y análisis contextual.

- *Análisis léxico-morfológico.*
  El conocimiento léxico-morfológico es el fundamento de cualquier sistema de comprensión del Lenguaje Natural, dado que las oraciones están constituidas por palabras, que son las que incorporan información morfológica, sintáctica y semántica. Un mecanismo para obtener esta información es a través de un analizador léxico-morfológico que puede proporcionar la siguiente información:

  – Categoría gramatical. Cada palabra tiene asociada una etiqueta representativa de su categoría. Algunas de las categorías son: nombre, verbo, adjetivo, artículo, etc.

  – Información morfológica. La información morfológica asociada a cada palabra está formada por información de concordancia y por las reglas de formación de las palabras. La *información de concordancia* de cada palabra depende de la categoría gramatical, por ejemplo, la categoría nombre no tienen información del tipo persona. Entre otras, las informaciones más usadas dentro de la información de concordancia son: género, número, persona, etc. Las *reglas de formación* de las palabras permiten proporcionar el lema o raíz de una palabra.

  – Información semántica. La información semántica tiene un papel muy importante en la resolución de diferentes problemas lingüísticos. Esta información está formada, entre otras, por la categoría semántica, los rasgos semánticos y la forma lógica asociada.

  Al incorporar o asociar a cada una de las palabras del texto la información léxico-morfológica se produce el fenómeno de la *ambigüedad léxica*. Se distinguen diferentes tipos de ambigüedades léxicas:

  a) Ambigüedad léxica categorial.

b) Ambigüedad léxica pura (semántica).

La *ambigüedad léxica categorial* se presenta cuando una palabra puede desempeñar diferentes categorías sintácticas. Como por ejemplo, blanco que puede ser:

– un *adjetivo* en la oración (el color blanco....)
– un *nombre común* en la oración (Juan dio en el blanco).

Algunas de las técnicas usadas en la literatura para resolver este fenómeno se basan en *aproximaciones lingüísticas*, como por ejemplo, el sistema EnCG (Votilainen, 1988; Votilainen & Järvinen, 1995) basado en gramáticas de restricciones, y otras se fundamentan en *aproximaciones basadas en aprendizaje automático* como son los trabajos desarrollados para el español de Pla (2000) y Pla *et al.* (2000), donde se utilizan técnicas basadas en *n-gramas*.
La *ambigüedad semántica* se presenta en las palabras polisémicas que en función del contexto pueden tener un sentido u otro. Por ejemplo, la palabra banco es semánticamente ambigua, ya que puede tener varios sentidos:

– banco como entidad financiera o
– banco como mobiliario para sentarse.

La tarea que se encarga del estudio y resolución del problema de la ambigüedad semántica se conoce como *desambiguación del sentido de las palabras* (WSD, *Word Sense Disambiguation*). Los trabajos desarrollados por Montoyo y Palomar (2000), Montoyo *et al.* (2001), Rigau *et al.* (1997; 1998) muestran un adecuado tratamiento para este problema.

- *Análisis sintáctico.*
  Las palabras se combinan formando constituyentes a un nivel sintáctico superior (sintagmas nominales, sintagmas preposicionales y sintagmas verbales). El sintagma nominal (SN)

puede estar formado por un simple pronombre personal, por un nombre propio o por cualquier combinación de palabras cuyo núcleo sea un nombre. El sintagma preposicional (SP) está formado por una preposición y un SN. El sintagma verbal (SV) está formado por un simple verbo o por un verbo y varios SNs y/o SPs. Otros posibles tipos de constituyentes son las cláusulas de relativo y los sintagmas adverbiales.
La función principal del análisis sintáctico consiste en determinar si una oración es gramaticalmente correcta y proporcionar su estructura asociada que refleje las relaciones sintácticas entre cada una de las palabras que forman la oración. De acuerdo con Moreno *et al.* (1999), la información necesaria para poder realizar el análisis sintáctico puede estar representada a través de redes de transición o de gramáticas. Las *redes de transición* se basan en la aplicación de la teoría de grafos y de autómatas finitos. Mientras que una *gramática*, de acuerdo con Allen (1995), es la especificación formal de las estructuras permitidas por el lenguaje, diferenciandola del algoritmo de análisis, que es el mecanismo para determinar la estructura de la oración de acuerdo con la gramática.
Existen dos tipos de analizadores o algoritmos de análisis:

- *Completo o Global.* Los analizadores completos aceptan aquellas oraciones que son gramaticalmente correctas y rechazan las incorrectas. Es decir, el analizador sintáctico debe ser capaz de agrupar todas las palabras que forman la oración basándose en la gramática. Los analizadores completos requieren de la utilización de una gramática muy extensa formada por todas las reglas del lenguaje.

- *Parcial.* Según Abney (1997), el análisis parcial tiene como objetivo recuperar información sintáctica de forma eficiente y fiable, desde texto no restringido, sacrificando la completitud y profundidad del análisis global.

Para el español cabe destacar el TACAT desarrollado por Atserias *et al.* (1998) y *Castellón et al.* (1998). También se destaca el SUPP desarrollado por Martínez-Barco *et al.* (1998) y Palomar *et al.* (1999). Ambos analizadores, TACAT y SUPP, permiten utilizar las dos estrategias anteriores.
Uno de los fenómenos a resolver por los analizadores es la *ambigüedad estructural*. Se dice que una oración es ambigua estructuralmente cuando posee más de un árbol sintáctico de derivación. Como por ejemplo en la oración

El policía vio al ladrón con el telescopio

el SP con el telescopio puede agruparse al SP al ladrón para formar el SP complejo al ladrón con el telescopio como muestra la figura 1.1 o bien puede relacionarse con el verbo ver para formar un SP con función de complemento circunstancial de modo como muestra la figura 1.2.

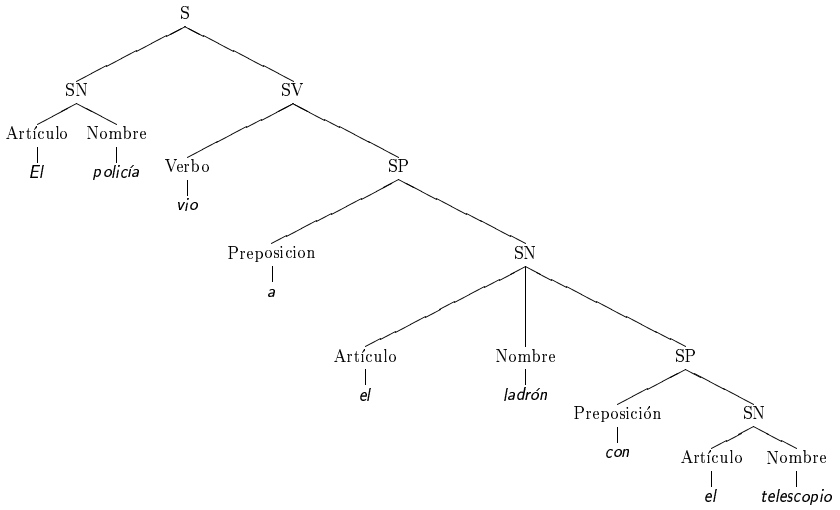

Figura 1.1. *Primer árbol de análisis de* El policía vio al ladrón con el telescopio

- *Análisis semántico.*
  A partir de la estructura generada en la fase de análisis anterior, genera otra estructura denominada *forma lógica* que

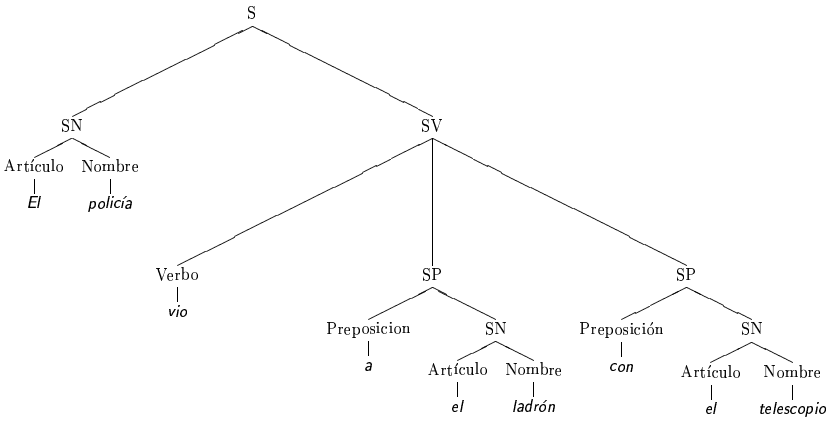

Figura 1.2. *Segundo árbol de análisis de* El policía vio al ladrón con el telescopio

representa el significado o sentido de la oración. El objetivo de este análisis es la obtención de una forma lógica independiente del contexto.

- *Análisis contextual.*
  El análisis contextual consiste en añadir a la forma lógica asociada a la oración información referente al contexto. Uno de los principales problemas a resolver en este análisis es la ambigüedad referencial, dentro de esta ambigüedad, distinguimos:

  – Elipsis. Este fenómeno consiste en la omisión de un constituyente debido a la aparición de ese constituyente en un sintagma previo. Los trabajos desarrollados por Palomar (1996) tratan de forma efectiva este problema.

  – Extraposición a izquierda. Este fenómeno aparece en las cláusulas de relativo, y ocurre cuando, en una oración, un subconstituyente de un constituyente que forma parte de la oración, se representa por otro a la izquierda del que está incompleto, tal y como se define en los trabajos de Martínez-Barco (1998) y Peral (1998).

– Ambigüedad referencial o Anáfora. La anáfora es uno de los fenómenos más comunes al hacer uso de las reglas para la combinación de palabras, del conocimiento del significado de las palabras y del conocimiento general del universo de discurso. La anáfora es la sustitución de determinadas palabras por otras, manteniendo el sentido de la oración y haciendo referencia a las primeras. La anáfora puede estar producida por diferentes tipos de palabras, las más usuales son los pronombres y las descripciones definidas (sintagmas nominales definidos introducidos por un artículo determinado o por un demostrativo).

3. *Fases dependientes del dominio.*
Una vez realizadas las fases de análisis anteriores se llevan a cabo dos tareas fundamentales para obtener el objetivo de los sistemas de EI. Estas tareas son:

- *Inferencia.* La tarea de inferencia consiste en detectar información que aparece de forma implícita en el texto mediante la aplicación de reglas específicas del contexto. Por ejemplo, en el siguiente texto que pertenece al dominio de las sucesiones de empresas:

  Steve Miller, presidente de la Transalp Corporation, ... John Grant releva del cargo a Steve Miller en Transalp Corporation.

  aparece de forma implícita que John Grant es el nuevo presidente. La información implícita en un texto puede aparecer en la misma oración o dispersa por el texto. Esta información puede ser reconocida a través de la aplicación de reglas del siguiente estilo:

  <pers1> + <verbo:[suceder, relevar, ...]> + <pers2>
  $\Rightarrow$ <pers1.atrib(cargo)>=<pers2.atrib(cargo)>
  $\wedge$ <pers2.atrib(cargo)>=$\emptyset$

- *Relleno de plantillas.* El relleno de plantillas consiste en asociar cada atributo de las plantillas, definidas previamente para el dominio de trabajo correspondiente, con la información explícita e implícita detectada en la tarea anterior.

Las tres fases (preproceso, análisis lingüísticos y dependiente del dominio) se han presentado desde el punto de vista del PLN, con el objetivo de identificar los principales problemas a resolver en el desarrollo de una aplicación de estas características. De los problemas identificados, uno de los más significativos e importantes en la EI es la resolución de las descripciones definidas (DD) ya que sin su resolución difícilmente se podrá relacionar información explícita con información implícita. Por tanto, y como consecuencia de los trabajos realizados en la EI, esta Tesis se centra en la resolución de la ambigüedad referencial producida por las DDs, tanto en el marco de los sistemas de EI como en textos de dominios no restringidos. Consideramos y así demostramos a lo largo de este trabajo que una resolución adecuada de las DDs mejora considerablemente la calidad de un sistema de EI.

1.1 Descripciones Definidas (DD)

El fenómeno de la *referencia* ha sido estudiado desde principios de siglo por Frege (1892) y Russell (1919), entre otros, habiéndose aceptado como un fenómeno semántico. Cuando se quiere comunicar o transmitir algún tipo de información acerca del mundo se necesita relacionar expresiones lingüísticas con entidades del mundo real. A la relación entre las expresiones lingüísticas y las entidades del mundo se le denomina *referencia* y al objeto o persona del mundo, como entidad física, de la que se quiere transmitir alguna información se le denominan *referente*. Las expresiones lingüísticas usadas para realizar este tipo de referencias (a entidades físicas) suelen ser sintagmas nominales y pronombres, aunque no todos los SNs siempre funcionan como expresiones referenciales.

Al fenómeno que ocurre cuando un pronombre refiere a un referente (entidad física) que no ha sido introducido previamente en

la comunicación se le conoce con el nombre de *deíxis*, tal y como define Vicente-Mateu (1994). En el ejemplo,

Dame éstas

cuando se dice la oración señalando al referente en cuestión. El pronombre éstas hace referencia al objeto que se señala con el dedo, pero no se ha hablado previamente de dicho objeto.

Por el contrario, si la referencia en lugar de hacerla a entidades físicas del universo de discurso se realiza a un objeto lingüístico, el fenómeno que se produce se conoce como *anáfora*. Este trabajo se centra en el estudio de la referencia a referentes lingüísticos, es decir, se centra en el estudio de la anáfora. A partir de ahora cuando se hable de *referencia* se referirá a *referencias lingüísticas*.

Tanto en el lenguaje escrito como en el hablado se usan normalmente diferentes tipos de expresiones para referirse a personas, a objetos, a un evento, a un lugar o a un proceso. Las expresiones lingüísticas más usadas son los pronombres y las descripciones definidas (sintagmas nominales introducidos por un artículo determinado o por un demostrativo).

En la literatura actual se pueden encontrar diversas definiciones del fenómeno de la anáfora. Entre todas ellas destacan las siguientes:

- Rico (1994) define la anáfora como *la relación de referencia que se establece entre una forma lingüística y un objeto, persona o una situación que ya han sido mencionados de manera explícita o implícita con anterioridad durante el proceso comunicativo.*

- Hirst (1981) define la anáfora como *el mecanismo que nos 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 desabreviar la referencia y por consiguiente determinar la entidad a la que se alude.*

En ocasiones, la solución propuesta a una expresión anafórica A, es el antecedente B que a su vez es otra expresión anafórica, la cuál tiene su propio antecedente C. Entonces se establece lo que

se conoce con el nombre de *cadena de correferencia*, A correfiere con B y B correfiere con C y además, se dice que A y C también *correfieren*.

- Allen (1995) define como correferentes a *cualquier par expresión anafórica - antecedente si ambos tienen el mismo referente físico (objeto del universo del discurso).*
  Esto lleva en ocasiones a hablar indistintamente de anáfora y correferencia.

- Alcaraz y Martínez (1997) entienden por correferencia *la relación que se establece entre dos expresiones referenciales (sintagmas nominales determinados, nombres propios, pronombres), entre dos unidades, coincidentes en un enunciado, con capacidad para aludir a entidades, cuando ambas se interpretan como alusivas a la misma entidad, cuando tienen el mismo referente.*

Sobre la base de las definiciones anteriores, podemos extraer las siguientes conclusiones:

i) la expresión anafórica puede ser de diversos tipos (entre otras, pronombres, descripciones definidas, adjetivos y adverbios) y

ii) la expresión lingüística a la que hace referencia una expresión anafórica también puede tener diferentes tipos gramaticales o entidades del discurso (sintagmas nominales, sintagmas verbales, oraciones completas, párrafos, etc.)

En este trabajo, nos centramos en las descripciones definidas como expresión anafórica y en los sintagmas nominales como expresión lingüística a los que se hace referencia. Así definimos descripción definida como:

"Un sintagma nominal introducido por un artículo definido o por un demostrativo"

como por ejemplo en:

... el coche derrapó en la curva debido a una mancha de aceite. La mancha no fue limpiada ...

donde el coche, la curva y la mancha son descripciones definidas.

Por otro lado, definimos la anáfora producida por una DD como:

"la relación de referencia que se estable entre una DD y un objeto, persona o situación expresada lingüísticamente con anterioridad"

Como en el ejemplo:

[Carlos Sainz]$_i$ pilotó [su Ford]$_j$ durante toda [la prueba]$_k$ sin problemas. [El piloto]$_i$ terminó exhausto pero [el coche]$_j$ no tuvo ningún problema mecánico. [El piloto de rallies]$_i$ sigue manteniendo una buena forma y [el vehículo]$_j$ le ofrece todas sus prestaciones.

Así, el objetivo de este trabajo es el tratamiento y resolución de la anáfora producida por las descripciones definidas, tomando como marco de aplicación los sistemas de extracción de información y como hipótesis de trabajo el desarrollo de un sistema de resolución de las DDs en textos no restringidos e independiente del dominio.

1.2 Organización de la Tesis

La presente memoria se ha estructurado en 7 capítulos.

En el capitulo 2 se presenta una revisión a la literatura referente a los sistemas de extracción de información. Además, en este capítulo se presenta el sistema de extracción de información EXIT en el que se ha integrado el módulo de resolución de las DDs propuesto en este trabajo.

En el capítulo 3, por un lado se realiza un profundo estudio de las características lingüísticas de las DDs y se presentan las clasificaciones de sus diferentes tipos realizadas por diversos autores. Y por otro, se realiza una exhaustiva revisión de los diferentes métodos existentes en la literatura para la resolución de las referencias producidas por las DDs.

En el capítulo 4 se presentan las diferentes fuentes de información que intervienen en la resolución de las referencias lingüísticas, así como su influencia en el tratamiento y resolución de las DDs. En este capítulo se presenta una propuesta de clasificación de los diferentes tipos de DDs que se encuentran en textos escritos en español, en función del tipo de información necesario para establecer la relación entre el antecedente y la DD.

En el capitulo 5 se presentan dos enfoques del problema de resolución de las DDs. El primero de ellos, se fundamenta en la búsqueda del antecedente mediante el uso de un sistema de reglas heurísticas aplicadas en forma de filtro. El segundo enfoque se basa por un lado en la generación automática de una red semántica utilizando WordNet español con el doble objetivo de la identificación previa de las DDs no anafóricas y la incorporación de información semántica en el proceso de resolución; y por otro lado, en un conjunto de reglas heurísticas que utilizan un sistema de pesos para la obtención del antecedente de la DD.

En el capitulo 6 se presenta el trabajo de experimentación realizado. Para llevar a cabo este trabajo, un entrenamiento de ambos enfoques se ha realizado independiente de la evaluación final.

En el capítulo 7, se recogen las conclusiones obtenidas con el desarrollo de este trabajo, así como diferentes líneas de trabajos a desarrollar en el futuro. Finalmente, se muestran las referencias bibliográficas utilizadas en el desarrollo de esta memoria.

2. Sistemas de Extracción de Información

En este capítulo se hace una breve introducción a los sistemas de extracción de información (EI). Se recoge la evolución a lo largo de la historia de los sistemas de EI, las características de las conferencias (MUC, *Message Understanding Conference*) que evalúan y definen las tareas a realizar por cualquier sistema de EI celebradas hasta el momento, así como las características de algunos sistemas. Finalmente, se presenta el sistema EXIT y la influencia que tiene la resolución de las DDs en los resultados obtenidos por el sistema.

2.1 Introducción

La *extracción de información* (EI) es una disciplina dentro del *Procesamiento del Lenguaje Natural* (PLN) cuyo objetivo es la extracción de determinada información, que se denomina *relevante*, a partir de unos textos. No debemos confundir esta disciplina con otra disciplina que también está dentro del Procesamiento del Lenguaje Natural como es la *recuperación de información* (RI). El objetivo de la RI es la obtención de aquellos textos completos en los que aparece la información relevante a partir de una colección de textos. La principal diferencia entre la EI y la RI radica en que mientras que la primera proporciona la información que exclusivamente interesa, la segunda proporciona los textos en los que aparece dicha información. Es por esto, que algunos autores, Cowie y Lehnert (1996), consideran la RI como una etapa previa a la EI. Las dos disciplinas usan teorías y técnicas de la lingüística computacional.

La *extracción de información* (EI) es una nueva disciplina pero no una nueva idea (Wilks, 1997). Ya en 1964 se podían encontrar artículos de investigación con títulos como *"Búsquedas de texto con plantillas"* desarrollados por Wilks (1987), pero esta idea no pudo ser probada por falta de capacidad computacional. Cowie (1983), realizó unos estudios sobre la extracción de estructuras canónicas a partir de unas guías predefinidas de plantas y animales.

Tal y como se indica en la MUC-7 (1998), la EI es una tecnología futurista desde el punto de vista de los usuarios en el mundo actual dirigido por la información. Más que indicar qué documentos necesitan ser leídos por el usuario, extrae trozos de información que son los que el usuario necesita conocer.

Otra definición de EI es la proporcionada por Gaizauskas y Wilks (1998) que definen la EI como la actividad de extraer automáticamente un tipo de información pre-especificada desde textos.

La información a extraer por estos sistemas se define a través de unas *plantillas*. Estas plantillas están formadas por una serie de atributos que caracterizan a cada una de ellas. La construcción de estas plantillas se realiza de antemano y dependen del contexto o dominio de trabajo del sistema y de la información que se desea obtener. A este contexto se le denomina *escenario*. Principalmente existen plantillas de dos tipos: *plantillas de entidades* y *de relaciones*.

El objetivo de la EI es construir sistemas que encuentren y unan información relevante mientras ignore otras informaciones que no sean relevantes para el ámbito de trabajo. La idea original era la de analizar textos no restringidos, sin embargo actualmente, los sistemas existentes trabajaban con textos restringidos en un dominio específico. Aunque los sistemas suelen estar diseñados para algún dominio concreto, pueden realizar algunas de sus tareas en textos no restringidos, como son el `reconocimiento de entidades` y la `resolución de correferencias`. Sobre esta última se centra el trabajo aquí propuesto.

Desde la perspectiva del PLN, los sistemas de EI son sistemas completos que deben trabajar en distintos niveles, desde el reco-

nocimiento de palabras hasta el análisis de oraciones, y desde el entendimiento al nivel de oración sobre el análisis de discurso al entendimiento sobre el texto completo (Lehnert, 1997).

De las numerosas aplicaciones que tiene la EI quizás la más interesante sea la introducción de datos en una base de datos (BD) a partir de textos no estructurados. Es decir, que al aplicar las técnicas de EI a un documento, sin ninguna estructura específica, se puede obtener una plantilla o formulario (información estructurada) con la información que se tenga que almacenar en la BD (Crawdford, 1997; University of Massachussetts, 1997).

2.2 Breve historia

Como se ha citado anteriormente la EI no es una nueva idea (Wilks, 1987), podemos encontrar sus raíces a mediados de los años 60. Pero es en los 80 cuando la EI empieza a crecer rápidamente. Esto es debido a la intervención de DARPA (*Defense Advanced Research Projects Agency*) que es la agencia de defensa de los Estados Unidos, la cual fomentó la competición entre diferentes grupos de investigación para que desarrollasen sistemas de EI.

Tal y como indican Gaizauskas y Wilks (1998) los trabajos realizados sobre EI los podemos dividir en tres etapas:

- antes de la intervención de DARPA,
- bajo las guías definidas por DARPA y
- últimos trabajos de extracción de información.

A continuación se detalla cada una de estas etapas:

a) Antes de la intervención de DARPA.
Los primeros trabajos que se realizaron consistían en el relleno de plantillas con información a partir de unos textos. Así, el sistema Sager (Sager, 1981), que trabaja en el dominio médico, constituye un proyecto ampliamente desarrollado combinando análisis sintáctico superficial y el uso de plantillas. Uno de los aspectos más interesantes de este trabajo consiste en que no hace una definición a priori del formato de la información por parte de un experto, sino que dado un conjunto de textos en

un sublenguaje del dominio, a través de un análisis, descubre las clases de palabras del dominio.
A finales de los 80 y principios de los 90 la inducción de plantillas se abandonó, pasándose a realizar una definición de las mismas por parte de un experto del dominio. El primer sistema de EI como resultado de un problema comercial fue JASPER desarrollado por Andersen *et al.* (1992) del grupo de la Carnegie-Mellon, los cuales no disponían de recursos lingüísticos externos como corpus, lexicón, diccionarios específicos, ni algoritmos de aprendizaje.

b) Bajo las guías definidas por DARPA.
A mediados de los 80 varios grupos de investigación que trabajaban en la EI a partir de mensajes navales empezaron a reunirse para entender y comparar el comportamiento de sus sistemas. Estas reuniones dieron lugar a las hoy conocidas como *Messages Understanding Conference* (MUC).
El objetivo de estas conferencias es fijar un régimen de evaluación cuantitativo para los sistemas de EI, estableciendo el conjunto de textos sobre los que todos los sistemas son evaluados. A continuación se muestra una breve descripción de las MUC celebradas hasta el momento:

- *MUC-1 (mayo-87).* Participaron seis sistemas. No hubo definición de tareas ni se estableció un criterio de evaluación.

- *MUC-2 (mayo-89).* Participaron ocho sistemas. El dominio de trabajo fue, al igual que en la MUC anterior, el de los textos sobre operaciones navales. Se definió una plantilla, así como, reglas para rellenar los *slots* (atributos que caracterizan a las plantillas). También se definió un criterio de evaluación, sin embargo, de forma consensuada, se llegó a la conclusión que ese criterio no era adecuado.

- *MUC-3 (mayo-91).* Participaron quince sistemas. El dominio en estas conferencias dejó de ser el de las operaciones navales para pasar a ser el de los atentados terroristas en

América Latina. En esta conferencia se adaptaron las medidas de evaluación utilizadas en recuperación de información (cobertura y precisión) para la extracción de información. Se entiende por *cobertura* (*recall*) el número de extracciones correctas realizadas con respecto al número de extracciones posibles existentes en el texto. *Precisión* (*precision*) es el número de extracciones correctas del total extraído.

- *MUC-4 (junio-92).* Participaron diecisiete sistemas. El dominio y las estructuras de las plantillas no cambiaron sustancialmente. Los cambios se produjeron en la definición de tareas, corpus, medidas de desarrollo y un test de protocolos para proporcionar control en la generación de información falsa, así como en la definición del mejor sistema de evaluación independiente de los textos, en hacer resultados más consistentes y en proporcionar mayor significado a los resultados de comparar dos sistemas. Como resultado de combinar la precisión y la cobertura se introdujo una nueva medida, se le denominó *medida-F* (*F-mesure*). Esta evaluación marcó el comienzo de la inclusión de estas conferencias en el programa TIPSTER (programa de investigación sobre recuperación de información y extracción de información del gobierno de los Estados Unidos).

- *MUC-5 (agosto-93).* Participaron diecisiete sistemas. Hubo dos dominios de trabajo: fusiones de empresas extraídas de artículos de revistas financieras y anuncios de productos microelectrónicos. También se trabajó en dos lenguas: inglés y japonés. Los corpus cada vez eran más grande. La precisión y cobertura perdieron la condición de medidas primarias para pasar a una nueva medida denominada *error de relleno* (*error for response fill*) que intenta medir la fracción incorrecta de respuesta del sistema. Se esperaba que esta medida permitiera a los investigadores centrarse más directamente sobre los problemas de sus sistemas, en particular sobre la información perdida y la información falsa.

- *MUC-6 (noviembre-95).* Participaron diecisiete sistemas. Los objetivos se centraron en la modularidad y portabilidad de los sistemas. Se definieron cuatro tareas fundamentales que, según las definiciones realizadas por Cunningham (Cunningham, 1997), consisten en:

  – Reconocimiento de entidades (RE). El objetivo de esta tarea es encontrar y clasificar las entidades. Se entiende por entidades los nombres de personas, organizaciones, lugares, expresiones temporales y expresiones numéricas relacionadas con monedas.

  – Resolución de correferencias (CO). El objetivo es identificar las expresiones en el texto que hacen referencia al mismo objeto. Las relaciones de correferencias son sólo identificadas entre ciertas clases de expresiones sintácticas (sintagmas nominales definidos y pronombres).

  – Plantillas de elementos (PE). El objetivo de esta tarea es añadir información descriptiva a los resultados de RE. Es decir, añadir información de atributos relacionados con las entidades que son interesantes para el dominio de trabajo.

  – Plantillas de escenario (ES). El objetivo de esta tarea es reunir los resultados de PE en el escenario específico.

  Existen otras cuatro tareas que no se desarrollaron debido a la falta de acuerdo sobre la definición de las características que debían tener dichas tareas y a la falta de tiempo y dinero para producir el desarrollo y las fuentes de prueba. Estas otras cuatro tareas fueron:

  – evaluación de la estructura de análisis (proporciona el análisis sintáctico canónico de cada oración),

– evaluación de la estructura predicado-argumento (proporciona un análisis semántico canónico de cada oración),

– desambiguación del sentido de las palabras (desambiguación del sentido de cada palabra con respecto algún recurso léxico, como WordNet (Miller *et al.*, 1993)) y

– correferencias entre documentos (determina las correferencias entre distintos documentos).

El dominio de esta conferencia fue el de las sucesiones de dirección en eventos financieros. En ella, las mediadas de evaluación precisión y cobertura recuperaron su categoría de métricas de primer orden. Dentro de esta MUC se celebró paralelamente la primera conferencia sobre *Reconocimiento de Entidades Multilinguales* (*MET-1, Multilingual Entity Task*) que se diferencia de la tarea de reconocimiento de entidades del MUC en que se utiliza textos en inglés y japonés, a partir de los cuales los sistemas tienen que extraer las entidades existentes.

- *MUC-7 (primavera-98).* Los dominios de trabajo fueron dos: el de accidentes de aviones, que se utilizó para entrenar los sistemas, y el de los lanzamientos de misiles y artefactos, que se utilizó para evaluar los sistemas. Se desarrolló paralelamente la segunda *Multilingual Entity Task* (MET-2). Se definió una tarea más que se añade a las cuatro anteriormente mencionadas (RE, CO, PE, ES) y que se denomina *relación de plantillas* (RP). El objetivo de esta nueva tarea (RP) es identificar las relaciones entre las diferentes plantillas de elementos. Existen tres tipos de relaciones: *empleado de*, *localizado en* y *producto de*.

c) Otros trabajos sobre EI.

Paralelos a los sistemas que participan en las MUC se han desarrollado numerosos proyectos. A continuación se muestra

algunos de estos proyectos:

- POETIC (*Portable Extendable Traffic Information Collator*) (Evans *et al.*, 1995) extrae información referente a los incidentes de tráfico que causan congestión del mismo. Este sistema trabaja en un dominio restringido, en este dominio se utiliza un sublenguaje caracterizado por una gramática diferente y un fuerte uso de la elipsis.

- SINTESI (*Sistems Integrato per TESti in Italiano*) (Ciravenga *et al.*, 1992) procesa pequeños textos que describen defectos de los coches.

- Las iniciativas de la Ingeniería del Lenguaje (LE) dentro del III y IV Programas Marco de la Comisión de la Comunidad Europea (CEC), han agrupado numerosos proyectos que aún hoy en día están desarrollándose. Algunos de ellos son: TREE (TREE System, 1998), FACILE (FACILE project, 1998), COBALT (COBALT system, 1998), AVENTINUS (AVENTINUS system, 1998), ECRAN (ECRAN project, 1998), entre otros.

A continuación se presentan las arquitecturas más usadas por los sistemas de EI.

2.3 Arquitectura de los sistemas de extracción de información

Cowie y Lehnert (1996) proponen una arquitectura que se puede considerar como el punto de partida de las arquitecturas actualmente más utilizadas. En esta arquitectura se distinguen los siguiente módulos o etapas:

- *Filtrado: NIVEL DE TEXTO.* Determina la relevancia del texto o partes del texto basándose en estadísticas de palabras o en

la ocurrencia de determinados patrones. Esta etapa constituye la tecnología que se conoce como recuperación de información.

- *Etiquetado léxico: NIVEL DE PALABRA.* Marca las palabras con la información léxica y morfológica a través de unas etiquetas. En muchos sistemas de extracción de información existe un preproceso previo durante el cual se hace uso de etiquetadores léxicos, para un análisis preliminar de unidades dentro de las oraciones, y reglas de propósito especial para reconocer las clases semánticas de las unidades (unidades monetarias, nombres propios, lugares, etc.). Por ejemplo en

  D. Quijote de la Mancha

  el sintagma preposicional de la Mancha puede formar parte del nombre o no.

- *Etiquetado semántico: NIVEL SINTAGMA NOMINAL.* Reconoce los sintagmas de la oración en el dominio y les añade información semántica.

- *Analizador sintáctico (parsing): NIVEL ORACIÓN.* A partir de la gramática se obtiene los sintagmas de la oración mostrando las relaciones entre dichos sintagmas.

- *Interpretación del discurso: NIVEL INTER-ORACIÓN.* Superpone y mezcla las estructuras producidas por el analizador. Reconoce y unifica las expresiones de referencia.

- *Generador de salida: NIVEL PLANTILLA.* Formatea la salida a la forma predefinida.

Posteriormente, Grishman (Grishman, 1997) propone una evolución de esta arquitectura, figura 2.1, en la que se observa que tiene una estructura modular, o mejor dicho tubular ya que consiste en una serie de módulos independientes que realizan las tareas básicas de los sistemas de EI y en la que la salida de un módulo sirve como entrada del módulo siguiente.

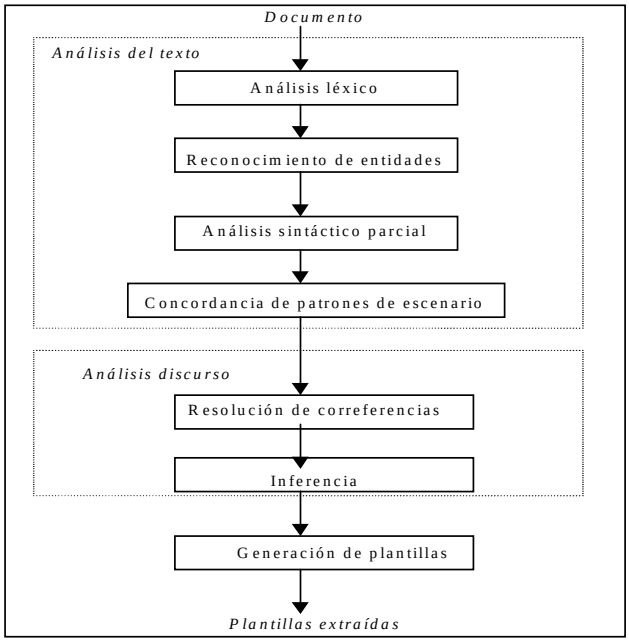

Figura 2.1. *Estructura de un sistema de extracción de información*

Los sistemas de EI Proteus/PET de la Universidad de Nueva York (Yangarber & Grishman, 1998) y LaSIE-II de la Universidad de Sheffield (Humphreys *et al.*, 1998), entre otros, utilizan este tipo de arquitectura modular (tubular), en el que cada módulo se corresponde con una o varias de las diferentes tareas que va a realizar.

A continuación se muestra cada uno de los módulos perteneciente a la arquitectura de Grishman:

- Análisis léxico. Para la realización del análisis léxico se propone realizar una serie de tareas previas, como son:
  - Segmentación del texto. División del texto en las diferentes oraciones que lo forman.

  - Tokenización. División de cada una de las oraciones en las mínimas unidades léxicas.

  - Etiquetado. Asignación de la categoría gramatical, información de concordancia y su raíz morfológica.

- Reconocimiento de entidades. Este módulo se encarga de realizar la tarea definida en el MUC-6. Para llevar a cabo este reconocimiento, Grishman propone las siguientes tareas:

  - Búsqueda en diccionarios específicos. Se buscan, en diccionarios del tipo de enciclopedias geográficas, bases de datos de nombres y de organizaciones, grupos de palabras que normalmente son nombres propios enlazados a través de la conjunción y o la preposición de, entre otros, para agruparlos en una única unidad léxica.

  - Reconocimiento de nombres propios (entidades). Se buscan algunas palabras claves que informen del tipo de entidad de que se traten. Estas palabras son del tipo de tratamientos de persona (Mr., Mrs., etc.), tratamiento de organizaciones (Inc., Inc., etc.), entre otros.

- Análisis sintáctico parcial. Se realiza el análisis sintáctico del texto utilizando las técnicas de análisis parcial. La información sintáctica que añade esta fase es utilizada por las fases posteriores.

- Concordancia de patrones de escenario. El objetivo de este módulo es extraer eventos y relaciones relevantes para el escenario. Como ocurre en el siguiente ejemplo, perteneciente a un sistema que trabaja en el dominio de las sucesiones de empresas:

  John Grant succeeds to Steve Miller.

  A través de la utilización de patrones de escenario se llega a la conclusión que el puesto en la empresa, ocupado por Steve

Miller, es ahora ocupado por John Grant.

- Resolución de correferencias. Se resuelven todas las referencias lingüísticas existentes en el texto producidas por cualquier tipo de expresión anafórica.

- Inferencia. En muchas ocasiones aparece información parcial, acerca de un evento, dispersa a lo largo del texto y necesita ser reconocida a través de la concordancia de patrones (fase anterior a la resolución de correferencias). Pero, en otros casos, la información aparece implícita y debe ser reconocida a través de un proceso de inferencia. En la mayoría de los casos la resolución de correferencias juega un papel muy importante ya que pone de manifiesto muchas relaciones entre diversas partes del texto.

- Generación de plantillas. Toda la información relevante se reconoce después de realizar las tareas de los módulos anteriores. Después de identificar la información relevante se debe rellenar los atributos correspondientes a la información detectada, para ello se aplica unas reglas de extracción que asocian el tipo de información con el atributo de plantilla y, posteriormente, en función del objetivo del sistema de EI, se almacenará en un base de datos o se utilizará esta información para otros fines.

Por otro lado, existe otra posible arquitectura para los sistemas de EI, del tipo de pizarra, es decir existe un módulo central que actúa como un repositorio con el que interaccionan el resto de módulos que son los que realizan las tareas principales del sistema. El sistema LOLITA (Garigliano *et al.*, 1998) (figura 2.2) utiliza como módulo central o repositorio una red semántica de más de 100.000 nodos. Los nodos se enlazan a través de un conjunto de enlaces y un conjunto de variables de control. Las variables de control añaden la información básica del nodo, como el tipo (evento, entidad, relación, etc.), la familia (humano, inanimado, alimento, organización, etc.), tipo léxico (nombre, preposición, ad-

verbio, etc.). Estos sistemas realizan las mismas tareas que los de arquitectura tubular, aunque pueden aparecer divididas en diferentes módulos o diferente orden.

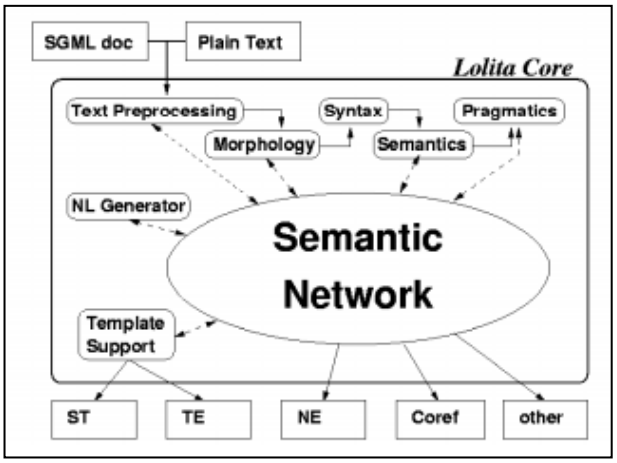

Figura 2.2. *Arquitectura del sistema LOLITA*

Siguiendo con las directrices de una arquitectura tubular, a continuación se describe el sistema EXIT, desarrollado por el Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI)de la Universidad de Alicante. En este sistema se ha integrado y se ha evaluado el módulo de resolución de las correferencias producidas por las DDs desarrollado por este trabajo, tal y como se describe en el capítulo de evaluación (Cap. 6) .

2.4 Sistema EXIT

El sistema de extracción de información EXIT (Llopis *et al.*, 1998) es un sistema que trabaja en el dominio restringido de las escrituras de compraventa. El objetivo principal de este sistema es el de extraer todos los datos relacionados con una compraventa inmobiliaria (comprador, vendedor, notario, inmueble, características

de la venta, etc.) e introducirlos en una base de datos (BD), cuyo esquema entidad-relación se muestra en la figura 2.3.

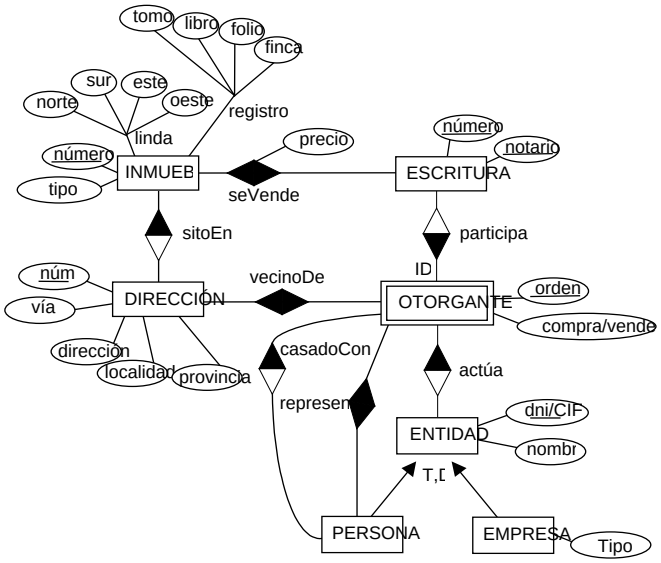

Figura 2.3. *Esquema entidad-relación de la BD utilizada por el sistema EXIT*

La arquitectura del sistema EXIT, figura 2.4, sigue las directrices de la arquitectura propuesta por Grishman. Un conjunto de módulos que desarrollan una tarea específica y que la salida de cada módulo es la entrada del módulo siguiente.

En el sistema EXIT se distinguen 5 etapas, en el capítulo 5 se presentan de forma más detallada todos los módulos que intervienen de forma directa en la resolución de las correferencias. Las etapas del sistema EXIT son:

1. *Análisis léxico.* La información de entrada al sistema será una serie de escrituras de compraventa que se encuentran en formato electrónico. Dentro de esta fase que incluye dos módulos se realizan las siguientes tareas:

   - Segmentación del texto en oraciones. Se utiliza el mecanismo desarrollado en Muñoz (1998) y Muñoz y Palomar (1999)

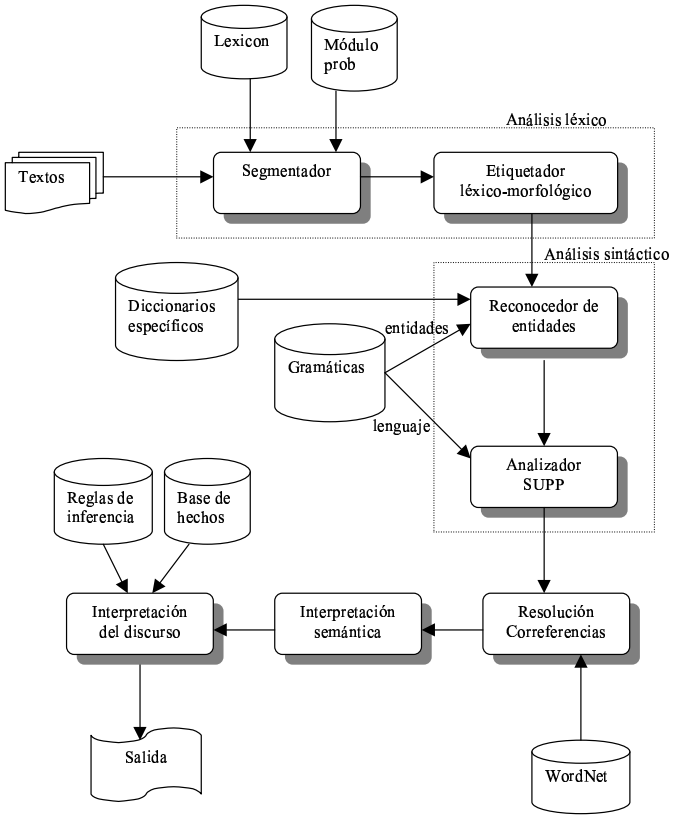

Figura 2.4. *Arquitectura del sistema EXIT*

para dividir la escritura (texto de entrada) en oraciones. Además, se procesa cada una de las oraciones de la escritura, para identificar palabras y asignar a cada una de ellas un identificador que permanece durante todo el proceso.

- Etiquetado léxico-morfológico. Se procesa cada palabra, asignándole las categorías gramaticales posibles que pueda tomar dicha palabra. Se utiliza el etiquetador desarrollado por Pla (2000).

2. *Módulos de Análisis Sintáctico.* Al igual que el sistema LaSIE-II, en esta fase se realizan dos tareas:

- Reconocimiento de entidades. En este módulo se utilizan unos diccionarios específicos para la búsqueda de palabras del tipo de nombres propios, nombres de organizaciones y nombres de lugares, como son:

  – Diccionario de nombres (de 4337 entradas)
  – Diccionario de apellidos (de 4657 entradas)
  – Diccionario de localidades (de 53000 entradas)
  – Diccionario de actividades (de 1500 entradas)

  Se utiliza una gramática específica para el reconocimiento de entidades que incorpora las características de las entidades que aparecen en este dominio. Aunque pueda parecer que mediante la tarea de búsquedas en los diccionarios se reconozcan entidades, esto no es así, ya que sólo reconoce grupos de nombres propios, sin indicar el tipo de entidad de la que se trata. Se utiliza el método desarrollado por Muñoz *et al.* (1998).

- Análisis sintáctico. Para realizar el análisis sintáctico se ha utilizado el analizador parcial SUPP presentado en Martínez-Barco *et al.* (1998) y Palomar *et al.* (1999). Este analizador usa una gramática que aporta las reglas del lenguaje utilizando el formalismo gramatical de las *gramáticas de unificación de huecos.*

3. *Resolución de fenómenos lingüísticos.* Se resuelven tanto el fenómeno de la elipsis como el de las correferencias. Para resolver la elipsis se propone utilizar el método presentado en Palomar (1996). Las correferencias que se resuelven son las producidas por los pronombres y por las descripciones definidas. Para las correferencias producidas por los pronombres se propone utilizar el método presentado por Ferrández *et al.* (1999) y para el tratamiento de las DDs se aplican el método desarrollado en este trabajo.

4. *Interpretación semántica.* La interpretación semántica consiste en la construcción de la forma lógica de cada una de las oraciones que forma la escritura. Mediante la interpretación semántica se consiguen formas independientes del contexto.

5. *Interpretación del discurso.* Por último, se propone realizar una interpretación del discurso basándose en un conjunto de reglas específicas del dominio de las compraventa y de unas reglas de inferencias necesarias para identificar la información que aparece de forma implícita. Las formas lógicas independientes del contexto de la etapa anterior, pasan a ser dependientes del contexto al realizar la interpretación del discurso. Una vez realizada la interpretación del discurso se realiza el relleno de plantillas mediante la utilización de las reglas de extracción. Las plantillas utilizadas por el sistema EXIT son las que se muestran en la figura 2.5

<PLANTILLA-0001> : =
DOC_NR :
CONTENIDO :
<COMPRA_VENTA-0001> : =
NOTARIO :
VENDEDOR :
COMPRADOR :
OBJETO_COMP :
COSTES :
<PERSONA-0001> : =
NOMBRE :
DNI :
DIRECCION :
TITULO :

<ORGANIZACION-0001> : =
NOMBRE :
CIF :
DIRECCION :
TITULO :
<INMUEBLE-0001> : =
IDENTIFICACION :
TIPO :
VIA :
DIRECCION :
LINDA :
REGISTRO :

Figura 2.5. *Plantillas del sistema EXIT*

2.5 Influencia de la resolución de las correferencias en la extracción de información

Como se ha visto anteriormente, una de las motivaciones que llevó a la MUC-6 a definir la tarea de la resolución de las correferencias fue el hecho de que en ocasiones una misma entidad aparecía referenciada o citada por medio de diferentes expresiones, lo cual

provocaba que los reconocedores de entidades extrajeran o rellenaran varias plantillas para una misma entidad. Para evitar este problema, que afecta a la efectividad del relleno de plantillas, se definió la tarea de la resolución de las correferencias, y en concreto la resolución de las DDs. Una vez se aplica la resolución de las DDs se identifica como una única entidad todas las apariciones de estas expresiones y sólo se rellena una única plantilla.

La resolución de las DDs no sólo soluciona el problema del relleno multiple de plantillas, sino que además soluciona el problema de la información implícita. La información denominada relevante se define a través del uso de unas plantillas que se proporcionan previamente al sistema de EI. A estas plantillas le acompañan una serie de reglas que definen las características de la información a asignar a cada uno de los atributos que forman la plantilla. A pesar de que los sistemas de EI trabajan en dominios restringidos, los textos sobre los que se aplica no tienen porque tener un formato fijo. Además, los textos sobre un dominio determinado escritos por diferentes autores pueden tener formatos y estructuras distintas. No es normal encontrar la información relevante presente en el texto de una manera explícita, sino que lo normal es que aparezca de forma implícita a través de relaciones entre diferentes partes del texto. La forma de extraer la información implícita es mediante la resolución de las correferencias ya que establece dichas relaciones entre las diferentes partes del texto.

El siguiente ejemplo muestra los problemas mencionados anteriormente.

Ante mí, José García Gómez, notario del Ilustre Colegio de Barcelona, comparecen D. Antonio Martínez López y D. José Pérez Pérez, ambos mayores de edad y con DNI 24.658.789 y DNI 12.345.678, respectivamente..... El Sr. Martínez, propietario del piso situado en la Avda. Diagonal número 4 piso 3 letra A, vende su propiedad al Sr. Pérez por un importe de trece millones (13.000.000) de pesetas..... El vendedor se hace cargo de todas las costas que originan esta transacción y el comprador hace entrega de talón bancario por el importe total de la venta.

La figura 2.6 muestra las plantillas generadas por el sistema de EI sin resolución de las correferencias. En las plantillas usadas por los sistemas de EI hay atributos que pueden ser opcionales y otros que son obligatorios ya que la información puede que no esté presente en los textos. En los resultados que se muestran en la figura 2.6, se observa que determinados atributos, cuya información está presente en el texto, no aparecen, como es el caso de costes, vendedor y comprador en la plantilla *compra-venta* al no saber detectar la información implícita, o el de los DNI en las plantillas *personas* al no resolver la referencia del adverbio respectivamente. También se observa que se generan más plantillas del tipo *persona* de las debidas ya que no reconoce como la misma entidad la plantilla de *persona 0002* y la *0004*, y la *0003* con la plantilla *persona 0005*.

<COMPRA_VENTA-0001>: =
NOTARIO : **Persona-0001**
VENDEDOR :
COMPRADOR :
*OBJETO_COMP :***inmueble-0001**
COSTES :
<PERSONA-0001>: =
NOMBRE : **José García Gómez**
DNI :
DIRECCION :
TITULO :
<PERSONA-0002>: =
NOMBRE : **Antonio Martínez López**
DNI :
DIRECCION :
TITULO : **D.**
<PERSONA-0003>: =
NOMBRE : **José Pérez Pérez**
DNI :
DIRECCION :
TITULO : **D.**

<PERSONA-0004>: =
NOMBRE : **Martínez**
DNI :
DIRECCION :
TITULO : **Sr.**
<PERSONA-0005>: =
NOMBRE : **Pérez**
DNI :
DIRECCION :
TITULO : **Sr.**
<INMUEBLE-0001> : =
IDENTIFICACION : **inmueble**
TIPO : **piso**
VIA : **avda.**
DIRECCION : **Diagonal número 4 piso 3 letra A**
LINDA :
REGISTRO :

Figura 2.6. *Salida del sistema EXIT sin módulo de CO*

La figura 2.7 muestra las plantillas generadas cuando el sistema de extracción de información EXIT incorpora el módulo de resolución de las correferencias producidas tanto por pronombres como por DDs y alias. En estas plantillas se observa que los problemas que anteriormente se daban desaparecen y la cantidad de atributos rellenos correctamente aumenta gracias a la aplicación de la resolución de las correferencias. Este no es un caso aisla-

<COMPRA_VENTA-0001>: =
NOTARIO : **Persona-0001**
VENDEDOR : **Persona-0002**
COMPRADOR : **Persona-0003**
OBJETO_COMP: **Inmueble-0001**
COSTES: **Persona-0001**
<PERSONA-0001>: =
NOMBRE : **José García Gómez**
DNI :
DIRECCION :
TITULO :
<PERSONA-0002>: =
NOMBRE : **Antonio Martínez López**
DNI: **24.658.789-R**
DIRECCION :
TITULO : **D.**
<PERSONA-0003>: =
NOMBRE : **José Pérez Pérez**
DNI : **12.345.678-J**
DIRECCION :
TITULO : **D.**
<INMUEBLE-0001> : =
IDENTIFICACION :**inmueble**
TIPO :**piso**
VIA : **Avda.**
DIRECCION : **Diagonal nº 4, piso 3 letra A**
LINDA :
REGISTRO :

Figura 2.7. *Salida del sistema EXIT con el módulo de CO*

do, sino que, por lo general, la resolución de las correferencias incrementa la precisión en la tarea de la EI.

2.6 Conclusiones

En este capítulo se ha presentado la evolución de los sistemas de EI desde sus orígenes hasta la actualidad. La intervención de DARPA mediante la organización de las diferentes MUC proporcionó la evolución de este tipo de sistemas y dictó las directrices a seguir y las tareas a desarrollar. Por otro lado, se ha presentado la arquitectura modular de Grishman que aplican la mayoría de sistemas. Ha quedado patente, la importancia de la resolución de las correferencias en los sistemas de EI, ya que una de las formas en las que más comúnmente aparece información dispersa y de forma implícita en un texto es a través de relaciones entre sintagmas nominales y más concretamente a través de DDs. Por esta razón, para cualquier sistema de EI es fundamental resolver adecuadamente las relaciones que producen este tipo de expresiones para realizar de una forma efectiva la tarea de extracción de información. En este capítulo ha quedado demostrado que la resolución de las referencias lingüísticas producidas por las DDs incrementa los resultados del sistema EXIT en particular y de cualquier otro en general.

3. Descripciones definidas: El ámbito del problema y métodos para su resolución

En este capítulo se presenta una profunda revisión bibliográfica del tratamiento y resolución de las descripciones definidas en español. Por una lado, se muestran las características de las DDs y las clasificaciones de los diferentes tipos realizadas por diversos autores, y por otro lado, se presentan los métodos desarrollados para su adecuado tratamiento y resolución.

3.1 Estudio Lingüístico.

Las expresiones lingüísticas más usadas para hacer referencia a entidades lingüísticas previamente introducidas en el discurso son los pronombres y las descripciones definidas.

Tal y como define Russell (1919) se entiende por *descripción definida* a cualquier sintagma nominal que es introducido por un artículo determinado (el, la, lo, las, los) o por un demostrativo (este, ese, aquel, estos, esos, aquellos, estas, esas, aquellas).

Las DDs tienen el problema inherente de que no siempre hacen referencia a una entidad lingüística previamente introducida, sino que a veces introducen una nueva entidad lingüística en el discurso. Otra diferencia que cabe destacar de la DD con respecto al pronombre es la distancia con su antecedente. Mientras que la distancia máxima entre el pronombre y la entidad a la que hace referencia suele ser de unas pocas oraciones, en el caso de la DD suele ser mucho mayor debido a la gran cantidad de información semántica que almacena. Otra diferencia entre el pronombre y la DD es que mientras el pronombre sólo puede hacer referencia a la totalidad de su antecedente (*identidad*), la DD puede hacerlo tanto a la totalidad de su antecedente como a una parte de él

(casa, tejado), estableciéndose diferentes tipos de relaciones como puede ser *parte de*, *conjunto - subconjunto* o *conjunto- miembro*. En estos casos, la DD hace referencia a su antecedente y a la vez introduce una nueva entidad.

La casa que se ha comprado Juan tiene goteras. El tejado está roto

En el ejemplo, el sintagma en el que aparece tejado hace referencia al sintagma en el que aparece casa y a su vez introduce la entidad tejado en el discurso.

El presente trabajo se centra en el estudio de las posibles referencias producidas por las descripciones definidas por lo que a continuación se muestra el estudio realizado sobre las propiedades referenciales o no de los sintagmas nominales.

En la figura 3.1 se muestra la clasificación de los diferentes tipos de sintagmas nominales que pueden hacer referencia a un objeto del mundo, realizada por Leonetti (1990).

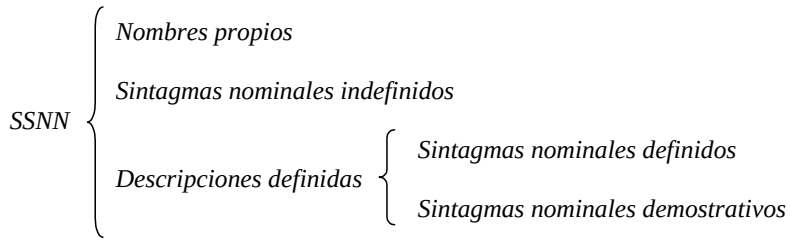

Figura 3.1. *Clasificación de los sintagmas nominales que pueden referenciar un objeto*

En la figura anterior se observa que existen tres tipos principales de sintagmas nominales que pueden referir a una entidad[1], como son los *nombres propios*, los *sintagmas nominales indefinidos* y las *descripciones definidas*. En algunos trabajos, como el de Bustos (1986), no se habla de descripciones definidas sino sólo de sintagmas nominales definidos, que los define como aquellos sin-

[1] Se entiende por entidad cualquier objeto o persona que aparece explícita o implícitamente en la comunicación.

tagmas nominales que están encabezados por un morfema definido (artículos definidos, adjetivos demostrativos).

Puede parecer extraño que los sintagmas nominales indefinidos puedan hacer referencia a un referente al igual que una descripción definida. En la tradición gramatical se han asociado las expresiones definidas con los usos referenciales y las indefinidas con usos no referenciales, sin embargo esto no siempre ocurre así. En el ejemplo:

Sonia está buscando un libro

el sintagma un libro puede referirse a un referente indeterminado, es decir hace referencia a un libro, el que sea. Existen estudios de otros autores, como Donnellan (1996), en los que se realiza una clasificación de los sintagmas nominales más específica en función del tipo de referencia. Donnellan distingue entre *uso referencial y atributivo*, también llamados *uso específico y no específico*. Los usos referenciales vienen asociados a referentes concretos (Sonia está buscando el libro de texto) y los usos no específicos cuando el referente no es uno concreto (Sonia está buscando un libro).

Hasta ahora, el término *referencia* hacía mención a la relación existente entre una expresión lingüística y un objeto del mundo (referente). A partir de ahora cuando se hable de referencia se hace mención a la relación existente entre dos expresiones lingüísticas (denominadas antecedente y expresión anafórica). A esta relación también se le denomina *anáfora*. Esta memoria se centra en las posibles relaciones entre las DDs y sus antecedentes, sin tener en cuenta las relaciones producidas entre otras expresiones anafóricas (pronombres, adverbios, adjetivos) y su antecedente correspondiente.

La relación entre la expresión anafórica y su antecedente no es una mera relación de sustitución sino que constituye un fenómeno de señalización-localización, según los criterios de proximidad/alejamiento. Cuando en los fenómenos anafóricos se habla de referencia a los antecedentes se realiza en el sentido de alusión cotextual. En la figura 3.2 se muestra una clasificación de las diferentes categorías gramaticales que pueden tener las expresiones referenciales o anafóricas:

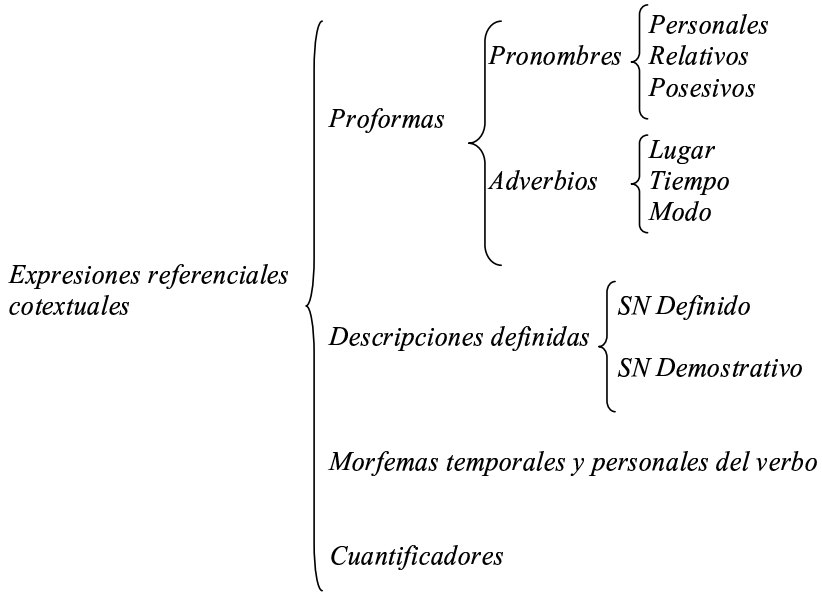

Figura 3.2. *Categorías gramaticales que pueden referenciar*

En la figura 3.3 se muestran las diferentes categorías gramaticales que pueden tomar las expresiones que desempeñan el papel de antecedente.

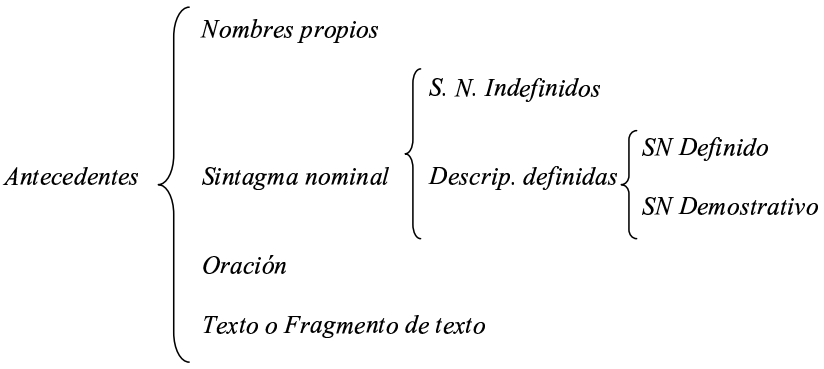

Figura 3.3. *Tipos de posibles antecedentes*

La relación anafórica establece una relación entre dos expresiones lingüísticas (A,B), dónde B es correferente con A. Si además

el antecedente B es una expresión anafórica cuyo antecedente es C, entonces se establece lo que se conoce con el nombre de *cadena de correferencia* y se dice que A y C son también *correferentes*. En la figura 3.4 se muestra un ejemplo en el que puede observarse las cadenas de correferencias.

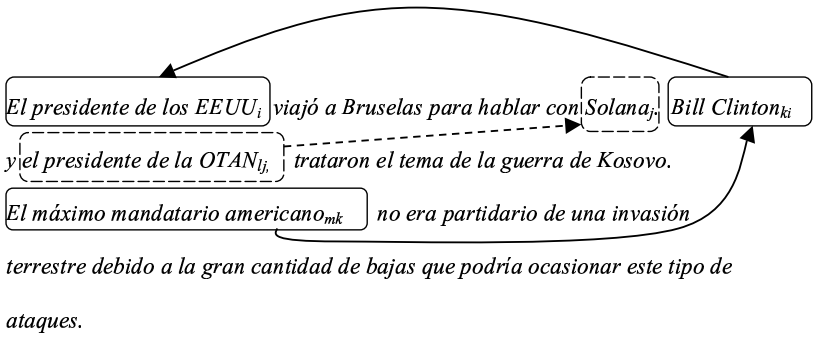

Figura 3.4. *Ejemplo de cadenas de correferencia en un texto*

En el ejemplo, se obtienen dos cadenas de correferencias:

1. La cadena de correferencia indicada con los subíndices (i,ki,mk). El sintagma nominal definido el máximo mandatario americano correfiere con el nombre propio Bill Clinton y éste, a su vez, correfiere con el sintagma nominal definido el presidente de los EEUU. Este último tiene un referente que es la persona física concreta.

2. La cadena de correferencia indicada con los subíndices (j,lj). El sintagma nominal definido el presidente de la OTAN correfiere con el nombre propio Solana que a su vez tiene como referente a una persona física concreta.

Este trabajo se centra en las cadenas de correferencias producidas por DDs o como otros autores denominan sintagmas nominales definidos. Según define Leonetti (1990), existen dos tipos de DDs, en función del artículo que introduce el sintagma nominal. Si el artículo es un artículo definido (el, la, los, las) tenemos

un sintagma nominal definido, pero si es un demostrativo (este, ese, aquel, estos, esos, aquellos, estas, esas, aquellas) la descripción definida se dice que es del tipo de sintagma nominal demostrativo.

En el trabajo de Bustos (1986) se presenta un estudio de las propiedades referenciales de los sintagmas nominales definidos. En ese trabajo se denominan los sintagmas nominales definidos con propiedades referenciales *descripciones* y aquellos que no tienen capacidad para referenciar un objeto *sintagmas nominales genéricos*. En el ejemplo

el hombre es un animal racional

el sintagma nominal definido el hombre no denota un individuo concreto, sino que denota a una clase o un tipo de individuos, por lo que no cumple el requisito de unicidad de las descripciones.

Como se ha visto, el artículo definido no siempre tiene un uso anafórico. Algunos trabajos, como los de Kempson (1975), distinguen entre el uso anafórico lingüístico y otro no lingüístico que referencia a un objeto del universo de discurso. Es decir, los sintagmas nominales definidos con uso anafórico no lingüístico introducen entidades en el discurso (hay una referencia a un referente físico), mientras que los sintagmas nominales definidos con un uso anafórico lingüístico correfieren con un antecedente introducido anteriormente en el discurso. Por esta razón, el uso de un artículo definido por parte de un hablante infiere al oyente que el hablante tiene conocimiento de lo referido por la expresión encabezada por dicho artículo definido.

En cuanto al artículo demostrativo, ocurre lo mismo que con los artículos definidos, es decir, puede tener dos usos anafóricos: uno lingüístico y otro no lingüístico. Por tanto, se puede concluir que los sintagmas nominales definidos y demostrativos cuando desempeñan un papel anafórico lingüístico correfieren con un antecedente, el cuál deberá de escogerse entre todas las expresiones lingüísticas que puedan desempeñar el papel de antecedente y hayan sido introducidas previamente. Cuando desempeñan un uso anafórico no lingüístico introducen una entidad en el discurso, por lo que no habrá que obtener ningún antecedente.

En relación con las ideas expuestas por Bustos se concluye con la existencia de dos categorías principales en las que se pueden dividir los sintagmas nominales definidos y demostrativos:

- Con usos anafóricos
- Con usos referenciales

En este trabajo nos centramos en el tratamiento y resolución de los sintagmas nominales definidos y demostrativos que llamaremos descripciones definidas. A continuación se realiza un estudio de los usos referenciales de las DDs. Para ello se muestran las diferentes clasificaciones de las mismas, realizadas por diversos autores en función de las diferentes relaciones entre las DDs y los antecedentes.

3.2 Ámbito de tratamiento del problema.

El objetivo de este trabajo es localizar en un texto el antecedente de una DD. Además, se pretende establecer las posibles cadenas de correferencia entre todas las DDs con propiedades anafóricas y los antecedentes.

Como se ha visto anteriormente, no todas las DDs tienen propiedades anafóricas. A partir de ahora nos vamos a centrar en las características de las DDs que actúan como expresiones anafóricas, mientras que las que no tienen esa capacidad anafórica lo único que hacen, para nuestros intereses, es introducir una entidad en el discurso.

Diversos autores como Christopherson (1939), Hawkins (1978), Prince (1981), y Poesio y Vieira (1998), han realizado diversas clasificaciones de los diferentes tipos de DDs en función de las relaciones entre expresión anafórica y antecedente.

A continuación se muestran algunas de estas clasificaciones, con ejemplos de cada uno de los tipos. En estos ejemplos la coincidencia en subíndice indica que existe una relación de correferencia entre los dos sintagmas.

3.2.1 Clasificación de Hawkins.

Hawkins (1978) presenta una extensión de la clasificación de las DDs propuesta por Christopherson (1939) en función de los diferentes usos que se le pueden dar en lengua inglesa a las DDs. Dicha clasificación es la siguiente:

a) **Usos anafóricos**. Este tipo de DDs son aquellas que hacen referencia a un antecedente que ha sido introducido previamente en el discurso. Por ejemplo:

- Fred was discussing *an interesting book*$_i$ in his class. I went to discuss *the book*$_i$ with him afterwards.
- Bill was working at *a lathe*$_i$ the other day. All of a sudden *the machine*$_i$ stopped turning

En esta categoría se engloba tanto a las DDs que hacen referencia a un antecedente con el mismo núcleo (como ocurría en el primer ejemplo), como las DDs que hacen referencia a un antecedente que es un sinónimo ó un hipónimo del núcleo (a lathe - machine).

b) **Usos en situaciones directas**. Hawkins engloba en esta categoría a aquellas DDs que se suelen dar en situaciones de conversación en las que se hace referencia a un objeto del mundo, que sin ser introducido anteriormente en el discurso es visible o fácilmente inferible por los que intervienen en la conversación. Por ejemplo:

- Please, pass me *the salt*$_i$
- Beware of *the dog*$_j$

En el primer ejemplo the salt es visible a todos los que están sentados a la mesa. Por el contrario, en el segundo ejemplo podemos no ver el objeto (the dog) pero rápidamente se deduce que existe un perro que es peligroso.

c) **Usos en situaciones generales**. También dentro de las situaciones de conversación, Hawkins identifica DDs con las que el hablante apela al conocimiento general del oyente. Por ejemplo:

- Have you seen *the bridesmaid*$_j$?

En el ejemplo, se sabe que si estás en una boda hay una dama de honor (bridesmaid).

d) **Usos anafóricos asociativos**. Esta categoría engloba aquellas DDs que comparten alguna relación entre objetos y sus componentes o atributos, siendo conocida esta relación tanto por el hablante como por el oyente. Por ejemplo:

- Bill bought *a new car*$_i$ to please Mary but she didn't like *the colour*$_i$

En este ejemplo existe una relación entre the car y the colour. Tanto el oyente como el hablante conocen que los coches pueden tener colores diferentes.

e) **Usos no comunes**. Hawkins engloba en esta categoría a las DDs que no hacen referencia a ningún antecedente introducido anteriormente. Por ejemplo:

- I don't like *the colour red*$_i$
- *The number seven*$_i$ is my lucky number

En estos casos, los sintagmas nominales the colour red y the number seven no hacen referencia a ningún sintagma previo, aunque hayan aparecido previamente sintagmas nominales cuyos núcleos sean colour y number, respectivamente.

f) **Usos de modificadores no explicativos**. Por último, Hawkins engloba en esta categoría a un pequeño número de modificadores que van acompañados del artículo definido (the) en

inglés. Por ejemplo:

- My wife and I share *the same secrets*$_i$
- *The first person to sail to America*$_i$ was an Icelander.

En ambos ejemplos, los modificadores same y first requieren del artículo definido the.

3.2.2 Clasificación de Prince.

Prince (1981; 1982) estudia detalladamente la conexión entre las suposiciones del escritor/hablante acerca del lector/oyente y los conocimientos lingüísticos de los sintagmas nominales. En estos trabajos se indica que hay una serie de factores que afectan a la elección de las DDs. Los tipos identificados son:

a) **Nuevo oyente / Viejo oyente** *(Hearer New/Hearer Old).* Uno de los factores que afecta a la elección de un sintagma nominal es si una entidad del discurso es vieja o nueva con respecto al conocimiento del oyente. Un hablante usará un nombre propio o una DD cuando suponga que el oyente tiene conocimiento del objeto o persona al que se refiere, en caso contrario hará uso de un sintagma nominal indefinido.

- I'm waiting for it to be noon so I can call *Sandy Thompson*$_i$
- I'm waiting for it to be noon so I can call someone in *California*$_i$

b) **Nuevo discurso / Viejo discurso** *(Discourse New /Discurse Old).* Las entidades del discurso pueden ser nuevas o viejas en el modelo de discurso en función de si es su primera aparición en el texto o no. Un sintagma nominal puede hacer referencia a una entidad ya introducida en el discurso o puede hacer referencia a una entidad no introducida previamente (fenómeno de la deíxis). La categoría de *viejo discurso* se corresponde con las categorías de usos anafóricos de la clasificación de Hawkins y la categoría de *nuevo discurso* con las

categorías de usos en situaciones generales y usos en situaciones directas de la clasificación de Hawkins.

c) **Inferibles** *(Inferrables)*. Algunas entidades no son ni viejas en el discurso ni viejas para el oyente, pero tampoco son totalmente nuevas. Son aquellas en las que el oyente es capaz de inferir la relación existente con una entidad ya introducida anteriormente, sin ser la misma. Se corresponde con la anáfora asociativa de Hawkins. Como por ejemplo en:

- A book... the author

d) **Sintagmas que contienen inferibles** *(Containing inferrables)*. Esta categoría podría considerarse como una subcategoría de la anterior. En ella se encuentran aquellos sintagmas nominales que, siendo inferibles, su parte de conexión se especifica como parte del nuevo sintagma nominal. Por ejemplo en:

- The door of the Bastille was painted purple.

Esta categoría se corresponde con la de los usos no comunes de Hawkins.

3.2.3 Clasificación de Poesio y Vieira.

En los trabajos desarrollados por Poesio y Vieira (1998) y por Vieira y Poesio (1998) se realiza un profundo estudio de las DDs. En estos trabajos se presenta una simplificación de la clasificación de Hawkins. La clasificación es la siguiente:

a) **Usos anafóricos con el mismo núcleo**. En esta clase se incluyen las DDs que hacen referencia a un antecedente que ha sido introducido en el discurso anteriormente y que tiene el mismo núcleo. Se trata de una parte de las DDs clasificadas en la categoría de usos anafóricos de la clasificación de Hawkins y a la categoría de viejo discurso de la clasificación de Prince.

Por ejemplo:

- Bill bought *a new car*$_i$ to please Mary. *The car*$_i$ is a new BMW model.

La descripción definida the car tiene el mismo núcleo que el sintagma nominal a new car.

b) **Usos asociativos**. En esta categoría se engloban todas las DDs que tienen un uso anafórico y que no han sido clasificadas en la categoría anterior. Estas DDs tienen usos anafóricos, pero tienen una relación con su antecedente en forma de sinónimos, hipónimos, hiperónimos, información acerca de eventos, etc. En Clark (1977) se denomina a este tipo como *referencias indirectas* (*bridging references*). En esta categoría se engloban tanto los usos anafóricos que no tienen el mismo núcleo, como los usos asociativos de Hawkins. Por ejemplo:

- Bill bought *a new car*$_i$ to please Mary. *The vehicle*$_i$ is a new BMW model.

La descripción definida the vehicle hace referencia al sintagma nominal a new car porque existe una relación de hiperonimia entre vehicle y car.

c) **Usos en situaciones generales/no comunes**. En esta clase se incluyen los usos en situaciones generales y no comunes de la clasificación de Hawkins. Las situaciones generales hacen referencia a una entidad o evento cuya existencia es de conocimiento común. Por ejemplo:

- About the same time, *the Iran-Iraq war*$_i$, which was rolling oil markets, ended.

En las situaciones no comunes la interpretación de la DD se basa en información adicional que aparece junto al sintagma

nominal definido. Esta información adicional aparece en forma de:

- Cláusulas relativas.
- Cláusulas asociativas
- Complemento de sintagmas nominales
- Cláusulas apositivas
- Construcciones copulativas

d) **Idiomáticos**. Esta clase incluye referencias indirectas, expresiones idiomáticas y usos metafóricos. Por ejemplo:

- It went back into *the soup*$_i$.

Adicionalmente a la clasificación anterior, en los trabajos desarrollados por Poesio *et al.* (1997), y por Vieira y Teufel (1997) se realiza un estudio detallado de las DDs clasificadas como descripciones con usos asociativos, es decir, referencias indirectas. A continuación se muestra este estudio:

a) **Sinónimo /Hipónimo /Merónimo**. Esta clase incluye aquellas DDs que tienen una relación de sinonimia, hiponimia o meronimia con su antecedente. Por ejemplo:

- the house y the chimney (Meronimia, parte de).
- the daily television show y the program (Hiponimia)
- new album y the record (Sinonimia).

b) **Nombres propios**. Esta clase incluye las DDs que hacen referencia a nombres propios, como pueden ser nombres de persona, compañías, lugares, etc. Por ejemplo:

- Mozart y the composer.
- Coca-Cola y the company.

c) **Nombres compuestos**. Esta clase incluye las DDs cuyo antecedente es un nombre que forma parte de un nombre compuesto a parte del núcleo. Por ejemplo:

- discount packages y the discounts.

d) **Eventos**. En esta clase se incluyen las DDs que no hacen referencia a un sintagma nominal que ha aparecido anteriormente, sino que hay ocasiones en las que se hace referencia a un verbo o a oraciones. Por ejemplo:

- Kadane Oil Co. is currently *drilling*$_i$ two wells and *putting*$_i$... *the activity*$_i$....

e) **Tópicos del discurso**. En esta clase se incluyen las DDs que hacen referencia a un tópico del discurso. Por ejemplo:

- the industry (en un texto acerca de compañías de petróleo).

f) **Inferencias**. En esta clase se incluyen las DDs cuya relación con su antecedente está basada en una relación inferencial compleja, por ejemplo relaciones causa, efecto, consecuencia, etc. Por ejemplo:

- the last week's earthquake y the suffering people.

Una vez presentadas las diferentes clasificaciones de las DD, a continuación se presenta el estado actual en la resolución de la referencia lingüística, centrándonos en la resolución de las DDs.

3.3 La resolución de la referencia lingüística

La resolución de las referencias lingüísticas producidas por los diferentes tipos de expresiones lingüísticas (pronombres, adjetivos, descripciones definidas, etc.) ha sido abordada por diferentes autores. La mayoría de ellos se han centrado en la resolución de

las referencias producidas por los pronombres, aunque algunos de los sistemas desarrollados por estos investigadores son capaces de tratar las referencias producidas por varios tipos de expresiones lingüísticas. Así, por ejemplo, el desarrollado por Kameyama (1997) trata los pronombres y las DDs dentro del sistema de extracción de información FASTUS (Hobbs *et al.*, 1996) o el sistema *SPAR* (Shallow Processing Anaphora Resolver) desarrollado por Carter (1987) que resuelve también los pronombres y las DDs. Pero también existen otros como el desarrollado por Poesio y Vieira (1998) que se centra exclusivamente en la resolución de las DDs.

Entre las diferentes aproximaciones existentes que tratan la resolución de las referencias producidas por los pronombres, es importante destacar los trabajos desarrollados por Hobbs (1978), Lappin y Leass(1994), las diferentes aproximaciones desarrolladas por Mitkov en (1995; 1997; 1998) y Mitkov y Stys (1995), las desarrolladas por Carbonell y Brown (1988), o las de Dagan e Itai (1990; 1991) y las de Ge *et al.* (1998) que están basadas en modelos probabilísticos. Para el español cabe destacar los trabajos de Ferrández *et al.* (1997; 1998b; 1998a; 1999; 2000) que centran sus trabajos en la resolución de la anáfora pronominal de tercera persona y en la resolución de la anáfora adjetiva en monólogos; y los trabajos de Martínez-Barco y Palomar (Martínez-Barco *et al.*, 1999; Martínez-Barco & Palomar, 2000) en la resolución de la anáfora pronominal en diálogos.

En concreto, los trabajos de Ferrández et al. están basados en el uso de restricciones y preferencias lingüísticas. Las restricciones morfológicas y sintácticas actúan como un filtro para rechazar candidatos que no cumplen la restricción y las preferencias, extraídas de estudios empíricos, seleccionan el candidato en base a estas preferencias mediante un tratamiento adecuado. Estos trabajos se han aplicado a textos no restringidos obteniendo una tasa de éxito del 72% para la anáfora pronominal y de un 76% para la adjetiva.

Por otro lado, los trabajos de Martínez-Barco y Palomar (Martínez-Barco *et al.*, 1999; Martínez-Barco & Palomar, 2000) proponen un espacio de accesibilidad anafórico basado en la es-

tructura del diálogo y los actos del habla que se usan como preferencias estructurales en la resolución de la anáfora.

Por último, los trabajos de Ferrández y Peral (2000) en la resolución de los pronombres omitidos detectan la omisión de l sujeto mediante la utilización de las técnicas de análisis y posteriormente lo tratan como si fuera una anáfora. Para ello utilizan un conjunto de restricciones y preferencias específico para este tipo de problema.

El algoritmo desarrollado por Hobbs (1978) se basa en la búsqueda, en un orden predefinido a través del árbol de análisis (análisis sintáctico completo), de un sintagma nominal que concuerde con el pronombre en género y número. Hobbs aplicó este algoritmo a la resolución de los pronombres ingleses he, she, it y they. A pesar de ser un algoritmo extremadamente sencillo, Hobbs obtuvo un 81.8% de éxito al aplicarlo sobre 100 ejemplos de diferentes fuentes.

Lappin y Leass (1994) desarrollaron un algoritmo para la resolución de las referencias producidas por los pronombre personales de tercera persona, los reflexivos y los recíprocos. En este algoritmo se proponen seis condiciones o restricciones sintácticas (*condiciones de no correferencialidad*). Si alguna de ellas es satisfecha por el sintagma nominal considerado como candidato a ser la referencia del pronombre, entonces dicho candidato es rechazado. Este algoritmo obtuvo un 85% de éxito en textos escritos en inglés.

Mitkov ha desarrollado diferentes aproximaciones al tratamiento de las referencias de los pronombres personales de tercera persona. En una primera aproximación, propone un algoritmo que usa técnicas de razonamiento con incertidumbre en el que se usan dos valores de certeza (Mitkov, 1995). En una segunda aproximación, propone un método basado en restricciones y preferencias que utilizan diferentes tipos de conocimientos (Mitkov, 1998).

Durante los últimos años la resolución de la anáfora pronominal ha sido el objetivo del trabajo de muchos investigadores, tanto en inglés como en español. Sin embargo, a la resolución de las DDs no se le ha dedicado tanto esfuerzo como a la pronominal,

principalmente por dos razones:

1. Mientras que los pronombres "casi siempre"hacen referencia a un antecedente que ha aparecido anteriormente en el texto (discurso), las DDs no siempre tienen propiedades referenciales. Ésta es la principal razón para pensar que el tratamiento de las DDs es distinto.

2. La resolución de las DDs, a priori, requiere el uso de más y mejores fuentes de información. El uso de la información semántica es fundamental en la resolución exitosa de la DD, tal y como se demostrará en la presente memoria.

Una vez presentados brevemente los principales trabajos en anáfora pronominal, y extraídas las principales diferencias entre los pronombres y las DDs, a continuación se presentan los diferentes métodos existentes en la literatura que estudian y tratan la resolución de las DDs. Se pueden clasificar dichos métodos en aproximaciones lingüísticas (basadas en el uso de información léxica, morfológica, sintáctica y semántica) y en aproximaciones basadas en aprendizaje automático (usan textos anotados con información lingüística para aprender reglas en la resolución de las DDs). A continuación se muestran las principales técnicas y métodos englobados en ambas aproximaciones para finalizar presentando sus resultados comparativos.

3.4 Aproximaciones lingüísticas

A lo largo de los últimos años, se han desarrollado diferentes técnicas y métodos para la resolución de las descripciones definidas, fundamentalmente en inglés. Los primeros trabajos fueron desarrollados por Kameyama (1997). Estos trabajos están integrados dentro del sistema de extracción de información FASTUS (Hobbs *et al.*, 1996). Otras propuestas también desarrolladas dentro de sistemas de EI tuvieron lugar en 1998 como consecuencia de

la celebración de la última MUC (MUC-7). Entre ellas cabe destacar las propuestas de los sistemas LaSIE-II (Humphreys *et al.*, 1998), LOLITA (Garigliano *et al.*, 1998) y Oki (Fukumoto *et al.*, 1998). Las propuestas integradas en los sistemas de EI anteriores resuelven tanto la anáfora pronominal como las producidas por las descripciones definidas.

Por último, destacar los trabajos realizados por Vieira y Poesio (1997; 1998; 1999) que se centran en el estudio de las DDs y en el desarrollo de un algoritmo para la correcta resolución en textos escritos en lengua inglesa.

La tabla 3.1, a modo d e resumen inicial, muestra las principales características de algunos sistemas de resolución de las DDs, sobre la base del tipo de referencia tratada, la información utilizada y los corpus sobre los que realizan las evaluaciones de los sistemas.

Algoritmo	Tipos	Información	Corpus	Referencia
Vieira y Poesio	DD	morfológica, sintáctica, semántica	WSJ	(Vieira, 1998)
Kameyama	pronombres y DD	morfológica, sintáctica, semántica	Geo-político	(Kameyama, 1997)
LaSIE-II	pronombres y DD	morfológica, sintáctica y semántica	MUC-7	(Humphreys *et al.*, 1998)
LOLITA	pronombres y DD	morfológica, sintáctica y semántica	MUC-7	(Garigliano *et al.*, 1998)
Oki	pronombres y DD	morfológica y sintáctica	MUC-7	(Fukumoto *et al.*, 1998)

Tabla 3.1. Sistemas basados en conocimiento

A continuación se describen de forma más detallada cada uno de los algoritmos de la tabla 3.1.

3.4.1 Algoritmo de Poesio y Vieira

Poesio y Vieira son dos de los investigadores que más estudios y trabajos han desarrollado por el tratamiento y resolución de las descripciones definidas en inglés. Como consecuencia de estos trabajos han presentado una clasificación de los diferentes tipos de DDs que se pueden encontrar en un texto (Poesio & Vieira, 1998; Vieira, 1998; Vieira & Poesio, 1998) y diferentes propuestas a la resolución de cada uno de los tipos de DDs (Poesio *et al.*, 1997; Vieira & Teufel, 1997; Vieira, 1998; Vieira, 1999).

En los trabajos desarrollados por Poesio y Vieira el corpus utilizado es el Penn Tree Bank (Marcus *et al.*, 1993) que está formado por artículos del Wall Street Journal (WSJ) anotados sintácticamente.

A continuación se muestran las reglas usadas para el tratamiento y resolución de las DDs. Las diferentes categorías pertenecientes a la clasificación de Poesio y Vieira, presentada anteriormente, se han agrupado en tres tipos: (1) *anáfora directa* que se corresponde con los usos anafóricos con el mismo núcleo de la clasificación, (2) identificación de las DDs que introducen una nueva entidad en el discurso, se corresponden con las no anafóricas y (3) anáforas indirectas (bridging reference) y el uso de WordNet (Miller *et al.*, 1993) en la resolución de las DDs, que se corresponden con los usos asociativos y usos en situaciones generales y no comunes.

1. **Anáfora directa**. Poesio y Vieira (1998) definen la anáfora directa como la relación que se establece entre una DD y un antecedente con el mismo núcleo. Para resolver la referencia de una DD se debe identificar los posibles antecedentes y compararlos con la DD, esto se realiza mediante la comparación de los núcleos de ambos. Todos los antecedentes con el mismo núcleo son seleccionados como potenciales candidatos a ser la solución. Los trabajos desarrollados por Poesio y Vieira utilizan una ventana de búsqueda[2] pero si no hay ningún antecedente en ella buscan en el resto del texto.

[2] La ventana de búsqueda está formada por las 4 oraciones anteriores a la oración en la que se encuentra la descripción definida.

No es suficiente la comparación entre los núcleos para resolver las referencias de la anáfora directa. De hecho, son pocos los casos en los que sólo comparando los núcleos se proponen el candidato correcto. En la mayoría de los casos hay que realizar un estudio de los pre y postmodificadores. Por ejemplo, the red car y the blue car no pueden correferir aunque tengan el mismo núcleo. En general, para poder comparar este tipo de casos se necesitaría información semántica. Poesio y Vieira proponen las siguientes dos reglas:

- Se permite como posible antecedente a aquellos cuyos modificadores incluyan a los modificadores de la descripción definida.

- Se permite como antecedente de cualquier descripción definida a un sintagma nominal sin ningún modificador.

2. **Identificación de nueva entidad**. Diversos autores han demostrado que no todas las DDs tienen propiedades referenciales, sino que introducen una nueva entidad en el discurso (más de la mitad de ellas). Debido al elevado número de ellas en los textos, una previa identificación del tipo de las DDs (anafóricas o no anafóricas) mejorarían la eficiencia del sistema. La identificación de estas DDs desarrolladas por Poesio y Vieira están basadas en características léxicas y sintácticas de los sintagmas nominales. Las características que proponen son:

   a) Predicados especiales. Se pueden diferenciar dos casos:

      - En ocasiones algunos premodificadores como first o best cuando van acompañados de una cláusula relativa introducen una nueva entidad en el discurso. Por ejemplo en

        the first person to sail to America

      - Algunos núcleos como fact, result, conclusion, idea, etc. normalmente introducen una nueva entidad en lugar de

hacer referencia a una entidad previa.

b) Modificadores restrictivos. Cuando la información del modificador es esencial para la identificación del antecedente se dice que es restrictivo, pero si no es esencial se dice que es no restrictivo. Normalmente, los postmodificadores restrictivos aparecen en la primera mención de la entidad, por lo que su aparición en una DD denota un uso no anafórico. Alguna de las construcciones que Poesio y Vieira encontraron en sus corpora de trabajo fueron:

- Cláusulas relativas introducidas por pronombres relativos como who, what, when, where, because o that o por ningún pronombre (en inglés una cláusula relativa puede estar introducida sin ningún pronombre). Por ejemplo,

  the girl who I met...
- Indefinidos. Cláusulas infinitivas y gerundios. Es propio de la lengua inglesa ya que en español se traduciría por un cláusula de relativo. Por ejemplo

  the man writing the letter is my friend
- Sintagmas preposicionales. La mayoría de los postmodificadores de un sintagma nominal son sintagmas preposicionales y suelen ser restrictivos. Por ejemplo

  the years before the war...

  no son unos años cualquiera, sino esos concretamente.
- Premodificadores restrictivos. Normalmente son nombres propios o un nombre. Por ejemplo

  the Iraq-Iran war...

  o

  the city payroll

c) Aposiciones. Las DDs que normalmente ocurren dentro de una aposición introducen una nueva entidad en el discurso.

Por ejemplo

Juan Antonio Samaranch, the president of COI, open the games

d) Construcciones copulativas. Las DDs que tienen lugar en construcciones copulativas no tienen que ser resueltas y son clasificadas como una nueva entidad del discurso. Por ejemplo

the bottom line is that he is ....

e) Nombres propios. Nombres propios que están precedidos por un artículo introducen una nueva entidad. Por ejemplo

the Wall Street Journal...

f) Expresiones temporales. Algunas expresiones que referencian tiempo como the morning introducen una nueva entidad y son tratados como los predicados especiales.

3. **Anáforas indirectas** (bridging reference). Las anáforas indirectas son los casos más complejos de tratar de las DDs. El principal problema es que una misma descripción indirecta puede relacionarse con varios antecedentes dentro de un mismo texto. Es obvio que para resolver este tipo de DD se necesita información semántica. A continuación se muestran las diferentes heurísticas usadas por este algoritmo para la resolución de las anáforas indirectas:

   - Para proporcionar la información acerca de sinónimos, hiperónimos, hipónimos, merónimo y holónimos se usa el recurso léxico WordNet.

   - Para resolver las anáforas indirectas que hacen referencia a nombres propios, como por ejemplo en

     Pinkerton Inc... the company

     se hace mediante la identificación del tipo de entidades. La forma en la que identifica las entidades en el texto es me-

diante la búsqueda en WordNet o por algún tipo de identificadores como Co, Inc, Mr.

- Nombres compuestos. A veces se suele usar una descripción indirecta para referir a un nombre compuesto. Como por ejemplo

  discount packages ... the discounts

  o

  stock market crash ... the markets

- Anáforas indirectas basadas en sintagmas verbales. En WordNet no hay información acerca de las relaciones entre nombre y verbos. Para solucionar esta falta de información lo que se hizo fue nominalizar los verbos y buscar relaciones en WordNet con el núcleo de la descripción. Como por ejemplo

  changes were proposed... the proposals

  o

  he plans to ... the plan

Las características anteriores se integraron en un algoritmo cuya estructura general es la siguiente:

1. Asignación de un número de sintagma.

2. Se selecciona la ventana de búsqueda, después de varios experimentos establecieron el tamaño de la ventana en las cuatro oraciones anteriores a la DD.

3. Si el sintagma nominal es una DD se aplican las siguientes reglas:

   a) Se examina la lista de predicados especiales para identificar las posibles DDs que introducen una nueva entidad

b) Se examina si la DD ocurre en una aposición. En caso afirmativo se clasifica como no anafórica.

c) Se intenta buscar un antecedente con el mismo núcleo (anáfora directa). Si se tiene éxito se clasifica como una anáfora directa y se introduce en el sistema una regla que indique la correferencia entre los dos indices de los dos sintagmas.

d) Se comprueba si el núcleo es un nombre propio. Si es nombre propio se clasifica como una nueva entidad.

e) Se comprueba si presenta postmodificadores restrictivos. Si es así se clasifica como una nueva entidad.

f) Se comprueba si hay un nombre propio como premodificador. Si es así se clasifica como una nueva entidad.

g) Se comprueba si la DD tiene lugar en una construcción copulativa. En ese caso se clasifica como nueva entidad.

h) Si las reglas anteriores no tienen ninguna éxito se clasifica como una anáfora indirecta y se aplican las reglas para los nombre propios y búsquedas en WordNet.

De forma general el algoritmo se define de la siguiente forma:

- las dos primeras reglas intentan clasificar las DDs como una nueva entidad,
- la siguiente regla pretende resolver la anáfora directa,
- las cuatro reglas siguientes vuelven a intentar clasificar la DD como una nueva entidad y

- la última regla intenta solucionar las anáforas indirectas.

3.4.2 Algoritmo de Kameyama

Kameyama (1997) desarrolló un algoritmo para la resolución de las referencias lingüísticas producidas por las expresiones gramaticales de tipo pronombre y descripciones definidas. Este algoritmo fue probado y evaluado en el seno del sistema de extracción de información FASTUS (Hobbs *et al.*, 1996). Este sistema sólo resuelve las relaciones de tipo identidad sin tener en cuenta otros tipos de relaciones. Las relaciones de tipo identidad son aquellas en las que la expresión anafórica hace referencia a la totalidad del antecedente y éste puede sustituir a la expresión anafórica sin que cambie el sentido de la oración.

El algoritmo de resolución de las referencias producidas por los pronombres y las DDs está basado en tres factores: *región de texto accesible*, *consistencia semántica* y *preferencias sintácticas dinámicas*. A continuación se muestran, de forma más detallada, las características de cada uno de estos factores.

1. *Región de texto accesible.* Establece la zona del texto donde se busca el posible antecedente. Esta región de texto o ventana de búsqueda se estableció de forma experimental. Se utilizó una ventana de tres oraciones anteriores a la oración en la que aparece el pronombre y de diez oraciones anteriores a la aparición de la DD.

2. *Consistencia semántica.* La consistencia semántica entre la expresión anafórica y su antecedente se estableció en función a los siguientes criterios:

   - Consistencia de número. La expresión anafórica debe tener el mismo número, `singular` o `plural`, que su antecedente.

   - Consistencia de tipo. El tipo de la expresión anafórica (pronombre, descripción definida, etc.) debe ser igual o un sub-

tipo del tipo del antecedente.

- Consistencia de los modificadores. En el caso de las descripciones definidas los posibles modificadores deben ser consistentes entre ellos.

3. *Preferencias sintácticas dinámicas.* Las preferencias se usan para elegir un candidato entre varios posibles que cumplen las consistencias semánticas y se encuentran dentro de la región del texto accesible. Las preferencias sintácticas dinámicas propuestas son:

   - Se prefiere como antecedente los que se encuentran en la misma oración de izquierda a derecha.

   - Si no se encuentra ningún antecedente en la misma oración, se busca en la oración inmediatamente anterior del mismo modo.

   - Si ningún antecedente es propuesto por las reglas anteriores, se recorre el resto de las oraciones que pertenecen al espacio de búsqueda en orden inverso, es decir de derecha a izquierda.

3.4.3 Resolución de correferencias en el sistema LaSIE-II

El mecanismo de resolución de las correferencias lingüísticas integrado en el sistema LaSIE-II (Humphreys *et al.*, 1998), basado en el sistema LaSIE-I, realiza una comparación entre las nuevas entidades del discurso y las ya existentes. El mecanismo es capaz de resolver diferentes tipos de expresiones gramaticales (pronombres, nombres propios y sintagmas nominales). En primer lugar intenta resolver las referencias que producen cualquier tipo de expresión anafórica dentro de la misma oración (intraoracional) y, posteriormente, resuelve las referencias a entidades de otras oraciones

(interoracional). Cuando el mecanismo detecta que una entidad refiere a otra, une ambas entidades en una misma instancia[3] y elimina del discurso la instancia menos específica.

El módulo de resolución de correferencias del sistema LaSIE está formado por cuatro componentes:

- **El modelo del mundo**. El sistema LaSIE (I y II) utiliza una representación del mundo a través de una ontología de conceptos. Esta ontología no trata de abarcar el mundo en su totalidad, sino que el sistema LaSIE utiliza una ontología específica para cada dominio sobre el que trabaja. Esta ontología es predefinida a mano en función del dominio. Cada entidad perteneciente a uno de los nodos de la ontología viene caracterizada por una estructura atributo-valor. Conforme se procesa el texto se genera el modelo del discurso.

- **El modelo del discurso**. El modelo del discurso se construye a partir de las QLF (*quasi-logical form*) de las diferentes oraciones, integrándolas en una ontología dependiente del dominio. Las QLF son simplemente conjunciones de términos lógicos de primer orden. Los predicados en la representación de las QLF se derivan o de la raíz de la propia palabra, o de clases cercanas de predicados relacionados. Por ejemplo la DD the company se representa en QLF a través de los predicados company(e22), det(e22,the)
  La figura 3.5 presenta la ontología que el sistema LaSIE usa en el dominio de las sucesiones en puestos de trabajo. La ontología está formada por una serie de nodos enlazados mediante relaciones del tipo *is a* o *is instance*. Cada nodo tiene asociado una estructura atributo-valor que son heredadas hacia abajo en el árbol. Existen atributos fijos y otros que pueden ser opcionales. Conforme se procesa el texto cada una de las entidades se añade a la ontología en la clase correspondiente (objeto, evento, propiedad), es decir hay una instanciación de los nodos. La

[3] Se denomina instancia de una entidad al conjunto de información (morfológica, sintáctica y semántica) correspondientes a la entidad que se usa para ver la compatibilidad entre dos entidades.

información referente a los pares atributos-valor es proporcionada por las QLF resultantes de la salida del analizador sintáctico.

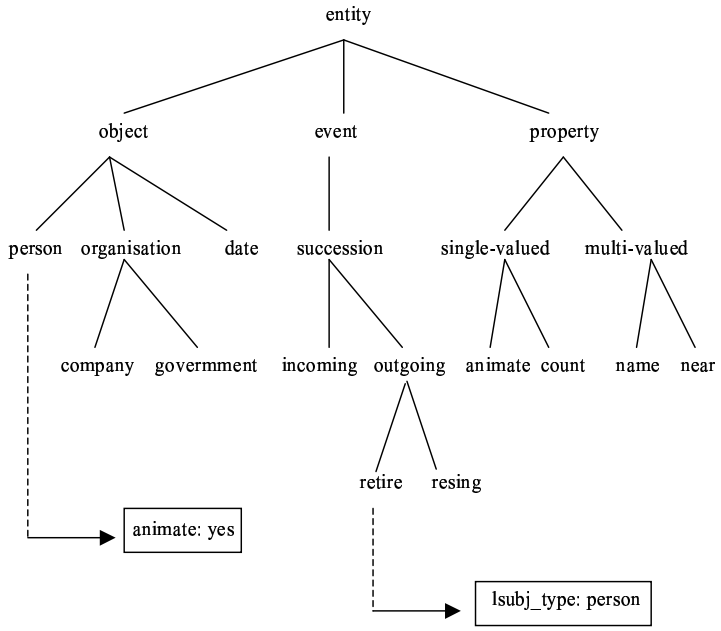

Figura 3.5. *Ontología usada en el dominio de las sucesiones en empresas por LaSIE*

- **El algoritmo de resolución**. El algoritmo de resolución de correferencias se aplica cada vez que se encuentra una expresión gramatical en el texto que puede tener propiedades referenciales (pronombres, nombres propios y sintagmas nominales). Una vez se encuentra este tipo de expresiones, se usa su instancia para ser comparada con instancias previas que han sido introducidas en el modelo de discurso. Este sistema sólo resuelve lo que denomina correferencia objeto que son referencias a objetos o entidades estáticas (objeto, personas, lugares) pero no trata la correferencias de eventos (acciones o eventos). La resolución de las referencias a objetos o entidades estáticas las lleva a cabo comparando el siguiente conjunto de instancias en este orden:

1. En primer lugar resuelve las referencias de los nombres propios. Compara cada instancia de un nombre propio de la oración que se está procesando con instancias de nombres propios presentes en el modelo de discurso.

2. Compara cada nueva instancia con instancias de la misma oración (resolución de las referencias intraoracionales).

3. Compara cada instancia de un pronombre con cada instancia existente en el modelo de discurso perteneciente al mismo párrafo. En el caso de que el pronombre sea la primera instancia del párrafo se usa el párrafo anterior.

4. Compara cada instancia de un sintagma nominal con cada instancia del modelo de discurso que pertenezca al párrafo actual o anterior.

Se puede observar que el espacio de búsqueda del antecedente correcto de cada tipo de expresión anafórica es diferente. Por ejemplo, para un nombre propio se busca en todo el texto anterior a su aparición, para los pronombres se busca su antecedente en el mismo párrafo y para los sintagmas nominales en el párrafo actual y anterior. Esta última regla fue introducida para mejorar la eficiencia del sistema y no tener que comparar con todas las instancias previas. Una vez comparado con las instancias pertenecientes a su espacio de búsqueda, un conjunto de instancias pueden ser candidatos a correferir con el sintagma nominal o cualquier expresión anafórica, entonces el algoritmo procede de la siguiente manera para cada par de instancia:

1. *Consistencia de atributo.* Existen una serie de atributos que sólo pueden tomar un único valor para todas las instancias pertenecientes a ese tipo. Por ejemplo, animate y time. Si dos instancias tienen en común este tipo de atributos, sus valores deben ser los mismos. En caso de que esto no se cumpla la resolución de este par se abandona.

2. *Consistencia semántica.* Para determinar la consistencia semántica se requiere establecer un camino en la ontología entre los tipos semánticos de cada una de las instancias. Si el camino existe, se calcula un valor de similitud semántica (C_s). Este valor se calcula en función de la inversa de la longitud del camino medida en números de nodos entre los dos tipos semánticos. Para instancias del *tipo evento*, se dice que un camino es válido si ambos tipos semánticos son sub-eventos de un evento superior en la ontología. Por ejemplo, hire y retire son sub-eventos de sucession. Para instancias del *tipo objeto*, un camino es válido si entre los dos tipos existe una relación de dominio, es decir, están en la misma rama de la ontología. Por ejemplo, company y organisation. Si no se encuentra un camino válido, no se acepta a la entidad ya perteneciente al modelo del discurso como antecedente de la nueva entidad.

3. *Calcular la similitud (S).* El valor de similitud semántica (C_s) se suma con un valor de consistencia de atributo que viene dado por el número de atributos con el mismo valor (A_v) dividido por el número de atributos compartidos (A_c).

$$S = C_s + \frac{\sum A_v}{\sum A_c}$$

Una vez calculada para cada par de instancias (expresión anafórica - posible candidato) su valor de similitud, el par con el valor de similitud más alto es el escogido como el antecedente correcto a la expresión anafórica. Si más de un par tienen el valor de similitud más alto entonces el par con la instancia más cercana en el texto a la expresión anafórica es el preferido. La instancia de la expresión anafórica se unifica con la instancia del antecedente en una nueva instancia que se añade al modelo del discurso y se elimina la menos específica (más alta en la ontología). Los valores de los atributos de esta instancia eliminada son heredados por las instancias que estén por debajo de ella en la ontología.

- **Restricciones adicionales**. Tal y como se ha mencionado anteriormente existe una serie de restricciones adicionales que se

tienen en cuenta a la hora de rechazar algún par de instancias que puedan ser incompatibles. El conjunto de restricciones básicas usadas son:

– Los sintagmas nominales indefinidos no son comparados con instancias previas, ya que se parte de la base que un sintagma nominal indefinido no puede referir a una entidad previa, más bien se usa para introducir una entidad por primera vez.

– Los pronombres no pueden ser antecedentes de otros tipos. En ocasiones, es posible que haya algún pronombre que no sea resuelto y entonces se introduce su instancia en el modelo de discurso. Esta instancia no puede ser antecedente de ninguna instancia. Con esto se garantiza que un pronombre no va a ser la raíz de ninguna cadena de correferencia.

– Los nombres propios no pueden referir a instancias de fecha. Las nuevas instancias de nombres propios que tienen asociada la clase semántica `objeto` se definen como distintas a todas las instancias con el atributo de clase semántica `fecha`.

– Se impide que nombres no propios usados como modificadores puedan correferir. Una instancia introducida como modificador de otra instancia se considera distinta a todas las demás. Por ejemplo, la instancia video extraída del sintagma the video manufactures no se permite que correfiera con otras instancias. Para esa instancia se usa la instancia manufactures.

– Se impide que un pronombre refiera a una fecha, a un número o a un lugar. Esta restricción, es quizás la más dependiente del dominio para el que se ha evaluado LaSIE. Sin embargo, existen adverbios que pueden hacer referencia a lugares o tiempo, como son allí y entonces, respectivamente.

– Una fecha que no contenga un nombre propio no puede referir a otra fecha sin nombre propio. Es decir, que un sintagma

nominal the month podrá referir a January.

– Una instancia de fecha que no sea introducida por un artículo definido no podrá referir a una instancia de fecha previa.

Algunas mejoras de este sistema fueron desarrolladas y evaluadas en la MUC-7. Las nuevas características que se usaron para ver la compatibilidad entre dos entidades fueron:

- **Distancia**. En el sistema LaSIE-I la resolución de pronombres y nombres buscaba el antecedente sólo en la oración en la que aparecía y en la anterior, ni siquiera para expresiones que siempre requieren un antecedente por ser siempre una expresión anafórica (pronombres). En LaSIE-II el mecanismo amplió su espacio de búsqueda hasta el párrafo anterior, obteniendo una mejora del 2% en la cobertura de los pronombres y un 7% en la cobertura de los nombres.

- **Construcciones copulativas**. Las construcciones del tipo SN_1 *es* SN_2 en las cuáles SN_2 correfiere con SN_1 son tratadas, pero se añaden una serie de reglas para tener en cuenta las posibles excepciones a otras reglas. En general un sintagma nominal indefinido no tiene antecedente, pero esto debe ser relajado para poder aplicar las reglas copulativas. Por ejemplo, en

  The F14 "Tomcat" is the Navy's first-line fighter aircraft...

  las dos descripciones definidas correfieren (The F14 "Tomcat" y the Navy's first-line fighter aircraft), pero en

  Bill is a president

  los dos sintagmas nominales no correfieren por las restricciones previas.

- **Catáfora**. El sistema trata dos tipos de catáforas: la primera de ellas cuando se da en oraciones entrecomilladas y la segunda en oraciones copulativas del estilo This is a mystery.

- **Coordinación de sintagmas nominales**. Cuando el sistema encuentra una coordinación de sintagmas nominales crea una instancia que añade al modelo del discurso por cada uno de los sintagmas que forman el sintagma coordinado más la instancia correspondiente al sintagma coordinado. Por ejemplo si se coordinan dos sintagmas nominales como en la oración

  Bruce and his boys ...

  en el modelo del discurso se añaden tres instancias, una para Bruce, otra para his boys y una última para el sintagma coordinado Bruce and his boys en el cual el atributo de número tiene el valor de `plural`. Evidentemente, parte del éxito de esta regla dependerá de la eficiencia del analizador sintáctico que es el encargado de proporcionar este tipo de información.

- **Correferencias con títulos**. Para poder tratar las correferencias en los títulos de los textos se deben relajar las reglas usadas en el texto normal debido a la falta de información sintáctica, que en la mayoría de casos tiene su origen en el estilo telegráfico usado en este tipo de oraciones. Suelen se oraciones incompletas, en las que habitualmente falta el verbo.

3.4.4 Resolución de correferencias en el sistema LOLITA

El módulo de resolución de las correferencias del sistema de extracción de información LOLITA (Garigliano *et al.*, 1998) procesa el texto oración a oración para buscar el antecedente correcto a cada expresión anafórica. Cada vez que se encuentra un posible referente (un sintagma nominal), se almacena en la *lista de contexto (context buffer)*. Cuando se identifica una expresión anafórica (pronombre o DD), el sistema busca un posible antecedente a esta expresión anafórica en la lista de contexto.

El mecanismo utilizado por el módulo de resolución de las correferencias es mediante la confrontación de una serie de reglas en función del tipo de expresión anafórica. Si el sistema no encuentra ningún candidato en la lista de contexto y la expresión anafórica es una DD, entonces la almacena en la lista de contexto. Si el sistema propone un antecedente como candidato a la expresión anafórica, entonces el sistema unifica ambas expresiones y lo introduce en la lista de contexto. Pero si el sistema propone más de un candidato, entonces el sistema construye una estructura especial para representar esta ambigüedad y la utiliza como entrada a un sistema de preferencias heurísticas para decidir cual de los candidatos propuestos es el escogido como solución de la expresión anafórica.

Las preferencias heurísticas están basadas en la teoría del *centering*, en conocimientos psicolingüísticos y en el sentido común. Estas reglas calculan la importancia de los candidatos basándose en características gramaticales (concordancia de género y número), semánticas (compatibilidad entre los modificadores), así como en su posición en la oración, su proximidad y su relación con el tópico del discurso.

Este algoritmo depende en gran medida de la realización de un buen análisis, tanto sintáctico (proporciona las características gramaticales y sintácticas) como semántico (proporciona las características semánticas de compatibilidad). Los resultados obtenidos en la evaluación de este sistema no fueron muy satisfactorios, una de las principales razones fue debida a la utilización de un análisis completo. En muchos casos, el análisis completo no puede realizarse y no siempre se proporciona la información suficiente para realizar un buen análisis sintáctico. Esto lleva a problemas en los que incluso relaciones simples no son resueltas correctamente. Otro de los problemas que se encontraron fue la falta de recursos que proporcionen la información necesaria para resolver las correferencias producidas por sintagmas nominales coordinados. Además de estos errores, se encontraron errores triviales en algunos procesos, como en el reconocimiento de entidades. La resolución de algunos de estos problemas triviales, posteriormente,

ocasionaron un aumento de los resultados al evaluar nuevamente el sistema con los textos utilizados en la MUC-7.

3.4.5 Resolución de correferencias en el sistema Oki

El módulo de resolución de las correferencias usado en el sistema de extracción de información Oki (Fukumoto *et al.*, 1998) está formado por tres módulos o componentes principales: (1) el módulo de reconocimiento de patrones superficiales, (2) el módulo de reconocimiento de patrones estructurales y (3) algunos programas de filtrado (etiquetado para el análisis y postprocesamiento). Para la resolución de la correferencia, el algoritmo de resolución necesita como entrada una estructura de árbol que aglutine todos los árboles de análisis de cada oración.

A continuación se muestran las características de cada módulo:

1. **Reconocimiento a nivel superficial**. El sistema de reconocimiento a nivel superficial se usa para el reconocimiento de entidades y abreviaturas. Además, este mecanismo se usa para reconocer elementos correferentes entre las entidades anteriormente detectadas. Por ejemplo, cuando el nombre de persona Mr. John Doe aparece en un texto, las palabras John y Doe se pueden usar como abreviaturas de esa persona, por lo tanto correfieren.

2. **Reconocimiento a nivel estructural**. Este es el módulo principal para la resolución de correferencias. En primer lugar, el sistema de resolución reconoce referencias de aposición y de construcciones del tipo *A es B* (construcciones copulativas). Después se extraen las expresiones anafóricas del tipo pronombre y descripciones definidas (solo las encabezadas por el artículo definido) y recorre la estructura de árbol de abajo a arriba y de izquierda a derecha para encontrar su antecedente.

3. **Programas de filtrado**. En este módulo se distinguen las tareas de:

- Etiquetado para el análisis. Toda la información referente a cada una de las palabras del texto se incorpora a él a través de etiquetas SGML (*Standard Generalized Markup Language*) (Goldfarb, 1990). Una vez añadida la información referente al reconocimiento de entidades previo y a las posibles correferencias entre ellas, se ocultan para poder realizar el análisis sintáctico.

- Post-proceso. La información de las etiquetas de correferencias se extraen de la estructura de árbol del texto generado por el análisis sintáctico y se añade al texto etiquetado. La forma de etiquetar las correferencias es mediante etiquetas SGML en las cuales cada relación se numera y su antecedente se identifica también mediante un número de antecedente.

3.5 Aproximaciones basadas en aprendizaje automático

Todos los intentos realizados en el tratamiento del problema de la referencia como un problema basado en aprendizaje han sido desarrollados como si fuese un problema de clasificación, es decir, el clasificar a dos sintagmas nominales (antecedente y descripción definida) como correferentes o no. Existen dos tipos de aproximaciones basadas en aprendizaje. La primera de ellas, se basa en el aprendizaje a partir de corpus etiquetados correferencialmente que se usan para que el sistema aprenda. A este tipo de sistemas se les denominan *sistemas supervisados*. La segunda aproximación se denominan *sistemas no supervisados* los cuales no requieren de corpus previamente etiquetados correferencialmente.

Parece obvio que el principal problema que presentan los mecanismos o sistemas supervisados es la gran cantidad de textos anotados correferencialmente necesarios para que los sistemas aprendan eficientemente y sean capaces de resolver correctamente otros textos no etiquetados. Dos de las aproximaciones más representativas de este tipo de algoritmos son el sistema MLR, desarrollado por Aone y Bennet (1995b), y el sistema RESOLVE, desarrollado

por McCarthy y Lehnert (1995). Ambos algoritmos aplican el algoritmo de inducción basado en árboles de decisión C4.5 (Quinlan, 1992) y un mecanismo para coordinar la colección de decisiones de correferencias correctas. Este mecanismo fuerza al algoritmo a cumplir la propiedad transitiva (si SN_i correfiere con SN_j y SN_j correfiere con SN_k entonces SN_i debe correferir con SN_k).

Por otro lado, las aproximaciones no supervisadas no requieren textos de entrenamiento. Una de las aproximaciones más destacadas que pertenece a esta categoría es la desarrollada por Cardie y Wagstaff (1999).

Algoritmo	Tipos	Estrategia	Corpus	Referencia
RESOLVE	Pronombre y DD	Árboles de decisión	Negocios	(McCarthy & Lehnert, 1995)
Aone & Bennet	Pronombres y DD	Árboles de decisión	Periódicos japoneses	(Aone & Bennet, 1996)
Bean & Riloff	Nueva entidad	Heurísticas	MUC-4	(Bean & Riloff, 1999)
Cardie & Wagstaff	Sintagmas nominales	Técnicas de agrupamiento	MUC-6	(Cardie & Wagstaff, 1999)

Tabla 3.2. Sistemas basados en aprendizaje

La tabla 3.2 presenta un resumen de las características principales de los diferentes sistemas basados en aprendizaje que se van a presentar de forma más detallada en los siguientes apartados.

3.5.1 Sistema RESOLVE

El sistema RESOLVE desarrollado por McCarthy y Lehnert (1995) utiliza árboles de decisión para aprender a clasificar dos sintagmas nominales como correferentes o no. El sistema utiliza el algoritmo de inducción C4.5 desarrollado por Quinlan (1992), basado en árboles de decisión, pero puede usar cualquier otro algoritmo que use vectores de características formados por pares atributo-valor. El sistema utilizó 8 atributos en el vector de características. De

las ocho características, dos contienen información de la primera referencia (antecedente), dos contienen información de la segunda (expresión anafórica) y cuatro atributos que contienen información conjunta de ambas entidades. A continuación se muestran las características usadas:

- *Nombre-i (Name-i).* ¿La referencia i contiene algún nombre? Los posibles valores son `si` y `no`. Existen dos características de este tipo, una para la expresión anafórica y otra para el antecedente.

- *Empresa resultante-i (JV-Child-i).* ¿La referencia i contiene o refiere a una empresa resultante de una fusión de empresas?, es decir, ¿refiere a una empresa formada como resultado de la fusión entre dos o más empresas o entidades? Los posibles valores son `si`, `no` y/o `desconocido`. Al igual que la característica anterior existe una para el antecedente y otra para la expresión anafórica.

- *Alias.* ¿La referencia contiene un alias de alguna de las empresas?, es decir, ¿contiene un nombre o una subcadena de alguna de las empresas? Los posibles valores son `si` y `no`.

- *Ambas forman parte de la empresa resultante (Both-JV-Child).* ¿Ambas referencias refieren a una empresa resultante de una fusión? Esta característica se define de la siguiente forma:

  – `si` cuando $\forall i$, JV-Child-i = si.
  – `no` cuando $\forall i$, JV-Child-i = no.
  – `desconocido` en cualquier otro caso.

- *Nombre común.* ¿La referencia comparte algún nombre común? Algunas referencias contienen sintagmas nominales complejos los cuales están formados por aposiciones, oraciones de relativo, etc. Esta característica compara constituyentes simples de los sintagmas nominales de cada referencia. Los posibles valores

son `si` y `no`.

- *Misma oración.* ¿Las referencias están en la misma oración? Los posibles valores que puede tomar son `si` y `no`.

3.5.2 Algoritmo de Aone y Bennet

Aone y Bennet (1996), (1995a) y (1995b) presentan un algoritmo entrenable de forma automática para la resolución de la anáfora en japonés, denominado MLR (Machine Learning-based Resolver) evolución del MDR (Machine Designed Resolver)[4]. Este algoritmo usa el método de los árboles de decisión desarrollado por Quinlan (1992) y utiliza como corpus de entrenamiento artículos de periódicos japoneses etiquetados con información del discurso. Como se ha visto anteriormente en el sistema RESOLVE (McCarthy & Lehnert, 1995), el método de árboles de decisión C4.5 usa una serie de características. Aone y Bennet usaron para el entrenamiento 66 características, agrupadas en función del tipo de información que usan:

- **Información léxica**, como por ejemplo la categoría.
- **Información sintáctica**, como por ejemplo el papel gramatical (sujeto, objeto directo, objeto indirecto, etc.).
- **Información semántica**, como por ejemplo la clase semántica proporcionada por el analizador semántico.
- **Información posicional**, como por ejemplo la distancia entre la anáfora y su antecedente.

Algunas de estas características son independientes o unarias, es decir que se usan dos características diferentes para representar

[4] MDR es un sistema de resolución de la anáfora multilingual basado en aprendizaje, el cual es robusto, extensible y entrenable manualmente. Además este sistema usa información del discurso seleccionada manualmente.

la información de la expresión anafórica y del posible antecedente. Como por ejemplo el valor del número morfológico de cada uno de ellas (`singular` o `plural`). Otras características son conjuntas o binarias, es decir muestran características concernientes a la relación entre la expresión anafórica y su antecedente. Como por ejemplo la relación posicional entre la anáfora y el antecedente.

Se han empleado diferentes métodos de entrenamiento que utilizan los siguientes tres parámetros:

- *Cadena anafórica.* Se usa en la selección de los ejemplos de entrenamiento. Cuando su valor es `ON` se selecciona el conjunto de ejemplos de entrenamiento, tanto positivos como negativos. Los ejemplos positivos son aquellos cuya expresión anafórica aparece directamente enlazada con su antecedente en el corpus etiquetado y aquellos cuya expresión anafórica aparece emparejada con uno de los antecedentes de su cadena anafórica[5] para mantener la propiedad transitiva. Los ejemplos negativos son los emparejamientos de la expresión anafórica con todos los antecedentes previos menos con los que aparecen en su cadena anafórica. Cuando el valor de la cadena anafórica es `OFF`, sólo se tienen en cuenta como ejemplos positivos los que están directamente enlazado en el corpus.

- *Identificación del tipo de anáfora.* El parámetro para la identificación del tipo de anáfora se usa para el entrenamiento del árbol de decisión. Cuando el valor es `ON`, el árbol de decisión se entrena para responder `NO` cuando un par expresión anafórica-antecedente no es correferencial y para responder el tipo de anáfora cuando es correferencial. Si el parámetro está en `OFF`, el árbol de decisión se entrena para responder `SI` o `NO` en función de si el par es correferencial o no.

- *Factor de confianza.* El factor de confianza puede tomar valores de `0` a `100` y se usa para podar el árbol de decisión. Cuanto más alto es el factor de confianza menos se poda el árbol. Cuanto

[5] En este trabajo se ha denominado cadena de correferencia

menor es el factor de confianza más se tiende a podar el árbol y por tanto conseguir árboles más generales. En estos trabajos se han utilizado los valores de 25, 50,75 y 100%.

3.5.3 Algoritmo de Bean y Riloff

Bean y Riloff (1999) proponen un algoritmo basado en el aprendizaje del corpus para la distinción entre las descripciones definidas que introducen una nueva entidad en el discurso de las que hacen referencia a una entidad lingüística previamente introducida. Este algoritmo usa métodos estadísticos para generar listas de descripciones definidas no anafóricas, llamadas por Bean y Riloff como *existenciales*, y patrones de descripciones definidas extraídos del corpus de entrenamiento. El algoritmo usa estas listas y patrones para el reconocimiento de las descripciones definidas existenciales en nuevos textos.

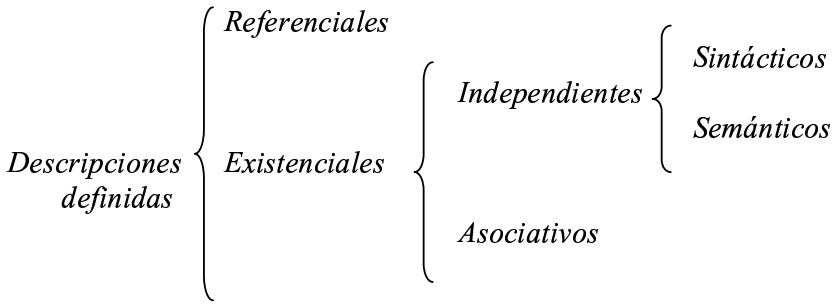

Figura 3.6. *Clasificación de las descripciones definidas (Bean & Riloff)*

Bean y Riloff presentan una clasificación de los diferentes tipos de descripciones definidas, tal y como muestra la figura 3.6. En esta clasificación se presenta la división de las descripciones definidas en dos categorías principales: *referenciales* y *existenciales*. Las descripciones definidas referenciales son aquellas que hacen referencia a una entidad lingüística previamente introducida. Las descripciones existenciales son aquellas que no hacen referencia a una entidad lingüística previamente introducida pero el lector tie-

ne de dicha entidad una representación cognitiva en su mente. Por ejemplo, cuando un lector lee el sintagma the F.B.I. o the White House, puede no hacer referencia a entidades previas pero el lector tiene una representación mental de la entidad. A su vez las descripciones definidas existenciales se dividen en dos subcategorías: *independientes* y *asociativas*. Las descripciones definidas asociativas son aquellas que aparecen inherentes a una acción, evento, objeto u otro contexto. Como por ejemplo, en un texto acerca de baloncesto podemos encontrar descripciones definidas del tipo the score, the hoop, aunque no deben de tener antecedentes previos en el texto. Las descripciones definidas independientes son aquellas que pueden entenderse por sí solas. A su vez se puede subdividir este tipo en *sintácticas* y *semánticas*. Las independientes semánticamente son aquellas que son comprendidas por el lector porque comparte un conocimiento del evento y un conocimiento del mundo. Por ejemplo, se entiende el significado de the F.B.I. sin necesidad de ninguna información adicional. Las descripciones definidas existenciales independientes sintácticamente son aquellas que están modificadas estructuralmente. Por ejemplo en

the man who shot Liberty Valence

el sintagma the man es existencial independiente sintácticamente porque la cláusula relativa identifica su único referente.

La clasificación desarrollada por Bean y Riloff tiene su correspondencia con la clasificación de Prince (1981). El tipo existencial independiente se corresponde con el tipo *nueva clase* de Prince, el tipo existencial asociativo con el tipo *inferible* y el tipo referencial con las de tipo *evocada*

El algoritmo desarrollado por Bean y Riloff se centra sólo en la identificación de las descripciones definidas existenciales independientes debido a que las asociativas son sólo un 10% de las descripciones definidas existenciales que suelen aparecer en un texto. El algoritmo utiliza los siguientes métodos para su reconocimiento:

- **Heurísticas sintácticas**. Se crean una serie de reglas sintácticas a partir de un estudio de los pre y postmodificadores de los sintagmas nominales existenciales. A menudo se encuentran nombres propios en premodificadores de un sintagma nominal. Por ejemplo, the U.S. president no es ambiguo, pero the president sí. A veces también aparecen restricciones en cláusulas de relativo, aposiciones y sintagmas preposicionales. Como por ejemplo

  the president of the United State

  es existencial debido al sintagma preposicional of the United State o en

  the president who governs U.S

  es existencial debido a la cláusula de relativo who governs U.S. También desarrollaron un conjunto de reglas para reconocer descripciones definidas referenciales. La mayoría de ellas son de la forma *the* <*número*> <*nombre*>, como por ejemplo la DD

  the 12 men

  tiene un antecedente, ya que en el contexto se habrá citado a un grupo de personas. También clasifican como referenciales aquellas descripciones definidas cuyo núcleo haya aparecido anteriormente, se corresponden con los usos anafóricos directo en la clasificación de Poesio y Vieira.

- **Extracción de la primera oración**. Como es obvio, una descripción definida referencial refiere a una entidad que ha aparecido en una posición anterior en el texto. Este hecho sirve como base para la identificación de las descripciones definidas existenciales. Si la descripción definida aparece en la primera oración de un texto es existencial. Se usa un corpus de entrenamiento para generar una lista de posibles DDs existenciales, mediante la extracción de todas las descripciones definidas de la primera

oración de cada texto y si no cumplen las reglas heurísticas para la identificación de referenciales se introducen en dicha lista de posibles existenciales.

- **Patrones de núcleos existenciales** (EHP, *Existencial Head Patterns*). Bean y Riloff estudiaron las descripciones definidas introducidas en la lista anterior y observaron una serie de características. Estas características se plasmaron en una serie de patrones con la siguiente forma *the* $<x+>$ $<nombre_1$ ... $nombre_N>$, donde $<x+>$ representa una o más palabras. Como por ejemplo

  the $<x+>$ government

  o

  the $<x+>$ Salvadoran government

- **Lista de sólo definidos** (DO, *Definite-only List*). En algunas ocasiones hay sintagmas existenciales que sólo aparecen en construcciones definidas, como por ejemplo the F.B.I. Para crear esta lista se recorre el texto y se crea una lista con todas las descripciones definidas y sus frecuencias. En un segundo recorrido se buscan todas las apariciones en construcciones indefinidas de los núcleos identificados en el primer recorrido. Finalmente se obtiene una lista ordenada de todas las descripciones definidas con su probabilidad de uso en una construcción definida y luego por su frecuencia de uso como definida.

Al estar basado en heurísticas, el método puede ser en ocasiones incorrecto. Dos errores que habitualmente tienen lugar son:

1. Incorrectas asunciones en la identificación de descripciones definidas en la primera oración de un texto. A veces una descripción definida de la primera oración puede ser referencial. Esta mala clasificación de una descripción definida puede ocasionar

la mala clasificación de otras muchas descripciones definidas.

2. Existencialismo ocasional. Algunas descripciones definidas existenciales en un texto pueden ser referenciales en otro. Por ejemplo, para un lector de periódicos salvadoreños la descripción definida the guerrillas puede ser existencial ya que todos conocen que se refiere a fuerzas contra-gubernamentales. Pero en otros textos puede aparecer

   a group of FLMN rebels attacked the capital...

   y posteriormente aparecer the guerrilla, en este caso es referencial.

Para solucionar estos problemas Bean y Riloff desarrollaron lo que denominaron una *vacuna (vaccine)*. Para detectar errores en la extracción de descripciones definidas de la primera oración usaron probabilidades definidas. Si la probabilidad era superior a un umbral, entonces se consideraba la descripción definida como existencial y en caso contrario no se clasificaba. Para solucionar los existenciales ocasionales se observaba dónde tenía lugar la primera aparición en el texto de la descripción definida, si aparecía en las primeras tres oraciones se clasificaba como existencial y en otro caso no se clasificaba.

3.5.4 Algoritmo de Cardie y Wagstaff

Cardie y Wagstaff (1999) presentan un algoritmo no supervisado para el resolución de las referencias lingüísticas producidas por los sintagmas nominales. La principal diferencia de este método con otros métodos similares es que trata el problema de la correferencia como un problema de *agrupamiento* (clustering). El algoritmo de agrupamiento proporciona un mecanismo flexible para la coordinación de restricciones y preferencias dependientes e independientes del contexto que permite la división de los sintagmas nominales en clases correferenciales equivalentes. Al igual que otros métodos (McCarthy & Lehnert, 1995) utiliza un vector de

características atributo-valor sobre el que se aplica el algoritmo de agrupamiento.

Todos los sintagmas nominales usados para describir un mismo concepto específico, una misma entidad del mundo real, suelen estar relacionados o estar muy *cerca* unos de otros . Es decir, existe una distancia muy pequeña al comparar los pares atributo-valor de ambos sintagmas. Por este motivo, se pueden usar técnicas de agrupamiento para agrupar en una misma clase todos los sintagmas que hacen referencia a la misma entidad. Los sintagmas con una distancia mayor que un valor umbral, r, no se agrupan en la misma clase o grupo. La distancia entre dos sintagmas nominales se mide en función a una serie de características que representan cada sintagma nominal. El algoritmo desarrollado por Cardie y Wagstaff usa un localizador de sintagmas nominales simples (Cardie & Pierce, 1998). Este localizador separa los sintagmas nominales complejos (sintagmas nominales con sintagmas preposicionales) en varios sintagma nominales simples.

Cada sintagma nominal se representa por el siguiente conjunto de 11 características:

- **Palabras individuales**. Las palabras que forman el sintagma nominal se almacenan como una característica.

- **Núcleo**. La última palabra del sintagma nominal se considera como el núcleo de dicho sintagma.

- **Posición**. Los sintagmas nominales se numeran secuencialmente empezando por el principio del documento.

- **Tipo de pronombre**. Si el sintagma nominal es un pronombre, se marca con el tipo del pronombre: `Nominativo`, `Acusativo`, `Posesivo` o `Ambiguo`. En cualquier otro caso, esta característica se marca como `Ninguna`.

- **Artículo**. Cada sintagma nominal se marca con el tipo del artículo que lo encabeza: `Indefinido` o `Definido`. En caso con-

trario se etiqueta como `Ninguno`.

- **Aposición**. Esta característica indica si el sintagma nominal tiene lugar en una construcción apositiva.

- **Número**. Esta característica indica si el sintagma nominal es `plural` o `singular`.

- **Nombre propio**. Indica si el sintagma nominal está formado por un nombre propio.

- **Clase semántica**. Para obtener la clase semántica, el algoritmo hace uso del recurso léxico WordNet (Fellbaun, 1998). Los núcleos de un sintagma nominal se caracterizan como `Tiempo`, `Ciudad`, `Animal`, `Humano` u `Objeto`. Si el núcleo del sintagma no pertenece a ninguna de estas clases se clasifica como su padre inmediato en la jerarquía de clases. Además, el sistema usa un algoritmo para reconocer `Numeros`, `Monedas` y `Empresas`.

- **Género**. Esta característica indica el género del sintagma nominal: `Masculino`, `Femenino` o `Neutro`. Para determinar el género se usa WordNet y una lista de nombres para los nombres propios.

- **Animado**. Para los sintagmas nominales cuya característica semántica sea `Animal` o `Humana` se clasifica como `Animado` y para los demás como `Inanimado`.

El sistema utiliza la siguiente fórmula para el cálculo de la distancia entre dos sintagmas nominales:

$$dist(NP_i, NP_j) = \sum_{f \in F} w_f \star incompatibility_f(NP_i, NP_j)$$

donde F representa el conjunto de características representada anteriormente; la función $incompatibility_f$ devuelve un valor entre 0 y 1 en función del grado de compatibilidad entre la característica

f de ambos sintagmas nominales, y w_f representa la importancia (peso) de la característica f en el conjunto de características. En la tabla 3.3 se muestra los valores usados por el algoritmo.

Característica	**Peso**	**Función incompatibilidad**
Palabras individuales	10.0	(N. palabras no iguales / N. palabras en el NP más grande)
Núcleo	1.0	1 si el núcleo difiere, en otro caso 0
Posición	5.0	diferencias en posición / máxima diferencia en el documento
Pronombre	r	1 si NP_i es un pronombre y NP_j no, en otro caso 0
Artículo	r	1 si NP_j es indefinido y no es aposición, sino 0
Palabra-Subcadena	$-\infty$	1 si NP_i está incluido en NP_j
Aposición	$-\infty$	1 si NP_j es una aposición y NP_i es su predecesor, en otro caso 0
Número	∞	1 si es no concuerda en número, sino 0
Nombre propio	∞	1 si ambos son nombres propios, sino 0
Clase semántica	∞	1 si no concuerda en clase, sino 0
Género	∞	1 si no concuerda en género, sino 0
Animado	∞	1 si no concuerda en animado, sino 0

Tabla 3.3. Incompatibilidades y pesos usados

El algoritmo de agrupamiento, una vez dado el umbral (r), asigna a cada sintagma nominal su clase. Compara cada uno de los sintagmas nominales con los anteriores en el texto mediante la aplicación de la formula de la distancia. Dos sintagmas son candidatos a agruparse a menos que exista alguna incompatibilidad entre alguno de los sintagmas nominales perteneciente a alguna de las dos clases (la de la expresión anafórica y la del antecedente). De este modo, el algoritmo de agrupamiento mantiene

automáticamente la propiedad transitiva de la relación de correferencia.

3.6 Comparación de resultados

La comparación de resultados no es una tarea sencilla debido a las diferentes características de cada uno de los algoritmos presentados. Por un lado, no todos los algoritmos tratan los mismos tipos de descripciones definidas, ni tratan los mismos tipos de relaciones entre la expresión anafórica y su antecedente, ni todos los algoritmos realizan las mismas tareas (identificación previa, resolución de la anáfora directa, indirecta, nuevas entidades en el discurso, etc.). Por otro lado, no todos los algoritmos utilizan los mismos corpus, ni están desarrollados para los mismos lenguajes.

Por todo ello, la realización de una comparación directa entre los resultados obtenidos por los algoritmos presentados anteriormente es imposible y se realizará una comparación indirecta.

A continuación se definen las medidas de evaluación usadas por los diferentes algoritmos que se han presentado, así como una comparación de los resultados alcanzados por cada uno de los sistemas pertenecientes a ambas aproximaciones.

3.6.1 Medidas de evaluación

Como se ha visto, se han desarrollado diferentes estrategias y métodos para resolver un mismo problema. Esto nos puede llevar a pensar qué estrategia es mejor y dentro de cada estrategia qué algoritmo es mejor utilizar. Desde el punto de vista de la efectividad, para evaluar los sistemas de una forma cuantitativa en la realización de la tarea y para poder comparar los diferentes sistemas entre ellos, se usan tres medidas del campo científico del Procesamiento del Lenguaje Natural como son la *precisión*, la *cobertura* y la *medida-F*. Estas medidas miden exclusivamente la efectividad de la realización de las tareas sin tener en cuenta otros factores como pueden ser el tiempo de computación, costes de los recursos a utilizar, etc.

Las tres medidas se definen de la siguiente manera:

- *Cobertura* (Recall). Número de soluciones correctas del total existentes.

$$R = \frac{N.\,de\,soluciones\,correctas}{N.\,de\,expresiones\,anaf.\,existentes}$$

- *Precisión* (Precision). Número de soluciones correctas del total tratadas.

$$P = \frac{N.\,de\,soluciones\,correctas}{N.\,de\,soluciones\,propuestas}$$

- *Medida-F* (F-measure). Es una medida, definida por Rijsbergen (1979) proporciona una medida combinada de precisión y cobertura.

$$Medida - F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R}$$

  donde β representa el peso de la cobertura frente a la precisión. El valor típico usado es de 1 para dar la misma importancia a la precisión que a la cobertura.

El problema que presenta la medida de la precisión y de la cobertura es que cualquier intento de mejorar una de ellas con lleva la disminución de la otra. A la hora de comparar dos sistemas surge la duda de cual elegir como mejor, si el que tiene mayor precisión o el que tiene mayor cobertura. Es decir, si se prefieren los sistemas que aunque resuelvan pocas cosas lo hagan de una forma correcta, o por el contrario, si se prefieren los sistemas que sean capaces de tratar muchos tipos de problemas aunque no los resuelvan todo lo eficientemente que fuera deseable. Para evitar este tipo de decisiones y para poder comparar sistemas con valores muy similares surge la medida-F. El problema que presenta la medida-F radica en el valor de β utilizado. Normalmente cada

autor usa el valor de β que mejor se ajuste a sus intereses, por lo que habrá que calcular de nuevo los valores de medida-F de todos los sistemas para poder realizar la comparación.

3.6.2 Resultados

En esta sección se presentan los resultados obtenidos por cada uno de los autores anteriores en el desarrollo de sus sistemas. Posteriormente se realiza una comparación de todos los resultados obtenidos.

Los resultados obtenidos por los autores estudiados son:

- *Poesio y Vieira.*
  La evaluación de su algoritmo se realizó en diversas etapas, tal y como se muestra en la tabla 3.4. En la resolución de las referencias producidas por la anáfora directa se obtuvo un 62% de cobertura, un 83% de precisión y un 71% de medida-F. En la tarea de la identificación de las nuevas entidades del discurso, los resultados obtenidos fueron de una cobertura del 69%, una precisión del 72% y una medida-F del 70%. En la resolución de las referencias producidas por las anáforas indirectas, los resultados obtenidos fueron de un 28% (30/106) de precisión y un 17.7% (30/204) de cobertura. En general, el algoritmo de Poesio y Vieira obtiene una cobertura del 53%, una precisión del 76% y una medida-F del 63%.

Tipo	**Precision**	**Cobertura**	**Medida-F**
Nueva entidad	72%	69%	70%
anáfora directa	83%	62%	71%
anáfora indirecta	28%	17.7%	
Total	76%	53%%	63%

Tabla 3.4. Resultados obtenidos por el algoritmo de Vieira y Poesio

- *Kameyama.*
  El algoritmo fue evaluado en el sistema de extracción de infor-

mación FASTUS sobre un corpus de 30 textos del dominio de negocios geo-políticos. Los resultados obtenidos, tabla 3.5, fueron de una cobertura del 59% y una precisión del 72% para el sistema completo. El algoritmo no sólo resuelve las descripciones definidas, sino también resuelve otras categorías gramaticales como los pronombres o los nombres propios. Los resultados obtenidos para cada uno de estos tipos fueron de una precisión del 69% para los nombres propios, un 62% de precisión para los pronombres y del 46% para las descripciones definidas. En la resolución de las descripciones definidas no se realiza una identificación previa de las descripciones definidas anafóricas de las no anafóricas, sino que intenta resolver todas y propone un antecedente para aquellas que son anafóricas. En los resultados presentados no se tiene en cuenta el éxito o fracaso obtenido en la identificación de las no anafóricas.

Tipo	**Cobertura**	**Precisión**
descripción definida		46%
pronombre		62%
nombre propio		69%
Total	59%	72%

Tabla 3.5. Resultados del algoritmo de Kameyama

- *LaSIE-II.*
  La evaluación del sistema LaSIE-I en la MUC-6 fue de una cobertura del 54.19% (801/1478) y una precisión del 69.83% (801/1147). La mejora de este mecanismo de resolución de correferencias dio origen al LaSIE-II que fue presentado en la MUC-7. La evaluación de este mecanismo se realizó sobre un total de 30 documentos, en la que se alcanzó una cobertura del 56.1%, una precision del 68.8% y una medida-F del 61.8%.
- *LOLITA.*
  La evaluación de este algoritmo se realizó en la MUC-7, después de aplicar el sistema a un total de 20 artículos diferentes se obtuvo una cobertura del 46.9%, una precisión del 57% y

Tipo	Precisión	Cobertura	Medida-F
MUC-7	68.8%	56.1%	61.8%
MUC-6	69.83%	54.19%	

Tabla 3.6. Resultados del algoritmo de correferencias del sistema LaSIE

una medida-F del 51.5%. Una de las principales razones de los bajos resultados obtenidos en la resolución de la correferencia fue debida a la utilización de un análisis completo. Otro de los problemas que se encontraron fue la falta de recursos que proporcionen la información necesaria para resolver las correferencias producidas por sintagmas nominales coordinados. Además de estos errores se encontraron errores triviales en algunos procesos, como en el reconocimiento de entidades que ocasionaron un aumento de los resultados al evaluar nuevamente el sistema corregido con los textos utilizados en la MUC-7 . En la tabla 3.7 se muestran ambos resultados.

Tipo	Precisión	Cobertura	F-medida
MUC-7	57%	46.9%	51.5%
Después de errores	58.6%	48%	52.8%

Tabla 3.7. Resultados del algoritmo de correferencias del sistema LOLITA

- *Oki.*
  Los resultados obtenidos por este sistema en la MUC-7 fueron de una cobertura del 32.9% (26/79), una precisión del 55.3% (26/47) y una medida-F del 41.3%.

- *RESOLVE.*
  Este sistema fue aplicado a textos pertenecientes al dominio de los negocios, concretamente al de las fusiones empresariales. En este dominio el sistema alcanzó resultados de precisión del 87.6% y de cobertura del 85.4%.

- *Aone y Bennet.*
  La evaluación del algoritmo MLR se hizo usando diferentes factores de confianza. Los mejores resultados se obtuvieron con un factor de confianza de 75. Se obtuvo una cobertura del 69.73%, una precisión del 86.73% y una medida-F del 77.30%.

- *Bean y Riloff.*
  Este algoritmo usó un corpus de entrenamiento formado por 1600 textos y un conjunto de 50 textos como corpus de evaluación. Los resultados obtenidos por este algoritmo fueron de un 81.8% de cobertura y un 85.65% de precisión.

- *Cardie y Wagstaff.*
  Para la evaluación de este método se usaron dos corpus, uno de entrenamiento y otro de evaluación. Cada uno de estos corpus estaba formado por 30 documentos. En el corpus de entrenamiento estaban anotadas sus cadenas de correferencias y se usó para el cálculo de la medida de distancia, así como para la selección del valor de agrupamiento r. Al aplicar el algoritmo sobre el corpus de entrenamiento se observó que los mejores resultados se obtenían con un $r = 4$ (cobertura = 48.8%, precisión = 57.4%, medida-F = 52.8%). Los resultados obtenidos al aplicar el algoritmo de agrupamiento al corpus de evaluación con el mismo valor r fue de una cobertura del 52.7% y una precisión del 54.6%y una medida-F del 54%.

En la tabla 3.8 se muestra un resumen de los resultados obtenidos por cada uno de los algoritmos presentados anteriormente. Como puede apreciarse, los valores de β utilizados para calcular la medida-F son distintos. Por tanto, el valor de la medida-F no puede usarse para comparar entre sí los diferentes algoritmos. Por esta razón y porque algunos de los autores no han realizado el cálculo de esta medida, en la tabla 3.9 se presentan los valores de medida-F obtenidos para cada uno de los algoritmos utilizando un valor de $\beta = 1$, lo que equivale a aplicar la siguiente formula:

$$Medida - F = \frac{2 \times P \times R}{P + R}$$

Algoritmo	C	P	F	β	Referencia
Vieira y Poesio	53.00%	76.00%	63.00%	0.95	(Vieira, 1998)
Kameyama	59.00%	72.00%			(Kameyama, 1997)
LaSIE-II	56.10%	68.80%	61.80%	1	(Humphreys *et al.*, 1998)
LOLITA	48.00%	58.60%	52.80%	0.70	(Garigliano *et al.*, 1998)
Oki	32.90%	55.30%	41.30%	1	(Fukumoto *et al.*, 1998)
RESOLVE	85.40%	87.60%			(McCarthy & Lehnert, 1995)
Bean y Riloff	81.80%	85.65%			(Bean & Riloff, 1999)
Aone y Bennet	69.73%	86.73%	77.30%	0.70	(Aone & Bennet, 1996)
Cardie y Wagstaff	52.70%	54.60%	54.00%	0.55	(Cardie & Wagstaff, 1999)

Tabla 3.8. Resultados obtenidos por los diferentes sistemas presentados

Los resultados mostrados en la tabla 3.9 parecen revelar que en general los métodos basados en aprendizaje automático obtienen mejores resultados que los basados en aproximaciones lingüísticas. La mayoría de los basados en aprendizaje obtienen valores de medida-F de alrededor de un 80%, mientras que la mayoría de los basados en aproximaciones lingüísticas alcanzan valores de alrededor de un 60%. Este dato puede hacer pensar que la mejor opción para la resolución de las descripciones definidas es optar por una aproximación basada en aprendizaje. Pero si se estudian los dos algoritmos que se centran en la resolución exclusiva de las descripciones definidas se observa que el algoritmo desarrollado por Vieira y Poesio alcanza unos valores de Medida-F del 62.45% mientras que el algoritmo de Cardie y Wagstaff alcanza un 53.63%. Esto nos lleva a pensar que las aproximaciones basadas en aprendizaje automático son más apropiadas para la resolución de los pronombres debido a que su información semántica es mínima y casi siempre para cada tipo de pronombres se comporta

de forma parecida. En cambio para la resolución de las descripciones definidas se obtienen mejores resultados con aproximaciones lingüísticas debido principalmente a la gran cantidad de información semántica necesaria para la resolución de todos sus tipos.

Algoritmo	F
Vieira y Poesio	62.45%
Kameyama	64.85%
LaSIE-II	61.80%
LOLITA	46.94%
Oki	41.30%
RESOLVE	86.48%
Bean y Riloff	83.68%
Aone y Bennet	77.29%
Cardie y Wagstaff	53.63%

Tabla 3.9. Resultados de Medida-F con $\beta = 1$

Otros factores a tener en cuenta a la hora de elegir qué algoritmo de resolución de las referencias lingüísticas es más adecuado, además de las medidas de precisión, cobertura y medida-F son:

- *tiempo computacional*, en ninguno de los algoritmos aquí presentados, se ha hablado del tiempo que tarda en resolver cada uno de los diferentes tipos de expresiones anafóricas. Los sistemas relacionados con consultas de bases de datos documentales o de búsquedas de información en internet necesitan que la resolución de las correferencias sea muy rápida. Por lo tanto, será un factor a tener en cuenta a la hora de elegir entre varios posibles algoritmos.

- disponibilidad de recursos, aunque como se ha visto en las tablas anteriores los algoritmos basados en aprendizaje automático son más eficaces que los basados en aproximaciones lingüísticas para la resolución de las correferencias en general, tienen el inconveniente de necesitar muchos textos etiquetados correferencialmente para que el algoritmo aprenda a resolver estos tipos de referencias. Además sólo pueden resolver aquellos tipos que

estén presentes en los ejemplos. La construcción de este tipo de textos etiquetados es muy costosa y no siempre se disponen de ellos comercialmente, por lo que a veces se tendrá que decidir otra aproximación por falta de los recursos necesarios. De igual forma, si se pretende usar una aproximación lingüística y no se dispone de los recursos necesarios (información morfológica, información sintáctica y semántica) no podrá emplearse esta estrategia.

- tipo de expresión anafórica, de los resultados anteriores parece extraerse que la resolución de la anáfora pronominal a través de aproximaciones basadas en el aprendizaje automático obtiene mejores resultados, posiblemente debido a su menor carga léxica. En cambio, para la resolución de las descripciones definidas se obtienen mejores resultados con aproximaciones lingüísticas.

3.7 Conclusiones

En este capítulo se ha presentado, por un lado, una revisión de las características de las descripciones definidas junto con las clasificaciones realizadas por diversos autores. Estas clasificaciones abarcan diferentes tipos de textos y orientadas a la lengua inglesa. Por otro lado, se ha presentado una revisión de diferentes algoritmos para la resolución de las descripciones definidas desarrollados por diferentes autores y evaluados sobre diferentes textos, por lo que la realización de una comparación directa ha sido imposible y se ha tenido que realizar una comparación indirecta.

4. Fuentes de información: El uso de la información en el tratamiento y resolución de las descripciones definidas

En este capítulo se presenta un análisis de la importancia e influencia de las diferentes fuentes de información (léxica, morfológica, sintáctica, semántica y pragmática) en el tratamiento y resolución de las DDs. Posteriormente, y sobre la base de este estudio, se propone una clasificación de las DDs para el español desde el punto de vista de la información que se considera necesaria para establecer la relación expresión anafórica-antecedente.

4.1 Introducción

Cualquier sistema de Procesamiento del Lenguaje Natural basado en una aproximación lingüística necesita la utilización de fuentes de información que proporcionen el conocimiento necesario para la realización de las tareas básicas de estos sistemas. Una de estas tareas, la resolución de las correferencias, necesita de la contribución de diferentes fuentes de información para la correcta resolución. La influencia de cada una de estas fuentes en la resolución depende del tipo de expresión anafórica. Los diferentes tipos o fuentes de información se pueden clasificar en cuatro grupos principales:

- Información léxico-morfológica
- Información sintáctica
- Información semántica
- Información pragmática

A continuación se estudian las características de cada fuente de información, así como su influencia en la resolución de las DDs.

4.2 Información léxico-morfológica

Existen diferentes formas y usos de la información léxico-morfológica. Desde nuestro punto de vista consideramos que un uso apropiado de esta información sería aquella que que proporciona: (1) categoría gramatical, (2) información de concordancia y (3) lema. A continuación se describen de forma más detallada cada una de estas características.

1. *Categoría gramatical.* La categoría gramatical proporciona la información del tipo o clase de palabra (*nombre, adjetivo, verbo,* etc.). Existe lo que se denominan clases cerradas (*determinantes, pronombres, preposición*, etc.) que no suelen incorporar nuevas palabras, y clases abiertas que son aquellas clases a las que pueden añadirse nuevas palabras como son *nombre, verbo, adjetivos, adverbios*, etc.
   A cada una de las categorías gramaticales le corresponde una etiqueta representativa de su clase. Existen diferentes conjuntos de etiquetas que representan cada una de las clases o tipos de palabras. Cabe destacar el conjunto de etiquetas del Xerox POS Tagger adaptado al español en el Proyecto CRATER (CRATER, 1994), y el juego de etiquetas PAROLE definido en el Proyecto ITEM (Martí *et al.*, 1998), (Proyecto ITEM, 1996-1999). Las etiquetas no sólo tienen información del tipo de palabra, sino que además incluyen la información de concordancia. Algunos conjuntos de etiquetas sólo disponen de una etiqueta muy genérica para la clase o tipo, pero en algunos casos es conveniente que estas etiquetas genéricas se subdividan en etiquetas más específicas, por ejemplo, la etiqueta verbo (V000000). En ocasiones es mejor utilizar etiquetas más completas que añadan información, por ejemplo, del tipo de verbo modal (VM00000) o auxiliar (VA00000).

2. *Información de concordancia.* En la información de concordancia se distinguen tres características principales:

   - Concordancia de género. Esta característica indica el *género* de la unidad léxica. Esta información es importante para la

realización de varias tareas, entre ellas la tarea de análisis sintáctico. Se habla de concordancia de género porque debe existir una concordancia con el género de las unidades léxicas pertenecientes al mismo constituyente sintáctico. Por ejemplo, en un sintagma nominal formado por un determinate y un nombre, no puede tener el determinante un género y el nombre otro. Es decir, las construcciones del tipo La coche son incorrectas ya que no hay concordancia de género. Los posibles valores del género son tres: `masculino`, `femenino` y `neutro`.

- Concordancia de número. Esta característica indica el *número* de la unidad léxica. Al igual que ocurre con la concordancia de género, debe haber una concordancia de número. Es decir, construcciones del tipo los coche son incorrectas. Los posibles valores que puede tomar esta característica son: `singular` y `plural`.

- Concordancia de persona. Esta característica indica la *persona* de la unidad léxica. También debe concordar con las unidades léxicas que le acompaña. Es decir, que no se puede tener, por ejemplo, un pronombre de una persona acompañado por un verbo de otra. Los posibles valores de esta características son: `primera persona`, `segunda persona`, `tercera persona`. Si esta característica se combina con la concordancia de número se tienen los mismos valores para el singular y para el plural (`primera persona del singular` y `primera del plural`, etc.).
  Esta característica es sólo utilizada por las clases *pronombre* y *verbo*.

La información de concordancia también es deseable que aparezca en la etiqueta, de modo que una palabra como casa tenga asociada la etiqueta NCFS000 que significa Nombre Común Femenino Singular. En cambio, la palabra casas tenga asociada la etiqueta NCFP000, que significa Nombre Común Femenino Plural. La etiqueta VAIP3S0 correspondiente al verbo auxiliar

ha significa Verbo Auxiliar Indicativo Presente 3 persona del Singular.

3. *Lema.* Las palabras aparecen en un texto de diversas formas. Las palabras, mediante la utilización de las reglas de formación, derivan en otras palabras que comparten la misma raíz o lema. El lema de una palabra es lo que sirve de entrada al lexicón, que proporciona el resto de información de concordancia en función de las reglas de formación utilizadas. Por ejemplo, casas, su lema es casa cuya entrada en el lexicón proporcionaría la etiqueta NCF0000, pero al utilizar la regla de formación del plural su etiqueta pasa a ser NCFP000.

4.2.1 Influencia de la información léxico-morfológica en la resolución de las DDs

El uso e influencia de esta información en el tratamiento de las DDs son claros: permiten identificar el tipo de las expresiones referenciales a través de la información de la categoría gramatical presente en la etiqueta (pronombres, para el caso de la anáfora pronominal, o determinantes que indican si el sintagma nominal es una descripción definida o no). Por otro lado, la información de concordancia formada por género, número y persona, es fundamental para poder elegir o descartar antecedentes como posibles soluciones a una expresión anafórica. En el tratamiento de las DDs, este tipo de información se puede utilizar para seleccionar entre varios posibles candidatos, en lugar de utilizarla para descartar candidatos como ocurre con los pronombres. El motivo de utilizar esta información para seleccionar en lugar de rechazar candidatos en la resolución de las DDs es que muchas palabras sinónimas pueden tener género distinto, como por ejemplo afueras y alrededores. En el ejemplo

La gente que vivía en *los alrededores*$_i$ de la ciudad no tenía ni para comer. *Las afueras*$_i$ no son más que poblados de chabolas donde los indígenas malviven.

el sintagma nominal Las afueras hace referencia al sintagma nominal los alrededores. Si se utiliza la información de concordancia para rechazar, entonces se eliminaría el candidato correcto. En cambio, si se utiliza como preferencia puede ser elegido si no hay otras circunstancias que lo desestimen.

La diferencia entre el pronombre y la DD radica en que esta última tiene mucha más información semántica y puede no concordar con su antecedente, pero en cambio un pronombre apenas tiene información semántica y por eso ha de concordar con su antecedente, tal y como muestra el siguiente ejemplo:

Juan y Ana se comieron una paella. Él prefería pescado, pero ella prefería el arroz. Ambos no comen carne.

un sistema que resuelva la anáfora pronominal al encontrar el pronombre él, en un primer momento tomaría como posibles candidatos todos los sintagmas nominales previos (tabla 4.1).

Antecedente	Género	Número
Juan	masculino	singular
Ana	femenino	singular
Juan y Ana	masculino	plural
una paella	femenino	singular

Tabla 4.1. Tabla de antecedentes

La utilización de la concordancia de género eliminaría los sintagmas nominales Ana y una paella. La lista de candidatos quedaría formada por los elementos Juan y Juan y Ana, como muestra la tabla 4.2.

Antecedente	Género	Número
Juan	masculino	singular
Juan y Ana	masculino	plural

Tabla 4.2. Tabla de candidatos del pronombre él

Al aplicar la concordancia de número se eliminaría el sintagma nominal Juan y Ana, quedando en la lista de candidatos el sintagma Juan (tabla 4.3) que es la solución a la referencia producida por el pronombre.

Antecedente	Género	Número
Juan	masculino	singular

Tabla 4.3. Tabla de candidatos después de aplicar restricciones morfológicas

En el mismo ejemplo, para el pronombre ella, al aplicar la restricciones de género se eliminarían los sintagmas Juan y Juan y Ana[1] (tabla 4.4)

Antecedente	Género	Número
Ana	femenino	singular
una paella	femenino	singular

Tabla 4.4. Tabla de candidatos del pronombre ella

y al aplicar la concordancia de número no habría ningún sintagma a eliminar, quedando como posibles candidatos los mismos sintagmas Ana y una paella. Sólo con información morfológica no puede resolverse este pronombre, ya que se necesita información adicional para saber que Ana es del tipo persona y que los pronombres personales siempre refieren a entidades del tipo persona. Para el pronombre ambos, al aplicar la concordancia de número nos dejaría en la lista de candidatos el sintagma nominal coordinado Juan y Ana.

Como se verá en el algoritmo propuesto para las DDs, la información de concordancia, tanto de género como de número, se utiliza sólo como preferencia en lugar de restricción.

[1] Cuando existe una coordinación de sintagmas y cada uno de los sintagmas tiene un género diferente, el sintagma coordinado toma el género no marcado, que en español es el masculino.

4.3 Información sintáctica

El análisis sintáctico, núcleo de cualquier sistema de Procesamiento del Lenguaje Natural, proporciona la información sintáctica. La principal función del análisis sintáctico consiste en determinar si la oración es gramaticalmente correcta o no. Los analizadores sintácticos, como se ha visto anteriormente, pueden ser completos o parciales y pueden utilizar diferentes técnicas de análisis (descendentes o ascendentes). Los analizadores, además de indicar si la oración es correcta, proporcionan información relativa a cómo se combinan las palabras para formar constituyentes sintácticos (sintagmas nominales, sintagmas verbales y sintagmas preposicionales). Para suministrar este tipo de información los analizadores se basan en las gramáticas que proporcionan las reglas de formación de los constituyentes sintácticos. Existen diversos tipos formalismos gramaticales, entre otros, Gramáticas de cláusulas definidas (DCG, Definite Clause Grammar), Gramáticas de unificación de huecos (SUG, Slot Unification Grammar), Gramáticas léxico-funcionales (LFG, Lexical-Function Grammar), que han sido utilizadas en los trabajos desarrollados en el Grupo de Procesamiento del Lenguaje de la Universidad de Alicante.

A continuación se estudia la influencia de la información sintáctica en la resolución de las DDs.

4.3.1 Influencia de la información sintáctica en la resolución de las referencias

La referencia lingüística o anáfora, como ya se ha comentado, es un fenómeno que ocurre tanto entre diferentes oraciones (*interoracional*) como dentro de la misma oración (*intraoracional*). La información sintáctica tiene mayor influencia en la referencia intraoracional que la interoracional. En la referencia interoracional, dicha información se utiliza para seleccionar entre varios candidatos, y no para rechazar. En los trabajos de Ferrández *et al.* (Ferrández *et al.*, 1998b), se recoge que en la resolución del pronombre en referencias interoracionales se utiliza para seleccionar candidatos información sintáctica del tipo: constituyentes en la

misma posición con respecto al verbo, o que acompañe al mismo verbo.

Sin embargo, en la literatura se encuentran diferentes estudios de la influencia de la información sintáctica en la resolución de las referencias intraoracional. En concreto, existen dos estudios que definen un conjunto de reglas que impiden la referencia entre dos constituyentes de la misma oración. Este tipo de reglas se usan para descartar candidatos. Estos estudios se centran principalmente en las condiciones de no correferencialidad de los pronombres. De todas las condiciones que proponen estos dos estudios sólo una hace referencia a los sintagmas nominales. Esto es debido a que la expresión gramatical normalmente utilizada para referir dentro de la misma oración es el pronombre, mientras que las DDs se utilizan normalmente para referir a objetos o entidades lingüísticas alejadas en el texto.

Los estudios a los que se hace referencia son debidos a Reinhart (1983) y a Lappin y Leass (1994). A continuación se muestra las reglas utilizadas por cada uno de ellos:

1. Reinhart (1983). Reinhart presenta las reglas de *C-dominio* para identificar la incompatibilidad referencial entre dos constituyentes de la misma oración. Se dice que un nodo A *c-domina* a otro B, si el nodo C que domina inmediatamente a A, o bien domina B, o es inmediatamente dominado por el nodo D que domina a B, y D es de la misma categoría que C. Se define el *dominio* del nodo A como el conjunto de todos los nodos que c-domina. Reinhart utiliza estas definiciones para establecer las siguientes reglas:

   a) Un sintagma nominal (SN) no pronominal es no correferencial con cualquier SN completo que lo c-domine.

      The house of the beach is very dirty

      En este ejemplo, los sintagmas the house y the beach no pueden correferir aunque la información de concordancia lo permita.

Rosa feels that Rosa has a severe headache

En este segundo ejemplo, el primer sintagma nominal Rosa domina al segundo sintagma nominal, que es no pronominal, Rosa. Por lo tanto según esta regla los dos sintagmas no correfieren, es decir debe de tratarse de dos personas distintas que se llaman Rosa.

b) Un pronombre debe ser interpretado como no correferencial con cualquier SN completo que lo c-domine.

John and him played tennis yesterday

En este ejemplo, el pronombre him no refiere al SN John al estar los dos dominados por el sintagma nominal coordinado John and him.

c) Un pronombre reflexivo o recíproco es correferencial con el SN que lo c-domina.

Andrew rubbed himself with the towel

En este ejemplo el pronombre reflexivo himself hace referencia a Andrew que es el SN que lo c-domina.

d) Un pronombre no reflexivo ni recíproco es no correferencial con cualquier SN que lo c-domine.

A cousin of mine come today

En este ejemplo, el pronombre posesivo mine no hace referencia al SN a cousin ya que lo c-domina.

2. Lappin y Leass (1994). Lappin y Leass definen un conjunto de reglas de no correferencialidad para los pronombres. Un pronombre P es no correferencial con un sintagma nominal N no reflexivo y no recíproco si se da alguna de las siguientes condiciones:

a) P y N tienen características de concordancia incompatibles. Las características de concordancia son su número,

género y persona. En el siguiente ejemplo,

The woman$_i$ said that *he*$_i$ is funny

el pronombre he no concuerda en género con el sintagma nominal The woman.

b) P está en dominio argumental (*argument domain*) de N. Se dice que P está en el dominio argumental de N si y solo si P y N son ambos argumentos del mismo núcleo. En el siguiente ejemplo,

John$_i$ seems to want to see *him*$_i$

el pronombre him está en el dominio argumental de John.

c) P está en el dominio adjunto (*adjunct domain*) de N. Se dice que P está en el dominio adjunto de N si y solo si N es un argumento del núcleo H, P es objeto de la preposición PREP y PREP es un adjunto de H. En el siguiente ejemplo,

She$_i$ sat near *her*$_i$

el pronombre her está en el dominio adjunto del sintagma nominal pronominal She.

d) P es un argumento del núcleo H, N no es un pronombre, y N no está contenido por H. En el siguiente ejemplo,

He$_i$ believes that *the man*$_i$ is amusing

el pronombre he es un argumento del verbo believe y el sintagma nominal the man está contenido en el sintagma verbal del verbo believe.

e) P está en el dominio del sintagma nominal N. Se dice que P está en el dominio del sintagma nominal (*NP domain*) N si y solo si N es el determinante del nombre Q y P es un argumento de Q, o P es el objeto de la preposición PREP

y PREP es un adjunto de Q. En el siguiente ejemplo,

John$_i$'s portrait of *him*$_i$ is interesting

el pronombre him está en el dominio del sintagma nominal John.

f) P es un determinante del nombre Q y N está contenido en Q. Se dice que P está contenido en el sintagma Q si y solo si P es un argumento o un adjunto de Q (P está inmediatamente contenido en Q), o P está inmediatamente contenido algún sintagma R y R está contenido en Q. En el siguiente ejemplo,

His$_i$ portrait of *John*$_i$ is interesting

el pronombre his es un determinante del nombre portrait y el sintagma nominal John está contenido en el sintagma nominal His portrait of John.

En los corpus estudiados para este trabajo, no se han encontrado ninguna DD que haga referencia a una entidad lingüística de la misma oración, y sólo se ha encontrado algunos casos de sintagmas nominales de tipo nombre propio que sí cumplía la condición de no correferencialidad referente a los sintagmas nominales propuesta por Reinhart. Por esta razón este tipo de reglas no han sido utilizadas en este trabajo.

4.4 Información semántica

En los estudios desarrollados por Frege (1892) y Russell (1919) se definía la anáfora como un fenómeno semántico, por lo tanto disponer de este tipo de información es esencial para una correcta resolución de las referencias lingüísticas. Un ejemplo de la importancia que tiene la información semántica se puede ver en:

Juan estaba con su perro en el parque. El animal se revolcaba por el césped

si se utiliza exclusivamente la información léxica-morfológica y sintáctica para solucionar la referencia lingüística de la descripción definida el animal no se podría resolver ya que no se tiene ninguna forma de relacionar la DD con los sintagmas nominales previos (Juan, su perro y el parque).

Si se añade la información semántica, se observa que existe una relación semántica entre las palabras perro y animal. Por tanto, si se dispone de esta información se descartan los sintagmas nominales Juan y el parque por inconsistencia semántica y se elige como solución el sintagma nominal su perro.

Para el uso de la información semántica como fuente de información en la resolución de las referencias lingüísticas producidas por las descripciones definidas es necesario contar con algún recurso que proporcione información sobre sinónimos, hiperónimos, hipónimos, etc. de las palabras. En este trabajo se propone como recurso semántico el WordNet español que es uno de los módulos de los que forman el EuroWordNet (Vossen, 1998). Además de usar el WordNet español como recurso léxico, se utiliza la ontología de conceptos del EuroWordNet que es la misma para todos los WordNets que integra. En los siguientes apartados se presentan de forma más detallada el WordNet español y la ontología del EuroWordNet.

4.4.1 WordNet español

WordNet (Miller *et al.*, 1993) es una base de datos léxica para el inglés que consiste en relaciones semánticas entre diferentes palabras. Las palabras con un significado similar se agrupan formando lo que se denominan *synsets*[2]. En esta base de datos léxica no se establecen sólo relaciones de sinonimia, sino que existen otros tipos de relaciones entre los diferentes *synsets* como son, entre otras, las relaciones de hiponimia, hiperonimia, meronimia, holonimia y antonimia. En el WordNet desarrollado por Miller, versión en

[2] Conjunto de sinónimos

inglés, no se establecen relaciones entre palabras pertenecientes a diferentes categorías (nombres, verbos, adjetivos y adverbios).

El EuroWordNet (Vossen, 1998) es un recurso léxico que está formado por diferentes WordNets de diferentes lenguas europeas (inglés, holandés, español, italiano, alemán, checoslovaco, estonio y francés) relacionados a través de un módulo denominado ILI (*Inter-Lingual Index*). Una de las principales diferencias del WordNet español[3] con el WordNet (versión para la lengua inglesa) es que proporciona algunas relaciones entre distintas categorías léxicas, como son nombres y verbos (Gonzalo *et al.*, 1998). Las relaciones semánticas entre las diferentes categorías que se muestran en ese trabajo son:

- Quasisinónimo (*xpos-near-synonym*). Por ejemplo del nombre caza obtenemos el verbo cazar.
- Hiperónimos (*has-xpos-hyperonym*). Eclipse y oscuro.
- Hipónimos (*has-xpos-hyponym*). Oscuro y eclipse.
- Quasiantónimo (*xpos-near-antonym*). Desconexión y conectar.
- Papel temático (*role & involved*). Dentro de esta categoría se encuentran diferentes tipos de relaciones, para nuestros intereses destacamos la relación *involved-agent* que dado un verbo nos proporciona el nombre asociado, cazar y cazador y *role-agent* que dado un nombre nos proporciona el verbo asociado, cazador y cazar.
- Causa-efecto (*cause & is-caused-by*). Matar y muerte.
- Subevento (*has-subevent & is-subevent-of*). Lavar y mojar.

[3] Desarrollado por los grupos de Procesamiento de Lenguaje Natural de la Universidad de Educación a Distancia de Madrid, de la Universidad Politécnica de Cataluña y de la Universidad de Barcelona dentro del proyecto europeo EuroWordNet

4.4.2 Ontología del EuroWordNet

La ontología del EuroWordNet (Vossen, 1998) es la misma para todos los WordNets que integran este proyecto. El principal propósito de esta ontología es mantener la uniformidad y la compatibilidad entre los diferentes WordNets. La ontología tiene 63 categorías diferentes, también denominados *conceptos ontológicos*. La ontología agrupa los 63 conceptos en diferentes niveles. En el nivel más alto se distinguen las siguientes categorías o conceptos ontológicos denominados *conceptos principales* (*Top Concepts*): *entidades de primer orden* (1stOrderEntity), *entidades de segundo orden* (2ndOrderEntity) y *entidades de tercer orden* (3rdOrderEntity) que a su vez agrupan el resto de conceptos denominados *conceptos base* (*Base Concepts*). Un visión general de la ontología se muestra en la figura 4.1. A continuación se muestran las características de los *conceptos principales*:

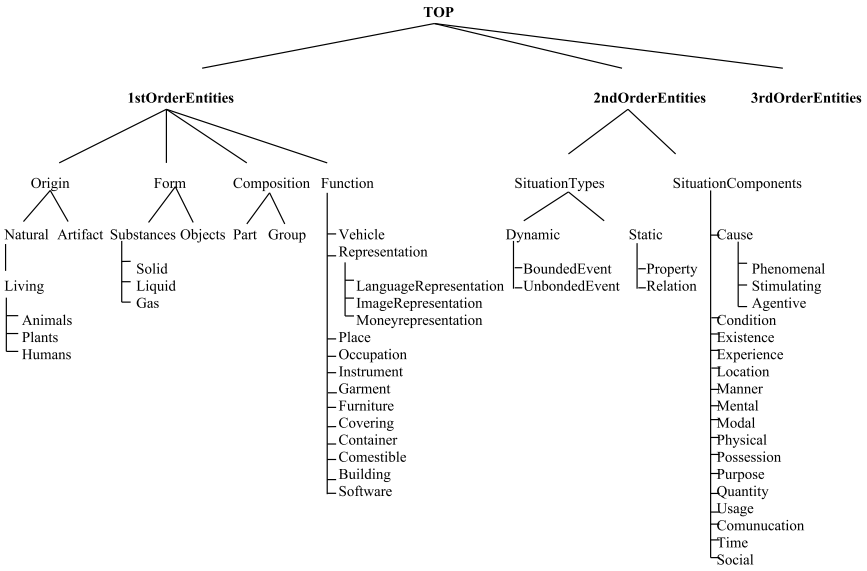

Figura 4.1. *Ontología de EuroWordNet*

1. Entidades de primer orden (*1stOrderEntity*). Dentro de las entidades de primer orden se encuadra cualquier entidad concreta

perceptible por los sentidos, localizada en cualquier punto del tiempo, o en un espacio tridimensional. Es decir, está formado por palabras referentes a personas, animales y algunos objetos y substancias físicas, como por ejemplo, vehículo, animal, substancia, objeto.
La entidades de primer orden se dividen en cuatro grupos principales:

- *Origen.* Indica la procedencia de la entidad, que puede dividirse en *natural* y *artefacto*. A su vez natural puede dividirse en *planta*, *humano* y *animal*.

- *Forma.* Indica la forma de la entidad, una substancia o un objeto. La substancia puede subdividirse en *líquido*, *sólido* y *gaseoso*.

- *Composición.* Indica si está formando un grupo o forma parte de algo. Las posibles divisiones son *grupo* y *parte*.

- *Función.* Representa la acción o actividad típica asociada a la entidad. Algunos de los posible valores son: *vehículo*, *representación*, *software*, etc.

2. Entidades de segundo orden (*2ndOrderEntity*). Dentro de esta categoría se encuadra cualquier entidad que sea una *situación estática* (*Static Situation*) o una *situación dinámica* (*Dynamic Situation*), que no pueda ser comprendida, oída, vista y sentida como una cosa físicamente independiente. Está formada por palabras referentes a eventos, procesos, o situaciones. Mientras que las entidades de primer orden existen, las de segundo orden ocurren o tienen lugar. Como por ejemplo:

   ocurrir, ser, haber, empezar, terminar, causar, resultar, continuar, ocurrir, etc.

Las entidades de segundo orden se dividen en dos grupos conceptuales:

- *Tipo de situación.* Representa la forma en la que una situación puede ser cuantificada y distribuida en función del tiempo y su dinamicidad. Podemos identificarla como una situación *estática* o *dinámica.* En general, una situación estática no supone cambios y una dinámica supone un cambio continuo o específico.

- *Componente de situación.* Es el componente semántico más importante que caracteriza una situación.

3. Entidades de tercer orden (*3rdOrderEntity*). Dentro de esta categoría se engloban las entidades referentes a cualquier proposición observable que exista de forma independiente en el tiempo y en el espacio. Son entidades que puede ser verdad o falso más que reales. También puede ser afirmada, negada, recordada u olvidada. Las palabras que forman esta categoría están relacionadas con ideas, pensamientos, teorías e hipótesis. Como por ejemplo:

   idea, pensamiento, información, teoría, plan, etc.

La primera división que realiza la ontología de los conceptos base, es una división disjunta. Es decir, una palabra con un determinado sentido no puede ser una combinación de estos conceptos de primer nivel. Las categorías pertenecientes a cada uno de los conceptos de esta primera división de la ontología son:

- Entidades de primer orden son siempre *nombres concretos*

- Entidades de segundo orden pueden ser *nombres, verbos* o *adjetivos*

- Entidades de tercer orden son siempre *nombres abstractos*

La mayoría de las subdivisiones, de segundo nivel, de la ontología son disjuntas, aunque pueden haber algunos casos en los que no. Por ejemplo, la subdivisión de las entidades de primer orden en sus cuatro grupos principales puede no ser disjunta, como se ve en la figura 4.2. En ese ejemplo, la palabra fruit[4] dentro de las entidades de primer orden puede clasificarse como una función ya que es comestible, como una forma ya que es un objeto físico, como una composición ya que esta formada por varias partes (piel, hueso, etc.) y natural ya que es una "planta". Pero en general, una entidad no puede ser, por ejemplo, natural y artefacto a la vez.

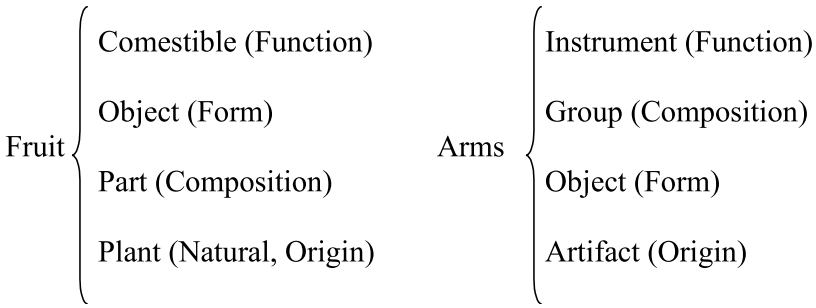

Figura 4.2. *Algunos conceptos base en la ontología*

4.4.3 Influencia de la información semántica en la resolución de las descripciones definidas

Tal y como Frege (1892) y Russell (1919) citan en sus trabajos, la información semántica juega un papel muy importante a la hora de resolver las referencias lingüísticas producidas por cualquier tipo de expresión anafórica. A continuación se muestra una serie

[4] Se utiliza el término en inglés debido a que la ontología, como es general para todas las lenguas integrantes del proyecto EuroWordNet, los utiliza en este idioma.

de ejemplos en los que la información semántica es fundamental para obtener el antecedente adecuado de una DD.

En la resolución de las DDs con el mismo núcleo puede parecer que no haga falta información semántica, ya que bastaría con estudiar los núcleos de la DD y del antecedente para ver que son iguales. Pero esto, no es realmente así, ya que en ocasiones estos núcleos están acompañados de una serie de modificadores que pueden estar relacionados semánticamente. Así, en los siguientes ejemplos,

- Preferencia entre dos candidatos. En el ejemplo

  El coche de carrera es más rápido que el coche utilitario. El coche de competición también es más caro

  Si se resuelve la referencia producida por la DD el coche de competición sin usar información semántica, se puede escoger entre los dos sintagmas nominales previos. Si se aplica alguna preferencia del tipo *más cercano* se escogería la DD el coche utilitario cuando en realidad ésta no es la solución correcta. Ahora bien, si se añade información semántica se observa que hay una relación semántica entre las palabras carrera y competición, por lo que se prefiere este antecedente como solución, siendo además la solución correcta.

- Descartar candidatos incompatibles. En el ejemplo

  Juan llevaba un sonotone en la oreja derecha. En la oreja izquierda llevaba un pendiente

  Si se resuelve sin usar información semántica se escogería como solución a la DD la oreja izquierda la otra descripción definida con el mismo núcleo la oreja derecha, siendo ésta una solución incorrecta. Al añadir información semántica se observa que entre izquierda y derecha hay una relación de antonimia, con lo cual el

candidato la oreja derecha se rechaza y se clasifica la DD como no anafórica

En el tratamiento de las descripciones definidas con distinto núcleo que su antecedente el uso de la información semántica toma mayor importancia, ya que no sólo se deben de estudiar posibles relaciones semánticas entre los modificadores, sino que también las relaciones semánticas entre ambos núcleos. En los siguientes ejemplos se presenta una muestra de las diferentes situaciones que se pueden dar en un texto.

- Núcleos relacionados por sinonimia. En el ejemplo

  El coche es uno de los medios de transporte más usados. El automóvil da libertad de movimiento

  En este ejemplo existe una relación de sinonimia entre coche y vehículo. Sin este tipo de información no podría resolverse la referencia de forma correcta.

- Núcleos relacionados por hiperonimia o hiponimia. En el ejemplo

  El perro es es el mejor amigo del hombre. El animal es dócil y leal

  En este ejemplo existe una relación de hiponimia. Perro es un hipónimo de animal. Sin este tipo de información no podría resolverse la referencia de forma correcta.

- Núcleos relacionados por meronimia u holonimia. En el ejemplo

  El coche se detuvo a los pocos metros de la meta. El motor estaba ardiendo

  En este ejemplo existe una relación de meronimia entre coche y motor. Sin este tipo de información no podría resolverse la

referencia de forma correcta.

- Papel temático. En el ejemplo

  Juan Pérez vende su casa a Fernando Gómez. El vendedor se hace cargo de las costas

  En este ejemplo existe una relación entre dos categorías gramaticales diferentes (nombre y verbo). El nombre vendedor y el verbo vender, por lo tanto el que realiza la acción de vender es el que se referencia a través de la descripción definida el vendedor. Este tipo de información está presente en el WordNet español a través de las relaciones de *involve-agent* y *role-agent*.

Como se ha visto en los ejemplos anteriores, en la mayoría de casos la información semántica es fundamental para poder elegir el candidato adecuado a cada DD o para rechazar candidatos incompatibles.

4.5 Información pragmática

Los tres tipos de información presentados en las secciones anteriores no son suficientes para la resolución de todos los tipos de DDs. Existen algunas DDs cuya resolución se basa en el conocimiento del mundo. Esta información, denominada *información pragmática*, proporciona un conocimiento variable con el tiempo y conocimiento asumido por todos los que intervienen en el discurso pero que no aparece explícitamente. Por ejemplo cuando se habla de

el presidente del Comité Olímpico Internacional (COI)

todo el mundo lo asocia en estos momentos con Juan Antonio Samaranch pero con el paso del tiempo esta información cambiará porque será otra persona la que ocupe ese cargo. También, cuando

se habla de la Giralda y nos encontramos con en esa ciudad andaluza, todo el mundo sabe que se está refiriendo a Sevilla que es la ciudad donde se encuentra la Giralda.

Tomando como base el análisis de las diferentes fuentes de información y la influencia de éstas en la resolución de las DDs, a continuación se muestra la clasificación que hemos desarrollado para el español de los diferentes tipos de descripciones definidas en función del tipo de información necesaria para poder establecer la relación entre el antecedente y la DD.

4.6 Propuesta de clasificación de las DDs para el español.

Se han desarrollado diferentes taxonomías por diferentes autores para la lengua inglesa que clasifican las DDs en diversas categorías en función a varios conceptos, como pueden ser el tipo de relaciones con su antecedente, los conocimientos del hablante y oyente, etc. En este trabajo se ha desarrollado una clasificación para las DDs en español en función del tipo de información necesaria para poder establecer la relación entre la DD y el antecedente, en el caso que exista. Aunque en principio no debe haber muchas diferencias con el inglés, en multitud de ocasiones el artículo determinado el no es traducido al artículo the inglés. Esto es debido a que las DDs en inglés se usan sobre todo para referenciar una entidad que ha aparecido anteriormente o para hablar de un único objeto. Por lo que el número de DDs en español es mucho mayor que en inglés.

En la clasificación presentada por Poesio y Vieira, aunque si bien se abarcan todas las posibles DDs que se pueden encontrar en un texto, no se tiene en cuenta el tipo de información necesaria para poder establecer la relación entre el antecedente y la expresión definida.

A continuación se presenta la clasificación de las DDs propuesta, obtenida a partir del estudio de un fragmento del corpus LE-

XESP[5] (Muñoz *et al.*, 2000a), de los estudios de otras clasificaciones de las DDs para el inglés, y del análisis de las fuentes de información presentado anteriormente. Dicho fragmento está formado por un total de 91370 palabras distribuidas en 2737 oraciones. Este estudio se ha realizado en función del nivel de representación lingüística al que pertenece la información necesaria para poder establecer la relación entre antecedente y descripción definida. Los niveles de representación lingüística son cuatro: *nivel fonológico, nivel morfológico, nivel sintáctico-semántico-textual* y *nivel pragmático*. De estos cuatro niveles, sólo los dos últimos se utilizan para poder establecer las relaciones entre antecedente y descripción definida.

La figura 4.3 recoge la clasificación propuesta para las DDs. A continuación se describe con detalle cada una de las categorías descritas en dicha clasificación:

Figura 4.3. *Clasificación propuesta de las descripciones definidas*

[5] LEXESP es un corpus en español que contiene 5 millones de palabras etiquetadas léxicamente perteneciente al proyecto del mismo nombre desarrollado por el Departamento de Psicología de la Universidad de Oviedo y por el Grupo de Lingüística Computacional de la Universidad de Barcelona, con la colaboración del Grupo de Tratamiento del Lenguaje de la Universidad Politécnica de Cataluña.

a) **Usos no anafóricos**. En esta categoría se encuadran aquellas DDs que no tienen ninguna relación de referencia con otras DDs, sintagmas nominales, nombres propios, etc. (posibles antecedentes) que han aparecido con anterioridad en el texto. Se dice que este tipo de descripciones introducen una nueva entidad en el discurso que posteriormente será considerada como un posible antecedente.

b) **Usos anafóricos de nivel sintáctico-semántico-textual**. En esta categoría se encuadran aquellas DDs que están relacionadas con una entidad introducida anteriormente. En función del tipo de relación entre el antecedente y la descripción distinguimos los siguientes tipos:

- **Mismo núcleo** *(anáfora directa)*. Son aquellas DDs cuyo núcleo y el núcleo del antecedente coinciden. Por ejemplo:

  – *La casa de la playa*$_i$ es muy grande. *La casa*$_i$ era de un médico muy bueno.

  En este ejemplo la descripción definida la casa hace referencia a la casa de la playa. Los dos sintagmas tienen el mismo núcleo, casa.

- **Relación semántica**. Son aquellas DDs cuyo núcleo es un sinónimo, hiperónimo-hipónimo o un merónimo-holónimo del núcleo del antecedente. Por ejemplo:

  – *Los soldados*$_i$ irrumpieron en el poblado. *Estos hombres*$_i$ rudos y despiadados dispararon contra todo.

  La DD estos hombres rudos hace referencia a los soldados. Los soldados es un hipónimo de hombres.

- **Papel temático**. Son aquellas DDs que hacen referencia a un antecedente que aparece en una oración cuyo verbo indi-

ca la acción desempeñada por la descripción. Por ejemplo:

– *Juan*$_i$ vende la casa a Fernando. *El vendedor*$_j$ pagará las costas.

La DD el vendedor hace referencia a Juan que aparece en una oración junto al verbo vender.

c) **Usos anafóricos de nivel pragmático**. En esta categoría se encuadran las DDs que tienen una relación indirecta de referencia con su antecedente. Para resolver este tipo de referencia es necesario tener conocimiento del mundo. Distinguimos los siguientes tipos:

- **Nombres propios**. En esta categoría se encuadran aquellas DDs cuyo antecedente es un nombre propio y además indica alguna cualidad de dicho antecedente. Por ejemplo:

  – Bill Clinton y El presidente de los EEUU.

- **Tópicos del discurso**. En esta categoría se encuadran aquellas DDs que hacen referencia al tópico del discurso del que trata el texto. Por ejemplo, en artículos deportivos cuando indicamos los equipos hacemos referencia a los equipos del deporte que se esté hablando.

- **Inferibles**. En esta categoría se encuadran aquellas DDs que tienen una relación de referencia compleja (relaciones causa-efecto, causa-consecuencia, etc.). Por ejemplo:

  – *Los ataques sandinistas*$_i$ eran frecuentes. *Las víctimas*$_i$ aparecían por cualquier lado.

  La DD las víctimas es una consecuencia de los ataque sandinistas.

En el nivel pragmático aparecen varios tipos de DDs (nombres propios, tópicos del discurso e inferibles) que producen parte de los usos asociativos descritos en Poesio y Vieira (1998). Un sistema que intente resolver las referencias producidas por estos tipos de descripciones definidas necesita recursos que proporcionen e infieran conocimiento del mundo (información pragmática). Estos recursos son costosos de construir, muy voluminosos y tienen que estar en continua actualización. Los sistemas que dependan de este tipo de recursos tendrán el inconveniente de la completitud y del manejo de grandes bases de datos que ralentizarán los sistemas de resolución de la anáfora.

En el nivel sintáctico-semántico-textual se encuadran las DDs que Poesio y Vieira denominan usos anafóricos con el mismo núcleo y parte de los usos asociativos (sinónimos, hiperónimos, merónimos y eventos). Para la resolución de las referencias producidas por las DDs con usos anafóricos de nivel sintáctico-semántico-textual no se necesita la utilización de recursos con conocimiento del mundo, ya que todas las referencias pueden resolverse con la información sintáctica producida por analizadores sintácticos y por recursos semánticos como, por ejemplo, el WordNet español que proporciona la información de los sinónimos, hipónimos e hiperónimos.

La figura 4.4 muestra la distribución de las descripciones definidas que se han encontrado en el fragmento del corpus LEXESP utilizado. En ese fragmento se observó que alrededor del 52% de las descripciones definidas no tenían propiedades referenciales y, por tanto, introducían una nueva entidad en el discurso; alrededor del 30% eran DDs que necesitaban de información textual-sintáctica y semántica para establecer la relación con su antecedente, y el restante 28% necesitaban de información pragmática para establecer la referencia con su antecedente.

En resumen, este trabajo se centra en la resolución de

- Anáfora directa (DDs con mismo núcleo que su antecedente)
- Anáfora indirecta (Relacionadas semánticamente y de papel temático)

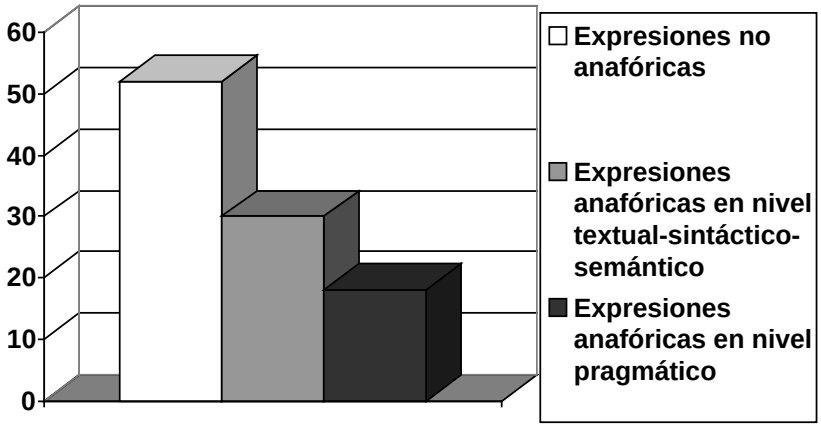

Figura 4.4. *Distribución de las descripciones definidas en el corpus LEXESP*

pertenecientes a la categoría de nivel textual-sintáctico-semántico. La distribución encontrada en esta categoría es la que recoge la figura 4.5. Alrededor del 80% de las DDs encontradas son anáforas directas y el 20% restante son DDs que producen anáfora indirecta.

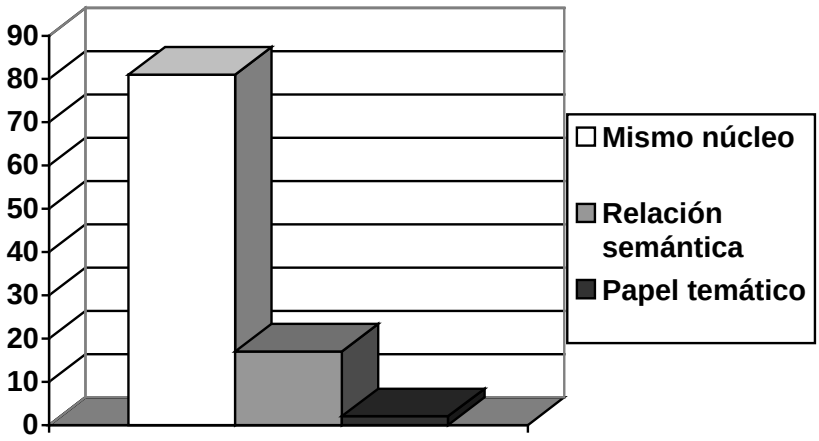

Figura 4.5. *Distribución de las descripciones definidas en el nivel sintáctico-semántico*

La tabla 4.5 presenta la correspondencia entre nuestra clasificación y la clasificación propuesta en la MUC-7. La clasificación

de la MUC-7 establece diferentes tipos en función del tipo de referencia que hace la DD al antecedente. Así, se distingue entre:

- *Identidad.* La expresión anafórica hace referencia a la totalidad de su antecedente (perro, can).

- *Parte de.* La expresión anafórica hace referencia a una parte del antecedente (casa, tejado).

- *Conjunto - subconjunto.* La expresión anafórica hace referencia a un subconjunto del antecedente (muebles, sillas).

- *Conjunto- miembro.* La expresión anafórica hace referencia a un miembro del antecedente (consejo de administración, presidente).

En los tres últimos casos, la descripción definida hace referencia a su antecedente y a la vez introduce una nueva entidad. En el ejemplo anterior, casa, tejado, el sintagma en el que aparece tejado hace referencia al sintagma en el que aparece casa y a su vez introduce la entidad tejado en el discurso.

La tabla 4.5 presenta una confrontación entre la clasificación utilizada en las MUC y la desarrollada en este trabajo. Se observa que determinados tipos de nuestra clasificación pueden pertenecer a varios tipos de la clasificación de las MUC. Por ejemplo, las de mismo núcleo pueden ser de tipo identidad y conjunto-subconjunto; o las relacionas semánticamente pueden ser de todos los tipos: identidad, parte de, conjunto-subconjunto y conjunto-miembro.

Como se observa en la tabla 4.5existe una compatibilidad entre nuestra clasificación y la utilizada en las MUC, pero debido a que uno de los objetivos de este trabajo se basa en el estudio de la influencia de las diferentes fuentes de información en la resolución de las referencias lingüísticas producidas por las DDs, se ha propuesto una clasificación de las DDs en función del tipo de información necesaria para establecer la relación DD-antecedente.

MUC-7	Identidad	Parte de	Conjunto-miembro	Conjunto-subconjunto
Muñoz *et al.* (2000a)	Mismo núcleo R. Semántica Papel temático Nombre propio	R. Semántica Tópico Inferible	R. Semántica Nombre propio	Mismo núcleo R. Semántica

Tabla 4.5. Comparación de las clasificaciones

4.7 Conclusiones

En este capítulo se han presentado los diferentes tipos de fuentes de información y sus características, las cuales intervienen en la resolución de la referencia lingüística. Además, se ha presentado la influencia de cada uno de estos tipos de información en la resolución de las descripciones definidas. El estudio se ha centrado en las informaciones de tipo: léxico-morfológica, sintáctica, semántica y pragmática. Así, como consecuencia del estudio de las diferentes fuentes de información utilizadas en la resolución de las referencias, se ha desarrollado una clasificación para las descripciones definidas en español en función del tipo de información necesaria para establecer la relación entre la DD y el antecedente a diferencia de otras clasificaciones, como la de la MUC-7, que se basa en el tipo de referencia que hace a su antecedente.

5. Tratamiento y resolución de las Descripciones Definidas

En este capítulo se presenta la propuesta de tratamiento y resolución de las descripciones definidas en español. Tomando como base la clasificación realizada sobre las DDs y el análisis de las fuentes de información necesarias para una adecuada resolución se presentan dos enfoques. En el primer enfoque se desarrolla un algoritmo de resolución de las DDs basado en un conjunto de heurísticas que se aplican como un sistema de filtrado. El segundo enfoque surge como solución a los problemas presentados por el primer enfoque. Este segundo enfoque se basa en la construcción automática de una red semántica y en la utilización de un conjunto de heurísticas a través de un sistema de pesos para seleccionar el antecedente correcto. Además, se muestran las tareas previas a la resolución que aportan la información necesaria para la adecuada resolución de las correferencias.

5.1 Introducción

Según se ha planteado en capítulos anteriores, los principales trabajos sobre DDs coinciden en que las dos principales tareas en el tratamiento y resolución de las DDs son:

- La clasificación de las DDs en anafórica o no anafórica y,
- La resolución de las referencias lingüísticas producidas por las DDs anafóricas.

Tomando como hipótesis de partida ambas tareas, se plantean dos enfoques para el tratamiento y resolución de las DDs.

El primer enfoque, enfoque inicial, presenta un algoritmo que realiza al mismo tiempo la resolución y la clasificación de las DDs. El algoritmo está basado en la utilización de información léxico-morfológica, sintáctica y semántica a través de una serie de heurísticas ordenadas extraídas del corpus de entrenamiento. La aplicación de estas heurísticas sobre los posibles antecedentes proporciona el antecedente adecuado a la descripción definida. El segundo enfoque resuelve algunos de los problemas detectados en la aplicación del primer enfoque sobre los corpus de evaluación. Éste consiste en dos pasos, el primero de ellos es la clasificación de las DDs en anafóricas o no anafórica mediante la generación automática de una red semántica que utiliza la ontología del recurso léxico WordNet español; cada una de las DDs se añade a la red en el nodo correspondiente a su clase semántica de la ontología del WordNet español. Y en segundo lugar, se aplica un sistema de heurísticas basadas en pesos para encontrar el antecedente correcto a la DD.

A continuación se presentan algunas características propias de las descripciones definidas y finalmente se presenta el sistema de resolución completo con una descripción detallada de cada una de las fases.

5.2 Características de las descripciones definidas

Las descripciones definidas pueden tener o no propiedades referenciales, por esta razón no siempre es necesario resolverlas, ya que no siempre hacen referencia a un antecedente previo. Ésta es una de las mayores desventajas que tienen los algoritmos de resolución de las DDs frente a otros algoritmos de resolución de las referencias producidas por otras categorías gramaticales, como por ejemplo los pronombres.

De los diferentes tipos de DDs en español que recoge la clasificación presentada en el capítulo 4, se tratarán las DDs no anafóricas y las de nivel textual-sintáctico-semántico. Las DDs de

nivel pragmático se consideran como si fueran no anafóricas, al no disponer de recursos que proporcionen este tipo de información.

Diferentes estudios realizados por diversos autores (Bean & Riloff, 1999; Poesio & Vieira, 1998; Muñoz *et al.*, 2000a) muestran que más de la mitad de las DDs que aparecen en un texto no tienen propiedades referenciales. Aquellos algoritmos que no realizan una identificación previa del tipo de descripción definida (anafórica o no anafórica) emplean tiempo computacional para intentar resolver algo que no tiene solución. Por tanto, la identificación previa de este tipo de DDs hace que el algoritmo sea más eficaz. Tal y como dice Donnellan (1996), (1998), el principal problema que se encuentra a la hora de la identificación previa de las descripciones definidas no anafóricas es que no se puede decir categóricamente que una descripción definida en una oración es una expresión anafórica o no. Según Donnellan, una DD tiene propiedades referenciales en función de las intenciones del hablante en ese caso particular. Por ejemplo, si en una oración se encuentra la descripción definida la casa por si sola no se sabe si será anafórica o no. Para identificar su tipo correctamente hay que ver la aparición anterior de otros sintagmas nominales que puedan estar relacionados con esta DD.

Otros autores, como Bean y Riloff (1999), y Vieira y Poesio (1998), desarrollan métodos para la realización de una identificación previa. El método desarrollado por Bean y Riloff (1999) consiste en un mecanismo basado en aprendizaje. El aprendizaje lo realiza mediante el procesamiento de corpus anotados con información de las referencias lingüísticas. El método desarrollado por Vieira y Poesio (1998), integrado en el propio mecanismo de resolución de las referencias de las DDs, se basa en una serie de reglas relacionadas con la construcción sintáctica de la DD y en la aparición de algunos modificadores restrictivos en la propia DD. Aunque ambos métodos presentan buenos resultados, el principal problema consiste en que la identificación es la primera tarea a realizar por el sistema de resolución y, por tanto, los errores cometidos pueden ser trasladados a tareas o etapas posteriores. No identificar una DD no anafórica no es un problema excesivamente grave puesto que el mecanismo o algoritmo de resolución no

proporcionará ningún antecedente como solución y la clasificará como no anafórica. El problema ocurre cuando una DD anafórica se identifica como no anafórica entonces no se le aplica el mecanismo de resolución.

Por otra parte, un aspecto a tener en cuenta en la resolución de las referencias producidas por las diferentes categorías gramaticales con propiedades referenciales es la clase o tipo de posibles antecedentes que pueden tener. En un texto podemos tener como posibles antecedentes diferentes tipos de expresiones, como son nombres propios, sintagmas nominales, oraciones completas, o incluso, textos o trozos de textos. A continuación se muestran algunos ejemplos de estos casos:

- Juan Antonio Samaranch presidió la apertura de los juegos olímpicos. Este hombre ha conseguido durante su mandato grandes logros.

  En este caso, el antecedente de la descripción definida este hombre es el nombre propio Juan Antonio Samaranch. En nuestro caso, los antecedentes de tipo nombre propio son tratados como los de sintagma nominal ya que el analizador sintáctico reconoce los nombres propios como un sintagma nominal.

- La victoria de Carlos Sainz en el rally de Córcega le acerca al liderato. Este triunfo le dará la moral suficiente para olvidar su último traspié.

  La descripción definida este triunfo tiene como antecedente el sintagma nominal La victoria de Carlos Sainz. En este ejemplo concreto, el antecedente es otra descripción definida, pero podría tratarse de cualquier tipo de sintagma nominal (definido o indefinido).

- Al que madruga Dios le ayuda. Este refrán suele ser cierto.

La descripción definida Este refrán hace referencia a toda la oración anterior.

En nuestra propuesta, se consideran como posibles antecedentes los sintagmas nominales y los nombres propios (que el analizador sintáctico los clasifica como sintagma nominal). Por lo tanto, las posibles DDs que no hagan referencia estos dos tipos se consideran como DDs no anafóricas. También, se consideran como no anafóricas las descripciones definidas que necesitan información pragmática o conocimiento del mundo para establecer la posible relación con su antecedente. Casos como, Juan Antonio Samaranch y el presidente del Comité Olímpico Internacional no son resueltos, y se toma la DD como no anafórica.

Por otro lado, en cuanto a los tipos de DDs anafóricas tratadas en esta memoria, nos centramos en las DDs que producen una anáfora directa (mismo núcleo) y en las que producen una anáfora indirecta (relacionadas semánticamente y de papel temático).

Una vez presentadas las principales características a tener en cuenta en la resolución de las DDs, a continuación presentamos el esquema general de resolución de las tareas previas necesarias para posteriormente centrarnos en los dos enfoques adoptados en la resolución de las DDs.

5.3 Resolución de las Descripciones Definidas

La figura 5.1 muestra el procedimiento general del sistema de resolución de las DDs utilizado por ambos enfoques.

En este esquema general se distinguen las siguientes etapas o tareas previas a la aplicación de la resolución de las DDs:

1. Segmentación del texto en oraciones. Una de las primeras tareas a realizar por cualquier sistema de Procesamiento de Lenguaje Natural es la división del texto en unidades más pequeñas, por ejemplo en oraciones. Procesos posteriores como el análisis léxico-morfológico y análisis sintáctico utilizan co-

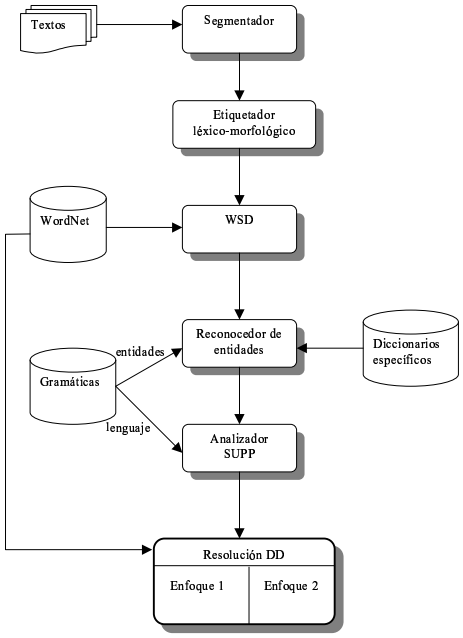

Figura 5.1. *Esquema general del sistema completo*

mo marco de actuación la oración.

2. Etiquetado léxico-morfológico. En esta etapa se añade información léxica-morfológica a cada una de las palabras en el texto, que se utiliza en los procesos posteriores como el análisis sintáctico, el reconocimiento de entidades y la resolución de las referencias producidas por cualquier tipo de expresión anafórica.

3. Desambiguación del sentido de las palabras (WSD). La incorporación de información semántica en cualquier etapa dentro de un sistema de PLN conlleva una tarea adicional como es la identificación del sentido correcto de las palabras polisémicas ya que en función del sentido que tengan éstas la información semántica asociada será diferente.

4. Análisis sintáctico. El análisis sintáctico se realiza oración a oración utilizando información tanto léxico-morfológica como de una gramática específica del lenguaje. Dentro de la etapa de análisis sintáctico se diferencian dos tareas:

   a) Reconocimiento de entidades. En esta etapa se reconocen todas las entidades que aparecen en la oración. Se entiende por entidades los nombres propios de personas, organizaciones, localizaciones, fechas y cantidades monetarias. Para ello, se hace uso de una gramática de entidades y de una serie de diccionarios especiales: de nombres, de apellidos, de empresas y de ciudades. El reconocimiento de este tipo de entidades es fundamental sobre todo para la resolución de los *alias*[1].

   b) Análisis sintáctico. En esta etapa se realiza el análisis sintáctico propiamente dicho. Para ello se apoya en la información léxico-morfológica del paso anterior, en la información proporcionada por el reconocedor de entidades y en la información que le proporciona la gramática del lenguaje sobre la que trabaja el analizador.

A continuación se explican de forma más detallada cada una de las tareas previas de las que consta el sistema de resolución.

5.3.1 Segmentación del texto

Hay que tener en cuenta que el punto de partida de cualquier sistema de Procesamiento de Lenguaje Natural es un texto electrónico, sin ningún tipo de información lingüística o con algunas etiquetas estructurales, en el caso de ser un documento HTML. Cualquier texto está formado por una colección de oraciones agrupadas en párrafos. Para la realización de esta tarea se utiliza el algoritmo desarrollado por Muñoz (1998) y Muñoz y Palomar (1999). Este

[1] Se define un alias como una subcadena de un nombre propio completo o bien una cadena que hace referencia mediante otra descripción a un nombre propio. Como por ejemplo, el Sr. Gómez es un alias de Juan Gómez.

algoritmo de segmentación del texto se basa en la utilización de un árbol de decisión que usa una ventana de palabras alrededor del punto, concretamente las dos palabras anteriores y la posterior al punto.

Según los estudios realizados sobre el dominio de trabajo, el símbolo del punto puede desempeñar las siguientes funciones:

1. *Abreviatura.* (Sr., D., S.A., Cía., etc.)

2. *Acrónimo.* (I.B.M., N.I.F., etc.)

3. *Separador de números.* Puede ser, tanto separador de la parte entera de la decimal, como separador de los millares de las centenas (3.25, 7.5, 1.000, etc.).

4. Finalizador de oraciones.

El algoritmo de segmentación estudia los diferentes papeles que puede desarrollar el punto, así como algunos caracteres especiales como la interrogación (?) o la admiración (!), que se consideran siempre como final de oración. Como ya se ha indicado, para decidir el papel que desempeña el punto (final de oración o no) se utilizan las dos palabras anteriores (*p-2* y *p-1*) y la posterior al símbolo del punto (*p+1*), tal y como se muestra en la figura 5.2

p-2 p-1 . p+1

El texto se analiza palabra a palabra, si alguna de esas palabras es un símbolo del tipo comillas o paréntesis se busca el símbolo de finalización (otras comillas o símbolo de cerrar paréntesis), sin tener en cuenta los posibles puntos que aparezcan en medio. Nuestro método define una serie de reglas de decisión para discernir si es final de oración o no. Además, utiliza un diccionario de propósito general y una lista de abreviaturas (sr., sra., cia, avda., etc).

A continuación se muestra un ejemplo del funcionamiento de este sistema de segmentación en oraciones.

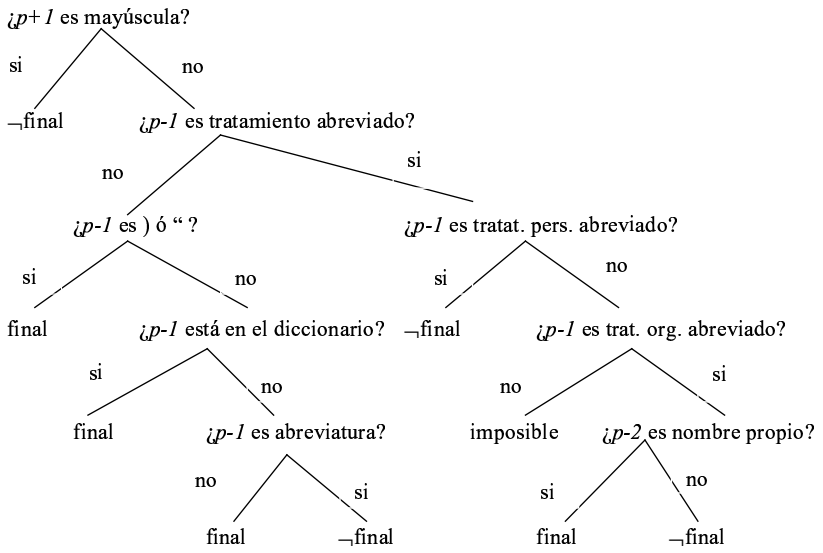

Figura 5.2. *Árbol de decisión usado para segmentar las oraciones de un texto*

Sainz vence en Chipre y acaba con su mala racha. El español se reencuentra con el triunfo tras dos años y recupera opciones en el Mundial. Dos años y 45 días llevaban Carlos Sainz y Luis Moya (Ford) conviviendo con el infortunio, rozando el éxito y sin poder celebrar una victoria en el campeonato del mundo. Pero ayer, por fin, quebraron la mala racha con un triunfo brillante en Chipre, en uno de los rallies más duros del campeonato, después de dominarlo con autoridad desde el primer kilómetro. Con esta victoria, la número 23ª de su historial, el piloto madrileño recupera las posibilidades de pelear por el título. Al campeonato le quedan cuatro pruebas y Sainz es cuarto, a sólo siete puntos del finés Marcus Gronholm (Peugeot), que abandonó en la primera etapa. Debió ser una sensación extraña para el hombre que ahora comparte con Juha Kankkunen el récord de triunfos en la historia del Mundial (23 en 132 participaciones). Aunque nadie ha ganado más carreras que él, se pierde la costumbre de mirar a los rivales desde arriba. Sainz no lo había experimentado desde finales de julio de 1998, cuando ganó en Nueva Zelanda, y por eso ayer era un hombre feliz. "Ha sido una de mis mejores victorias", reconoció, "porque hacía mucho tiempo que la esperaba".

La salida del sistema es:

1. Sainz vence en Chipre y acaba con su mala racha.
2. El español se reencuentra con el triunfo tras dos años y recupera opciones en el Mundial.
3. Dos años y 45 días llevaban Carlos Sainz y Luis Moya (Ford) conviviendo con el infortunio, rozando el éxito y sin poder celebrar una victoria en el campeonato del mundo.
4. Pero ayer, por fin, quebraron la mala racha con un triunfo brillante en Chipre, en uno de los rallies más duros del campeonato, después de dominarlo con autoridad desde el primer kilómetro.
5. Con esta victoria, la número 23 de su historial, el piloto madrileño recupera las posibilidades de pelear por el título.
6. Al campeonato le quedan cuatro pruebas y Sainz es cuarto, a sólo siete puntos del finés Marcus Gronholm (Peugeot), que abandonó en la primera etapa.
7. Debió ser una sensación extraña para el hombre que ahora comparte con Juha Kankkunen el récord de triunfos en la historia del Mundial (23 en 132 participaciones).
8. Aunque nadie ha ganado más carreras que él, se pierde la costumbre de mirar a los rivales desde arriba.
9. Sainz no lo había experimentado desde finales de julio de 1998, cuando ganó en Nueva Zelanda, y por eso ayer era un hombre feliz.
10. "Ha sido una de mis mejores victorias", reconoció, "porque hacía mucho tiempo que la esperaba".

Cada una de las oraciones obtenidas son las que serán procesadas de forma independiente por el resto de módulos del sistema. Si el texto de entrada es una página HTML, entonces las etiquetas del lenguaje de marcado (HTML) no se tienen en cuenta.

Este segmentador fue evaluado en el sistema EXIT, presentado anteriormente, alcanzando una precisión del 98.8%.

5.3.2 Etiquetado léxico-morfológico

A cada una de las oraciones del texto se les aplica un etiquetador léxico-morfológico para proporcionar la etiqueta gramatical correspondiente, la información de concordancia y el lema de la palabra. En este trabajo se ha utilizado el etiquetador léxico-morfológico de Pla (2000) y Pla et *al.* (2000) que utiliza el conjunto de etiquetas PAROLE[2]. En Pla (2000) se puede ver una descripción completa de cada uno de los tipos de etiquetas pertenecientes a cada categoría gramatical.

La salida del etiquetador para la primera oración del ejemplo anterior es la siguiente:

```
Sainz sainz NP00000
vence vencer VMIP3S0
en en SPS00
Chipre chipre NP00000
y y CC00
acaba acabar VMIP3S0
con con SPS00
su su DP3CS00
mala malo AQ0FS00
racha racha NCFS000
. . Fp
```

Como se puede observar, para cada palabra de la oración, el etiquetador proporciona la propia palabra, el lema y su etiqueta PAROLE correspondiente. Por ejemplo, la etiqueta del nombre

[2] PAROLE es un proyecto europeo para proporcionar un conjunto de etiquetas gramaticales para 14 lenguas europeas con un formato uniforme y reusable. En este proyecto participaron: Consorzio Pisa Ricerche (IT), Institiúid Teangeolaíochta Éireann (IE), ERLI (FR), Göteborgs Universitet Department of Swedish (SE), Institut d'Estudis Catalans (ES), University of Helsinki (FI), University of Birmingham (UK), University of Sheffield (UK), Instituut voor Nederlandse Lexicologie (NL), Centro de Linguística da Universidade de Lisboa (PT), Center for Sprogteknologi (DK), Institute for Language and Speech Processing (GR), Universidad de Barcelona (ES), Instituto de Engenharia de Sistemas e Computadores (PT), Université de Liège (BE), Institut National de la Langue Française (FR), Institut für Deutsche Sprache (DE) y Det Danske Sprog-og Litteraturselskab (DK).

racha que es NCFS000 indica que es un nombre común femenino singular, o para el verbo acaba indica que su lema es acabar y que la etiqueta VMIP3S0 es la de verbo modal impersonal tercera persona del singular. Mediante la etiqueta propuesta por el etiquetador se proporciona toda la información léxico-morfológica.

5.3.3 Desambiguación del sentido de las palabras (WSD)

La desambiguación del sentido de las palabras es una de las tareas necesarias en cualquier sistema de PLN que incorpore información semántica. La incorporación de la información semántica a cualquier tarea y en particular a la resolución de las DDs tiene el problema de la determinación del sentido correcto de las palabras polisémicas. Dos métodos de desambiguación del sentido de las palabras sirven de referencia en nuestro trabajo. Ambos trabajos están siendo desarrollados en el grupo de investigación de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante. El primero de ellos, es el desarrollado por Montoyo y Palomar (2000) que obtiene unos resultados de un 65.7% de cobertura y una precisión del 66.6% en la desambiguación de los nombres. Este método denominado de *Marcas de Especificidad* se basa en el uso de la taxonomía de nombres del WordNet español y un conjunto de heurísticas. El contexto de trabajo de este método es el de todas las palabras que se encuentran en la misma oración, para ello estudia las posibles relaciones entre los sentidos de las palabras que se encuentran en ella. El segundo de ellos, desarrollado por Suárez (2001), se basa en la desambiguación léxica utilizando modelos de máxima entropía. Este método utiliza un aprendizaje supervisado basado en la distribución de probabilidad de ciertas características lingüísticas observada en un corpus de entrenamiento. El sistema aplica modelos de Máxima Entropía para construir clasificadores para cada palabra a desambiguar léxicamente. Los modelos de máxima entropía buscan una distribución de probabilidad óptima que permita predecir, después de una fase de entrenamiento, la clase a la que pertenece cierto contexto, en nuestro caso lingüístico. El entrenamiento se realiza mediante el cálculo de ciertas funciones binarias

que representan la presencia o no de una característica concreta dentro cada contexto que acompaña a una palabra previamente etiquetada con su significado correcto. Se obtienen, mediante un procedimiento de optimización, una serie de coeficientes que son los que se utilizan finalmente como base para la clasificación de nuevos contextos, y representan la importancia de cada característica medida en la distribución de probabilidad. El método propuesto utiliza WordNet como recurso donde obtener las definiciones de los significados de cada palabra.

5.3.4 Análisis sintáctico

El analizador sintáctico utilizado en este trabajo, denominado SUPP (Ferrández *et al.*, 1998b; Martínez-Barco *et al.*, 1998; Palomar *et al.*, 1999), ha sido desarrollado en el seno del grupo de Procesamiento de Lenguaje y sistemas de Información (GPLSI) de la Universidad de Alicante. Este analizador es capaz de utilizar tanto técnicas de análisis parcial como completa. Para este trabajo se han utilizado exclusivamente técnicas de análisis parcial. Dentro de la etapa del análisis sintáctico se diferencian dos tareas diferentes:

1. Reconocimiento de entidades
2. Análisis sintáctico

Estas dos tareas se basan en la utilización de la información léxico-morfológica, añadida en la etapa anterior. Además, para el reconocimiento de entidades se basa en una gramática de entidades y, para el análisis sintáctico, se basa en otra gramática específica del lenguaje para el reconocimiento de todos los constituyentes sintácticos de la oración. Estas dos gramáticas se definen mediante la terminología usada por las gramáticas de unificación de huecos (SUG, *Slot Unification Grammar*) (Ferrández, 1998).

Reconocimiento de entidades. La tarea a realizar por el reconocedor de entidades consiste en la identificación en el texto

de los nombres de personas, organizaciones o empresas, lugares, fechas y cantidades. Para llevar a cabo esta tarea se desarrolló una gramática de entidades (Muñoz, 1998; Muñoz *et al.*, 1998; Muñoz & Palomar, 1999). Las entidades en el texto se identifican de forma similar a como se hace en el análisis sintáctico parcial, mediante una gramática que contempla los constituyentes de tipo entidad.

Las reglas que forman la gramática de entidades se han definido de forma manual a partir de un estudio del dominio del sistema de extracción de información EXIT (Llopis *et al.*, 1998), escrituras de compraventa. Además, para la realización de esta tarea se utilizan varios diccionarios específicos, como son:

- Diccionario de nombres
- Diccionario de apellidos
- Diccionario de localidades
- Diccionario de actividades

Las características fundamentales que cumplen las palabras que forman parte de las entidades objeto de estudio son:

- Palabras en mayúsculas,

- Palabras que hayan sido catalogadas como nombre propio,

- Palabras que pueden contener lo que denominamos un *disparador* (Sr., Don, Avenida, etc.). Por ejemplo, la palabra Don precede a una entidad persona, o la palabra avenida precede a una dirección.

- Palabras formadas, en parte o en su totalidad, por algún número. Como por ejemplo, 15 de septiembre ó 2.500.000.

Los pasos a seguir para el reconocimiento son:

1. *Identificación de los disparadores.* La identificación de disparadores se realiza recorriendo las oraciones en busca de algunas palabras que introduzcan o finalicen una posible entidad. Por ejemplo, las palabras Don, Doña, S.A., S.L., etc. informan de la posible existencia de una entidad. Si no existe disparador entonces se consideran como entidad los nombres propios.

2. *Reconocimiento de nombres, organizaciones o direcciones.* Una vez identificada la palabra que actúa de disparador se necesita conocer la entidad completa. Es decir, se necesita conocer el principio y el final de la entidad. La figura 5.3 muestra un pequeño subconjunto de las reglas SUG[3] utilizadas para el reconocimiento de entidades, en la que la regla *entidad-persona* es la utilizada como punto de partida.

entidad_persona ++> *trat_per* , *persona.*

trat_per ++> [D. | DON | DOÑA | Dña. | SEÑOR | Sr. | SEÑORA | Sra. | EXCMO. | ILMO. | V.I. | etc].

persona ++> *nombre* , << *nombre* >> , *apellido* , [y] , *apellido.*

persona ++> *nombre* , << *nombre* >> , *apellido* , << *apellido* >>.

nombre ++> NP0000.

apellido ++> *apell2* , << *apell2* >>.

apellido ++> *apell2*, [-] , *apell2.*

apell2 ++> << prep >> , << art >> , *apell.*

apell ++> NP0000

Figura 5.3. *Pequeña muestra de las reglas SUG usadas para el reconocimiento de entidades*

3. *Desambiguación.* En el reconocimiento de entidades nos encontramos con el problema de la identificación del inicio y final de las entidades en casos como,

[3] La gramática SUG utiliza el símbolo ++> símbolo de producción, el símbolo de la coma (,) para representar la unión, los símbolos << >> para representar que el constituyente que engloba es opcional y los corchetes ([]) para indicar la elección de uno de los elementos

Clínica San Carlos y Telefónica S.A..

Para resolver estos problemas se aplican las siguientes heurísticas, que tratan por separado las cadenas que aparecen a la izquierda y a la derecha de la conjunción:

- Siempre que aparezcan disparadores de personas u organizaciones en las dos subcadenas, se dividen en dos entidades.

- Si la cadena de la izquierda contiene un disparador y la otra no, se divide en dos entidades ya que la parte de la izquierda se sabe que es una entidad por si sola.

- Si la cadena de la izquierda no contiene un disparador y la de la derecha sí. En este caso, si en la subcadena de la izquierda se identifica alguna actividad (clínica, restaurante, etc.) entonces forma una entidad propia, en caso contrario forma parte del nombre de la entidad de la derecha. Pueden existir casos en los que sea necesario realizar una posterior desambiguación semántica.

- Si ninguna de las dos subcadenas dispone de disparadores, se aplican las mismas heurísticas pero teniendo en cuenta la actividad en lugar de los disparadores.

- Si hay un acrónimo a la izquierda o derecha del operador ambiguo, siempre se divide en dos entidades.

A continuación se muestra, a través de una estructura de rasgos, la entidad identificada y las reglas necesarias para su identificación.

D. José Gómez García

$$persona \begin{bmatrix} \text{Tratamiento} & [\text{D0 'D.'}] \\ \text{Nombre} & [\text{NP0000 'José'}] \\ \text{Apellido1} & [\text{NP0000 'Gómez'}] \\ \text{Apellido2} & [\text{NP0000 'García'}] \end{bmatrix}$$

Las reglas SUG usadas para la identificación fueron:

entidad-persona ++> trat-per , persona
trat-per ++> [D.]
persona ++> nombre,<<nombre>>,apellido,<<apellido>>
nombre ++> NP0000.
apellido ++> apell2 , << apell2 >>
apell2 ++> NP0000

Analizador sintáctico. Como se ha citado anteriormente, el analizador utilizado es el analizador SUPP. La técnica de análisis empleada realiza un análisis parcial ascendente, es decir, identifica pequeñas unidades sintácticas que forman parte de unidades sintácticas mayores.

La figura 5.4 muestra la salida del analizador SUPP para la frase, este coche ha sido utilizado por corredores profesionales.

Este analizador se fundamenta en la utilización de una gramática propia del lenguaje (español ó inglés) definida a partir de las características de las reglas SUG. Estas reglas SUG se han implementado a través de la utilización de unas estructuras denominadas *estructuras de huecos*. Existe una estructura de huecos característica para cada tipo de constituyente. En la estructura de huecos se almacena la información relativa al constituyente, de tipo léxico-morfológica (género, número, persona, lema). La estructura de huecos de cualquier constituyente tiene como mínimo tres argumentos.

```
** ORACION ANALIZADA PARCIALMENTE:
  ** SINT.NOMINAL:
    ** SINT.NOMINAL SIMPLE:
      ** DETERMINANTE 1:
        ** ADJETIVO SIMPLE:
          este
        ** SUSTANTIVO:
          coche
  ** NUCLEO VERBAL:
      ** VERBO:
        haber
        ** VERBO:
          ser
          ** VERBO:
            utilizar
  ** SINT.PREPOSICIONAL:
    ** SINT.PREPOSICIONAL SIMPLE:
      ** PREPOSICION:
        ** PREPOSICION SIMPLE:
          por
      ** SINT.NOMINAL:
        ** SINT.NOMINAL SIMPLE:
            ** SUSTANTIVO:
              corredor
          ** ADYACENTE ADJETIVO:
              ** ADJETIVO SIMPLE:
                profesional
```

Figura 5.4. *Análisis de una frase*

- El primer argumento almacena la información de concordancia del constituyente. Se almacena información relativa al género, número, persona e información semántica.

- El segundo argumento es un identificador que lo diferencia de otras estructuras en la misma oración. Se usa el número de constituyente en la oración.

- El resto de argumentos serán o bien otras estructuras de huecos correspondientes a otros subconstituyentes que están presentes en este constituyente, o bien el lema del propio constituyente si se trata de un constituyente final.

La figura 5.5 muestra un breve ejemplo de la oración:

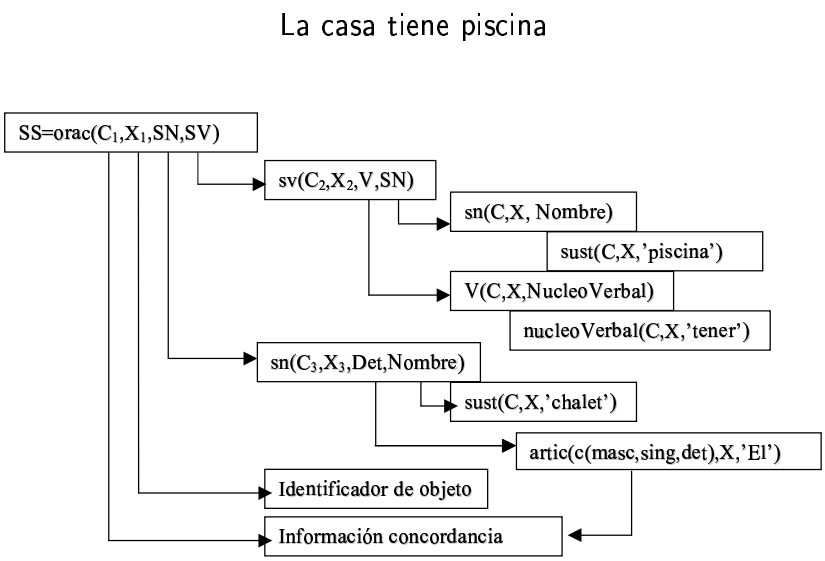

Figura 5.5. *Estructura de huecos de una oración completa*

La representación de la estructura de huecos anterior, utilizando una estructura de rasgos, se muestra en la figura 5.6.

Todas las etapas o procesos que se han aplicado al texto tienen como objetivo incorporar información necesaria para el proceso de resolución de las referencias lingüísticas. Por otro lado, en cuanto al uso de información semántica, ésta es proporcionada por el WordNet español y forma parte de las propuesta de resolución de las DDs presentadas a continuación.

5.4 Tratamiento y resolución de las descripciones definidas

Como han demostrado diversos estudios (Bean & Riloff, 1999), (Poesio & Vieira, 1998) y (Muñoz *et al.*, 2000a), no es necesario la resolución de las referencias de todas las DDs, ya que más de la mitad de las DDs son no anafóricas. Diversos trabajos desarrollados por varios autores, (Bean & Riloff, 1999) y (Vieira,

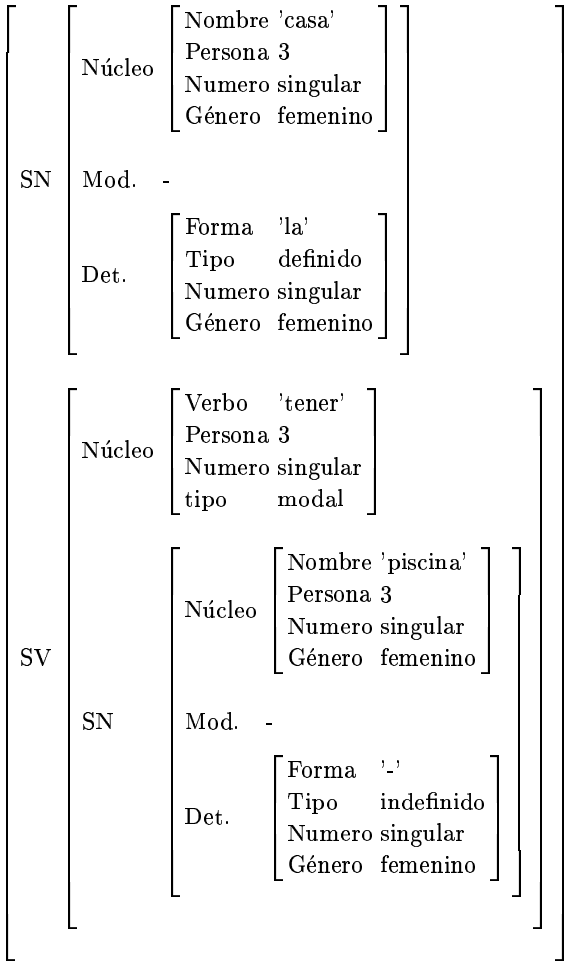

Figura 5.6. *Estructura de rasgos de la oración*

1998), proponen métodos para la identificación previa del tipo de DD que se trata (anafórico o no anafórico). El principal problema que presentan estos métodos es que los errores cometidos en esta fase inicial se propagan durante la tarea de resolución de las referencias.

El modo de clasificar la DD en anafórica o no anafórica, previo o no a la clasificación, lo consideramos primordial en el tratamiento de las DDs. Así y tomando como base lo anteriormente

comentado, se desarrollo un primer enfoque que busca un antecedente para todas las DDs existentes en el texto y para aquellas que no encuentra ningún antecedente las clasifica como no anafóricas. Una vez desarrollado este enfoque, y tomando como punto de partida sus principales carencias (incorrecta clasificación, coste computacional, etc.) se desarrolló un segundo enfoque que sí que realiza una identificación previa de algunas de las descripciones definidas no anafóricas. Este segundo enfoque toma como punto de partida la idea de filósofos de primeros de siglo, como Russell y Frege, que definen el problema de la referencia como un problema semántico. Esta idea lleva a pensar que una DD tendrá como antecedente, en el caso que lo tenga, un sintagma nominal perteneciente a una clase semántica equivalente. El desarrollo de este segundo enfoque se basa en la utilización de la ontología del WordNet español[4] para la generación de una red semántica. Esta identificación previa de las DDs, aunque no identifica un gran número de descripciones definidas de un texto, lo hace de manera correcta. Por lo tanto no propaga errores de identificación a la fase de resolución.

Ambos enfoques utilizan información léxico-morfológica, sintáctica y semántica para el tratamiento de las DDs, tienen en cuenta como posibles antecedentes sintagmas nominales que han aparecido en las oraciones previas; y el objetivo de ambos enfoques consiste en el tratamiento adecuado de cada tipo de DD. Es decir, identificar las DDs no anafóricas y proporcionar las cadenas de correferencias de las DDs anafóricas en el texto. No se trata de establecer relaciones dos a dos entre la descripciones definidas y su antecedente, sino que el objetivo es el establecer cadenas de correferencias, como se muestra en la figura 5.7.

La experimentación realizada sobre el primer enfoque mostró algunas deficiencias que se trataron de subsanar en el segundo enfoque. Los dos enfoques desarrollados fueron:

[4] La ontología utilizada por el EuroWordNet es la misma para todos los WordNets que lo forman.

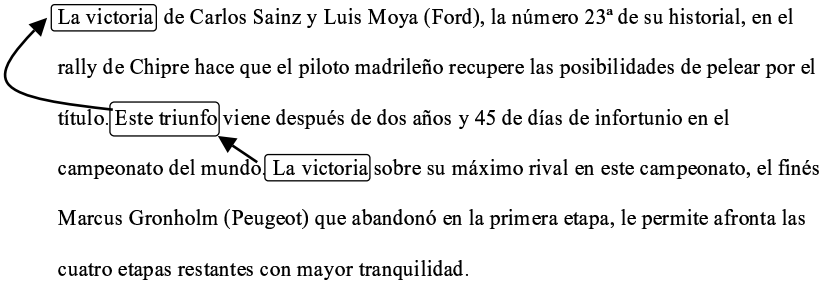

Figura 5.7. *Ejemplo de una cadena de correferencia en un texto*

- *Enfoque 1: Algoritmo de resolución de las descripciones definidas basado en heurísticas (ARH).* En este primer enfoque se desarrolló un algoritmo para el tratamiento de las DDs basado en un conjunto de heurísticas extraídas del corpus de entrenamiento y aplicadas como un sistema de filtrado.

- *Enfoque 2: Algoritmo de resolución de las DDs basado en la generación automática de una red semántica (AGSN).* En este enfoque se desarrolló un algoritmo para el tratamiento de las DDs basado en la construcción automática de una red semántica y la aplicación de un conjunto de heurísticas basadas en pesos.

A continuación se muestra de forma más detallada cada enfoque.

5.5 Enfoque 1: Algoritmo de resolución de las descripciones definidas basado en heurísticas (ARH)

A pesar de la naturaleza dual de las descripciones definidas, este primer enfoque no realiza una identificación previa del tipo de DD (anafórico o no anafórica), sino que aplica el algoritmo de resolución a todas las DDs que aparecen en el texto. Si el algoritmo de resolución no encuentra ningún antecedente a la DD, entonces la clasifica como *no anafórica*. Por el contrario, si es capaz de

proporcionar un antecedente entonces se clasifica como *anafórica* y dicho antecedente es propuesto como solución a la DD.

El algoritmo de resolución de las referencias producidas por las DDs está basado en la aplicación de una serie de heurísticas extraídas del estudio del corpus de entrenamiento (LEXESP). Estas heurísticas se aplican a todos los sintagmas nominales pertenecientes a todas las oraciones previas a la aparición de la descripción definida. Es decir, no establece un espacio de búsqueda de la solución limitado a un número de oraciones anteriores, sino que su espacio de búsqueda va desde la oración inicial hasta la actual (la que contiene a la DD a resolver).

```
procedimiento TRATAR_ANAFORA(NP, Lista_Ant, Lista_Corref)
   Sea NP el sintagma nominal a procesar,
   Sea Lista_Ant la lista que contiene todos los NP previos,
   Sea Lista_Corref la lista que contiene las correferencias
        encontradas

        IDENTIFICAR_DD(NP, Lista_Ant, TipoDD)
        si TipoDD="descripción definida" entonces
            RESOLVER_MISMO_NUCLEO(NP, Lista_Ant, Sol)
            si NO_ENCONTRADA(Sol) entonces
                RESOLVER_ANAFORA_INDIRECTA(NP, Lista_Ant, Sol)
            si ENCONTRADA(Sol) entonces
                AÑADIR(NP, Sol, Lista_Corref)
            fin si
        fin si
        AÑADIR(NP, Lista_Ant)

fin procedimiento
```

Figura 5.8. *Esquema del primer enfoque: algoritmo ARH*

La figura 5.8 muestra un esquema general del procedimiento de resolución propuesto en este primer enfoque (Muñoz & Ferrández, 2000). Se observan 3 pasos principales:

1. Identificación de las DDs. Realiza una búsqueda en el texto de las DDs mediante el uso la información sintáctica proporcionada por el analizador.

2. Resolución de las DDs con el mismo núcleo que su antecedente (anáfora directa).

3. Resolución de las DDs con distinto núcleo que su antecedente (anáfora indirecta).

5.5.1 Identificación de las descripciones definidas.

El punto de partida del algoritmo es la identificación de las DDs en el texto. El mecanismo utilizado para la identificación de las DDs consiste en recorrer el texto analizado sintácticamente con anterioridad. Cuando se encuentra un sintagma nominal se realiza una comprobación automática del primer constituyente de dicho sintagma. Si el primer constituyente del sintagma nominal es un artículo definido o un demostrativo entonces se clasifica dicho sintagma nominal como una DD. Si el sintagma nominal no se clasifica como una DD, entonces dicho sintagma se almacena en una lista denominada *lista de candidatos*. Esta lista de candidatos está formada por todos los sintagmas nominales (definidos e indefinidos) previos a la aparición en el texto de la DD. Una vez detectada en el texto la DD se le aplica el algoritmo de resolución. Pero en función del tipo de DD habrá que aplicar, o bien el algoritmo de resolución de las anáforas directas, o bien el de las anáforas indirectas. El problema que se plantea es que por su construcción exclusivamente, no se puede decir si una DD produce una anáfora directa o indirecta. Tal y como se ha visto en el capítulo 4, la distribución de las descripciones definidas que necesitan de información léxico-morfológica, sintáctica y semántica para establecer la relación entre el antecedente y la DD es:

- *mismo núcleo* alrededor de un 81%,
- *relacionadas semánticamente* cerca de un 17% y
- *papel temático*, el 2% restante.

Debido a esta distribución, el proceso de resolución de las DDs se ha dividido en dos fases:

1. Resolución de la anáfora directa.

2. Resolución de la anáfora indirecta.

El proceso es el siguiente, una vez detectada la DD, se clasifica inicialmente como una anáfora directa y el algoritmo de resolución de la anáfora directa intenta proporcionar una solución. Si no es capaz de encontrar una solución entonces se clasifica como anáfora indirecta y el algoritmo correspondiente intenta proporcionar la solución. Si este algoritmo de resolución de la anáfora indirecta tampoco es capaz de proporcionar una solución, se clasifica la DD como no anafórica, tal y como muestra la figura 5.9, que recoge un esquema de la aplicación del algoritmo.

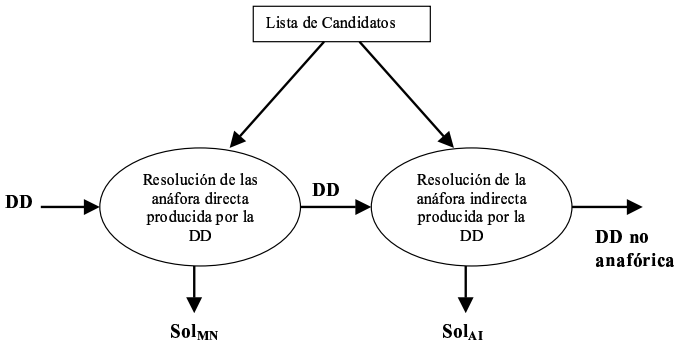

Figura 5.9. *Esquema de aplicación del algoritmo de resolución*

5.5.2 Resolución de anáfora directa producida por las DDs.

El proceso de resolución de las DDs que tienen el mismo núcleo que el antecedente (Muñoz & Palomar, 2000) aplica una serie de heurísticas para reducir la lista de antecedentes y, finalmente, proporcionar uno de ellos como la solución más adecuada. Estas heurísticas están enfocadas al estudio de posibles relaciones entre los modificadores de ambos sintagmas nominales (antecedente y DD).

Se realizó un estudio sobre el corpus de entrenamiento LEXESP para establecer las relaciones existentes entre los diferentes modificadores de ambos sintagmas nominales. De este estudio, tal y como se muestra en la tabla 5.1, se obtuvo que alrededor de un 44% de las veces se trata de una repetición de la propia descripción definida. Un 40% de las veces, la expresión anafórica está incluida en el antecedente o al revés. Es decir, que uno de los dos tiene los mismos modificadores más alguno más. Y un 16% de las veces existen relaciones semánticas entre los modificadores.

Tipo	**Porcentaje**
repetición	44%
incluida	40%
relación de modificadores	16%

Tabla 5.1. Distribución de los antecedentes con el mismo núcleo que la descripción definida

A continuación se muestran algunos ejemplos de los tres tipos posibles de relaciones:

1. Repetición de la DD.

   - *La victoria*$_i$ se consiguió con mucho sacrificio.... *La victoria*$_i$ provocó una pasión descontrolada por parte de sus seguidores.

2. DD incluida en otra DD. Se pueden encontrar dos casos:

   - La descripción definida es más general que el antecedente. Este caso se puede observar en el siguiente ejemplo,

     – *La casa de madera*$_i$ que está en la montaña sirve de refugio a muchos esquiadores. *La casa*$_i$ está permanentemente en condiciones por si alguien queda atrapado.

- El antecedente es más general que la descripción definida. Como por ejemplo en la oración,

  – *El edificio*$_i$ fue construido usando las últimas técnicas en soportar seísmos. *El moderno edificio*$_i$ ha resistido los dos últimos terremotos.

3. Relación semántica entre los modificadores. Entre los diferentes modificadores que pueden acompañar al núcleo de una DD y los de su antecedente se pueden establecer tres casos diferentes de relaciones:

   - Relación positiva. En este caso entre ambos modificadores puede existir una relación de sinonimia, hiperonimia o hiponimia. En el siguiente ejemplo hay una relación de sinonimia entre rápido y veloz

     – Juan se compró *el coche más rápido del mercado*$_i$.... *El veloz coche*$_i$ le costó diez millones de pesetas.

   - Relación negativa. Existe una relación de antonimia entre ambos modificadores. En el caso que se de este tipo de relación entre los modificadores, el candidato es rechazado. Como por ejemplo en el siguiente ejemplo hay una relación de antonimia entre derecha e izquierda

     – Juan se colocó un pendiente en *la oreja derecha*$_i$. Mientras llevaba un sonotone en *la oreja izquierda*$_j$.

   - No existe ninguna relación entre ambos modificadores. Se pueden dar casos como en el siguiente ejemplo donde no hay ninguna relación entre madera y blanca

     – *La casa de madera*$_i$ es más acogedora para el campo....*Esa casa blanca*$_j$ me gustaría más con el techo en rojo.

Este estudio sirve como base para la definición del esquema general de este algoritmo presentado en la figura 5.10 y de las heurísticas para seleccionar el antecedente más adecuado.

```
procedimiento RESOLVER_MISMO_NUCLEO(NP, Lista_Ant, Sol)
   Sea NP la descripción definida a procesar,
   Sea Lista_Ant la lista de candidatos a antecedentes,
   Sea Sol el NP propuesto como solución

   para cada ANT perteneciente a Lista_Ant hacer
        COMPARA_NUCLEOS(NP, ANT, Lista_Cand)
   fin para
   si ELEMENTOS(Lista_Cand) = 0 entonces
        tipo_DD = "indirecta"
   sino
     si ELEMENTOS(Lista_Cand) > 0 entonces
        APLICAR_HEURISTICAS(Lista_Cand,Sol)
        si ENCONTRADA(Sol) entonces
             DEVOLVER Sol
        sino
          tipo_DD = "indirecta"
        fin si
     fin si
   fin si

fin procedimiento
```

Figura 5.10. *Resolución de las DDs con el mismo núcleo*

Este algoritmo de resolución utiliza como datos de entrada dos argumentos:

- La lista de candidatos a antecedentes, que está formada por todos los sintagmas nominales previos que han aparecido en el texto.

- La propia descripción definida que tiene que resolver.

Como datos de salida el algoritmo devuelve la solución (Sol) de la DD en el caso de que la encuentre. Si no encuentra ninguna

solución no devuelve nada y se activará el módulo de resolución de la anáfora indirecta.

En el esquema anterior se observan dos pasos principales:

1. *Restricción de la lista de antecedentes.* Como ya se ha indicado, la lista de candidatos a antecedentes está formada por todos los sintagmas nominales que han aparecido previamente en el texto. Este algoritmo proporciona una solución si existe un antecedente con el mismo núcleo que la DD y sin modificadores incompatibles con su antecedente. Para ello, selecciona de la lista de candidatos a antecedentes todos aquéllos que tienen el mismo núcleo y los almacena en una nueva lista denominada *lista de candidatos a antecedentes con el mismo núcleo.* La selección de estos candidatos se realiza mediante la confrontación o comparación del lema del núcleo de la DD con el lema de cada uno de los posibles antecedentes, almacenando en la lista de candidatos todos aquellos que tengan el mismo lema.

2. *Aplicación de heurísticas.* El método de aplicación de las heurísticas se muestra en la figura 5.11. Se aplica la primera heurística, si esta heurística no la satisface ningún candidato, se continúa con la siguiente regla utilizando la misma lista de candidatos. Si la regla aplicada se satisface por un sólo candidato éste es el elegido como solución de la descripción definida. En el caso que más de un candidato satisfaga la regla heurística, entonces se les aplica a esos candidatos el resto de reglas heurísticas. Es decir, se utiliza un sistema de filtrado de los posibles antecedentes mediante la sucesiva aplicación de heurísticas a los antecedentes que satisfacen las heurísticas previas.

   Las heurísticas usadas por el algoritmo de resolución de las DDs del mismo núcleo son:

   - **Heurística H1: (Repetición)**. Se seleccionan de la lista de candidatos, que está exclusivamente formada por candi-

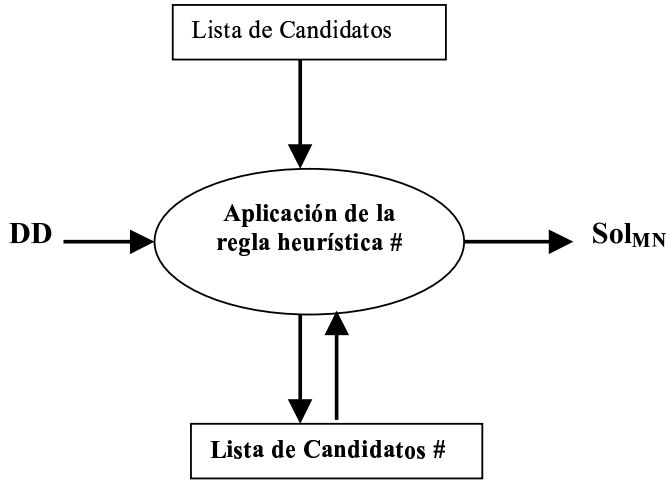

Figura 5.11. *Esquema de aplicación de las reglas heurísticas*

datos con el mismo núcleo, aquellos candidatos que tengan los mismos modificadores. Es decir, se prefiere como candidato la aparición de la misma DD en el texto.

– *La victoria*$_i$ se consiguió con mucho sacrificio.... *La victoria*$_i$ provocó una pasión descontrolada por parte de sus seguidores

- **Heurística H2: (Incluida)**. En ocasiones, el antecedente de una DD tiene los mismos modificadores que la DD y algunos más. En estos casos se dice que la DD está incluida en su antecedente. El siguiente ejemplo muestra este caso:

  – *El coche rojo de Juan*$_i$ es una auténtica maravilla. *Estos coches tan caros*$_j$ da pena dejarlos en la calle. *El coche de Juan*$_i$ lo vieron aparcado en Alicante

  En este ejemplo tenemos tres DDs cuyo núcleo es coche. Si se intenta resolver la última DD, el coche de Juan, la lista de candidatos estaría formada por dos DDs (el coche rojo de Juan y estos coches tan caros). Al no existir ningún sintagma nominal igual que la DD, la heurística H1 no se

satisface por ningún candidato por tanto, la lista de candidatos se mantiene igual y se le aplica la heurística H2. Esta heurística seleccionará el primer candidato de la lista ya que tiene un modificador (de Juan) que también aparece en la DD. Si la última oración hubiera sido el coche lo vieron aparcado en Alicante, al aplicar la heurística H2 habría seleccionado los dos candidatos de la lista, (el coche rojo de Juan y estos coches tan caros), ya que la DD a resolver no tiene modificador alguno y, por tanto, está incluida en ambos candidatos. El algoritmo de resolución continuaría aplicando el resto de heurísticas.

- **Heurística H3: (Relación entre modificadores)**. En ocasiones el núcleo de la DD está acompañado por modificadores que están relacionados semánticamente (relación semántica positiva) con los modificadores de los antecedentes a través de relaciones de sinonimia, hiperonimia y antonimia. Para proporcionar este tipo de información se utiliza el recurso léxico *WordNet español* perteneciente al proyecto *EuroWordNet*. A continuación se muestran algunos ejemplos:

  – *El coche más rápido del mercado*$_i$ alcanza los 300 km/h con un consumo de 18 litros.... *El veloz coche*$_i$ provoca en los amantes de la velocidad sensaciones indescriptibles.

  En este ejemplo el coche más rápido y el veloz coche están relacionados ya que existe una relación entre rápido y veloz. Pero también nos encontramos casos en los que los modificadores tienen un relación de antonimia (relación semántica negativa) y pueden ser rechazados, como por ejemplo:

  – *El coche más rápido del mercado*$_i$ alcanza los 300 km/h con un consumo de 18 litros.... *El coche más lento*$_j$ apenas alcanza los 40 km/h.

En este ejemplo el coche más rápido y el coche más lento no están relacionados ya que existe una relación de antonimia entre lento y rápido. Por otro lado, en otras ocasiones no se puede establecer ningún tipo de relación entre los modificadores, como por ejemplo:

– *El coche más rápido del mercado*$_i$ alcanza los 300 km/h con un consumo de 18 litros.... *El coche rojo*$_j$ es más llamativo

La heurística H3 selecciona aquellos candidatos cuyos modificadores tengan alguna relación semántica positiva.

- **Heurística H4: (Concordancia de género y número)**. En el caso de que las heurísticas anteriores proporcionen más de un candidato, entonces se prefieren aquellos que tengan el mismo género y el mismo número. Si se aplican las reglas al siguiente ejemplo:

  – *El coche rojo de Juan*$_i$ es una auténtica maravilla. *Estos coches tan caros*$_j$ da pena dejarlos en la calle. *El llamativo coche*$_i$ lo vieron aparcado en Alicante

  Al resolver la última DD el llamativo coche, la lista de candidatos estaría formada por las DDs el coche rojo de Juan y estos coches tan caros, al aplicar H4 seleccionaría sólo la primera, el coche rojo de Juan.

- **Heurística H5: (Frecuencia)**. Si una vez aplicada la concordancia de género y número más de un candidato permanece en la lista de candidatos, entonces se elige como candidato aquel que más veces haya sido seleccionado como antecedente de DDs previas.

  – *El coche de Carlos Sainz*$_i$ tiene 300 cv. *Este coche*$_i$ vale más de 100 millones de pesetas. *Los coches de carreras*$_j$ están dotados con los últimos avances tecnológicos. *El coche de*

Sainz$_i$ tiene un sistema de suspensión inteligente mientras que *el coche de los pilotos que no corren este mundial*$_k$ tienen una suspensión de inferior categoría. *Este potente coche*$_i$ le permite seguir soñando con ser campeón del mundo.

Una vez aplicada las heurísticas anteriores, la DD Este potente coche podría tener como posible antecedente las DDs El coche de Carlos Sainz ó el coche de los pilotos que no corren este mundial. Al aplicar la heurística H5, se selecciona el El coche de Carlos Sainz ya que es el antecedente propuesto en más ocasiones.

- **Heurística H6: (Más cercano)**. Si más de un candidato satisface la heurística H5, entonces se elige como candidato correcto a la DD el candidato más cercano en el texto.

En el caso que ningún candidato se seleccione como solución de la DD se le aplicará a ésta el algoritmo de resolución de las anáforas indirectas que se presenta en el siguiente apartado.

A continuación se muestra un ejemplo completo de la aplicación de este algoritmo de resolución de las descripciones definidas del mismo núcleo.

[El coche de [competición]$_1$]$_2$ que [Juan López]$_3$ se compró le costó [diez millones de [pesetas]$_4$]$_5$. [Este coche]$_6$ había sido usado por [corredores profesionales]$_7$. [El coche de [competición]$_8$]$_9$ quedó destrozado en [un accidente]$_{10}$ en [el circuito del [Jarama]$_{11}$]$_{12}$. [El coche de [carrera]$_{13}$]$_{14}$ fue vendido como [chatarra]$_{15}$. [Los coches de [competición]$_{16}$]$_{17}$ para poder competir en [carreras del [campeonato del [mundo]$_{18}$]$_{19}$]$_{20}$ tienen que cumplir [unas normas muy estrictas]$_{21}$. [Estos coches]$_{22}$ son revisados continuamente para que [ningún piloto]$_{23}$ pueda aprovecharse de [ciertas ventajas mecánicas]$_{24}$ que le proporcione [algunos segundos de [ventaja]$_{25}$]$_{26}$ frente a [otros pilotos]$_{27}$. [La mayoría de [fabricantes de [coches de [competición]$_{28}$]$_{29}$]$_{30}$]$_{31}$ lo son también de [coches utilitarios]$_{32}$. [Estos coches]$_{33}$ no pueden ser comprados por [cualquier persona]$_{34}$. [Carlos

Sainz]$_{35}$ es [un piloto de [rallies]$_{36}$]$_{37}$ que ha ganado [el campeonato del [mundo de [rallies]$_{38}$]$_{39}$]$_{40}$. [Marc Gené]$_{41}$ y [Pedro de la Rosa]$_{42}$ son [pilotos de [Formula I]$_{43}$]$_{44}$. [Estos pilotos]$_{45}$ no han ganado [ningún campeonato]$_{46}$.

La tabla 5.2 muestra qué antecedentes satisfacen cada una de las heurísticas para cada una de las DDs que se encuentran en el texto. En cada columna aparece un número que corresponde al número de sintagma nominal en el texto anterior. Si en una columna *Hn* aparece un conjunto de números significa que esos sintagmas nominales satisfacen la heurística n para esa DD. La regla denominada *H0* representa la selección de los antecedentes que tienen el mismo núcleo que la descripción definida.

DD	H0	H1	H2	H3	H4	H5	H6	Sol.
6	2		2					2
9	2,6	2						2
12	-							-
14	2,6,9			2,9	2,9	2,9	9	9
17	2,6,9	2,9				2,9	9	9
18	-							-
19	-							-
22	2,6,9,14,17		2,6,9,14,17		17			17
31	-							-
33	2,6,9,14,17		2,6,9,14,17		17,29	17,29	29	29
39	19			19				19
40	18			18				18
45	23,27,37,44		23,27,37,44		44			44

Tabla 5.2. Heurísticas aplicadas

La DD con el número *6* (Este coche) está incluida en el sintagma nominal con el número *2* (El coche de competición), ya que no tiene modificadores. Al aplicar H2, se elige el antecedente correcto. La descripción definida *9* (El coche de competición) tiene como posibles candidatos dos sintagmas nominales (*2* (El coche de competición) y *6* (Este coche)). Al aplicar H1 se selecciona como antecedente correcto *2* (El coche de competición).

La resolución de la DD *14* (El coche de carrera) se realiza al aplicar H3 y H6 que proporcionan como solución el sintagma *9* (El coche de competición). H3 selecciona los sintagmas *2* (El coche de competición) y *9 (El coche de competición)* de la lista de candidatos porque existe una relación de sinonimia entre carrera y competición, rechazando el sintagma número *6*. H4 y H5 no modifican la lista de candidatos al concordar en género y número, y tener la misma frecuencia. Finalmente, el sintagma más cercano, *9* (El coche de competición), se elige como solución.

Para la DD *22* (Estos coches) la aplicación de H4 es la que proporciona el antecedente correcto *17* (Los coches de competición). Lo mismo ocurre al resolver la DD *45* (Estos pilotos) que, al aplicar H4, selecciona como solución el sintagma nominal *44* (pilotos de Formula I) en lugar del *37* (un piloto de rallies).

Se observa que las DDs *12*(el circuito del Jarama), *19* (el campeonato del mundo), *18* (el mundo) y *31* (La mayoría de fabricantes de coches de competición) introducen nuevas entidades en el discurso.

Finalmente, a las DDs *39* (el mundo de rallies) y *40* (el campeonato del mundo de rallies) sólo se les aplica H3 para comprobar que no existe una relación de antonimia entre los modificadores. Al tener la lista de candidatos sólo un elemento se podría pensar que éste es la solución, pero hay que comprobar que no existe esa posible relación de antonimia entre los modificadores.

5.5.3 Resolución de las anáforas indirectas producidas por las DDs.

En esta categoría se tratan exclusivamente las DDs con una determinada relación semántica con su antecedente (sinonimia, hiperonimia e hiponimia) y las de papel temático (agente y paciente).

Para resolver este tipo de relaciones entre la DD y su antecedente hace falta conocer toda la información semántica relacionada con los núcleos de ambos sintagmas nominales. Para proporcionar este tipo de información se utiliza el recurso léxico *WordNet español*.

Este algoritmo de resolución utiliza como datos de entrada la lista de todos los antecedentes previos en el texto y la propia DD, y proporciona como salida la solución a la DD si existe. La figura 5.12 muestra el esquema general del algoritmo desarrollado por Muñoz *et al.* (2000b).

```
procedimiento RESOLVER_ANAFORA_INDIRECTA(NP, Lista_Ant, Sol)
   Sea NP la descripción definida a procesar,
   Sea Lista_Ant la lista que contiene todos los NP previos,
   Sea Sol el NP propuesto como solución

   OBTENER_INFORMACION_SEMANTICA(NP, Listas_SEM)
   para cada ANT perteneciente a Lista_Ant hacer
        PERTENECE(NP, Lista_SEM,Lista_Cand)
   fin para
   si ELEMENTOS(Lista_Cand) = 0 entonces
        tipo_DD = "no anafórica"
   sino
     si ELEMENTOS(Lista_Cand) > 0 entonces
        APLICAR_HEURISTICAS(Lista_Cand,Sol)
        si ENCONTRADA(Sol) entonces
            DEVOLVER Sol
        fin si
     fin si
   fin si

fin procedimiento
```

Figura 5.12. *Resolución de las DDs con distinto núcleo (anáfora indirecta)*

Se observa que el algoritmo propuesto tiene, prácticamente, la misma estructura que el algoritmo de resolución de las DDs del mismo núcleo. En este algoritmo se observan tres etapas o pasos en lugar de dos. Dichos pasos son los siguientes:

1. *Obtención de la información semántica.* Para obtener la información semántica relacionada con la DD se extrae su núcleo y se consulta en el WordNet español todos los tipos de relaciones semánticas. Con toda la información semántica se construye una lista semántica para cada tipo de relación. Las listas están formada por todas las palabras sinónimas, hiperónimas

e hipónimas. Una de las principales diferencias entre el WordNet español y el inglés radica en que en la versión para el español incorpora relaciones entre elementos de diferentes categorías (nombres y verbos), tal y como se muestra en los trabajos de Gonzalo *et al.* (1998). La incorporación de este tipo de información al recurso léxico permite resolver las referencias denominadas de papel temático. Para ello, se construye una lista de papel temático utilizando las relaciones denominadas por el recurso léxico como *role-agent* y *involved-agent* que muestran las relaciones entre verbos y nombres. Por ejemplo, cazar-cazador.

2. *Obtención de la lista de candidatos.* Una vez generadas las listas semánticas relacionadas con la DD se recorre toda la lista de posibles antecedentes y aquéllos cuyo núcleo pertenezca a una de las listas semánticas se seleccionan y pasan a formar parte de la lista de candidatos. Además, para resolver las DDs del tipo denominado *papel temático*, se incorporan a la listas de candidatos aquellos sintagmas nominales de las oraciones cuyo verbo esté relacionado con el núcleo de la DD.
   Si la lista de candidatos está formada por más de un candidato se aplican el resto de pasos. Si sólo está formada por un candidato se aplicará la heurística H5 para comprobar que no existe ningún modificador antónimo que haga que se rechace como solución. Y si la lista de candidatos no tiene ningún candidato se clasifica la descripción definida como no anafórica y se almacena dicha DD en la lista de candidatos a antecedentes.

3. *Elección del candidato correcto.* Cuando la lista de candidatos está formada por más de un candidato, la elección del antecedente correcto se realiza aplicando heurísticas a modo de filtro. A continuación se presentan cada una de estas heurísticas:

   - **Heurística Hi1: (Sinónimo).** Esta heurística selecciona todos los candidatos cuyos núcleos son sinónimos al núcleo de la DD. El siguiente ejemplo muestra la aplicación de esta

heurística:

– *El coche oficial*$_i$ tiene un blindaje para evitar ataques terroristas.... *Este automóvil*$_i$ es capaz de resistir las bombas anticarro

La DD Este automóvil hace referencia a El coche oficial porque existe una relación de sinonimia entre coche y automóvil.

- **Heurística Hi2: (Hiperónimo-hipónimo).** La segunda heurística utilizada, es la selección entre los posibles candidatos aquellos cuyos núcleos tengan una relación de hiperonimia o hiponimia con el núcleo de la descripción definida.

  – *El coche oficial*$_i$ tiene un blindaje para evitar ataques terroristas.... *Este vehículo*$_i$ es capaz de resistir las bombas anti-carro

  La DD Este vehículo hace referencia a El coche oficial porque existe una relación de hiperonimia entre coche y vehículo.

- **Heurística Hi3: (Papel temático).** La heurística H3 selecciona los candidatos que tenga alguna relación de papel temático con la DD. Como en el siguiente ejemplo:

  – *Juan*$_i$ caza todos los fines de semana.... *Es el típico cazador*$_i$ que fanfarronea ....

  La DD el cazador hace referencia a Juan porque existe una relación de *role-agent* entre cazador y cazar. Cuando se forma la lista de candidatos a antecedentes se seleccionan todos los sintagmas nominales previos cuyos núcleos tienen una relación de sinonimia, hiperonimia o hiponimia, más aquellos que están relacionados a través de un papel temático. Como se observa en el ejemplo, (Juan y cazador)

no tienen ninguna relación entre sí, pero sí que existe una relación entre la DD y el verbo de la oración en la que aparece el antecedente (cazador y cazar). Cuando existe esta relación el sintagma nominal que hace la función de sujeto (Juan) se añade a la lista de candidatos a antecedentes. El analizador SUPP aplica unas heurísticas para establecer que sintagma nominal hace la función de sujeto o no.

- **Heurística Hi4: (Mismos modificadores).** Esta heurística selecciona aquellos candidatos que tengan los mismos modificadores que la descripción definida.

  – *La casa de campo*$_i$ es más entrañable que *la casa de ciudad*$_j$... *La vivienda de campo*$_i$ suele tener menos comodidades.

  La DD que aparece en la última oración, La vivienda de campo, hace referencia a la DD la vivienda de campo en lugar de hacer referencia a la DD la vivienda de ciudad. El algoritmo es capaz de elegir el candidato correcto debido a que el antecedente correcto tiene el mismo modificador, de campo, que la DD.

- **Heurística Hi5: (Modificadores relacionados).** Esta regla selecciona aquellos candidatos que tengan por lo menos uno de sus modificadores relacionados semánticamente con algún modificador de la DD. Si la relación entre los modificadores es una relación de antonimia el candidato se rechaza.

  – *El coche más rápido*$_i$ es fabricado en EE.UU..... *Este veloz vehículo*$_i$ es capaz de alcanzar los 300Km/h en pocos segundos

  La DD Este veloz vehículo hace referencia a la descripción definida El coche más rápido ya que los núcleos están relacionados a través de una relación de hiperonimia (vehículo y coche) y además existe una relación de sinonimia entre

los modificadores (veloz y rápido)

Una vez aplicadas las heurísticas anteriores, si más de un candidato pertenecen a la lista de candidatos se aplican las siguientes tres mismas heurísticas que en la resolución de la descripción definida del mismo núcleo.

- **Heurística Hi6: (Concordancia de género y número).** Se prefiere aquellos que tengan el mismo género y el mismo número. Existen, por ejemplo, sinónimos (afueras -alrededores) con distinto género. Por esta razón, esta heurística en la resolución de las DDs se utiliza como preferencia en lugar de una restricción como ocurre en la resolución de la anáfora pronominal.

- **Heurística Hi7: (Frecuencia).** Si una vez aplicada la concordancia de género y número más de un candidato permanece en la lista de candidatos, entonces se elige como candidato aquel que más veces haya sido seleccionado como antecedente de DDs previas.

- **Heurística Hi8: (Más cercano).** Si más de un candidato satisface la heurística Hi7 entonces se elige como candidato correcto de la DD el candidato más cercano en el texto.

A continuación se muestra un ejemplo completo del proceso de resolución de las anáforas indirectas producidas por las DDs.

[Las familias de [hoy en día]$_1$]$_2$ disponen de [[una vivienda en [el campo]$_3$]$_4$ y [otra en [la ciudad]$_5$]$_6$]$_7$. [Las casas de [campo]$_8$]$_9$ suelen ser [más entrañables]$_{10}$ que [las casas de [la ciudad]$_{11}$]$_{12}$ ya que [el ambiente rural]$_{13}$ invita a pasar [momentos muy agradables]$_{14}$ frente a [una chimenea]$_{15}$. [Las construcciones urbanas]$_{16}$ suelen ser [grandes edificios]$_{17}$ que aprovechan todo [el espacio disponible]$_{18}$. [Juan]$_{19}$ se compró [una casa]$_{20}$ en [el campo]$_{21}$ para descansar. [Esta casa]$_{22}$ estaba embargada y le causó [muchos problemas]$_{23}$. Finalmente, le

dijeron que [el comprador de [la casa]$_{24}$]$_{25}$ tenía que pagar todas [las deudas]]$_{26}$.

En la tabla 5.3 se muestra el resultado de aplicación del algoritmo de resolución de las anáforas indirectas producidas por la DD al ejemplo anterior. En esta tabla se presentan los antecedentes que satisfacen cada una de las heurísticas para cada una de las DDs. Algunas de las DDs de esta tabla, como por ejemplo *11, 21, 22, 24*, se resuelven con el algoritmo de resolución de las DDs del mismo núcleo, por eso su solución en la tabla contiene el subíndice *MN*. El resto de descripciones definidas o bien introducen una nueva entidad (NE en la solución), o hacen referencia a un antecedente previo. La DD *9* las casas de campo y la DD *12* las casas de la ciudad utilizan la misma heurística, Hi4, para obtener el candidato correcto *4* una vivienda en el campo y *6* otra vivienda en la ciudad, respectivamente. Se observa que la DD número *6* otra en la ciudad hay una elipsis del núcleo. El analizador parcial utilizado incorpora un módulo para la resolución de la elipsis que proporciona el núcleo correcto casa, por lo que, para el algoritmo de resolución de las DDs es como si apareciese el sintagma completo. Para resolver la referencia producida por las DDs número *16* las construcciones urbanas se necesitan aplicar las heurísticas Hi2, Hi5 y Hi6. La primera heurística proporciona cuatro posibles antecedentes cuyos núcleos están relacionados a través de una relación de hiperonimia (construcción - vivienda y construcción - casa). La heurística Hi5 selecciona aquellos que tienen modificadores relacionados (*6* otra vivienda en la ciudad y *12* las casas de la ciudad). Finalmente, la heurística de concordancia de género y número selecciona la DD número *12* las casas de la ciudad. La heurística Hi3 se aplica para resolver la referencia de la DD *25* el comprador proponiendo como solución la DD *19* Juan.

5.5.4 Conclusiones del algoritmo ARH

Este algoritmo de resolución de las DDs, basado en el uso de heurísticas en forma de filtro, aunque, en general, presenta bue-

DD	Hi0	Hi1	Hi2	Hi3	Hi4	Hi5	Hi6	Sol.
3								NE
5								NE
9	4,6	4,6			4			4
11								5_{MN}
12	4,6,7,9	4,6,7,9			6			6
13								NE
16	4,6,7,9,12		4,6,7,9,12			6,12	12	12
18								NE
21								3_{MN}
22								20_{MN}
24								22_{MN}
25	19			19				19

Tabla 5.3. Heurísticas aplicadas en la resolución de la anáfora indirecta

nos resultados como se verá en el capítulo de experimentación, también tiene algunos problemas.

Los principales problemas detectados son:

- El algoritmo de resolución ARH intenta resolver las referencias de todas las DDs sin realizar una identificación previa del tipo de DD (anafórica o no anafórica). Evidentemente este algoritmo emplea un elevado tiempo computacional en intentar resolver algo que no tiene solución, sobre todo sí se tienen en cuenta los estudios realizados por diversos autores en los que se demuestra que más de la mitad de las DDs que están presenten en un texto son no anafóricas.

- Otro problema relacionado con el coste computacional del algoritmo es que el espacio de búsqueda del antecedente correcto está formado por todas las oraciones previas a la oración en la que aparece la DD. En textos grandes el número de comparaciones se hace muy elevado lo que supone una excesiva lentitud del algoritmo. La reducción del espacio de búsqueda ocasiona una disminución en la precisión obtenida.

- Por otro lado, se establece un orden de prioridad para aplicar los procesos de resolución de cada tipo de anáfora. Primero se aplica la resolución de las DDs con el mismo núcleo y posteriormente, si no se ha obtenido solución, se aplica el algoritmo de resolución de las anáforas indirectas. Este orden de aplicación del algoritmo se establece al realizar un estudio del corpus de entrenamiento. Aplicar este orden de resolución provoca que, en ocasiones, no se escoja el antecedente más adecuado aunque el escogido pertenezca a la misma cadena de correferencia que el correcto.

- La forma de aplicar las heurísticas, forma de filtro, provoca que el orden en el que se establecen las heurísticas puede rechazar candidatos que satisfagan primeras heurísticas pero sean más apropiados que la solución propuesta.

Como consecuencia de los problemas detectados en este primer enfoque se planteó un segundo enfoque con los siguientes objetivos:

- Realización de una fase previa de identificación de las DDs no anafóricas.

- Reducción del número de comparaciones a realizar para la búsqueda del antecedente correcto.

- Optimización del proceso de resolución mediante la aplicación de heurísticas basadas en pesos.

El apartado siguiente recoge este segundo enfoque.

5.6 Enfoque 2: Algoritmo de resolución de las DDs basado en la generación automática de una red semántica (AGSN)

Para el desarrollo de este enfoque se partió de las ideas de Frege (1892) y Russell (1919) que definen la anáfora como un fenómeno semántico. Este enfoque se basa en la generación automática de una red semántica (AGSN, *Automatic Generation of a Semantic Network*) conforme se procesa el texto utilizando la ontología usada por el recurso léxico EuroWordNet.

Este segundo enfoque soluciona los problemas que presentaba el enfoque anterior:

- *Espacio de búsqueda.* El problema del elevado número de comparaciones que había que realizar en el enfoque anterior para la obtención del antecedente adecuado, se soluciona en este enfoque puesto que una DD que tenga una relación de correferencialidad de tipo *identidad* con otro sintagma nominal sólo puede existir si ambos pertenecen a la misma clase semántica. Por lo tanto, la DD sólo se compara con los sintagmas nominales pertenecientes a la misma clase, aunque el espacio de búsqueda sigue estando formado por todas las oraciones previas desde el inicio del texto. En este enfoque se tratan las relaciones de sinonimias, hiperonimias e hiponimias. Un estudio de las características de las diferentes palabras relacionadas mediante los tres tipos anteriores de relaciones semánticas en el recurso léxico (WordNet español) demuestra que todas ellas pertenecen a la misma clase semántica, como puede verse en los siguientes ejemplos:

  – sinonimia. La palabra casa tiene como posibles sinónimos las palabras domicilio, hogar y morada, entre otros y sus conceptos bases de la ontología son 1stOrderEntity Artifact Building Form Function Group Object Origin para todas las palabras pertenecientes al mismo *synset*.

– hiperonimia-hiponimia. Para la misma palabra anterior, casa, las relaciones de hiperonimia nos proporcionan vivienda cuya clase en la ontología está formada por los conceptos base 1stOrderEntity Artifact Building Form Function Group Object Origin que como se puede observar son los mismos que para la palabra casa. Lo mismo ocurre con la relación de hiponimia, una palabra hipónima de casa es la palabra chalet cuya clase semántica también está formada por los conceptos bases 1stOrderEntity Artifact Building Form Function Group Object Origin.

Este hecho permite agrupar todas las DDs pertenecientes a la misma clase y, por tanto, reducir el número de comparaciones a realizar.

- *Identificación previa.* La identificación previa del tipo de DD realizada por otros autores se basan en la identificación de algunos modificadores denominados restrictivos y de algunas construcciones sintácticas. El principal problema de estos trabajos es la propagación de errores a la fase de resolución. En este segundo enfoque, se presenta un método para la identificación previa de las DDs no anafóricas basado en la construcción de una red semántica. Como se ha visto cuando se comentaba el problema del *elevado número de comparaciones*, todas las DDs relacionadas semánticamente (relación de sinonimia, hiperonimia e hiponimia) pertenecen a la misma clase o concepto ontológico. Por tanto, se puede identificar una DD como no anafórica porque será la primera que active dicho concepto en la red pues no existe ningún sintagma nominal con el que pueda ser comparada. El número de DD que se consigue identificar con este método no es muy elevado pero no se propaga ningún tipo de error a la fase de resolución.

- *Solución correcta.* La utilización de la propiedad transitiva de la referencia en la evaluación, encubre el problema de la elección de un antecedente que no es el estrictamente correcto, pero que pertenece a la misma cadena de correferencia que el anteceden-

te correcto. En ocasiones, el sistema de filtrado empleado en la aplicación de las heurísticas, puede no seleccionar el candidato correcto y, por tanto, cualquier otra heurística posterior propone un candidato diferente. Para solucionar este problema, en este segundo enfoque se ha optado por utilizar las heurísticas con un sistema de peso en lugar del sistema de filtrado del primer enfoque. A cada heurística se le asigna un peso de forma experimental. Se aplican todas las heurísticas a todos los posibles antecedentes. Por cada heurística que satisfagan se le suma el peso correspondiente a esa heurística. La suma total de los pesos forma lo que se ha denominado su *factor de importancia*. Finalmente, el algoritmo escoge el antecedente con mayor *factor de importancia*.

```
procedimiento AGSN (Texto)
    Sea Texto el texto a procesar,

  CREAR_RED(Red)
  para cada NP del Texto hacer
    OBTENER_CLASE_SEMANTICA(NP,Clase)
    Si EXISTE_CLASE(Red, Clase) entonces
        TIPO_NP(NP, TipoNP)
        si TipoNP="definido" entonces
            RESOLVER_ANAFORA_PESOS(Red, NP, Clase, TipoDD, Sol)
            si TipoDD="Anafórico" entonces
                ALMACENAR_CORREFERENCIA(Lista_COrref,NP,Sol)
            fin si
        fin si
    fin si
    ALMACENAR_NP(Red, Clase, NP)
  fin para
  DEVOLVER Red, Lista_COrref
fin procedimiento
```

Figura 5.13. *Mecanismo AGSN (segundo enfoque)*

El esquema general de este segundo enfoque se muestra en la figura 5.13. En este esquema se observan dos fases principales que a continuación se detallan:

- Construcción de la red semántica y clasificación del tipo de DD.

- Resolución de las referencias lingüísticas.

5.6.1 Construcción de la red semántica y clasificación de la DD

El proceso de generación automática de la red semántica por el mecanismo AGSN (Muñoz & Palomar, 2001) es el siguiente:

- **Paso 1. Búsqueda de sintagmas nominales**. El mecanismo recorre el texto analizado sintácticamente y cuando encuentra un sintagma nominal extrae su núcleo.

- **Paso 2. Obtención de los conceptos bases**. Se consulta la glosa del lema del núcleo en el WordNet español. En la glosa se encuentran todos los conceptos bases de dicha palabra. Dichos conceptos bases se utilizan para obtener la clase de la palabra. Se identifican tantas clases como posibles combinaciones de los conceptos se puedan dar. Es decir, dos palabras con cuatro conceptos base cada una, donde tres de ellos son iguales en ambas y uno distinto, pertenecen a clases distintas. Para las palabras polisémicas se debe escoger el sentido correcto. Para ello, se debe aplicar un desambiguador del sentido de las palabras.

- **Paso 3. Almacenamiento del sintagma nominal en la red semántica y clasificación de las DDs**. Una vez se obtiene la clase a la que pertenece la palabra, se busca dicha clase en la red semántica. Pueden ocurrir dos casos:

  1. *La clase semántica no existe en la red semántica.* En este caso se crea el nuevo concepto ontológico (clase) en la red y se almacena el sintagma nominal en la lista de candidatos de esta nueva clase. Si además el sintagma nominal es una DD entonces se clasifica como *no anafórica.*

2. *La clase existe en la red semántica.* Si la clase semántica a la que pertenece la palabra ya existe en la red, entonces, si se trata de una DD, se le aplica el algoritmo de resolución. Si se trata de cualquier otro tipo de sintagma nominal se almacena en la lista de candidatos de esa clase. El algoritmo de resolución de las DDs se encargará de clasificarla finalmente como anafórica o no anafórica y de proporcionar el antecedente adecuado si es anafórica.

La figura 5.14 muestra la estructura que tiene una de las clases de la red semántica. En ella se observa que varios conceptos (palabras) están relacionados dentro de la misma clase a través de relaciones de sinonimia, hiperonimia e hiponimia.

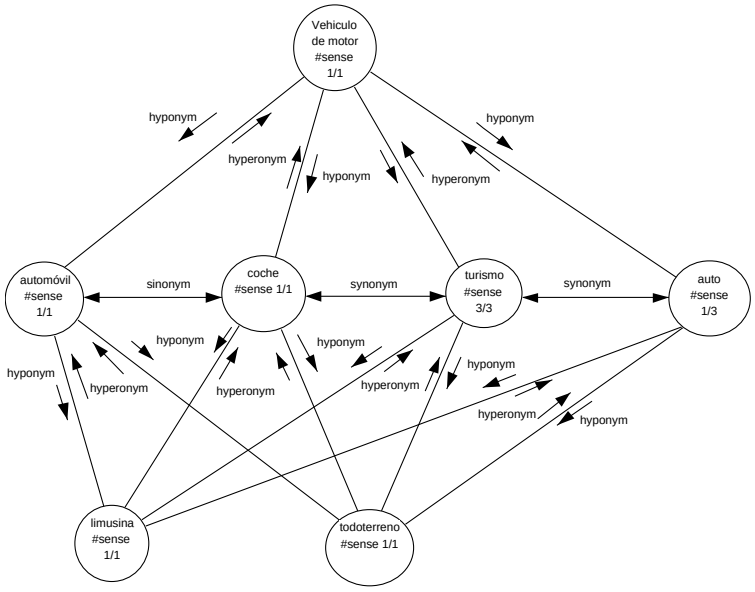

Figura 5.14. *Porción de una clase de la red semántica*

Después de presentar la generación de la red semántica, se presenta el algoritmo de resolución que utiliza dicha red.

5.6.2 Resolución de las referencias usando la red semántica

A diferencia de como se hace en el primer enfoque, la resolución de las DDs dentro de la red semántica no hace distinción entre las anáforas directas e indirectas producidas por las DDs. Este algoritmo establece un conjunto de heurísticas pesadas que se aplican a todos los antecedentes que pertenecen a la misma clase semántica que la DD, para obtener el factor de importancia de cada antecedente. El algoritmo selecciona como antecedente correcto el antecedente con mayor factor de importancia. Los valores de los pesos utilizados por este algoritmo se han ajustado experimentalmente, un resumen de ellos se observa en la tabla 5.4.

Heurística	**Tipo**	**Pesos**
HP1	Mismo núcleo	50
HP2	Núcleo sinónimo	40
	Núcleo hiperónimo	35
	Núcleo hipónimo	35
HP3	Mismo modificador	10
	Modificador sinónimo	9
	Modificador hiperónimo	8
	Modificador hipónimo	8
	Modificador antónimo	-1000
HP4	Concordancia gen. y num.	2
HP5	Frecuencia	2

Tabla 5.4. Pesos utilizados por cada heurística

El esquema de este algoritmo se muestra en la figura 5.15. Se distinguen los siguientes pasos o etapas principales:

1. *Obtención de la lista de candidatos.* Dos palabras pertenecientes a la misma clase semántica podrán ser correferentes si tienen el mismo núcleo o si entre ellas existe una relación de sinonimia (coche - auto), de hiperonimia (coche - vehículo) o de hiponimia (coche - ambulancia) y no existe una relación

```
procedimiento RESOLVER_ANAFORA_PESOS(Red, NP, Clase, TipoDD, Sol)
   Sea Red la red semántica,
   Sea NP la descripción definida,
   Sea Clase la clase a la que pertenece el NP,
   Sea TipoDD el tipo de descripción definida,
   Sea Sol el NPc solución al NP

    OBETENER_ANTECEDENTES(Red, Clase, Lista_ANT)
    para cada A perteneciente a Lista_ANT hacer
        RELACIONAR_NUCLEOS(NP, A, Valor)
        si Valor > 0 entonces
            ALMACENAR(NP, Lista_CAND)
        fin si
    fin para
    si ES_VACIA(Lista_CAND) entonces
        TipoDD="no anafórico"
    sino
        para cada C perteneciente a Lista_CAND hacer
            para Contador = 1 hasta Max_Heurística hacer
                APLICAR_HEURISTICA(Contador, C, Valor)
                Valor_c = Valor_c + Valor
            fin para
        fin para
        OBTENER_MAXIMO(Lista_CAND, Maximo)
        si Maximo <= 0 entonces
            TipoDD="no anafórico"
            DEVOLER TipoDD
        sino
            SELECCIONAR_MAXIMOS(Lista_CAND, Maximo, Lista_Max)
            si ELEMENTOS(Lista_Max) > 1 entonces
                MAS_CERCANO(Lista_MAX, Sol)
            sino
                Sol = Lista_MAX
            fin si
            DEVOLVER Sol
        fin si
    fin si
    DEVOLVER TipoDD
fin procedimiento
```

Figura 5.15. *Algoritmo de resolución de las DDs usando pesos*

semántica negativa entre algún modificador de ambos sintagmas. Estas relaciones se usan para rechazar todos los elementos (sintagmas nominales) que pertenecen a la misma clase semántica que la DD y no satisfacen ninguna de estas relaciones. Para ello, se aplican las siguientes heurísticas:

- **Heurística HP1: (Mismo núcleo).** Si el candidato tiene el mismo núcleo que la DD se le suma un valor de 50 al factor de importancia.

- **Heurística HP2: (Relación semántica entre núcleos).** Si el núcleo de la DD es:

  – *Sinónimo.* Si la relación es la de sinonimia se le añade al factor de importancia un valor de 40.

  – *Hiperónimo-hipónimo.* Si la relación es de hiperonimia o de hiponimia se le añade al factor de importancia un valor de 35.

Todos aquellos antecedentes con un valor del factor de importancia mayor que cero pasan a formar parte de la lista de candidatos. Un ejemplo de aplicación de las heurísticas HP1 y HP2 podría ser el siguiente:

En la clase ontológica 1stOrderEntity Artifact Form Function Instrument Object Origin Vehicle podemos tener sintagmas con núcleos como coche, vehículo, automóvil, avión, biplano, jet, reactor. Si se intenta resolver una descripción definida del tipo el coche de carrera, al aplicar esta heurística los *factores de importancia* de cada antecedente quedarían del siguiente modo: coche (50), vehículo (35), automóvil (40), avión (0), biplano (0), jet (0) y reactor (0). Por tanto, los sintagmas nominales coche (50), vehículo (35), automóvil (40) pasarían a formar parte de la lista de candidatos.

2. *Selección del candidato adecuado.* Si la lista de candidatos no contiene ningún elemento, entonces la DD se clasifica como no anafórica. Si por el contrario la lista de candidatos tiene algún elemento hay que estudiar las posibles relaciones con la DD. Para la elección del candidato más adecuado también se utilizan heurísticas pesadas que añaden un valor determinado al factor de importancia en función del tipo de heurísticas que se satisfagan. Las heurísticas utilizadas son:

   - **Heurística HP3: (Modificadores).** Si el candidato tiene modificadores que están relacionados con los modificadores de la DD se añaden los siguientes valores al factor de importancia:

     – Si es el mismo modificador se añade un valor de 10.

     – Si hay una relación de sinonimia un valor 9. Por ejemplo rápido y veloz en las descripciones el coche más rápido y el veloz automóvil.

     – Si hay una relación de hiperonimia o hiponimia entre los modificadores se añade un valor de 8 al factor de importancia. Como por ejemplo, roble y madera en las DDs la caja de roble y la caja de madera.

     – Si hay una relación de antonimia se le añade un valor de -1000 para que el antecedente sea rechazado. Como por ejemplo, izquierda y derecha en las descripciones la oreja derecha y la oreja izquierda.

   - **Heurística HP4: (Concordancia género y número).** Se añade un valor de 2 al factor de importancia de cada uno de los candidatos que concuerden en género y número.

   - **Heurística HP5: (Frecuencia).** Se añade un valor de 2 al factor de importancia del candidato o de los candidatos

más frecuentes en el texto.

Una vez aplicadas todas las heurísticas a cada uno de los posibles candidatos, si más de uno tiene el valor máximo se elige el más cercano. Si el antecedente con valor máximo tiene un valor negativo o igual a 0 entonces la DD se clasifica como no anafórica.

A continuación se muestra el funcionamiento del algoritmo a través de un ejemplo. En el siguiente texto encontramos diferentes entidades pertenecientes a diferentes conceptos ontológicos. Conforme se procesa el texto se consultan cada uno de los núcleos de los sintagmas nominales en la ontología de EuroWordNet para extraer el concepto al que pertenece.

[El coche de [competición]$_1$]$_2$ que [Juan López]$_3$ se compró le costó [diez millones de [pesetas]$_4$]$_5$. [Este coche]$_6$ había sido usado por [corredores profesionales]$_7$. [El coche de [competición]$_8$]$_9$ quedó destrozado en [un accidente]$_{10}$ en [el circuito del [Jarama]$_{11}$]$_{12}$. [El coche de [carrera]$_{13}$]$_{14}$ fue vendido como [chatarra]$_{15}$. [Los coches de [competición]$_{16}$]$_{17}$ para poder competir en [carreras del [campeonato del [mundo]$_{18}$]$_{19}$]$_{20}$ tienen que cumplir [unas normas muy estrictas]$_{21}$. [Estos coches]$_{22}$ son revisados continuamente para que [ningún piloto]$_{23}$ pueda aprovecharse de [ciertas ventajas mecánicas]$_{24}$ que le proporcione [algunos segundos de [ventaja]$_{25}$]$_{26}$ frente a [otros pilotos]$_{27}$. [La mayoría de [fabricantes de [coches de [competición]$_{28}$]$_{29}$]$_{30}$]$_{31}$ lo son también de [coches utilitarios]$_{32}$. [Estos coches]$_{33}$ no pueden ser comprados por [cualquier persona]$_{34}$. [Carlos Sainz]$_{35}$ es [un piloto de [rallies]$_{36}$]$_{37}$ que ha ganado [el campeonato del [mundo de [rallies]$_{38}$]$_{39}$]$_{40}$. [Marc Gené]$_{41}$ y [Pedro de la Rosa]$_{42}$ son [pilotos de [Formula I]$_{43}$]$_{44}$. [Estos pilotos]$_{45}$ no han ganado [ningún campeonato]$_{46}$.

A continuación mostramos de modo intuitivo y mediante un ejemplo el funcionamiento del algoritmo propuesto.

1. Se recorre el texto hasta que se encuentra un sintagma nominal. De cada sintagma nominal se extrae su núcleo (figura

5.16) que será la entrada utilizada en la búsqueda de su concepto ontológico en el recurso WordNet español.

** SINT.NOMINAL: ** SINT.NOMINAL SIMPLE: ** DETERMINANTE 1: ** ADJETIVO SIMPLE: este ** SUSTANTIVO: coche	** SINT.NOMINAL: ** SINT.NOMINAL SIMPLE: ** SUSTANTIVO: corredor ** ADYACENTE ADJETIVO: ** ADJETIVO SIMPLE: profesional
Núcleo = **coche** Tipo = **descripción definida** Semántica= **1stOrderEntity Artifact Form Function Instrument Object Origin Vehicle**	Núcleo = **corredor** Tipo = **sintagma nominal** Semántica = **1stOrderEntity Animal Form Living Natural Object Origin**

Figura 5.16. *Sintagmas nominales extraídos de la segunda oración*

2. Se busca la clase semántica o concepto ontológico de cada núcleo en la ontología del EuroWordNet. La primera vez que se encuentre un determinado concepto ontológico se activará su nodo en la red creada para ello (figura 5.17) creando una lista de candidatos para esa clase que contiene sólo el sintagma nominal tratado. Si la aparición de ese sintagma nominal supone la activación del nodo y se trata de una DD, se clasificará como no anafórica. Si por el contrario, el nodo ya estaba activado y se trata de una descripción definida se clasificará, en un primer momento, como anafórica y se le aplicará el algoritmo de resolución de las correferencias lingüísticas. Si este algoritmo no proporciona ninguna solución, entonces finalmente se clasificará como no anafórica.

La figura 5.18 muestra la red semántica obtenida para el procesamiento del ejemplo anterior. Cada concepto ontológico dispone de una lista en la que se almacenan cada uno de los sintagmas nominales que pertenecen a ese concepto. En la figura se representan con el número del subíndice que tiene el sintagma nominal en el ejemplo. Además, se representa el antecedente propuesto por el algoritmo de resolución con su número de sintagma, en el ejemplo, entre paréntesis. Se puede observar que hay siete descripciones definidas anafóricas, seis de ellas están en la misma clase semántica (vehículo). Estas descripciones definidas se comparan sólo con los

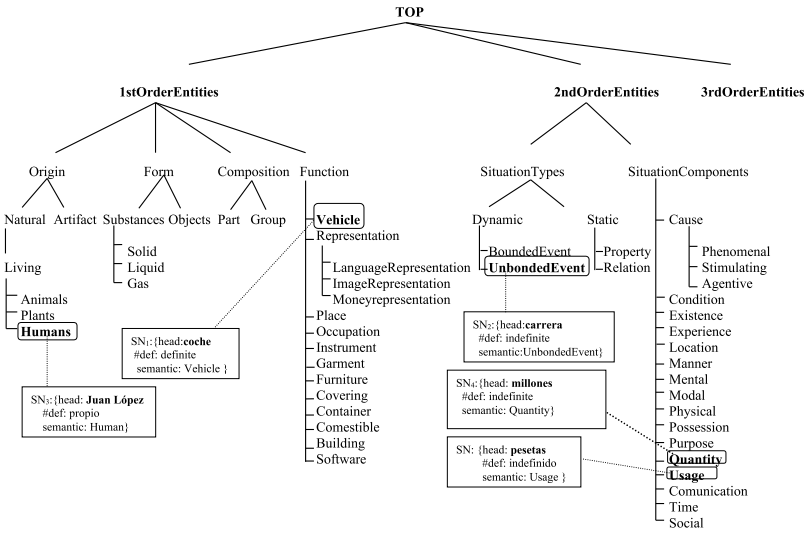

Figura 5.17. *Ejemplo de generación automática de la red semántica*

sintagmas nominales pertenecientes a esa clase, reduciendo así el número de comparaciones. Por ejemplo, la descripción definida etiquetada como 33 (Estos coches) sólo se compara con 9 sintagmas nominales en lugar de hacerlo con los 32 sintagmas anteriores en el texto. También, la descripción definida etiquetada con 40 (El campeonato del mundo de rallies) se compara sólo con seis sintagmas previos en lugar de con los 39 anteriores. Se presenta un método que no sólo distingue entre las descripciones definidas anafóricas de las que no lo son, sino que, además, presenta un algoritmo de resolución de las referencias que reduce el número de comparaciones a realizar para proporcionar el antecedente correcto manteniendo el tamaño del espacio de búsqueda (todas las oraciones previas). Por lo tanto, mantiene la eficacia del algoritmo y mejora la eficiencia temporal del mismo.

La tabla 5.5 muestra los valores de cada candidato al resolver cada DD. Por ejemplo, existen tres antecedentes posibles para resolver correctamente la DD número *14* (El coche de carrera). De estos tres antecedentes, dos tienen el mismo factor de importancia (59), *2* (El coche de competición) y *9* (El coche de competición), por lo que finalmente hay que aplicar la heurística del más cerca-

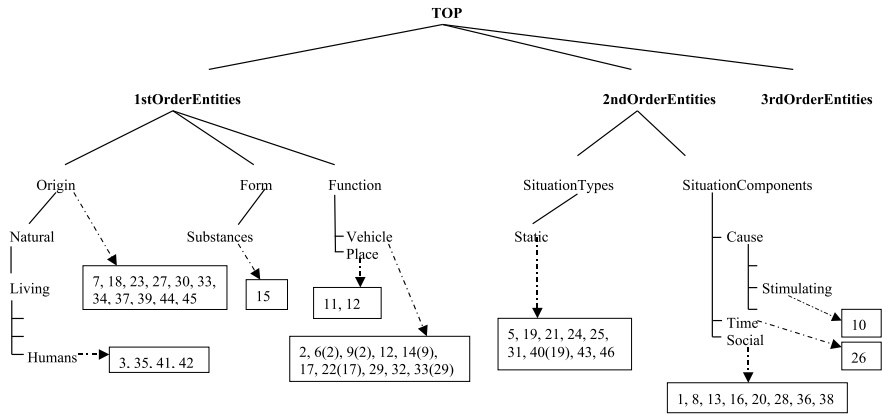

Figura 5.18. *Ejemplo de la red semántica*

no, que elige el antecedente *9* (El coche de competición). En este caso los dos antecedentes son iguales, por eso es lógico que los dos obtengan el mismo factor de importancia. En el caso de las DDs número *22* (Estos coches) y número *45* (Estos pilotos), de los 5 posibles antecedentes, sólo uno tiene el factor de importancia más alto, siendo este el elegido como antecedente correcto, *17* (Los coches de competición) y *44* (pilotos de Formula I), respectivamente.

5.6.3 Conclusiones del AGSN

Como ya ha citado anteriormente, este algoritmo de resolución de las referencias producidas por las DDs, basado en la construcción automática de una red semántica, resuelve los problemas encontrados con el primer enfoque:

- Identifica previamente algunas de las DDs no anafóricas sin propagar errores a la fase de resolución.

- Reduce el número de comparaciones a realizar sin reducir el espacio de búsqueda puesto a que la lista de candidatos a antecedentes sólo está formada por los sintagmas nominales pertenecientes a la misma clase.

N. sintagma	N. candidato	Pesos	F.Impor	Solución
2	-			-
6	2	50	50	2
9	2	50+10	60	2
	6	50	50	
11	-			-
12	-			-
14	2	50+9	59	
	6	50	50	
	9	50+9	59	9
17	2	50+10	60	
	6	50	50	
	9	50+10	60	9
	14	50+9	59	
18	-			-
19	-			-
22	2	50	50	
	6	50	50	
	9	50	50	
	14	50	50	
	17	50+2	52	17
31	-			-
33	2	50	50	
	6	50	50	
	9	50	50	
	14	50	50	
	17	50+2+2	54	
	22	50+2+2	54	22
	29	50+2	52	
	32	50+2	52	
39	18	50	50	18
40	19	50	50	19
45	7	45	45	
	23	50	50	
	27	50+2	52	
	37	50	50	
	44	50+2+2	54	44

Tabla 5.5. Pesos aplicados al ejemplo anterior

- Al utilizar un sistema de pesos en lugar del sistema de filtro (primer enfoque) no se descartan antecedentes en etapas previas, sino que a todos los antecedentes se les aplican todas las heurísticas y se selecciona aquél que tenga mayor factor de importancia (peso).

- Otra ventaja que proporciona el sistema de pesos con respecto al sistema de filtrado es que se puede tratar las DDs del mismo núcleo y las anáforas indirectas al mismo tiempo sin tener que aplicar dos conjuntos de heurísticas distintos.

5.7 Conclusiones

En este capítulo se han desarrollado dos enfoque para el tratamiento y resolución de las descripciones definidas. El primer enfoque consiste en la aplicación de un algoritmo basado en heurísticas (ARH) cuyas características principales son:

- No realiza identificación previa del tipo de DD (anafórica o no anafórica). por lo que trata de resolver todas las DDs.

- Resolución independiente de las anáforas directas e indirectas producidas por las DDs.

- Espacio de búsqueda de la solución formado por todas los sintagmas nominales previos.

- Utilización del sistema de heurísticas como un filtro.

- Resolución de las anáforas directas (DDs con el mismo núcleo) e indirectas (relacionadas semánticamente y papel temático).

Por otro lado, el segundo enfoque basado en la generación automática de una red semántica (AGSN) tiene las siguientes características principales:

- Identificación previa de las DDs no anafóricas.
- Resolución conjunta de las anáforas directas e indirectas.
- Mantenimiento del espacio de búsqueda de la solución pero con reducción del número de comparaciones a realizar.
- Aplicación de las heurísticas con un sistema de pesos.
- Resolución de las anáforas directas y parte de las indirectas, concretamente las relacionadas semánticamente.

En el siguiente capítulo se presenta la experimentación realizada sobre los dos enfoques para la obtención óptima del mejor conjunto de heurísticas.

6. Experimentación

En este capítulo se presenta la evaluación realizada a los dos enfoques presentados en el capítulo anterior. En primer lugar, se presentan las características de los corpus utilizados, tanto para el entrenamiento como para la evaluación de ambos enfoques. También se presentan los resultados obtenidos al aplicar el test de fiabilidad a cada corpus después de la anotación de dichos corpus con información de correferencialidad por parte de varios anotadores. Además, se presentan las características del conjunto de etiquetas usadas para la anotación de los corpus. Finalmente, se muestran los experimentos realizados con el objetivo de mejorar los algoritmos hasta conseguir una versión final que ha sido evaluada con el corpus de evaluación.

6.1 Descripción de los corpus

El entrenamiento y evaluación de los algoritmos, anteriormente presentados, se han realizado sobre diversos corpus. Se han utilizados tres corpus diferentes sobre los que se ha realizado toda la experimentación. Uno de los principales objetivos, marcados en los comienzos de esta Tesis fue el desarrollo de un sistema de resolución de las referencias lingüísticas producidas por las DDs en textos no restringidos, aunque la necesidad que originó este trabajo fue la integración del algoritmo de resolución de este tipo de expresiones anafóricas en el sistema de extracción de información EXIT que trabaja en el dominio restringido de las escrituras notariales. Por ello, se utilizan dos corpus de dominio no restringido (LEXESP y Periódicos) y otro de dominio restringido (escrituras de compraventa).

A continuación se presentan las principales características de cada uno de los corpus utilizados:

- **LEXESP (L)**. Es un corpus en español, que contiene 5 millones de palabras etiquetadas léxicamente, perteneciente al proyecto del mismo nombre desarrollado por 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 Tratamiento del Lenguaje de la Universidad Politécnica de Cataluña. Los diferentes textos que forman este corpus tienen distintos estilos y son de diversos autores. Una parte del etiquetado léxico-morfológico fue revisado manualmente por varios etiquetadores. El etiquetado consiste en añadir a cada palabra de este corpus la raíz de la palabra, una etiqueta identificativa propia de su tipo gramatical (nombre, adjetivo, verbo, pronombre, etc.) e información de concordancia (género, número, persona). El conjunto de etiquetas utilizado fue el juego de etiquetas PAROLE. Para este trabajo, se ha utilizado un pequeño fragmento del corpus formado por unas 60.000 palabras. Una parte de este fragmento se ha utilizado para realizar un estudio a partir del cual se han obtenido los diferentes tipos de DDs existentes en un texto, tal y como se ha presentado en el capítulo 4, y para el entrenamiento de los diferentes algoritmos. Otra parte diferente del fragmento se utiliza para la evaluación de los algoritmos.

- **Escrituras notariales (E)**. Este corpus está formado por diversas escrituras de compraventa de inmuebles. No contiene ningún tipo de información léxico-morfológica supervisada, por lo que los errores que pudieran ser cometidos por el analizador léxico utilizado son arrastrados a los módulos de procesamiento posteriores, entre ellos, el algoritmo de resolución de las DDs. Al tratar sobre un dominio restringido y tan estricto como son los documentos legales de compraventa (escrituras notariales), todos los textos que forman el corpus tienen la misma estructura y características. Este corpus es el utilizado por el sistema de

extracción de información EXIT, y sobre él se realizó el estudio de la resolución de las DDs en sistemas de extracción de información, (Palomar & Muñoz, 2000) La aplicación del algoritmo de resolución de las DDs puede mejorarse para dominios restringidos con algunas heurísticas adicionales propias del dominio. Sin embargo, dado que el objetivo de este trabajo es desarrollar un algoritmo de tratamiento y resolución de las DDs en textos no restringidos, se ha evaluado este corpus usando el mismo conjunto de heurísticas que para el resto de los corpus.

- **Periódicos digitales (P)**. Está formado por artículos de prensa, deportiva principalmente, extraídos de las páginas web de diversos periódicos digitales. Por lo tanto, el punto de partida, al igual que el corpus anterior, son textos sin ningún tipo de información léxico-morfológica adicional, aunque sí posee algunas etiquetas HTML. Estas etiquetas HTML son etiquetas de formato, principalmente, tales como etiquetas para cursivas (`<i>...</i>`), negrillas (`<b>...</b>`) y encabezados (`<h1>, <h2>`, etc). Además, tiene otras etiquetas de hiperenlaces (`<a href=.....> .... </a>`) y de inserción de imágenes (`<img scr=...>`). Al igual que para el corpus de las escrituras de compraventa se necesita un analizador léxico que proporcione la información léxico-morfológica necesaria para los procesos posteriores, antes de procesar este tipo de textos, se necesita de un preproceso que elimine estas etiquetas para poder procesar exclusivamente el texto.

El analizador léxico utilizado para incorporar la información léxico-morfológica ha sido el desarrollado por Pla (2000) y Pla *et al.* (2000).

La tabla 6.1 presenta el número de DDs que se han encontrado, de forma manual, en cada uno de los corpus y los diferentes tipos de DDs encontradas. Los corpus tienen un total de 2132 descripciones definidas, de las que 1315 son descripciones definidas no anafóricas, es decir, introducen una nueva entidad en el discurso, y 817 descripciones definidas tienen propiedades referenciales. La

distribución de las descripciones definidas anafóricas es la siguiente: 566 son DDs tienen el mismo núcleo que su antecedente, 178 DDs cuyos núcleos tienen una relación de sinonimia, hiponimia o hiperonimia con el núcleo de su antecedente, y 73 DDs que tienen una relación de papel temático con su antecedente.

Corpus	**Total**	**Nueva entidad**	**Mismo núcleo**	**Relación semántica**	**Papel temático**
LEXESP	1164	785	262	90	27
Escrituras	607	331	205	43	28
Periódicos	361	199	99	45	18
Total	2132	1315	566	178	73

Tabla 6.1. Distribución de las DDs en los corpus

Una parte de estos corpus se ha utilizado como *corpus de entrenamiento* (47%) y otra parte como *corpus de evaluación* (53% restante). La distribución final de los corpus queda tal y como se muestra en la tabla 6.2. Por una parte, el corpus de entrenamiento está formado por un total de 915 descripciones definidas de las que 558 no tienen propiedades referenciales y 357 sí tienen propiedades referenciales. La distribución de las DDs con propiedades referenciales en el corpus de entrenamiento es: 248 DDs tienen el mismo núcleo que su antecedente, el núcleo de 83 DDs tienen una relación semántica (sinonimia, hiperonimia o hiponimia) con el núcleo de su antecedente y 26 DDs desempeñan un papel temático. Por otra parte, el corpus de evaluación está formado por un total de 1217 DDs divididas de la siguiente forma: 757 descripciones definidas que no tienen propiedades referenciales y 460 con propiedades referenciales. La distribución en el corpus de evaluación de las DDs con propiedades referenciales es: 318 DDs con el mismo núcleo que su antecedente, 95 DDs relacionadas semánticamente y 47 de papel temático.

Los corpus utilizados contienen información de correferencialidad para realizar la evaluación automática. Esta información es proporcionada en un archivo diferente a través de hechos prolog con la siguiente estructura para cada frase del corpus que conten-

Tipo	Corpus	Total	DD no anafórica	DD Anafórica			
				AD	RS	PT	Total
Entrena. 43%	L	490	329	114	41	6	161
	D	255	137	85	21	12	118
	N	170	92	49	21	8	78
	Total	915	558	248	83	26	357
Evaluación 57%	L	674	456	148	49	21	218
	D	352	194	120	22	16	158
	N	191	107	50	24	10	84
	Total	1217	757	318	95	47	460

Tabla 6.2. Distribución de las DDs en los corpus de entrenamiento y evaluación (revisión manual)

ga alguna referencia lingüística producida por una DD:

oSol(6, [anaf([el,alrededores,de,lima],[la,afueras],comun,<),anaf([el, suburbio,de,la,madre,coraje], [el,suburbio,de,la,madre,coraje], comun, <),anaf([la,madre,coraje],[la,madre,coraje,peruano],propio,<)]).

En esta estructura se observa que en la frase 6 del corpus existen 3 DD, la primera de ellas, las afueras hace referencia a la DD los alrededores de Lima. La segunda, el suburbio de la Madre Coraje hace referencia a una anterior aparición de la misma DD y por último la DD la Madre Coraje peruana hace referencia a la DD la Madre Coraje.

A continuación se presentan cada uno de los experimentos realizados hasta obtener las dos versiones finales del algoritmo de resolución de las referencias lingüísticas producidas por las descripciones definidas.

6.1.1 Metodología de evaluación

La metodología utilizada para la evaluación y la obtención de los algoritmos que forman los dos enfoques presentados anteriormente es la siguiente. Se divide cada uno de los corpus en dos partes: entrenamiento y evaluación. Un algoritmo base se aplica al corpus de entrenamiento, los errores y problemas detectados

se solventan con la modificación o adición de nuevas heurísticas y la incorporación de nuevas fuentes de información en diferentes experimentos posteriores. Cada experimento tienen el objetivo de solucionar los problemas encontrados en los experimentos previos para obtener una versión final que será evaluada con el corpus de evaluación, por lo que la evaluación del algoritmo de cada enfoque es independiente.

Los experimentos realizados fueron los siguientes:

1. **Experimentación en el entrenamiento de los algoritmos de resolución de las DDs.**

   - *Experimento 1.* Se aplica un algoritmo inicial sobre cada uno de los fragmentos de los corpus de entrenamiento. Este algoritmo inicial utiliza exclusivamente información léxico-morfológica y sintáctica, y sólo trata y resuelve las referencias producidas por las DDs del mismo núcleo. El algoritmo inicial es una adaptación para la resolución de las DDs del algoritmo de resolución de los pronombres utilizado en nuestro grupo de investigación. Los resultados obtenidos mostraron la necesidad de añadir una nueva heurística al algoritmo inicial para resolver uno de los problemas.

   - *Experimento 2.* La modificación del algoritmo base con la adicción de la nueva heurística y su aplicación sobre los corpus de entrenamiento pone de manifiesto la necesidad de añadir información semántica. La información semántica permite tratar todos los casos de las DDs con el mismo núcleo que no es capaz de resolver ni el algoritmo inicial ni este segundo algoritmo. Además, permite resolver algunos tipos de anáforas indirectas producidas por DDs (núcleos relacionados semánticamente y papel temático).

   - *Experimento 3.* La adquisición de un recurso léxico, como WordNet español, permite añadir información semántica al algoritmo de tratamiento y resolución de las DDs. Esta in-

formación se utiliza a través de la incorporación de nuevas heurísticas relacionadas con la información semántica. En este algoritmo se resuelven las anáforas directas (DDs con el mismo núcleo que su antecedente) y algunos tipos de las anáforas indirectas (relacionadas semánticamente y las de papel temático).

- *Experimento 4.* Después de resolver el problema relacionado con la falta de información semántica, los esfuerzos se centran en reducir el tiempo computacional necesario para el tratamiento y resolución de las DDs en textos grandes. La solución propuesta es la generación automática de una red semántica y la aplicación del mismo conjunto de heurísticas para la adecuada resolución. La aplicación de este mecanismo sobre los corpus de entrenamiento muestra que la generación de la red semántica no influye en la efectividad del algoritmo y sí reduce, considerablemente, el tiempo computacional.

- *Experimento 5.* Al aplicar cualquiera de los algoritmos de los experimentos anteriores sobre los corpus de entrenamiento, en ocasiones, se obtiene como solución correcta un antecedente que en realidad no es el más adecuado, aunque sí pertenece a la misma cadena de correferencia que el antecedente correcto. En estos casos, se toma el resultado propuesto como correcto porque se puede establecer la cadena de correferencia a lo largo del texto (propiedad transitiva). Este experimento consiste en aplicar el mismo algoritmo del experimento 4 (que usa las mismas heurísticas que el experimento 3) pero sin tener en cuenta la propiedad transitiva al obtener los resultados.

- *Experimento 6.* Para obtener más antecedentes correctos de forma directa se cambia el sistema de aplicación de las heurísticas. En lugar de usar un sistema de filtrado, se usa un sistema de pesos que aplica todas las heurísticas a todos los antecedentes seleccionados. Se aplica este nuevo algorit-

mo, que utiliza el sistema ponderado de aplicación de las heurísticas, a los corpus de entrenamiento y tampoco se tiene en cuenta la propiedad transitiva a la hora de evaluar los resultados, para poder compararlos con el experimento 5. Además, se obtienen los resultados utilizando la propiedad transitiva para comparar con el experimento 4.

2. **Experimentación en la evaluación final de los algoritmos propuestos**

   - *Experimento 7.* Una vez obtenidos los algoritmos de cada uno de los enfoques (experimento 3 y 6, respectivamente) se aplican sobre el corpus de evaluación.

Métrica. Las métricas utilizadas por los algoritmos presentados a lo largo de esta memoria han sido la precisión, la cobertura y la medida-F. La cobertura alcanzada por nuestros algoritmos depende de la cobertura alcanzada por el analizador (SUPP), utilizado en este trabajo, ya que es el que identifica los diferentes sintagmas nominales en el texto. Si el analizador parcial SUPP los reconoce entonces el algoritmo de resolución de las referencias producidas por las DDs los tratará.

Se define la cobertura (C) como el cociente entre el número de tratamiento correctos (Cor) y el número total de descripciones definidas existentes (DDe), mientras que la precisión (P) se define como el cociente entre el número de tratamientos correctos (Cor) y el número de descripciones definidas tratadas (DDt).

$$P = \frac{Cor}{DDt} \qquad C = \frac{Cor}{DDe} \tag{6.1}$$

6.2 Experimentación

La experimentación realizada sobre el algoritmo inicial ha dado lugar a la evolución, de dicho algoritmo inicial, hasta obtener los dos algoritmos representativos de cada uno de los dos enfoques presentados. Los experimentos realizados han servido para añadir

o extraer heurísticas al algoritmo inicial, así como para establecer el orden adecuado de aplicación de las heurísticas. Los experimentos mostrados en esta memoria están orientados a la mejora de los resultados obtenidos en cada experimento anterior.

En la experimentación se pueden diferenciar dos fases: (1) Fase de entrenamiento cuyo objetivo es obtener el algoritmo óptimo para cada enfoque y (2) Evaluación final que muestra los resultados obtenidos.

El número de DDs detectadas por el sistema, como se ha mencionado anteriormente, depende exclusivamente del analizador parcial SUPP. Por tanto, el número de DDs tratadas por cada uno de los experimentos es el mismo, la tabla 6.3 muestra la distribución de los diferentes tipos de DDs que el sistema ha detectado en cada uno de los corpus (tanto para entrenamiento como para evaluación). En el corpus de entrenamiento se han detectado un total de 870 DDs de las que 533 son no anafóricas y 337 tienen propiedades referenciales. La distribución de las DDs con propiedades referenciales es: 233 DDs tienen el mismo núcleo que su antecedente (anáfora directa), 80 DDs su núcleo tiene una relación semántica (sinonimia, hiperonimia o hiponimia) con el núcleo de su antecedente y 24 DDs desempeñan un papel temático. En el corpus de evaluación se han detectado un total de 1180 DDs de las que 715 son no anafóricas y 435 tienen propiedades referenciales. De estas 435 DDs con propiedades referenciales, 301 producen una anáfora directa, 90 DDs están relacionas semánticamente con el núcleo de su antecedente y 44 DDs desempeñan un papel temático.

A continuación se muestra detalladamente cada uno de los experimentos realizados que han dado origen a los dos enfoques anteriormente presentados.

6.2.1 Experimento 1: Algoritmo inicial

Este experimento es el punto de partida de la fase de entrenamiento. El algoritmo consiste en un conjunto de heurísticas (H1-H6) que se aplican, en ese orden, para descartar antecedentes hasta alcanzar la solución. Este algoritmo es una adaptación del algo-

Tipo	Corpus	Total	DD no anafórica	DD Anafórica			
				AD	RS	PT	Total
Entrena. 43%	L	466	313	108	40	5	153
	D	243	131	80	20	12	112
	N	161	89	45	20	7	72
	Total	870	533	233	80	24	337
Evaluación 57%	L	634	430	137	48	19	204
	D	335	184	115	20	16	151
	N	181	101	49	22	9	80
	Total	1180	715	301	90	44	435

Tabla 6.3. Distribución de las DDs en los corpus de entrenamiento y evaluación (detectadas por el sistema)

ritmo base utilizado en la resolución de la anáfora pronominal desarrollado por nuestro grupo de investigación para la resolución de las DDs. El algoritmo termina cuando un antecedente es propuesto como solución o no existe ningún antecedente, con lo cuál se clasifica como no anafórica. La información utilizada por este algoritmo es léxico-morfológica y sintáctica. El conjunto de heurísticas inicial que se ha tomado como punto de partida se ha extraído de un primer estudio lingüístico de las características de las DDs del mismo núcleo que producen la anáfora directa en el corpus LEXESP.

El algoritmo esta formado por las siguientes heurísticas (H):

- **H1:** *Concordancia en género y número.* Se prefieren aquellos antecedentes que concuerden en género y número con la DD.

- **H2:** *Mismo núcleo.* Se prefieren aquellos antecedentes cuya raíz o lema del núcleo sea la misma que la raíz del núcleo de la descripción definida.

- **H3:** *Repetición.* Se prefieren como antecedente las apariciones anteriores de la misma DD.

- **H4:** *Indefinidos.* Se prefieren los sintagmas nominales indefinidos a los definidos como antecedentes.

- **H5:** *Frecuentes.* Se prefieren los antecedentes que más aparezcan en el texto.

- **H6:** *Más cercano.* En el caso de que las heurísticas anteriores proporcionen más de un antecedente, esta heurística asegura que el algoritmo sólo propone una única solución.

En el algoritmo de resolución de las DDs se realizan dos tareas principales:

- *Clasificación de las DDs.* Las DDs que no se resuelven se clasifican como no anafóricas, el resto son anafóricas.

- *Resolución de referencias.* Se proporciona un antecedente a una DD anafórica.

Los resultados obtenidos con este *algoritmo inicial* sobre los corpus de entrenamiento se muestran en las tablas 6.4 y 6.5. La tabla 6.4 muestra los resultados de la clasificación de las DDs (anafóricas o no anafóricas), se observa que los resultados medios obtenidos por este algoritmo inicial es de una cobertura del 75% y una precisión del 78.5%. Hay que tener en cuenta que el elevado número de fallos en la identificación de nuevas entidades es debido a que aquellas DDs anafóricas que el algoritmo no encuentra su solución se clasifican como no anafórica. Como no se utiliza información semántica todas las DDs que producen una anáfora indirecta (relacionadas semánticamente y de papel temático) se clasifican como no anafóricas cuando en realidad no lo son. Por esta razón la precisión no llega a alcanzar el 80%.

En cuanto a la resolución de las DDs, la tabla 6.5 muestra los resultados obtenidos para cada uno de los diferentes tipos de descripciones definidas anafóricas. La resolución de la anáfora directa alcanza unos resultados medios de una cobertura del 63.3% y una

	Nueva entidad			
C	A	F	C%	P%
L	300	58	79.8	83.8
E	121	42	71.2	74.2
P	79	37	65.3	68.1
	500	137	75	78.5

Tabla 6.4. Resultados del experimento 1 sobre los corpus de entrenamiento: Clasificación

precisión del 67.3%. Al no utilizar este algoritmo más información que la léxico-morfológica y la sintáctica los resultados obtenidos para resolver la anáfora indirecta es del 0%.

	Mismo núcleo				Relación semántica				Papel temático			
C	A	F	C%	P%	A	F	C%	P%	A	F	C%	P%
L	73	35	64	67.6	0	40	0	0	0	5	0	0
E	53	27	62.3	66.2	0	20	0	0	0	12	0	0
P	31	14	63.3	68.8	0	20	0	0	0	7	0	0
	157	76	63.3	67.3	0	80	0	0	0	24	0	0

Tabla 6.5. Resultados del experimento 1 sobre los corpus de entrenamiento: Resolución

En este experimento se detectaron los siguientes problemas:

- No se tienen en cuenta los modificadores de las DDs.
- No se tratan las anáforas indirectas producidas por las DDs.
- No se utiliza información semántica.

6.2.2 Experimento 2: Resolución de la anáfora directa

En este segundo experimento se intenta solucionar el problema de los modificadores presentado en el experimento anterior, para ello, se introduce la heurística (H4). En este experimento no se añade ningún tipo de información nueva, sólo se usa información léxico-morfológica y sintáctica. El proceso de resolución es igual que el presentado anteriormente.

El algoritmo esta formado por las siguientes heurísticas (H):

- **H1:** *Concordancia en género y número.* Se rechazan los antecedentes que no concuerden en género y número.

- **H2:** *Mismo núcleo.* Se rechazan aquellos antecedentes cuya raíz del núcleo no sea la misma que la raíz del núcleo de la descripción definida.

- **H3:** *Repetición.* Se prefieren como antecedente las apariciones anteriores de la misma descripción definida.

- **H4:** *Incluido.* Si no existe una descripción definida exactamente igual que la expresión anafórica, se busca un antecedente que incluya a la expresión anafórica. Es decir, que tenga los mismos modificadores que ella y alguno más.

- **H5:** *Indefinidos.* Se prefieren como antecedentes aquellos antecedentes que sean sintagmas nominales indefinidos a los definidos.

- **H6:** *Frecuentes.* Se prefieren los antecedentes que más aparezcan en el texto.

- **H7:** *Más cercano.* En el caso de que las heurísticas anteriores proporcionen más de un antecedente, esta heurística asegura que el algoritmo sólo propone una única solución.

Los resultados obtenidos con este *segundo experimento* sobre los corpus de entrenamiento se muestran en las tablas 6.6 y 6.7. En la tarea de clasificación o de identificación de las DDs no anafóricas, se obtienen unos resultados medios de una cobertura del 75.4% y una precisión del 78.9% (tabla 6.6). Como se puede ver, mejora muy levemente los resultados producidos por el experimento 1. Siguen obteniéndose resultados bajos al no tratar tampoco, igual que el experimento 1, la anáfora indirecta. Los resultados alcanzados por la tarea de resolución de las referencias se muestran en la tabla 6.7, se observa que hay un incremento en la resolución de la anáfora directa (mismo núcleo), se alcanza una cobertura del 64.9% y una precisión del 69%. La heurística H4 corrige algunos de estos problemas asignando el antecedente correcto.

	Nueva entidad			
C	A	F	C%	P%
L	302	56	81.3	84.3
E	121	42	71.2	74.2
P	80	36	66.1	68.9
	503	134	75.4	78.9

Tabla 6.6. Resultados del experimento 2 sobre los corpus de entrenamiento: Clasificación

	Mismo núcleo				**Relación semántica**				**Papel temático**			
C	A	F	C%	P%	A	F	C%	P%	A	F	C%	P%
L	76	32	66.7	70.3	0	40	0	0	0	5	0	0
E	53	27	62.4	66.2	0	20	0	0	0	12	0	0
P	32	13	65.3	71.1	0	20	0	0	0	7	0	0
	161	72	64.9	69	0	80	0	0	0	24	0	0

Tabla 6.7. Resultados del experimento 2 sobre los corpus de entrenamiento: Resolución

Los problemas detectados en este experimento son:

- No se contemplan todos los casos de los modificadores de una DD.
- No se utiliza la información semántica.
- No se resuelve la anáfora indirecta producida por las DDs.
- El orden de las preferencias rechazan candidatos correctos.

6.2.3 Experimento 3: Incorporación de información semántica

El objetivo de este experimento es la incorporación de información semántica para el tratamiento de todas las relaciones entre modificadores y de la anáfora indirecta producida por las DDs. Para llevar a cabo este experimento se cuenta con el WordNet español que permite añadir información semántica al proceso de resolución[1]. También se hace una división del algoritmo en dos partes en este tercer experimento. Por un lado, el tratamiento de la anáfora directa producido por las DDs (Muñoz & Palomar, 2000) que, además de la información léxico-morfológica y sintáctica, necesitan de información semántica para resolver posibles relaciones entre los modificadores de la DD y del antecedente. Y por otro lado, el tratamiento de la anáfora indirecta producido por las DDs (Muñoz *et al.*, 2000b) que usan un conjunto de heurísticas distintas a las anteriores basadas en su mayoría en información semántica.

El proceso de resolución de la *anáfora directa* producida por las DDs de este experimento modifica el orden de las heurísticas para evitar que candidatos correctos sean rechazados por heurísticas

[1] La librería para extraer la información semántica del WordNet español fue desarrolladas por Armando Suárez y las DDL que incorpora la información al programa prolog las desarrolló Maximiliano Saíz-Noeda

iniciales y añade la heurística que estudia la relación entre modificadores debida a la incorporación de información semántica. El algoritmo de la anáfora directa está formado por las siguientes heurísticas (H):

- **H1:** *Mismo núcleo.* Se prefieren aquellos antecedentes cuya raíz del núcleo sea la misma que la raíz del núcleo de la descripción definida.

- **H2:** *Repetición.* Se prefieren aquellos antecedentes que tengan los mismos modificadores.

- **H3:** *Incluida.* Se prefieren aquellos antecedentes que contienen todos los modificadores de la DD y alguno más.

- **H4:** *Relación entre modificadores.* Esta heurística prefiere aquellos antecedentes que tengan modificadores relacionados semánticamente con los modificadores de la DD. Pero existen casos en los que los modificadores tienen un relación de antonimia y deben rechazarse.

- **H5:** *Concordancia de género y número.* En el caso que las heurísticas anteriores proporcionen más de un candidato entonces se prefieren aquéllos que tengan el mismo género y el mismo número.

- **H6:** *Frecuencia.* Si una vez aplicada la concordancia de género y número más de un candidato permanece en la lista de candidatos, entonces se prefiere como candidato aquel que más veces haya sido seleccionado como antecedente de descripciones definidas previas.

- **H7:** *Más cercano.* Si más de un antecedente satisface la heurística H5 entonces se prefiere, como antecedente correcto, el antecedente más cercano en el texto.

Por otro lado, el proceso de resolución de la *anáfora indirecta* producido por las DDs se realiza sólo si el algoritmo de las DDs con el mismo núcleo no encuentra una solución a la DD. Este algoritmo está formado por las siguientes heurísticas (Hi):

- **Hi1:** *Relacionados semánticamente.* Se prefieren aquellos antecedentes cuya raíz del núcleo tenga una relación de sinonimia, hiperonimia, hiponimia o de papel temático con la descripción definida.

- **Hi2:** *Sinónimo.* Esta heurística prefiere los antecedentes sinónimos a la descripción definida.

- **Hi3:** *Hiperónimo-hipónimo.* Se prefieren aquellos que tengan una relación de hiperonimia o hiponimia con la descripción definida.

- **Hi4:** *Papel temático.* Se prefieren los que pertenezca a la lista de papel temático.

- **Hi5:** *Mismos modificadores.* Esta heurística prefiere aquellos antecedentes que tengan los mismos modificadores que la descripción definida.

- **Hi6:** *Modificadores relacionados.* Esta heurística prefiere aquellos antecedentes que tengan alguno de sus modificadores relacionados semánticamente con algún modificador de la DD. Si la relación entre los modificadores es de antonimia el candidato se rechaza.

- **Hi7:** *Concordancia de género y número.* En el caso que las heurísticas anteriores proporcionen más de un candidato entonces se prefieren aquéllos que tengan el mismo género y el mismo número.

- **Hi8:** *Frecuencia.* Se elige como antecedente aquel que más veces se haya el seleccionado como antecedente de previas descripciones definidas.

- **Hi9:** *Más cercano.* Si más de un antecedente satisface *Hi8* entonces se elige como candidato correcto el candidato más cercano en el texto a la DD.

Los resultados obtenidos con este experimento sobre los corpus de entrenamiento se muestran en las tablas 6.8 y 6.9. La tabla 6.8 muestra los resultados en la identificación de DD no anafóricas. En esta tabla se observa un incremento tanto en la cobertura como en la precisión, cobertura media del 90.1% y precisión media del 94.2%. La incorporación de información semántica ha permitido resolver las anáforas indirectas y, consecuentemente, no se incrementan los fallos en la tarea de la identificación de las no anafóricas.

	Nueva entidad			
C	A	F	C%	P%
L	302	26	91.8	96.5
E	121	22	88.3	92.4
P	80	23	86.9	89.9
	503	71	90.1	94.2

Tabla 6.8. Resultados del experimento 3 sobre los corpus de entrenamiento: Clasificación

La tabla 6.9 muestra los resultados obtenidos en la tarea de resolución para cada tipo de DD. Para la anáfora directa (DD con el mismo núcleo) se obtiene una cobertura media del 80.2% y una precisión media del 85.4%. En la resolución de las relacionadas semánticamente se alcanza una cobertura media del 61.4% y una precisión media del 63.7%. En la resolución de la anáfora de papel temático se alcanza una cobertura media del 46.1% y

una precisión media del 50%. En general, la tarea de resolución alcanza una cobertura media del 73.4% y una precisión media del 78.2%.

	Mismo núcleo				**Relación semántica**				**Papel temático**			
C	A	F	C%	P%	A	F	C%	P%	A	F	C%	P%
L	94	14	82.5	87	27	13	65.9	67.5	3	2	50	60
E	67	13	78.8	83.7	13	7	61.9	65	7	5	58.3	58.3
P	38	7	77.6	84.4	11	9	52.4	55	2	5	25	28.5
	199	34	80.2	85.4	51	29	61.4	63.7	12	12	46.1	50
	General											
	A				F				C%			P%
	262				75				73.4			78.2

Tabla 6.9. Resultados del experimento 3 sobre los corpus de entrenamiento: Resolución

Toda la experimentación anterior pone de manifiesto algunos problemas, como por ejemplo, que en textos grandes el algoritmo de resolución ARH era excesivamente lento debido a dos factores:

- Intentar solucionar todas las DDs (anafóricas y no anafóricas).
- Muchos antecedentes en el espacio de búsqueda de la solución (todos los sintagmas nominales de todas las oraciones previas del texto).

Para resolver estos problemas se desarrolló un nuevo algoritmo y se llevaron a cabo los siguientes experimentos.

6.2.4 Experimento 4: Construcción automática de una red semántica

Como ya se ha citado, el algoritmo AGSN se desarrolló para resolver, entre otros, los problemas de lentitud encontrados en el

algoritmo ARH. Este segundo enfoque se basa en la construcción automática de una red semántica. Las ventajas que proporciona este algoritmo son dos:

1. Dado que realiza una identificación previa de algunas DDs no anafóricas, no se consume tiempo computacional en resolver algo que no tiene solución.

2. Al agrupar las DDs en nodos de la red semántica en función de su concepto ontológico, se reduce el número de comparaciones a realizar, ya que no se compara con todos los sintagmas nominales que aparecen en todas las oraciones previas, sino que únicamente se compara con todos los sintagmas nominales de todas las oraciones anteriores que pertenecen al mismo nodo de la red semántica.

Este algoritmo no trata la resolución de las DDs de papel temático, por lo que las va a considerar como si se tratase de una DD no anafórica. Debido a este hecho la distribución de las DDs en los corpus de entrenamiento y evaluación quedan tal y como se muestra en la tabla 6.10. Las DDs detectadas por el sistema se distribuyen tal y como se muestran en la tabla 6.11.

Al igual que el resto de experimentos presentados anteriormente, este experimento:

- Identifica las DDs no anafóricas.

- Resuelve las referencias de las DDs anafóricas.

La *identificación* de las descripciones definidas no anafóricas se realiza en dos fases:

1. *Identificación previa.* Antes de aplicar el mecanismo de resolución.

Tipo	Corpus	Total	DD no anafórica	DD anafórica		
				AD	RS	Total
	L	1164	812	262	90	352
	D	607	359	205	43	248
	N	361	217	99	45	144
	Total	2132	1388	566	178	744
Entrena.	L	490	335	114	41	155
43%	D	255	149	85	21	106
	N	170	100	49	21	70
	Total	915	584	248	83	331
Evaluación	L	674	477	148	49	197
57%	D	352	210	120	22	142
	N	191	117	50	24	74
	Total	1217	804	318	95	413

Tabla 6.10. Distribución de las DDs en los corpus sin tener en cuenta las de papel temático (revisión manual)

Tipo	Corpus	Total	DD no anafórica	DD Anafórica		
				AD	RS	Total
Entrena.	L	466	318	108	40	148
43%	D	243	143	80	20	100
	N	161	96	45	20	65
	Total	870	557	233	80	313
Evaluación	L	634	449	137	48	185
57%	D	335	200	115	20	135
	N	181	110	49	22	89
	Total	1180	759	301	90	391

Tabla 6.11. Distribución de las DDs en los corpus sin tener en cuenta las de papel temático (detectadas por el sistema)

2. *Identificación posterior.* Aquéllas que el algoritmo de resolución no es capaz de proporcionar un antecedente.

La identificación previa de las DDs no anafóricas mejora el coste computacional del algoritmo ya que no se intenta resolver una DD no anafórica clasificada previamente. Los resultados obtenidos

en la identificación se muestran en la tabla 6.12. Se observa que sobre los corpus de entrenamiento la identificación previa de las DDs no anafóricas, a través de la red semántica, obtiene una cobertura del 43% con una precisión del 100%. Aunque la cobertura es baja, la precisión es máxima. Es decir, se identifican pocas DDs no anafóricas, pero aquellas que se identifican, se hace de forma correcta y, por tanto, no se propagan errores a la fase de resolución. Por otro lado, la identificación de las DDs no anafóricas por parte del algoritmo de resolución se realiza a través de la aplicación del propio algoritmo de resolución. Si no encuentra ningún antecedente válido, entonces clasifica la DD como no anafórica. En la tabla 6.12 se observa que se alcanza una cobertura del 82.9% y una precisión del 90.2% en la identificación a posteriori. Para realizar el cálculo de las métricas, en esta segunda etapa, se han descontado las DDs no anafóricas tratadas por la identificación a previa. Es decir, el cálculo de las métricas para la identificación a posteriori, por ejemplo en el corpus de entrenamiento del LEXESP, se ha realizado de la siguiente manera. De un total de 335 DDs no anafóricas se han descontado las que se han identificado de forma correcta por la fase de identificación previa (142), por lo que nos queda un total de 193 (335-142) DDs no anafóricas. El algoritmo de resolución ha tratado 176 DDs (165 aciertos y 11 fallos). Para realizar el cálculo de la cobertura se han dividido los aciertos (165) entre las existentes (193), dando una cobertura del 85.5%. La precisión se ha obtenido al dividir los aciertos (165) ente las tratadas (176) obteniendo un 93.7%.

En general, la identificación de las DDs no anafóricas, combinando la identificación previa y la posterior, alcanza una cobertura media del 90.2% y una precisión media del 94.6%.

Después de aplicar el mecanismo AGSN sobre el corpus se detectaron algunos problemas en la clasificación previa de las descripciones definidas no anafóricas. Dichos problemas son:

- El núcleo de la DD no se encuentra en el recurso WordNet español. Las clases de los núcleos que no se suelen encontrar en este recurso son nombres propios y lugares, entre otras. Todas

	Identificación previa				Identificación posterior			
Corpus	A	F	C%	P%	A	F	C%	P%
LEXESP	142	0	42.4	100	165	11	83.5	93.7
Escrituras	64	0	43	100	69	10	81.2	87.3
Periódicos	45	0	45	100	42	9	76.4	82.3
Total	251	0	43	100	276	30	82.9	90.2

Identificación general			
A	F	C%	P%
527	30	90.2	94.6

Tabla 6.12. Identificación de las DDs no anafóricas por el experimento 4

aquellas DDs cuyos núcleos no se encuentran en el WordNet español se almacenan en una misma clase genérica.

- A veces los premodificadores y postmodificadores de los dos sintagmas nominales (descripción definida y su posible antecedente) tienen sentidos opuestos. (oreja derecha y oreja izquierda). Para el mecanismo AGSN es imposible detectar esta clase de DDs no anafóricas porque sólo se usa el núcleo del sintagma para generar la red semántica y las comparaciones con los demás sintagmas nominales pertenecientes a la misma clase las realiza el algoritmo de resolución de las referencias.

Después del estudio del corpus, se observó que las DDs introducidas por un demostrativo, la mayoría de las veces, hacían referencia a una entidad lingüística introducida previamente. Parece obvio, que los demostrativos se usan para hacer referencias a entidades o bien lingüísticas o físicas que están presentes en el discurso o que conocen tanto hablante como oyente. Por tanto, la identificación previa de las DDs no anafóricas realizadas por el algoritmo AGSN, se realiza sólo de las DDs que no están introducidas por un demostrativo ya que estas siempre se resuelven.

El algoritmo de resolución usa el mismo conjunto de heurísticas (excluyendo las relacionadas con el papel temático) que el algoritmo ARH del experimento 3 y como muestra la tabla 6.13 se

obtienen los mismos resultados para la resolución de las anáforas directas (precisión del 85.4%) y para la resolución de las DDs cuyo núcleo está relacionado semánticamente con el núcleo de su antecedente (precisión del 63.7%). En general, los resultados de la resolución de las anáforas producidas por las DDs en este experimento 4 aumenta ligeramente (cobertura del 75.6% y precisión del 79.9%) con respecto a las obtenidas en el experimento 3 (cobertura del 73.4% y precisión del 78.2%) dado que no se tienen en cuenta las de papel temático, que se resuelven en el experimento 3 con una precisión del 50%.

En cuanto a la identificación final de las DDs que introducen una nueva entidad, los resultados también aumenta al no tratarse las de papel temático (precisión del 94.6% frente al 85.4% del experimento 3). Por tanto, la utilización de una red semántica de este estilo mantiene la eficiencia en la tareas de resolución del algoritmo y reduce el tiempo de computación considerablemente.

	Mismo núcleo				**Relación semántica**			
C	A	F	C%	P%	A	F	C%	P%
L	94	14	82.5	87	27	13	65.9	67.5
E	67	13	78.8	83.7	13	7	61.9	65
P	38	7	77.6	84.4	11	9	52.4	55
	199	34	80.2	85.4	51	29	61.4	63.7
	General							
	A		F		C%		P%	
	250		63		75.6		79.9	

Tabla 6.13. Resultados del experimento 4 sobre los corpus de entrenamiento: Resolución

6.2.5 Experimento 5: Sin propiedad transitiva

Los resultados que se muestran en la experimentación anterior se usa la propiedad transitiva la cual permite establecer cadenas de correferencia. Un antecedente propuesto se considera correcto,

si es el correcto o si pertenece a la misma cadena de correferencia que el antecedente correcto. La utilización de un sistema de aplicación de heurísticas en forma de filtro provoca que, en ocasiones, se seleccione como antecedente uno que no sea el más adecuado, aunque sí pertenece a la misma cadena de correferencia. Para solucionar este hecho, se desarrolló un sistema de aplicación de heurísticas basados en pesos (sistema ponderado). Todas las heurísticas se aplican a todos los candidatos y se escoge aquél que tenga mayor peso. Para poder realizar una adecuada comparación de los dos sistemas de aplicación (sistema de filtrado y sistema ponderado) se realizaron dos experimentos más que son: a) la aplicación del mismo algoritmo del experimento 4 sin utilizar la propiedad transitiva, que aquí se presenta, y b) la experimentación con un conjunto de pesos que se presenta en el experimento 6 sin utilizar la propiedad transitiva y con ella, para poder comparar con los resultados del experimento 5 y 4, respectivamente.

	Mismo núcleo				**Relación semántica**			
C	A	F	C%	P%	A	F	C%	P%
L	87	21	76.3	80.6	25	15	61	62.5
E	63	17	74.1	78.8	11	9	52.4	55
P	35	10	71.4	77.8	10	10	47.6	50
	185	48	74.6	79.4	46	34	55.4	57.5
	General							
	A		F		C%		P%	
	231		82		69.8		73.8	

Tabla 6.14. Resultados del experimento 5 sobre los corpus de entrenamiento sin propiedad transitiva: Resolución

Como se ha comentado, en este quinto experimento, se usa el mismo conjunto de heurística que en el experimento 4. Los resultados obtenidos al aplicar el algoritmo a los corpus de entrenamiento, y sin considerar la propiedad transitiva, se muestran en la tabla 6.14. No se muestran los resultados de la tarea de identificación de las DDs no anafóricas ya que la utilización o no de la propiedad transitiva no afecta a esta tarea. La tabla 6.14 muestra

un decremento en los resultados de la resolución de todos los tipos de anáfora. En general, en la tarea de resolución de referencias se obtiene una cobertura media del 69.8% y una precisión media del 73.8% frente a una cobertura del 75.6% y una precisión del 79.9% obtenidos con el experimento 4. En la resolución de la anáfora directa se obtiene una cobertura de 74.6% y una precisión del 79.4% frente a una cobertura del 80.2% y una precisión del 85.4% del experimento 4. En la resolución de las referencias entre DDs con relaciones semánticas entre sus núcleos se obtienen una cobertura del 55.4% y una precisión del 57.5% frente a una cobertura del 61.4% y una precisión del 63.7% del experimento 4.

6.2.6 Experimento 6: Utilización de pesos en las heurísticas

En este experimento se utiliza un sistema ponderado de aplicación de las heurísticas basado en pesos. Todas las heurísticas se aplican a todos los posibles antecedentes, si un antecedente satisface una heurística se le suma a su factor de importancia el peso o valor correspondiente a esa heurística, y así con todas las demás. El antecedente con valor más alto es el escogido como solución a la descripción definida. Se experimentó con varios tipos de pesos, escogiendo finalmente los mostrados en la tabla 6.15. En estos experimentos se observó que la diferencia de pesos utilizada entre las preferencias relacionadas con los modificadores y las preferencias relacionadas con los núcleos tenía mayor importancia que los propios valores en si. Es decir, que otros pesos con diferencias de pesos similares obtienen resultados similares.

Se efectúan dos pruebas sobre el corpus de entrenamiento:

1. *Aplicación sin la propiedad transitiva.* Se aplica el algoritmo sin tener en cuenta la propiedad transitiva a la hora de considerar un antecedente propuesto como correcto o no. Los resultados de este experimento (tabla 6.16), pone de manifiesto al compararlos con los del experimento 5 (sistema de filtro sin propiedad transitiva) que se obtienen mejores resultados. En general, se obtienen una cobertura del 73.1% y una precisión

Heurística	Peso
Mismo núcleo	50
Núcleo sinónimo	40
Núcleo hiper/hipónimo	35
Mismo modificador	10
Modificador sinónimo	9
Modificador hiper/hipónimo	8
Modificador antónimo	−1000
Concordancia género y número	2
Frecuencia	2

Tabla 6.15. Pesos usados

del 77.3% frente a una cobertura del 69.8% y una precisión del 73.8%.

2. *Aplicación con la propiedad transitiva.* Se experimenta con el mismo algoritmo pero teniendo en cuenta la propiedad transitiva. Los resultados de este experimento (tabla 6.17) muestran una cobertura media del 81.5% y una precisión media del 86.7% en la resolución de las DDs con el mismo núcleo y una cobertura media del 63.9% y una precisión media del 66.2% en las relacionadas semánticamente que también son superiores a los resultados obtenidos en el experimento 4.

La primera prueba realizada al experimento 6 mejora los resultados del experimento 5 y la segunda prueba del experimento 6 mejora los resultados del experimento 4. Por esta razón se escoge el sistema de aplicación basado en pesos utilizando la propiedad transitiva como el método representativo del segundo enfoque para aplicarle el corpus de evaluación.

6.2.7 Experimento 7: Evaluación

Los algoritmos resultantes de los experimentos 3 (ARH) y 6 (AGSN) se eligieron como los algoritmos representativos de cada uno de los dos enfoques presentados en esta memoria. Para realizar esta evaluación se ha utiliza el corpus de evaluación que

	Mismo núcleo				Relación semántica			
C	A	F	C%	P%	A	F	C%	P%
L	91	17	79.8	84.3	26	14	63.4	65
E	65	15	76.5	81.3	13	7	61.9	65
P	36	9	73.5	80	11	9	52.4	55
	192	41	77.4	82.4	50	30	60.2	62.5
	General							
	A		**F**		**C%**		**P%**	
	242		71		73.1		77.3	

Tabla 6.16. Resultados del experimento 6 sobre los corpus de entrenamiento sin propiedad transitiva: Resolución

	Mismo núcleo				Relación semántica			
C	A	F	C%	P%	A	F	C%	P%
L	95	13	83.3	88	28	12	68.3	70
D	69	11	81.1	86.3	13	7	61.9	65
N	38	7	77.6	84.4	12	8	57.1	60
	202	31	81.5	86.7	53	27	63.9	66.2
	General							
	A		**F**		**C%**		**P%**	
	255		58		77		81.5	

Tabla 6.17. Resultados del experimento 6 sobre los corpus de entrenamiento con propiedad transitiva: Resolución

es independiente del de entrenamiento y se ha tenido en cuenta la propiedad transitiva.

La identificación de las DDs no anafóricas tienen la particularidad que en el primer enfoque, ARH, se realiza después de aplicar el algoritmo de resolución y en el segundo enfoque se realiza en dos etapas: antes de aplicar el algoritmo de resolución y después de aplicarlo.

La tabla 6.18 muestra los resultados de la identificación de las DDs no anafóricas por el algoritmo ARH. En esta tabla se observa que, de un total de 757 DDs no anafóricas existentes en los corpus de evaluación, el sistema es capaz de reconocer 666 de forma correcta y 49 de forma errónea. Es decir, trata un total de

715 DDs no anafóricas, en la tarea de identificación de las DDs no anafóricas por lo que se obtienen una cobertura media del 88% y una precisión media del 93.1%.

	DD no anafórica				
C	E	A	F	C%	P%
L	456	410	20	89.9	95.3
D	194	166	18	85.6	90.2
N	107	90	11	84.1	89.1
	757	666	49	88	93.1

Tabla 6.18. Identificación de las DDs no anafóricas por ARH en el corpus de evaluación

	Identificación previa					**Identificación posterior**				
Corpus	R	A	F	C%	P%	R	A	F	C%	P%
LEXESP	449	179	0	39.9	100	270	250	20	92.6	92.6
Escrituras	200	81	0	40.5	100	119	101	18	84.9	84.9
Periódicos	110	44	0	40	100	66	55	11	83.3	83.3
Total	759	304	0	40	100	455	406	49	89.2	89.2
	Identificación general									
	E		A		F		C%		P%	
	804		710		49		88.3		93.5	

Tabla 6.19. Identificación de las DDs no anafóricas por AGSN en el corpus de evaluación

La tabla 6.19 muestra los resultados de la identificación las DDs por el algoritmo AGSN en cada una de las fases de las que se compone el proceso. Para cada una de las fases, el cálculo de la cobertura no se realiza sobre el total de DDs no anafóricas existentes en el corpus de evaluación sino sobre el total de DDs no anafóricas a tener en cuenta por cada fase. De un total de 804 DDs no anafóricas, el sistema trata 759 por lo que, en general, la cobertura del sistema es de un 88.3%, identificando de forma correcta 710 de ellas, por lo que alcanza una precisión del 93.5%.

En la primera fase, el sistema debe tratar 759 de ellas e identifica 304, todas de forma correctas y el resto no sabe si es anafórica o no, por lo que alcanza una cobertura del 40% y una precisión del 100%. En la segunda fase, se tratan todas aquellas que la primera no es capaz de tratar (455), identifica de forma correcta 406 y de forma incorrecta 49, esto es, alcanza una precisión y una cobertura del 89.2%.

La tabla 6.20 muestra los resultados obtenidos en la resolución de las referencias producidas por las DDs anafóricas por parte del algoritmo ARH en los corpus de evaluación. En general, se obtiene una cobertura del 71.5% (329/460) y una precisión del 75.6% (329/435). Dentro de los diferentes tipos de DDs anafóricas se observa que los mejores resultados se obtienen en la resolución de la anáfora directa. En esta tarea se obtienen una cobertura del 78.9% (251/318) y una precisión del 83.4% (251/301). En la resolución de las anáforas indirectas los resultados son más bajos que en la anáfora directa debido a la gran cantidad de información semántica necesaria para resolver de forma satisfactoria cada una de las DDs, se obtiene una cobertura del 57.9% (55/95) y una precisión del 61.1% (55/90) en la resolución de las DDs cuyos núcleos están relacionados semánticamente con su antecedente y se obtienen una cobertura del 48.9% (23/47) y una precisión del 52.3% (23/44) en la resolución de las DDs de papel temático.

	Mismo núcleo				**Relación semántica**				**Papel temático**			
C	A	F	C%	P%	A	F	C%	P%	A	F	C%	P%
L	117	20	79	85.4	31	17	63.3	64.6	11	8	52.4	57.9
E	94	21	78.3	81.7	12	8	54.6	60	9	7	56.2	56.2
P	40	9	80	81.6	12	10	50	54.5	3	6	30	33.3
	251	50	78.9	83.4	55	35	57.9	61.1	23	21	48.9	52.3
	Resolución general											
	A				F				C%			P%
	329				106				71.5			75.6

Tabla 6.20. Resultados del algoritmo ARH sobre los corpus de evaluación: Resolución

La tabla 6.21 muestra los resultados obtenidos por el algoritmo AGSN en la resolución de las DDs anafóricas. En general, el sistema obtiene una cobertura del 75.3% (311/413) y una precisión del 79.5% (311/391). En la resolución de la anáfora directa se obtiene una cobertura del 79.9% (254/318) y una precisión del 84.4% (254/301). En la resolución de las DDs cuyo núcleo está relacionado semánticamente con el de su antecedente se obtiene una cobertura del 60% (57/95) y una precisión del 63.3% (57/90).

	Mismo núcleo				**Relación semántica**			
C	A	F	C%	P%	A	F	C%	P%
L	117	20	79	85.4	32	16	65.3	66.6
E	95	20	79.2	82.6	13	7	59.1	65
P	42	7	84	85.7	12	10	50	54.5
	254	47	79.9	84.4	57	33	60	63.3
	Resolución general							
	A		F		C%		P%	
	311		80		75.3		79.5	

Tabla 6.21. Resultados del algoritmo AGSN sobre los corpus de evaluación: Resolución

6.2.8 Comparación de los algoritmos

La experimentación y evaluación realizada de los algoritmos ARH y AGSN muestran buenos resultados (tabla 6.22). Aunque los resultados de AGSN son mejores que los de ARH hay que tener en cuenta que AGSN no trata las DDs anafóricas de papel temático. También, hay que hacer constar que para textos grandes, el mecanismo AGSN es más rápido que ARH debido a la identificación previa de algunas DDs no anafóricas y a que se reduce el número de comparaciones a realizar, aunque se mantiene como espacio de búsqueda de la solución todas las oraciones anteriores.

La comparación de ambos algoritmos (ARH y AGSN) se muestra en la tabla 6.23. El algoritmo AGSN mejora los resultados de

Algoritmo	Nueva entidad					Resolución anáfora				
	E	A	F	C%	P%	E	A	F	C%	P%
ARH	757	666	49	88	93.1	460	329	106	71.5	75.6
AGSN	804	710	49	88.3	93.5	391	311	80	75.3	79.5

Tabla 6.22. Comparativa de ARH y AGSN

ARH debido a que su sistema de aplicación de las heurísticas (sistema basado en la aplicación de pesos) garantiza que todas las heurísticas se aplican a todos los candidatos, mientras que el sistema de filtrado usado por ARH, en ocasiones, rechaza un candidato correcto por no cumplir alguna de las heurísticas previas.

Algoritmo	Mismo núcleo					Relación semántica				
	E	A	F	C%	P%	E	A	F	C%	P%
ARH	318	251	50	78.9	83.4	95	55	35	57.9	61.1
AGSN	318	254	47	79.9	84.4	95	57	33	60	63.3

Tabla 6.23. Comparativa de ARH y AGSN en tareas comunes

La comparación de estos dos algoritmos con los desarrollados por otros autores, estudiados en este trabajo, tiene que ser una comparación indirecta debido a la utilización de diferentes corpus y diferentes lenguajes. Además, la mayoría de los trabajos presentados no hacen una evaluación de forma diferenciada de cada una de las tareas (introducción de una nueva entidad, resolución de la anáfora directa e indirecta), excepto la evaluación realizada por Vieira y Poesio.

La tabla 6.24 presenta los resultados de todos los algoritmos aquí presentados para la tarea de resolución. La primera parte de la tabla 6.24 presenta los resultados obtenidos por los algoritmos basados en aproximaciones lingüísticas y la segunda parte los basadas en aprendizaje automático. Aunque puede parecer que los algoritmos basados en aprendizaje presentan mejores resultados

que los basados en aproximaciones lingüísticas hay que tener en cuenta que los resultados presentados son sobre la resolución de todos los tipo de referencias (pronominal, adjetiva, de descripciones definidas) y que el único algoritmo basado en aprendizaje que exclusivamente trata los sintagmas nominales (Cardie y Wagstaff) presenta resultados de la misma magnitud que los basados en aproximaciones lingüísticas.

Algoritmo	Cobertura (%)	Precisión (%)
Kameyama	46	59
Viera y Poesio	53	76
LaSIE-II	56.1	68.8
LOLITA	48	58.6
OKI	32.9	55.3
ARH	71.5	75.6
AGSN	75.3	79.5
RESOLVE	85.4	87.6
Bean y Riloff	81.8	85.6
Aone y Bennet	69.7	86.7
Cardie y Wagstaff	52.7	54.6

Tabla 6.24. Comparación con otros algoritmos.

Dos de los algoritmos presentados hacen una identificación de las descripciones definidas que introducen una nueva entidad sin realizar una resolución previa. El algoritmo de Bean y Riloff sólo identifica este tipo de descripciones, pero no resuelve ninguna referencia. Los resultados obtenidos por AGSN en la identificación previa es de sólo un 40% de cobertura y una precisión del 100%, pero al finalizar la aplicación del algoritmo de resolución, la cobertura alcanza un valor del 88.3% y la precisión un 93.5%. La tabla 6.25 muestra los resultados en la identificación de las no anafóricas. Se observa que el algoritmo AGSN obtiene los mejores resultados en la identificación de las DDs no anafóricas.

De los algoritmos estudiados, sólo los algoritmos de Vieira y Poesio, Cardie y Wagstaff y los desarrollados en este trabajo, se centran exclusivamente en la resolución de las DDs. Pero, sólo

Algoritmo	Precisión
Viera y Poesio	72%
ARH	93.1%
AGSN	93.5%
Bean y Riloff	85.6%

Tabla 6.25. Identificación de las descripciones definidas no anafóricas.

Vieira y Poesio presentan una evaluación de cada una de las tareas (identificación del tipo, resolución de la anáfora directa e indirecta). Por este motivo nuestros algoritmos se comparan con el desarrollado por ellos, tabla 6.26. En la resolución de las anáforas directas se obtienen resultados similares y es en la resolución de la anáfora indirecta (en este trabajo se diferencia en dos partes: relacionada semánticamente y papel temático) y en la identificación de las descripciones definidas que introducen una nueva entidad donde se aprecia una mayor diferencia. En el caso de las descripciones definidas que producen una anáfora indirecta, la diferencia entre los algoritmos radica en que el algoritmo de Vieira y Poesio:

- no realiza desambiguación del sentido de las palabras
- las relaciones encontradas en WordNet no siempre reflejan la referencia correcta. Los corpus utilizados por estos autores ocasionaban frecuentes desacuerdos entre los anotadores para el etiquetado del antecedente correcto de una anáfora indirecta.

Algoritmo	Nueva entidad	Mismo Núcleo	Anáfora Indirecta
Viera y Poesio	72%	83%	28%
ARH	93.1%	83.4%	52.3%%
AGSN	93.5%	84.4%	63.3%

Tabla 6.26. Comparación de con el algoritmo de Vieira y Poesio.

6.3 Conclusiones

La experimentación realizada sobre los dos algoritmos presentados, ARH y AGSN, muestran la influencia de las diferentes fuentes de información en la resolución de las correferencias. La incorporación de diversas fuentes de información no sólo permite tratar más tipos de descripciones definidas, sino que también permite mejorar la eficiencia de los algoritmos.

Además, la experimentación realizada pone de manifiesto la conveniencia de usar un sistema de aplicación de las heurísticas con pesos en lugar de un sistema de filtro, ya que propone en más ocasiones el antecedente correcto. La comparación con otros autores muestra resultados similares en la resolución de la anáfora directa, pero mejores resultados en la identificación del tipo y en la resolución de la anáfora indirecta.

7. Conclusiones y trabajos futuros

En este trabajo se ha desarrollado un estudio, tratamiento y resolución de la anáfora producida por las descripciones definidas (sintagmas nominales introducidos por un artículo definido o por un demostrativo) cuyo referente sea un sintagma nominal.

Se ha presentado un sistema de resolución de las DDs desde dos enfoques. El primer enfoque realiza en un mismo proceso la identificación de las DDs no anafóricas y la resolución de las DDs anafóricas basado en un conjunto de heurísticas aplicadas en forma de filtro. Por otro lado, en el segundo enfoque, se plantean dos procesos, uno de ellos para la identificación de algunas de las DDs no anafóricas basado en la generación automática de una red semántica y otro para la resolución de las DDs anafóricas basado en un conjunto de heurísticas aplicadas en base a su factor de importancia (sistema de pesos). Consideramos que ambos enfoques por sus características son independientes del dominio y aplicables a textos no restringidos. Las principales fuentes de información utilizadas en ambos enfoques han sido la información léxico-morfológica, sintáctica y semántica.

Las principales contribuciones de este trabajo son:

- Estudio y clasificación de los diferentes tipos de descripciones definidas que se encuentran en un texto escrito en español. A partir del estudio realizado sobre el corpus de entrenamiento, se desarrolló una clasificación de los diferentes tipos de descripciones definidas. Esta clasificación, realizada para el español, es la primera que existe desde un punto de vista de la información necesaria para poder establecer una relación anafórica entre la descripción definida y su posible antecedente. En esta clasifica-

ción se distinguen dos grandes grupos: las descripciones definidas *no anafóricas* y las descripciones definidas *anafóricas*. En las primeras de ellas no es posible establecer ningún tipo de relación anafórica con ningún sintagma nominal que haya aparecido previamente en el texto. Se dice que introducen una nueva entidad. Mientras que en el segundo tipo, descripciones definidas anafóricas, se establece una nueva división en función del tipo de información necesaria para su resolución: *uso anafórico de nivel sintáctico-semántico-textual* y uso *anafórico a nivel pragmático*. En este trabajo se han tratado y resuelto las pertenecientes al nivel sintáctico-semántico-textual.

- Se ha realizado un estudio de las diferentes fuentes de información y su influencia en la resolución de las DDs. Las fuentes de información analizadas han sido la información léxico-morfológica, la sintáctica y la semántica. Con este estudio, se ha llegado a la conclusión que la información léxico-morfológica influye notablemente ya que proporciona información sobre la categoría de la palabra, así como información morfológica (género, número, tiempo). Por otro lado, la información sintáctica nos permite identificar las DDs en el texto y sus posibles antecedente; y por último la influencia de la información semántica es considerable ya que la incorporación de un recurso léxico como WordNet español nos ha permitido la resolución de la gran número de DDs.

- Desarrollo de dos enfoques diferentes en la resolución de las DDs. El primer enfoque realiza la identificación de las DDs no anafóricas y la resolución de las DDs anafóricas en un único proceso mientras que el segundo enfoque realiza dos procesos. El primer proceso realiza la identificación de parte de las DDs no anafóricas y la generación automática de una red semántica donde se almacena cada sintagma nominal que aparece en el texto en su clase correspondiente y el segundo proceso resuelve las DDs anafóricas y clasifica el resto de DDs no anafóricas que el primer proceso no fue capaz de clasificar.

- El primer enfoque se basa en la aplicación de un conjunto de heurísticas aplicadas en forma de filtro para obtener el antecedente adecuado a la DD. Este conjunto de heurísticas utiliza los diferentes tipos de información analizados en este trabajo y se fundamenta en las características observadas en las DDs que aparecen en el corpus de entrenamiento. Todas aquellas DDs que el sistema de resolución no es capaz de encontrar una solución el sistema las clasificas como no anafóricas. Por tanto en un único proceso el sistema realiza las dos tareas, identificación de las DDs no anafóricas y resolución de las anafóricas.

- El segundo enfoque realiza las dos tareas que realiza el primer enfoque en dos procesos. El primero de los procesos realiza una clasificación previa a la aplicación del algoritmo de resolución de parte de las DDs no anafóricas y un segundo proceso que realiza la resolución de las referencias de las DDs anafóricas y la clasificación de las DDs no anafóricas que no han podido ser resueltas por el primer proceso. La clasificación previa de las DDs no anafóricas se basa en la construcción de una red semántica utilizando como base la ontología del WordNet español. Una DD se clasifica como no anafórica si no existe un sintagma nominal previo en el texto que pertenezca a la misma clase de la red. Si existe un sintagma nominal que pertenece a la misma clase no puede clasificarse como tal y habrá que aplicarle el método de resolución. El método de resolución empleado utiliza un conjunto de heurísticas en función a su factor de importancia. Es decir, cada una de las heurísticas tiene asociada un peso que es añadido al factor de importancia general del candidato si la satisface, y finalmente se escoge el candidato con mayor factor de importancia.

- Además, la utilización de esta red semántica nos ha permitido reducir el número de comparaciones a realizar para encontrar el antecedente correcto sin reducir el espacio de búsqueda de la solución. Ambos enfoques utilizan todas las oraciones anteriores, desde el inicio del texto, para buscar la solución correcta. En el primer enfoque la lista de candidatos a antecedentes está

formada por todos los sintagmas nominales previos en el texto, con lo que para textos grandes el número de comparaciones es muy elevado. En el segundo enfoque la lista de candidatos a antecedentes está formada exclusivamente por los sintagmas nominales previos que pertenecen al mismo concepto base (misma clase) , con lo que el número de comparaciones se reduce considerablemente.

- Los algoritmos han sido entrenados sobre un corpus de entrenamiento formado por tres fragmentos de diferentes corpus con el objetivo de obtener de una forma experimental las diferentes heurísticas a aplicar y el orden de las mismas. Se han realizado un total de seis experimentos sobre ambos algoritmos, en cada uno de ellos se han ido incorporando diferentes tipos de información lingüística (léxico-morfológica, sintáctica y semántica) a través de la aplicación de heurísticas apropiadas, las cuales han permitido, en cada experimento, poder tratar y resolver nuevos tipos de descripciones definidas. El corpus de entrenamiento utilizado (47%) son fragmentos diferentes de los usados para la evaluación final. A la hora de considerar si un antecedente propuesto es correcto o no, se ha tenido en cuenta la propiedad transitiva de las cadenas de correferencia. Por lo que un antecedente propuesto es considerado como correcto si pertenece a la misma cadena de correferencia que el correcto. Es decir, si el antecedente propuesto a su vez hace referencia al correcto. Esta propiedad se ha tenido en cuenta en todos los experimentos salvo en los experimento 5 y 6 para mostrar los beneficios del sistema de aplicación de las heurísticas utilizando pesos en lugar del sistema de filtro.

- La evaluación del sistema se realizó sobre tres fragmentos independientes del corpus de evaluación (53%). En la tarea de identificación de las descripciones definidas no anafóricas, el primer enfoque alcanza una cobertura del 88% y una precisión del 93.1%, mientras que el segundo enfoque obtiene unos resultados de una cobertura del 88.3% y una precisión del 93.5%. En la tarea de resolución de las referencias, el primer enfoque obtiene

una cobertura del 71.5% y una precisión del 75.6%, mientras que el segundo enfoque alcanza una cobertura del 75.3% y una precisión del 79.5%.

- La evaluación llevada a cabo ha permitido realizar una comparación indirecta con los resultados obtenidos por otros autores. Dicha comparación ha sido indirecta debido a que cada uno de los diferentes autores han utilizado un corpus de entrenamiento y de evaluación distinto. La mayoría de los trabajos se ha realizado sobre el inglés y nuestro sistema trabaja en español. Además, cada uno de los autores han hecho la evaluación de sus sistemas teniendo en cuenta diferentes tareas realizar. Se ha hecho una comparación con los resultados mostrados por Vieira y Poesio debido a que sus trabajos presentan una evaluación de las diferentes tareas (identificación de las DDs no anafóricas, resolución de la anáfora directa e indirecta) por separado, lo cual permite realizar la comparación indirecta con nuestros algoritmos. En la resolución de la anáfora indirecta (mismo núcleo) se obtienen resultados similares, pero es en la identificación de las no anafóricas y en la resolución de la anáfora indirecta donde se encuentran sustanciales mejoras.

- La aplicación de este módulo a sistemas de extracción de información incrementa la efectividad de dichos sistemas. Los textos utilizados por los sistemas de extracción de información no suelen tener un formato fijo, aunque trabajen en un dominio restringido, y la información a extraer aparece de manera implícita (dispersa) en el texto. Es decir, que no es normal encontrar en los textos a tratar la información a extraer en el formato exacto. La resolución de las correferencias, en general, y la de las descripciones definidas, en particular, ayudan a establecer las relaciones necesarias entre todas las partes del texto, de forma que la información implícita se transforma en información explícita que faciliten el relleno de las plantillas correspondientes.

7.1 Trabajos futuros

Con este trabajo no se pretende hacer un punto y final, sino todo lo contrario, existen líneas de investigación abiertas que pueden complementarlo y mejorarlo. Algunas de estas nuevas líneas de investigación son:

- Resolver otros tipos de referencias a parte del tipo *identidad*, como son: *parte de*, *conjunto-subconjunto*, *conjunto-miembro*. Para ello hay que añadir otros tipos de relaciones semánticas como son las relaciones de meronimias u holonimias. Estos tipos de relaciones se extraen del recurso WordNet español.

- Añadir información pragmática para poder resolver las descripciones definidas del nivel pragmático de la clasificación presentada en esta memoria.

- Resolución de expresiones temporales desde el punto de vista de proporcionar una fecha exacta. Se continuarán los trabajos desarrollados en esta línea por Saquete y Martínez-Barco (2000), tanto para la resolución de expresiones referenciales finitas (una fecha exacta) como la resolución de expresiones temporales cuya referencia sea un rango de fechas.

- Aplicación de estos algoritmos a otra clase de discursos como pueden ser los diálogos. Para ello, habrá que estudiar la influencia de la estructura del discurso, los tópicos de propio discurso y el espacio de accesibilidad en la resolución de las descripciones definidas.

- En la actualidad se está desarrollando un sistema completo de procesamiento de lenguaje natural orientado a la resolución de la anáfora producida por diversos tipos de expresiones anafóricas (pronombres, nombres) sobre diferentes tipos de textos (textos dialogados y textos no dialogados). Este sistema se ha denominado PHORA (Palomar *et al.*, 2001) y utiliza el formalismo gramatical de las gramáticas léxico-funcionales (LFG)

en lugar de las gramáticas de unificación de huecos (SUG).

- Realización de un etiquetado estándar de la información de correferencialidad de cada corpus utilizando etiquetas SGML. Estas etiquetas nos permite incorpora información para cada sintagma presente en una cadena de correferencia cual es el número de sintagma en el texto, la identificación de la cadena de correferencia a la que pertenece a través de un número único, número del sintagma al que hace referencia, tipo de relación con su antecedente (directa, relacionada semánticamente y papel temático) y toda la información necesaria que ayude a realizar una exhaustiva evaluación automática.

7.2 Producción científica

Este trabajo de investigación ha tenido como consecuencia la publicación de diferentes trabajos relacionados con la resolución de la anáfora y con diversas tareas relacionadas con la extracción de información. A continuación se muestra cada una de estas contribuciones:

1. Publicaciones en revistas nacionales e internacionales

   - Palomar, M.; Ferrández, A.; Moreno, L.; Martínez-Barco, P.; Peral, J.; Saiz-Noeda, M.; Muñoz, R. An Algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics*. 2000 (pendiente de aceptación).
   - Bia, A.; Muñoz, R. Aplicación de técnicas de extracción de información a Bibliotecas Digitales. *Procesamiento del Lenguaje Natural* **26**:207-214. 2000. ISSN:1135-5948.
   - Muñoz, R.; Martínez-Barco, P. y Ferrández, A. Método para la resolución de correferencias de sintagmas nominales definidos incluyendo alias y acrónimos en el sistema de extracción de información EXIT. *Procesamiento del Lenguaje Natural* **25**:143-149. 1999. ISSN:1135-5948.

- Ferrández, A.; Palomar, M.; Martínez-Barco, P.; Peral, J.; Muñoz, R.; Saiz-Noeda, M. Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística *Procesamiento del Lenguaje Natural* **25**:217-218, 1999. ISSN:1135-5948.
- Llopis, F.; Muñoz, R.; Suárez, A. and Montoyo, A. EXIT: Propuesta de un sistema de extracción de información de textos notariales. *Novática* **133**:26-30. 1998. ISSN 0211-2124.
- Muñoz, R.; Montoyo, A.; Llopis, F.; Suárez, A. Reconocimiento de Entidades en el sistema EXIT. *Procesamiento del Lenguaje Natural* **23**: 47-53. 1998. ISSN:1135-5948.

2. Comunicaciones a congresos internacionales

- Muñoz, R.; Palomar, M. Semantic-driven Algorithm for Definite Description Resolution. In *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics joint with the 10th Meeting of the European Chapter of the Association for Computational Linguistics, ACL-EACL'2001.* Toluse (France). July 2001. (pendiente de aceptación).
- Palomar, M. and Muñoz, R. Definite Descriptions in an Information Extraction System. In: Maria Carolina Monard and Jaime Simao Eds. *The International Joint Conference IBERAMIA'2000-SBIA'2000.* Lecture Notes in Artificial Intelligence **1952**:320-328. Sao Paulo (Brazil). November 2000. ISBN: 3-540-41276-X.
- Muñoz, R., Saiz-Noeda, M., Suárez, A., Palomar, M. Semantic Approach to Bridging Reference Resolution In *Proc of International Conference Machine Translation and Multilingual Applications in the New Millennium. MT2000.* Exeter (UK). 17-1 a 17-8 November 2000.
- Muñoz, R. and Palomar, M. Procesing of Spanish Definite Descriptions with the Same Head. In D.N. Christodoulakis Ed. *Natural Language Processing - NLP 2000.* Lecture No-

tes in Artificial Intelligence **1835**:212-220. Patras (Greece). June, 2000. ISBN: 3-540-67605-8.

- Muñoz, R.; Palomar, M. y Ferrández, A. Procesing of Spanish Definite Descriptions. In O. Cairo, L.E. Sucar and F.J. Cantú Eds. *MICAI 2000: Advances in Artificial Intelligence.* Lecture Notes in Artificial Intelligence **1793**:526-537. Acapulco (México). Abril, 2000. ISBN: 3-540-67354-7.
- Muñoz, R. and Ferrández, A. Definite Description Resolution in Spanish In *Proc. of International Conference on Artificial and Computational Intelligence For Decision, Control and Automation in Engineering and Industrial Applications. ACIDCA2000.* Monastir (Túnez). pp. 140-145. Marzo 2000.
- Palomar, M.; Ferrández, A.;Moreno, L.; Saiz-Noeda, M.; Muñoz, R.; Martínez-Barco, P.; Peral, J.; Navarro, B. A Robust Partial Parsing Strategy based on the Slot Unification Grammars. In *Proc. of The Sixth Conference on Natural Language Processing, TALN99.* Corsica, France. pp. 263-272. July 1999.
- Peral, J.; Martínez-Barco, P.; Muñoz, R.; Ferrández, A; Moreno, L.; Palomar, M. Una técnica de análisis parcial sobre textos no restringidos (SUPP) aplicada a un Sistema de Extracción de Información (EXIT). In *Proc. of VI Simposio Internacional de Comunicación Social.* Santiago de Cuba (Cuba). pp 662-669. January 1999. ISBN: 959-11-02050-X. (mención especial).
- Muñoz, R.; Palomar, M. Sentence boundary and named entity recognition in EXIT system: information extraction system of notarial text. In *Proc. of IV International Meeting on Artificial Intelligence and Emerging Technologies in Accounting, Finance and Tax.* Huelva, Spain. pp. 129-142. December 1998.

3. Capítulos de libro

- Muñoz, R.; Palomar, M. Sentence boundary and named entity recognition in EXIT system: information extraction system of notarial text. In Bonson, E. and Vasarhelyi, M. Eds.

Emerging Technologies in Accounting and Finance. pp. 129-142. 1999. ISBN:84-921883-2-4.

4. Informes técnicos

- LLopis, F.; Muñoz, R.; Suárez, A.; Montoyo, A.; Palomar, M.; Ferrández, A.; Martínez-Barco, P.; Peral, J.; Romero, R. and Saiz, M. Propuesta de un sistema de extracción de información de textos notariales: EXIT. *Report Interno.* Departamento de Lenguajes y sistemas Informáticos. Universidad de Alicante.

Otros trabajos relacionados que han complementado los trabajos realizados para esta memoria son:

1. Congresos internacionales

- Palomar, M.; Saiz-Noeda, M.; Muñoz, R.; Suárez, A.; Martínez-Barco, P.; Montoyo, A. PHORA: NLP System for Spanish. In: Gelbukh, A. Ed. *2nd International conference on Intelligent Text Processing and Computational Linguistics CICLing-2001.* Lecture Notes in Computer Science **2004**. México D.C. (México). pp. 126-139 February 2001.
- Palomar, M.; Saiz-Noeda, M.; Muñoz, R.; Suárez, A.; Martínez-Barco, P. PHORA: A system to solve the Anaphora in Spanish. In *Proc. of Third International Conference on Discourse Anaphora and Anaphor Resolution, DAARC2000.* Lancaster (UK). pp. 206-211. November, 2000.
- Martínez-Barco, P.; Muñoz, R.; Azzam, S.; Palomar, M.; Ferrández, A. Evaluation of pronoun resolution algorithm for Spanish dialogues. In *Proc. of International Conference Venezia per il Trattamento Automatico delle Lingue, VEXTAL.* Venice, Italy. pp. 325-332. November 1999. ISBN: 88-8098-112-9.

Referencias

Abney, S. 1997. *Tagging and Partial Parsing.* Kluwer Academic Publishers.

Alcaraz, E., & Martínez, M. A. 1997. *Diccionario de lingüística moderna.* Editorial Ariel S.A.

Allen, J. 1995. *Natural Language Understanding.* The Benjamin / Cummings Publishing Company, Inc.

Andersen, P.M., Hayes, P.J., Huettner, A.K., Nirenburg, I.B., Schmandt, L.M., & Weinstein, S.P. 1992. Automatic extraction of facts from press releases to generate news stories. *Pages 170–177 of: Proceedings of the Third Conference on Applied Natural Language Processing.*

Aone, C., & Bennet, S.W. 1995a. Automated acquisition of anaphora resolution strategies. *Pages 1–7 of: Proceedings of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.*

Aone, C., & Bennet, S.W. 1995b. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. *Pages 122–129 of: Proceedings of the 33rd Annual Meeting of the ACL.*

Aone, C., & Bennet, S.W. 1996. Applying Machine Learning to Anaphora Resolution. *Connectionist, statistical and simbolic approaches to learning for Natural Language Processing*, 302–314.

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. *In: Proceedings of*

First International Conference on Language Resources and Evaluation (LREC'98).

AVENTINUS system. 1998. *AVENTINUS: advanced information system for multilingual drug enforcement.* `http://www2.echo.lu/langeng/projects/aventinus`. Visitada 2-12-98.

Bean, D. L., & Riloff, E. 1999. Corpus-based Identification of Non-Anaphoric Noun Phrases. *Pages 373–380 of: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic.*

Bustos, E. 1986. *Pragmática del español.* Madrid: UNED. Ministerio de Educación y Ciencia.

Carbonell, J.G., & Brown, R.D. 1988. Anaphora resolution: a multi-strategy approach. *Pages 96–101 of: Proceedings of 12th International Conference on Computational Linguistics (COLING'88).*

Cardie, C., & Pierce, D. 1998. Error-Driven Pruning of Tree-Bank Grammars for Base Noun Phrase Identification. *Pages 218–224 of:* ACL (ed), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Cardie, C., & Wagstaff, K. 1999. Noun Phrase coreference as Clustering. *Pages 82–89 of: Proceeding of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Carter, D. M. 1987. *Interpreting Anaphors in Natural Language Texts.* Chester, UK: Ellis Horwood.

Castellón, I., Civit, M., & Atserias, J. 1998. Syntactic Parsing of Unrestricted Spanish Text. *In: Proceedings of First International Conference on Language Resources and Evaluation (LREC'98).*

Christopherson, P. 1939. *The Articles: A study of their theory and use in English.* Copenhagen: E. Munksgaard.

Ciravenga, F., Campia, P., & Colognese, A. 1992. Knowledge extraction from texts by SINTESI. *Pages 1244–1248 of: Proceedings of COLING.*

Clark, H. H. 1977. Bridging. *Pages 411–420 of:* Johnson-Laird, P., & Wason, P (eds), *Thinking: readings in cognitive science.* London New York: Cambridge University Press.

COBALT system. 1998. *COBALT: Construction, augmentation and use of knowledge bases from natural language documents.* `http://www2.echo.lu/langeng/projects/cobalt`. Visitada 2-12-98.

Cowie, J. R. 1983. Automatic analysis of descriptive texts. *Pages 117–123 of: ACL Proceedings, Conference on Applied Natural Language Processing.*

Cowie, J. R., & Lehnert, W. 1996. Information Extraction. *Comunications of ACM*, **39**(1), 81–91.

CRATER, Proyecto. 1994. *Corpus Resources and Terminology ExtRaction.* MLAP-93/20. `http://www.lllf.uam.es/proyectos/crater.html`.

Crawdford, M. 1997. *Information Extraction.* `http://www.dcs.shef.ac.uk/research/groups/extraction`. Visitada el 10-12-1997.

Cunningham, H. 1997. *Information Extraction a User Guide.* Tech. rept. Research memo CS-97-02. Institute for Language, Speech and Hearing (ILASH), and Department of Computer Science, University of Sheffield. UK.

Dagan, I., & Itai, A. 1990. Automatic processing of large corpora for the resolution of anaphora references. *In: Proceedings of 13th International Conference on Computational Linguistics (COLING'90).*

Dagan, I., & Itai, A. 1991. A statistical filter for resolving pronoun references. *Artificial Intelligence and Computer Vision*, 125–135.

Donnellan, K. 1996. Reference and definite description. *Philosophical Review*, **75**, 281–304.

Donnellan, K. 1998. *Definite descriptions: A Reader.* Gary ostertag edn. Chap. Reference and definite description, pages 173–194.

ECRAN project. 1998. *ECRAN: extraction of content: research at near-market.* `http://www2.echo.lu/langeng/projects/ecran`. Visitada 2-12-98.

Evans, R., Gaizauskas, R., Cahill, L., Walker, J., Richardson, J., & Dixon, A. 1995. POETIC: a system for gathering and disseminating traffic information. *Journal of Natural Language Engineering*, **1**(4), 363–387.

FACILE project. 1998. *FACILE: Fast and Accurate Categorisation of Information by Language Engineering.* `http://www2.echo. lu/langeng/projects/facile`. Visitada 2-12-98.

Fellbaun, C. 1998. *WordNet, an electronic lexical database.* MIT Press.

Ferrández, A. 1998. *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gram´áticas de unificación de huecos.* Ph.D. thesis, Universidad de Alicante.

Ferrández, A., & Peral, J. 2000. A Computational Approach to Zero-propouns in Spanish. *Pages 166–172 of: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistic (ACL).*

Ferrández, A., Palomar, M., & Moreno, L. 1997. Slot Unification Grammar. *Pages 523–532 of: Proceedings of the Joint Conference on Declarative Programming. APPIA-GULP-PRODE.*

Ferrández, A., Palomar, M., & Moreno, L. 1998a. A computational approach to pronominal anaphora, one-anaphora and surface count anaphora. *Pages 117–128 of:* Botley, Simon, & Eds, Tony McEnery (eds), *Proceedings of Second Colloquium on Discurse Anaphora and Anaphor Resolution (DAARC2).*

Ferrández, A., Palomar, M., & Moreno, L. 1998b. Anaphora resolution in unrestricted texts with partial parsing. *Pages 385–391 of:* ACL (ed), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Ferrández, A., Palomar, M., & Moreno, L. 1999. An empirical approach to Spanish anaphora resolution. *Machine Traslation*, **14**(2-3).

Frege, G. 1892. Sobre sentido y referencia. *Pages 24–45 of:* Villanueva, Luis Ml. Valdés (ed), *La búsqueda del significado:*

Lecturas de filosofía del lenguaje. Tecnos.

Fukumoto, J., Masui, F., Shimohata, M., & Sasaki, M. 1998. *Oki Electric Industry: Description of the Oki System as used for MUC-7*. `http://www.muc.saic.com/proceedings/`.

Gaizauskas, R., & Wilks, Y. 1998. Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, **54**(1), 70–105.

Garigliano, R., Urbanowicz, A., & Nettleton, D. J. 1998. University of Durham: Description of the LOLITA System as used in MUC-7. *In:* (Morgan Kaufman Publishers, 1998).

Ge, N., Hale, J., & Charniak, E. 1998. A statistical approach to anaphora resolution. *Pages 161–170 of:* Charniak, E. (ed), *Proceedings of 6th WorkShop on Very Large Corpora*.

Goldfarb, C.F. 1990. *The SGML Handbook*. Oxford University Press.

Gonzalo, J., Verdejo, F., Chugur, I., López, F., & nas, A. Pe1998. Extracción de relaciones seánticas entre nombres y verbos en EuroWordNet. *Procesamiento del Lenguaje Natural*, **23**, 97–103.

Grishman, R. 1997. Information Extraction: Techniques and Challanges. *Lecture Notes in Computer Science*, **1299**, 10–27.

Hawkins, J. A. 1978. *Definiteness and indefiniteness*. Atlantic Highlands, New Jersey: Humanities Press.

Hirst, G. 1981. *Anaphora in Natural Language Understanding*. Berlin: Springer-Verlag.

Hobbs, J. 1978. Resolving pronoun references. *Lingua*, **44**, 311–338.

Hobbs, J. R., Appelt, D. E., Bear, J., Israel, D., Kameyama, M., Stickel, M., & Tyson, M. 1996. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. *In:* Roche, E., & Schabes, Y. (eds), *Finite State Devices for natural Language Processing*. Cambridge, Massachussetts: MIT Press.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., & Mitchell, B. 1998. University of Sheffield: Description of the LaSIE-II System as used for MUC-7. *In:* (Morgan Kaufman Publishers, 1998).

Kameyama, M. 1997. Recognizing Referential Links: An Information Extraction Perspective. *In:* (Mitkov, R. and Boguraev, B., 1997).

Kempson, R. M. 1975. *Presupposition and the delimitatation of Semantics.* Crambidge University Press.

Lappin, S., & Leass, H.J. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

Lehnert, W. 1997. *Information Extraction.* `http://www.cs.buffalo.edu/g̃herrera/inf-ext.html`. Visitada el 10-12-1997.

Leonetti, M. 1990. *El artículo y la referencia. Colección Gramática del español.* Taurus Universitaria.

Llopis, F., Muñoz, R., Suárez, A., & Montoyo, A. 1998. EXIT: Propuesta de un sistema de extracción de información de textos notariales. *Novática*, **133**, 26–30.

Marcus, Mitchell P., Santorini, Beatrice, & Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational linguistics*, **19**, 313–330.

Martí, M.A., Rodríguez, H., & Serrano, J. 1998. *Declaración de categorías morfosintácticas.* Proyecto ITEM. Doc. nº2. `http://sensei.ieec.uned.es/item`.

Martínez-Barco, P. 1998. *Aportaciones a la resolución de extraposición de elementos en lenguaje natural utilizando técnicas incrementales.* Trabajo de Investigación, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

Martínez-Barco, P., & Palomar, M. 2000. Dialogue structure influence over anaphora resolution. *In:* Cairo, O., Sucar, E. L., & Cantu, F. J. (eds), *Proceeding of Mexican International Conference on Artificial Intelligence.* Lectures Notes in Artificial Intelligence, vol. 1793. Acapulco, Mexico: Springer-Verlag.

Martínez-Barco, P., Peral, J., Ferrández, A., Moreno, L., & Palomar, M. 1998. Analizador Parcial SUPP. *Pages 329–341 of: Proceedings of VI biennial Iberoamerican Conference on Artificial Intelligence, IBERAMIA'98.*

Martínez-Barco, P., Muñoz, R., Azzam, S., Palomar, M., & Ferrández, A. 1999. Evaluation of pronoun resolution algorithm for Spanish dialogues. *Pages 325–332 of: Proceedings of the Venezia per il Trattamento Automatico delle Lingue (VEXTAL'99).*

McCarthy, J. F., & Lehnert, W. G. 1995. Using Decision Trees for Coreference Resolution. *Pages 1050–1055 of: Proceedings of the Fourteenth International Conference on Artificial Intelligence.*

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. 1993. Five Papers on WordNet. *International journal of lexicography*, **Special Issue**(3).

Mitkov, R. 1995. An uncertainty reasoning approach to anaphora resolution. *In: Proceedings of the Natural Language Pacific Rim Symposium.*

Mitkov, R. 1997. Factors in anaphora resolution: they are not the only things the matter. A case study based on two different approaches. *Pages 14–21 of:* ACL (ed), *Proceedings of ACL / EACL.*

Mitkov, R. 1998. Robust pronoun resolution with limited knowledge. *Pages 869–875 of:* ACL (ed), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).*

Mitkov, R., & Stys, M. 1995. 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, RANLP95.*

Mitkov, R. and Boguraev, B. (ed). 1997. *Proceedings of ACL / EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.*

Montoyo, A., & Palomar, M. 2000. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. *Pages 103–107 of: Proc. 11th International Workshop on Database and Expert Systems Applications (DEXA 2000).* Greenwich, London, UK: IEEE Computer Society.

Montoyo, A., Palomar, M., & Rigau, G. 2001. WordNet Enrichment with Classification Systems. *In:* ACL (ed), *Proceedings of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations.*

Moreno, L., Palomar, M., Molina, A., & A.Ferrández. 1999. *Introducción al Reconocimiento del Lenguaje Natural.* Alicante: Servicio de Publicaciones de la Universidad de Alicante.

Morgan Kaufman Publishers (ed). 1998. *Proceedings of Seventh Message Understandig Conference.*

MUC-7. 1998. *SAIC:Science Aplications International Corporation.* `http://www.muc.saic.com/index.html`. Visitada 27-11-98.

Muñoz, R. 1998. *Propuesta de un método para la resolución del límite de la frase y el reconocimiento de entidades mediante SUG aplicado en el sistema de extracción de información EXIT.* Trabajo de Investigación, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

Muñoz, R., & Ferrández, A. 2000. Definite Descriptions Resolution in Spanish. *Pages 140–145 of: Proceeding of ACIDCA'2000: Corpora and Natural Language Processing.*

Muñoz, R., & Palomar, M. 1999. *Emerging Technologies in Accounting and Finance.* E. Bonsón and M. Vasarhelyi edn. Chap. Sentence Boundary and Named Entity Recognition in EXIT System: Information Extraction System of Notarial Texts, pages 129–142.

Muñoz, R., & Palomar, M. 2000. Processing of Spanish Definite Descriptions with the Same Head. *Pages 212–220 of:* Christodoulakis, DimitrisÑ. (ed), *Proceeding of NLP2000: Filling the gap between theory and practice.* Lectures Notes in Artificial Intelligence, vol. 1835. Patras, Greece: Springer-Verlag.

Muñoz, R., & Palomar, M. 2001. Semantic-driven Algortithm for Definite Description Resolution. *In:* ACL (ed), *Proceedings of ACL / EACL'01.* (pendiente de aceptación).

Muñoz, R., Montoyo, A., Llopis, F., & Suárez, A. 1998. Reconocimiento de entidades en el sistema EXIT. *Procesamiento del Lenguaje Natural*, **23**, 47–53.

Muñoz, R., Palomar, M., & Ferrández, A. 2000a. Processing of Spanish Definite Descriptions. *Pages 526–537 of:* Cairo, O., Sucar, E. L., & Cantu, F. J. (eds), *Proceeding of Mexican International Conference on Artificial Intelligence.* Lectures Notes in Artificial Intelligence, vol. 1793. Acapulco, Mexico: Springer-Verlag.

Muñoz, R., Saiz-Noeda, M., Suárez, A., & Palomar, M. 2000b. Semantic approach to bridging reference resolution. *Pages 17.1–17.8 of: Proc. 11th International Conference on Machine Translation and Multilingual Applicantions in the New Millenium (MT'2000).*

Palomar, M. 1996. *Aportaciones a la resolución de la elipsis en lenguaje natural utilizando técnicas incrementales.* Ph.D. thesis, Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.

Palomar, M., & Muñoz, R. 2000. Definite Descriptions in Information Extraction System. *In:* Monmard, Carolina, & Simao, Jaime (eds), *Proceeding of Iberamia-SBIA'2000.* Lectures Notes in Artificial Intelligence. Atibaia, São Paulo, Brazil: Springer-Verlag.

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: Proceeding of 6e Conférence annuelle sur le Traitement Automatique des Langues Naturelles. TALN'99.*

Palomar, M., Saiz-Noeda, M., Muñoz, R., Suárez, A., Martínez-Barco, P., & Montoyo, A. 2001. PHORA: A NLP aystem for Spanish. *Pages 126–139 of:* Gelbukh, A. (ed), *Computational Linguistics and Intelligent Text Processing.* Lecture Notes in Computer Science, vol. 2004. Mexico City: Springer-Verlag.

Peral, J. 1998. *Adaptación de las gramáticas de unificación de huecos para tratamiento del fenómeno lingüístico de la extraposición.* Trabajo de Investigación, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

Pla, F. 2000. *Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos.* Tesis doctoral, Departamen-

to de Sistemas Informáticos y Computación. Universidad de Politécnica de Valencia.

Pla, F., Molina, A., & Prieto, N. 2000. Tagging and Chunking with Bigrams. *In: Proceeding of 19th International Conference on Computational Linguistics (COLING'2000).*

Poesio, M., & Vieira, R. 1998. A Corpus-Based Investigation of Definite Description Use. *Computational Linguistics*, **24**, 183–216.

Poesio, M., Vieira, R., & Teufel, S. 1997. Resolving Bridging references in Unrestricted Text. *In:* (Mitkov, R. and Boguraev, B., 1997).

Prince, E. 1981. Toward a taxonomy of given-newinformation. *Pages 223–256 of:* Cole, P. (ed), *Radical Pragmatics.* New York: Academic Press.

Prince, E. 1982. The ZPG letter: subjects, definiteness, and information status. *In S Thompson and W. Mann, eds., Discourse description:diverse analyses of a fund-raising text. John Benjamins*, 295–325.

Proyecto ITEM. 1996-1999. *Recuperación de Información Textual en un Entorno Multilíngüe con Técnicas de Lenguaje Natural.* Comisión Interministerial de Ciencia y Tecnología TIC96-1243-C03. `http://sensei.ieec.uned.es/item`.

Quinlan, J. R. 1992. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

Reinhart, T. 1983. *Anaphora and Semantic Interpretation.* Croom Helm linguistics series. Beckenham, Kent, BR3 1AT: Croom Helm Ltd.

Rico, C. 1994. *Aproximación estadístico-algebraica al problema de la resolución de la anáfora en el discurso.* Ph.D. thesis, Universidad de Alicante.

Rigau, G., Atserias, J., & Agirre, E. 1997. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. *In:* ACL (ed), *Proceedings of ACL / EACL.*

Rigau, G., Rodríguez, H., & Agirre, E. 1998. Building Accurate Semantic Taxonomies from Monolingual MRDs. *In:* ACL (ed), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Interna-*

tional Conference on Computational Linguistics (COLING-ACL'98).

Russell, B. 1919. Descripciones. *Pages 46–56 of:* Villanueva, Luis Ml. Valdés (ed), *La búsqueda del significado: Lecturas de filosofía del lenguaje.* Tecnos.

Sager, N. 1981. Natural Language Information Processing. *Reading.*

Saquete, E., & Martínez-Barco, P. 2000. Grammar Specification for the Recognition of Temporal Expressions. *Pages 21.1–21.7 of: Proc. 11th International Conference on Machine Translation and Multilingual Applicantions in the New Millenium (MT'2000).*

Suárez, A. 2001. A Word Sense Disambiguation system applying Maximum Entropy Models. *In:* ACL (ed), *Proceedings of ACL / EACL'01.* (pendiente de aceptación).

TREE System. 1998. *TREE:Trans European Employement.* `http: //www2.echo.lu/langeng/projects/tree`. Visitada 2-12-98.

University of Massachussetts. 1997. *Information Extraction.* `http://www.dcs.shef.ac.uk/research/groups/`. Visitada el 10-12-1997.

van Rijsbergen, C.J. 1979. *Information retrieval.* London: Butterworths.

Vicente-Mateu, J. A. 1994. *La deíxis. Egocentrismo y subjetividad en el lenguaje.* Murcia: Servicio de Publicaciones de la Universidad de Murcia.

Vieira, R. 1998. *Definite Description Processing in Unrestricted Text.* Ph.D. thesis, University of Edinburgh.

Vieira, R. 1999. Co-reference resolution of definite descriptions. *Pages 497–503 of: Proceedings of VI Simposio Internacional de comunicación Social.*

Vieira, R., & Poesio, M. 1998. Processing definite descriptions in corpora. *In:* Botley, S., & McEnery, T. (eds), *Corpus-based and computational aproach to anaphora.* London: UCL Press.

Vieira, R., & Teufel, S. 1997. Towards resolution of bridging descriptions. *Pages 522–524 of:* ACL (ed), *Proceedings of ACL / EACL.*

Vossen, P. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, **3**(1).

Votilainen, A. 1988. NPTool, a Detector of English Noun Prhase. *In: Proceedings of the Workshop on Very Large Corpora, ACL.*

Votilainen, A., & Järvinen, T. 1995. Specifying a Shallow Grammatical Representation for Parsing Purposes. *In: Proceedings of the 7th European Chapter of the Association for Computational Linguistics (EACL).*

Wilks, Y. 1987. *Text searching with templates.* Tech. rept. Technical report ML 162. Cambridge Language Research Unit.

Wilks, Y. 1997. Information Extraction as a Core Language Technology. *Lecture Notes in Computer Science*, **1299**, 1–9.

Yangarber, R., & Grishman, R. 1998. NYU: Description of the Proteus/PET system used for MUC-7. *In:* (Morgan Kaufman Publishers, 1998).