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

Ir directamente al índice

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

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

Anar directament a l'index

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

Universitat d'Alacant
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos

Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español

Maximiliano Saiz Noeda

Alicante, junio de 2002

Esta Tesis Doctoral presentada por Maximiliano Saiz Noeda para la obtención del título de Doctor Ingeniero en Informática ha sido desarrollada bajo la dirección conjunta del Dr. Manuel Palomar Sanz, de la Universidad de Alicante, y de la Dra. Lidia Moreno Boronat, de la Universidad Politécnica de Valencia.

Agradecimientos

Al llegar al final de este trabajo (que es, a la vez, el principio de otros muchos), sería ingrato no recordar el esfuerzo de muchas personas que han contribuido a su satisfactoria realización.

En primer lugar, quiero agradecer a mis directores, Manuel Palomar y Lidia Moreno, su estímulo y orientación en mis tareas investigadoras, su comprensión y sus siempre valiosos consejos, sin los cuales habría sido imposible llevar a cabo esta Tesis.

Al Departamento de Lenguajes y Sistemas Informáticos, por su apoyo institucional que ha respaldado mi trayectoria profesional y contribuido a mi desarrollo como investigador, y a todos mis colegas del Departamento, sin dejar de mencionar especialmente a mis compañeros del Grupo de Procesamiento del Lenguaje y Sistemas de Información, con quienes he compartido, además de tareas de docencia e investigación, enriquecedores y gratos momentos durante los últimos años.

Al Grup de Processament del Llenguatge Natural de la Universitat Politècnica de València y al Research Group in Computational Linguistics de la Universidad de Wolverhampton, grupos hermanados con el nuestro, con los que he disfrutado de charlas y reuniones científicas de gran interés y cuyas contribuciones, sin duda, han quedado reflejadas en el presente trabajo. Quiero mencionar especialmente a Ruslan Mitkov, cuyo saber y aportación documental han sido fundamentales en esta Tesis.

Agradezco a Armando su ayuda con la "maquineta", que tan útil ha sido para el desarrollo de toda la programación presentada en este trabajo. A mi hermana, Belén, cuya implicación en muchas fases de esta Tesis ha supuesto un sacrificio en ocasiones

superior al mío propio; a ella y a Helena "con hache" agradezco su inestimable ayuda en temas lingüísticos.

A mis padres, que me han transmitido su aliento en todo momento, a mis amigos y, por supuesto, a Pepa, a la que nunca podré compensar (aunque prometo intentarlo) por tantos años de paciencia y apoyo generoso y constante.

A todos ellos, GRACIAS.

M. S. N.
Alicante, abril de 2002

Índice general

Índice de cuadros

Índice de figuras

1. Introducción

El Procesamiento del Lenguaje Natural (en adelante PLN), desde sus comienzos en los años 50, ha intentado poner a prueba a investigadores de todo el mundo en la resolución de tareas que, bajo una aparente simplicidad desde el punto de vista humano, han escondido una elevada complejidad de resolución desde la perspectiva computacional.

La cinematografía y la novela de ciencia ficción de las últimas décadas del siglo XX predecían la existencia de engendros mecánicos y grandes "superordenadores" que a comienzos del presente siglo tendrían completamente superadas las barreras de comunicación hombre-máquina a través del lenguaje natural. Desgraciadamente, muchas de estas barreras siguen siendo retos aparentemente inalcanzables.

El PLN, que intenta simular el comportamiento lingüístico humano, se define como "una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre-máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales" (Moreno et al., 1999).

Esta flexibilidad de los lenguajes naturales frente a los formales va acompañada del fenómeno de la ambigüedad, que es uno de los principales problemas que un sistema de PLN ha de resolver. De entre los fenómenos lingüísticos que plantean ambigüedad en la comprensión de la información destaca por su relevancia la anáfora. El fenómeno de la anáfora, elemento fundamental de la cohesión entre oraciones y marcador de la coherencia del texto (Rigau, 1981), se enmarca en el ámbito de la ambigüedad referencial (Moreno et al., 1999) y se fundamenta en el principio de

la economía lingüística (ver capítulo 2). Su estudio y resolución, que interesó especialmente con la llegada de la gramática textual, requiere una perspectiva lingüística amplia que, junto al análisis morfosintáctico, exige, tal y como se mostrará en el presente trabajo, la consideración semántica y la contextual o pragmática.

Podríamos, así, diferenciar varios niveles de conocimiento, correspondientes a los distintos niveles lingüísticos, necesariamente implicados en el proceso de resolución de la anáfora: nivel léxico, referente al vocabulario de una lengua; nivel morfológico, relativo esencialmente a los morfemas de género, número y persona; nivel sintáctico, que analiza estructuras de secuencias de unidades léxicas; nivel semántico, que trata el significado o sentido de los elementos y estructuras oracionales; nivel pragmático, que pone en relación las unidades lingüísticas con el contexto extralingüístico.

Algo tan simple para un lector humano como relacionar las entidades del discurso en función de la evolución de un texto o resolver la ambigüedad introducida por elementos textuales sin carga semántica, como el caso del pronombre, tiene ocupados a muchos grupos de investigación a lo largo de la última década. Uno de los grupos más activos en el territorio español dedicados a la resolución de problemas lingüísticos es el Grupo de Procesamiento del Lenguaje y Sistemas de Información (en adelante GPLSI) de la Universidad de Alicante. Este grupo, desde su creación a principios de los noventa, ha venido trabajando en la resolución de fenómenos lingüísticos como la elipsis y la anáfora para su aplicación a tareas clásicas del PLN como la extracción de información, la recuperación de información, la traducción automática, los sistemas de búsqueda de respuesta[1], los resúmenes de texto, los sistemas de diálogo, etc.

[1] Estos sistemas son conocidos en la bibliografía como sistemas de *Question Answering*.

1.1 Motivación

En lo referente a la resolución de la anáfora, tarea que fundamenta la escritura de este trabajo, los distintos estudios realizados por el GPLSI han cubierto un amplio espectro de tareas que han culminado en la publicación de un conjunto de tesis doctorales. Estas tareas han sido la resolución de la elipsis (Palomar, 1996), la resolución de la anáfora pronominal y adjetiva (Ferrández, 1998), la resolución de descripciones definidas (Muñoz, 2001), la resolución de anáforas en diálogos (Martínez-Barco, 2001) y la resolución y generación de anáforas en sistemas multilingües (Peral, 2001). Todos estos trabajos han proporcionado buenos resultados y una base fundamental de nuevas líneas de investigación, que se mencionarán en esta memoria. Todos ellos[2] han partido del uso de un análisis sintáctico parcial y un conjunto de restricciones y preferencias que, con el uso de información morfológica y sintáctica, han proporcionado resultados enormemente interesantes.

Una de las tareas pendientes, a la cual los mencionados trabajos hacen referencia, es la incorporación de información semántica y de conocimiento del mundo para resolver este tipo de fenómenos. La información semántica se ha revelado como uno de los factores más importantes que influyen en los procesos humanos de resolución de la correferencia. Es evidente que en oraciones como (1), el oyente identifica sin ningún problema el referente del pronombre sujeto omitido a través del concepto semántico asociado al verbo y al atributo que le acompaña.

(1) El mono subió al árbol a coger un *plátano*$_i$ cuando el sol salía. $\emptyset_i$ estaba maduro.

Esta idea era difícil de llevar a la práctica en sistemas computacionales y especialmente, en aquellos de propósito general, por el elevado coste que suponía dotar a los elementos textuales de características semánticas. Sin embargo, con el nacimiento de

[2] A excepción de la estrategia de resolución de descripciones definidas (Muñoz et al., 2000; Muñoz, 2001), en la que se describe una propuesta basada en los sentidos de las palabras y las relaciones semánticas para incrementar la eficiencia del algoritmo.

recursos lingüísticos como WordNet[3] comenzó a plantearse la posibilidad de incorporar este tipo de información a tareas de PLN sin un coste prohibitivo.

Es por ese motivo por el que se abrió una nueva línea de trabajo para incorporar la información semántica al proceso de resolución de la anáfora. Esta nueva línea trataría de usar una técnica de resolución similar a la utilizada hasta el momento en los distintos trabajos del grupo de investigación, pero incorporando una nueva fuente de información que enriqueciera el proceso de resolución y, por tanto, incrementara la eficacia de un sistema global de PLN. Para ello se contó, en principio, con el recurso léxico WordNet para su aplicación en inglés y, con el posterior nacimiento del proyecto EuroWordNet, su aplicación pudo ser extendida al tratamiento de textos en español.

Esta incorporación de la información semántica a los procesos de resolución de la anáfora llevaba implícito el uso de los papeles sintácticos de los componentes oracionales del corpus de entrada. Así, bien a través de análisis completos o bien partiendo de análisis sintácticos parciales con enriquecimientos, el etiquetado de papeles sintácticos posibilitaría la propuesta de nuevos métodos de resolución basados en información sintáctico-semántica.

De este modo, en esta Tesis se presenta un estudio sobre la influencia que sobre el proceso de resolución de la anáfora tienen tanto la información sobre papeles sintácticos como la información semántica basada en conceptos ontológicos extraídos de WordNet. Además del estudio realizado sobre esta influencia, se propone un método de resolución basado en la aplicación de estas fuentes de información. Este trabajo complementa los aspectos relativos a la resolución de la anáfora en español tratados y viene a llenar el hueco que hasta ahora existía en los trabajos que sobre esta tarea se han desarrollado tanto en el seno del GPLSI como en muchas otras aproximaciones a la resolución computacional de la anáfora.

[3] El apartado 4.3.4 explica con detenimiento los aspectos más relevantes de este recurso en lo que atañe al presente trabajo.

1.2 Objetivos de este trabajo

De acuerdo con lo expuesto en la sección anterior, el objetivo fundamental de este trabajo es demostrar la relevancia de la información basada en papeles sintácticos y combinada con la información semántica en el proceso de resolución computacional de la anáfora.

Para conseguir este objetivo, es necesario un estudio previo de la importancia que tiene la información sobre papeles sintácticos (a la que llamaremos información sintáctica enriquecida) y la información semántica en relación con otras fuentes que se aplican en el proceso de resolución de la anáfora.

Se tratarán distintos ejemplos en los que se podrá comprobar la relevancia de estas fuentes de información y se planteará una serie de estrategias de resolución basadas, entre otras fuentes, en la información de origen sintáctico enriquecido y semántico. Estas estrategias se definirán en el marco de un conjunto de sistemas que aplican fuentes de información de distinta índole a la resolución de la anáfora. Estos sistemas serán considerados *de conocimiento limitado* cuando utilicen fuentes puramente morfosintácticas (ver sección 3.1) y *enriquecidos* cuando incorporen información semántica y de la estructura del discurso (ver sección 3.2).

El objetivo último es mostrar los resultados positivos que ofrece la incorporación de la semántica y el análisis sintáctico enriquecido en la resolución computacional de la anáfora.

1.3 Organización y estructura de la Tesis

Esta Tesis consta de siete capítulos, siendo el primero de ellos esta introducción, encargada de realizar una breve descripción del trabajo realizado y de la propia organización de su exposición.

El segundo capítulo se ocupa de contextualizar el fenómeno de la anáfora, así como de establecer una serie de clasificaciones fundamentadas en diferentes criterios. El objetivo perseguido en este capítulo es el de cincunscribir el problema de la anáfora dentro del conjunto de fenómenos en los que se encuadra y con los que se

relaciona, así como el de establecer la base lingüística sobre la que se trabajará a lo largo de los diferentes capítulos que conforman esta Tesis.

El tercer capítulo realiza un repaso de las diferentes estrategias, métodos y sistemas propuestos para la resolución computacional de la anáfora. El capítulo organiza la exposición de estos métodos siguiendo el mismo esquema de la propuesta de esta Tesis, es decir, los divide en tres grandes grupos en función de las fuentes de información que utilizan. Así, se distingue entre métodos de conocimiento limitado, cuando la resolución se basa en criterios puramente morfosintácticos, métodos enriquecidos, cuando se aplica además algún tipo de información semántica o de discurso al proceso de resolución y, finalmente, métodos alternativos, cuando los procesos de resolución, están basados en técnicas extra-lingüísticas.

El cuarto capítulo expone la propuesta principal del método enriquecido de resolución de la anáfora pronominal en español (ERA). Para llegar a esta propuesta, se trata, en primer lugar, el origen de las fuentes de conocimiento que intervienen en el proceso de resolución de la anáfora. A continuación, se expone el método de resolución de la anáfora pronominal en español (Palomar et al., 2001a), basado en información puramente morfosintáctica y que sirve como base metodológica para el método enriquecido. Tanto el método de conocimiento limitado como el método enriquecido van acompañados de una descripción del conjunto de herramientas y recursos usados para su aplicación, así como de la definición de las restricciones y preferencias que utilizan en el proceso de resolución.

El quinto capítulo muestra los resultados de la evaluación realizada, por un lado, con el método de conocimiento limitado y, por otro, con el método ERA. Para el primero, se muestran datos cuantitativos obtenidos de la evaluación sobre un corpus extenso y de la comparación de los resultados de este método con los de otros métodos clásicos que han sido implementados. En el caso del método enriquecido, se muestran los resultados obtenidos en la evaluación con el uso de un banco de pruebas diseñado para tal fin. Además de los datos puramente cuantitativos, se realiza un es-

tudio detallado de la influencia que cada conjunto de restricciones y preferencias tiene sobre el proceso global de resolución.

El sexto capítulo trata el marco de aplicación de la resolución de la anáfora en general y del método ERA en particular. Para ello, describe los aspectos que relacionan el método enriquecido con las tareas de desambiguación léxica, justificando la problemática que WordNet plantea en estas tareas de desambiguación. Asimismo, el capítulo expone dos propuestas de aplicación del método basadas en el uso de herramientas de desambiguación y campos temáticos de WordNet, así como en la extensión de WordNet con terminología del sector público definida en el proyecto EuroTerm. Por último, el capítulo describe, a través de la propuesta del proyecto TUSIR, la integración de las tareas de resolución de la anáfora en sistemas de comprensión de textos aplicada a la Recuperación de Información.

El séptimo y último capítulo recoge las conclusiones del trabajo, así como de plantear algunas líneas de trabajo en progreso, cuyo objetivo es el de mejorar la propuesta.

Tras las referencias bibliográficas usadas para el desarrollo de este trabajo de investigación que se incluyen al final de esta Tesis, se presenta un anexo en el que se incluyen las tablas resumen de los datos sobre la evaluación del método ERA.

2. Ámbito del problema

2.1 Contextualización y definición de la anáfora

El lenguaje natural obedece a tres principios básicos, de los que se derivan las propiedades que lo definen: economía, creatividad y simbolismo[1]. La primera de ellas, la economía, sirve como punto de partida para presentar el fenómeno de la anáfora.

El principio de la economía fundamenta las propiedades de intercambiabilidad, dualidad y eficiencia del lenguaje.

- *Intercambiabilidad.* Los participantes en la comunicación pueden transmitir y recibir mensajes, sin que cada una de estas actividades requiera el conocimiento y dominio de reglas gramaticales distintas.
- *Dualidad.* El lenguaje natural se organiza en dos niveles: uno integrado por un número limitado de unidades mínimas carentes de significado o fonemas, y otro en el que esas unidades se agrupan, de acuerdo con un número limitado de reglas combinatorias, formando un número ilimitado de unidades con significado (morfemas, oraciones y discursos).
- *Eficiencia.* El lenguaje natural consta además de elementos que cambian su denotación de acuerdo con la situación comunicativa en la que se empleen. Ciertamente, una misma unidad lingüística puede emplearse para hacer referencia a determinadas entidades del mundo (objetos e ideas, reales o imaginarios) en función

[1] Hockett (1971) propuso en la década de los cincuenta una serie de propiedades definitorias de las lenguas y del lenguaje humanos. Dicha caracterización sigue siendo todavía punto de partida indiscutible sobre tema. Véase también Moreno-Cabrera (1991).

de los participantes, el lugar y el tiempo en que se produce el acto comunicativo.

Es, pues, evidente que estas tres propiedades de orden estructural y de funcionamiento redundan en la economía del sistema lingüístico.

2.1.1 Elipsis

Uno de los fenómenos más característicos y complejos que afectan a todo sistema lingüístico está vinculado al principio de economía. Se trata de la supresión o elisión de unidades lingüísticas. La elisión se manifiesta en todos los niveles lingüísticos (Abad, 1980):

- Elisión fonética: puede producirse al principio (aféresis), al final (apócope) o en mitad de la palabra (síncopa). Tales son los casos de "*norabuena*" por "*enhorabuena*", "*labstracción*" por "*la abstracción*" y "*algún*" por "*alguno*", respectivamente.
- Elisión morfológica: presente en algunos procesos de sufijación, como el sincretismo de "*tenista*" por "*tenis+ista*".
- Elisión sintáctica: este tipo de elisión recibe el nombre de elipsis, como en "*Juan es rico, pero su hermano no*" por "*Juan es rico, pero su hermano no es rico*".

Conviene distinguir dos tipos de elipsis (Lyons, 1971):

- Elipsis contextual: tal es el caso de exclamaciones como "*¡Gracias!*" o "*¡Buenos días!*" por "*le deseo buenos días*" o "*le doy las gracias*".
- Elipsis gramatical: como sucede en "*–¿De quién es este coche? –De Pedro, si no lo ha vendido todavía.*" por "*–¿De quién es este coche? –Este coche es de Pedro, si Pedro no ha vendido este coche todavía.*".

Es obvio que el fenómeno de la elipsis tiene mucho que ver con el hecho de que en los seres humanos la capacidad de memoria y procesamiento de la información es limitada. La elipsis, al no hacer explícitos elementos innecesarios, constituye un medio no sólo de

aligerar la expresión, sino también de facilitar el procesamiento de la información recibida.

Por otra parte, si los fenómenos de elisión fonética y morfológica pertenecen al ámbito de estudio de la competencia gramatical de los hablantes[2], la elisión sintáctica o elipsis (ya sea contextual o gramatical) nos obliga a considerar la existencia en el hablante de una competencia discursiva, que hace posible que éste sea capaz de generar e interpretar un discurso como un todo y no sólo como resultado de una mera sucesión de frases.

En el marco de la lingüística del discurso (escrito u oral) suele hacerse una distinción entre coherencia y cohesión (Moeshler y Reboul, 1991). La coherencia o interpretabilidad tiene que ver con las propiedades de orden temático (unidad del tema y dependencias lógicas entre los distintos subtemas) que hacen posible la interpretación de un discurso a partir de la puesta en funcionamiento de mecanismos lógico-interpretativos como la inferencia. Por ejemplo, de "*El agua está muy fría, me quedaré en la arena*" se inferirían enunciados implícitos como "*podría resfriarme*" o "*no me gusta el agua fría*". La cohesión discursiva o continuidad informativa, por su parte, es resultado de las relaciones proposicionales entre frases. Un discurso estará cohesionado si en él se mantienen las transiciones textuales, la progresión temática o el cambio de enfoque, que marcadores como "*finalmente*", "*así pues*" o "*en primer lugar*" se encargan de señalar.

Cabe señalar, no obstante, que el mecanismo que garantiza las relaciones intraoracionales, interoracionales y extraoracionales, y, por tanto, la coherencia y la cohesión discursivas, es la referencia, esto es, la relación entre determinadas unidades lingüísticas y los objetos o entidades del mundo real o de un mundo posible. Se trata, pues, de una relación palabra-mundo.

2.1.2 Deixis

Una forma eficaz de producir la referencia es introducir en el discurso elementos que remitan al marco en el que se produce el

[2] Se entiende por competencia el conocimiento que el hablante tiene de su lengua (Chomsky, 1965).

acto comunicativo, es decir, a los participantes y a las coordenadas espacio-temporales tanto de éstos como del propio acto. Este procedimiento es conocido con el nombre de deixis (del griego *δεικνυμι* ó *δεικνυω*, 'indicar, señalar').

Los pronombres personales (*yo*, *tú*, *él*, ...) son una manifestación de carácter universal del mecanismo lingüístico de la deixis. No obstante, existen otras unidades lingüísticas que forman parte también del grupo de elementos deícticos. Tal es el caso de adjetivos/pronombres demostrativos (*este/éste*, *ese/ése*, *aquel/aquél*, ...), manifestaciones de la denominada deixis de persona, es decir, la que indica o señala la identidad de los interlocutores presentes en el acto comunicativo; de adverbios de lugar (*aquí*,*allí*, *allá*, ...) y de tiempo (*ahora*, *mañana*, *hoy*, *más tarde*, ...), manifestaciones de la deixis de lugar y de tiempo respectivamente, esto es, que sirven para indicar o señalar las coordenadas espacio-temporales del acto comunicativo. Algunos verbos tales como la pareja *ir/venir* constituyen también oposiciones deícticas: *venir* se emplea para denotar movimiento hacia el hablante o hacia algún lugar relacionado con éste, mientras que *ir* hace referencia a un movimiento de alejamiento del hablante. El tiempo verbal tiene también una fuerte carga deíctica. Por ejemplo, la afirmación "*estoy trabajando*" deja claro que el acto de trabajar coincide temporalmente con el acto de comunicación.

Llamamos, pues, deixis al procedimiento lingüístico mediante el cual introducimos en el discurso unidades lingüísticas (pronombres, adverbios y verbos) que hacen referencia a distintos elementos del acto comunicativo o de la enunciación.

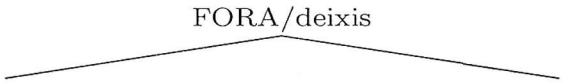
2.1.3 *Fora* y anáfora

Cabe señalar que estas relaciones palabra-mundo pueden ser de naturaleza extraoracional o extradiscursiva (si las unidades lingüísticas denotan deícticamente entidades del mundo exterior al mensaje lingüístico), o bien intraoracional o intradiscursiva (si éstas remiten a una entidad interna al mensaje o, dicho de otra manera, cuando la referencia está dentro del contexto lingüístico mismo). En el primer caso decimos que se trata de una deixis

no textual o exofórica, y en el segundo, de una deixis textual o endofórica. Por ejemplo, en "*Mira esto*", el pronombre *esto* es manifestación de una deixis de naturaleza exofórica, mientras que en "*Pedro dice que él no lo hará*", *Pedro* y *él* mantienen una relación deíctica de naturaleza endofórica.

En las relaciones endofóricas se establece una correferencia entre las unidades lingüísticas implicadas. En este sentido cabe señalar que dos o más elementos son correferentes cuando hacen referencia a la misma entidad (individuo, objeto o idea). Convencionalmente los índices *i/j* se emplean para señalar la lectura correferente o no. Así por ejemplo, en "*Pedro dijo que él no vendría*", indicaríamos la lectura correferente como "*Pedro$_i$ dijo que él$_i$ no vendría*" y la no correferente como "*Pedro$_i$ dijo que él$_j$ no vendría*".

Según Moreno-Cabrera (1991), si empleamos el término *fora* para designar aquellas unidades del discurso que remiten a otro elemento (interno o externo al mismo mensaje), podríamos proponer el siguiente esquema que resume todos los fenómenos de naturaleza deíctica mencionados anteriormente:

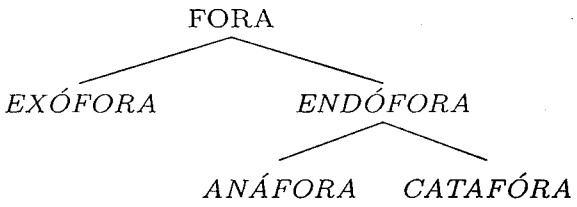


Además, en ocasiones la relación se establece entre un elemento generalmente pronominal y otro denominado antecedente, que aparece en el contexto lingüístico inmediato, es decir, en la misma frase o en otra anterior, como en "*Juan$_i$ cree que no lo$_i$ llamarán*", o en "*[Ya he comprado el libro]$_i$. No se lo$_i$ diré.*". Dicha relación se denomina deixis anafórica o simplemente anáfora (del griego $\alpha\nu\alpha\phi o\rho\alpha$, 'referencia, remisión').

Cuando la relación de correferencia se establece entre un elemento, generalmente pronominal, y otro que aparece a continuación (consecuente), decimos que se trata de una deixis catafórica

o catáfora[3], como sucede en "*Todos los que la$_i$ conocen dicen que María$_i$ es muy simpática*".

Podemos incluir estas definiciones en el esquema anterior de la siguiente manera:



Así pues, tanto la anáfora como la catáfora se consideran categorías de endófora (Moreno et al., 1999), la cual viene definida por su dependencia del contexto lingüístico, en oposición a la exófora, que se desarrolla en el contexto situacional.

2.2 Clasificación de la anáfora

Definido el marco en el que se produce el fenómeno de la anáfora en general, proponemos a continuación una serie de clasificaciones que pretender dar cuenta de la complejidad lingüística de dicho fenómeno.

Si bien el espectro de clasificaciones de la anáfora es enormemente amplio, en el presente trabajo se ha optado por tres clasificaciones que aluden, por una parte, a la relación entre el elemento anafórico y su antecedente y, por otra, a la categoría gramatical tanto del antecedente como de la anáfora. Este planteamiento recoge la esencia del fenómeno lingüístico de la anáfora en sus diferentes vertientes y sirve como punto de partida óptimo para el desarrollo de este trabajo.

[3] Merece la pena señalar que en el ámbito de la gramática generativo-transformacional (Haegeman, 1994) se ha sustituido el término *antecedente* por el de *backwards anaphora* ('anáfora hacia atrás') y el *consecuente*, por *following anaphora* ('anáfora hacia delante'). Es decir, esta teoría lingüística interpreta la catáfora como un tipo de anáfora con el fin de unificar las condiciones que regulan la anáfora y la catáfora tradicionales.

2.2.1 Según la relación entre el elemento anafórico y su antecedente

Atendiendo a la relación entre el término anafórico y su antecedente, cabe distinguir dos tipos básicos de anáfora:

- Anáfora de referencia (Rigau, 1981) o profunda (Moreno et al., 1999). Se da cuando dos o más elementos que mantienen una relación anafórica comparten referente, entendiendo por referente la entidad del mundo a la que estos elementos remiten, como puede verse en (2).

(2) Luisa cortó el *vestido*$_i$ y María *lo*$_i$ cosió.

- Anáfora de sentido (Rigau, 1981) o superficial (Moreno et al., 1999). Se da cuando dos o más elementos que mantienen una relación anafórica tienen el mismo significado pero distinto referente, como sucede en (3).

(3) Andrés perdió su *pasaporte*$_i$ y a Luis se *lo*$_i$ robaron.

2.2.2 Según la categoría gramatical del antecedente

Atendiendo a la categoría gramatical del antecedente anafórico, podemos realizar la siguiente clasificación:

- *Sintagma nominal.* El antecedente tiene como núcleo un nombre, común (4) o propio (5).

(4) Arturo se ha puesto *gafas*$_i$. *Las*$_i$ ha comprado en la óptica de Pedro.

(5) *Arturo*$_i$ se ha puesto gafas. *Le*$_i$ quedan muy bien.

- *Sintagma verbal.* El antecedente tiene como núcleo un verbo.

(6) Mi mujer quiere *conducir durante toda la noche*$_i$ pero yo no quiero que *lo haga*$_i$.

- *Sintagma adverbial.* El antecedente anafórico está representado por un adverbio, como ocurre en (7).

(7) María está *arriba*$_i$. *Allí*$_i$ se trabaja mejor.

Con frecuencia, un sintagma preposicional (SP) tiene valor adverbial en la oración, desempeñando la función de complemento circunstancial de tiempo, lugar, modo, ... Cuando la anáfora tiene también ese valor adverbial, como en (8), estos casos de sintagama preposicional son similares a los ya mencionados de sintagma adverbial, por lo que se pueden incluir en el mismo grupo.

(8) María está trabajando *en la buhardilla*$_i$. *Allí*$_i$ hay más luz.

- *Oración completa, hecho o idea.* Un antecedente anafórico puede estar representado por una oración completa, como en (9), así como por un conjunto de ellas, texto o fragmento de texto, por lo que la anáfora hará alusión a un hecho o una idea mencionados anteriormente.

(9) *Marisa está embarazada*$_i$. Su marido no *lo*$_i$ sabe.

2.2.3 Según la categoría gramatical del elemento anafórico

La clasificación propuesta a continuación se fundamenta en la categoría gramatical asociada al elemento o elementos anafóricos de la oración. Así, las anáforas se agrupan según se trate de pronombres, sintagmas nominales, verbos o adverbios, con la excepción de la llamada anáfora superficial numérica que, pudiendo incluirse tanto en las anáforas pronominales como en las de sintagma nominal (SN), precisamente por este motivo constituye por sí misma un grupo.

Cada tipo de anáfora se acompaña de dos ejemplos. El primero de ellos corresponde a un caso que podría resolverse con los sistemas basados en criterios morfosintácticos (de conocimiento limitado) expuestos en el capítulo 3 (sección 3.1, pág. 30). El segundo plantea un caso en el que la relación entre la anáfora y

su antecedente viene determinada por rasgos semánticos y cuya resolución implicaría, por tanto, el uso de criterios de naturaleza semántica.

Anáfora pronominal. La anáfora pronominal, objeto primordial de nuestro estudio, es la más frecuente de todas, también la de mayor complejidad, por la amplitud y complejidad de la categoría misma del pronombre.

Los distintos tipos de anáfora pronominal responden a los distintos tipos de pronombres tradicionalmente establecidos por la gramática, ocupando un lugar central los pronombres personales. Muy vinculados a ellos, otro grupo lo constituyen los pronombres reflexivos y recíprocos, así como los demostrativos y posesivos. Los pronombres de relativo representan un apartado especial por reunir la doble cualidad de conjunción (introducen una oración subordinada adjetiva o de relativo) y pronombre. Asimismo, dada la naturaleza pronominal de la palabra inglesa *one*, se incluye entre las anáforas pronominales la llamada *one-anaphora*.

1. Anáfora de pronombre personal

   En la siguiente clasificación se asume la diferenciación tradicional entre pronombres personales de sujeto y de objeto o complemento, así como entre pronombres tónicos y átonos. La noción que reúne un número limitado de pronombres en este grupo es la de persona gramatical. Los pronombres son sintagmas nominales, pertenecen a la clase del sustantivo, al que, contrariamente a la idea contenida en la propia palabra pronombre, no siempre se puede decir que sustituyan. Los pronombres personales tónicos (*yo / tú / él / ella / ello / nosotros / nosotras / vosotros / vosotras / ellos / ellas*) son un claro ejemplo de ello, por lo que algunos autores prefieren llamarlos "sustantivos personales" (Alarcos, 1994).

   - *De pronombre personal sujeto* (*yo / tú / él / ella / ello / nosotros / nosotras / vosotros / vosotras / ellos / ellas*).

Como señala Gili Gaya (1961), en español se hace poco empleo de este tipo de pronombres, dada la claridad de las distintas desinencias de las formas verbales[4]. En este sentido, el pronombre personal en primera y segunda persona es enfático, mientras que en tercera persona puede haber ambigüedad, puesto que mientras primera y segunda persona sólo hay una, las terceras personas pueden ser varias[5]. Es propio de los pronombres de primera y segunda persona del singular remitir a los participantes en el acto comunicativo; de ahí su consideración como deícticos. Esta característica no es propia de los pronombres de tercera persona, que pudiendo presentar un uso deíctico, pueden designar cualquier individuo u objeto distinto del oyente y del hablante, tanto si está presente en el acto de habla como si no. Ello les confiere un valor referencial (Fernández, 1999) o anafórico (Hernanz y Brucart, 1987), ya que su interpretación se realiza a través de la presencia en el contexto lingüístico inmediato de una palabra con la que el pronombre mantiene relación de correferencia. Por ello, nuestro estudio se centrará en los pronombres de tercera persona.

La anáfora de (10) puede ser resuelta sin problemas aplicando criterios puramente morfológicos.

Por otro lado, a pesar de que en (11) existe concordancia morfológica completa entre el pronombre omitido de tercera persona y los tres sintagmas nominales de la oración anterior, sólo uno de ellos (*el plátano*) puede ser relacionado con el pronombre anafórico por ser el único al que se puede asociar el rasgo de estar *maduro*.

---

[4] A diferencia de los que ocurre en inglés o francés, lenguas en las que las desinencias personales se han perdido u oscurecido obligando a anteponer el pronombre (a no ser que el sujeto aparezca junto al verbo).

[5] Si bien esta observación puede parecer excesivamente trivial, justifica el hecho de que la resolución de la anáfora en la rama del procesamiento del lenguaje natural se haya centrado fundamentalmente en las terceras personas pronominales y no en la primera o la segunda.

(10) *Andrés*$_i$ sabe la *combinación*$_j$ de la *caja*$_k$ fuerte. *Él*$_i$ está hoy de viaje.

(11) El *mono*$_k$ subió al *árbol*$_j$ a coger un *plátano*$_i$. ($\emptyset_i$) estaba maduro.

El neutro pronominal (*ello*, *le* dativo y *lo* acusativo) representa un caso especial, por cuanto, al no existir en español sustantivos neutros, hace referencia a conceptos antes mencionados que no son, lógicamente, sustantivos morfológicos.

- *De pronombre personal complemento*:
  - Las *formas tónicas (mí / ti / sí / usted / él / ella / ello / nosotros / nosotras / vosotros / vosotras / ustedes / ellos / ellas / conmigo / contigo / consigo)*, acompañadas siempre por una preposición, pueden desempeñar función de objeto directo, indirecto o complemento circunstancial.
  - Las *formas átonas (me / nos / te / os / le / la / lo / les / las /los / se)* se emplean siempre sin preposición. La primera y segunda persona se usan como formas únicas de los complementos directos e indirectos sin preposición. En los de tercera persona, los pronombres *lo*, *la*, *los* y *las* funcionan como complemento directo, mientras que *le*, *les* y *se* funcionan como complemento indirecto[6].

  De nuevo, se muestra en la misma manera que en el apartado anterior, en (4) hay tres sintagmas nominales que pueden ser antecedentes del pronombre anafórico *la*, pero únicamente la televisión puede ser apagada.

(12) No tengo noticias de *Luis*$_i$. No *lo*$_i$ veo desde octubre.

(13) La *televisión*$_i$ está encendida cuando *Luisa*$_j$ llega a la *cocina*$_k$. Ella *la*$_i$ apaga cuando se acuesta.

Los pronombres átonos, a diferencia de los tónicos, especialmente los de complemento indirecto, pueden co-aparecer

[6] Las alteraciones sufridas en el uso correcto de estas formas pronominales conducen a los fenómenos del leísmo y el laísmo.

también con sintagmas nominales plenos, en lo que se conoce como "reduplicación" o "doblado" de clíticos (Fernández, 1999): "*Le$_i$ di las llaves a ella$_i$*".

2. Anáfora de pronombre omitido

En español es extremadamente frecuente la anáfora producida por la omisión del sujeto o anáfora cero, tal y como se puede ver en los ejemplos (14) y (15).

(14) *Luis$_i$* entregó los papeles a los asesores. $\emptyset_i$ Estaba preocupado por los plazos de presentación.

(15) *Isabel$_i$* llamó a la *empresa$_j$* de mudanzas. $\emptyset_i$ Deseaba marcharse cuanto antes.

Si bien este fenómeno puede ser considerado como un tipo de elipsis[7], en este trabajo será tratado como un tipo de anáfora pronominal, bajo la suposición de que el elemento elidido es un pronombre que concuerda morfológicamente con el verbo al que acompaña. Una vez determinada esta sustitución se establecerán criterios de selección del antecedente similares a los propuestos para el resto de los pronombres tratados.

3. Anáfora de pronombre demostrativo

La correcta utilización de un pronombre demostrativo con función anafórica, por el carácter deíctico de éste, permite una clara identificación del antecedente con el que correfiere. Por ejemplo, en "*Luis está enfadado con Antonio. Éste no le habla desde hace años*", se asociaría de manera natural *Antonio* con el pronombre *éste*, puesto que es el más cercano.

[7] Aparecen vinculadas al fenómeno de la elipsis tanto la denominada anáfora cero (conocida también por el término inglés *zero-anaphora*) como la llamada anáfora de complemento nulo (del inglés *null complement anaphora*): "*Luis fue al acto; María, en cambio, no pudo Ø*" (Brucart, 1999). Para un tratamiento exhaustivo del fenómeno de la elipsis y su resolución computacional, véase (Palomar, 1996).

(16) De entre los asistentes destacaba una *joven*$_i$ con rasgos orientales. *Ésta*$_i$ parecía ausente.

(17) Antonio conoce el *nombre*$_i$ del *pintor*$_j$. *Éste*$_i$ se pronuncia con dificultad.

Una vez más, en (17), un rasgo semántico (la posibilidad de ser pronunciado) selecciona de entre los tres antecedentes posibles el único que se puede asociar a dicho rasgo (*el nombre*).
También en este apartado habría que destacar el caso de las formas neutras del pronombre demostrativo (*esto*, *eso*, *aquello*), en el mismo sentido que se ha señalado en el apartado de los pronombres personales neutros.

4. Anáfora de pronombre posesivo

En (19) la selección del antecedente correcto de la anáfora está basada de nuevo en la información semántica contenida en su sintagma verbal: sólo el coche puede estar estropeado (al menos en sentido literal).

(18) Tus *ojos*$_i$ son azules. Los *suyos*$_i$ son verdes.

(19) Este *coche*$_i$ es del *hermano*$_j$ de su *amigo*$_k$. El *mío*$_i$ está estropeado.

5. Anáfora de pronombre reflexivo

El pronombre reflexivo correfiere por definición con el sujeto del verbo del que depende.

(20) *Marta*$_i$ *se*$_i$ pinta mucho.

(21) *Luis*$_i$ fue de excursión al río$_j$. *Se*$_i$ bañó con sus amigos.

En (21) no es posible asociar al río la capacidad de bañarse, por lo que resulta evidente que Luis es el antecedente del pronombre anafórico[8].

Habría que mencionar en este apartado el caso en el que el elemento anafórico es un pronombre recíproco. Si en el pronombre reflexivo la acción recae sobre el sujeto del verbo al que acompaña, en el caso del pronombre recíproco el antecedente (así como el sujeto de la oración) es plural y expresa una acción que cada integrante de dicho sujeto ejerce sobre el otro y recibe de él: "*Luisa se casa con Juan en septiembre. Se quieren mucho.*"

En el caso de las oraciones recíprocas, el pronombre se suele acompañar de palabras o frases que eviten la ambigüedad ("*entre sí*", "*uno a otro*", "*mutuamente*", "*recíprocamente*", ...) para distinguirlas de acciones comunes que afectan a más de un sujeto pero no son recíprocas ("*Luis y Miguel se quejan mucho*").

6. Anáfora de relativo

Como se ha apuntado en la introducción de este apartado, el pronombre de relativo se caracteriza por ser conjunción además de pronombre con una función sintáctica determinada en la oración que introduce.

Dejando a un lado la ambigüedad que en (23) el sintagma nominal *El perro de mi amigo* puede plantear en español, el antecedente del relativo *que* es seleccionado por la información semántica del sintagma verbal: mi amigo puede trabajar en un banco, el perro no.

(22) *Los discos*$_i$ *que*$_i$ te presté son muy antiguos.

[8] La resolución previa de la anáfora correspondiente al sujeto omitido permitiría resolver el pronombre reflexivo a partir de la información sintáctica que relaciona dicho pronombre con el sujeto.

(23) El *perro*$_i$ de *mi amigo*$_j$, *que*$_j$ trabaja en un banco, es de pura raza.

7. One anaphora

Esta tipo de anáfora, estudiado exclusivamente en el caso del inglés[9], sustituye el sustantivo antecedente por el pronombre anafórico *one*. En (25) la anáfora plantea una relación semántica entre *negro* (*black*) y *oscuro* (*dark*).

(24) *I have washed* all my skirts$_i$ *and* the blue one$_i$ *has shrunk.*

He lavado *todas mis camisas*$_i$ y *la azul*$_i$ ha encogido.

(25) *I have* a black bicycle$_i$ *and* a white bicycle$_j$, *but I prefer* the dark one$_i$.

Tengo *una bicicleta negra*$_j$ y *una bicicleta blanca*$_j$, pero prefiero *la oscura*$_i$.

Anáfora de sintagma nominal (descripciones definidas). La clasificación de los tipos de anáfora de sintagma nominal está basada en el tipo de determinante del SN que cumple la función anafórica (artículo determinado, demostrativo o posesivo).

1. SN con artículo determinado
Mientras en (26) la resolución de la anáfora se puede realizar a través de mecanismos exclusivamente léxicos, en (27) es necesario establecer relaciones de carácter semántico entre el antecedente y la anáfora. En este caso existe una relación de sinonimia entre *empresa* y *compañía*.

[9] Ferrández (1998) llama anáfora de tipo adjetivo a la correspondiente a la *one-anaphora* en español, en la que aparece un sintagma nominal con el núcleo nominal elidido cuya función es realizada por el adjetivo. La diferente consideración del adjetivo como modificador del nombre (núcleo) elidido o como el propio núcleo del sintagma nominal marca la diferencia entre anáfora y elipsis. A nuestro parecer, esta traslación de la one-anaphora al español es una elipsis y no una anáfora.

(26) De entre los asistentes destacaba una *joven*$_i$ con rasgos orientales. *La joven*$_i$ parecía ausente.

(27) Luis tiene una *empresa*$_i$ de exportación. *La compañía*$_i$ cuenta con 200 empleados.

2. SN con determinante demostrativo
Como en el caso anterior, puede comprobarse la relación semántica existente entre antecedente y anáfora. En (29) la relación definida es de hiperonimia/hiponimia: "*bambú* ***es una*** *planta*"

(28) De entre los asistentes destacaba una *joven*$_i$ con rasgos orientales. *Esta joven*$_i$ parecía ausente.

(29) El *bambú*$_i$ es la base de nuestros productos. Oriente nos proporciona *esta planta*$_i$.

3. SN con determinante posesivo
Es importante mencionar que en el caso del posesivo, a diferencia de los anteriores, el elemento anafórico siempre será una entidad perteneciente a (poseída por) su antecedente. Esta situación condiciona el tipo de relación semántica existente entre antecedente y anáfora, quedando excluidas relaciones como la sinonimia. En (20) la relación existente es de meronimia/holonimia: "*salón* ***es parte de*** *casa*".

(30) De entre los asistentes destacaba una *joven*$_i$. *Su indumentaria*$_i$ era muy llamativa.

(31) La *casa*$_i$ de *María*$_j$ es enorme. *Su salón*$_i$ tiene 30 metros cuadrados.

Las relaciones semánticas existentes entre la anáfora de sintagma nominal y su antecedente definen una diferenciación entre anáforas directas e indirectas, que, si bien no responde a criterios basados en la categoría gramatical del elemento anafórico,

requiere una mención por su interés desde el punto de vista de la información semántica. La distinción entre anáforas directas e indirectas se basa en que los núcleos de la anáfora y del antecedente sean iguales o no. Así, en (32), antecedente y anáfora coinciden en el núcleo *casa*, mientras que en (33), sendos núcleos (*casa* y *piso*) mantienen una relación de sinonimia entre sí que les hace correferentes.

(32) Luis e Isabel están reformando *su casa*$_i$. *La casa*$_i$ es muy pequeña para los dos.

(33) Luis e Isabel están reformando *su casa*$_i$. *El piso*$_i$ es muy pequeño para los dos.

Anáfora superficial numérica. Como se apuntaba al principio de esta sección, si bien este tipo de anáfora puede ser incluido en cualquiera de los dos grupos anteriores, al poder estar representado tanto por un adjetivo sustantivado como por un pronombre, tiene entidad suficiente para ser tratado de forma independiente. En la medida en que este tipo de anáfora alude al orden establecido por sus antecedentes, en (21) la resolución pasa únicamente por la elección del primero de los antecedentes enumerados. Sin embargo, en (35) es necesario el conocimiento del mundo para extraer, de todas las ciudades mencionadas, aquellas que son españolas.

(34) *Luis*$_i$ y *Mariano*$_j$ tienen una tienda. *El primero*$_i$ trabaja sólo por la mañana.

(35) *Roma*$_j$, *Milán*$_k$, *Madrid*$_m$, *Barcelona*$_i$ y *París*$_n$ presentan sus colecciones de otoño. *La segunda de las ciudades españolas*$_i$ amplía el número de diseñadores.

Los pronombres distributivos pueden tener, como se ve en (36), una función anafórica próxima a la anáfora superficial numérica y desde ese punto de vista pueden incluirse en este grupo, dado que implican un orden sin referirse a él de manera numérica.

(36) *Alumnos*$_i$ y *profesores*$_j$ comparten la misma opinión. *Los unos*$_i$ la defienden desde sus pupitres y *los otros*$_j$ lo hacen desde la tarima.

En ese mismo sentido podríamos referirnos a usos anafóricos del pronombre demostrativo, como en (37).

(37) Los *rusos*$_i$ y los *americanos*$_j$ han llegado a la luna. *Éstos*$_i$ lo hicieron en 1969 y *aquéllos*$_j$ poco tiempo después.

Anáfora verbal. En la anáfora verbal la forma pronominal *lo* se refiere a un verbo o a un sintagma verbal (sin complemento directo) al que se alude mediante un verbo auxiliar o similar (pro-verbo). Así, en (38), *lo hagas* representa a *fumar*. Este verbo anafórico no proporciona rasgos semánticos específicos, por lo que la aplicación de esta fuente de información no resulta especialmente útil para su resolución.

(38) No se puede *fumar*$_i$ en este recinto, así que no *lo hagas*$_i$.

Como caso de anáfora verbal, cabe citar en este apartado la denominada anáfora de complemento nulo (*null complement anaphora*), en la que elipsis y anáfora coinciden: el núcleo del SV está ocupado por un verbo en forma personal que selecciona una oración de infinitivo elíptica, cuyo contenido está presente en el contexto anterior (Brucart99), como en "*Le gusta bailar, pero no sabe (∅ = bailar)*".

Anáfora adverbial. Dividimos este grupo de anáforas en temporales y locativas según la circunstancia temporal (39) o espacial (40) descrita por el antecedente. Como en el caso de la anáfora verbal, la información semántica contenida en estos adverbios es muy general y su incorporación es costosa y no facilita la resolución de la anáfora.

(39) No acabaré mis estudios hasta *el año que viene*$_i$. *Entonces*$_i$ haré unas prácticas en una empresa.

(40) Frente a la *oficina*$_j$ hay un *taller*$_i$. *Ahí*$_i$ encontrarás los recambios para tu coche.

2.3 Ámbito del presente trabajo

De acuerdo con lo expuesto en la sección anterior, y en referencia a la relación existente entre la anáfora y su antecedente, el interés de este trabajo se centra en la anáfora de referencia o anáfora profunda. Respecto a la categoría gramatical del antecedente, se tratarán únicamente aquellas anáforas que hacen referencia a un sintagma nominal. De esta manera, en la propuesta de resolución de este trabajo, el conjunto de potenciales antecedentes de una anáfora vendrá representado por una lista de nombres correspondientes a los núcleos de los sintagmas nominales candidatos.

En lo referente a la categoría gramatical de la expresión anafórica, esta Tesis se circunscribe a la anáfora pronominal. En particular, se tratarán las anáforas generadas por pronombres personales, demostrativos, reflexivos y omitidos, todas ellas de tercera persona.

3. Trabajos sobre la resolución de la anáfora

La resolución de la anáfora ha sido durante las últimas dos décadas una preocupación de lingüistas e informáticos. Esta tarea, considerada por muchos como una de las más importantes dentro del tratamiento de la ambigüedad en el Procesamiento del Lenguaje Natural, ha sido abordada desde distintos puntos de vista por los sistemas más variados.

Realizar una clasificación de estos sistemas no es una tarea fácil, ya que muchos de ellos han desarrollado estrategias combinadas para mejorar sus resultados. Dado que el trabajo aquí presentado plantea la incorporación de fuentes de información de carácter semántico para la resolución de la anáfora, esta clasificación distribuye estos trabajos en tres grandes grupos:

- *Métodos de conocimiento limitado*: aproximaciones que resuelven la anáfora con el uso de información morfológica y/o sintáctica.
- *Métodos enriquecidos*: estrategias que incorporan, junto a las anteriores, fuentes de información adicional como la semántica (bien basada en etiquetados o en el uso de ontologías) o la pragmática (a través del análisis del discurso o el conocimiento del mundo).
- *Métodos alternativos*: este grupo incluye aquellas estrategias no catalogadas en los dos anteriores. Usan técnicas basadas en la estadística o modelos de inteligencia artificial.

Tal y como se ha señalado anteriormente, algunos de los trabajos podrían encajar en más de uno de los grupos definidos, si bien se ha elegido la característica más relevante de cada estrategia para su clasificación.

3.1 Métodos de conocimiento limitado

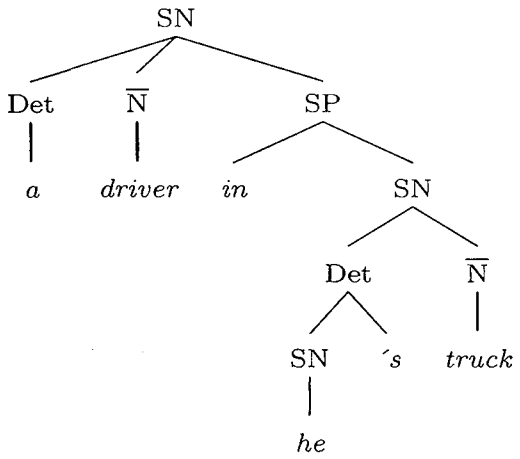
El paralelismo morfosintáctico existente entre la expresión anafórica y el antecedente ha sido usado tradicionalmente como uno de los principales recursos en la resolución de la anáfora. Los sistemas presentados en esta sección tratan de resolver la anáfora bien con mecanismos pobres en conocimiento (algunos de ellos no usan análisis sintáctico) o bien con el uso del paralelismo sintáctico a partir de análisis parciales o completos. Son, en definitiva, propuestas que utilizan información morfológica y sintáctica y que resultan de interés por su bajo coste computacional al proporcionar interesantes resultados que, en casos como el algoritmo clásico de Hobbs, han sido difíciles de superar.

3.1.1 El algoritmo clásico de Hobbs

Hobbs (1976, 1978) plantea uno de los primeros y más importantes métodos para la resolución de la anáfora. Realiza dos enfoques del problema. El primero de ellos, el que nos ocupará en esta sección y que ha convertido este algoritmo en uno de los más importantes y referenciados de la historia, es el que Hobbs llama *algoritmo ingenuo (naif) de resolución de la anáfora.* Esta aproximación utiliza conocimiento morfosintáctico para la selección del antecedente correcto de una anáfora producida por un pronombre personal. El conocimiento sintáctico queda representado por árboles de análisis de superficie que definen perfectamente la estructura sintáctica de la oración. Hobbs plantea que estas representaciones eliminan determinada ambigüedad sintáctica, tal y como muestran los siguientes ejemplos extraídos de Hobbs (1978):

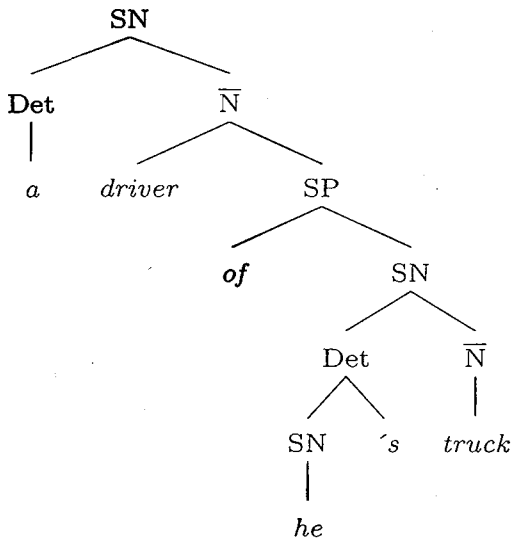
(41) *Mr. Smith saw a driver* ***in*** *his truck.*

El Sr. Smith vio a un conductor **en** su camión.



(42) *Mr. Smith saw a driver* ***of*** *his truck.*

El Sr. Smith vio a un conductor **de** su camión.



Tal y como muestran los árboles de análisis para cada ejemplo, en (41), el posesivo *his* (*su*) parece referirse a *driver* (*conductor*), mientras que en (42) podría no hacerlo.

El algoritmo planteado por Hobbs recorre el árbol de análisis buscando un sintagma nominal (SN) con el género y el número adecuados. La búsqueda se realiza siguiendo los siguientes pasos:

1. Comienza por el sintagma nominal (SN) que domina de forma más inmediata al pronombre.
2. Sube por el árbol al primer nodo del sintagma nominal (SN) u oración (S) encontrado. Llama X a este nodo y p al camino utilizado para llegar a él.

3. Recorre todas las ramas por debajo del nodo X a la izquierda del camino p con un recorrido por niveles de izquierda a derecha y de arriba a abajo. Propone como antecedente cualquier SN que tenga un nodo SN o S entre él y p.
4. Si el nodo X es el nodo más alto de la oración, recorre los árboles de las oraciones anteriores en el texto de la más reciente hacia atrás. Cada árbol se recorre por niveles de izquierda a derecha y de arriba a abajo y cuando se encuentra un SN, se propone como antecedente. Si X no es el nodo más alto de la oración, se continúa con el paso 5.
5. Desde el nodo X, sube por el árbol hasta el primer SN o S encontrado. Llama X a este nuevo nodo y p al camino atravesado hasta llegar a él.
6. Si X es un SN y si el camino p a X no pasa a través del nodo $\overline{N}$ que domina inmediatamente X, propone X como el antecedente.
7. Recorre todas las ramas por debajo de X a la izquierda del camino p por niveles de izquierda a derecha y de arriba a abajo. Propone como antecedente cualquier SN encontrado.
8. Si X es un nodo S, recorre todas las ramas de X a la derecha del camino p por niveles, de izquierda a derecha y de arriba a abajo, sin llegar a ir por debajo de cualquier SN o S encontrado. Propone como antecedente cualquier nodo SN encontrado.
9. Vuelve al paso 4.

Gráficamente, y siguiendo una vez más el ejemplo propuesto por el autor, la figura 3.1 ilustra el recorrido realizado por el algoritmo en la oración (43):

(43) *The Castle in Camelot remained the* residence$_i$ *of the king until 536 when he moved* it$_i$ *to London.*

El Castillo de Camelot siguió siendo la *residencia*$_i$ del rey hasta el 536 cuando él *la*$_i$ trasladó a Londres.

Adicionalmente, Hobbs incorpora algunas restricciones de selección del tipo 'las fechas no se mueven', 'los lugares no se mueven' o 'los objetos muy grandes no se mueven'. Este tipo de res-

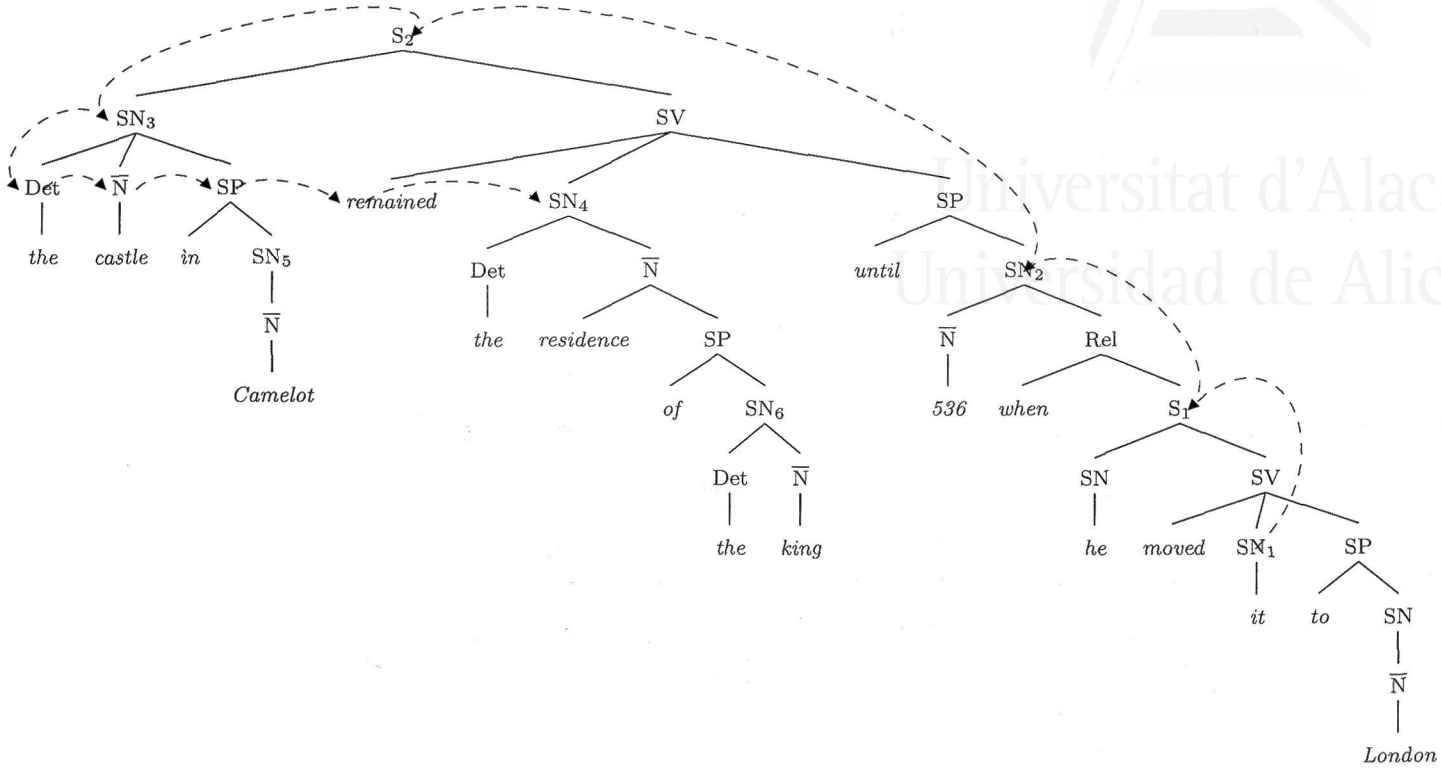


Figura 3.1. Ejemplo de recorrido de árbol sintáctico en el algoritmo de Hobbs (1978) para el ejemplo (43)

tricciones evitan que se escoja *536* o *the castle* (*el castillo*) como antecedentes.

Hobbs utiliza también dos condiciones de no correferencia propuestas por Langacker (1969):

- Un pronombre no reflexivo y su antecedente no pueden aparecer en la misma oración simple[1].
- El antecedente de un pronombre debe preceder o dominar al pronombre.

El sistema de Hobbs fue evaluado sobre un corpus compuesto por tres textos procedentes de un libro de arqueología, una novela y una publicación semanal con un total de 300 pronombres (100 pronombres en cada texto). El algoritmo obtiene índices de éxito del 88,3 % y afirma que aumenta hasta el 91,7 % con la incorporación de restricciones de selección como las mencionadas anteriormente. En los datos de su evaluación, el autor afirma que más de la mitad de las anáforas tienen un único antecedente posible, con lo que hace un cálculo adicional del sistema aplicado a las anáforas con más de un antecedente. De un total de 132 pronombres, las restricciones de selección resuelven 12 y el algoritmo resuelve 96, lo que hace un total de 108 anáforas resueltas, es decir, un 81,8 % de tasa de éxito. En cuanto a los resultados obtenidos, es muy importante tener en cuenta que éstos proceden de una evaluación manual del sistema que parte de un análisis perfecto del texto tratado, con lo que el porcentaje de error sólo puede ser atribuido a las características propias del sistema y no a errores de etapas de preproceso y análisis previos.

En cualquier caso, sea cual sea el sistema de evaluación elegido, este algoritmo proporciona un enfoque simple pero de una gran eficacia, que lo ha convertido, a lo largo de los años, en un clásico dentro de los sistemas de referencia y comparación de aproximaciones a la resolución de la anáfora (Walker, 1998; Dagan y Itai,

[1] El concepto de oración simple, tal y como se entiende por su autor, coincide con el concepto de cláusula, cuya definición será fundamental en el método propuesto en esta Tesis. Entendemos por cláusula, y así lo haremos a lo largo de todo este trabajo, toda estructura oracional introducida por un único verbo (en forma personal). De esta manera, la diferencia conceptual entre oración y cláusula es que la primera podrá contener tantas unidades de la segunda como verbos existan en ella.

1991; Lappin y Leass, 1994; Baldwin, 1997; Ge et al., 1998; Byron y Allen, 1999; Tetreault, 1999; Ge, 2000; Palomar et al., 2001a).

3.1.2 El algoritmo de Lappin y Leass basado en la sintaxis

Lappin y Leass (1994) definen un algoritmo basado en información exclusivamente morfo-sintáctica para la resolución de los pronombres de tercera persona y las anáforas reflexivas y recíprocas cuyos antecedentes son sintagmas nominales.

El algoritmo *RAP* (*Resolution of Anaphora Procedure* – Procedimiento de Resolución de la Anáfora –) trabaja sobre representaciones sintácticas generadas con el analizador sintáctico basado en gramáticas de huecos de McCord (1990, 1993) y selecciona el antecedente correcto de un pronombre a partir de medidas de relevancia derivadas de la estructura sintáctica.

El algoritmo *RAP* incorpora:

- Dos filtros para eliminar aquellos antecedentes con incompatibilidad morfológica (género, número y persona) y sintáctica.
- Un procedimiento que identifica pronombres no anafóricos (pleonásticos).
- Un algoritmo de enlace anafórico para determinar el antecedente de un pronombre reflexivo o recíproco dentro de la misma oración.
- Un procedimiento que asigna valores a distintos parámetros como el rol gramatical, el paralelismo de roles gramaticales, la frecuencia de aparición o la proximidad. De esta manera, se asignan pesos de importancia a los candidatos para que posteriormente un procedimiento de decisión seleccione el elemento preferido de la lista. Se dota de mayor importancia (peso) a los sintagmas nominales con función de sujeto (frente a los que no la tienen), a objetos directos (frente a otros complementos), a argumentos de un verbo (frente a adjuntos y objetos de sintagmas preposicionales del verbo) y a núcleos del sintagma nominal (frente a complementos del núcleo).
- Un procedimiento para identificar sintagmas nominales enlazados anafóricamente como una clase de equivalencia para la que

el valor de importancia se calcula como la suma de los valores de importancia de sus elementos.

- Un procedimiento de selección del elemento preferido de una lista de candidatos.

La propuesta de Lappin y Leass es una de las referencias más importantes del trabajo realizado en esta Tesis, en particular en lo referente al sistema de restricciones sintácticas y morfológicas tanto del método de conocimiento limitado como del método enriquecido. Es por ello que estudiaremos con más detenimiento todos y cada uno de los elementos que conforman el sistema de Lappin y Leass.

Los filtros morfosintácticos. Los filtros morfosintácticos de correferencia entre un pronombre y un sintagma nominal se componen de seis condiciones de no-correferencia dentro de una oración. Para definir estos filtros llamaremos P al pronombre y SN al sintagma nominal. Asimismo, ilustraremos las condiciones con los mismos ejemplos que proporcionan los autores utilizando subíndices para expresar la existencia o no de correferencia, siendo dos elementos correferentes o no según sus índices coinciden o no. Las seis condiciones de no-correferencia son:

1. P y SN tienen características morfológicas (género, número y persona) incompatibles.

   (44) *The* woman$_i$ *said that* he$_j$ *is funny.*
   La *mujer*$_i$ dijo que *él*$_j$ es divertido.

2. P está en el dominio de argumentos[2] de SN.

   (45) She$_i$ *likes* her$_j$.
   Ella$_i$ *la*$_j$ ama.

   (46) John$_i$ *seems to want to see* him$_j$.
   John$_i$ parece querer ver-*le*$_j$.

[2] P está en el dominio de argumentos de N si y sólo si P y N son argumentos del mismo núcleo.

3. P está en el dominio de adjuntos[3] de SN.

(47) She$_i$ *sat near* her$_j$.

Ella$_i$ se sentó cerca de *ella*$_j$.

4. P es un argumento del núcleo H, SN no es un pronombre y SN está contenido[4] en H.

(48) He$_i$ *believes that the* man$_j$ *is amusing.*

Él$_i$ cree que el *hombre*$_j$ es divertido.

(49) *This is the* man$_i$ he$_j$ *said* John$_k$ *wrote about.*

Éste es el *hombre*$_i$ sobre el que *él*$_j$ dijo que *John*$_k$ escribió.

5. P está en el dominio de sintagma nominal[5] de SN.

(50) John$_i$*'s portrait of* him$_j$ *is interesting.*

El retrato de *John*$_i$ de *él*$_j$ es interesante.

6. P es un determinante de un nombre Q, y SN está contenido en Q.

(51) His$_i$ *portrait of* John$_j$ *is interesting.*

Su$_i$ retrato de *John*$_j$ es interesante.

(52) His$_i$ *description of the portrait by* John$_j$ *is interesting.*

Su$_i$ descripción del retrato de *John*$_j$ es interesante.

Identificación de pronombres no anafóricos (pleonásticos). La identificación del *it* pleonástico se realiza con un procedimiento de carácter tanto sintáctico como léxico. Los autores definen, por un lado, un conjunto de adjetivos modales (*neccesary*,

[3] P está en el dominio de adjuntos de N si y sólo si N es un argumento de un núcleo H, P es el objeto de una preposición PREP y PREP es un adjunto de H.

[4] P está contenido en Q si y sólo si *a)* P es un argumento o un adjunto de Q, es decir, P está contenido inmediatamente en Q, o *b)* P está inmediatamente contenido en R y R está contenido en Q.

[5] P está en el dominio de sintagma nominal de N si y sólo si N es el determinante de un nombre Q y *a)* P es el argumento de Q o *b)* P es el objeto de una preposición PREP y PREP es el adjunto de Q.

important, *desirable*, ...) y, por otro, un conjunto de verbos cognitivos (*recommend*, *assume*, *expect*, ...). El procedimiento utiliza unas reglas estructurales que hacen uso de estos conjuntos para determinar si un pronombre es anafórico o no. Por ejemplo, la construcción "*It is [adjetivo modal] that [oración]*" indica que el pronombre *it* es pleonástico.

Posibles antecedentes de pronombres reflexivos y recíprocos. Para la identificación de posibles antecedentes de pronombres reflexivos y recíprocos en la misma oración, *RAP* incorpora un mecanismo de enlace anafórico basado en la siguiente jerarquía de argumentos:

sujeto > agente(pasiva) > O.D. > O.I. y circunstancial

A partir de esta jerarquía y las definiciones dadas anteriormente sobre dominio de argumentos, dominio de adjuntos y dominio de sintagma nominal, un sintagma nominal N es un posible antecedente de una pronombre reflexivo o recíproco A si no tiene características morfológicas incompatibles y se da una de las siguientes condiciones[6]:

1. A está en el dominio de argumentos de N y N ocupa una posición superior a la de A en la jerarquía de argumentos.

   (53) They$_i$ *wanted to see* themselves$_i$.

   Ellos$_i$ querían ver-*se*$_i$.

   (54) *Mary knows the* people$_i$ *who John introduced to* each other$_i$.

   María conoce a *la gente*$_i$ que John presentó *entre sí*$_i$.

2. A está en el dominio de adjuntos de N.

[6] Dadas las diferencias de uso del pronombre reflexivo o recíproco en inglés y en español, algunas de las traducciones de los siguientes ejemplos, por intentar representar el carácter reflexivo o recíproco de los ejemplos originales, pueden resultar algo forzadas tanto gramatical como estilísticamente.

(55) He$_i$ *worked by* himself$_i$.

Él$_i$ trabaja por *sí mismo*$_i$.

(56) *Which* friends$_i$ *plan to travel with* each other$_i$*?*

¿Qué *amigos*$_i$ planean viajar *unos con otros*$_i$?

3. A está en el dominio de sintagma nominal de N.

(57) *John likes* Bill$_i$*'s portrait of* himself$_i$.

A Juan le gusta el retrato de *Bill*$_i$ de *sí mismo*$_i$.

4. N es un argumento del verbo V, existe un sintagma nominal Q en el dominio de argumentos o de adjuntos de N de tal manera que Q no tiene ningún determinante nominal y *a)* A es un argumento de Q o *b)* A es un argumento de una preposición $PREP$ y $PREP$ es un adjunto de Q.

(58) They$_i$ *told old stories about* themselves$_i$.

Ellos$_i$ contaron viejas historias de *ellos mismos*$_i$.

5. A es un determinante de un nombre Q y *a)* Q está en dominio de argumentos de N y N ocupa una posición superior a la de Q en la jerarquía de argumentos, o *b)* Q está en el dominio de adjuntos de N.

(59) John and Mary$_i$ *like* each other$_i$*'s portraits.*

A *John y Mary*$_i$ les gustan los retratos al *uno del otro*$_i$.

Pesos de importancia. *RAP* define un conjunto de propiedades o factores de importancia a los que se les asigna un peso. Cada uno de estos factores contribuirá al peso total de cada uno de los candidatos. Así, se le da un mayor peso a los candidatos que están en la misma oración que el pronombre (100), a los sujetos (80), a los predicados nominales que se encuentran en estructuras existenciales (75), a los objetos directos (50), a los objetos indi-

rectos y complementos oblicuos[7] (40), a los sintagmas nominales no contenidos en otro sintagma nominal (80) y a los sintagmas nominales no contenidos en un sintagma adverbial (50).

Los pesos de cada factor, correspondientes a los valores que aparecen entre paréntesis en el párrafo anterior, han sido definidos experimentalmente por los autores.

Estos pesos, en el proceso de resolución de la anáfora, pueden ser alterados en función de distintos criterios. Así, la catáfora está fuertemente penalizada, por lo que el peso de importancia de un candidato que está después del pronombre es reducido sustancialmente. Por otro lado, si el candidato tiene el mismo papel sintáctico que el pronombre, su peso aumenta.

Asimismo, se determina un umbral, de manera que cualquier candidato cuyo peso no lo supere será rechazado.

Sintagmas nominales enlazados anafóricamente. Se define también un conjunto de clases de equivalencia de candidatos, es decir, un conjunto de cadenas anafóricas o cadenas de correferencia. En estas clases de equivalencia quedan agrupados todos aquellos candidatos que hacen referencia al mismo elemento del discurso. Cada clase de equivalencia (que puede estar formada por un único elemento) lleva asociada un peso que resulta de la suma de los pesos de aquellos factores de importancia que cumplen al menos un elemento de la clase.

Estas clases de equivalencia constituyen un mecanismo dinámico de la importancia de los sintagmas nominales en el texto.

Selección del candidato preferido. Para la selección del candidato antecedente, *RAP* aplica los filtros sintácticos y los factores de importancia, aumentando o disminuyendo el peso de cada candidato en función de los criterios detallados anteriormente.

Para las anáforas reflexivas y recíprocas aplica el algoritmo de enlace anafórico ya detallado. Para los pronombres de tercera

[7] Se entiende por complemento oblicuo aquel sintagma nominal que es complemento de una preposición.

persona, tras la aplicación del método, escoge como el antecedente de la anáfora aquel con el mayor peso de todos .

Cuando existen candidatos con el mismo peso, se prefieren aquellos que se encuentran en la misma oración. Los valores de importancia de los candidatos en las oraciones anteriores se degradan progresivamente en favor de los de la oración actual. Ante un caso de "empate" entre candidatos, se escogerá aquel más cercano al pronombre.

La evaluación. *RAP* fue entrenado con un corpus compuesto por cinco manuales de informática con aproximadamente 82000 palabras. Se extrajeron 560 pronombres de tercera persona (reflexivos y recíprocos incluidos) y sus correspondientes antecedentes. Manualmente, el sistema se entrenó para determinar el valor más adecuado de los factores de importancia.

Una vez entrenado, el algoritmo fue evaluado sobre 360 pronombres, seleccionados aleatoriamente del corpus de manuales de informática, anteriomente mencionado, con 1.25 millones de palabras. *RAP* proporcionó un índice de éxito del 86 %, con un 72 % para los 70 casos intersentenciales y un 89 % para los restantes 290 casos intrasentenciales. Asimismo, realizaron otros muchos experimentos activando y desactivando algunos de los factores de importancia para evaluar sus repercusiones, determinando que unos de los que más influencia tenían en la correcta resolución de la anáfora (aproximadamente un 20 %) eran los factores que relacionan la cercanía oracional entre pronombre y antecedente.

Comparación con otros trabajos. Lappin y Leass comparan *RAP* con el algoritmo de Hobbs. Dado que el algoritmo de Hobbs no las resuelve, se excluyen del experimento las anáforas reflexivas y recíprocas y los pronombres pleonásticos.

En la comparación, el algoritmo de Hobbs demuestra resolver con mayor éxito las anáforas intersentenciales (un 87 % frente al 74 % de *RAP*). Sin embargo, el hecho de que *RAP* obtenga mejores resultados en las anáforas intrasentenciales (un 89 % frente al 81 % de Hobbs) y de que el número de anáforas intersentenciales

sea muy bajo en el corpus, hace que el factor de éxito global de *RAP* sea superior al de Hobbs en aproximadamente un 4 %.

Uno de los aspectos más interesantes de esta comparación es la reflexión que los autores hacen sobre el comportamiento de ambos algoritmos. Existe un alto grado de convergencia entre ambos algoritmos, a pesar de que las estrategias son muy diferentes. Esto es debido a que en inglés los papeles sintácticos pueden ser identificados a través del orden oracional. Los autores afirman que, por el contrario, para idiomas de orden libre, como el español, existirá una clara divergencia en el comportamiento de ambos algoritmos.

Los patrones de Dagan en el *RAP*. Dagan (1992) incorpora al *RAP* un procedimiento de patrones similar a otro propuesto anteriormente por Dagan y Itai (1990)[8] que asigna estadísticamente un valor a los patrones de co-ocurrencia de nombres y verbos en un corpus. Este sistema, denominado *RAPSTAT*, permite resolver anáforas que *RAP* no resolvía correctamente. Veamos el siguiente ejemplo propuesto por el autor:

(60) *The Send Message display is shown, allowing you to enter your* message$_i$ *and specify where* it$_i$ *will be sent.*

El indicador de Enviar Mensaje se muestra, permitiéndole introducir el *mensaje*$_i$ y especificar dónde será enviado *éste*$_i$.

RAP asigna a los dos posibles antecedentes del pronombre *it* (*éste*), *display* (*indicador*) y *message* (*mensaje*), un peso de 345 y 315 respectivamente. Por otro lado, en el corpus usado por *RAPSTAT*, el par verbo-objeto *display-send* (*indicador-enviar*) aparece una sola vez, mientras que el par *message-send* aparece 289 veces, con lo que éste patron consigue una puntuación estadística considerablemente mayor. De esta forma, mientras que *RAP* elige el candidato incorrecto, *RAPSTAT* resuelve la anáfora correctamente.

En la comparación entre *RAP* y *RAPSTAT*, el segundo proporciona un porcentaje de éxito del 89 %, aproximadamente un

[8] Este trabajo se describe en profundidad en 3.3.1.

3 % superior al primero. En un total de 41 casos, ambos sistemas discrepan en la solución, siendo la correcta la proporcionada por *RAPSTAT* en un 61 % de los casos y resolviéndola correctamente *RAP* en el restante 39 %.

Tal y como se ha indicado previamente, el trabajo de Lappin y Leass ha servido de inspiración en la definición de las estrategias de resolución planteadas en esta Tesis, especialmente en lo referente al sistema de restricciones y preferencias propuesto en el capítulo 4.

3.1.3 La resolución de Kennedy y Boguraev sin análisis sintáctico completo

A partir del algoritmo de Lappin y Leass (1994) anteriormente descrito, que hace uso de un análisis sintáctico completo, la propuesta de Kennedy y Boguraev (1996) usa únicamente la salida de un etiquetador de categorías gramaticales enriquecida con algunas anotaciones de papel sintáctico[9].

Uno de los principales objetivos perseguido por los autores en el desarrollo de un sistema de resolución de la anáfora a partir de un análisis no completo viene dado, precisamente, por las limitaciones tecnológicas del análisis sintáctico, que, tanto en el momento en el que se ubica el trabajo de los autores como en el momento actual, sigue sin proporcionar una salida lo suficientemente robusta y fiable. Por otra parte, este enfoque permite la aplicación de la resolución de la anáfora en un entorno de trabajo más general, que no incluya necesariamente análisis completo.

El sistema de Kennedy y Boguraev identifica los sintagmas nominales a través de un conjunto de reglas gramaticales que definen la composición de un SN y, de la misma forma que el de Lappin y Leass, elimina en primer lugar aquellos candidatos que no pueden correferir con el pronombre, bien por restricciones morfológicas (concordancia de género y número) o bien por restricciones sintácticas. Este último es uno de los puntos que más le diferencian del sistema de Lappin y Leass. Dado que no cuentan con

[9] El concepto de papel sintáctico, denominado por los autores función gramatical de elementos léxicos, será un concepto fundamental en el desarrollo de esta Tesis.

análisis completo, y por tanto no pueden aplicar el filtro sintáctico intrasentencial del *RAP*, usan tres condiciones sintácticas de no correferencia:

1. Un pronombre no puede correferir con un co-argumento: se eliminan todos los complementos directos e indirectos que siguen a un pronombre identificado como sujeto u objeto (se supone que el sujeto marca el comienzo de la siguiente claúsula).
2. Un pronombre no puede correferir con un constituyente no pronominal al que domina y precede: se eliminan los referentes no pronominales que están en la misma oración que el pronombre y le siguen. La relación de dominio se indica por la relación de precedencia y por el entorno sintáctico (un argumento que no está contenido en un adjunto o incluido en otro sintagma nominal domina a aquellas expresiones que precede).
3. Un pronombre no puede correferir con un constituyente que lo contiene: esta restricción elimina la correferencia entre un pronombre posesivo[10] y el sintagma nominal que modifica.

A partir de la lista de candidatos reducida tras la aplicación de los filtros morfológico y sintáctico, y de forma muy similar al *RAP*, el algoritmo incrementa o reduce el valor de importancia de cada candidato en función de su proximidad, situación o paralelismo con respecto a la expresión anafórica. El candidato con el mayor valor de relevancia es el elegido como antecedente. En caso de "empate", se escoge el más cercano.

Es importante destacar el hecho de que el sistema propuesto por Kennedy y Boguraev no plantea simplemente un "recorte" del *RAP* en lo referente al tipo de análisis, sino que más bien supone una extensión del mismo, incorporando algunos factores de importancia propios. Estos factores dan pesos de importancia

[10] Es muy importante, en este punto, mencionar que, si bien los autores hablan de pronombres posesivos, en realidad tratan los adjetivos posesivos como pronombres, tal y como se puede comprobar en el ejemplo extraído del propio artículo (Kennedy y Boguraev, 1996):

(61) *For 1995, the company set up* its *headquarters in Hall 11...*
En 1995, la compañía establece *su* cuartel general en el Hall 11...

a los candidatos con función gramatical de posesivo y también a aquellos que aparecen en el mismo segmento de discurso de la anáfora[11].

El cuadro 3.1 muestra una comparación entre los factores de importancia usados por Lappin y Leass y los utilizados por Kennedy y Boguraev, con sus pesos iniciales asociados.

Lappin y Leass		*Kennedy y Boguraev*	
Misma oración	100	Misma oración	100
		Contexto	50
Sujeto	80	Sujeto	80
Estructura existencial	70	Estructura existencial	70
		Posesivo	65
Objeto directo	50	Objeto directo	50
Objeto indirecto y oblicuo	40	Objeto indirecto	40
		Complemento oblicuo	30
Sintagma nominal	80	Sintagma nominal	80
Sintagma adverbial	50	Sintagma adverbial	50

Cuadro 3.1. Comparación entre factores de importancia de los trabajos de Lappin y Leass (1994) y Kennedy y Boguraev (1996)

En lo referente a la evaluación de su sistema, Kennedy y Boguraev utilizaron un conjunto de 27 textos de distinta índole, seleccionados aleatoriamente de recortes de prensa, publicidad, artículos de revista y otros documentos disponibles en la red. Los textos contenían un total de 306 pronombres[12].

De los 306 pronombres, 231 fueron resueltos correctamente, lo que supone un índice de éxito del 75 %. Si bien el resultado obtenido es inferior al proporcionado por el método de Lappin y Leass, los autores ponen de manifiesto el hecho de que dicha evaluación se ha realizado sobre un conjunto de textos muy variado, mientras que Lappin y Leass efectúan su evaluación sobre manuales de informática, textos mucho más estables, poniendo en duda la capacidad de *RAP* para conseguir los mismos resultados en textos menos normalizados.

[11] Este segmento de discurso se calcula mediante el algoritmo de segmentación definido por Hearst (1994).

[12] En realidad, se eliminaron manualmente un total de 30 pronombres *it* no anafóricos (pleonásticos) no detectados por el sistema y otros 6 más que hacían referencia a sintagmas verbales.

En lo referente al análisis de fallos, Kennedy y Boguraev revelan un 35 % de errores debidos a problemas de incompatibilidad de género[13] y un 14 % debido al estilo indirecto usado en algunos pasajes.

Uno de los puntos de divergencia entre el estudio de Kennedy y Boguraev (1996), por una parte, y los de Lappin y Leass (1994) y Dagan et al. (1995), por otra, es que los primeros hablan de una reducida importancia de los filtros sintácticos en la resolución de la anáfora (sólo dos de los 75 errores), mientras que los demás sugieren una relevancia mucho mayor.

3.1.4 El sistema CogNIAC de Baldwin

Baldwin (1997) presenta el sistema CogNIAC para la resolución de la correferencia con el uso de recursos y conocimiento limitados. CogNIAC es un sistema que, a diferencia de otros, no resuelve el pronombre en caso de ambigüedad, es decir, cuando no está lo suficientemente seguro del antecedente propuesto. Si el sistema no devuelve un único candidato como solución, la respuesta se considera ambigua y el pronombre no resuelto. Esto da como resultado un sistema de gran precisión, pero de baja cobertura.

El algoritmo se basa en la salida de un etiquetador de categorías gramaticales para identificar los sintagmas nominales simples. Con un conjunto de expresiones regulares, identifica sujeto, verbo y objeto de las cláusulas definidas manualmente.

La resolución de los pronombres se efectúa de izquierda a derecha en el texto. Para cada pronombre se aplica un conjunto de reglas en el orden expuesto a continuación. Cada regla, tal y como se va a enunciar, va acompañada por el número entre paréntesis de pronombres correcta e incorrectamente resueltos, respectivamente, en un corpus de entrenamiento con un total de 200 pronombres:

[13] Este asunto será tratado en el sistema propuesto en esta Tesis. De hecho, una de las restricciones morfológicas incluye un conjunto de factores morfo-semánticos que intentan eliminar este problema. Ver apartado 4.3.8 (pág. 137).

1. Si existe un único candidato posible[14], se escoge como antecedente (8,0).
2. Si el pronombre es reflexivo, se escoge el candidato posible más cercano en la oración actual (16,1).

   (62) *Mariana motioned for* Sarah$_i$ *to seat* herself$_i$ *on a two-seater lounge.*

   Mariana hizo señas a *Sarah*$_i$ para que *se*$_i$ sentara en un asiento de dos plazas.

3. Si es el único candidato posible en las oraciones anterior y actual, se escoge como antecedente (114,2).

   (63) Rupert Murdock$_i$*'s News Corp. confirmed his interest in buying back the ailing New York Post. But analysts said that if* he$_i$ *winds up bidding for the paper...*

   La News Corp. de *Rupert Murdock*$_i$ confirmó su interés por comprar de nuevo al "enfermo" New York Post. Pero los analistas dijeron que si *él*$_i$ cierra la oferta para el periódico...

4. Si el pronombre es posesivo y hay una expresión coincidente en la anterior oración, se escoge como antecedente (4,1).

   (64) *After he was dry, Joe carefully laid out the damp towel in front of* his locker$_i$*. Travis went over to* his locker$_i$*, took out a towel and started to dry off.*

   Cuando estuvo seco, Joe puso cuidadosamente la toalla mojada frente a *su taquilla*$_i$. Travis cruzó hacia *su taquilla*$_i$, sacó una toalla y comenzó a secarse.

5. Si sólo hay un candidato posible en la oración actual, se elige como antecedente (21,1).

[14] Los autores entienden por candidato posible aquel que es compatible, tanto morfológicamente (género y número) con la anáfora como con las restricciones de correferencia (es decir, los pronombres no reflexivos no pueden correferir con otros argumentos de su verbo/preposición, etc.).

(65) *After a week* Constantin$_i$ *tired of reading the old novels in the bottom shelf of the bookcase –somewhere among the gray well thumbed pages* he$_i$ *had hoped to find a message of one of his predecessors...*

Después de una semana *Constantin*$_i$ cansado de leer las viejas novelas del estante inferior de la estantería –en algún sitio entre las páginas grises bien manoseadas *él*$_i$ había esperado encontrar un mensaje de uno de sus antepasados...

6. Si el sujeto de la oración anterior contiene un único candidato posible y la anáfora es sujeto de la oración actual, se escoge como antecedente (11,0).

(66) *Besides, if he provoked* Malek$_i$, *uncertainties were introduced, of which there were already far too many.* He$_i$ *noticed the supervisor enter the lounge...*

Además, al provocar a *Malek*$_i$, surgieron dudas, de las que había ya demasiadas. *Él*$_i$ se dio cuenta de que el supervisor entraba al salón...

Las reglas se aplican para resolver uno a uno cada pronombre. Si una lo resuelve, no se aplica la siguiente. Si ninguna lo resuelve, la anáfora queda sin resolver.

Para la evaluación, Baldwin realiza dos experimentos con sus sistema. En primer lugar, compara CogNIAC con el algoritmo de Hobbs (1976)[15] y, en segundo lugar, lo evalúa con un corpus del *Wall Street Journal* sobre un conjunto de textos narrativos.

Para la comparación de CogNIAC con el algoritmo de Hobbs, Baldwin trata únicamente el pronombre personal de tercera persona. Además los errores no se enlazan (si un pronombre está mal resuelto en una oración, el error se corrige para resolver el siguiente) por lo que no se arrastran errores de resolución. Dado que el algoritmo de Hobbs resuelve todos los pronombres, Baldwin añade dos reglas de baja precisión a las seis originales:

[15] Este algoritmo se detalla en 3.1.1.

7. Si hay un centro C_b que mira hacia atrás[16] en la cláusula actual que es también candidato a antecedente, se escoge como antecedente.
8. Escoge el candidato más cercano como antecedente.

Estas reglas hacen que CogNIAC resuelva todos los pronombres, eliminando la posible ambigüedad mencionada anteriormente (si todas fallan, la regla 8 selecciona el más cercano).

La comparación de ambos algoritmos revela que, para un total de 298 pronombres de tercera persona, el algoritmo de Hobbs proporciona un 78.8 % de éxito (235 pronombres resueltos correctamente), mientras que el algoritmo de CogNIAC con las 8 reglas (baja precisión) obtiene un 77.9 % de éxito (232 pronombres resueltos correctamente). Por otro lado, y para los mismos textos, el CogNIAC de 6 reglas (alta precisión) proporciona un 92 % de precisión[17](190/206) y un 64 % de cobertura[18](190/298).

Por otro lado, en la evaluación sobre textos del *Wall Street Journal*, se añade un conjunto de módulos, como un analizador parcial para identificar cláusulas finitas, un detector del *it* pleonástico, un patrón de selección de sujeto, reglas para procesar estilo indirecto, reglas que buscan un único antecedente ocho oraciones antes, doce oraciones antes, etc. Además, se eliminan las reglas 4, 7 y 8. CogNIAC proporciona en este caso una precisión del 73 % con una cobertura del 75 %.

3.1.5 Aproximación pobre en conocimiento de Mitkov

Mitkov (1998) presenta una aproximación pobre en conocimiento para resolver los *it* anafóricos. A partir de la salida de un etiquetador gramatical, el algoritmo forma una lista de candidatos utilizando un conjunto de reglas de sintagma nominal. Con los sintagmas nominales en una distancia de dos oraciones y mediante el uso de la concordancia morfológica (género y número) elimina los candidatos incompatibles con la expresión anafórica. A la lista resultante, el sistema aplica un conjunto de preferencias

[16] Para unas nociones básicas sobre *centering*, ver 3.2.6.
[17] $precisión(P)$ = pronombres correctos/pronombres tratados
[18] $cobertura(C)$ = pronombres correctos/total pronombres

asignando una puntuación a cada candidato a través de los llamados indicadores de antecedente (*antecedent indicators*) y aplica una serie de prioridades en el caso de que más de un candidato obtenga la misma puntuación.

Los indicadores de antecedente tienen su fundamento en estudios empíricos e integran información de relevancia, de situación estructural, de distancia o de preferencia de términos. Cada indicador de antecedente asigna una puntuación a cada candidato (-1, 0, 1 ó 2). El candidato con la mayor puntuación tras la aplicación de todos los indicadores será el propuesto por el sistema como antecedente. Mitkov (1998) muestra con ejemplos los siguientes indicadores de antecedente:

- Se prefieren los sintagmas nominales definidos a los indefinidos. Se consideran definidos aquellos sintagmas nominales introducidos por un artículo definido, un pronombre demostrativo o un posesivo.
- Se prefieren aquellos sintagmas nominales que representan el "tema". Una simple heurística define el tema como el primer sintagma nominal de una oración no imperativa[19].
- Se prefieren los sintagmas nominales que siguen inmediatamente a un conjunto de verbos denominados verbos de indicación (*discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyze, synthesize, study, survey, deal, cover*).
- Se prefieren los sintagmas nominales que se repiten, tanto de manera idéntica como con el mismo núcleo o sinónimos.
- Se prefieren los sintagmas nominales que aparecen en el encabezado de una sección.
- Se prefieren los sintagmas nominales que no forman parte de un sintagma preposicional. Esta preferencia parte de la teoría del *centering* (véase 3.2.6).

[19] Esto se basa en el hecho de que, en un texto coherente, la información conocida (o tema) aparece primero, y por lo tanto forma un enlace correferencial con el texto anterior, mientras que la nueva información (o rema) amplía la información sobre el tema.

- Se prefieren los sintagmas nominales con un patrón de situación idéntico al del pronombre. Dada la ausencia de información sintáctica, esta preferencia sólo trata patrones que identifican la posición del sintagma nominal con respecto al verbo (antes o después).
- Se prefieren los sintagmas nominales dentro de una estructura de referencia inmediata. Estas estructuras, similares a los patrones mencionados en la preferencia anterior, tienen la siguiente forma:

  You V_1 SN ... *con* (you) V_2 it (*con* (you) V_3 it)

  donde *con* es una conectiva –*and* (*y*), *or* (*o*), *before* (*antes*), *after* (*después*), etc.–. El sintagma nominal que sigue a V_1 se considerará el candidato más adecuado para el pronombre (*it*).
- Se prefieren los sintagmas nominales con mejor distancia referencial. En oraciones complejas, los sintagmas nominales que se encuentran en la cláusula anterior[20] son los mejores candidatos para una anáfora en la siguiente cláusula, seguidos por los sintagmas nominales en la oración anterior y seguidos por los sintagmas nominales que se encuentran dos oraciones antes y por último los que están tres oraciones antes. Para oraciones simples se consideran sólo los sintagmas nominales situados una, dos o tres oraciones antes.
- Se prefieren los sintagmas nominales relacionados con el dominio del texto.

Las puntuaciones asignadas por cada indicador quedan recogidas en el cuadro 3.2.

Mitkov realiza dos experimentos sobre textos pertenecientes a manuales de informática. Como medida global de la eficiencia del método, el sistema pobre de conocimiento proporciona un 89,7 % de tasa de éxito en una evaluación manual sobre 196 pronombres.

Además, compara su sistema con el CogNIAC (Baldwin, 1997) por ser un sistema de concepción muy similar (ambos son aproximaciones pobres en conocimiento y ambos utilizan un etiquetador

[20] La identificación de cláusulas en oraciones complejas se realiza con reglas obtenidas experimentalmente.

Preferencia	+	−
SN definido	0	-1
SN tema	1	0
SN con verbo de indicación	1	0
SN repetido (2 o más veces)	2	
SN repetido (1 vez)	1	
SN en encabezado	1	0
SN no incluido en un SP	0	-1
Situación con respecto al verbo	2	0
Referencia inmediata	2	0
Misma cláusula	2	
Oración anterior	1	
2 oraciones antes	0	
3 oraciones antes	-1	
SN del dominio	1	0

Cuadro 3.2. Valores asignados por los indicadores de antecedente (Mitkov, 1998)

gramatical como entrada). En su evaluación manual, el sistema de Mitkov revela mejores resultados que el de Baldwin, teniendo en cuenta, tal y como precisa el autor, que dicha mejoría puede manifestarse en los textos tratados por el primero (manuales de informática), pero que podría cambiar en otro tipo de textos.

3.1.6 La unificación de huecos de Ferrández

Ferrández (1998) integra un módulo basado en restricciones y preferencias morfosintácticas para la resolución de la anáfora pronominal, adjetiva y superficial numérica en un sistema de PLN, a partir de un análisis parcial con el uso del formalismo gramatical SUG (*Slot Unification Grammar*).

La principal aportación de este trabajo es el enfoque basado en las gramáticas SUG de unificación de huecos (Ferrández et al., 1998) y en el analizador parcial SUPP (Ferrández et al., 1999).

Para la resolución de la anáfora se propone la aplicación de un conjunto de restricciones y preferencias, basadas fundamentalmente en criterios morfológicos –concordancia en género y número– y sintácticos –basados en las reglas *c-dominio* definidas por Reinhart (1983)–.

La evaluación del sistema se realiza sobre el corpus *The Blue Book*[21] (manual técnico de telecomunicaciones, International Telecommunications Union CCITT handbook), en sus versiones en español e inglés.

Si bien el sistema trata la resolución de la anáfora pronominal, adjetiva y superficial numérica, sólo la primera resulta de especial interés dada la reducida representación de las otras en el corpus tratado.

Para la resolución en español, sobre un total de 100 pronombres personales (53 de complemento, 26 en sintagmas preposicionales y 21 no incluidos en sintagmas preposicionales), el éxito medio obtenido es del 83 % (85 %, 85 % y 76 % respectivamente) en la detección del antecedente correcto (Ferrández et al., 1998).

En la adaptación del algoritmo para el inglés (Ferrández et al., 1999), se obtienen mejores resultados sobre la versión inglesa del corpus. Exactamente, sobre un total de 81 pronombres personales, el algoritmo detecta el antecedente correcto en un 87 % de las ocasiones. Los autores achacan esta mejoría a la inferior longitud de las oraciones en inglés que en español.

Es interesante apuntar que en todos estos trabajos se destaca, como uno de los principales factores de error, la ausencia de información semántica. De hecho, Ferrández (1998) propone, si bien no la utiliza para su evaluación, la incorporación de información semántica al análisis sintáctico con el uso de ontologías de dominio propuesta por el método IRSAS (Moreno, 1993).

Con el uso del mismo formalismo gramatical, trabajos de investigación más recientes han logrado mejorar los resultados en la resolución de la anáfora para dominios más generales en su evaluación con corpus menos restringidos y enriquecidos con nuevos módulos de detección de sujetos pronominales omitidos (Peral, 2001).

[21] Corpus incluido en el Proyecto CRATER (Corpus Resources and Terminology Extraction). Proyecto financiado por la Comisión de las Comunidades Europeas (DG-XIII). Investigadores principales: F. Marcos y F. Sánchez. Laboratorio de Lingüística Informática. Facultad de Filosofía y Letras. Universidad Autónoma de Madrid. Para más información visitar las páginas `http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html` y `http://www.lllf.uam.es/ fernando/projects/CRATER.html` (última visita agosto de 2001).

3.1.7 Conclusiones sobre los métodos de conocimiento limitado

En esta sección hemos estudiado los principales sistemas de resolución de la anáfora basados en el uso de información de origen morfológico y sintáctico. Estos sistemas, incluidos los más clásicos, han conseguido resultados excelentes con niveles de coste computacional, en general, reducido. Uno de los factores fundamentales de su éxito ha sido la evaluación sobre corpus de dominio restringido. El cuadro 3.3 resume los datos principales sobre todos y cada uno de los métodos tratados.

3.2 Métodos enriquecidos

La resolución de la anáfora con conocimiento limitado ha demostrado a través de sus distintas aproximaciones ser un método robusto y computacionalmente asequible. No obstante, no es difícil intuir que la incorporación de nuevas fuentes de conocimiento a la resolución de la anáfora pueden mejorar los resultados obtenidos con conocimiento más limitado. Así, Hobbs, que con su sistema de conocimiento limitado (Hobbs, 1976) obtiene muy buenos resultados (véase 3.1.1), plantea la necesidad de información semántica como un complemento imprescindible en los procesos de resolución de la anáfora, ampliando, para ello, su propuesta con la representación semántica del texto (Hobbs, 1978), de la que extrae reglas adicionales para el proceso de resolución.

Esta sección presenta sistemas de resolución de la anáfora que, si bien, al igual que los sistemas de la sección anterior, pueden hacer uso de información de naturaleza morfosintáctica, incorporan estrategias adicionales con nuevas fuentes de conocimiento.

Aunque estos métodos suelen suponer un mayor consumo de recursos y tiempo con respecto a los del grupo anterior, la introducción de nuevas fuentes de conocimiento proporciona nuevos e interesantes criterios adicionales de selección del antecedente correcto de una expresión anafórica.

	Autores (año)	Tipo de anáfora	Idioma	Morfol.	Sintác.	Restr.	Pref.	Corpus usado	Nº pron.	Evaluación
1976	Hobbs	Pers y pos.	INGLÉS	☑	☑	☑		Manual de arqueología Novela Prensa	300	81'6%
1994	Lappin y Leass	Pers. Reflexivas y recíprocas	INGLÉS	☑	☑	☑	☑	Manuales de informática	560	86%
1996	Kennedy y Boguraev	Pronominal (3ª persona) Anáforas reflexivas y recíprocas	INGLÉS	☑	☑	☑	☑	Textos variados	306	75%
1997	Baldwin	Pronominal (3ª persona)	INGLÉS	☑	☑	☑		Hobbs WSJ	298	P=92% C=64% P=73% C=75%
1998	Mitkov	Pronombre "it"	INGLÉS	☑	☑	☑	☑	Manuales de informática	196	87,7%-89,7%
1999	Ferrández	Pronominal	ESPAÑOL INGLÉS	☑	☑	☑	☑	Bluebook (ES) Bluebook (IN)	100 81	83% 87%

Cuadro 3.3. Resumen de métodos de resolución de la anáfora con conocimiento limitado

3.2.1 Restricciones y preferencias de Carbonell y Brown

Carbonell y Brown (1988) realizan una aproximación "multiestrategia" para la resolución de la anáfora, asumiendo que la combinación de un conjunto de estrategias proporciona mejores resultados. Se concentran en la anáfora intersentencial por considerarla la más frecuente y la más importante en el diseño de interfaces de lenguaje natural. El sistema integra distintas fuentes de conocimiento: sintaxis oracional, concordancia semántica, estructura del diálogo y conocimiento general del mundo, para cuya aplicación los autores proponen un marco general de resolución de la anáfora basado en restricciones y preferencias. Mientras que las restricciones no pueden ser transgredidas, las preferencias seleccionarán a través de un sistema de ponderación el antecedente anafórico de entre todos los candidatos que cumplan todas las restricciones.

Las restricciones son:

- Concondancia local: concordancia en género y número entre la anáfora y el candidato.
- Semántica caso-rol[22]: restringe los rasgos semánticos del candidato a los rasgos correspondientes a la anáfora. Así, en el ejemplo (67), el pronombre *lo* no puede hacer referencia a *Juan* ni a *mesa*, ya que ninguno de los dos es comestible.

(67) Juan$_j$ cogió el *pastel*$_i$ de la *mesa*$_k$ y se *lo*$_i$ comió.

  Estas restricciones son similares a las restricciones definidas por otros autores, como *restricciones de selección* (Hobbs, 1976, 1978) –ver 3.1.1– o *consistencia del tipo semántico* (Rich y Luperfoy, 1998) –ver 3.2.2–.
- Precondición–postcondición: eliminan los candidatos envueltos en acciones cuya postcondición viola la precondición impuesta por la anáfora. Por ejemplo, en (68), el pronombre *él* hace referencia a Antonio, ya que Juan no posee la manzana en ese momento.

[22] Traducción libre de término en inglés *case-rol semantics*.

(68) Juan le dio a *Antonio*$_i$ una manzana. *Él*$_i$ se comió la manzana.

Evidentemente, y tal y como reconocen los autores, la aplicación de estas restricciones requiere una enorme cantidad de conocimiento para poder ser aplicada de forma general.

Por otro lado, las preferencias son:

- Paralelismo sintáctico: se da prioridad a aquellos candidatos con el mismo papel sintáctico que la anáfora, tal y como se muestra en los siguientes ejemplos, en los que anáfora y antecedente comparten el mismo índice.

(69) (a) El programador combinó *Prolog*$_i$ con *C*$_j$, aunque ya *lo*$_i$ había combinado anteriormente con Pascal.

(b) El programador combinó *Prolog*$_i$ con *C*$_j$, aunque ya había combinado anteriormente Pascal con *éste*$_i$.

- Alineamiento semántico: se da prioridad a aquellos candidatos que se alinean semánticamente con la anafora, tal y como se puede comprobar en (70a) y (70b).

(70) (a) María condujo del *parque*$_j$ al *club*$_i$. Pedro fue *allí*$_i$ también.

(b) María condujo del *parque*$_i$ al *club*$_j$. Pedro salió de *allí*$_i$ también.

- Topicalización sintáctica: se prefieren los candidatos topicalizados[23] y se proponen como antecedentes si no incumplen ninguna restricción. Se detecta la topicalización a través de determinadas estructuras lingüísticas.
- Proximidad intersentencial: se prefieren los candidatos más próximos a la anáfora recorriendo el texto hacia atrás.

[23] Si bien no existe un acuerdo en la denominación del fenómeno, el concepto de topicalización coincide con lo que algunos autores denominan tematización, entendida como "aquel mecanismo sintáctico en virtud del cual el tema –sea o no sujeto– aparece en un lugar periférico dentro de la oración, que suele coincidir (aunque no necesariamente) con la posición inicial" (Hernanz y Brucart, 1987). Sobre los conceptos de tema y rema, véase nota 19 (pág. 50).

Este sistema fue evaluado sobre un texto con 31 oraciones. Se realizó un análisis completo mediante un formalismo basado en una gramática léxico-funcional. De un total de 30 anáforas (27 pronominales y 3 descripciones definidas), el sistema resuelve correctamente todas menos cuatro, con lo que el porcentaje de éxito se puede establecer en un 86,6 %. Si bien es un resultado bastante interesante, la muestra es demasiado pequeña y el dominio excesivamente restringido para asegurar el mismo comportamiento en una evaluación de mayor envergadura.

3.2.2 La arquitectura distributiva de Rich y Luperfoy

Rich y Luperfoy (1998) describen una arquitectura distributiva para la resolución de la anáfora pronominal. Un analizador proporciona un conjunto de características que representan las propiedades sintácticas de los constituyentes de la oración, mientras que un procesador semántico produce una lista de referentes discursivos y hechos relacionados con ellos. El módulo de resolución de la anáfora añade a este conjunto de hechos otros relativos a relaciones de correferencia entre referentes del discurso.

Los autores afirman que no existe una teoría coherente sobre la que se pueda construir un sistema de resolución de la anáfora, sino que existen muchas teorías parciales cada una de las cuales explica un conjunto de fenómenos que influyen en el uso e interpretación de la anáfora pronominal. Por ello, al igual que Carbonell y Brown (1988) definen una arquitectura calificada como distribuida por la integración de un conjunto de módulos que representan cada teoría parcial con el fin de cubrir un mayor espectro de tratamiento de la anáfora pronominal.

Los módulos que integran el sistema forman un conjunto de fuentes de restricción (tal y como las denominan los autores) que aplican conocimiento morfológico, sintáctico y semántico a través de los siguientes factores:

- Proximidad: propone candidatos del discurso más reciente. No afecta a lo que otros factores puedan determinar.
- Concordancia en género y número: la anáfora debe concordar en género y número con su antecedente. Este factor no propo-

ne antecedentes, sino que actúa de filtro para los antecedentes propuestos por otros factores.

- Animación[24]: los pronombres neutros se refieren a cosas inanimadas, mientras que los pronombres masculinos y femeninos hacen alusión a cosas animadas[25]. Este factor tampoco propone antecedentes, sino que filtra antecedentes propuestos.
- Referencia inconexa: este factor hace uso de las restricciones basadas en la sintaxis (Reinhart, 1983) aplicadas a pronombres reflexivos y recíprocos. Propone antecedentes a pronombres reflexivos y sirve de filtro para pronombres no reflexivos.
- Consistencia del tipo semántico: este factor no considera como válidos aquellos candidatos que no satisfacen las restricciones impuestas por la interpretación semántica de la oración. Para poder aplicar esta restricción, los autores definen manualmente una jerarquía de tipos y un conjunto de interpretaciones de los verbos. Así, a la frase "*The system$_i$ created an error log$_j$. It$_i$ printed it$_j$.*" se le aplica interpretación manualmente creada del verbo *print* (*imprimir*) que sería:

  agente: humano/ordenador
  paciente: estructura de información

  y que resolvería correctamente ambos pronombres. Esta restricción es similar a las propuestas por otros autores[26] en las que se define el tipo de restricciones asociadas a cada rol semántico (agente y paciente) de un verbo determinado (el verbo *imprimir* requiere un agente de tipo 'humano' y un paciente de tipo 'es-

[24] Entendemos por animación el rasgo que define si un sustantivo es o no animado.

[25] Es preciso recordar que este sistema ha sido planteado originalmente para el inglés, en el que, de manera general, la relación género-animación es válida. Este mismo planteamiento para el español no sería adecuado, si bien en este trabajo se propone una adaptación de este tipo de reglas al tratamiento de la resolución de la anáfora en español. Para más información sobre esta adaptación, puede consultarse la sección 4.3 (pág 113).

[26] Este tipo de filtro, tratado de forma similar por otros autores como restricciones de selección (Hobbs, 1978, 1986) (Carter, 1986, 1987a) o restricciones caso-rol (Carbonell y Brown, 1988), tiene una analogía inmediata con uno de los objetivos principales de esta Tesis, que es el de determinar, de manera automática, cuáles son los rasgos semánticos relacionados con los papeles sintácticos de una oración y que definen su comportamiento a través de un conjunto de patrones.

tructura de información'). De esta manera, aquellos candidatos que no cumplen estas restricciones pueden ser eliminados.
- Foco global: propone como antecedentes aquellas entidades del discurso que forman parte del foco global.
- Catáfora: en algunos casos, el sistema propone como antecedente un sintagma nominal que aparece después del pronombre.
- Accesibilidad lógica: impone un conjunto de reglas basadas en la accesibilidad de referentes como cuantificadores o negadores (Kamp, 1981).

Cada uno de estos factores proporcionan para cada candidato, a través de una fórmula, una puntuación (entre -5 y 5) y una medida de confianza (entre 0 y 1).

$$valor = \frac{\sum_{i=1}^{n} puntuación(i) \cdot confianza(i)}{\sum_{i=1}^{n} confianza(i)} \tag{3.1}$$

El valor final es un número entre 0 y 1 que combina un conjunto de pares (puntuación-confianza) con la formula mostrada en (3.1). A partir de este valor se selecciona el antecedente correcto.

En lo referente a la evaluación, los autores no proporcionan en su documentación información sobre los resultados de este algoritmo.

3.2.3 El algortimo de Kameyama

Kameyama (1997b) propone un algoritmo para la resolución de la anáfora nominal[27]. El algoritmo utiliza un conjunto de entradas incompletas sintácticamente, que son todavía más pobres que las entradas del sistema de Kennedy y Boguraev (1996). El algoritmo de Kameyama trabaja con tres factores principales:

- Regiones de texto accesible: definidas como el texto precedente completo para los nombres propios, 10 oraciones para las descripciones definidas y 3 oraciones para los pronombres.

[27] La anáfora nominal es aquella introducida por pronombres, sintagmas nominales definidos y nombres propios que hacen referencia a un sintagma nominal antecedente.

- Consistencia semántica: consistencia de número, consistencia de tipo (las anáforas deben ser del mismo tipo o contener el tipo de su antecedente)[28]. Para aplicar este factor el algoritmo requiere la definición de una jerarquía que, según el propio autor, es escasa e incompleta y está definida *ad hoc* para la aplicación de este factor.
- Consistencia de modificador: *española* y *francesa* son inconsistentes, mientras que *francesa* y *multinacional* no lo son. El autor afirma que el sistema no tiene conocimiento suficiente para aplicar este factor adecuadamente.

Dado que no se dispone de información sobre papeles sintácticos, el algoritmo realiza una aproximación realizando una ordenación lineal de la oración de izquierda a derecha[29].

En la evaluación para el sistema de extracción de información MUC-6 FASTUS (Kameyama, 1997a), el algoritmo reveló uno de los resultados más exitosos: 59 % de cobertura y 72 % de precisión.

3.2.4 Combinación de técnicas lingüísticas y estadísticas de Mitkov

Mitkov (1994, 1996) define un modelo integrado de resolución de la anáfora basado en la combinación de métodos lingüísticos tradicionales con una aproximación estadística.

El método integra módulos asociados a diferentes fuentes de conocimiento:

- El módulo sintáctico (que incluye también información morfológica) asegura la concordancia en género, número y persona entre el antecedente y la anáfora, así como que ambos no son incompatibles según las restricciones *c-comando* (Reinhart, 1983).

[28] En realidad, lo que plantea este factor es una relación semántica de sinonimia o hiperonimia/hiponimia entre anáfora y antecedente. Estas relaciones semánticas se tratan más ampliamente en la 4.3.4 (pág. 120).

[29] Esta técnica es muy habitual para asignar papeles sintácticos cuando no se dispone de información a través de un analizador. Sin embargo, esto es sólo posible en lenguajes de orden fijo, como el inglés, ya que permiten hacer aproximaciones fiables de papeles sintácticos. En el caso de lenguajes de orden libre, como el español, el uso de esta técnica es mucho menos fiable.

- El módulo semántico comprueba la consistencia entre la anáfora y el posible antecedente, eliminando los candidatos incompatibles según la semántica del verbo principal o la animación del candidato, dando preferencia a aquellos candidatos con el mismo rol semántico que la anáfora. Esta información semántica ha sido añadida previamente de forma manual.
- El módulo de conocimiento del dominio es una base de conocimiento de los conceptos del dominio tratado.
- El módulo de conocimiento del discurso puede localizar el centro del segmento de discurso actual, para lo que utiliza un motor bayesiano estadístico que sugiere el centro más probable ante una nueva evidencia. Este módulo desempeña un papel muy importante y suele proponer el centro localizado como el antecedente más probable.

Los módulos sintácticos y semánticos (excepto el paralelismo sintáctico y semántico) sólo filtran los candidatos sin proponer ninguno, mientras que los módulos de dominio, heurístico y de discurso son los que proponen el antecedente.

Mitkov realiza dos pruebas del método. La primera, activando los módulos sintáctico, semántico y de dominio, y la segunda incorporando además el de discurso. Los resultados demuestran una mejora de resolución de la anáfora cuando se combinan las estrategias lingüísticas tradicionales con la aproximación estadística propuesta (el grado de éxito va del 87.7 % al 89.7 % en la primera prueba, y del 86.7 % al 91.6 % en la segunda).

3.2.5 El sistema SPAR de Carter

Carter (1986, 1987a) utiliza fuentes de conocimiento basadas en la sintaxis, la semántica y el foco local para el sistema SPAR (*Shallow Processing Anaphor Resolver* – Resolutor de anáfora con procesamiento superficial), un sistema que resuelve la anáfora nominal. Para ello combina un conjunto de teorías, especialmente la teoría del foco local (Sidner, 1979)[30], la teoría de preferencia

[30] Una de las evoluciones más interesantes del trabajo de Sidner en esta línea de investigación, junto con el trabajo de Grosz, culminó en la teoría del *centering* (Grosz et al., 1995), ampliamente utilizada por sistemas de resolución de la anáfora (ver 3.2.6).

semántica (Wilks, 1975) y la inferencia del sentido común (Carter, 1987b).

SPAR trabaja sobre la salida de un analizador sintáctico en inglés (Boguraev, 1979) que resuelve la ambigüedad estructural.

A continuación, el sistema aplica la interpretación de pronombres (PI) definida por Sidner (1979, 1983), mientras que para los sintagmas nominales léxicos[31] se aplican otras reglas basadas en el foco local. Las reglas PI proponen un único candidato. Para cada pronombre propuesto, el sistema usa una fórmula para calcular la densidad semántica de cada palabra y establece también los rasgos semánticos del pronombre. Para ello, Carter define unas primitivas a partir de las expuestas por Wilks (1975), del tipo:

```
((MAN-SUBJ) ((MAN-OBJE) (TELL FORCE)));
```

que quiere decir que el verbo *interrogar* en el sentido de 'forzar a alguien a decir algo' es preferiblemente hecho por una persona (sujeto) a una persona (objeto directo). Si se produce la coincidencia semántica, el candidato se propone como antecedente[32].

SPAR proporciona uno de los mejores resultados obtenidos por un sistema de resolución de la anáfora. Fue evaluado sobre dos grupos de textos correspondientes a historias en inglés. El primer grupo, con un total de 65 pronombres, fue escrito para ser evaluado con el SPAR, y el sistema resolvió todos los pronombres. El segundo, escrito por personas sin conocimiento de la forma de trabajar del SPAR, contenía un total de 242 pronombres, de los que 226 (93 %) fueron correctamente resueltos. El autor afirma que el porcentaje puede elevarse hasta el 96 % con un procedimiento de recuperación de errores. Estos resultados tan sorprendentes que superan a los de la mayoría de los sistemas tratados, responden, según el autor, a evaluaciones de corpus muy concretos, definidos *ad-hoc* y con situaciones ideales de análisis. Si bien el lector no debe dejarse impresionar por dichas cifras desde un punto de vista puramente computacional, estos resultados revelan las posibili-

[31] El concepto de sintagma nominal léxico, traducido del inglés *lexical noun phrase*, se corresponde con el concepto de sintagma nominal que hace referencia a una entidad previa del discurso.

[32] Este enfoque refiere una vez más al tipo de restricciones y preferencias semánticas planteadas en este trabajo (ver 4.3, pág. 113).

dades que quedan abiertas para sistemas que incorporen nuevas fuentes de información.

3.2.6 Algoritmos basados en la estructura del discurso

La forma de construir el discurso supone una herramienta de interesantes beneficios como fuente de información estructural. Este apartado recoge, por un lado, algunas teorías basadas en la estructura del discurso y, por otro lado, algunas estrategias de resolución que toman como punto de partida las teorías anteriores.

El *centering*. Una de las teorías más populares, en la que se basan una gran cantidad de estas estrategias, es el *centering* (Grosz et al., 1983, 1995). El *centering* se define como un marco global para modelar la coherencia local en el discurso. El marco conceptual del *centering* explica la coherencia local que relaciona el foco local (entidad más relevante en el contexto actual y, por tanto, principal candidato anafórico) y la forma de las expresiones anafóricas.

Este marco se basa en tres afirmaciones principales:

1. Dado un enunciado U_i, el modelo predice qué entidad del discurso será el foco de U_{i+1}.
2. Cuando el foco local es el mismo entre un enunciado y el siguiente, el modelo predice que se hará referencia a dicho foco mediante un pronombre.
3. Cuando se encuentra un pronombre, el modelo proporciona un orden de preferencia sobre los antecedentes posibles del enunciado anterior.

Para ello, en cada U_i se crean las siguientes estructuras de datos:

La lista $Cf(U_i)$ de "centros que miran hacia adelante" (*forward-looking centers*), ordenada, que incluye todas las entidades del discurso del enunciado U_i. Su primer elemento es el "centro" preferido, $Cp(U_i)$, y será el candidato que se espera encontrar en $Cb(U_{i+1})$.

El elemento $Cb(U_{i+1})$ o "centro que mira hacia atrás" (*backward-looking center*), que es el elemento mejor posicionado de $Cf(U_i)$, al que se hará referencia en el siguiente enunciado U_{i+1}.

El criterio de ordenación usado en Grosz *et al.* (1995) ordena los elementos de la lista Cf mediante papeles sintácticos. De esta forma, las entidades con papel de sujeto se prefieren a aquellas que lo tienen de objeto y los objetos se prefieren a los otros (complementos circunstanciales, etc.).

El *centering* define un orden de preferencia basado en técnicas para efectuar un cambio de foco, como se muestra en el cuadro 3.4

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

Cuadro 3.4. Tipos de transición en el *Centering*

Al partir de las estructuras de datos previamente definidas y del criterio de ordenación anterior, el núcleo de la teoría se basa en dos reglas de *centering*:

Regla 1: Si cualquier miembro de $Cf(U_i)$ es referenciado por un pronombre en U_{i+1}, entonces $Cb(U_{i+1})$ debe ser un pronombre.

Regla 2: las secuencias de continuaciones se prefieren a las secuencias de retenciones, y las secuencias de retenciones se prefieren sobre las secuencias de desplazamientos.

En las siguientes secciones se tratarán algunas propuestas que utilizan la teoría del *centering* como base de métodos de resolución de la anáfora (Brennan et al., 1987; Tetreault, 1999).

El *centering* funcional. Uno de los problemas que plantean los idiomas de orden libre es la dificultad del análisis de los roles gramaticales. Basándose en este hecho, la teoría del *centering funcional* (Strube, 1998; Strube y Hahn, 1999) usa un criterio de ordenación diferente, basado en lo que los autores denominan la

familiaridad de las entidades del discurso, información extremadamente relevante para lenguajes de orden libre.

Strube define, según este criterio de ordenación, dos conjuntos de expresiones: las entidades del discurso conocidas para el oyente (*hearer-old*) y las entidades del discurso nuevas para el oyente (*hearer-new*). Así, en el conjunto de entidades conocidas se incluyen las entidades del discurso mencionadas previamente y ya resueltas (anáforas pronominales, nombres propios ya aparecidos, pronombres relativos, aposiciones, etc.) y las conocidas pero no usadas (nombres propios y títulos). El resto de entidades se asignan al conjunto de entidades nuevas. De este modo, el criterio básico para resolver el pronombre es la preferencia de entidades conocidas frente a entidades nuevas.

Así, Strube (1998) propone la siguiente adaptación al modelo del *centering*:

- La lista Cf se sustituye por la lista de entidades de discurso relevantes (S-list), que contiene aquellas entidades del discurso que han sido referidas en el enunciado actual y en el previo.
- Los elementos de la lista S-list se ordenan de acuerdo con criterio básico definido anteriormente y con la información sobre la posición:

  Si $x \in Old$ y $y \in New$, entonces x precede a y.
  Si $x, y \in Old$ o $x, y \in New$,
  entonces si $enunciado(y)$ precede a $enunciado(x)$,
  entonces x precede a y,
  si $enunciado(y) = enunciado(x)$ y $pos(x) < pos(y)$,
  entonces x precede a y.

- Puesto que no hay una definición clara de los que se considera como enunciado, los autores adoptan el siguiente criterio: las cláusulas verbales se definen como enunciados por sí mismas, mientras que las cláusulas no verbales se procesan con la principal, constituyendo un único enunciado.

De esta forma, Strube propone el algoritmo siguiente:

1. Si se encuentra una expresión de referencia,

a) si es un pronombre, comprobar los elementos de la lista S-list por orden hasta que alguno sea válido.
b) actualizar la S-list con la información de la expresión de referencia.

2. Si se termina el análisis del enunciado U, eliminar todas las entidades del discurso de la lista S-list que no hayan sido referidas en U.

La evaluación del algoritmo obtuvo una precisión del 85,4 %, mejorando los resultados del algoritmo de *centering* propuesto por Brennan *et al.* (1987), que sólo alcanzó el 72,9 % cuando fue aplicado al mismo corpus.

La teoría del centering funcional ha sido aplicada también a sistemas de resolución de la anáfora en diálogos (Eckert y Strube, 2001).

El algoritmo BFP. El algoritmo BFP (Brennan et al., 1987), basado en la teoría del *centering* (Grosz et al., 1983, 1995) descrita en 3.2.6, aplica dos tipos de restricciones. Por un lado, incorpora las llamadas restricciones de "contra-índices"[33], de naturaleza muy similar a las restricciones *c-dominio* (Reinhart, 1983).

Por otro lado, y por lo que respecta a la estructura del discurso, el BFP distingue entre cambio suave y cambio severo[34] para identificar la entidad central del discurso a la que se refiere el hablante. Ambos cambios representan un cambio de entidad central del discurso, aunque el cambio suave indica la intención del hablante de continuar hablando de la entidad de cambio, algo que no ocurre en el cambio severo.

En una evaluación posterior (Walker, 1998), el comportamiento del BFP fue comparado con el algoritmo clásico de Hobbs (1976) a través de una simulación manual de ambos algoritmos sobre tres textos distintos[35]. Dos de ellos, los mismos que había usado Hobbs en sus experimentos (una novela y una publicación semanal) contenían 100 pronombres cada uno. El tercer texto era un

[33] Del inglés *contra-indexing*.

[34] Traducciones libres de los términos originales *smooth-shift* y *rough-shift*.

[35] Conviene destacar que este hecho parte de una situación ideal de los textos en los que los errores sólo pueden ser debidos a fallos en el módulo de resolución y no a fallos de análisis o errores acumulados por incorrecciones anteriores.

conjunto de diálogos hombre-hombre transcritos con un total de 81 pronombres. De los textos procedentes de la novela, el algoritmo de Hobbs resolvió correctamente 88 pronombres, mientras que el BFP resolvió 90. Los pronombres resueltos por el algoritmo de Hobbs y el BFP en los textos procedentes de la publicación semanal fueron de 89 y 79, respectivamente, mientras que en el diálogo los pronombres correctamente resueltos fueron de 49 y 51, respectivamente. En esta comparación, el autor concluye que no se puede dar una diferencia importante entre ambos algoritmos, si bien en el segundo grupo de textos el algoritmo de Hobbs supera al BFP con holgura.

El algoritmo BFP ha sido citado en numerosas ocasiones en la bibliografía sobre la resolución de la anáfora y ha servido como sistema base para algunas aproximaciones de interés, como el algoritmo LRC (Tetreault, 1999) o la adaptación para la resolución de la anáfora en diálogos de Byron y Stent (1998).

El modelo LRC de Tetreault. El LRC (*Left-Right Centering*, *centering* Izquierda-Derecha) de Tetreault (1999) es un algoritmo de resolución de pronombres basado en la teoría del *centering*. Este algoritmo es, en realidad, una alternativa al BFP (Grosz et al., 1983, 1995) y su principal ventaja, tal y como señala el autor, es que las intervenciones[36] del hablante se procesan de manera acumulativa, además de un inferior coste computacional. El funcionamiento del algoritmo es básicamente el siguiente: en primer lugar, se busca en la intervención actual el posible antecedente. Si no se encuentra, se continúa la búsqueda en la lista Cf de las anteriores intervenciones, siguiendo un recorrido de izquierda a derecha.

El LRC se evalúa comparándolo con otros tres algoritmos: el BFP (Brennan et al., 1987), el algoritmo S-list de Strube (Strube, 1998) y el algoritmo de Hobbs (Hobbs, 1976). Los cuatro algoritmos se ejecutan sobre un fragmento del corpus Pen TreeBank anotado (Marcus et al., 1993), formado por 195 artículos de pren-

[36] Aunque el término intervención puede ser interpretado de forma ambigua, Tetreault lo simplifica considerando cada nueva oración como una nueva intervención.

sa[37]. De los 2096 pronombres contenidos en el texto, se eliminan aquellos contenidos en lenguaje citado, dado que dos de los cuatro algoritmos a comparar (BFP y S-list) no soportan la resolución en textos citados, con lo que el total de pronombres a tratar es de 1696. Para el análisis, los algoritmos se dividen en dos grupos: aquellos que buscan un antecedente intersentencialmente a través de las listas Cf (grupo "N") y aquellos que sólo pueden buscar en la oración inmediatamente anterior (grupo "1").

En el grupo "N", formado por el algoritmo de Hobbs, el de Strube y el LRC-N[38], el de Hobbs obtiene el mejor resultado, un 72,8 %, seguido por el LRC-N con un 72,4 % y finalmente por el de Strube con un 68,8 %. En el grupo "1", formado por el LRC-1, el de Strube y el BFP, el mejor resultado es el conseguido por el LRC-1, un 71,2 %, seguido del Strube-1 y el BFP con un 66 % y un 56,7 % respectivamente[39].

La estructura del discurso en los sistemas de restricciones y preferencias. Martínez-Barco (2001) realiza un estudio sobre cómo definir un espacio en el que se estima que puede estar el antecedente correcto de la anáfora (el llamado *espacio de accesibilidad anafórica*). Como muestra el autor, la mayoría de los sistemas estiman este espacio de accesibilidad anafórica utilizando o bien todo el discurso (*espacio completo*) o bien un número determinado de oraciones que se extrae de la observación de corpus (*ventana de oraciones*) y que evidentemente varía de un tipo de anáfora a otra, pero también de un corpus a otro. Sin embargo, la estimación del adecuado espacio de accesibilidad anafórica se convierte en una tarea crítica, ya que un fallo en la estimación por defecto puede provocar que el verdadero candidato quede fuera de la lista inicial de candidatos posibles, con lo cual el sistema generaría una respuesta errónea. Por otra parte, una estimación por exceso generaría grandes listas de candidatos, multiplicando

[37] El mismo corpus fue utilizado en (Ge et al., 1998).

[38] El LRC fue incluido en ambos grupos adaptado a cada uno de ellos con los nombres LRC-N y LRC-1.

[39] Es importante tener en cuenta que la evaluación de Tetreault toma en consideración los algoritmos y no los sistemas, debido a la no disponibilidad de corpus anotado. Las diferencias fundamentales entre ambas estrategias de evaluación se discuten en (Mitkov, 2002).

no sólo el tiempo de respuesta del sistema, sino también la posibilidad de devolver una respuesta errónea.

A partir de estas dos ideas, en (Martínez-Barco, 2001; Palomar y Martínez-Barco, 2001) se presenta un sistema para la resolución de la anáfora en diálogos basado en restricciones y preferencias, que incorpora no sólo información lingüística (morfosintáctica) sino también información de la estructura del diálogo. Para ello se basan en las teorías de (Fox, 1987), en las que se expone que la primera mención a un referente en una secuencia de contextos se realiza con sintagma nominal. Después de esto, el hablante utilizará una anáfora para dar a entender que la secuencia aún no ha sido cerrada. Por lo tanto, las anáforas se usan para mantener secuencias abiertas. Así, los autores identifican dos secuencias diferentes capaces de generar la mayoría de las anáforas en un diálogo: el par adyacente y el ámbito del tópico. El primero genera referencias a antecedentes locales, mientras que el segundo genera referencias al propio tópico del diálogo.

El conocimiento de las diferentes estructuras que generan dichas secuencias en el discurso permite al sistema: a) estimar un espacio de accesibilidad anafórico coherente con las intenciones y el conocimiento de los hablantes, que además, al tener un fundamento estructural, no depende del corpus, sino únicamente del tipo de anáfora, y por otra parte, b) incluir nuevas preferencias, basadas en la posición que ocupan los candidatos en esta estructura, que ayudarán en la búsqueda del mejor candidato.

La evaluación de esta propuesta fue realizada sobre un corpus formado por 200 diálogos. De las 392 anáforas contenidas en dicho corpus, el sistema detectó 365, de las que resolvió correctamente un 81

3.2.7 Resolución de descripciones definidas

El tratamiento computacional del fenómeno de la anáfora se ha centrado fundamentalmente en la resolución de pronombres, exceptuando algunos trabajos importantes en la resolución de descripciones definidas.

Uno de los trabajos más recientes para el inglés en la resolución de descripciones definidas[40] anafóricas en esta línea es el de Vieira y Poesio (2000), quienes plantean un sistema de procesamiento superficial con el uso de información estructural, información procedente de recursos léxicos como WordNet, así como información general bien codificada manualmente o bien adquirida de forma automática a partir de un corpus.

Vieira y Poesio clasifican las descripciones definidas en *anáforas directas* (usan el mismo núcleo que el sintagma nominal con el que correfieren), *descripciones puente*[41] (tienen un núcleo distinto al del sintagma nominal con el que correfieren) y *de nuevo discurso* (introducen una nueva entidad del discurso). El método resuelve los tres tipos de referencias con el uso de un árbol de decisión que proporciona cada una de las tres categorías en función de un conjunto de reglas léxicas, morfológicas, sintácticas y semánticas.

El sistema propuesto por Vieira y Poesio utiliza el fragmento anotado del corpus Pen TreeBank I (Marcus et al., 1993), que contiene artículos del *Wall Street Journal*. La evaluación revela un 62 % de cobertura y un 83 % de precisión para la resolución de la anáfora directa, mientras que en la identificación de descripciones de nuevo discurso la cobertura y la precisión son de un 69 % y un 72 % respectivamente. El sistema general que reconoce la primera aparición y las siguientes apariciones de una descripción definida obtuvo un 53 % de cobertura y un 76 % de precisión. Por otro lado, la resolución de descripciones puente fue mucho más baja. El índice de éxito en la interpretación de relaciones semánticas entre descripciones definidas (sinonimia, hiperonimia, meronimia)[42] fue del orden del 28 %, debido a la necesidad de conocimiento del mundo que tienen este tipo de descripciones definidas para su resolución.

[40] Los autores consideran descripción definida el sintagma nominal con el artículo definido inglés *the*. No incluyen otros tipos de sintagmas nominales como las construcciones pronominales, demostrativas o posesivas.

[41] Traducción libre del término en inglés *bridging descriptions*.

[42] Para más información sobre estas relaciones y sobre el recurso léxico WordNet, véase 4.3.4 (pág. 120).

Para el caso del español, cabe destacar el trabajo realizado por Muñoz (2001), en el que se propone un sistema de resolución de las descripciones definidas en español basado en restricciones y preferencias. También propone un método de clasificación de descripciones definidas en anafóricas y no anafóricas, basado en la generación de una red semántica desde WordNet (Muñoz et al., 2000; Muñoz y Palomar, 2001). Este método proporciona resultados similares al anterior en lo referente a las anáforas directas y mejora sensiblemente los resultados en las descripciones puente gracias al uso de la red semántica combinada con el recurso WordNet español.

3.2.8 Otros métodos enriquecidos

Aparte de los métodos anteriormente expuestos, se describen a continuación otros métodos, de los que se destacan sus características principales.

Las preferencias semánticas de Wilks. Wilks (1975) utiliza un módulo de resolución de la anáfora dentro de un sistema de traducción inglés-francés que integra cuatro niveles de resolución de pronombres dependiendo del tipo de anáfora y del mecanismo necesario para resolverla. El nivel inferior, denominado "anáfora A" utiliza conocimiento de sentidos individuales de palabras para resolver casos como el expuesto en el ejemplo (71), donde cada pronombre se interpreta de forma correcta haciendo uso del conocimiento de que los monos no pueden estar maduros y los plátanos no pueden estar hambrientos o, lo que es lo mismo, que los monos son mejores candidatos a estar hambrientos que los plátanos y que éstos son mejores candidatos a estar maduros que los primeros[43].

(71) *Give the* bananas$_i$ *to the* monkeys$_j$ *although* they$_i$ *are not ripe, because* they$_j$ *are very hungry.*

[43] El sistema propuesto en esta Tesis deriva en deducciones de este tipo por lo que esta misma oración será utilizada posteriormente para ejemplificar su funcionamiento. Ver sección 4.3 (pág. 113).

Dale los *plátanos*$_i$ a los *monos*$_j$ aunque $\emptyset_i$ no estén maduros, porque $\emptyset_j$ están hambrientos.

Si el sentido de las palabras falla en la búsqueda de un único antecedente, se utilizarán métodos de inferencia para las "anáforas B" (aquellas que necesitan inferencia analítica) o para las "anáforas C" (aquellas que requieren conocimiento del mundo real más allá del simple significado). Si la anáfora sigue sin ser resuelta, un conjunto de reglas basadas en el "foco de atención" intentará encontrar el tópico de la oración para usarlo como antecedente.

Las reglas de Guenthner y Lehmann. Guenthner y Lehmann (1983) proponen un conjunto de reglas para la resolución de la anáfora en el contexto de diálogos con preguntas a bases de datos relacionales. El sistema construye una estructura de representación del discurso y aplica un conjunto de factores a los candidatos hasta que uno es propuesto como antecedente. Estos factores son morfológicos (concordancia en género y número), sintácticos (similares a las reglas *c-dominio*), semánticos (el antecedente no es incompatible con una consulta a base de datos bien formada) y pragmáticos (un conjunto de reglas que prefieren candidatos en oraciones más recientes en vez de en menos recientes, pronombres en vez de sintagmas nominales léxicos, sintagmas nominales no incluidos en otros, sujetos en vez de no sujetos, objetos en vez de no objetos y anáfora en vez de catáfora). Como se puede comprobar, esta aproximación aplica un conjunto de preferencias muy similar al de otros métodos basados en restricciones y preferencias previamente expuestos.

El producto escalar de vectores de Rico. Rico (1994) propone un método que incorpora información morfológica, sintáctica, semántica y pragmática a la resolución de la anáfora. Para ello, el método asigna un valor numérico a cada uno de los atributos lingüísticos de la expresión anafórica y de sus candidatos a antecedente según la relevancia de cada una de las fuentes de información. La lista de los valores asignados para cada sintagma nominal forma un vector.

$$v \cdot w = \sum_{i=1}^{n} v_i \cdot w_i \tag{3.2}$$

Siguiendo la fórmula (3.2), el método utiliza el producto escalar entre los vectores del antecedente y la expresión anafórica (v y w) siendo n el número de elementos del vector y v_i el elemento que ocupa la posición i del vector v. Este producto escalar proporciona un valor de distancia, permitiendo al método ordenar los candidatos de acuerdo a la cercanía de su vector con la del vector de la expresión anafórica.

La aproximación de Nasukawa. Nasukawa (1994) realiza una sencilla aproximación a la resolución de la anáfora basada en dos tipos de preferencias básicas:

- La frecuencia de repetición en oraciones anteriores: la frecuencia en oraciones anteriores de un sintagma nominal con el mismo lema es un indicativo de preferencia para la selección del antecedente correcto.
- La posición sintáctica: el autor utiliza una regla heurística que favorece a los sujetos frente a los objetos.

Para su método, Nasakawa utiliza un diccionario de sinónimos bajo la premisa de que un candidato es tan válido como su sinónimo para ser antecedente. A pesar de definir las dos preferencias anteriores, en su implementación final el autor tiene en cuenta preferencias de tipo estructural o posicional únicamente, dado que la preferencia del sujeto sobre el objeto requiere un análisis sintáctico más profundo.

Para la evaluación utiliza un corpus procedente de dos manuales de informática, con 1904 oraciones y 112 pronombres de tercera persona, de los que el autor trata el caso del *it* y obtiene un 93'8 % de éxito.

La resolución de la anáfora en el sistema de extracción de información multilingual de Azzam *et al.*. Azzam et al. (1998a,b) desarrollan un módulo de resolución de la anáfora en el marco de M-LaSIE, un sistema de extracción de información multilingüe. Para la resolución utilizan conocimiento morfo-sintáctico y una red que define características semánticas de las palabras. Esta red semántica está adecuada al dominio de los textos tratados, con lo que proporciona información tan valiosa para un dominio concreto como carente de utilidad para un sistema genérico.

En la resolución de la anáfora pronominal el sistema obtiene una precisión y cobertura del 78 % y el 47 % respectivamente para el francés, y del 86 % y el 63 % para el inglés.

El sistema COCKTAIL. COCKTAIL (Harabagiu y Maiorano, 1999) es un sistema de resolución de la correferencia que usa un conjunto de heurísticas adquiridas del estudio del corpus y basadas en información sintáctica, semántica y de discurso.

El sistema trata tanto la anáfora pronominal como la nominal, pero dispone de distintas reglas para el tratamiento de cada tipo de anáfora (reflexiva, posesiva, de relativo, de 3ª persona, de 1ª persona, sintagmas nominales definidos y sintagmas nominales indefinidos).

COCKTAIL hace comprobaciones de carácter semántico entre anáfora y antecedente. La información semántica requerida para esta tarea es extraída de WordNet y del corpus anotado TreeBank.

Los antecedentes pueden ser encontrados no sólo en el fragmento de texto accesible, sino también en las cadenas de correferencia.

Las heurísticas de COCKTAIL tienen en cuenta la lexicalización (por ejemplo, cuando la anáfora es un adjunto de un verbo de comunicación) y algunas reglas de coherencia simples (por ejemplo, cuando la anáfora es el sujeto del verbo *add* (*añadir*), el antecedente puede ser un sujeto anterior de un verbo de comunicación.

En recientes trabajos (Harabagiu y Maiorano, 2000), que toman como base los anteriores, los autores hacen uso de un corpus bilingüe inglés-rumano para mejorar la resolución de la anáfora en ambos idiomas.

3.2.9 Conclusiones sobre los métodos enriquecidos

En esta sección se ha realizado un repaso de los principales sistemas de resolución de la anáfora basados en el uso de información de origen morfológico y sintáctico y enriquecidos con fuentes de conocimiento adicionales, como la semántica o la pragmática, el conocimiento del dominio e información del discurso.

Si bien el coste computacional de estos métodos puede exceder al de los expuestos en la sección anterior, los resultados obtenidos

por estos métodos, no sólo superan en algunos casos a los anteriores, sino que abren una línea de trabajo que intenta simular con mayor fidelidad la resolución natural del problema lingüístico de la anáfora. El mayor desarrollo de herramientas y recursos en otras áreas del procesamiento del lenguaje natural (análisis sintácticos mejorados, desambiguación del sentido de las palabras, ...) permitirá llevar a cabo técnicas de resolución más adecuadas y completas.

El cuadro 3.5 resume los datos de los principales métodos vistos en esta sección.

3.3 Métodos alternativos

Esta sección mostrará métodos de resolución de la anáfora que no han podido ser incluidos en los anteriores, bien por tratarse de métodos mixtos que combinan fuentes de conocimiento lingüístico con otro tipo de fuentes, bien por resolver la anáfora con el uso de datos estadísticos, patrones de co-ocurrencia, algoritmos genéticos u otras fuentes de información.

3.3.1 Los patrones de co-ocurrencia de Dagan e Itai

Ante la costosa implementación de las estrategias a gran escala basadas en restricciones y preferencias, Dagan y Itai (1990, 1991) presentan una estrategia alternativa de resolución del pronombre de tercera persona en oraciones seleccionadas aleatoriamente de un corpus.

Para resolver la anáfora, el modelo utiliza patrones de co-ocurrencia formados por el antecedente y el verbo de la expresión anafórica, de manera que se preferirá aquel patrón que más se repita en el corpus y, por lo tanto, el candidato que lo forme será elegido como el antecedente correcto.

Veamos el ejemplo propuesto por los propios autores. En (72) aparece el pronombre *it* en dos ocasiones, una como sujeto del verbo *recolectar* (*collect*) y otra como su objeto directo. Los candidatos a antecedente de esas anáforas son *money* (*dinero*), *collection* (*recolección*) y *government* (*gobierno*).

	Autores (sistema)	Tipo de anáfora	Idioma	Morfol.	Sintác.	Semán.	Pragm.	Discur.	Restr.	Pref.	Corpus usado	Nº pron.	Evaluación
1988	Carbonell y Brown	Pronominal DDs	INGLÉS	☑	☑	☑	☑		☑	☑	Frases sueltas	31	86,6%
1988	Rich y Luperfoy	Pronominal	INGLÉS	☑	☑	☑			☑	☑	-	-	-
1996	Mitkov	'It' pronominal	INGLÉS	☑	☑	☑	☑		☑	☑	Manuales de informática	-	86,7%-91,6%
1987	Carter	Nominal	INGLÉS		☑	☑		☑			-	242	93%
1987	Brenan et. al. (BFP)	Pronominal	INGLÉS		☑			☑			Hobbs + diálogos	281	77,3%
1997	Kameyama	Nominal	INGLÉS	☑		☑					MUC-6	-	59%-72%
1999	Tetreault (LRC)	Pronominal	INGLÉS					☑			Artículos de prensa	1696	72,4%
2000	Vieira y Poesio	DDs	INGLÉS	☑	☑	☑	☑				Pen Tree Bank	-	Cob.: 62% Prec.: 83%
2001	Martínez-Barco (ARIADNA)	Pronominal	ESPAÑOL	☑	☑			☑	☑	☑	Diálogos	392	Cob.: 75% Prec.: 81%

Cuadro 3.5. Resumen de los métodos enriquecidos

(72) *They knew full well that the companies held tax* money$_i$ *aside for collection later on the basis that the* government$_j$ *said* it$_j$ *was going to collect* it$_i$.

Ellos sabían bien que las compañías retenían *dinero*$_i$ de impuestos para su posterior recolección basándose en que el *gobierno*$_j$ dijo que (*éste-esta*)$_j$ iba a recolectar(*lo-la*)$_i$.[44]

Los patrones de co-ocurrencia que produce cada candidato con el verbo *collect,* así como el número de veces que aparecen en el corpus dichos patrones de palabras se muestran en el cuadro 3.6. Según los datos obtenidos, se preferirá el candidato *government* como antecedente del primer pronombre (sujeto) y el candidato *money* como antecedente del segundo pronombre (objeto).

sujeto	*verbo*	*objeto*	*apariciones*
collection			0
money	*collect*		5
government			198
		collection	0
	collect	money	149
		government	0

Cuadro 3.6. Estadística sobre co-ocurrencia de patrones del ejemplo (72)

El modelo de Dagan e Itai se compone de dos fases fundamentales: la primera fase, denominada fase de adquisición, en la que se procesa el corpus y se obtienen los datos estadísticos; la segunda fase, denominada fase de desambiguación, que es la fase de resolución de la anáfora en la que los pronombres se "desambiguan" a través de la detección de su antecedente.

Se llevó a cabo un experimento para resolver el pronombre personal neutro *it* en el corpus *Hansard*, formado por las actas del parlamento canadiense. El corpus de prueba se seleccionó manualmente. Para ello, algunas oraciones con el pronombre *it* se extrajeron de forma aleatoria del corpus. De esas se tomaron en consideración únicamente los candidatos contenidos en la mis-

[44] Nótese que en la traducción al español se ha mantenido el carácter neutro del pronombre inglés *it*.

ma oración que la anáfora. Para asegurar un número aceptable de candidatos, se escogieron aquellas ocurrencias del pronombre posteriores a la palabra decimoquinta de la oración, con lo que el número medio de candidatos por anáfora era de 2,8. Asimismo, se eliminaron aquellos casos en los que el pronombre no tenía un sintagma nominal como antecedente, los pronombres pleonásticos[45] y las anáforas que no estaban involucradas en una relación del tipo *sujeto-verbo*, *verbo-objeto* y *adjetivo-nombre*. También se eliminaron los casos en los que la anáfora tenía un único candidato. En total se suprimieron las dos terceras partes del texto original y quedaron un total de 59 ejemplos.

Los datos estadísticos se recuperaron de un texto de 28 millones de palabras al que se le aplicó un análisis sintáctico para detectar los pares de co-ocurrencia. El método no pudo resolver 21 de los 59 casos por no llegar ninguno de los patrones generados al umbral estadístico de 5 apariciones. En los restantes 38, el método de Dagan e Itai resolvió correctamente el pronombre en 33 casos (87 %).

Otro de los experimentos de los autores fue la incorporación de este método como complemento de otros métodos ya desarrollados. En particular, se incorporó el método en el algoritmo de Hobbs (1976)[46]. En la combinación de ambos métodos se prefería un candidato distinto al propuesto por el algoritmo de Hobbs siempre que su co-ocurrencia fuera muy superior (el doble en este experimento). Para este nuevo experimento se extrajeron los datos estadísticos de tres corpora distintos: artículos del Washington Post (40 millones de palabras), artículos del Associated Press News Wire (24 millones de palabras) y el corpus Hansard (85 millones de palabras). Se extrajeron las oraciones con no más de 25 palabras que contenían el pronombre *it*, así como la oración

[45] El *it* pleonástico se corresponde con el pronombre no anafórico, es decir, no es referente de ninguna entidad de discurso previa, sino que es parte de construcciones del tipo "*it is raining*" (*llueve*) o "*it was John who bought it*" (*fue John quien lo compró*).

[46] El algoritmo de Hobbs original proponía un único candidato. Dagan y Itai (1991) modificaron el algoritmo con el fin de que continuara la búsqueda y propusiera más de un candidato.

inmediatamente anterior a cada una de ellas[47]. Además de los casos eliminados en el experimento anterior, en éste se eliminaron también aquellos en los que el analizador no producía un árbol de análisis aceptable. Tampoco se consideraron aquellos casos en los que la relación del pronombre con el verbo no proporcionaba información semántica de interés (por ejemplo, como sujeto del verbo *to be*, *ser-estar*). También se eliminaron los nombres propios como candidatos, así como los casos en los que una anáfora formaba parte de los candidatos a antecedente.

Tras el filtrado, el método trató 74 casos de anáfora pronominal de 3ª persona. El algoritmo de Hobbs resolvió correctamente un 64 % de los casos, porcentaje que fue elevado al 74 % al combinarlo con el método estadístico[48].

El método de Dagan et al. (1995) fue utilizado también para mejorar el algoritmo RAP propuesto por Lappin y Leass (1994) sobre manuales técnicos y se consiguió una leve mejora de un 3 % de éxito aproximadamente. Esta propuesta se detalla en 3.1.2.

Dagan e Itai hacen notar en su trabajo que el modelo que proponen utiliza palabras y no clases semánticas. Desde su punto de vista, el uso de palabras específicas proporciona restricciones más precisas. Asimismo, están de acuerdo en que el uso de clases semánticas favorece la generalización en aquellos casos en los que no existen suficientes datos para patrones específicos y los patrones generales pueden aportar datos adicionales. En este sentido, y tal y como se mostrará más adelante, esta Tesis propone un mecanismo similar al descrito por Dagan e Itai, en el que los datos estadísticos se obtienen de patrones formados por clases semánticas y no sólo por palabras específicas. El siguiente capítulo desarrollará más exhaustivamente esta idea en su sección 4.3 (pág. 113).

---

[47] Hobbs (1976) realiza un estudio sobre la distribución de los pronombres y sus antecedentes y concluye que el 98 % de los antecedentes se encuentran en la misma oración que el pronombre o en la oración anterior.

[48] De los 74 casos, 38 no superaban el umbral para aplicar el método estadístico. Tomados los 36 restantes, el 64 % de éxito del algoritmo de Hobbs ascendió hasta el 86 % al combinarlo con el método estadístico.

3.3.2 La aproximación probabilística de Ge et al.

Ge et al. (1998); Ge (2000) definen un marco estadístico para la resolución del pronombre anafórico de tercera persona.

Para la selección del antecedente, se utiliza una probabilidad fruto de la combinación de distintos factores de resolución anafórica. Estos factores son:

- Distancia: cuanto mayor es la distancia entre el candidato y la anáfora, menor es la probabilidad de que sea el candidato. Como medida de distancia se usa la denominada "distancia de Hobbs", ya que se calcula de la siguiente manera: se ejecuta el algoritmo de Hobbs hasta que propone quince candidatos. El k-ésimo candidato propuesto se dice que se encuentra a una distancia de Hobbs igual a k.
- Género y número: este factor responde a la característica ya utilizada por otros autores de que la anáfora y el antecedente coinciden en género y número.
- Animación: el rasgo de 'animado' o 'no animado' del candidato aporta también información sobre la probabilidad de ser antecedente.
- Información de núcleo dominante[49]: se calcula la probabilidad de que un candidato específico adquiera el mismo papel sintáctico que la anáfora.
- Número de apariciones: con este factor se favorece a aquellos sintagmas nominales que más se repiten a lo largo del texto.

Las probabilidades asociadas a estos factores se multiplican y combinan para cada candidato. Se propondrá como antecedente aquel con la mayor probabilidad.

La evaluación se realizó sobre el 90 % (el restante 10 % se usó para entrenamiento del sistema) de un fragmento del corpus Pen TreeBank, formado por textos procedentes del *Wall Street Journal* con 93931 palabras con 2477 pronombres, de los que 1371 eran pronombres personales en singular. El corpus fue etiquetado

[49] Este término traducido del inglés *governing head information* es análogo al concepto de patrones de co-ocurrencia empleado por Dagan y Itai (1991); Dagan (1992). Para ampliar información sobre éstos, véase 3.3.1.

manualmente con los índices de referencia y el número de apariciones de cada sintagma nominal. Se excluyeron también las apariciones del *it* pleonástico. En la evaluación, los autores comprueban la eficacia de cada uno de estos factores por separado. La distancia de Hobbs resuelve un 65,3 % de los casos. Este porcentaje es incrementado por la información de género, número y animación hasta el 75,7 %. El factor de información de núcleo dominante (patrones de co-ocurrencia) sólo incrementó el porcentaje hasta el 77,9 %. Por último, el número de apariciones incrementó el porcentaje de éxito global hasta el 82,9 %.

3.3.3 La resolución de Cardie y Wagstaff basada en agrupamientos

Cardie y Wagstaff (1999) describen la resolución de la anáfora como un problema de agrupamientos[50]. Cada sintagma nominal queda definido a través de un vector formado por once características y sus valores. El algoritmo agrupará los sintagmas nominales en clases de equivalencia según los valores de esas características. Esta aproximación no utiliza fuentes de conocimiento tal y como lo hacen los sistemas vistos hasta ahora, ya que trabaja sobre la salida de un simple detector de sintagmas nominales y utiliza algunas heurísticas combinadas con WordNet[51] y listas de palabras, sin requerir análisis sintáctico de ningún tipo.

Para la detección de sintagmas nominales los autores usan el localizador de sintagmas nominales Empire (Cardie y Pierce, 1998), que sólo extrae sintagmas nominales simples[52]. Las características asociadas a cada sintagma nominal son:

- Palabras individuales: se almacena el número de palabras que contiene el sintagma nominal.

[50] Término que traduce el inglés *clustering*.

[51] Información más detallada sobre este recurso léxico puede encontrarse en 4.3.4 (pág. 120).

[52] Entendemos por sintagmas nominales simples (frente a los compuestos) aquellos que no contienen otro sintagma nominal en su interior. Por ejemplo, "*El teléfono de Luis*" es un sintagma nominal compuesto que contiene dos sintagmas nominales simples: *el teléfono* y *Luis*.

- Núcleo: la última palabra de cada sintagma nominal es considerada el núcleo[53].
- Tipo de pronombre: los pronombres se marcan como nominativos (*he*, *she*, ...), acusativos (*him*, *her*, ...) o ambiguos (*you*, *it*).
- Artículo: cada sintagma nominal se marca como definido –si está introducido por un artículo definido (*the*)–, como indefinido –si el artículo es indefinido (*a*, *an*)– o queda sin marca.
- Si el sintagma nominal está entre comas, se considera apositivo[54].
- Número: si el núcleo termina en "s", el sintagma nominal es considerado plural.
- Nombre propio: Utilizando una heurística basada en la situación de mayúsculas y minúsculas se determina si un sintagma nominal es un nombre propio.
- Clase semántica: Los autores usan WordNet para extraer, a través de su núcleo, una de las siguientes características del sintagma nominal: *time* ('tiempo'), *city* ('ciudad'), *animal* ('animal'), *human* ('humano') y *object* ('objeto'). Otro algoritmo asigna las clases semánticas de *number* ('número'), *money* ('dinero') y *company* ('compañía').
- Género: el género masculino, femenino o neutro se obtiene de WordNet. Una lista de nombres comunes sirve para asignar género a nombres propios.
- Animación: los sintagmas nominales etiquetados como 'humano' o 'animal' se anotan como 'animado'. Los demás se etiquetan como 'inanimado'.

Para realizar los agrupamientos, se utiliza el concepto de distancia, que define a partir de qué umbral dos sintagmas nominales pueden formar parte de la misma clase de equivalencia. Esta distancia se calcula con la fórmula siguiente:

[53] Conviene recordar que se trata de sintagmas nominales simples y en inglés, por lo que esta característica, aunque no aplicable a otros idiomas (como el español, en el que los modificadores del nombre suelen ir después de éste) es perfectamente válido en este caso.

[54] Los autores reconocen que esta forma de determinar si un sintagma nominal está dentro de una estructura de aposición es muy restrictiva.

$$dist(SN_i, SN_j) = \sum_{f \in F} w_f \times incompatibilidad_f(SN_i, SN_j)$$

donde F es el conjunto de características de cada sintagma nominal, la función $incompatibilidad_f$ indica el grado de incompatibilidad entre la característica f de SN_i y de SN_j, w_f muestra la importancia (peso) relativa de la compatibilidad con respecto a la característica f. En realidad, estas características son un conjunto de restricciones (marcadas por pesos con valores ∞ y $-\infty$) y preferencias (marcadas con valores enteros). De esta forma, si el peso asociado es ∞, la compatibilidad entre los dos sintagmas nominales es imposible (distinta clase semántica, distintos rasgos morfológicos y de animación). Por otro lado, si el peso asociado es $-\infty$, entonces la pertenencia de ambos a la misma clase de equivalencia es clara (si uno incluye al otro como subcadena o si uno es aposición y viene a continuación del otro), siempre que no se de una condición contraria (∞). En cuanto a las preferencias, el algoritmo asigna una serie de valores en otros casos en los que compara la posición, los núcleos o el número de palabras de los sintagmas nominales.

La evaluación fue realizada para la tarea MUC-6 (MUC-6, 1995) en ambos modos *dry run* ("ejecución seca") y *formal evaluation* ("evaluación formal"). En el primero, el algoritmo obtuvo 48,8 % y 57,4 % de cobertura y precisión, respectivamente, con una medida F de 52,8 %. En el segundo, los resultados de cobertura y precisión fueron de 52,7 % y 54,6 %, respectivamente, con una medida-F de 53,6 %.

3.3.4 Las técnicas automáticas de Aone y Benett

Dentro de las aplicaciones de aprendizaje automático[55] y sobre la base de un trabajo previo de resolución de la anáfora multilingüe (Aone y McKee, 1993), Aone y Bennett (1995) describen un sistema de resolución de la anáfora en japonés que trabaja sobre un corpus de artículos de prensa etiquetados con información del discurso (Aone y Bennett, 1994).

[55] Conocido en la bibliografía en inglés como *Machine Learning*.

El sistema resolutor de aprendizaje automático *MLR* (*Machine Learning Resolver*) utiliza un árbol de decisión entrenado con un conjunto de vectores de características asociadas a la anáfora y a los antecedentes. Estos vectores de pueden ser unarios (representando características individuales de anáfora o candidato, como el género o el número) o binarios (representando relaciones entre anáfora y candidato, como la distancia entre ellos).

MLR usa un conjunto de 66 vectores de características que incluyen información léxica (p.ej. categoría), sintáctica (p.ej. papel sintáctico), semántica (p.ej. clase semántica) y posicional (p.ej. distancia entre anáfora y antecedente). El tratamiento del corpus a través del análisis léxico, sintáctico y semántico del sistema de PLN proporciona los valores de estas características y crea los marcadores del discurso para cada sintagma nominal y oración.

Para el entrenamiento utilizan métodos diferentes basados en los tres parámetros siguientes:

- *Cadenas anafóricas*: este parámetro se usa para seleccionar, por un lado, un conjunto de ejemplos positivos y, por otro, un conjunto de ejemplos negativos. Cuando el parámetro está activado, los ejemplos positivos para cada anáfora son todos los pares formados por la anáfora y cualquier sintagma nominal anterior que se encuentre en la misma cadena de correferencia que la anáfora. Los ejemplos negativos corresponden a los pares formados por la anáfora y cualquier sintagma nominal anterior no incluido en la cadena de correferencia.
- *Identificación del tipo de anáfora*: este parámetro se usa para entrenar los árboles de decisión. Cuando este parámetro está activado, el árbol de decisión se entrenará para que dé una respuesta negativa en el caso de que la anáfora y el candidato no correfieran o para devolver el tipo de anáfora cuando son correferentes. Si está desactivado, un árbol binario de decisión será entrenado únicamente para dar una respuesta positiva o negativa sin indicar el tipo de anáfora.
- *Factor de confianza*: con un valor de 0 a 100 (en concreto, valores de 25 %, 50 %, 75 % y 100 %), sirve para realizar podas en el árbol de decisión. Un factor de confianza mayor realiza menos

podas. Un factor de confianza menor, realiza más podas y genera un árbol más pequeño y generalizado.

El entrenamiento se realizó con un corpus sobre fusiones de empresas con un total de 1971 anáforas, de las que 929 eran nombres propios, 546 casi-cero-pronombres[56], 282 cero-pronombres y 82 descripciones definidas.

La evaluación, realizada sobre un corpus de fusiones de empresas, se lleva a cabo sobre seis modos diferentes del sistema, cada uno de ellos con diferentes valores en los parámetros antes mencionados. Los resultados de la evaluación se proporcionan en función de las anáforas detectadas por el sistema, y no en función de todas las anáforas del texto[57]. Las medidas utilizadas para la evaluación son la precisión, la cobertura y la *medida-F*[58] definidas de la siguiente forma:

$$\text{precisión} = \frac{N_c}{N_t} \qquad \text{cobertura} = \frac{N_c}{N_a} \qquad F = \frac{(\beta^2 + 1{,}0) \times P \times R}{\beta^2 \times P + R}$$

donde N_a es el número de anáforas detectadas por el sistema, N_c es el número de anáforas resueltas correctamente, N_t es el número de anáforas tratadas, P es la precisión, R es la cobertura y β es el índice de importancia dado a la cobertura sobre la precisión (en este caso $\beta = 1$).

Utilizando F como la medida global de comportamiento, los mejores resultados[59] sobre 1139 anáforas del corpus de evaluación fueron los correspondientes al modo del sistema en el que el parámetro de cadenas de correferencia estaba activado y el de identificación de tipo, desactivado. Los índices de cobertura resultantes se encuentran entre el 67,53 % y el 70,20 %, los de precisión

[56] La diferencia que establecen los autores entre los cero-pronombres y los casi-cero-pronombres reside en que los segundos se refieren al sujeto de la cláusula inicial de una oración compleja con más de una cláusula y uno o más cero-pronombres.

[57] Este concepto hace que la precisión y cobertura definidas por Aone y Benett difieran sensiblemente de las tratadas por otros autores en sus trabajos (Baldwin, 1997; Gaizauskas y Humphreys, 1996).

[58] Traducción literal del término inglés *F-measure* (Aone y Bennett, 1995).

[59] El resto de los resultados de los diferentes experimentos puede consultarse en (Aone y Bennett, 1995, 1996).

entre el 83,49 % y el 88,55 % y los de la *medida-F* entre el 76,27 % y el 77,27 %.

3.3.5 El algoritmo genético de Byron y Allen

Byron y Allen (1999) definen un enfoque de resolución de la anáfora basado en un conjunto de módulos inspirados en estudios previos para conseguir un factor de relevancia (*salience*) para cada antecedente:

- Incrementan la relevancia del candidato seleccionado por el algoritmo de Hobbs (1986).
- Disminuyen la relevancia del estilo indirecto (Kameyama, 1997a).
- Disminuyen la relevancia de los sintagmas nominales indefinidos (Mitkov, 1998).
- Incrementan la relevancia del primer sintagma nominal en la oración (Mitkov, 1998).
- Disminuyen la relevancia si está en una oración de relativo (Kennedy y Boguraev, 1996).
- Disminuyen la relevancia si está en un sintagma preposicional (Mitkov, 1998).
- Incrementan el valor de los sujetos
- Incrementan el valor del candidato más reciente

Lo que diferencia este algoritmo de los que le inspiran es la forma en que se asigna el peso a los diferentes factores. Para ello los autores utilizan un algoritmo genético que usa números aleatorios en la primera generación, mutación estándar, cruces y operaciones de réplica para las siguientes. Cada forma individual es el porcentaje de pronombres resueltos correctamente. La población inicial es quince y después de cada generación los cinco individuos más fuertes se pueden reproducir, parando después de veinte generaciones.

Para la evaluación utiliza 3900 oraciones del corpus Treebank, anotado anafóricamente, usando un 70 % para el aprendizaje y un 30 % para el entrenamiento. El índice de éxito obtenido (69.1 %) mejora muy ligeramente los del algoritmo de Hobbs (67.8 %), aunque los autores plantean la posible mejora de su algoritmo con la

incorporación de dos módulos más basados en el índice de aparición y restricciones seleccionales (Ge et al., 1998).

3.3.6 Conclusiones sobre los métodos alternativos

En esta sección se han presentado algunas estrategias de resolución de la anáfora que, aunque pueden hacer uso de información lingüística de origen similar a los métodos anteriores, utilizan técnicas distintas para su aplicación.

Si bien estas aproximaciones presentan ideas interesantes, algunas de las cuales han servido como base de determinadas propuestas de esta Tesis (Dagan y Itai, 1990, 1991), la utilización de estas técnicas no ha sido definitivas en la tarea de resolución de la anáfora al proporcionar resultados similares, tal y como se puede comprobar en el cuadro resumen 3.7.

3.3.7 Conclusiones del capítulo

Antes de concluir este capítulo, parece necesario reflexionar acerca de los resultados proporcionados por los diferentes sistemas aquí presentados. Es muy difícil comparar resultados de unos y otros métodos fundamentalmente porque cada uno de ellos ha seguido procesos de definición, implementación y evaluación absolutamente dispares. De hecho, la aplicación de diferentes análisis, diferentes corpus e incluso diferentes implementaciones hace imposible una comparación entre sistemas para decidir cuál es "el mejor" o simplemente qué sistema es mejor que otro. Trabajos como el de Mitkov (2001) ofrecen un retrato muy acertado de esta situación proponiendo plataformas comunes (Barbu y Mitkov, 2000, 2001) para la evaluación de estos métodos. La creación de concursos internacionales para la resolución de la anáfora a partir de estándares daría algo más de luz sobre los resultados reales de cada uno de los métodos.

No obstante, tal y como se ha visto en este capítulo, el campo de la resolución de la anáfora permite la aplicación de una gran cantidad de técnicas de naturaleza muy variada. Parece claro, a tenor de los datos proporcionados por sus autores, que las

	Autores (sistema)	Idioma	Morfol.	Sintác.	Semán.	Pragm.	Discur.	Corpus usado	Nº pron.	Evaluación
1991	Dagan e Itai	INGLÉS	☑	☑	☑	☑		Actas parlamento canadiense	59	Cob.: 64% Prec.: 87%
1995	Aone y Benett (MLR)	JAPONÉS	☑	☑	☑		☑	Fusiones	1139	Cob.: 67%-70% Prec.: 86%88%
1999	Cardie y Wagstaff	INGLÉS	☑	☑	☑			Pen Tree Bank (WSJ)	-	Cob.: 52,7% Prec.: 54,6%
1999	Byron y Allen	INGLÉS	☑	☑				Pen Tree Bank	-	67,8%
2000	Ge y Charniak	INGLÉS	☑		☑			Pen Tree Bank (WSJ)	1371	82,9%

Cuadro 3.7. Resumen de métodos alternativos

aproximaciones para la resolución de la anáfora han obtenido buenos resultados. De hecho, parece que las mejoras que se pueden conseguir en este campo pasan por la incorporación de nuevas estrategias y recursos adicionales que proporcionen una información más cercana al proceso mental de resolución seguido por el oyente.

Según algunos autores citados (Hobbs, 1978; Mitkov, 2002; Palomar et al., 2001a), y siguiendo también los dictados del sentido común, parece que podemos encontrar en la semántica y en las relaciones ontológicas algunos de estos recursos que, aunque utilizados de forma natural por el oyente humano, resultan algo más complejos de aplicar por su mayor dificultad de representación.

La definición de la anáfora como fenómeno no sólo lingüístico sino específicamente semántico ayuda a comprender trabajos como el presentado en esta Tesis, en la cual la información semántica y ontológica, en combinación con otras estrategias, puede llevar a la consecución de sistemas de resolución de la anáfora con índices de error muy bajos.

4. Método de resolución de la anáfora

En este capítulo trataremos el problema de la resolución de la anáfora desde el punto de vista lingüístico y computacional.

En primer lugar, se realizará un análisis de las fuentes de conocimiento que intervienen en el proceso de resolución de la anáfora, ilustrando cada una de ellas con ejemplos de su aplicación. Asimismo, se repasarán los recursos y herramientas que aportan estas fuentes de información al proceso de resolución.

En esta Tesis se propone un método enriquecido de resolución de la anáfora pronominal en español (ERA). Para plantear este método, se parte de un estudio detallado del propuesto en Palomar et al. (2001a), basado en conocimiento limitado. Este estudio ha llevado a la simplificación y optimización del conjunto de restricciones y preferencias planteado originalmente. A partir de este conjunto de restricciones y preferencias basadas exclusivamente en información morfológica y sintáctica, se planteará la incorporación de fuentes de información adicional que servirán como base metodológica del ERA.

Ambos métodos tratarán las anáforas producidas por los pronombres personales, demostrativos, reflexivos y omitidos de tercera persona, tanto en anáforas intrasentenciales como intersentenciales.

4.1 Origen de las fuentes de información en la resolución de la anáfora

Desde un punto de vista lingüístico, el proceso de resolución de la anáfora pasa por la aplicación de conocimiento procedente de distintas fuentes. En esta sección se tratarán todas y cada una de

las fuentes de conocimiento lingüístico que intervienen en dicho proceso.

Asimismo, se hace necesaria la presencia de un conjunto de recursos y herramientas que proporcionen las distintas fuentes de conocimiento y posibiliten el tratamiento computacional de la resolución de la anáfora.

Para cada una de estas fuentes de conocimiento se tratará la forma en que interviene en el proceso de resolución y, por otro lado, el tipo de recursos o herramientas que permiten su instrumentación dentro de un sistema de procesamiento del lenguaje natural.

4.1.1 Información léxica

La información léxica está contenida en el lexicón, esto es, en el conjunto de unidades léxicas pertenecientes a un sistema lingüístico. Dicha información consta de: la etiqueta relativa a la categoría gramatical de cada unidad lingüística (nombre, verbo, pronombre,...) y de una o varias etiquetas correspondientes a cada uno de los rasgos de subcategorización o de selección que hacen posible que cada unidad lingüística seleccione otra u otras a la hora de combinarse formando las distintas oraciones posibles de una lengua ($\pm$concreto, $\pm$transitivo, ...).

La necesidad de esta información para cualquier tarea de PLN, incluida, naturalmente, la resolución de la anáfora, es evidente. Esta información proporcionada por los lexicones, cuya cobertura depende de su implementación, resulta un valioso recurso para obtener las unidades léxicas que forman el texto.

A partir de un texto a procesar, un analizador léxico se encarga de transformar las secuencias de símbolos en unidades léxicas y, a través de un conjunto de reglas, resolver posibles ambigüedades léxicas categoriales. Estos analizadores se denominan etiquetadores gramaticales[1].

[1] Del inglés *POS taggers* o *Part-of-speech taggers.*

Algunos ejemplos de estos etiquetadores son *relax*[2] (español, catalán e inglés), *TreeTagger*[2] (español e inglés), *Brill's tagger*[3] (inglés) o el propuesto por Pla (2000); Pla y Molina (2001).

4.1.2 Información morfológica

La morfología trata las palabras tomadas independientemente de sus relaciones en la oración y estudia su forma. Por tanto, la información morfológica que proporciona una palabra incluye datos sobre su flexión (género, número, persona, ...), derivación (sufijos, prefijos, ...) y composición (palabras simples, palabras compuestas). Asimismo, es objeto del estudio morfológico la categoría gramatical de las palabras (nombre, verbo, adverbio, ...).

En el proceso de la resolución de la anáfora, todos los rasgos morfológicos de los elementos oracionales intervienen en la selección del antecedente. En (73), la información de género, número y persona del pronombre decide por sí misma a la hora de relacionarlo referencialmente con sus posibles antecedentes.

(73) *Andrés*$_i$ sabe la *combinación*$_j$ de la *caja*$_k$ fuerte. *Él*$_i$ está hoy de viaje.

En (73) pueden descartarse todos los candidatos a antecedente del pronombre *él* excepto *Andrés*, que es el único con el que concuerda en género y número.

Sin embargo, en ocasiones la concordancia morfológica entre la anáfora y su antecedente no se cumple. Tal es el caso de sintagmas nominales con carácter de grupo:

---

[2] Desarrollado por el Grupo de Investigación de Lenguaje Natural del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Politécnica de Cataluña en colaboración con el Laboratorio de Lingüística Computacional de la Universidad de Barcelona. Demostración del etiquetador disponible en `http://nipadio.lsi.upc.es/cgi-bin/demo/demo.pl` (última visita en diciembre 2001).

[3] El etiquetador está disponible en `http://www.cs.jhu.edu/ brill/` y una demostración del mismo se puede encontrar en `http://rayuela.ieec.uned.es/cgi-bin/ircourse/brill.perl` (última visita en diciembre 2001).

(74) La *armada*$_i$ necesita jóvenes con ambición. $\emptyset_i$ Te ofrecen una especialización laboral y un buen sueldo.

En (74) el pronombre personal plural omitido tiene como antecedente un sintagma nominal que, siendo morfológicamente singular, tiene carácter colectivo o de grupo y puede ser referido en plural, como así ocurre. Este fenómeno hace que los sistemas de resolución de la anáfora que aplican restricciones morfológicas estrictamente de concordancia en género y número eliminen el antecedente correcto[4].

La correcta identificación de unidades morfológicas es esencial para cualquier proceso posterior. El análisis morfológico trata de establecer las cadenas de morfemas que forman una palabra, identificando sus rasgos de flexión, composición y derivación.

Si se combina el análisis morfológico con el léxico se puede obtener información morfológica más completa sobre las unidades léxicas ya desambiguadas.

Algunos analizadores morfológicos son *maco+*[5] (español, catalán e inglés) y *PC-KIMMO*[6] (inglés).

4.1.3 Información sintáctica

La sintaxis trata la combinación de las palabras en la frase (Ducrot y Schaffer, 1998). Los problemas principales de los que se ocupa la sintaxis se refieren al orden de las palabras, a los fenómenos de rección (es decir, la manera en que ciertas palabras imponen a otras variaciones de número, género, ...) y a las funciones que las palabras pueden cumplir en la oración.

En la resolución de la anáfora, es esencial contar con las relaciones sintácticas que se establecen, tanto entre el pronombre

[4] En el método propuesto en esta Tesis, este fenómeno es tenido en cuenta para enunciar las llamadas condiciones morfosemánticas (ver apartado 4.3.8).

[5] Desarrollado por el Grupo de Investigación de Lenguaje Natural del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Politécnica de Cataluña en colaboración con el Laboratorio de Lingüística Computacional de la Universidad de Barcelona. Demostración del analizador disponible en `http://nipadio.lsi.upc.es/cgi-bin/demo/demo.pl` (última visita en diciembre 2001).

[6] Disponible en `http://www.sil.org/pckimmo/ntnlp94.html` (última visita en diciembre 2001)

y su antecedente como entre cada uno de ellos y el resto de los elementos sintácticos de la oración.

Con respecto a las relaciones entre el antecedente y la anáfora, éstas se engloban dentro de la denominada anáfora intraoracional, y se fundamentan en un conjunto de teorías que parten de la teoría de rección y ligamento (Chomsky, 1981)[7]. Este tipo de teorías evitan la relación entre un SN y un pronombre al que domine, como en el caso de (75).

(75) *Isabel*$_i$ comió con *ella*$_j$ ayer.

En lo referente a la anáfora interoracional, es decir, aquellas que relacionan dos elementos situados en oraciones distintas, se pueden tener en cuenta algunos rasgos sintácticos, como el del papel desempeñado por el antecedente o la propia anáfora con respecto al verbo al que acompaña.

La obtención de la información sintáctica para las tareas computacionales de PLN supone el uso de un analizador sintáctico. Podemos distinguir dos clases de análisis sintáctico, el análisis parcial o superficial y el análisis completo.

En el análisis superficial, se identifican constituyentes sintácticos aislados. No se establecen relaciones sintácticas entre ellos, con lo que el coste computacional es bajo, a costa de disminuir la profundidad y la compleción. Son analizadores rápidos, fiables y robustos.

El análisis completo, por su lado, es menos robusto y fiable, ya que rechaza cualquier oración que no sea capaz de analizar de forma global. Sin embargo, proporciona información mucho más valiosa, ya que establece enlaces oracionales entre los diferentes elementos sintácticos.

[7] Algunos autores han enunciado teorías fundamentadas en la de Chomsky, como es el caso de las reglas *c-comando* (Reinhart, 1983), que sirven como base en la propuesta de esta Tesis (ver 4.2.2).

Algunos analizadores sintácticos son *SUPP* (Palomar et al., 1999) (análisis parcial en español), *tacat*[8] (parcial y completo en español y catalán) y *Conexor*[9] (Tapanainen y Järvinen, 1997) (análisis completo en inglés y español).

4.1.4 Información semántica

La semántica proporciona el significado de las palabras según el contexto. Gran parte de la información semántica de una unidad léxica se encuentra contenida ya en forma de rasgos semánticos en la descripción de dicha unidad. Esto es, la información semántica es responsable de la correcta combinación de unidades léxicas en un discurso. Por lo que a la relación anafórica se refiere, estos rasgos determinan preferencias y/o restricciones en relación a la correferencia.

(76) El *mono*$_i$ subió al *árbol*$_j$ a coger un *plátano*$_k$ porque $\emptyset_i$ estaba hambriento.

En (76) puede verse un ejemplo de anáfora generada por un pronombre omitido para cuya resolución es necesario aplicar información semántica: la condición de estar hambriento sólo puede estar asociada a un antecedente con el rasgo semántico 'animado'.

Para la aplicación de esta información semántica a la resolución de la anáfora es necesario contar con un recurso léxico que proporcione los sentidos posibles de las palabras, así como con una herramienta de desambiguación del sentido de las palabras (*Word Sense Disambiguation*), que seleccione el correcto de todos los posibles.

---

[8] Desarrollado por el Grupo de Investigación de Lenguaje Natural del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Politécnica de Cataluña en colaboración con el Laboratorio de Lingüística Computacional de la Universidad de Barcelona. Demostración del analizador disponible en `http://nipadio.lsi.upc.es/cgi-bin/demo/demo.pl` (última visita en diciembre 2001).

[9] Demostración del analizador disponible en `http://www.conexor.fi` (última visita en diciembre 2001).

El nacimiento de recursos como WordNet[10] o Mikrokosmos[11] han posibilitado la incorporación de esta fuente de conocimiento a las tareas de PLN.

4.1.5 Información pragmática

Hay que tener en cuenta que en una relación anafórica la correcta interpretación de la misma puede en ocasiones no depender de factores relacionados con el discurso en el que se da, sino con el universo sociocultural previo. Es evidente, por tanto, que la información pragmática, esto es, según Moreno et al. (1999), la relativa al conocimiento general del mundo, a la situación comunicativa concreta y a las presuposiciones e inferencias que conlleva, es fundamental para la resolución de la anáfora.

(77) El *Santo Padre*$_i$ se reunió con *Fidel*$_j$ en La Habana. Al bajar del avión $\emptyset_i$ se arrodilló y besó suelo cubano.

La resolución de la anáfora que plantea el pronombre omitido en el ejemplo (77) requiere del conocimiento de distintos aspectos sociales, culturales, políticos y geográficos (el Santo Padre es el Papa, el Papa siempre besa el suelo del lugar que visita, Fidel es Fidel Castro, Jefe del Gobierno de Cuba, ...).

---

[10] WordNet es una base de datos formada por relaciones semánticas entre los significados de las palabras (llamadas *synsets*), a las cuales se accede como si fuera un tesauro, donde las palabras están agrupadas por sus significados. Dada la importancia de WordNet en este trabajo, el apartado 4.3.4 detalla los aspectos fundamentales de este recurso.

[11] Mikrokosmos es un proyecto orientado a la representación del significado de los textos en lenguaje natural usando un formato multilingüe denominado TMR (*text meaning representation*), que representa el resultado del análisis de un texto de entrada dado en cualquiera de los idiomas soportados y sirve de entrada para el proceso de generación. El sentido del texto de entrada, derivado por el análisis de su información léxica, sintáctica, semántica y pragmática, se representa en el TMR como elementos a interpretar en términos de un modelo del mundo u ontología, tal y como se muestra en (Mahesh y Nirenburg, 1995). El proyecto Mikrokosmos ha sido desarrollado por el Laboratorio de Investigación Computacional (*CRL, The Computing Research Laboratory*) de la Universidad del Estado de Nuevo México. Para más información, puede visitarse `http://crl.nmsu.edu/Research/Projects/mikro/index.html` (última visita en diciembre 2001).

Por otro lado, la información pragmática incluye cierta información referente a la construcción del discurso en el que se desarrolla la anáfora (Moreno et al., 1999).

(78) *Andrés*$_j$ regaló un perro a *Pepe*$_i$ por su cumpleaños. Nuria *le*$_i$ trajo un coche teledirigido.

(79) *Andrés*$_i$ regaló un perro a *Pepe*$_j$ por su cumpleaños. Nuria *le*$_i$ reprendió enfadada.

Así, tanto en (78) como en (79) se define la misma acción inicial, mientras que la diferente interpretación del pronombre anafórico *le* que se infiere permite mantener la cohesión discursiva.

La aplicación de información pragmática en la resolución computacional de la anáfora es una tarea difícil de afrontar. Si bien se pueden definir algunas reglas específicas para resolver casos concretos, el uso de este tipo de conocimiento es una línea de investigación completamente abierta.

4.2 Resolución de la anáfora con conocimiento limitado para el español

4.2.1 Introducción

Los métodos basados en restricciones y preferencias de naturaleza morfológica y sintáctica han sido ampliamente utilizados en la bibliografía sobre la resolución de la anáfora dentro del procesamiento del lenguaje natural (Hobbs, 1976, 1978; Carbonell y Brown, 1988; Rich y Luperfoy, 1998; Lappin y Leass, 1994; Mitkov, 1994; Kennedy y Boguraev, 1996; Baldwin, 1997; Ferrández, 1998; Palomar et al., 2001a). Si bien existen ciertas diferencias en la forma de aplicación de estas restricciones y preferencias, podemos definir básicamente las restricciones como un conjunto de reglas que, a partir de una lista de candidatos, rechazan o eliminan aquellos que son incompatibles con la anáfora, esto es, que no pueden correferir con ella por motivos claros (por ejemplo, diferencia de género). Del mismo modo, podemos definir las preferencias

como un conjunto de reglas que se aplican a los candidatos que, siendo compatibles con la anáfora, tendrán que competir para ser el antecedente de la misma. La aplicación de preferencias intenta establecen un orden en el que el candidato que ocupa la primera posición resulta elegido como el antecedente correcto.

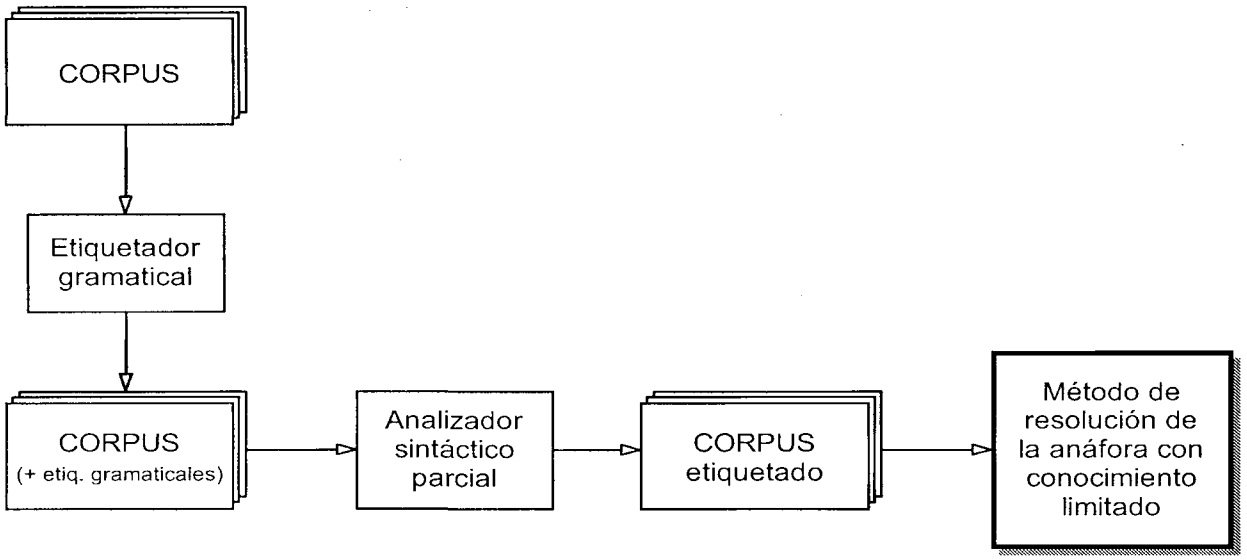


Figura 4.1. Sistema de resolución de la anáfora basado en conocimiento limitado

El método expuesto en esta sección ha sido elaborado a partir de la revisión y optimización del original publicado en Palomar et al. (2001a), que hace uso de fuentes de información morfológicas y sintácticas para la selección del sintagma nominal antecedente de un pronombre. El método se compone de tres fases fundamentales:

- Identificación del pronombre anafórico y de sus candidatos a antecedente.
- Aplicación de restricciones para eliminar candidatos incompatibles.
- Aplicación de preferencias para determinar cuál de los candidatos compatibles es el antecedente.

Este método se encuadra en un sistema de resolución de la anáfora en el que el corpus de entrada ha sido etiquetado tanto con información morfológica (con el uso de un etiquetador gra-

matical), como con información sintáctica (proporcionada por un analizador sintáctico parcial). La figura 4.1 muestra el esquema básico de este sistema.

Los siguientes puntos de esa sección tratarán con detenimiento los factores que intervienen en la definición de restricciones y preferencias, así como los mecanismos usados para su aplicación hasta completar el proceso de selección del antecedente. El esquema general de aplicación de restricciones y preferencias, incluido en el método de conocimiento limitado queda gráficamente representado en la figura 4.2.

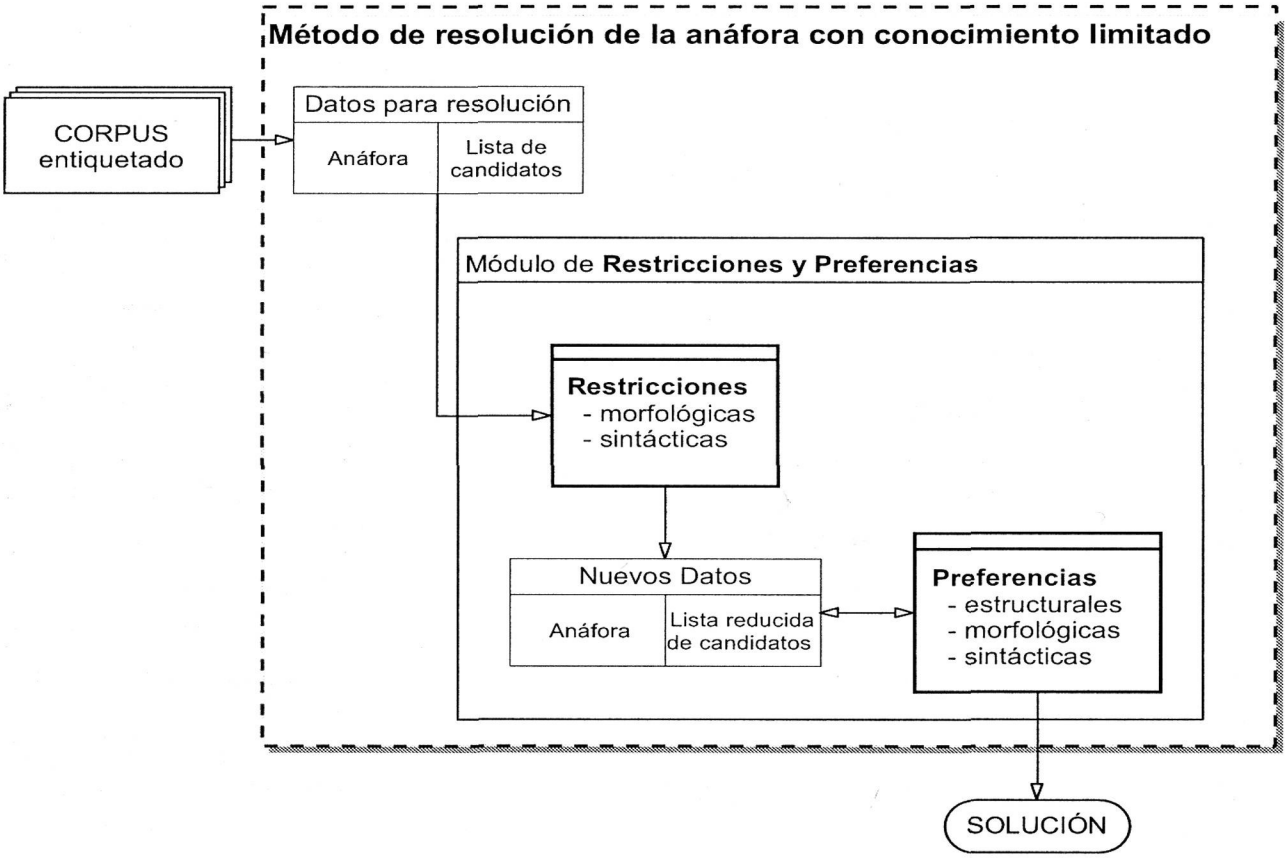


Figura 4.2. Módulo de restricciones y preferencias en el método basado en conocimiento limitado

4.2.2 Restricciones: eliminación de candidatos incompatibles

Tal y como se ha comentado en la introducción, las restricciones son un conjunto de reglas que se aplican para eliminar candidatos no compatibles con la anáfora. En primer lugar, y antes de comenzar con el enunciado de las restricciones, es conveniente detallar las condiciones que hacen que un candidato, o precisando más, un sintagma nominal, sea incompatible con un pronombre. Estas condiciones las llamaremos *condiciones de no correferencia pronombre-SN*, y procederán de dos fuentes de información de distinta naturaleza, la morfológica y la sintáctica.

Condiciones morfológicas de no correferencia pronombre-SN. Según estas condiciones, un SN y un pronombre no serán correferentes si no concuerdan en género, número y persona.

(80) *Andrés*$_i$ sabe la combinación de la *caja*$_j$ fuerte. *Él*$_i$ está hoy de viaje.

Tal y como se puede ver en (80), el SN candidato *la caja fuerte* cuyo núcleo *caja* posee rasgos morfológicos de femenino y singular, no puede correferir con el pronombre masculino singular *él*, mientras que el otro candidato, *Andrés*, sí posee rasgos morfológicos compatibles con el pronombre[12].

Condiciones sintácticas de no correferencia pronombre-SN. Para la definición de estas condiciones, se han tomado como referencia dos fuentes: por un lado, la teoría de la rección y ligamiento (Chomsky, 1981) y, por otro lado, las condiciones de no correferencia definidas en el trabajo sobre la resolución de la anáfora de Lappin y Leass (1994)[13] .

Antes de comenzar a enunciar estas restricciones, se hace necesario aclarar que el punto de partida de este método de conocimiento limitado es el análisis sintáctico parcial del texto, mientras que las teorías sobre las que se sustentan los conceptos sintácticos

[12] Existen algunas excepciones a estas condiciones morfológicas que involucran el uso de semántica y serán tratadas en el apartado 4.3.8.

[13] Véase 3.1.2 (pág. 35) para una exposición detallada de este trabajo.

utilizados suponen un análisis completo sobre el que se establecen las relaciones de comando o dominio. Esto, lógicamente, limita la definición de reglas y, por tanto, se pone de manifiesto en el enunciado de algunas de las condiciones de no correferencia. Dado que el análisis realizado es parcial, no se cuenta con la información sintáctica necesaria para afirmar, por ejemplo, si el SN es el sujeto o no de un verbo, algo que, por tanto, se ha de suponer en función de su posición con respecto al verbo. Así, se entiende que si un SN aparece antes del verbo, puede ser su sujeto, y que si aparece después, no será el sujeto a no ser que no exista ningún SN antes del verbo.

Cada pronombre se enmarca en un contexto sintáctico distinto. Esto hace que las condiciones de no correferencia varíen en función del tipo de pronombre:

1. Un SN no correfiere con un **pronombre reflexivo** si:

   a) El SN está en la misma cláusula[14] e incluido en otro constituyente.

   (81) El *primo*$_j$ de *Luis*$_i$ no *se*$_j$ peina desde los 25; está completamente calvo desde entonces.

   Tal y como se puede ver en (81), el SN *Luis* no puede correferir con el pronombre reflexivo *se* por encontrarse dentro de un sintagma preposicional (introducido por la preposición *de*) e incluido a su vez en un sintagma nominal (*El primo de Luis*).

   b) El SN está en una cláusula u oración diferente a la del pronombre.

   (82) *Lucía*$_j$ entró en la *habitación*$_k$ y *Juan*$_i$ *se*$_i$ miró aterrado en el espejo.

[14] Sobre el concepto de cláusula en el presente trabajo, véase nota 1 (pág. 34).

c) El SN aparece después del verbo y existe otro SN en la misma cláusula antes del verbo[15].

(84) El *pequeño*$_i$ *se*$_i$ lava la *cara*$_j$ cada mañana

2. Un SN no correfiere con un **pronombre personal o demostrativo** si:

a) El SN está en la misma cláusula que el pronombre y está incluido en un SP.

(85) Con *Luisa*$_i$ *la*$_j$ saqué a pasear.

Existe una posible excepción de esta condición de no correferencia que es el doble clítico [16], tal y como se muestra en el ejemplo (86). En este caso, el sintagma nominal correfiere con el pronombre a pesar de estar incluido en un SP. Este tipo de SP son en realidad los complementos directos e indirectos que se ven duplicados con el pronombre, y se introducen por la preposición *a*. Sin embargo, no se puede añadir esta restricción a la condición 2*a* ya que la preposición *a* puede introducir también otro tipo de SP con SN que no sean clíticos duplicados, tal y como se muestra en el ejemplo (87).

(86) A *Luisa*$_i$ *la*$_i$ saqué a pasear.

---

[15] Teniendo en cuenta que el español es un idioma de orden libre, condiciones como ésta podrían no ser operativas ante casos en los que un sintagma nominal no se encuentre en su posición habitual, como en (83), oración que tiene un sentido análogo al de (84) pero que presenta diferente orden de construcción.

(83) La *cara*$_i$ *se*$_j$ la lava el *pequeño*$_j$ cada mañana

Este tipo de problemas sólo puede ser resuelto con un análisis del texto completo o en el que se marquen las relaciones sintácticas entre los componentes oracionales.

[16] Tal y como se ha dicho en el apartado 2.2.3 (pág. 20), los pronombres átonos, a diferencia de los tónicos, especialmente los de complemento indirecto, pueden co-aparecer también con sintagmas nominales plenos, en lo que se conoce como reduplicación o doblado de clíticos (Fernández, 1999): "*Le*$_i$ *di las llaves a ella*$_i$".

(87) A la *calle*$_i$ *la*$_j$ saqué a pasear.

Algunos estudios, procedentes sobre todo de la gramática generativa, tratan el doblado de clíticos como un fenómeno no anafórico (Aoun, 1981), algo que debe ser tenido en cuenta a la hora de aplicar las mismas condiciones de no correferencia.

b) El SN está en la misma cláusula que el pronombre y el pronombre aparece antes del verbo[17].

(88) Bajo el centenario *abedul*$_j$ *él*$_i$ la besó en la mejilla.

Esta regla se justifica por la suposición de que si el pronombre aparece antes del verbo, entonces es el sujeto de dicho verbo.
De nuevo, hay que tener en cuenta que las condiciones se enuncian desde un análisis parcial y que el orden libre del español dificulta especialmente los mecanismos de definición de reglas basados exclusivamente en la posición de los elementos oracionales. Así, la condición de no correferencia que acabamos de enunciar puede no ser válida en ejemplos como el siguiente:

(89) Al propio padre de *Luis*$_i$ *él*$_i$ le grita con frecuencia.

Si bien este ejemplo podría considerarse como falto de naturalidad (sería más natural la frase "*A su propio padre él le grita con frecuencia*", es perfectamente válido desde el punto de vista gramatical y demuestra que, a pesar de estar contenido en un sintagma preposicional, el SN *Luis* puede correferir (no es que lo haga necesariamente, pero puede hacerlo) con el pronombre personal de sujeto *él*.

[17] Debido a que el análisis realizado es parcial, el hecho de que el pronombre aparezca antes del verbo supone que es el sujeto de dicho verbo.

c) El SN está en la misma cláusula que el pronombre, el pronombre aparece después del verbo[18] y el SN no está incluido en otro SN.

(90) El *padre*$_i$ de Germán siempre *le*$_j$ llama a *él*$_j$ cuando hay problemas.

En (90) se puede comprobar el funcionamiento de esta condición. El SN introducido por el núcleo *padre*, que no está incluido en otro SN y está en la misma cláusula que el pronombre[19], no puede correferir con éste. Obsérvese que el SN *Germán* sí podría correferir con el pronombre ya que, a pesar de estar en la misma cláusula, está contenido en otro SN (*el padre de Germán*).

d) El SN está en la misma cláusula que el pronombre, el pronombre está incluido en un SP que no está incluido en otro constituyente y el SN tampoco está incluido en otro constituyente.

(91) La *madre*$_i$ de Isabel trabaja con *ella*$_j$ en la empresa familiar.

En (91) el SN introducido por el núcleo *madre* no puede correferir con el pronombre *ella*, mientras que el SN *Isabel* podría hacerlo al no cumplir la condición por estar incluido en otro SN (*La madre de Isabel*).

e) El SN contiene al pronombre.

(92) En la fiesta apareció súbitamente un *primo*$_i$ de *él*$_j$.

En (92) el SN introducido por el núcleo *primo* (*el primo de él*) no puede correferir con el pronombre contenido en

[18] Por la misma razón aludida en la nota anterior, el hecho de que el pronombre aparezca después del verbo supone que es un complemento (directo, indirecto, circunstancial,...) de dicho verbo.

[19] En este ejemplo también se puede ver el fenómeno del doble clítico anteriormente mencionado.

dicho sintagma *él*.

f) El SN está coordinado con el pronombre.

(93) *Julia*$_i$ y *ella*$_j$ salieron a la misma hora hacia la fiesta.

Evidentemente, la coordinación establece en su enunciado un conjunto de elementos que son disjuntos y que, por tanto, no pueden correferir entre sí, tal y como se puede comprobar en (93).

g) El pronombre está incluido en una oración de relativo introducida por el SN.

(94) Luis tiene una *mujer*$_i$ que *le*$_j$ ama profundamente.

Tal y como se muestra en (94), el SN cuyo núcleo es *mujer* y que introduce a su vez la oración de relativo no puede correferir con el pronombre *le*.
No obstante, puede ocurrir que en la oración de relativo se incluya otra oración de relativo, en cuyo caso el SN que introduce la primera y el pronombre que aparece en la segunda podrían correferir:

(95) Luis es un *hombre*$_i$ que tiene una *mujer*$_j$ que *le*$_i$ ama profundamente.

Todas estas condiciones de no correferencia para pronombres personales y demostrativos son aplicables de forma análoga a pronombres omitidos[20].

4.2.3 Preferencias: la selección del antecedente

Las preferencias son un conjunto de reglas que intentarán discernir cuál de los candidatos que han superado la fase de restricciones resulta ser el antecedente del pronombre.

[20] La detección de pronombres omitidos en este método se ha realizado con el algoritmo definido en Ferrández y Peral (2000).

Gestión de preferencias. La aplicación de las preferencias se puede realizar utilizando dos métodos diferentes:

- *Filtrado*: El sistema de preferencias con filtrado aplica las preferencias en un orden preestablecido. Cada una de las preferencias decide qué candidatos pasarán a la aplicación de la preferencia siguiente (Carbonell y Brown, 1988; Ferrández et al., 1998). Aún cuando esta estrategia puede confundirse con un sistema de restricciones, la diferencia fundamental radica en que, mientras que al aplicar una restricción se eliminan todos los candidatos que no la cumplen, al aplicar una preferencia, si ésta no es satisfecha por ningún candidato, se pasa a la siguiente manteniendo intacta la lista de candidatos.
  Este sistema de aplicación de preferencias se fundamenta principalmente en el orden establecido para la aplicación de las mismas, siendo este orden fundamental en la eficacia del sistema.
  Si tras la aplicación de todas las preferencias queda un único candidato en la lista, éste será considerado el antecedente de la anáfora. En caso de que la lista contenga más de un candidato, entonces se decidirá entre ellos con una preferencia excluyente como, por ejemplo, la de mayor cercanía al pronombre anafórico.
- *Ponderado*: El sistema ponderado de aplicación de preferencias no establece ningún orden concreto de aplicación de las mismas, sino que asigna un peso a cada una de ellas (Mitkov, 1998; Cardie y Wagstaff, 1999). Este peso puede ser positivo, cero e incluso negativo, y contribuye a una puntuación global de cada candidato, de manera que el que obtenga una mejor puntuación será elegido como el antecedente de la anáfora.
  En caso de empate, se podrá usar alguna preferencia que resuelva el conflicto, como la de cercanía a la anáfora.

Parece lógico pensar que la aplicación de preferencias por filtrado podría resultar algo más limitada, ya que un candidato que no supere una de las preferencias será eliminado sin tener posibilidad de comprobar el resto de ellas. Este tipo de aplicación de preferencias ha sido defendida, sobre todo, por su bajo coste

computacional (Ferrández et al., 1998), mientras que el sistema ponderado parece resultar algo más flexible en su aplicación. Esta flexibilidad se puede justificar en la simplicidad de ajuste de pesos en un conjunto de preferencias, así como en la posibilidad de simular de forma inmediata el comportamiento de un sistema filtrado con el uso de un sistema ponderado que dote a cada preferencia, según su orden en el sistema filtrado, de un peso mayor que la suma del de todas las siguientes en dicho orden. Esta capacidad de simulación no es tan evidente en el caso contrario (realizar un sistema de preferencias filtrado que simule cualquier combinación de pesos en un sistema ponderado no es una tarea trivial).

En lo referente a las preferencias propuestas en este método de conocimiento limitado, se ha elegido el sistema de filtrado por sus implicaciones positivas en el coste computacional.

Aprendizaje de preferencias. Las preferencias que se enunciarán a continuación están basadas en el estudio de la importancia de cada tipo de conocimiento que el hombre aplica de forma natural para resolver la ambigüedad y seleccionar el antecedente de un pronombre. Adicionalmente, estas preferencias provienen del propio comportamiento del pronombre. Tal y como ya se ha comentado, el pronombre proporciona una cantidad de información semántica nula, por lo que es necesario, para una correcta resolución de la ambigüedad, que el antecedente no se encuentre demasiado alejado del pronombre[21]. De hecho, algunos pronombres como los reflexivos y los recíprocos, requieren que su antecedente se encuentre en la misma cláusula.

Así, el conjunto de preferencias definido a partir de este estudio es el siguiente:

A) El SN antecedente está en la misma cláusula.
B) El SN antecedente está en otra cláusula.
C) El SN está incluido en otro SN.
D) El SN es un nombre propio.
E) El SN es un SN indefinido.

[21] Este hecho ha llevado a algunos autores a definir una "ventana" o espacio de búsqueda del antecedente para evitar complicaciones computacionales.

F) El SN se ha repetido más de una vez en el texto.
G) El SN ha aparecido más de una vez con el verbo de la anáfora en el texto.
H) El SN ocupa la misma posición que la anáfora con respecto al verbo (antes o después).
I) El SN aparece antes del verbo.
J) El SN no es de tiempo.
K) El SN no es de cantidad.
L) El SN no es de dirección.
M) El SN no es abstracto.

Dado el diferente comportamiento de las distintas clases de pronombre, cada preferencia tiene una influencia distinta en función del tipo de anáfora tratado. Para establecer esta influencia se ha realizado un estudio del corpus de entrenamiento con el objetivo de asociar cada preferencia a cada tipo de anáfora en función de su influencia en el proceso de resolución. El cuadro 4.1 muestra esta relación, en la que cada factor (marcado con la letra correspondiente en la lista anterior) aparece acompañado del número de casos que lo cumplen dentro del corpus de entrenamiento[22], formado por 575 pronombres.

	Personales y Demostrativos	*Pronombres Omitidos*	*Pronombres Reflexivos*
A	74	57	100
B	26	43	0
C	24	4	3
D	27	63	53
E	6	7	0
F	62	79	66
G	18	20	94
H	50	89	84
I	59	89	91
J	100	100	100
K	99	100	100
L	99	100	100
M	100	100	100

Cuadro 4.1. Distribución porcentual de cada factor de preferencia en el corpus de entrenamiento para el método de conocimiento limitado

[22] Los datos relativos al tipo del corpus y su tamaño serán tratados en profundidad en el capítulo de evaluación (apartado 5.2.1, pág 154).

Esta distribución porcentual ha permitido decidir las preferencias que son relevantes según el tipo de pronombre, así como su orden de aplicación.

Conjunto de preferencias. A partir de la distribución porcentual de estos factores para cada tipo de anáfora, se define un conjunto de preferencias a aplicar que, en función del tipo de pronombre, variarán de orden según el estudio de la mencionada distribución porcentual:

- Se prefieren los SN candidatos que aparecen en la misma oración frente a los que aparecen en oraciones anteriores, siendo la preferencia mayor cuanto mayor es la proximidad entre candidato y anáfora. En el caso de los pronombres reflexivos, el candidato debe estar en la misma cláusula, por lo que ya se ha tratado este caso en las restricciones y no aparecerá como preferencia.
- Se prefieren los SN candidatos que ocupan la misma posición que la anáfora con respecto al verbo.
- Se prefieren los SN candidatos que se han repetido más veces en el texto.
- Se prefieren los SN candidatos que no están incluidos en otro SN.
- Se prefieren los SN que no son de tiempo, dirección, cantidad o tipo abstracto[23] (“*las ocho menos cuarto*”, “*calle primavera*”, “*cuarenta*”, “*una cosa*”, ...).

Una vez definidas estas preferencias, se expondrá a continuación la aplicación y el orden de las mismas según el tipo de pronombre a resolver.

Preferencias para pronombres personales o demostrativos.

1. SN que no son de tiempo, dirección, cantidad ni tipo abstracto.
2. SN en la misma oración que el pronombre.

[23] Estos factores, por su contenido semántico, parecen contradecir el carácter puramente morfosintáctico del método. Sin embargo, la detección de este tipo de características se realiza con el uso de reglas y no con ninguna clase de conocimiento semántico adicional a las fuentes ya expuestas.

3. SN en la oración anterior.
4. SN no incluidos en otro SN (por ejemplo, si aparecen en una cláusula de relativo o una aposición).
5. SN que se han repetido más de una vez en el texto.
6. SN que ocupan la misma posición (antes o después) que la anáfora con respecto al verbo.
7. SN que aparecen con el verbo de la anáfora más de una vez.

Preferencias para pronombres omitidos.

1. SN que no son de tiempo, dirección, cantidad ni tipo abstracto.
2. SN en la misma oración que el pronombre.
3. SN en la misma oración que el pronombre y que además ha sido solución para otro pronombre omitido.
4. SN en la oración anterior.
5. SN no incluidos en otro SN (por ejemplo, si aparecen en una cláusula de relativo o una aposición).
6. SN que aparecen antes del verbo.
7. SN que se han repetido más de una vez en el texto.

Preferencias para pronombres reflexivos.

1. SN que no son de tiempo, dirección, cantidad ni tipo abstracto.
2. SN no incluidos en otro SN (por ejemplo, si aparecen en una cláusula de relativo o una aposición).
3. SN que aparecen antes del verbo.

Preferencias comunes. En el caso de que la aplicación del conjunto de preferencias anteriormente expuestas genere un "empate" entre dos o más candidatos y, por tanto, no proporcione el antecedente del pronombre, es necesario aplicar alguna clase de preferencia de carácter más genérico y excluyente. Estas preferencias han sido establecidas empíricamente y se aplican en el orden dado para determinar el antecedente:

1. SN más repetido en el texto.
2. SN que ha aparecido más con el verbo de la anáfora.
3. SN más cercano al pronombre.

Como puede verse, en el caso extremo en el que, tras haber aplicado las dos primeras preferencias comunes todavía haya más de

un candidato en la lista, se seleccionará como antecedente anafórico el candidato más cercano al pronombre.

4.2.4 La aplicación del método de conocimiento limitado

Una vez expuestas las condiciones de no correferencialidad que permitirán la eliminación de candidatos a antecedente y el conjunto de preferencias que intervienen en la selección del candidato más apropiado, veamos el modo en que estas restricciones y preferencias se aplican en el proceso de resolución de la anáfora. Para ello, definiremos un sencillo algoritmo que muestre las etapas que intervienen en el método de resolución. Este algoritmo se muestra en la figura 4.3.

```
------------------------------------------------------------------------
Para cada oración O
  L = L + Almacenar los SN de O
  Para cada pronombre P en O
    Identificación de tipo del pronombre P
    L' = Aplicación de restricciones a L
    Si |L'| = 0 entonces P es exofórico
    Si |L'| = 1 entonces L[1] es el antecedente de P
    Si |L'| >1 entonces
      L''=Aplicación de preferencias a L' según el tipo de P
      Si |L''| = 1 entonces L[1] es el antecedente de P
      Si |L''| >1 entonces
        A=Aplicación de preferencias comunes a L''
        A es el antecedente de P
      finSi
    finSi
  finPara
finPara
------------------------------------------------------------------------
```

Figura 4.3. Algoritmo de aplicación del método de conocimiento limitado (Palomar et al., 2001a).

Este algoritmo no tiene en cuenta el espacio de búsqueda del candidato. Este espacio se define de forma diferente en función del tipo de pronombre. La definición de este espacio de búsqueda es vital para establecer un equilibrio entre la eficacia del sistema de resolución y el coste computacional asociado al mantenimiento de

la lista de candidatos. Tal y como se ha indicado, los pronombres reflexivos tienen su antecedente en la misma cláusula, mientras que los pronombres demostrativos, personales u omitidos podrán buscar su antecedente en la misma oración o incluso en oraciones anteriores. Así, diferentes autores proponen distintos espacios de búsqueda a partir de estudios sobre los textos tratados (Hobbs, 1976; Baldwin, 1997; Mitkov, 1998). En este trabajo, y a partir de un exhaustivo estudio del corpus, se ha definido un espacio de un máximo de cuatro cláusulas para la búsqueda del antecedente anafórico.

4.3 ERA: método enriquecido de resolución de la anáfora para el español

4.3.1 Introducción

Tal y como se verá en la fase de evaluación, el método propuesto anteriormente, al igual que otros enfoques basados en conocimiento limitado, ha demostrado obtener buenos resultados[24]. No obstante, la mayoría de los trabajos relevantes en esta línea concluyen con la necesidad de incorporar información semántica al proceso de resolución. En este sentido, la resolución de la anáfora pronominal en español no ha contado hasta ahora con una estrategia que integre de manera automática la semántica dentro de sus fuentes de información.

El método propuesto en la sección anterior está basado en información puramente morfosintáctica obtenida del uso de un etiquetador gramatical y un analizador sintáctico parcial. El método que proponemos en esta sección requiere, además de la anterior, de un conjunto de fuentes de información adicionales que mejoren los resultados de la resolución anafórica. Estas fuentes de información proceden, por un lado, de un enriquecimiento del análisis sintáctico parcial y, por otro, del uso de información semántica en el proceso de resolución.

[24] La sección 3.1 (pág. 30) explica con detalle el conjunto de estrategias para la resolución de la anáfora basadas en conocimiento limitado.

En lo referente al enriquecimiento del análisis sintáctico parcial, proponemos un conjunto adicional de etiquetas de carácter sintáctico y semántico. Las etiquetas sintácticas marcarán los papeles que los elementos oracionales analizados tienen con respecto al verbo. Esto permitirá redefinir las restricciones con información del papel sintáctico eliminando las conjeturas (a veces fallidas debido al propio orden libre del lenguaje que provoca dislocación o movimiento de elementos oracionales) basadas en la posición del sintagma nominal y del pronombre con respecto al verbo. Las etiquetas semánticas indicarán los sentidos correctos de los componentes textuales. Este sentido correcto permitirá el uso de la semántica en el proceso de resolución anafórica.

El método elaborará la información semántica usando dos técnicas diferentes:

- *Semántica basada en corpus*: se utilizarán los conceptos ontológicos asociados a los candidatos anafóricos y se relacionarán con el verbo de la anáfora. De esta manera, se definirá un conjunto de patrones semánticos u ontológicos que aportarán información de compatibilidad semántica para la resolución de la anáfora en la fase de aplicación de preferencias.
- *Semántica basada en conocimiento*: se definirán un conjunto de reglas de incompatibilidad semántica entre el antecedente y el pronombre que se aplicarán en la fase de restricciones para eliminar candidatos incompatibles.

Así, esta sección desarrollará las siguientes propuestas:

- Etiquetado morfológico, sintáctico, semántico y anafórico necesario para la aplicación del método.
- Obtención de reglas de compatibilidad y de incompatibilidad anafórica basadas en la semántica.
- Método enriquecido de resolución de la anáfora (ERA) basado en restricciones y preferencias.

En primer lugar, se detallará la propuesta de anotación sintáctica y semántica adicional. En segundo lugar se hablará de las posibilidades que EuroWordNet brinda como recurso utilizado para la extracción de información semántica así como la forma en que

esta información es usada en el proceso de resolución de la anáfora. Los últimos apartados de esta sección se dedicarán al método en sí, exponiendo el conjunto de restricciones y preferencias que utiliza, así como su esquema de aplicación.

4.3.2 Requisitos de aplicación del método

El método ERA se encuadra dentro de un sistema completo compuesto por un conjunto de elementos que le proporcionan la entrada. El esquema básico de este sistema queda recogido en la Figura 4.4

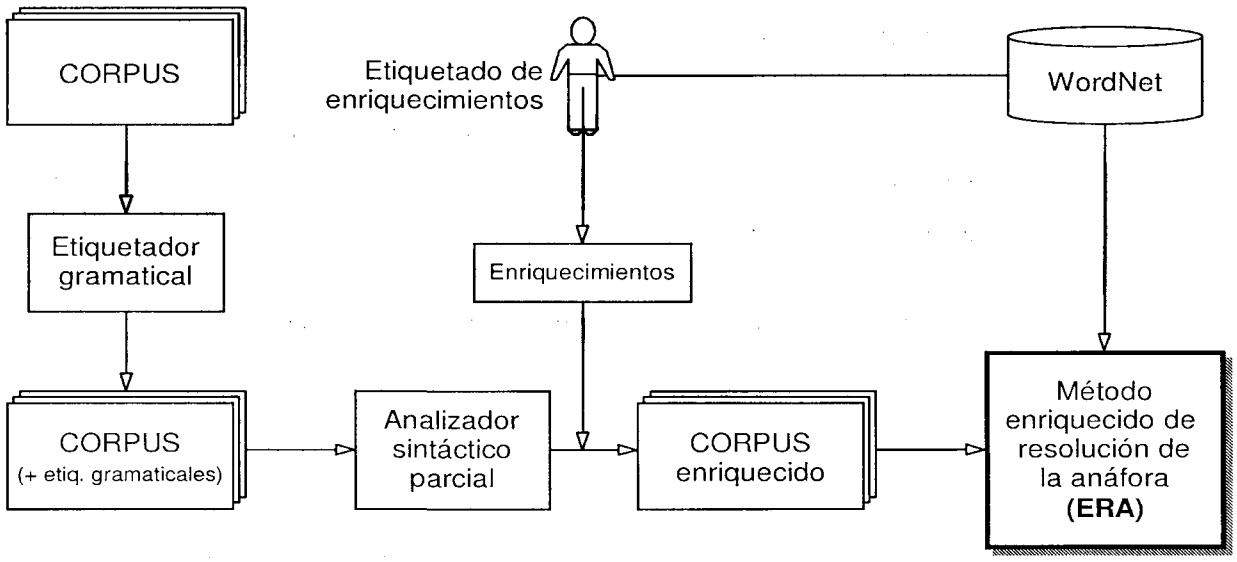


Figura 4.4. El sistema de resolución de la anáfora basado en el método enriquecido

Los requisitos de aplicación del método ERA proceden básicamente de dos fuentes:

- *Corpus enriquecido*: el corpus de entrada atravesará, en primera instancia, una fase de análisis morfológico en el que a cada palabra se le asignará su categoría gramatical así como una etiqueta de rasgos morfológicos. A continuación, el corpus será procesado por un analizador parcial que, a partir de un conjunto de reglas definidas por una gramática, etiquetará las estructuras sintácticas. A este análisis sintáctico se agregará un conjunto adicional de etiquetas para marcar, por un lado, los papeles sintácticos de

los elementos oracionales y por otro, los sentidos correctos de las palabras a partir del recurso léxico WordNet. El apartado 4.3.3 expone con mayor detalle en qué consiste esta propuesta de etiquetado.

- *WordNet*: El recurso léxico WordNet será consultado por distintos módulos del método ERA para la incorporación de la información semántica y ontológica. El apartado 4.3.4 explica en profundidad los aspectos más relevantes de WordNet en lo referente al contenido de esta Tesis mientras que los apartados 4.3.5 y 4.3.6 detallan el proceso de integración de la semántica en el método ERA.

4.3.3 Propuesta de etiquetado del corpus

Dadas las características de este método, es necesario contar con información adicional a la proporcionada por el análisis sintáctico parcial[25]. La necesidad de este etiquetado surge ante la escasa disponibilidad de recursos en español (y de corpus en particular) que incluyan este tipo de información.

Además de la morfología de cada palabra (género, número y persona) y la sintaxis que agrupa las palabras en componentes oracionales más complejos (sintagmas nominales, sintagmas preposicionales, sintagmas verbales,...), el método ERA requiere de un conjunto de fuentes de información adicionales. A continuación se describe cada una de las fuentes de información que el método ERA requiere, propuestas en diferentes niveles y que conforman el etiquetado requerido para la resolución adecuada de la anáfora:

1. *Nivel morfológico*: cada palabra va acompañada de una etiqueta que especifica su información léxico-morfológica relativa a su categoría gramatical, a su lema y a sus rasgos morfológicos de género número y persona. Para ello se ha usado el conjunto de etiquetas PAROLE definido en el proyecto ITEM (Martí et al., 1998). Este etiquetado se realiza de forma automática (Padró, 1997; Atserias et al., 1998) .

[25] Uno de los objetivos de este etiquetado es simular el la salida de un analizador completo en lo referente al árbol de dependencias sintácticas del verbo. Siguiendo la línea planteada en esta Tesis, en (Saiz-Noeda et al., 2000a, 2001a) pueden encontrarse propuestas basadas en este tipo de análisis.

2. *Nivel sintáctico*: en este nivel se proponen dos conjuntos de etiquetas que representan la forma sintáctica y el papel sintáctico de los constituyentes:

   a) La forma sintáctica, referida al tipo de sintagma que se etiqueta, recoge todo el conjunto de sintagmas reconocidos por la gramática en el análisis:
   - Sintagma nominal omitido (NP*)
   - Sintagma nominal (NP). Este tipo de sintagma nominal puede tener como núcleo:
     - nombre (NUCL NOUN)
     - pronombre (NUCL PRON)
     - verbo (NUCL VERB)
   - Sintagma verbal elidido o no. Si está elidido (VP* REF) hará referencia a un verbo aparecido con anterioridad. Si no está elidido, podrá tratarse de un sintagma verbal en activa (VP) o en pasiva (VP_PASS). El núcleo de un sintagma verbal puede ser:
     - verbo en activa simple (NUCL VERB)
     - verbo omitido (NUCL VERB* REF)
     - verbo en pasiva simple (NUCL PASS)
     - perífrasis verbal (NUCL VPER)
     - verbo pronominal (NUCL VERB PRON)
   - Sintagma preposicional simple (PP) o compuesto (PPC).
   - Sintagma adverbial (ADVP) cuyo núcleo puede ser:
     - adverbio (NUCL ADV)
     - verbo en gerundio (NUCL GER)
   - Sintagma adjetivo (ADJP) cuyo núcleo puede ser:
     - adjetivo (NUCL ADJ)
     - verbo en participio (NUCL PART)

   La base de este etiquetado sintáctico la proporciona el analizador parcial SUPP (Ferrández et al., 1998).

   b) El papel sintáctico de los componentes oracionales, en particular los subcategorizados por el verbo:
   - Sujeto (SUBJ)
   - Sujeto paciente (SPAC)
   - Atributo (ATRB)

- Objeto directo (OD)
- Objeto indirecto (OI)
- Complemento de régimen preposicional (CPREP)
- Complemento agente (CAGT)

3. *Nivel semántico*: la propuesta de anotación semántica engloba dos niveles de etiquetado diferentes:
   a) Etiquetado léxico-semántico: cada núcleo nominal, adjetival, verbal y adverbial se acompaña de su sentido correcto en WordNet. Dada el fino granulado que presenta este recurso, en ocasiones un término puede encajar con más de un sentido en cuyo caso la etiqueta contendrá a todos ellos. Si por el contrario la palabra no está en WordNet o el sentido que toma en el texto no esta recogido en ninguno de los de ese término, la palabra permanece sin etiqueta.
   b) Etiquetado sintáctico-semántico: los sintagmas preposicionales y adverbiales se acompañan de etiquetas semánticas de tipo localización (LOC), temporal (TIME) o modal (MOD).
4. *Nivel anafórico*: orientado fundamentalmente a la evaluación del método, se propone la inclusión de etiquetas de referencias anafóricas de manera que cuando un elemento es anafórico, se acompaña del identificador del SN al que hace referencia (REF *id*).
5. *Nivel estructural*: un conjunto de etiquetas adicionales delimitan unidades estructurales como la oración (S), la cláusula (C) o el párrafo (P).

El enriquecimiento manual del etiquetado del corpus se ha realizado sobre fragmentos del corpus Lexesp previamente etiquetado morfológicamente y analizado sintácticamente[26]. Para realizar en etiquetado adicional, se ha adaptado la salida del analizador parcial a un formato estilo TreeBank. El cuadro 4.2 muestra una comparación entre el análisis generado por el analizador sintáctico y el resultado del enriquecimiento sobre un fragmento del corpus Lexesp, en el que se han resaltado los cambios realizados.

[26] El apartado 5.3.1 describe con detalle todas las herramientas usadas para el preproceso del corpus.

```
(S 1
 (C
  <Cuando> "cuando" CS00 WNx
  (NP*:1,1 ROL:X
   (PRON ROL:X REF:R)
  )
  (VP:2
   (VERB
    <escribo> "escribir" VMIP1S0 WNx
   )
  )
  (NP:3,1 ROL:X
   (PRON:4,1 ROL:X REF:R
    <esto> "esto" PD3CS000 WNx
   )
  )
 )
 (C
  (NP:5,1 ROL:X
   (DET
    <la> "la" TDFS0 WNx
   )
   (NOUN
    <Madre_Coraje> "madre_coraje" NP00000 WNx
   )
   (ADJ
    <peruana> "peruano" AQ0FS00 WNx
   )
  )
  (VP:6
   (VERB
    <acaba> "acabar" VMIP3S0 WNx
   )
   (VPER
    <de> "de" SPS00 WNx
    (VERB
     <ser> "ser" VAN0000 WNx
    )
   )
  )
  (NP:7,1 ROL:X
   (VERB
    <reventada> "reventar" VMPP0SF WNx
   )
   (PP
    (PREP
     <por> "por" SPS00 WNx
    )
    (NP:8,1 ROL:X
     (DET
      <los> "el" TDMP0 WNx
     )
     (NOUN
      <senderistas> "senderista" NCCP000 WNx
     )
    )
   )
  )
  <.> "." Fp WNx
 )
)
```

```
(S 1
 (C
  <Cuando> "cuando" CS00 WN1
  (NP*:1,1 ROL:SUBJ
   (PRON ROL:X REF:R)
  )
  (VP:2
   (VERB
    <escribo> "escribir" VMIP1S0 WN2,3
   )
  )
  (NP:3,1 ROL:OD
   (PRON:4,1 ROL:OD REF:R
    <esto> "esto" PD3CS000 WNx
   )
  )
 )
 (C
  (NP:5,1 ROL:SPAC
   (DET
    <la> "la" TDFS0 WNx
   )
   (NOUN
    <Madre_Coraje> "madre_coraje" NP00000 WNx
   )
   (ADJ
    <peruana> "peruano" AQ0FS00 WN1
   )
  )
  (VP:6
   (VPER
    (AUX
     (VERB
      <acaba> "acabar" VMIP3S0 WNx
     )
     (PREP
      <de> "de" SPS00 WNx
     )
     (VERB
      <ser> "ser" VAN0000 WN1
     )
    )
    (PRN
     (VERB
      <reventada> "reventar" VMPP0SF WNx
     )
    )
   )
   (PP: ROL:AG
    (PREP
     <por> "por" SPS00 WNx
    )
    (NP:8,1 ROL:X
     (DET
      <los> "el" TDMP0 WNx
     )
     (NOUN
      <senderistas> "senderista" NCCP000 WNx
     )
    )
   )
  )
 <.> "." Fp WNx
 )
)
```

Cuadro 4.2. Comparación entre el etiquetado sintáctico parcial (izquierda) y el etiquetado enriquecido (derecha)

4.3.4 La información semántica desde WordNet y EuroWordNet

Una de las características fundamentales del método enriquecido es que se trata de una propuesta fundamentada no sólo en la sintaxis, sino también en el uso de ontologías y relaciones semánticas como una fuente de información adicional para el proceso de resolución de la anáfora.

La información semántica agregada será extraída de WordNet, un recurso léxico ampliamente extendido en los trabajos de investigación y utilizado en tareas de Procesamiento del Lenguaje Natural. En este apartado se expondrán las características más relevantes de este recurso.

Introducción. WordNet, tal y como describe Miller (1993), es un diccionario electrónico que almacena conjuntos de sinónimos denominados *synsets*. Cada *synset* describe un concepto semántico y contiene una lista de pares palabra-sentido así como punteros a otros *synsets* en forma de relaciones semánticas. De esta manera, los distintos sentidos de una palabra se almacenan en WordNet en *synsets* distintos. Además, cada *synset* puede ir acompañado de una definición o glosa como ocurre en los diccionarios convencionales.

EuroWordNet, desarrollo más reciente basado en el WordNet inglés (versión 1.5), es una base de datos léxica multilingüe que representa las relaciones semánticas entre conceptos básicos de idiomas europeos (Vossen, 2000). Consiste en un conjunto de WordNets para varios idiomas (inglés, holandés, español, italiano, alemán, francés, checo y estonio) y un módulo inter-lenguas (ILI-*Inter Lingual Index*) que enlaza los *synsets* de cada idioma con los del WordNet inglés. La importancia y las repercusiones que un recurso de este tipo tiene en los trabajos de investigación queda patente en el desarrollo de otros WordNets para otros idiomas. Tal es el caso del ya finalizado proyecto WordNet en catalán (Benítez et al., 1998) o del todavía en progreso proyecto Balkanet (Stamou et al., 2002b), cuyo objetivo es el de desarrollar una base de datos léxica multilingüe formada por WordNets en griego, turco, rumano, búlgaro, checo y serbio.

Al igual que en el caso de WordNet 1.5, EuroWordNet mantiene un conjunto de punteros entre *synsets* para representar relaciones semánticas entre ellos conformando así un recurso semántico en forma de red y de gran potencia. Asimismo, el árbol generado por las relaciones de hiponimia e hiperonimina establece en sus raíces un conjunto de conceptos ontológicos comunes para todos los lenguajes y que clasifican los *synsets* en categorías conceptuales.

En nuestro trabajo, EuroWorNet será usado como un recurso básico en la obtención de información semántica relacionada con un candidato a antecedente anafórico que permitirá establecer criterios adicionales de compatibilidad entre candidato y anáfora. Muchas otras tareas de procesamiento del lenguaje, en particular trabajos orientados a la desambiguación del sentido de las palabras[27], hacen uso de este valioso recurso como un sistema de representación semántica y conceptual del texto.

Por otro lado, aunque el WordNet español se encuentra dentro del proyecto global EuroWordNet, para este trabajo no haremos uso de las características multilingües del recurso, centrándonos únicamente en el WordNet español de forma aislada, a excepción de la ontología definida de forma común en EuroWordNet.

Las relaciones semánticas en WordNet. Si bien EuroWordNet añade un conjunto adicional de relaciones semánticas entre *synsets*, existen un conjunto de ellas que son comunes a todas las versiones de WordNet. Estas relaciones se muestran con algunos ejemplos en el cuadro 4.3.

Relación	*Nombre WN*	*Categorías*	*Ejemplo*	*EWN*
Antonimia	ANTONYM	nombre/nombre verbo/verbo	marido/mujer entrar/salir	SI SI
Hiponimia	HYPONYMY	nombre/nombre	cuchillo/navaja	SI
Meronimia	MERONYMY	nombre/nombre	casa/dormitorio	SI
Implicación	ENTAILMENT	verbo/verbo	comprar/pagar	**SUBEVENT o CAUSE**
Troponimia	TROPONYM	verbo/verbo	caminar/pasear	**HYPONYMY**
Causa	CAUSE	verbo/verbo	matar/morir	SI

Cuadro 4.3. Relaciones semánticas definidas en WordNet

[27] Conocida por el término inglés *Word Sense Disambiguation* y las siglas WSD.

La ontología de EuroWordNet. La ontología de EuroWordNet (Vossen et al., 1998) consta de 63 conceptos principales y distingue tres tipos de entidades:

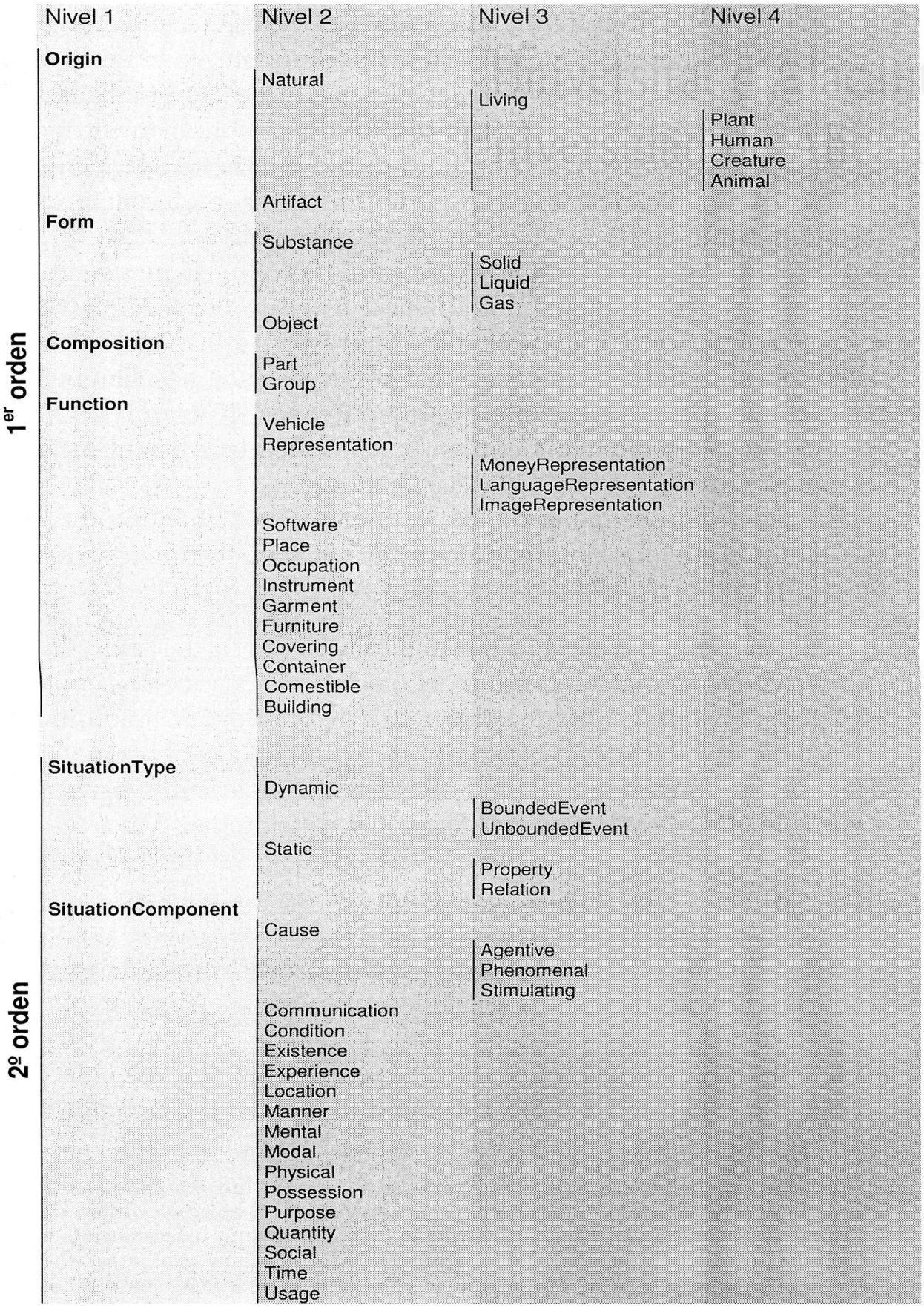
- Entidades de primer orden (*1stOrderEntity*): cualquier entidad concreta perceptible por los sentidos y localizada en cualquier punto del tiempo o del espacio tridimensional, p. ej.: *vehículo*, *animal*, *substancia*, ...
- Entidades de segundo orden (*2ndOrderEntity*): cualquier situación estática (propiedad, relación) o situación dinámica, que no puede ser tocada, escuchada o vista como una cosa física independiente. Puede ser localizada en el tiempo y "ocurre" más que "existe", p. ej.: *ocurrir*, *ser*, *comenzar*, *continuar*, *terminar*, ...
- Entidades de tercer orden (*3rdOrderEntity*): cualquier proposición no observable que existe independientemente del espacio y el tiempo. Puede ser 'falsa' o 'verdadera' más que 'real'. Puede ser afirmada o negada, recordada u olvidada, p. ej. *idea*, *pensamiento*, *información*, *teoría*, *plan*, ...

Estos conceptos ontológicos, asociados a cada *synset* de EuroWordNet, proporcionan propiedades semánticas que pueden ser usadas, tal y como veremos en las siguientes secciones, como fuente de conocimiento para aportar nuevos criterios y mejorar los resultados de la resolución de la anáfora. El cuadro 4.4 muestra los distintos niveles de la esta ontología de conceptos.

4.3.5 Reglas de compatibilidad semántica: los patrones semánticos

Tal y como se ha comentado, el método ERA propuesto en esta sección se caracteriza por el uso de la semántica como fuente de información esencial en la resolución de la anáfora. Esta semántica la proporcionan los conceptos ontológicos asociados a los candidatos a antecedente de un pronombre junto con el verbo de la anáfora[28].

[28] En este punto del trabajo es necesario señalar que la información semántica obtenida a partir de los conceptos ontológicos extraídos de EuroWordNet y la



Cuadro 4.4. Ontología principal definida en EuroWordNet

La figura 4.5 muestra cómo WordNet sirve de entrada para diferentes módulos del método ERA. Básicamente, la combinación de la ontología de EuroWordNet, el sentido de las palabras y la información referente al papel sintáctico de los constituyentes oracionales da como resultado un conjunto de patrones de compatibilidad semántica que servirán como factor de preferencia en la fase de resolución de la anáfora.

Uno de los módulos clave en este método es el generador semántico, cuyo objetivo fundamental es el de proporcionar una representación semántica del texto a través de la generación de colecciones de datos semánticos así como de patrones semánticos u ontológicos. Las colecciones y los patrones de compatibilidad conforman la base de conocimiento semántico que usa el método ERA en la fase de resolución de la anáfora.

El generador semántico, tal y como muestra la figura 4.6, está compuesto por dos módulos que realizan la función de adquisición de patrones en dos etapas:

- La extracción de colecciones semánticas: a partir del texto de entrada con el formato requerido, el módulo de extracción semántica construye un grupo de colecciones de ontologías, sinónimos y frecuencias asociadas a las palabras contenidas en el texto y consultadas en WordNet. Este proceso, completamente automático, consultará cada una de las palabras (nombres y verbo) en WordNet y extraerá sus elementos ontológicos correspondientes, realizando así mismo un conteo de apariciones en el texto para computar su frecuencia. Estas colecciones, por un lado, serán la base de la generalización de patrones y, por otro, serán consultadas en diferentes fases de aplicación de restricciones y preferencias.
- La generación de patrones de compatibilidad: con las colecciones previamente extraídas, este módulo se encarga de construir

sintáctica proporcionada por los papeles de los elementos oracionales van estrechamente unidas en la propuesta. No obstante, las estrategias que usan como base la combinación de ambas se han agrupado bajo el epígrafe común de información semántica, por ser ésta la fuente de conocimiento más relevante en el marco de esta aproximación. Delimitando así la información semántica, se hace una distinción entre ésta y la denominada sintáctico-semántica, que combina la sintaxis oracional y la semántica de rasgos.

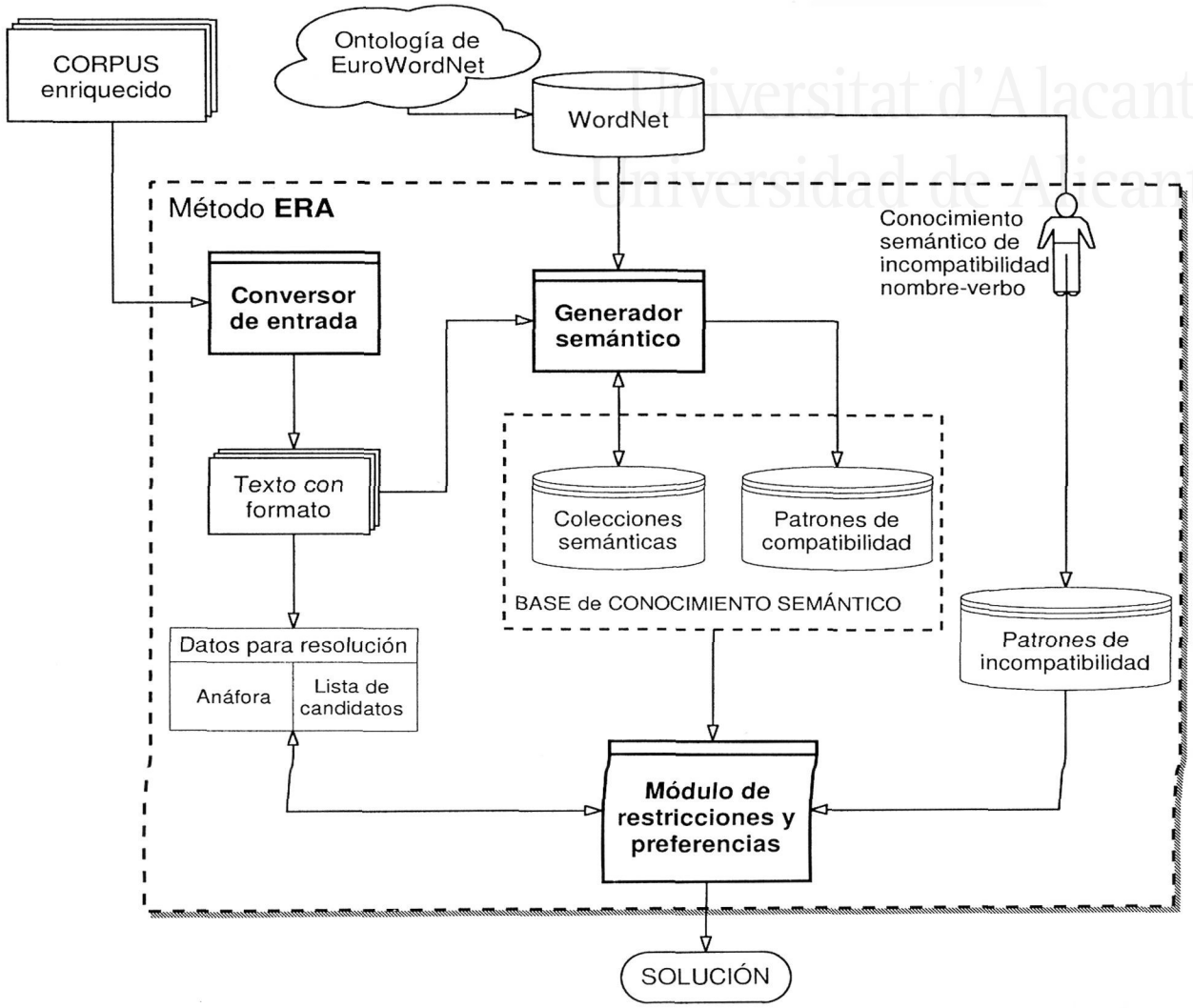


Figura 4.5. Detalle de los módulos integrantes del método ERA

automáticamente un conjunto de patrones semánticos nombre-verbo. Para ello, tomará los conceptos ontológicos asociados a cada nombre y los combinará con el verbo al que acompañan. Calculará su grado de compatibilidad en función del nivel de cada uno de los conceptos ontológicos y, finalmente, lo almacenará en el conjunto correspondiente (sujeto-verbo, verbo-objeto directo o verbo-objeto indirecto). Estos patrones se usarán en la fase de resolución como una fuente adicional de conocimiento que aportará criterios de preferencia de selección de candidatos.

Adquisición de patrones de compatibilidad. Cada patrón extraído del corpus se incorpora a un conjunto de patrones de compatibilidad semántica que sirve como base de conocimiento

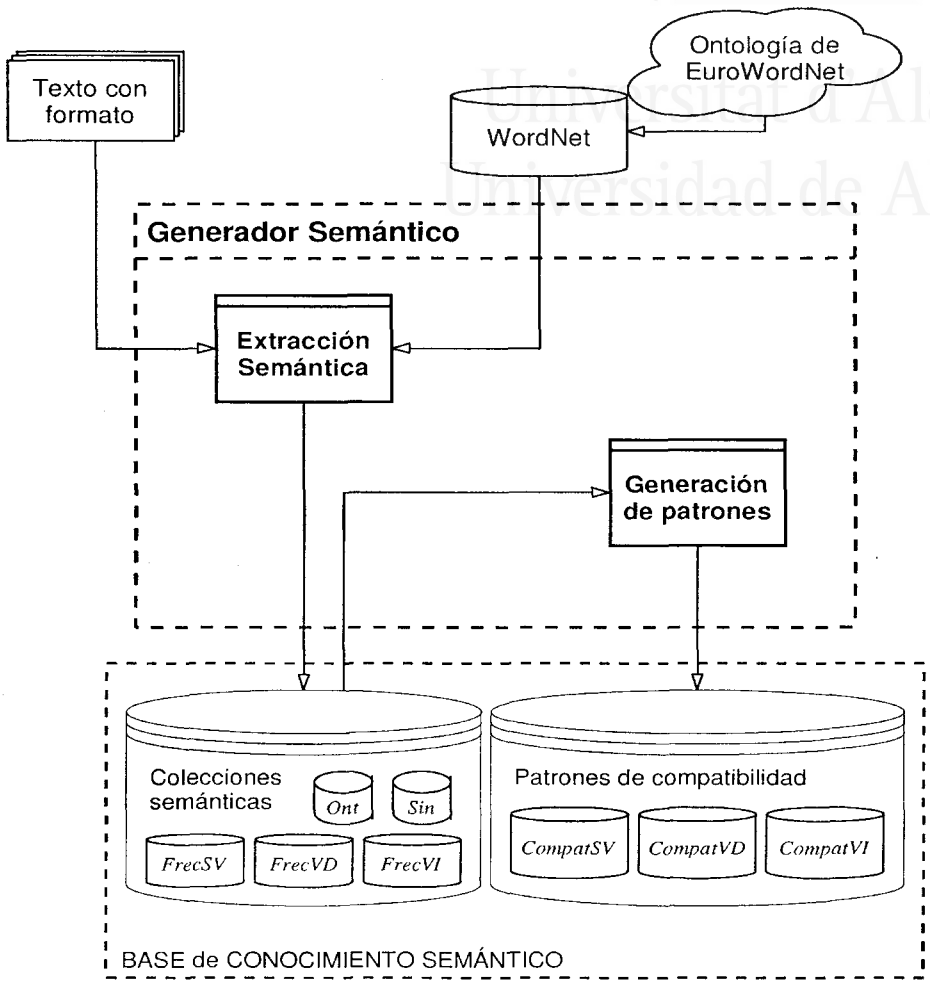


Figura 4.6. Generación de la base de conocimiento semántico para la adquisición de patrones

en la fase de resolución de la anáfora. Cada patron está formado por un concepto ontológico asociado a un nombre con función de sujeto, objeto directo u objeto indirecto y el verbo al que acompañan.

El módulo de extracción semántica construye las siguientes *colecciones semánticas*:

- *Ont*: colección de conjuntos ontológicos asociados a los términos nominales, denotando cada conjunto ontológico como $Ont_{(n\#s)}$ siendo n el nombre y s su sentido. Por ejemplo, para el nombre *mono* en su primer sentido de WordNet, el conjunto de conceptos ontológicos sería:

$$Ont_{(mono\#1)} = [\text{Animal, Form, Living, Natural, Object, Origin}]$$

- *Sin*: colección de conjuntos de sinónimos asociados a los términos nominales y verbales, denotando cada conjunto de sinónimos (integrantes de su mismo *synset* en WordNet) como $Sin_{(p\#s)}$ siendo p la palabra (nombre o verbo) y s su sentido. Así, para el nombre *jarrón* en su primer sentido de WordNet y para el verbo *lanzar* en su sentido décimo el conjunto de sinónimos sería:

$$Sin_{(jarrón\#1)} = [\text{jarrón\#1, florero\#2, vaso\#2, búcaro\#1}]$$
$$Sin_{(lanzar\#10)} = [\text{lanzar\#10, tirar\#17, arrojar\#11}]$$

  Dado que la lista de sinónimos representa al *synset* de WordNet, el propio término está incluido también en dicha lista.

- $FrecSV$: colección de frecuencias de aparición de pares sujeto-verbo, denotando la frecuencia de aparición de un par sujeto-verbo concreto como $FrecSV_{(n\#sentn,v\#sentv)}$ donde n y $sentn$ son el nombre y su sentido y v y $sentv$ son el verbo y su sentido. Cada nombre $n\#s$ procesado genera un par para cada uno de los nombres contenidos en el conjunto de sinónimos $Sin_{(n\#s)}$.

- $FrecVD$: colección de frecuencias de aparición de pares verbo-OD, denotando la frecuencia de aparición de un par verbo-OD concreto como $FrecVD_{(n\#sentn,v\#sentv)}$ donde n y $sentn$ son el nombre y su sentido y v y $sentv$ son el verbo y su sentido. Cada nombre $n\#s$ procesado genera un par para cada uno de los nombres contenidos en el conjunto de sinónimos $Sin_{(n\#s)}$.

- $FrecVI$: colección de frecuencias de aparición de pares verbo-OI, denotando la frecuencia de aparición de un par verbo-OI concreto como $FrecVI_{(n\#sentn,v\#sentv)}$ donde n y $sentn$ son el nombre y su sentido y v y $sentv$ son el verbo y su sentido. Cada nombre $n\#s$ procesado genera un par para cada uno de los nombres contenidos en el conjunto de sinónimos $SIN_{(n\#s)}$.

Con estas colecciones semánticas extraídas a partir del corpus y de WordNet, el módulo de generación de patrones construye la base de conocimiento formada por los siguientes *patrones de compatibilidad*:

- Conjunto de relaciones de compatibilidad sujeto-verbo, compuesto por patrones formados por cada uno de los conceptos

ontológicos asociados a un nombre con función de sujeto y cada uno de los sinónimos del verbo con el que aparecen $Sin_{(v\#sentv)}$. A este conjunto le llamaremos $CompatSV$. A la compatibilidad entre un concepto ontológico con función de sujeto y un verbo la llamaremos $CompatSV_{(c,v\#sentv)}$ donde c es el concepto ontológico, v es el verbo y $sentv$ es su sentido.

- Conjunto de relaciones de compatibilidad verbo-OD, compuesto por patrones formados por cada uno de los conceptos ontológicos asociados a un nombre con función de objeto directo y cada uno de los sinónimos del verbo con el que aparecen $Sin_{(v\#sentv)}$. A este conjunto le llamaremos $CompatVD$. A la compatibilidad entre un concepto ontológico con función de OD y un verbo la llamaremos $CompatVD_{(c,v\#sentv)}$ donde c es el concepto ontológico, v es el verbo y $sentv$ es su sentido.
- Conjunto de relaciones de compatibilidad verbo-OI, compuesto por patrones formados por cada uno de los conceptos ontológicos asociados a un nombre con función de objeto indirecto y cada uno de los sinónimos del verbo con el que aparecen $Sin_{(v\#sentv)}$. A este conjunto le llamaremos $CompatVI$. A la compatibilidad entre un concepto ontológico con función de OI y un verbo la llamaremos $CompatVI_{(c,v\#sentv)}$ donde c es el concepto ontológico, v es el verbo y $sentv$ es su sentido.

La figura 4.7 muestra un ejemplo sencillo de cómo actúa el generador semántico sobre un conjunto de nombres pertenecientes a un corpus de entrada.

El grado de compatibilidad asociado a cada uno de los patrones contenidos en estos conjuntos ha de tener una relación directa con el tipo de información que proporcionan. Así, en nuestra propuesta, se considera que cuanto más general sea el concepto ontológico, menos información semántica aporta y por tanto resulta menos relevante. Así, para asignar esta compatibilidad, se toma como referencia el nivel del concepto ontológico que forma el patrón (ver cuadro 4.4 en la pág. 123), dotando de mayor relevancia a aquellos patrones formados por conceptos ontológicos más concretos. Por ejemplo, un patrón formado por el concepto ontológico 'Living' tendrá un grado de compatibilidad 3 (correspondiente a su

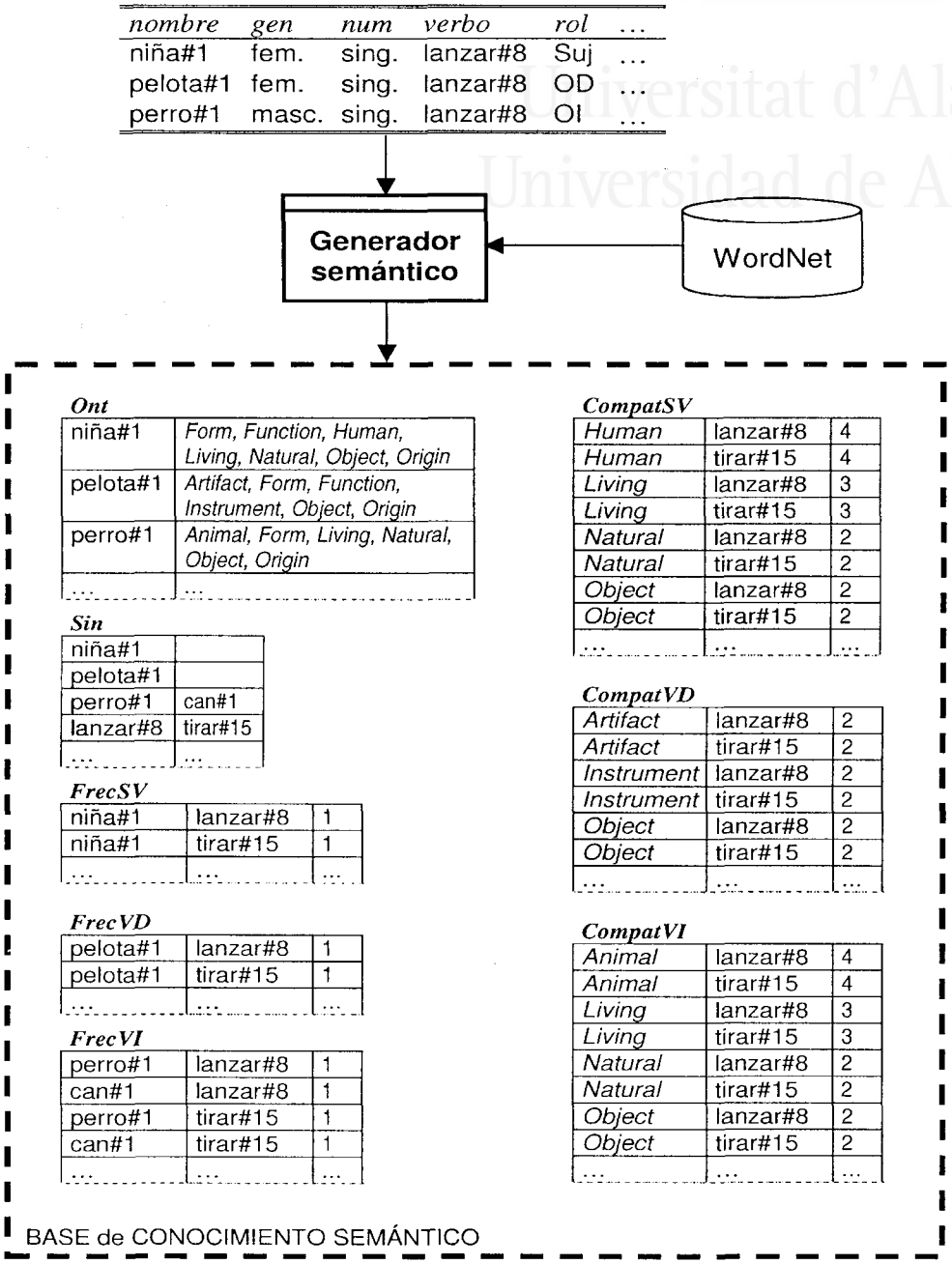


Figura 4.7. Ejemplo de adquisición de patrones

nivel), mientras que un patrón formado por el concepto ontológico 'Human' tendrá un grado de compatibilidad igual a 4. Además, la aparición repetida de un patrón determinado incrementa su grado de compatibilidad en el conjunto de relaciones correspondiente, con lo que, la compatibilidad y por tanto, la relevancia de

un patrón, será mayor cuanto más representado esté el patrón en el texto.

Por otra parte, el uso de la relación de sinonimia para la construcción de los patrones de compatibilidad establece un método cooperativo en el que la idea de palabra deja paso a la de concepto, incrementando así el alcance del método de resolución.

Los patrones semánticos en la resolución de la anáfora. En la fase de resolución de la anáfora, el generador semántico extrae todos los pares formados por los conceptos ontológicos de los candidatos y su verbo correspondiente y los incorpora a la base de conocimiento semántico.

El módulo de restricciones y preferencias combina los conceptos ontológicos de los candidatos con el verbo de la anáfora en función del papel sintáctico que ésta realice. Esta combinación da como resultado el conjunto de patrones semánticos asociados a la anáfora que tendrá que ser contrastado con los patrones de compatibilidad aprendidos del corpus con el fin de establecer un criterio adicional de preferencia sobre la lista de candidatos.

Para establecer este criterio de preferencia, se proponen un conjunto de reglas que indican si un nombre es compatible o no con un verbo, denominadas *reglas de compatibilidad semántica*.

En primer lugar, se definen las reglas de compatibilidad entre un verbo y un nombre:

Regla 1 *Un verbo* v *con sentido* $sentv$ *es compatible con un nombre* n *con sentido* $sentn$ *como sujeto de* $v\#sentv \iff$
$\exists c \in Ont(n\#sentn) \mid CompatSV_{(c,v\#sentv)} > 0$

Regla 2 *Un verbo* v *con sentido* $sentv$ *es compatible con un nombre* n *con sentido* $sentn$ *como objeto directo de* $v\#sentv \iff$
$\exists c \in Ont(n\#sentn) \mid CompatVD_{(c,v\#sentv)} > 0$

Regla 3 *Un verbo* v *con sentido* $sentv$ *es compatible con un nombre* n *con sentido* $sentn$ *como objeto indirecto de* $v\#sentv \iff$
$\exists c \in Ont(n\#sentn) \mid CompatVI_{(c,v\#sentv)} > 0$

Por otro lado se definen las reglas de preferencia semántica de un candidato frente a otro:

Regla 4 *Un candidato anafórico con núcleo* $n1\#sentn1$ *es preferido semánticamente frente a otro candidato con núcleo* $n2\#sentn2$ *como sujeto de un verbo* $v\#sentv \iff$

$\forall c_i \in Ont(n1\#sentn1)$,
$\forall d_i \in Ont(n2\#sentn2)$,
$\sum(CompatSV_{(c_i)}) > \sum(CompatSV_{(d_i)})$

Regla 5 *Un candidato anafórico con núcleo* $n1\#sentn1$ *es preferido semánticamente frente a otro candidato con núcleo* $n2\#sentn2$ *como objeto directo de un verbo* $v\#sentv \iff$

$\forall c_i \in Ont(n1\#sentn1)$,
$\forall d_i \in Ont(n2\#sentn2)$,
$\sum(CompatVD_{(c_i)}) > \sum(CompatSV_{(d_i)})$

Regla 6 *Un candidato anafórico con núcleo* $n1\#sentn1$ *es preferido semánticamente frente a otro candidato con núcleo* $n2\#sentn2$ *como objeto indirecto de un verbo* $v\#sentv \iff$

$\forall c_i \in Ont(n1\#sentn1)$,
$\forall d_i \in Ont(n2\#sentn2)$,
$\sum(CompatVI_{(c_i)}) > \sum(CompatSV_{(d_i)})$

El hecho de que estas reglas sean aplicadas como preferencia es porque el no cumplimiento de estas reglas de compatibilidad con un verbo en cualquiera de las posibles funciones sintácticas no implica una incompatibilidad sino tan sólo la ausencia del patrón tras la adquisición de patrones de compatibilidad.

Veamos un ejemplo de aplicación. Supongamos que en la fase de resolución de la anáfora el sistema ha de resolver la siguiente referencia pronominal:

(96) El mono subió al árbol a coger un *plátano*$_i$ cuando el sol salía.
$\emptyset_i$ maduraba lentamente.

El pronombre omitido, de tercera persona del singular puede correferir con cualquiera de los SN anteriores, con lo que la lista formada por los núcleos de los SN candidatos sería L=[mono#1, árbol#2, plátano#1, sol#2]. El pronombre omitido tiene función de sujeto del verbo *madurar#1*. A la hora de seleccionar el

candidato más compatible con el verbo, se realizará la búsqueda en el conjunto de relaciones de compatibilidad sujeto-verbo ($CompatSV$).

Los conjuntos de elementos ontológicos asociados a cada nombre son:

$Ont_{(mono\#1)}$ = [Animal, Living, Natural, Object]
$Ont_{(árbol\#2)}$ = [Group, Living, Natural, Object, Plant]
$Ont_{(plátano\#1)}$ = [Comestible, Group, Living, Natural, Object, Plant, Substance]
$Ont_{(sol\#2)}$ = [Natural, Object]

Supongamos que los patrones relacionados con este verbo y extraídos en la etapa de adquisición son:

Natural	madurar#1	24	(12 apariciones)
Living	madurar#1	36	(12 apariciones)
Plant	madurar#1	16	(4 apariciones)
Human	madurar#1	12	(3 apariciones)
Creature	madurar#1	4	(1 aparición)
Animal	madurar#1	12	(3 apariciones)
Substance	madurar#1	6	(3 apariciones)
Object	madurar#1	6	(3 apariciones)
Comestible	madurar#1	8	(4 apariciones)

Para determinar la mayor compatibilidad hay que aplicar la Regla 4:

$$\forall c_i \in Ont_{(mono\#1)},$$
$$\sum(CompatSV_{(c_i)}) = CompatSV_{(Animal,madurar\#1)} + CompatSV_{(Living,madurar\#1)} + CompatSV_{(Natural,madurar\#1)} + CompatSV_{(Object,madurar\#1)} = 12 + 36 + 24 + 6 = 78$$

$$\forall c_i \in Ont_{(árbol\#2)},$$
$$\sum(CompatSV_{(c_i)}) = CompatSV_{(Group,madurar\#1)} + CompatSV_{(Living,madurar\#1)} + CompatSV_{(Natural,madurar\#1)} + CompatSV_{(Object,madurar\#1)} + CompatSV_{(Plant,madurar\#1)} = 0 + 36 + 24 + 6 + 16 = 82$$

$$\forall c_i \in Ont_{(plátano\#1)},$$
$$\sum(CompatSV_{(c_i)}) = CompatSV_{(Comestible,madurar\#1)} + CompatSV_{(Group,madurar\#1)} + CompatSV_{(Living,madurar\#1)} + CompatSV_{(Natural,madurar\#1)} + CompatSV_{(Object,madurar\#1)} + CompatSV_{(Plant,madurar\#1)} + CompatSV_{(Substance,madurar\#1)} = 8 + 0 + 36 + 24 + 6 + 16 + 6 = \mathbf{96}$$

$$\forall c_i \in Ont_{(sol\#2)},$$
$$\sum(CompatSV_{(c_i)}) = CompatSV_{(Natural,madurar\#1)} + CompatSV_{(Object,madurar\#1)} = 24 + 6 = 30$$

Según este proceso, el candidato preferido es el sintagma nominal cuyo núcleo es *plátano*. Algo que también se observa en este ejemplo es la forma en que la aplicación de las reglas de compatibilidad establecen la lejanía entre el SN *sol* y el verbo *madurar#1* y, por el contrario, la proximidad entre los SN *árbol* y *mono* y el mismo verbo, algo que corrobora su grado real de compatibilidad. La aplicación de esta preferencia es uno de los elementos característicos del método enriquecido de resolución de la anáfora.

4.3.6 Reglas de incompatibilidad semántica

Además de la información semántica procedente de la adquisición de patrones y usada como criterio preferencial para seleccionar el candidato más compatible, el método ERA incorpora un conjunto de reglas basadas en conocimiento de incompatibilidad semántica. El objetivo de estas reglas es el de establecer criterios de eliminación de candidatos incompatibles con la anáfora.

Estas reglas se aplican a partir de un conjunto de patrones de incompatibilidad que siguen una estructura similar a la de los patrones usados para la compatibilidad semántica. La supervisión de estos patrones de incompatibilidad garantiza que su aplicación elimina únicamente aquellos candidatos que son realmente incompatibles con la anáfora.

Estas reglas están inspiradas en las restricciones de selección definidas por otros autores a partir de la subcategorización del verbo (Rich y Luperfoy, 1998; Hobbs, 1986; Carter, 1987a; Rich y Luperfoy, 1998; Carbonell y Brown, 1988). El enriquecimiento del corpus permite la definición de estas reglas sobre conceptos, en lugar de sobre palabras, evitando posibles problemas con las palabras polisémicas.

Se han definido para el método dos tipos de reglas sobre dos tipos de patrones distintos:

- *Reglas "no"*: este tipo de regla define lo que podríamos llamar incompatibilidad obligatoria de un verbo con un concepto ontológico determinado. Si un concepto ontológico de un nombre

está asociado a un verbo determinado a través de un patrón "no", entonces el nombre es incompatible con ese verbo. Formalmente, se podría enunciar de la siguiente manera:

Regla 7 *La regla* $NO(v\#sentv, c, r)$ *define la incompatibilidad del verbo* $v\#sentv$ *con cualquier nombre* $n\#sentn$ *que contenga a* c *en su lista de conceptos ontológicos* $Ont_{(n\#sentn)}$ *siendo* r *la función sintáctica que les relaciona.*

Por ejemplo, la regla de incompatibilidad aplicada a partir del patrón $NO(\mathsf{vivir\#1}, \mathsf{Artifact}, \mathsf{S})$ permitiría eliminar a todos aquellos candidatos que tengan *Artifact* en su lista de conceptos ontológicos como posibles antecedentes de una anáfora que es sujeto del verbo *vivir* con su primer sentido de WordNet.

- *Reglas "debe"*: este tipo de regla define lo que podríamos llamar compatibilidad obligatoria de un verbo con un concepto ontológico determinado. Si un nombre no contiene el concepto ontológico asociado a un verbo determinado a través de un patrón "debe", entonces el nombre es incompatible con ese verbo. Formalmente, se podría enunciar de la siguiente manera:

  Regla 8 *La regla* $DEBE(v\#sentv, c, r)$ *define la incompatibilidad del verbo* $v\#sentv$ *con todos los nombres* $n\#sentn$ *que no contengan a* c *en su lista de conceptos ontológicos* $Ont_{(n\#sentn)}$ *siendo* r *la función sintáctica que les relaciona.*

  Por ejemplo, la regla de incompatibilidad aplicada a partir del patrón $DEBE(\mathsf{Comestible}, \mathsf{comer\#2}, \mathsf{D})$ permitiría eliminar a todos aquellos candidatos que no tengan '**Comestible**' en su lista de conceptos ontológicos como posibles antecedentes de una anáfora que es objeto directo del verbo *comer* con su segundo sentido de WordNet.

La obtención de los patrones de incompatibilidad se puede realizar con diferentes técnicas:

- A partir de un conjunto de patrones definidos manualmente con el uso de conceptos ontológicos extraídos de WordNet (el

verbo *comer* no puede tener como objeto algo 'no comestible' o como sujeto algo 'no animado'). Este tipo de definición de incompatibilidad puede ser especialmente útil en la aplicación del método de resolución de la anáfora a dominios restringidos (Moreno et al., 1991).

- A partir de un proceso automático de adquisición de patrones. Este proceso debe garantizar que el conjunto de patrones extraído de la adquisición reúne las condiciones mínimas para ser representativo y, por tanto, los patrones no incluidos en la lista de los generados son realmente patrones de incompatibilidad. Esta segunda opción es más general, pero a la vez puede resultar algo más arriesgada, ya que puede que un antecedente sea eliminado por su incompatibilidad tan sólo porque el patrón que genera no ha aparecido previamente.
- A partir de un proceso mixto, en el que se generan los patrones y, durante el entrenamiento del módulo de resolución, se supervisan los supuestamente incompatibles. Este proceso permitirá que el sistema "aprenda" patrones incompatibles durante su propia evolución.

En nuestra propuesta, se ha optado por el uso de la primera y la tercera técnica. Por un lado, se han propuesto patrones de incompatibilidad a partir del "sentido comun", en los que se establecen incompatibilidades que resultan evidentes como las existentes entre sujetos de tipo no animado y verbos relacionados con procesos mentales (*pensar*, *deducir*, *reflexionar*, ...). Para completar este conjunto de reglas con algunas que, aunque evidentes, podrían no haberse tenido en cuenta, se han estudiado los resultados de la adquisición de patrones desde el corpus para determinar reglas adicionales que quedaban patentes ante su elevado índice de compatibilidad (tal es el caso del patrón *comer-comestible*). El hecho de haber desestimado la segunda opción es debido a la dificultad de encontrar en un corpus información adecuada para garantizar la efectividad de patrones generados de forma automática.

Siguiendo con el esquema de aplicación del método de conocimiento limitado, para el método ERA se ha usado un sistema de resolución de la anáfora basado en restricciones y preferencias. Así,

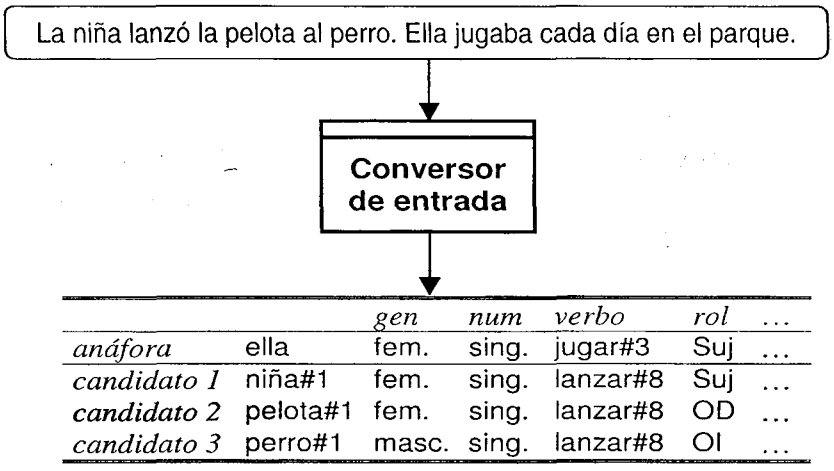
las primeras eliminarán candidatos claramente incompatibles con la anáfora mientras que las segundas establecerán criterios conjuntos que permitirán seleccionar un único antecedente del conjunto de candidatos compatibles.

Las reglas de incompatibilidad semántica, aplicadas como restricción y las reglas de compatibilidad, usadas como preferencia, configuran la información semántica incorporada en el proceso de resolución de la anáfora.

4.3.7 Módulo conversor de entrada

Como se puede ver en la figura 4.5, un conversor de entrada se encarga de facilitarle al módulo de restricciones y preferencias los datos necesarios para la resolución anafórica. Este módulo se encarga de transformar el corpus anotado y enriquecido en una estructura fija compuesta por los datos del pronombre anafórico y de sus candidatos a antecedente.

La figura 4.8 muestra un ejemplo de generación de esta estructura a partir de una oración del corpus supuestamente analizada[29].



		gen	*num*	*verbo*	*rol*	...
anáfora	ella	fem.	sing.	jugar#3	Suj	...
candidato 1	niña#1	fem.	sing.	lanzar#8	Suj	...
candidato 2	pelota#1	fem.	sing.	lanzar#8	OD	...
candidato 3	perro#1	masc.	sing.	lanzar#8	OI	...

Figura 4.8. Ejemplo de funcionamiento del módulo conversor de entrada

[29] Para facilitar su comprensión, en el gráfico de la figura 4.8 no se muestra el análisis y enriquecimiento realizado sobre la oración de entrada al módulo, aunque dicho análisis y enriquecimiento se supone realizado para la aplicación del módulo conversor de entrada.

La estructura generada por el conversor de entrada será con la que trabajen tanto el generador de patrones visto anteriormente como el módulo de aplicación de restricciones y preferencias.

A continuación veremos en qué consisten las restricciones y preferencias propuestas en el método ERA, cuyo esquema se puede ver en la figura 4.9.

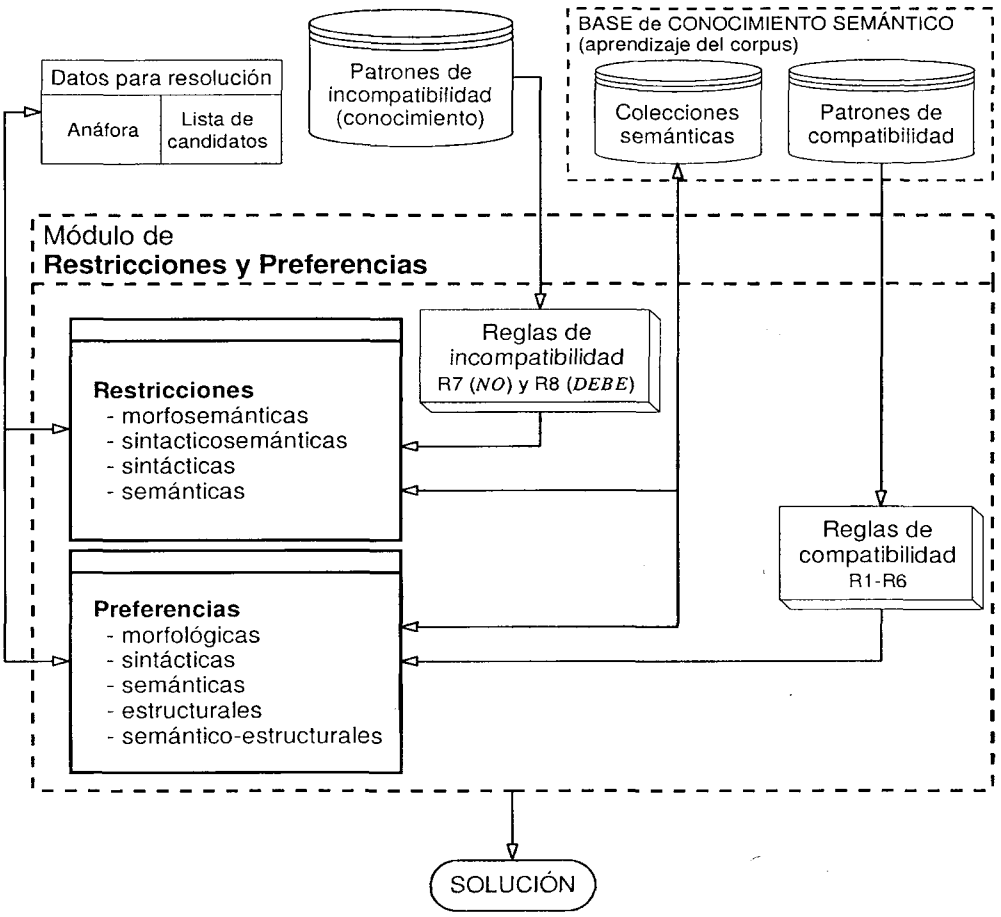


Figura 4.9. Esquema del módulo de Restricciones y Preferencias

4.3.8 Módulo de aplicación de restricciones

Para la eliminación de candidatos incompatibles, se proponen en esta ocasión un conjunto de restricciones de carácter morfológico, sintáctico y semántico. Estas restricciones, tal y como ocurre en el método anterior, definen las condiciones que hacen que un pronombre y un SN antecedente no puedan correferir.

Condición morfosemántica de no correferencia pronombre-SN . Conforme al tratamiento tradicional en los métodos de resolución de la anáfora, un SN y un pronombre no correferirán si no concuerdan en género, número y persona.

No obstante, existen algunos matices que mezclan características morfológicas y semánticas de antecedente y pronombre que pueden enriquecer considerablemente la eliminación o no de candidatos:

1. Un SN y un pronombre no serán correferentes si no concuerdan en género, número y persona, excepto si el pronombre es plural y el SN tiene el rasgo de 'grupo':

   (97) El *cuerpo*$_i$ de policía vela por su seguridad. *Ellos*$_i$ están siempre alerta.

Esta condición morfosemántica, enunciada para el español, puede enriquecerse considerablemente cuando se aplica a otros idiomas en los que los pronombres aportan mayor información morfológica. En particular, para el caso del inglés, el pronombre personal neutro *it* no correferirá con un SN de tipo persona, mientras que los pronombres masculino y femenino *he* y *she* nunca correferirán con un SN que no sea de tipo persona. En trabajos anteriores (Peral et al., 1999; Saiz-Noeda et al., 2000b) se puede ver la aplicación de estas restricciones al proceso de resolución de la anáfora en inglés y comprobar su interesante repercusión en los resultados de resolución anafórica (Peral, 2001).

Condiciones sintáctico-semánticas de no correferencia pronombre-SN . Por otro lado, la combinación de consideraciones asociadas a rasgos semánticos del sustantivo con criterios sintácticos asociados al pronombre nos permite definir dos condiciones que hemos denominado condiciones sintáctico-semánticas de no correferencia pronombre-SN:

1. Un SN con rasgo de 'no animado' no puede correferir con un pronombre personal de sujeto no neutro.

(98) El $coche_j$ de $Mario_i$ está averiado. $Él_i$ está muy preocupado por la reparación.

Esta condición permite eliminar los SN que no cuentan el rasgo de 'animado' cuando el pronombre sea personal de sujeto.

2. Los pronombres personales *le* y *les* con función de objeto directo sólo pueden correferir con un SN masculino con rasgo de 'humano'.

(99) $Luis_i$ ganó el $premio_j$ al mejor $cortometraje_k$. Le_i vi muy contento.

La forma *le* de objeto directo sólo se usa para referir a personas de género masculino[30].

Condiciones sintácticas de no correferencia pronombre-SN. En este punto se definirán las condiciones sintácticas a aplicar para rechazar candidatos que no correfieran con el pronombre. Estas condiciones, a diferencia del método anterior, se enunciarán desde el análisis enriquecido expuesto anteriormente, contando de esta manera con información no sólo de los constituyentes oracionales sino también de los papeles sintácticos que estos constituyentes tienen en la oración[31].

Esta nueva información va a permitir además encontrar puntos en común entre las condiciones aplicadas para distintos tipos de pronombre y, en la mayoría de los casos, definir conjuntos de condiciones más simplificados con respecto a métodos basados en conocimiento limitado.

Veamos las condiciones sintácticas de no correferencia según sea el pronombre anafórico:

1. Un SN no correfiere con un **pronombre reflexivo** si el SN no es el sujeto del mismo verbo al que acompaña el pronombre. El pronombre reflexivo hace que la acción del verbo recaiga sobre el sujeto de dicho verbo, con lo que siempre correferirá con

[30] El incumplimiento de esta norma da lugar al conocido fenómeno del leísmo.

[31] El apartado 4.3.3 (pág. 116) detalla el tipo de enriquecimiento aplicado sobre el corpus original y el 5.3.1 (pág. 160) muestra algunos aspectos relevantes del corpus usado para la evaluación.

él. Esta condición reúne en una sola todas las condiciones expuestas en el método anterior:

- La condición 1*a* (pág. 102) del método anterior rechaza todo SN que no tiene un papel sintáctico principal (sujeto, complemento directo, complemento indirecto, ...) por estar incluido en otro SN.
- La condición 1*b* (pág. 102) alude a la característica ya mencionada, propia de los pronombres reflexivos, de que han de encontrar su antecedente en la misma cláusula.
- La condición 1*c* (pág. 103) en realidad plantea que si existe un SN antes del verbo, éste debe ser el sujeto, con lo que los que aparecen después del verbo pueden ser rechazados. Esto, tal y como se ha comentado y visto en el ejemplo (83), puede no funcionar, con lo que la condición de no correferencia enunciada en este método resulta ser mucho más precisa.

La condición de no correfencia que se acaba de enunciar tiene un carácter tan restrictivo que no es necesario plantear en este método preferencias para este tipo de pronombres ya que la aplicación de la restricción proporciona el antecedente correcto.

2. Un SN no correfiere con un **pronombre personal o demostrativo** si:

   a) El SN y el pronombre modifican al mismo verbo y desempeñan papeles sintácticos diferentes.
   De nuevo, esta condición de no correferencia resume algunas de las expuestas en el método anterior para el mismo tipo de pronombres:

   - Según la condición 2*a* (pág. 103), el SN y el pronombre no pueden correferir si estando en la misma cláusula, el SN está dentro de un SP. Si es así, el SN desempeñará una función sintáctica distinta (complemento directo o indirecto, circunstancial, ...), por lo que no podrá correferir con el pronombre si éste tiene otra función sintáctica[32].

[32] Si bien el fenómeno del doblado de clíticos plantea una excepción (ver nota 16, pág 103), esta excepción en este caso se resuelve con facilidad ya que ambos elementos comparten el papel sintáctico con respecto al mismo verbo.

- Según la condición $2b$ (pág. 104) el SN y el pronombre no pueden correferir si éste aparece antes del verbo, intentando predecir que el pronombre es el sujeto de la oración y, por tanto, no podrá correferir con otro complemento del mismo verbo.
- Esta misma idea se enuncia en la condición $2c$ (pág. 105) pero esta vez en el caso de que el pronombre esté después del verbo, es decir, suponiendo que es un complemento del mismo, y por tanto no podrá coreferir con cualquier SN de la misma oración que tenga un papel sintáctico con respecto al mismo verbo (para lo que es necesario que no esté incluído en otro SN).
- Por otro lado, la condición $2d$ (pág. 105) trata el caso en el que el pronombre forme parte de un SP que modifique directamente al verbo (complemento) y el SN también modifique directamente al verbo. Una vez más, se establece una no-correferencia entre dos complementos diferentes del mismo verbo.
- Siguiendo con el mismo planteamiento, y según la condición $2e$ (pág. 105), si el pronombre está contenido en el SN, no tendrá ningún papel sintáctico con respecto al verbo y por tanto la condición enunciada antes también recoge este caso.
- Por último, y en lo referente a la condición $2g$ (pág. 106), si una oración de relativo es introducida por un SN, éste tendrá una función con respecto al verbo de la cláusula de relativo, con lo que todas las condiciones anteriores pueden ser aplicadas para cualquier pronombre que aparezca dentro de esta cláusula.

b) El SN está coordinado con el pronombre.
Ésta es en realidad una excepción de la condición anterior. Si el SN y el pronombre están coordinados, ambos tendrán el mismo papel sintáctico con respecto al verbo que modifiquen (en el caso de que modifiquen directamente a algún verbo). Sin embargo, y tal y como se comentaba en el método anterior (ver condición $2f$ en página 106), am-

bos representarán elementos disjuntos que nunca podrán correferir.

Al igual que en el método anterior, las condiciones enunciadas para los pronombres personales y demostrativos, son aplicables de forma análoga a los pronombres omitidos.

Condiciones semánticas de no correferencia pronombre-SN. La adquisición de reglas de incompatibilidad, detallada en el apartado 4.3.6 (pág. 133), tiene como objetivo descartar los candidatos que son claramente incompatibles con el verbo de la anáfora, según el papel sintáctico que ésta representa con respecto a aquél. Así, podemos establecer la siguiente condición semántica de no correferencia pronombre-SN:

1. Un SN no correfiere con un pronombre si queda definida su incompatibilidad a través de una regla "no" o una regla "debe" de incompatibilidad semántica con respecto a la anáfora y al papel sintáctico que ésta representa en relación a su verbo.

Veamos un ejemplo de aplicación de restricciones sobre la anáfora de la figura 4.8.

Como puede verse en el ejemplo de la figura 4.10, las restricciones morfológicas descartan el SN cuyo núcleo es *perro#1* por la no concordancia en género con el pronombre anafórico. Por otro lado, la aplicación de las reglas de incompatibilidad a partir de los patrones aprendidos establecen que sólo aquellos nombres que contengan el rasgo 'Living' podrán ser candidatos de un pronombre sujeto del verbo *jugar#3*. De esta manera, se elimina el candidato de núcleo *pelota#1* por no contener el mencionado rasgo.

En este caso, la aplicación de restricciones daría como resultado el antecedente correcto sin necesidad de aplicar preferencias. No obstante, en la mayoría de los casos el número de candidatos que superan la fase de restricciones es mayor que uno y por tanto se hace necesaria la aplicación de factores de preferencia que permitan la selección de considerado como mejor candidato.

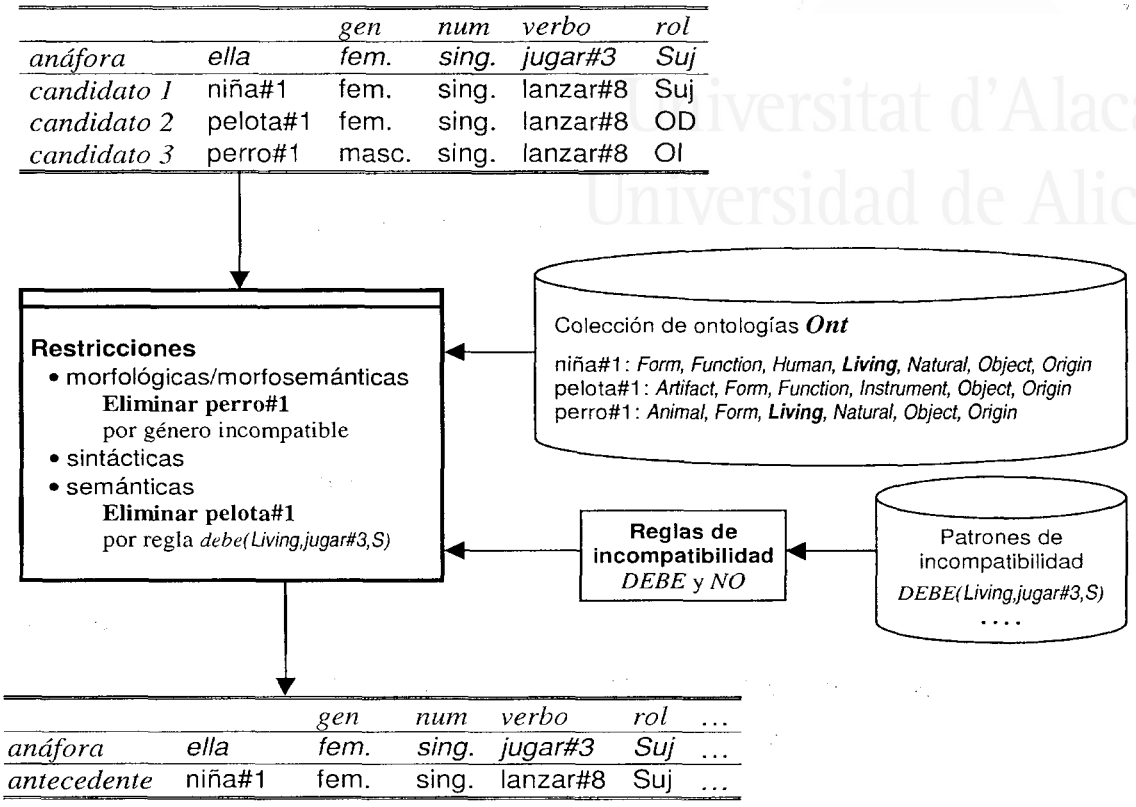


Figura 4.10. Ejemplo de aplicación de restricciones en el método ERA

4.3.9 Módulo de aplicación de preferencias

Las nuevas fuentes de información con las que cuenta el método ERA permitirán la incorporación de nuevas condiciones de preferencia a la hora de decidir el antecedente correcto de la anáfora.

El conjunto de preferencias ha sido seleccionado a partir del estudio del corpus. Así, las preferencias propuestas serán clasificadas en función de la fuente de información usada para su aplicación.

1. Preferencias de carácter morfológico
   a) Se prefieren los SN que concuerdan en número con la anáfora.

   Por la aplicación de las condiciones morfosemánticas de no correferencia es posible que queden antecedentes con información morfológica de número distinta a la de la anáfora. En las preferencias finales, se preferirán los que concuerden en número con ésta.

2. Preferencias de carácter sintáctico
   a) Se prefieren los SN candidatos que realizan la misma función sintáctica que la anáfora con respecto al verbo.
   Esta preferencia es similar a la utilizada en el método de conocimiento limitado, con la particularidad de que no tiene en cuenta la posición del SN y del pronombre (antes o después del verbo) sino su papel sintáctico (sujeto, complemento directo o complemento indirecto).
   b) Se prefieren los SN candidatos que no están incluidos en otro SN.
   Por tratarse de un fenómeno discursivo, es más común que la anáfora aluda a un SN principal más que a uno subordinado.

3. Preferencias de carácter semántico
   a) Se prefieren los SN que no son de tiempo, dirección, cantidad o tipo abstracto (*las ocho menos cuarto*, *calle primavera*, *cuarenta*, *una cosa*, ...).
   Si bien esta preferencia es idéntica en su enunciado a la propuesta en el método anterior, la forma de obtener el rasgo semántico asociado al SN es claramente distinta. Mientras que en el primer método se utilizan un conjunto de reglas para conjeturar dicho rasgo, en el segundo se obtiene a partir de la información proporcionada por el recurso léxico que contiene la clasificación ontológica requerida.
   b) Se prefieren los SN que son semánticamente más compatibles con el verbo del pronombre.
   Esta preferencia aplica un criterio semántico basado en el conjunto de relaciones de compatibilidad generado en la adquisición de patrones[33]. Se preferirá aquel candidato cuya compatibilidad sea la mayor.

4. Preferencias de carácter estructural
   a) Se prefieren los SN candidatos que aparecen en la misma oración frente a los que aparecen en oraciones anteriores,

[33] En el apartado 4.3.5 (pág 122), se detalla el proceso de obtención de estos patrones semánticos.

siendo la preferencia mayor en función de la cercanía entre candidato y anáfora.

b) Se prefieren, en el caso de los pronombre omitidos, los SN candidatos que han sido solución anteriormente de un pronombre omitido.

5. Preferencias de carácter semántico-estructural
   a) Se prefieren los SN candidatos que se han repetido más veces en el texto. Para la valoración de estas repeticiones se tendrá en cuenta en lugar de la palabra, el concepto formado por la palabra y su sentido en el texto, así como las apariciones de sinónimos de dicho concepto.
   b) Se prefieren los SN candidatos que se han repetido más veces en el texto con el mismo verbo de la anáfora. Se valorará positivamente la aparición repetida de un concepto (o cualquiera sus sinónimos) con un verbo (o con cualquiera de sus sinónimos) teniendo en cuenta además el papel sintáctico que tiene con respecto a dicho verbo.

Según estas preferencias, el conjunto propuesto para cada tipo de pronombre es el siguiente:

Preferencias para pronombres personales o demostrativos.

1. SN que no son de tiempo, dirección, cantidad ni tipo abstracto.
2. SN en la misma oración que el pronombre.
3. SN en la oración anterior.
4. SN no incluidos en otro SN (por ejemplo, si aparecen en una cláusula de relativo o una aposición).
5. SN que tienen el mismo papel sintáctico (sujeto o complemento directo) que la anáfora con respecto al verbo.
6. SN que se han repetido más de una vez en el texto.
7. SN que aparecen con el verbo de la anáfora más de una vez

Preferencias para pronombres omitidos.

1. SN que no son de tiempo, dirección, cantidad ni tipo abstracto.
2. SN en la misma oración que el pronombre.

3. SN en la misma oración que el pronombre y que además han sido solución para otro pronombre omitido.
4. SN en la oración anterior.
5. SN con función de sujeto.
6. SN no incluidos en otro SN (por ejemplo, si aparecen en una cláusula de relativo o una aposición).
7. SN que se han repetido más de una vez en el texto.

Preferencias para pronombres reflexivos. Como se ha comentado en la sección anterior dedicada a las condiciones de no correferencia, no es necesario aplicar ninguna clase de preferencia a los candidatos de un pronombre reflexivo, ya que la restricción aplicada sobre este tipo de pronombres proporciona el antecedente correcto de la anáfora.

Preferencias comunes. Si tras la aplicación de las preferencias enunciadas anteriormente asociadas a cada tipo de pronombre, no se ha conseguido obtener el antecedente correcto, es necesaria la aplicación de una serie de preferencias comunes que resuelvan el problema con la determinación de un único candidato como antecedente. Estas preferencias comunes son:

1. SN que concuerda en número con el antecedente.
2. SN más repetido en el texto.
3. SN más cercano al pronombre.

Como puede verse, se ha eliminado la segunda preferencia común propuesta en el método anterior que seleccionaba el candidato en función de su frecuencia de aparición con el verbo de la anáfora. En este nuevo conjunto de preferencias, se establece, a través de la primera de ellas, un criterio basado en la semántica obtenida de los patrones de compatibilidad que aporta, no sólo una co-ocurrencia de términos, sino una compatibilidad de conceptos semánticos u ontológicos asociados a esos términos[34].

Así, la primera de las preferencias establece un criterio basado en un conjunto de patrones sujeto-verbo y verbo-objeto formados por los conceptos ontológicos asociados a los candidatos y el verbo

[34] El apartado 4.3.5 (pág. 122) detalla la obtención y el uso de esta información ontológica.

anafórico. Siguiendo los criterios detallados en la descripción del método ERA, se dotará a cada candidato de un peso asociado a su compatibilidad semántica con el verbo del pronombre según el papel sintáctico que éste realice.

Si tras la aplicación de esta preferencia más de un candidato tiene el peso máximo se escogerá el candidato más repetido en el texto y, si la anáfora permanece todavía sin resolver, el candidato más cercano al pronombre será elegido como el antecedente de la anáfora.

En este sentido, si bien el conjunto de restricciones y preferencias para los pronombres demostrativos y personales ha sido definido de manera común para ambos, es preciso mencionar la existencia de una diferencia importante entre ellos[35]: la función señaladora de los pronombres demostrativos establece criterios de cercanía o lejanía[36]. Así, el uso de pronombres demostrativos que refieren a elementos lejanos (*aquel, aquella, aquellos, aquellas*) obligan a hacer una excepción en la preferencia relativa a la selección del candidato más cercano, ya que, en estos casos, la cercanía con el pronombre entra en contradicción con el carácter de 'lejanía' antes mencionado.

4.3.10 La aplicación del método ERA

La figura 4.11 muestra el algoritmo de aplicación definido para el método enriquecido de resolución de la anáfora en español (ERA).

El cuadro 4.5 resume el conjunto de reglas de compatibilidad e incompatibilidad semánticas, así como el conjunto de restricciones y preferencias usado por el método ERA según el tipo de información que proporcionan.

---

[35] Agradezco a Joaquím Moré López sus comentarios sobre el algoritmo de conocimiento limitado para la resolución de la anáfora (Palomar et al., 2001a), tanto en lo relativo a este aspecto en particular como a otros de interés para este trabajo.

[36] Halliday y Hassan (1976) definen los pronombres demostrativos como pronombres que seleccionan a un participante de un evento o una circunstancia que está lejos o cerca en el espacio o en el tiempo.

Reglas de compatibilidad semántica	Reglas de incompatibilidad semántica	Condiciones de no-correferencia (restricciones)	Preferencias
Regla 1: Un verbo *v#sentv* es compatible con un nombre *n#sentn* como sujeto de *v#sentv* $\Leftrightarrow \exists c \in Ont(n\#sentn) \mid CompatSV(c,v\#sentv)>0$ **Regla 2**: Un verbo *v#sentv* es compatible con un nombre *n#sentn* como OD de *v#sentv* $\Leftrightarrow \exists c \in Ont(n\#sentn) \mid CompatVD(c,v\#sentv)>0$ **Regla 3**: Un verbo *v#sentv* es compatible con un nombre *n#sentn* como OD de *v#sentv* $\Leftrightarrow \exists c \in Ont(n\#sentn) \mid CompatVI(c,v\#sentv)>0$ **Regla 4**: Un candidato anafórico con núcleo *n1#sentn1* es preferido semánticamente frente a otro candidato con núcleo *n2#sentn2* como sujeto de un verbo *v#sentv* $\Leftrightarrow$ $\forall c_i \in Ont(n1\#sentn1)$, $\forall d_i \in Ont(n2\#sentn2)$, $\Sigma(CompatSV(c_i)) > \Sigma(CompatSV(d_i))$ **Regla 5**: Un candidato anafórico con núcleo *n1#sentn1* es preferido semánticamente frente a otro candidato con núcleo *n2#sentn2* como OD de un verbo *v#sentv* $\Leftrightarrow$ $\forall c_i \in Ont(n1\#sentn1)$, $\forall d_i \in Ont(n2\#sentn2)$, $\Sigma(CompatVD(c_i)) > \Sigma(CompatVD(d_i))$ **Regla 6**: Un candidato anafórico con núcleo *n1#sentn1* es preferido semánticamente frente a otro candidato con núcleo *n2#sentn2* como OI de un verbo *v#sentv* $\Leftrightarrow$ $\forall c_i \in Ont(n1\#sentn1)$, $\forall d_i \in Ont(n2\#sentn2)$, $\Sigma(CompatVI(c_i)) > \Sigma(CompatVI(d_i))$	**Regla 7**: La regla *NO(v#sentv,c,r)* define la incompatibilidad del verbo *v#sentv* con cualquier nombre *n#sentn* que contenga a *c* en su lista de conceptos ontológicos *Ont(n#sentn)* siendo *r* la función sintáctica que les relaciona. **Regla 8**: La regla *DEBE(v#sentv,c,r)* define la incompatibilidad del verbo *v#sentv* con cualquier nombre *n#sentn* que no contenga a *c* en su lista de conceptos ontológicos *Ont(n#sentn)* siendo *r* la función sintáctica que les relaciona.	❖ **Condiciones morfosemánticas** ➢ Un SN y un pronombre no son correferentes si no concuerdan en género, número y persona, excepto si el pronombre es plural y el SN tiene el rasgo de `grupo' ❖ **Condiciones sintáctico-semánticas** ➢ Un SN con rasgo de `no animado' no puede correferir con un pronombre personal de sujeto no neutro. ➢ Los pronombres personales `le' y `les' con función de objeto directo sólo pueden correferir con un SN masculino con rasgo de `humano'. ❖ **Condiciones sintácticas** ➢ Pronombres reflexivos: ▪ El SN no es el sujeto del mismo verbo al que acompaña el pronombre. ➢ Pronombres personales, demostrativos y omitidos: ▪ El SN y el pronombre modifican al mismo verbo y desempeñan roles sintácticos diferentes. ▪ El SN está coordinado con el pronombre ❖ **Condiciones semánticas** ➢ Un SN no correfiere con un pronombre si queda definida su incompatibilidad a través de una regla ``no'' o una regla``debe'' de incompatibilidad semántica con respecto a la anáfora y al papel sintáctico que ésta representa en relación a su verbo.	❖ **Preferencias morfológicas** ➢ SN que concuerdan en número con la anáfora. ❖ **Preferencias sintácticas** ➢ SN con la misma función sintáctica que la anáfora. ➢ SN no incluidos en otro SN. ❖ **Preferencias semánticas** ➢ SN que no son de tiempo, dirección, cantidad o abstracto. ➢ SN semánticamente más compatibles con el verbo del pronombre. ❖ **Preferencias estructurales** ➢ SN en la misma oración, siendo la preferencia mayor en función de la cercanía entre candidato y anáfora. ➢ SN solución de pronombres omitidos anteriores ❖ **Preferencias semántico-estructurales** ➢ SN (o sinónimos) que se han repetido más veces en el texto, especialmente si se han repetido con el mismo verbo de la anáfora (con su misma función sintáctica).

Cuadro 4.5. Resumen de reglas de compatibilidad e incompatibilidad semántica, restricciones y preferencias usadas en el método ERA

```
-----------------------------------------------------------------------
Para cada oración O
  L = L + Almacenar los SN de O con sus datos de enriquecimiento
  Adquisición de patrones de compatibilidad con los SN de L
  Para cada pronombre P en O
    Identificación de tipo del pronombre P
    Aplicación de restricciones a L en función del tipo de P
      L'=Aplicar restricciones morfosemánticas a L
      L'=Aplicar restricciones sintáctico-semánticas a L
      L'=Aplicar restricciones sintácticas a L
      L'=Aplicar restricciones de reglas de incompatibilidad a L
    Si |L'| = 0 entonces P no es anafórico
    Si |L'| = 1 entonces L[1] es el antecedente de P
    Si |L'| >1 entonces
      Aplicación de preferencias a L' según el tipo de P
        L' = Aplicar preferencias estructurales y semántico-estructurales a L'
        L' = Aplicar preferencias morfológicas a L'
        L' = Aplicar preferencias sintácticas a L'
        L' = Aplicar preferencias semánticas a L'
      L'' = Mejor(L')
      Si |L''| = 1 entonces L[1] es el antecedente de P
      Si |L''| >1 entonces
        L' = Aplicar preferencias comunes
        Mejor(L') es el antecedente de P
      finSi
    finSi
  finPara
finPara
-----------------------------------------------------------------------
```

Figura 4.11. Algoritmo de aplicación del método ERA.

4.4 Conclusiones

La necesidad del uso de información semántica en los procesos de resolución de la anáfora ha sido un reto que siempre ha preocupado a los investigadores en este área. Una de las razones fundamentales que ha frenado las propuestas de aproximaciones en esta línea ha sido la falta de recursos lingüísticos que proporcionen las fuentes de información requeridas para su desarrollo.

En este trabajo se han propuesto dos aproximaciones a la resolución de la anáfora pronominal en español. Una, basada en conocimiento limitado puramente morfosintáctico y otra, basada en conocimiento enriquecido con semántica y con papeles sintácticos.

El enfoque con conocimiento limitado parte del estudio y simplificación del algoritmo presentado en Palomar et al. (2001a) y define un método basado en un conjunto de restricciones y preferencias de carácter morfológico, sintáctico y estructural. Mientras que las restricciones definen condiciones de no correferencia y eliminan candidatos incompatibles con la anáfora, las preferencias ponderan cada candidato en función de su cumplimiento y establecen los criterios necesarios para la selección del más adecuado como antecedente anafórico.

Tanto las ventajas como los inconvenientes que presenta este método tienen el mismo origen: el uso limitado de recursos que precisa. Esto hace que el método de conocimiento limitado sea computacionalmente más eficiente al tener menos información que extraer y consultar. Sin embargo, parece claro que un porcentaje de los fracasos de este método se podría subsanar con la aplicación de nuevas fuentes de información.

Por ello, y a pesar de que los resultados obtenidos por la propuesta basada en conocimiento limitado han sido satisfactorios, se ha propuesto un método (ERA) basado en un enriquecimiento de las fuentes de información.

Varias son las propuestas del método ERA:

- Propuesta del etiquetado necesario para la aplicación de una resolución de la anáfora que incluya la semántica entre sus fuentes de información.
- Propuesta de un módulo generador semántico que elabora la información semántica previamente etiquetada en el corpus con el uso del recurso léxico WordNet. El generador semántico es un núcleo fundamental del método ERA y tiene una doble función. Por un lado genera, a través de su módulo de extracción semántica, una serie de colecciones de datos semánticos relativos a las palabras aparecidas en el texto y, por otro lado, construye, con el módulo generador de patrones, un conjunto de relaciones o patrones formados por los conceptos ontológicos de un nombre con función de sujeto, objeto directo u objeto indirecto y el verbo al que acompañan. Durante la fase de resolución de la anáfora los patrones que combinan los conceptos ontológicos asociados

a cada uno de los candidatos a antecedente con el verbo de la anáfora serán contrastados con los patrones previamente adquiridos para determinar cuál de ellos es el más adecuado aplicando un conjunto de reglas de compatibilidad semántica.

- Propuesta de dos tipos de reglas de incompatibilidad semántica ("no" y "debe") que determinan las condiciones semánticas de no correferencia entre un sintagma nominal y un pronombre permitiendo al módulo de restricciones eliminar candidatos no deseados.
- Propuesta de un método de resolución de la anáfora pronominal en español que integra la semántica, basada en corpus (patrones de compatibilidad semántica) y en conocimiento (patrones de incompatibilidad semántica) con la información morfológica, sintáctica y semántica de los elementos oracionales. Este método define un conjunto de restricciones morfosemánticas y sintácticas que eliminan candidatos incompatibles y un conjunto de preferencias morfológicas, sintácticas, semánticas y estructurales que ponderan cada uno de los candidatos compatibles con la anáfora para determinar cuál es el antecedente correcto.

A lo largo del siguiente capítulo se mostrarán y discutirán los resultados, tanto cuantitativos como cualitativos, de la aplicación de ambos métodos sobre un corpus de evaluación.

5. Evaluación

5.1 Introducción

Los métodos anteriormente presentados muestran dos estrategias de resolución de la anáfora claramente diferentes. Por un lado, el método basado en conocimiento limitado partirá de un análisis parcial y aplicará criterios puramente morfosintácticos. Por otro lado, el método enriquecido partirá de un análisis parcial enriquecido y usará criterios morfológicos, sintácticos y semánticos. Para estos últimos, hará uso del recurso léxico WordNet así como de un conjunto de patrones semánticos generados en la fase de aprendizaje.

Para cada uno de los métodos propuestos se definirán el conjunto de herramientas y recursos utilizados para la evaluación, tanto en lo referente a los módulos implementados como a los corpus utilizados en las distintas fases y experimentos de la evaluación.

Se expondrá un conjunto de datos tanto cuantitativos como cualitativos referentes al comportamiento de estos métodos en el proceso de resolución, ilustrando dicho comportamiento con ejemplos extraídos del corpus.

Primero se tratarán los resultados obtenidos para el método de resolución de conocimiento limitado y a continuación los del método enriquecido. Para la evaluación de este último se realizarán varios experimentos en los que se estudiará el comportamiento de cada una de las restricciones y preferencias propuestas por el sistema, tanto de forma aislada como conjunta. Este estudio permitirá medir la influencia que tiene cada una de las fuentes de información en la resolución de la anáfora.

Este capítulo finalizará con una reflexión acerca del método propuesto.

5.2 Evaluación del uso de conocimiento limitado en la resolución de la anáfora en español

A lo largo de esta sección se expondrán los resultados obtenidos en la evaluación del método de conocimiento limitado propuesto en el capítulo anterior y basado en información morfosintáctica.

Además de la definición de recursos y herramientas que han sido utilizados para esta evaluación, se expondrá la estrategia seguida para llevarla a cabo así como los resultados obtenidos sobre los corpus de evaluación. Estos resultados se compararán con los proporcionados por las distintas implementaciones de conocidos algoritmos basados también en conocimiento limitado.

5.2.1 Herramientas y recursos utilizados

El corpus. Para la evaluación de este método, tal y como se ha comentado en la sección anterior, se usaron textos pertenecientes a dos corpus. Por un lado, se extrajeron textos del corpus BlueBook[1], que contiene el manual de la Union de Telecomunicaciones Internacional CCITT, publicado en inglés, francés y español. Este corpus contiene unos 5 millones de palabras etiquetadas automáticamente por el etiquetador léxico-morfológico Xerox (Cutting et al., 1998) adaptado al español. Por otro lado, se utilizó el corpus Lexesp[2] que ha sido anotado léxico-morfológicamente por los analizadores *maco* (Atserias et al., 1998) y *relax* (Padró, 1997) con el conjunto de etiquetas PAROLE (Martí et al.,

[1] El corpus BlueBook pertenece al proyecto CRATER (*Corpus Resources and Terminology Extraction Project*) financiado por la Comisión Europea (DG-XIII) y desarrollado por el Laboratorio de Lingüística Computacional de la Facultad de Filosofía y Letras de la Universidad Autónoma de Madrid, España (1994-1995). Más información en `http://www.lllf.uam.es/docs_en/final_report/drep22.html` (última visita marzo 2002).

[2] El corpus Lexesp pertenece al proyecto del mismo nombre llevado a cabo por el Departamento de Psicología de la Universidad de Oviedo (España), y desarrollado por el Grupo de Lingüística Computacional de la Universidad de Barcelona (España), con la colaboración del Grupo de Procesamiento del Lenguaje de la Universidad Politécnica de Cataluña (España).

1998). Este corpus contiene textos muy variados, escritos por diferentes autores y sobre distintos dominios: novela, política, noticias, viajes, religión, ...

El etiquetado léxico-morfológico. Tal y como se ha dicho, los dos corpus usados para la evaluación del sistema de conocimiento limitado han sido preprocesados por dos etiquetadores diferentes.

En el caso del Bluebook, el etiquetador léxico-morfológico Xerox se encargó de añadir a cada palabra del corpus su lema o raíz correspondiente así como una etiqueta con rasgos morfológicos. El etiquetador Xerox cuenta con un lexicón de 440000 formas completas derivadas de 40000 lemas y un conjunto de 475 etiquetas. El cuadro 5.1 muestra un ejemplo de salida del etiquetador a partir de una frase del corpus Bluebook en español.

Estos protocolos permiten controlar los bucles y las pruebas de diagnóstico...

```
Estos este DMPXMP
protocolos protocolo NCMP
permiten permitir VLPI3P
controlar controlar VLINF
los el ARTDMP
bucles bucle NCMP
y y CC
las el ARTDFP
pruebas prueba NCFP
de de PREP
diagnóstico diagnóstico NCMS
```

Cuadro 5.1. Ejemplo de etiquetado léxico morfológico del etiquetador Xerox (Cutting et al., 1998)

La etiqueta que acompaña a la palabra y su raíz proporciona información sobre la categoría gramatical (ART-artículo, NC-nombre común, PREP-preposición,...) y el resto de datos sobre su morfología (M-masculino, F-femenino, P-plural, I-indicativo, 3P-tercera persona,...).

En lo referente al corpus Lexesp, cada palabra va acompañada de una etiqueta perteneciente al conjunto de etiquetas PAROLE (Martí et al., 1998). La información básica que proporciona este conjunto de etiquetas es similar al anterior, contando con un total

de 230 etiquetas que representan la categoría gramatical así como los rasgos morfológicos e información de modo, tiempo y persona para los verbos. En el cuadro 5.2 se muestra un ejemplo de este etiquetado sobre una frase extraída del corpus Lexesp.

La igualdad en el destino determinaba un igual acento en la diversidad de rostros y expresiones.

```
La la TDFS0
igualdad igualdad NCFS000
en en SPS00
el el TDMS0
destino destino NCMS000
determinaba determinar VMII3S0
un un TIMS0
igual igual AQ0CS00
acento acento NCMS000
en en SPS00
la la TDFS0
diversidad diversidad NCFS000
de de SPS00
rostros rostro NCMP000
y y CC00
expresiones expresión NCFP000
. . Fp
```

Cuadro 5.2. Ejemplo de etiquetado léxico morfológico con etiquetas PAROLE (Martí et al., 1998)

El analizador sintáctico. A partir del corpus anotado por el etiquetador léxico-morfológico, el analizador sintáctico SUPP (Ferrández et al., 1998), basado en el formalismo gramatical SUG (Ferrández et al., 1997), genera un análisis parcial del texto. Este análisis está formado por las estructuras de huecos que almacenan toda la información necesaria, así como información del discurso a través de una lista de antecedentes aparecidos con anterioridad para su uso en la resolución de la anáfora. El cuadro 5.3 muestra un ejemplo de generación de análisis parcial a partir de una oración extraída del corpus Lexesp.

Los árboles estaban pelados y, en la desnudez de las ramas, los gorriones parecían más gordos.

```
** ORACION ANALIZADA PARCIALMENTE:
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** DETERMINANTE 1:
            ** ARTICULO (pl,masc,det): Los
            ** SUSTANTIVO (pl,masc,comun): árboles
   ** NUCLEO VERBAL:
         ** VERBO (pl,terc,imperfecto,noCopul): estaban
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** ADYACENTE ADJETIVO:
               ** ADJETIVO SIMPLE (pl,masc,cal): pelados
   ** CONJUNCION: y
   ** CONJUNCION: ,
   ** SINT.PREPOSICIONAL:
      ** SINT.PREPOSICIONAL SIMPLE:
         ** PREPOSICION:
            ** PREPOSICION SIMPLE: en
         ** SINT.NOMINAL:
            ** SINT.NOMINAL SIMPLE:
               ** DETERMINANTE 1:
                  ** ARTICULO (sing,fem,det): la
                  ** SUSTANTIVO (sing,fem,comun): desnudez
               ** SINT.PREPOSICIONAL:
                  ** SINT.PREPOSICIONAL SIMPLE:
                     ** PREPOSICION:
                        ** PREPOSICION SIMPLE: de
                     ** SINT.NOMINAL:
                        ** SINT.NOMINAL SIMPLE:
                           ** DETERMINANTE 1:
                              ** ARTICULO (pl,fem,det): las
                              ** SUSTANTIVO (pl,fem,comun): ramas
   ** CONJUNCION: ,
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** DETERMINANTE 1:
            ** ARTICULO (pl,masc,det): los
            ** SUSTANTIVO (pl,masc,comun): gorriones
   ** NUCLEO VERBAL:
         ** VERBO (pl,terc,imperfecto,noCopul): parecían
      ** ADVERBIO: más
   ** SINT.NOMINAL:
      ** SINT.NOMINAL SIMPLE:
         ** ADYACENTE ADJETIVO:
               ** ADJETIVO SIMPLE (pl,masc,cal): gordos
   ** CONJUNCION: .
```

Cuadro 5.3. Ejemplo de análisis sintáctico parcial SUPP (Ferrández et al., 1998)

5.2.2 Resultados del método de conocimiento limitado

Se seleccionaron para la evaluación un subconjunto de ambos corpus y se anotaron anafóricamente. La fase de anotación se realizo de la siguiente manera:

1. Se seleccionaron dos anotadores.
2. Se establecieron las normas de anotación.
3. Los anotadores realizaron su tarea en paralelo sobre el corpus.
4. Sobre la anotación, se realizó un test de confianza (Carletta, 1996; Carletta et al., 1997) para garantizar los resultados[3].

En lo referente a la medida de evaluación utilizada, en los resultados hablaremos de tasa de éxito, tasa resultante del cociente entre el número de pronombres correctamente resueltos y el número total de pronombres.

El cuadro 5.4 muestra los resultados para cada tipo de pronombre, resultando una tasa de éxito del 76,8 %.

	Personales	*Demostrativos*	*Omitidos*	*Reflexivos*	TOTAL
Total	429	69	1 099	80	1677
Resueltos	296	51	868	74	1289
Éxito	69,0 %	73,9 %	78,9 %	92,5 %	76,8 %

Cuadro 5.4. Resultados de la evaluación del método de conocimiento limitado

Como puede comprobarse, los resultados obtenidos por el método basado en conocimiento limitado son globalmente satisfactorios. Si bien el método falla en la selección del antecedente correcto en un 23,2 % de los casos, tras realizar un análisis de los errores estos pueden ser atribuidos a los siguientes factores:

- Errores en el etiquetador gramatical: los errores provocados por etiquetados incorrectos de la categoría gramatical ascienden a un 3 % de los errores totales.
- Errores en el análisis parcial: los errores provocados por la incorrecta identificación de sintagmas nominales complejos ascienden a un 7 % aproximadamente.
- Ausencia de información semántica: se ha considerado que la incorporación de información semántica podría ayudar aproximadamente en un 32 % de los casos en los que el método de conocimiento limitado no fue capaz de resolver la anáfora correctamente.

[3] Para más información sobre esta estrategia de anotación y verificación, consultar Palomar et al. (2001a).

- Excepciones en las preferencias: aproximadamente un 43 % de los errores se debían a casos especiales que las preferencias no tenían en cuenta.
- El resto de los errores se puede atribuir a antecedentes mal divididos (10 %), catáforas (2 %) y exóforas (3 %).

El siguiente apartado enmarcará los resultados obtenidos por este método en un conjunto de resultados obtenidos en la implementación de métodos de resolución de la anáfora basados también en conocimiento limitado.

5.2.3 Comparación directa con otros métodos implementados

El método de conocimiento limitado ha sido comparado con otros métodos clásicos recogidos en la bibliografía (ver sección 3.1 en la pág. 30).

Para llevar a cabo esta comparación de nuestro método con otros métodos basados en conocimiento limitado, se realizaron implementaciones de algunos algoritmos conocidos. Así, se implementaron el algoritmo *naif* de Hobbs (1978), el algoritmo de Lappin y Leass (1994) y una aproximación basada en la teoría del *centering* (Strube, 1998). Además, se utilizó como caso base el propuesto con sus mismas restricciones, eliminando las preferencias y usando el criterio de selección del candidato más cercano como medida de "desempate". El cuadro 5.5 muestra los resultados obtenidos por cada implementación.

	Personales	*Demostrat.*	*Omitidos*	*Reflexivos*	TOTAL
Pronombres	429	69	1099	80	1677
Base	60,3 %	75,0 %	47,0 %	86,0 %	53,4 %
Hobbs	63,0 %	51,0 %	62,0 %	85,0 %	62,9 %
Lappin y Leass	66,0 %	60,0 %	67,0 %	86,0 %	67,4 %
Centering	61,0 %	59,0 %	62,0 %	85,0 %	62,7 %
Método CL	68,0 %	77,0 %	79,0 %	92,0 %	76,7 %

Cuadro 5.5. Comparación de resultados de la evaluación del método de conocimiento limitado (CL) con respecto a otros métodos implementados

Como puede verse en esta comparación de datos, el método propuesto supera los resultados proporcionados por el resto de los métodos en un número de anáforas que oscila entre el 9 % y el 14 %. Es importante destacar a este respecto que, para realizar esta comparación, ha sido necesario adaptar los algoritmos implementados al español, ya que su concepción original se fundamentaba en el inglés. Es por ello que, en algunos casos, las implementaciones difieren ligeramente de los planteamientos originales. Este problema es un obstáculo insalvable cuando se desea comparar métodos desarrollados para idiomas diferentes, especialmente en idiomas con características tan dispares como el inglés y el español en lo que a la resolución de la anáfora se refiere.

5.3 Evaluación del método ERA

En esta sección se tratarán los aspectos relativos a la evaluación del método ERA, cuya principal aportación es la incorporación tanto de información asociadas a los papeles sintácticos como de información semántica.

Siguiendo una estructura similar a la de la sección anterior, se hará un repaso de los recursos y herramientas usados para la evaluación así como la explicación de la estrategia utilizada para llevarla a cabo.

La sección finalizará con un cuadro que recoge los resultados globales obtenidos en el proceso de evaluación, siendo la siguiente sección la encargada de mostrar el estudio detallado de estos datos para determinar la influencia de las diferentes fuentes de información que intervienen en el proceso de resolución.

5.3.1 Herramientas y recursos utilizados

El corpus. El corpus utilizado para la evaluación del método ERA está formado mayoritariamente por fragmentos extraídos del corpus Lexesp (ver 5.2.1). Este corpus está formado por textos complejos que, por su riqueza lingüística, suponen un reto para los sistemas de resolución anafórica. Por otro lado, es un corpus

variado, cuya diversidad en los temas contenidos es un punto muy importante para tareas de PLN orientadas a dominios no restringidos. Del Lexesp se han extraído los dos primeros bloques del corpus de evaluación, correspondientes a un artículo de opinión (L009) y a un texto narrativo (L065).

Adicionalmente a los fragmentos escogidos del Lexesp, se ha incorporado al corpus de evaluación un bloque de oraciones que han servido a lo largo de este trabajo como ejemplos de aplicación de los distintos criterios de restricción y preferencia (E001) .

El cuadro 5.6 muestra algunos datos relativos a los tres bloques mencionados.

	L009	L065	E001	TOTAL
N° oraciones	36	92	27	155
N° palabras	861	1951	187	2999
N° de anáforas	31	72	18	121

Cuadro 5.6. Composición del corpus de evaluación para el método ERA

Etiquetado y análisis. Las características tanto del etiquetado léxico-morfológico como del análisis parcial base del corpus coinciden con las propuestas en 5.2.1 (pág. 154) para el método de conocimiento limitado.

Dados los requisitos de este método, ampliamente tratados en 4.3.2 (pág. 115), ha sido necesario etiquetar el conjunto de oraciones que forman el corpus, por un lado, con información adicional sobre los papeles sintácticos de los sintagmas nominales (sujeto, objeto directo y objeto indirecto) y, por otro, con los sentidos correctos consultados en WordNet español. Esta tarea, completamente manual, establece limitaciones evidentes tanto en la extensión del corpus como en la propia estrategia de evaluación.

El recurso semántico: WordNet. La descripción y características generales de este recurso han sido previamente detalladas en 4.3.4 (pág. 120). En lo referente a las especificaciones particulares del WordNet utilizado, cabe mencionar que la versión escogida ha sido la del WordNet español, distribuida por la Asociación de Recursos de Lenguajes Europeos (*ELRA*). El WordNet

español consta de 23370 *synsets* con un total de 50526 sentidos. Entre estos synsets se han establecido un total de 55163 relaciones internas y 21236 relaciones de equivalencia. El cuadro 5.7 resume estos datos y establece la comparación de éstos con los del resto de los idiomas[4].

Idioma	nº de *synsets*	nº de sentidos	Relaciones internas	Relaciones de equivalencia
Inglés	16361	40588	42140	0
Holandés	44015	70201	111639	53448
Español	23370	50526	55163	21236
Italiano	48529	48499	117068	71789
Alemán	15132	20453	34818	16347
Francés	22745	32809	49494	22730
Checo	12824	19949	26259	12824
Estonio	9317	13839	16318	9004

Cuadro 5.7. Distribución de *synsets* y relaciones para los distintos WordNets de idiomas europeos

La distribución del WordNet español incluye además el conjunto de registros inter-lenguas (*ILI*) así como la ontología principal (*Top Ontology*) usada en el método ERA para los patrones de compatibilidad e incompatibilidad semántica.

Si bien existe en dicha distribución una interfaz (*Periscope*) para poder consultar las palabras y sus sentidos contenidos en WordNet, este recurso no viene acompañado de herramientas adecuadas para su consulta y manipulación desde un lenguaje de programación. Por ello, ha sido necesario desarrollar un conjunto de módulos y librerías para instrumentar el acceso a las bases de datos de synsets y facilitar así su gestión. Estas librerías, al igual que el resto de los módulos que integran la implementación, han sido desarrolladas en C++.

5.3.2 Entorno de evaluación: el banco de pruebas

Para llevar a cabo la evaluación del método ERA se ha diseñado un banco de pruebas que integra la implementación de dicho método. La interfaz del banco de pruebas permite hacer un seguimiento

[4] Datos extraídos de la página de la Agencia de Distribución de recursos de Lenguajes Europeos (*ELDA*). `http://www.elda.fr/` (última visita en marzo de 2002).

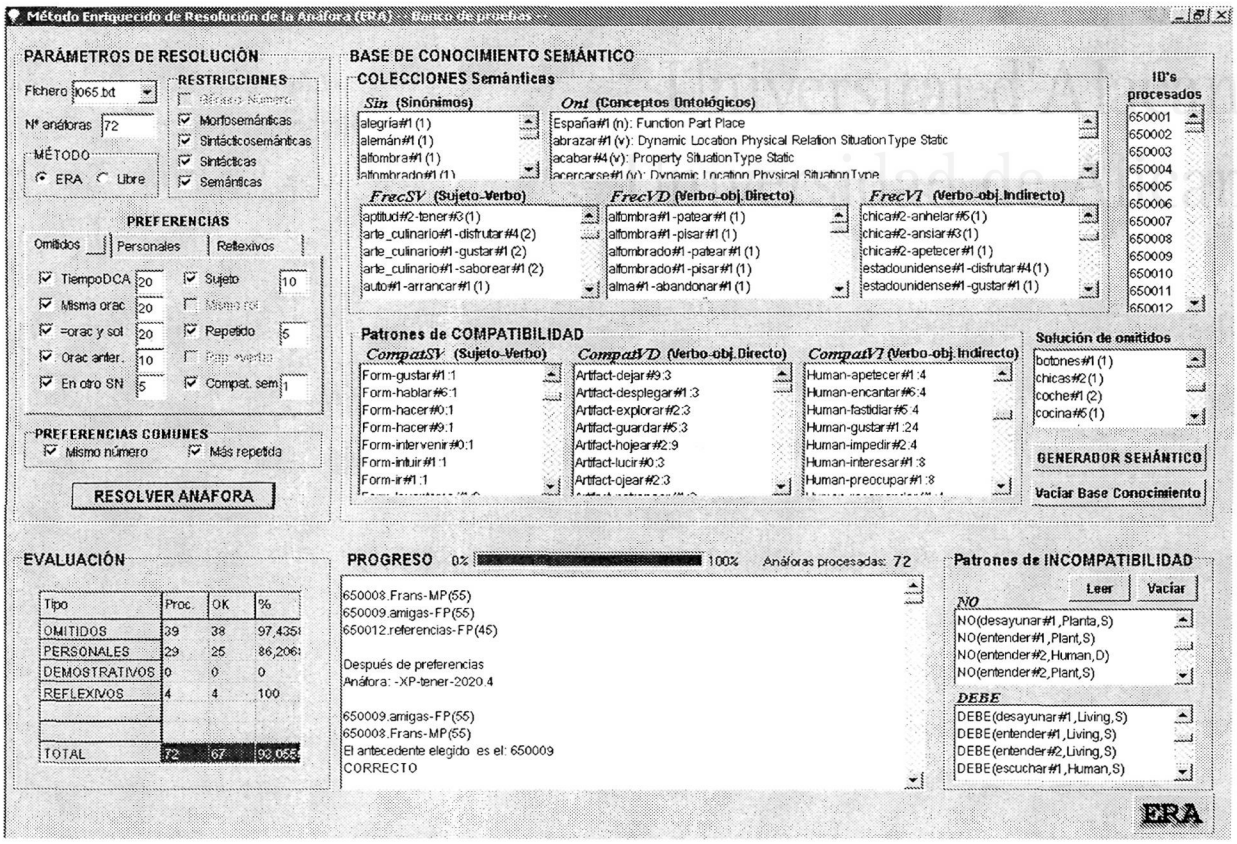


Figura 5.1. Interfaz del banco de pruebas de evaluación del método ERA

completo de los mecanismos asociados a la aplicación del método ERA en el corpus de evaluación. La figura 5.1 muestra una captura de dicha interfaz.

Una de las características esenciales del banco de pruebas es su capacidad total de configuración, permitiendo la posibilidad de activar y desactivar cualquiera de las restricciones y preferencias definidas en el método ERA. La figura 5.2 muestra una ampliación del módulo de configuración de parámetros en el que se puede comprobar su flexibilidad en lo relativo a la selección individual de cada restricción y preferencia.

Con el fin de poder establecer en todo momento un control sobre el proceso de resolución de la anáfora, la interfaz cuenta con una serie de indicadores de progreso: una barra porcentual, un contador de anáforas resueltas y una ventana de salida en la que se muestran los resultados de cada una de las fases de aplicación de restricciones y preferencias, la selección de los candidatos

Figura 5.2. Parámetros de configuración en el banco de pruebas

escogidos y un resumen final de pronombres mal resueltos y de los datos de evaluación. La figura 5.3 muestra un detalle de estos indicadores de progreso mientras que en el cuadro 5.8 aparece un ejemplo de una posible salida de la interfaz para una anáfora extraída del corpus de evaluación.

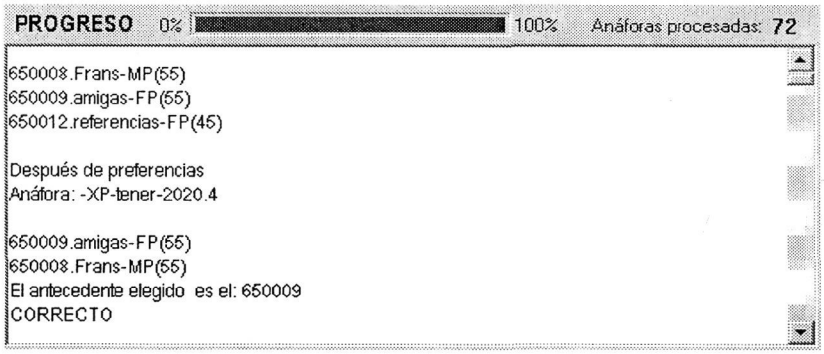


Figura 5.3. Indicadores de progreso en el banco de pruebas

Además, se ha incorporado en la interfaz un conjunto de módulos que muestran tanto los patrones de incompatibilidad semántica *NO* y *DEBE* como el conjunto de elementos pertenecientes a

Lisbeth, que patroneaba entonces el yate, se dirigió a Frans: - - Por favor, ¿quieres determinar la posición exacta del barco? Creo que estamos aproximándonos ya a la costa española. Van Steen, que estaba a su lado en el puente, extrajo el sextante de una caja de madera barnizada, y se dirigió con **él** a cubierta.

```
-----------
Anáfora 5.
-----------

Anáfora: él-MS-dirigirse-2023.3

Antecedentes:
650028.Frans-MP
650029.mar-FS
650030.Lisbeth-FS
650031.yate-MS
650032.Frans-MS
650033.posición-FS
650034.barco-MS
650035.costa-FS
650036.Van_Steen-MS
650037.lado-MS
650038.puente-MS
650039.sextante-MS
650040.caja-FS
650041.madera-FS
650042.Van_Steen-MS
650044.cubierta-FS

Restricciones morfosemánticas:
Elimino 'Frans' por GEN-NUM
Elimino 'mar' por GEN-NUM
Elimino 'Lisbeth' por GEN-NUM
Elimino 'posición' por GEN-NUM
Elimino 'costa' por GEN-NUM
Elimino 'caja' por GEN-NUM
Elimino 'madera' por GEN-NUM
Elimino 'cubierta' por GEN-NUM

Después de restricciones morfosemánticas
650031.yate-MS
650032.Frans-MS
650034.barco-MS
650036.Van_Steen-MS
650037.lado-MS
650038.puente-MS
650039.sextante-MS
650042.Van_Steen-MS

Después de restricciones sintáctico-semánticas
650031.yate-MS
650032.Frans-MS
650034.barco-MS
650036.Van_Steen-MS
650037.lado-MS
650038.puente-MS
650039.sextante-MS
650042.Van_Steen-MS

Restricciones sintácticas:
Elimino 'Van_Steen' por ROLES diferentes
Después de restricciones sintácticas
650031.yate-MS
650032.Frans-MS
650034.barco-MS
650036.Van_Steen-MS
650037.lado-MS
650038.puente-MS
650039.sextante-MS

Después de restricciones semánticas
650031.yate-MS
650032.Frans-MS
650034.barco-MS
650036.Van_Steen-MS
650037.lado-MS
650038.puente-MS
650039.sextante-MS

Preferencias
650031.yate-MS(40)
650032.Frans-MS(40)
650034.barco-MS(35)
650036.Van_Steen-MS(50)
650037.lado-MS(45)
650038.puente-MS(45)
650039.sextante-MS(55)

Después de preferencias
650039.sextante-MS(55)
El antecedente elegido es el: 650039
CORRECTO
```

Cuadro 5.8. Ejemplo de salida de la implementación del método ERA en la aplicación de restricciones y preferencias.

la base de conocimiento construida por el generador semántico e integrada en el proceso de resolución de la anáfora.

Los patrones *NO* y *DEBE* sirven como base para la aplicación de las reglas 7 y 8 con el mismo nombre y quedan representados[5] en el módulo de patrones de incompatibilidad semántica cuyo detalle aparece en la figura 5.4.

[5] Si bien los datos proporcionados por estas ventanas de la interfaz son meramente informativos y tan sólo permiten comprobar qué patrones se están aplicando en la fase de resolución, versiones futuras permitirán la incorporación de nuevos patrones a través de la propia interfaz.

Figura 5.4. Representación de patrones de incompatibilidad semántica en el banco de pruebas

La base de conocimiento semántico, tanto en lo referente a las colecciones semánticas como a los patrones de compatibilidad construidos a partir de ellas, tiene también representación en esta interfaz (ver figura 5.5).

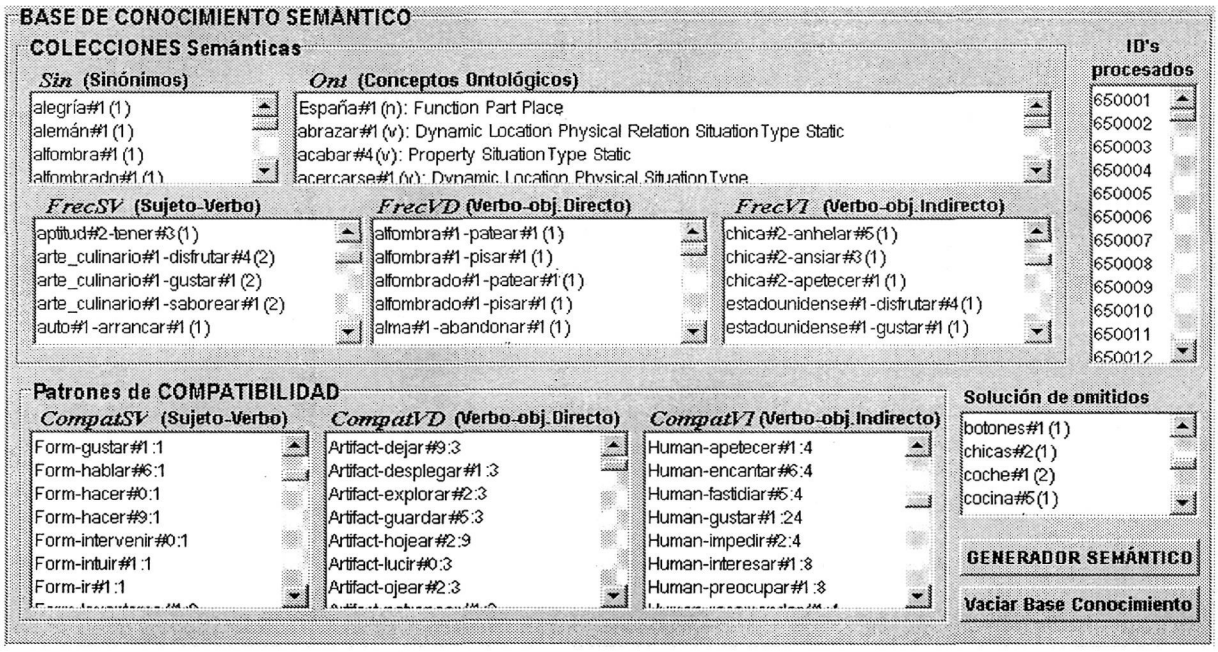


Figura 5.5. Representación de la base de conocimiento semántico en el banco de pruebas

Además, el proceso de generación de patrones de compatibilidad es independiente del de resolución de la anáfora, permitiendo así la adquisición previa de patrones descrita en el capítulo anterior. Esta independencia ha permitido evaluar la influencia de

la adquisición previa de patrones semánticos sobre los resultados globales de la resolución de la anáfora.

Las características visuales de este banco de pruebas, así como sus posibilidades de configuración, han permitido evaluar el comportamiento del método y la influencia de las distintas fuentes de conocimiento en la resolución de la anáfora, seleccionando o eliminando la aplicación de las distintas restricciones y preferencias y comprobando los resultados finales de la evaluación (ver figura 5.6). Todas y cada una de las pruebas realizadas en la evaluación del método serán convenientemente descritas en el siguiente apartado.

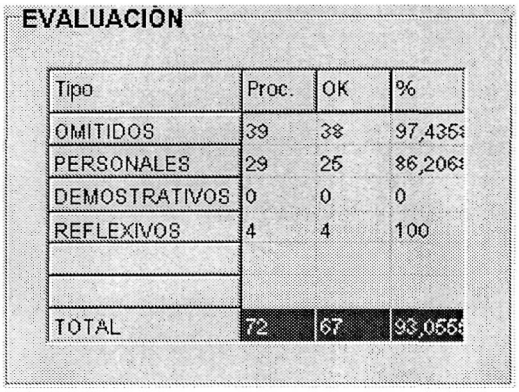


Figura 5.6. Ventana de evaluación en el banco de pruebas

5.3.3 Base de experimentación

A lo largo de este apartado se detallará el proceso seguido para la evaluación del método ERA, mientras que en los apartados que siguen a éste se relacionarán los resultados obtenidos con las fuentes de información integrantes del método, con el fin de medir la influencia de cada una de éstas en el proceso de resolución de la anáfora.

Como ya se ha dicho, el corpus de evaluación está formado por un conjunto de oraciones previamente analizadas morfosintácticamente y etiquetadas manualmente con los enriquecimientos necesarios para la aplicación del método. Este etiquetado manual adicional ha sido necesario al no disponer de ningún corpus que

cuente con dicha información o de recurso alguno que la proporcione de manera automática. Así pues, las necesidades adicionales del método ERA dificultan la comparación de sus resultados con los de otros métodos. Debido a esta dificultad se ha preferido un enfoque basado menos en los resultados globales de la aplicación del método y más en los resultados parciales de la incorporación o no de cada una de las fuentes de conocimiento que intervienen en el proceso de resolución.

Se han establecido unos pesos iniciales basados en el comportamiento de cada uno de los criterios estudiados en el método, con lo que ha sido posible utilizar la totalidad de las oraciones etiquetadas como corpus de evaluación. Estos pesos (ver cuadro 5.9) se han mantenido inamovibles durante todo el proceso.

	Peso
NO tiempo/dirección/cantidad/abstracto	20
Misma oración	20
Misma oración y solución de pron. omit.	20
Oración anterior	10
Sujeto	10
Mismo papel sintáctico	10
No en otro SN	5
Repetido	5
Repetido con el verbo de la anáfora	5

Cuadro 5.9. Pesos asignados a cada preferencia en el método ERA

Aprovechando la flexibilidad del banco de pruebas para la configuración de los parámetros de resolución, se han realizado diferentes pruebas con el fin de obtener datos relativos a la influencia de las distintas fuentes de conocimiento en el proceso de resolución de la anáfora sobre el corpus seleccionado.

El proceso de evaluación se ha realizado a partir de cuatro experimentos. En los experimentos primero y segundo se ha estudiado de forma independiente el comportamiento de, por una parte, las distintas condiciones de no correferencia (restricciones) y, por otra, las distintas preferencias definidas en el método. En el tercer experimento se han aplicado las restricciones y las preferencias de forma conjunta. Estos tres experimentos recogen resultados asociados a las distintas fuentes de información (morfológica,

sintáctica, semántica y estructural) en las que se agrupan las restricciones y las preferencias.

El cuarto experimento, orientado fundamentalmente al componente semántico de la propuesta, ha tenido en cuenta la influencia de la adquisición previa de patrones de compatibilidad semántica en el proceso de resolución.

A lo largo de esta sección se detallará el procedimiento seguido para el desarrollo de cada experimento. El objetivo de este capítulo es mostrar la metodología y la base de la evaluación. El anexo A (pág. 255) reúne los datos con los resultados de todos los experimentos realizados. Como ya se ha dicho, en los siguientes apartados se presentará la interpretación de todos estos datos.

Experimento 1. Estudio de las restricciones. El objetivo de este experimento es determinar la influencia que en el proceso de resolución tiene cada una de las restricciones propuestas en el método (detalladas en el apartado 4.3.8):

- Restricciones morfológicas (género y número).
- Restricciones morfosemánticas.
- Restricciones sintáctico-semánticas.
- Restricciones sintácticas.
- Restricciones semánticas (patrones de incompatibilidad).

La medición de esta influencia se ha realizado desde dos puntos de vista: la adición y la supresión de restricciones.

Adición de restricciones. Con el fin de obtener los resultados que cada fuente de información proporciona de manera individual se han tomado, como caso base, los resultados de la resolución atendiendo únicamente a la selección del candidato más cercano. A partir de este caso base se han ido incorporando de forma individual las restricciones asociadas a las diferentes fuentes de información. Cada resultado, por tanto, revela la influencia que tiene por separado cada tipo de restricción.

La adición de restricciones ha constado de las siguientes pruebas:

- Caso base: selección del candidato más cercano.

- Adición únicamente de restricciones morfológicas.
- Adición únicamente de restricciones morfosemánticas.
- Adición únicamente de restricciones sintáctico-semánticas.
- Adición únicamente de restricciones sintácticas.
- Adición únicamente de restricciones semánticas.

El cuadro de la sección A.1.1 (pág. 256) muestra los resultados parciales y globales de la adición de las distintas restricciones sobre el caso base.

Supresión de restricciones. Una vez medida la relevancia de cada restricción por separado, se han aplicado todas las restricciones de forma conjunta. El resultado obtenido ha servido como caso base para la eliminación individual de cada restricción.

El objetivo de esta prueba es comprobar de qué manera influye cada restricción al aplicarla conjuntamente con el resto. Después se han ido eliminando cada una de ellas de forma individual y se han medido los resultados de dicha eliminación.

La supresión de restricciones ha constado de las siguientes pruebas:

- Caso base: aplicación de todas las restricciones.
- Supresión únicamente de restricciones morfológicas y morfosemánticas.
- Supresión únicamente de restricciones sintáctico-semánticas.
- Supresión únicamente de restricciones sintácticas.
- Supresión únicamente de restricciones semánticas.

El cuadro de la sección A.1.2 (pág. 257) muestra los resultados del caso base y los asociados a la eliminación de cada una de las restricciones.

Experimento 2. Estudio de las preferencias. Siguiendo una estrategia similar a la aplicada en el primer experimento relativo a las restricciones, se ha realizado una valoración de la influencia de las preferencias propuestas en el método en función de la fuente de información que las agrupa. Así se han considerado los siguientes grupos de preferencias (detalladas en el apartado 4.3.9):

- Preferencias morfológicas:
  - SN con el mismo número que el pronombre (preferencia común).
- Preferencias sintácticas:
  - SN que no están en otro SN.
  - SN Sujeto.
  - SN con el mismo papel sintáctico que el pronombre.
- Preferencias semánticas:
  - SN que no son de tiempo, dirección cantidad ni tipo abstracto.
  - SN semánticamente compatibles con el pronombre (patrones de compatibilidad semántica).
- Preferencias estructurales:
  - SN en la misma oración que el pronombre.
  - SN en la misma oración que el pronombre y solución de un pronombre omitido anterior.
  - SN en la oración anterior a la del pronombre.
- Preferencias semántico-estructurales:
  - SN repetido en el texto.
  - SN repetido con el verbo de la anáfora en su mismo papel sintáctico.

Al igual que en el caso anterior, las pruebas sobre la influencia de las preferencias en la resolución de la anáfora se han realizado en función de su adición o su supresión.

Adición de preferencias. Partiendo de la idea de que las preferencias se aplican sobre aquellos candidatos que han superado la fase de restricciones y, por tanto, son potenciales antecedentes anafóricos, se ha considerado como base inicial de la adición de preferencias el resultado obtenido de la aplicación de todas las restricciones.

A partir de esta base, se han aplicado los grupos de preferencias antes comentados de forma individual para obtener los resultados de la aplicación de cada una de las fuentes de información por separado.

La adición de preferencias ha constado de las siguientes pruebas:

- Caso base: aplicación de todas las restricciones.
- Adición únicamente de preferencias morfológicas.
- Adición únicamente de preferencias sintácticas.
- Adición únicamente de preferencias semánticas.
- Adición únicamente de preferencias estructurales.
- Adición únicamente de preferencias semánticas y semántico-estructurales.

El cuadro de la sección A.2.1 (pág. 258) muestra los resultados de la adición de los distintos grupos de preferencias al caso base.

Supresión de preferencias. Para la supresión de preferencias se ha escogido como base la la aplicación de todas las restricciones y preferencias, por ser esta combinación la que proporciona los mejores resultados.

A partir de esta base se han ido suprimiendo grupos de preferencias de manera individual, comprobando el comportamiento del método con la ausencia de cada uno ellos.

La supresión de preferencias ha constado de las siguientes pruebas:

- Caso base: aplicación de todas las restricciones y todas las preferencias.
- Supresión únicamente de preferencias morfológicas.
- Supresión únicamente de preferencias sintácticas.
- Supresión únicamente de preferencias semánticas.
- Supresión únicamente de preferencias estructurales.
- Supresión únicamente de preferencias semánticas y semántico-estructurales.

El cuadro de la sección A.2.2 (pág. 259) muestra los resultados de la supresión de cada grupo de preferencias con respecto al caso base.

Experimento 3. Estudio conjunto de restricciones y preferencias. Con el fin de comprobar la influencia global de las distintas fuentes de información que intervienen en la resolución de la anáfora (morfológica, sintáctica, semántica y estructural) se

han realizado un conjunto de pruebas agrupando restricciones y preferencias en función de cada una de estas fuentes de información.

De nuevo, la estrategia seguida está basada en la adición y la supresión de cada conjunto de restricciones y preferencias.

Adición de restricciones y preferencias. El punto de partida de la adición de restricciones y preferencias es de nuevo el método de resolución basado en la selección del candidato más cercano.

Sobre esta base se han aplicado sucesivamente, y de forma independiente, cada uno de los grupos de restricciones y preferencias asociados a cada fuente de información.

Asimismo, dada la existencia de determinadas preferencias que integran, junto con el semántico, conocimiento de distintos tipos (morfosemánticas, sintáctico-semánticas y semántico-estructurales) se han realizado agrupaciones de restricciones y preferencias que combinan fuentes de información afines.

La adición de restricciones y preferencias ha contado con las siguientes pruebas:

- Caso base: selección del candidato más cercano.
- Adición únicamente de restricciones y preferencias morfológicas.
- Adición únicamente de restricciones y preferencias sintácticas.
- Adición únicamente de restricciones y preferencias semánticas.
- Adición de restricciones y preferencias semánticas combinadas (semánticas, morfosemánticas, sintáctico-semánticas y semántico-estructurales).
- Adición de restricciones y preferencias sintácticas combinadas (sintácticas y sintáctico-semánticas).
- Adición de restricciones y preferencias sintácticas y semánticas combinadas.

El cuadro de la sección A.3.1 (pág. 260) muestra los resultados de la adición de los diferentes grupos de restricciones y preferencias definidos sobre el caso base.

Supresión de restricciones y preferencias. Dado que los mejores resultados los proporciona la combinación de todas las fuentes de información, estos resultados definen la base de la supresión de los distintos conjuntos de restricciones y preferencias definidos.

Cada conjunto de restricciones y preferencias ha sido eliminado de forma individual del conjunto total, obteniendo los resultados asociados a la ausencia de cada uno de ellos en la resolución global del método.

La supresión de restricciones y preferencias ha contado con las siguientes pruebas:

- Caso base: aplicación de todas las restricciones y todas las preferencias.
- Supresión únicamente de restricciones y preferencias morfológicas.
- Supresión únicamente de restricciones y preferencias sintácticas.
- Supresión únicamente de restricciones y preferencias semánticas.
- Supresión de restricciones y preferencias semánticas combinadas (semánticas, morfosemánticas, sintáctico-semánticas y semántico-estructurales).
- Supresión de restricciones y preferencias sintácticas combinadas (sintácticas y sintáctico-semánticas).
- Supresión de restricciones y preferencias sintácticas y semánticas combinadas.

El cuadro de la sección A.3.2 (pág. 261) muestra los resultados de la supresión de cada grupo de restricciones y preferencias a partir del caso base completo.

Experimento 4. Estudio sobre la adquisición de patrones. Dado que uno de los objetivos fundamentales de este trabajo es estudiar la influencia de la información semántica en la resolución de la anáfora y dado que el método ERA incorpora esta información desde un conjunto de patrones de compatibilidad semántica extraídos automáticamente del corpus, se han realizado un conjunto

de pruebas para comprobar la influencia que la adquisición previa de patrones tiene sobre el proceso de resolución de la anáfora.

Estas pruebas han consistido en la evaluación independiente de cada uno de los tres bloques que forman el corpus. Para este experimento se han tomado dos casos base distintos. Por un lado, se ha partido de la aplicación de todas las fuentes de conocimiento y, por otro lado, de la aplicación de todas las restricciones y sólo las preferencias puramente semánticas.

Para cada uno de estos dos casos base, se han realizado dos experimentos. En primer lugar, se ha obtenido el resultado de evaluación de cada uno de los bloques a partir de los patrones de compatibilidad semántica adquiridos de los otros dos bloques. En segundo lugar, se ha realizado la adquisición de patrones a partir del corpus completo y se han obtenido los resultados para cada uno de los bloques.

El cuadro de la sección A.4 (pág. 262) muestra los resultados de estos dos experimentos sobre ambos casos base.

Interpretación de la experimentación. Si bien en los siguientes apartados se tratará con detenimiento la influencia de las fuentes de conocimiento a partir de los resultados obtenidos, uno de los puntos más importantes a destacar, en una primera reflexión, es el hecho de que los mejores resultados proporcionados por la implementación del método ERA, tanto en lo referente a restricciones como a preferencias, son los correspondientes a la aplicación conjunta de todas las fuentes de conocimiento. Esto parece indicar claramente que todas ellas contribuyen positivamente y de forma global a la obtención de mejores resultados.

A partir de los datos extraídos de la evaluación del método **ERA**, los siguientes apartados expondrán los puntos considerados más relevantes en la interpretación de los resultados obtenidos en las diferentes pruebas realizadas sobre el corpus de evaluación. El objetivo de dicha interpretación es el de determinar la influencia que tiene en el proceso de resolución de la anáfora cada una de las fuentes de información que intervienen. Así, cada una de estas secciones agrupará las interpretaciones relativas a cada fuente de

información, bien provenga de restricciones, de preferencias o de la combinación de ambas.

5.3.4 Influencia de la información morfológica

Restricciones morfológicas. La información morfológica ha demostrado ser una de las más relevantes como fuente de restricción, tanto cuando actúa de forma individual como cuando lo hace conjuntamente con el resto de las fuentes de información.

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	**27,27%**
Personales	53	12	**22,64%**
Demostr.	3	0	**0,00%**
Reflexivos	10	5	**50,00%**
	121	**32**	**26,45%**

Adición restricción Morfológica (gen. y núm.)

TOTAL	Anaf	OK	
Omitidos	55	20	**36,36%**
Personales	53	25	**47,17%**
Demostr.	3	2	**66,67%**
Reflexivos	10	7	**70,00%**
	121	**54**	**44,63%**

BASE de supresión: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**10000%**
Reflexivos	10	10	**100,00%**
	121	**69**	**58,68%**

Supresión restricción Morfológica

TOTAL	Anaf	OK	
Omitidos	55	17	**30,91%**
Personales	53	21	**39,62%**
Demostr.	3	1	**33,33%**
Reflexivos	10	10	**100,00%**
	121	**49**	**40,50%**

Cuadro 5.10. Adición y supresión de restricciones morfológicas en la evaluación

Como puede verse en el cuadro 5.10, al aplicarla de forma individual sobre el caso base, se produce un importante incremento en el porcentaje de éxito (+18,18 %), idéntico porcentaje al del decremento producido al eliminarla del conjunto total de restricciones.

No obstante, cabe mencionar el hecho de que al aplicar restricciones morfológicas de género y número, es posible que se elimine algún antecedente potencial. Tal es el caso de los ya mencionados nombres colectivos que no concuerdan necesariamente con el pronombre en su información morfológica de número, como ocurre

en el ejemplo (100) extraído del bloque L065 del corpus de evaluación:

(100) El espectáculo que se ofrecía al *trío*$_i$ holandés al rebasar la punta de La Guía era maravilloso. $\emptyset_i$ Estaban acostumbrados a ver mundo...

Sólo aplicando la condición morfosemántica de no correferencia definida en el método ERA se puede evitar la eliminación de *trío* (singular) como posible antecedente del pronombre omitido plural. Las ventajas e inconvenientes de este filtro morfosemántico serán discutidos más adelante en el apartado dedicado a la información semántica.

BASE de adición: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**71**	**58,68%**

Adición preferencia Morfológica (mismo núm.)

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	35	**66,04%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**72**	**59,50%**

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión preferencia Morfológica

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Cuadro 5.11. Adición y supresión de la preferencia morfológica de número en la evaluación

Preferencias morfológicas. La preferencia morfológica común aplicada de forma aislada tiene una relevancia muy débil dentro del proceso de resolución (ver cuadro 5.11). De hecho, si bien en su adición al caso base se percibe una ligera mejoría del resultado

final[6] (+0.82), puede verse como su eliminación del conjunto total de preferencias en la fase de supresión no altera el resultado final.

El objetivo de esta preferencia es complementar al filtro morfosemántico simulando el comportamiento de la restricción de género y número pero estableciendo un criterio más permisivo en el rasgo de número (no elimina, sólo prefiere).

Al tratarse de una preferencia común, los criterios sintácticos y semánticos suelen resolver la anáfora correctamente antes de su aplicación, por lo que su eficacia dentro del conjunto de preferencias parece estar unido a casos muy concretos. No obstante, no se ha detectado ningún ejemplo en el que esta preferencia provoque una solución incorrecta por lo que podríamos concluir que se trata de una preferencia de coste de aplicación muy bajo y que, si bien no origina un importante incremento en los resultados de resolución correcta, no entorpece la resolución y, por tanto, resulta útil en el conjunto global.

Combinación de restricciones y preferencias morfológicas. Del estudio de los resultados de la combinación de las restricciones y las preferencias morfológicas en la evaluación (ver cuadro 5.12) se extraen dos ideas principales.

Por un lado, la adición individual de restricciones y preferencias morfológicas proporciona los mismos resultados que la adición sólo de las restricciones, algo que corrobora la débil influencia de la preferencia de número de forma aislada.

Por otro lado, tanto la adición como la supresión de información morfológica muestran su positiva influencia (+18,18 %, −9,1 %) en el proceso de resolución global, con lo que se puede concluir que la morfología juega un papel importante en la resolución de la anáfora y que, además, su aplicación resulta de utilidad por sus buenos resultados y su bajo coste computacional.

[6] Esta mejoría, en realidad, es anecdótica y responde a un caso concreto relacionado con la restricción morfosemántica y que se comentará más adelante.

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	**27,27%**
Personales	53	12	**22,64%**
Demostr.	3	0	**0,00%**
Reflexivos	10	5	**50,00%**
	121	**32**	**26,45%**

Adición restr. y pref.. Morfológicas

TOTAL	Anaf	OK	
Omitidos	55	20	**36,36%**
Personales	53	25	**47,17%**
Demostr.	3	2	**66,67%**
Reflexivos	10	7	**70,00%**
	121	**54**	**44,63%**

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión restr. y pref. Morfológicas

TOTAL	Anaf	OK	
Omitidos	55	46	**83,64%**
Personales	53	41	**77,36%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**100**	**82,64%**

Cuadro 5.12. Adición y supresión de restricciones y preferencias morfológicas en la evaluación

5.3.5 Influencia de la información sintáctica

Restricciones sintácticas. Las restricciones sintácticas se fundamentan en teorías de rección que restringen su análisis a los componentes de una cláusula. Analizando los resultados obtenidos tras la adición y supresión de las restricciones sintácticas (ver cuadro 5.13) puede verse cómo, efectivamente, la influencia positiva de estas restricciones está asociada a aquellos casos en los que el antecedente se encuentra en la misma cláusula del pronombre.

En el caso de la adición (+7,43 %), las restricciones evitan que el método escoja antecedentes que, por estar en su misma cláusula, se encuentran más cerca del pronombre y, por tanto, se resuelven incorrectamente por el caso base de selección del más cercano, como ocurre en el ejemplo (101) extraído del bloque L065 en el que, al eliminar los candidatos *día* y *paliza* de su misma cláusula, el método resuelve el pronombre omitido correctamente.

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Adición restricciones Sintácticas

TOTAL	Anaf	OK	
Omitidos	55	16	29,09%
Personales	53	15	28,30%
Demostr.	3	0	0,00%
Reflexivos	10	10	100,00%
	121	41	33,88%

BASE de supresión: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	43,64%
Personales	53	34	64,15%
Demostr.	3	3	10000%
Reflexivos	10	10	100,00%
	121	69	58,68%

Supresión restricciones Sintácticas

TOTAL	Anaf	OK	
Omitidos	55	23	41,82%
Personales	53	31	58,49%
Demostr.	3	3	100,00%
Reflexivos	10	7	70,00%
	121	64	52,89%

Cuadro 5.13. Adición y supresión de restricciones sintácticas en la evaluación

(101) Al día siguiente *Frans*$_i$ se levantó temprano. Después de la paliza del día anterior, $\emptyset_i$ había convenido muy estratégica y cortésmente ...

Por otro lado, la adición de las restricciones sintácticas cubre varios casos de dislocación que no sería correctamente resuelta por un sistema que no hiciera uso de la información proporcionada por los papeles sintácticos. Un ejemplo muy común de estos casos es el de los dobles clíticos, como el del ejemplo (102) extraído de L065, donde una restricción basada en conocimiento sintáctico limitado elegiría un SN previo y nunca uno posterior.

(102) El hotel Reconquista resultaba muy agradable, aunque ciertamente no demasiado democrático por los precios, pero esto no *les*$_i$ preocupaba a los tres *holandeses*$_i$ ...

Por otra parte, uno de los casos más evidentes de la influencia de las restricciones sintácticas que revela la evaluación se percibe en los pronombres reflexivos, cuyo índice de resolución asciende al 100 % cuando se aplica el conocimiento sintáctico enriquecido. Las restricciones sintácticas basadas en el necesario papel de sujeto del antecedente de un reflexivo cubre la totalidad de los

casos, incluidos los de dislocación en los que el sujeto se encuentra después del verbo. Los ejemplos (103) y (104), extraídos de los bloques L065 y L009 respectivamente, muestran resoluciones de reflexivos[7] con y sin dislocación del sujeto.

(103) Los *tres*$_i$ en la *cubierta*$_j$ *se*$_i$ abrazaron...

(104) ¿por qué las mujeres al conducir, *se*$_i$ preguntaba *Barnes*$_i$, mueven todo el cuerpo hacia un lado o hacia el otro cuando toman las curvas?

La influencia de la supresión de las restricciones sintácticas sobre la aplicación global de todas las restricciones muestra resultados similares (aunque algo inferiores) a los proporcionados por su adición (−5,79 %). En este caso se percibe un decremento menos brusco en la resolución de reflexivos, reforzada por el resto de las fuentes de conocimiento, pero queda patente, en los datos obtenidos, la positiva influencia de esta restricción basada en los fundamentos antes mencionados.

Preferencias sintácticas. La adición de preferencias sintácticas (ver cuadro 5.14) es, junto con la de la semántica combinada, la que proporciona mejores resultados de forma aislada (+28,1 %) una vez aplicadas las restricciones y eliminados todos los candidatos potencialmente incompatibles.

Este balance tan positivo de la influencia de las preferencias sintácticas está lógicamente relacionado con la información relativa al papel sintáctico que proporcionan tanto los candidatos como el pronombre anafórico. Esta información refuerza las preferencias propuestas reduciendo la lista de candidatos a los más relevantes sintácticamente (sujetos, mismo papel sintáctico, ...) y seleccionando el correcto en gran parte de las ocasiones.

[7] El ejemplo (103) es en realidad un caso de pronombre recíproco. Si bien a lo largo de este trabajo se ha tratado la distinción existente entre los pronombres reflexivos y recíprocos, en el método propuesto no se hacen distinciones entre ambos por realizar un tratamiento computacional común en el que ambos se agrupan bajo el denominador común de reflexivo.

BASE de adición: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	43,64%
Personales	53	34	64,15%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	71	58,68%

Adición preferencias Sintácticas

TOTAL	Anaf	OK	
Omitidos	55	48	87,27%
Personales	53	44	83,02%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	105	86,78%

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Supresión preferencias Sintácticas

TOTAL	Anaf	OK	
Omitidos	55	46	83,64%
Personales	53	42	79,25%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	101	83,47%

Cuadro 5.14. Adición y supresión de preferencias sintácticas en la evaluación

Si bien este incremento en la correcta resolución es general, parece afectar más a los pronombres omitidos que a los personales. Una razón para esto podría ser el efecto positivo que tiene en el corpus de evaluación la preferencia exclusiva de los pronombres omitidos (se prefieren los candidatos con papel de sujeto). El ejemplo (105) corrobora esta afirmación, dándole una mayor relevancia al SN *la cocina española* por ser el único antecedente con función de sujeto.

(105) La *cocina*$_i$ española les gustaba también de lo lindo, como se demostró después en el comedor. Bueno, $\emptyset_i$ les gustaba a los holandeses, a los ingleses, a los alemanes, a los americanos y a los marcianos.

Un punto interesante, digno de comentario, es el contraste existente entre el elevado incremento en el porcentaje de éxito al añadir aisladamente las preferencias sintácticas (+28,1 %) y el bajo decremento del éxito en la resolución cuando se suprimen dichas preferencias (−8,27 %). Esto indica que muchos de los casos correctamente resueltos por las preferencias sintácticas quedan cubiertos por el resto de las fuentes de conocimiento aplicadas en la resolución. En el caso del ejemplo (105), debido a la informa-

ción basada patrones de compatibilidad semántica se preferiría el antecedente con núcleo *cocina* por haber aparecido previamente el patrón *gustar-cocina* con las mismas referencias semánticas que en el caso de la anáfora.

Combinación de restricciones y preferencias sintácticas. A lo largo de los dos apartados anteriores se ha detallado la influencia de las restricciones, por un lado, y de las preferencias, por otro, estudiada a partir de los resultados de la evaluación. Si esta influencia era ya positiva aplicando restricciones y preferencias por separado, la adición y la supresión conjunta de la sintaxis refleja un comportamiento todavía mejor (ver cuadro 5.15).

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Adición restr. y pref.. Sintácticas

TOTAL	Anaf	OK	
Omitidos	55	41	74,55%
Personales	53	31	58,49%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	85	70,25%

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Supresión restr. y pref. Sintácticas

TOTAL	Anaf	OK	
Omitidos	55	44	80,00%
Personales	53	40	75,47%
Demostr.	3	3	100,00%
Reflexivos	10	5	50,00%
	121	92	76,03%

Cuadro 5.15. Adición y supresión de restricciones y preferencias sintácticas en la evaluación

Así, la combinación de restricciones y preferencias en la adición al caso base de selección del candidato más cercano proporciona un incremento en el porcentaje de éxito enormemente satisfactorio (+43,8 %) fruto de la combinación de restricciones y preferencias cuya repercusión en los resultados ha sido muy positiva por separado. Por otro lado, aunque su supresión no genera un descenso

comparable (−15,71 %), es mayor que el provocado por la ausencia individual de cualquier otra fuente de información.

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Adición restr. y pref.. Sintácticas combinadas

TOTAL	Anaf	OK	
Omitidos	55	41	74,55%
Personales	53	33	62,26%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	87	71,90%

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Supresión restr. y pref. Sintácticas combinadas

TOTAL	Anaf	OK	
Omitidos	55	43	78,18%
Personales	53	40	75,47%
Demostr.	3	3	100,00%
Reflexivos	10	5	50,00%
	121	91	75,21%

Cuadro 5.16. Adición y supresión de restricciones y preferencias sintácticas combinadas en la evaluación

Además, si la información procedente de restricciones y preferencias sintácticas se combina con la que aportan las restricciones sintáctico-semánticas (ver cuadro 5.16), los resultados mejoran aún más (+45,45 %,−16,53 %). Podríamos concluir, por tanto, que el conocimiento sintáctico enriquecido con los papeles sintácticos de los componentes oracionales, es una de las fuentes de información más valiosas aplicada a la resolución de la anáfora, especialmente al combinarla con la información semántica.

5.3.6 Influencia de la información semántica

En este apartado se tratarán tanto las restricciones y las preferencias basadas en los patrones semánticos (compatibilidad e incompatibilidad) como las que combinan la semántica con otras fuentes de conocimiento (morfosemánticas, sintactico-semánticas

y semántico-estructurales)[8].

Restricciones semánticas. Los datos del cuadro 5.17 muestran los resultados obtenidos en la incorporación y la eliminación de las restricciones semánticas basadas en el uso de patrones de incompatibilidad semántica.

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	**27,27%**
Personales	53	12	**22,64%**
Demostr.	3	0	**0,00%**
Reflexivos	10	5	**50,00%**
	121	**32**	**26,45%**

Adición restricciones Semánticas

TOTAL	Anaf	OK	
Omitidos	55	16	**29,09%**
Personales	53	14	**26,42%**
Demostr.	3	1	**33,33%**
Reflexivos	10	5	**50,00%**
	121	**36**	**29,75%**

BASE de supresión: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**10000%**
Reflexivos	10	10	**100,00%**
	121	**69**	**58,68%**

Supresión restricciones Semánticas

TOTAL	Anaf	OK	
Omitidos	55	21	**38,18%**
Personales	53	32	**60,38%**
Demostr.	3	2	**66,67%**
Reflexivos	10	10	**100,00%**
	121	**65**	**53,72%**

Cuadro 5.17. Adición y supresión de restricciones semánticas en la evaluación

Aunque la influencia de las restricciones semánticas de forma independiente puede parecer algo débil (+3,3 %, −4,96 %), queda patente en los datos recogidos que su aplicación es extremadamente eficaz y resulta de clara utilidad en casos como el del ejemplo (106), tomado del bloque de evaluación L065, donde todos los candidatos (*dedo*, *hoteles*, *estrellas*, *guías*, *turismo*, *dibujo*, *edificio*, *guía*, *acierto*, *previsiones*, *mérito* y *bolsillo*) excepto los que representan personas (*Van Steen* y *Lisbeth*, el antecedente correcto) son eliminados por no corresponder con los rasgos de 'humano' que exige el sujeto del verbo saber en su primer sentido de WordNet definido a través del patrón $DEBE(\text{saber\#1}, \text{Human}, S)$.

[8] En este punto es necesario recordar las razones que han llevado a enunciar los distintos tipos de restricciones y preferencias dentro del marco de una fuente de información concreta. Para una explicación detallada, ver nota 28 (pág. 122).

(106) *Lisbeth*$_i$ sabía de sobra que era fácil acertar cuando se ponía el dedo sobre los hoteles de cinco estrellas de las guías de turismo españolas, o sobre los que venían precedidos con el pequeño dibujo de un edificio rojo en la guía francesa de Michelín; de modo que en este sentido el acierto de sus previsiones no tenía mucho mérito y había que atribuirlo más justamente al potente bolsillo de Van Steen, pero *ella*$_i$ se calló zorrunamente...

A pesar de eliminar todos los candidatos semánticamente incompatibles, la selección del más cercano en la aplicación individual de las restricciones semánticas resuelve la anáfora incorrectamente.

Un ejemplo de resolución inmediata es el mostrado en (107), usado en capítulos anteriores y tomado del bloque E001 del corpus de evaluación.

(107) El mono subió al árbol a coger un *plátano*$_i$. $\emptyset_i$ Maduraba al sol.

Este es un ejemplo claro de uso de información semántica específica. Para el verbo madurar WordNet proporciona cuatro sentidos distintos:

- *madurar#1, envejecer#2*: "She aged gracefully". 2ndOrderEntity 30 Dynamic Quantity SituationType
- *madurar#2*: "The plums ripen in July". 2ndOrderEntity 30 Dynamic Quantity SituationType
- *sazonar#1, madurar#3*: "The sun ripens the fruit". 2ndOrderEntity 30 Cause Dynamic SituationType
- *hacerse#1, madurar#4, crecer#4*: "He matured fast". 2ndOrderEntity 30 Dynamic Quantity SituationType

De ellos, sólo el segundo sentido está asociado al madurar de frutas, por lo que es posible definir el patrón:

$$DEBE(\text{Comestible}, \text{madurar}\#2, S)$$

Además, de los sentidos 1 y 2, que definen respectivamente el concepto de envejecimiento y el acto de hacer que algo madure, se pueden generar los siguientes patrones:

$$DEBE(\mathsf{Living}, \mathsf{madurar}\#1, S)$$
$$DEBE(\mathsf{Comestible}, \mathsf{madurar}\#3, D)$$

Aplicando la regla de incompatibilidad sobre el primero de los patrones, el método elimina *mono* y *árbol* por no contener ninguno de los dos el rasgo de 'comestible'.

Se deduce de todo esto que un índice claro de relevancia de estas restricciones semánticas lo proporciona el conjunto de patrones de incompatibilidad semántica definido en el método. En el momento de la evaluación se contaba con un conjunto total de 66 patrones de incompatibilidad, que incluían 24 formas verbales con un total de 54 conceptos (*synsets*) diferentes. El cuadro 5.18 muestra una lista con las definiciones de estos patrones.

Estos patrones recogen algunos de los verbos contenidos en el corpus de evaluación. La ampliación de este conjunto de patrones de incompatibilidad contribuiría positivamente a la mejora de la influencia de las restricciones semánticas en el proceso de resolución.

Uno de los problemas encontrados a la hora ampliar el conjunto de patrones de incompatibilidad es precisamente el conjunto de verbos que proporcionan poca o nula información semántica (*hacer*, *haber*, *tener*, *poder*, *pasar*, los copulativos *ser* y *estar*, ...), así como los verbos que no están contenidos en WordNet, bien porque la forma verbal no aparece o bien porque el sentido que toma no está dentro de los contenidos en WordNet para ese verbo. Algunos casos extraídos del corpus de evaluación resultan bastante significativos en lo referente a las carencias de WordNet en este sentido y quedan representados por la ausencia de verbos tan comunes como *comprar*, *exclamar*, *resultar*, *desesperar*, *brindar* o *atracar* así como la ausencia de acepciones de verbos como *tomar* (una curva), *adelantar* (un reloj) o *intervenir* (en una conversación).

$DEBE$(abrumar#1, Human, D)	$DEBE$(hablar#3, Human, S)
$DEBE$(abrumar#2, Human, D)	$DEBE$(hablar#4, Human, S)
$DEBE$(abrumar#3, Human, D)	$DEBE$(hablar#5, Human, S)
$DEBE$(apagar#3, Artifact, D)	$DEBE$(hablar#6, Human, S)
NO(apagar#3, Location, D)	$DEBE$(hablar#7, Human, S)
NO(apagar#3, Place, D)	$DEBE$(madurar#1, Living, S)
NO(apagar#3, Occupation, D)	$DEBE$(madurar#2, Comestible, S)
NO(apagar#3, Comestible, D)	$DEBE$(madurar#3, Comestible, D)
NO(apagar#3, Building, D)	$DEBE$(ojear#2, Human, S)
$DEBE$(apetecer#1, Human, I)	$DEBE$(preguntar#1, Human, S)
$DEBE$(aterrar#1, Human, I)	$DEBE$(preguntar#3, Human, S)
$DEBE$(bañar#5, Living, S)	$DEBE$(preocupar#1, Human, I)
$DEBE$(bañar#5, Living, D)	$DEBE$(preocupar#2, Human, I)
NO(bañar#5, Plant, S)	$DEBE$(preocupar#3, Human, I)
NO(bañar#5, Plant, D)	$DEBE$(preocupar#4, Human, I)
$DEBE$(callarse#1, Human, S)	NO(pronunciar#1, Living, D)
$DEBE$(comer#1, Comestible, D)	NO(pronunciar#2, Living, D)
$DEBE$(comer#2, Comestible, D)	NO(pronunciar#3, Living, D)
$DEBE$(comer#3, Comestible, D)	$DEBE$(saber#1, Human, S)
$DEBE$(comer#4, Comestible, D)	$DEBE$(saber#2, Human, S)
$DEBE$(decidir#1, Human, S)	$DEBE$(sentir#1, Human, S)
$DEBE$(decir#3, Human, S)	$DEBE$(sentir#2, Human, S)
$DEBE$(desayunar#1, Living, S)	$DEBE$(sentir#3, Human, S)
NO(desayunar#1, Planta, S)	$DEBE$(sentir#4, Human, S)
$DEBE$(entender#1, Living, S)	$DEBE$(sentir#5, Human, S)
NO(entender#1, Plant, S)	$DEBE$(sentir#6, Human, S)
$DEBE$(entender#2, Living, S)	$DEBE$(sentir#7, Human, S)
NO(entender#2, Plant, S)	$DEBE$(sentir#8, Human, S)
NO(entender#2, Human, D)	$DEBE$(ver#5, Living, S)
$DEBE$(escuchar#1, Human, S)	NO(vivir#1, Artifact, S)
$DEBE$(fastidiar#5, Human, I)	NO(vivir#2, Artifact, S)
$DEBE$(gustar#1, Human, I)	NO(vivir#3, Artifact, S)
$DEBE$(hablar#1, Human, S)	NO(vivir#4, Artifact, S)

Cuadro 5.18. Patrones de incompatibilidad semántica usados en la evaluación del método ERA

Otro problema de la aplicación de estos patrones de incompatibilidad es la ausencia en WordNet de los sustantivos que son núcleos de los SN antecedentes y sin cuyo sentido no es posible comprobar la potencial incompatibilidad.

Restricciones morfosemánticas. Los resultados de la evaluación (ver cuadro 5.19) muestran, en lo referente a la adición y la eliminación de restricciones morfosemánticas, un comporta-

miento muy similar al de las restricciones morfológicas (+17,35 %, −18,18 %).

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Adición restricciones Morfosemánticas

TOTAL	Anaf	OK	
Omitidos	55	20	36,36%
Personales	53	24	45,28%
Demostr.	3	2	66,67%
Reflexivos	10	7	70,00%
	121	53	43,80%

BASE de supresión: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	43,64%
Personales	53	34	64,15%
Demostr.	3	3	10000%
Reflexivos	10	10	100,00%
	121	69	58,68%

Supresión restricciones Morfosemánticas

TOTAL	Anaf	OK	
Omitidos	55	17	30,91%
Personales	53	21	39,62%
Demostr.	3	1	33,33%
Reflexivos	10	10	100,00%
	121	49	40,50%

Cuadro 5.19. Adición y supresión de restricciones morfosemánticas en la evaluación

En realidad, las condiciones morfosemánticas de no correferencia son menos restrictivas que las morfológicas y eso se refleja positiva y negativamente en la evaluación.

Por un lado, garantizan que, si un antecedente no concuerda en número con el pronombre anafórico, no será eliminado si dicho antecedente tiene el rasgo semántico de 'grupo'. De esta manera se evita la eliminación de candidatos que son potencialmente antecedentes del pronombre, como ocurre en el ejemplo (108), extraído del bloque E001, en el que el SN de núcleo *armada* sería eliminado directamente por las restricciones de carácter puramente morfológico.

(108) La *armada*$_i$ necesita jóvenes con ambición. $\emptyset_i$ Te ofrecen una especialización laboral y un buen sueldo.

El método, sin embargo, al aplicar únicamente información morfosemántica, no resuelve este ejemplo correctamente ya que escoge *jóvenes* como el antecedente correcto por ser el más cer-

cano a la anáfora. No obstante, ejemplos como el de (109) se resolverían directamente con el uso de esta fuente de información.

(109) La *policía*$_i$ vela por su seguridad. $\emptyset_i$ Están siempre alerta.

Por otro lado, el carácter menos restrictivo de las condiciones morfosemánticas de no correferencia plantea algunos inconvenientes como el del ejemplo (110) extraído del bloque L065 del corpus y que es el que marca la diferencia de resultados entre la aplicación de restricciones morfológicas y morfosemánticas.

(110) ...porque los *navegantes*$_i$ estaban aburridos de utilizar la piscina en su casa de La Haya. *Les*$_i$ encantaba la marcada salinidad del agua, y lo fácil que resultaba flotar.

En este ejemplo, todos los candidatos iniciales del pronombre *les* excepto su antecedente deberían ser eliminados tanto por las restricciones morfológicas como por las sintácticas. Sin embargo, la palabra *casa* etiquetada con su primer sentido tiene en WordNet la entrada

- *domicilio#1, habitación#3, hogar#1, morada#1, vivienda#3, casa#1*: a physical structure (e.g., a house) that someone is living in; "he built a modest dwelling near the pond"; "they raise money to provide homes for the homeless"03 06 1stOrderEntity Artifact Building Form Function Group Object Origin

de la que se extrae la lista de conceptos ontológicos:

$Ont_{(casa\#1)}$ = [Artifact, Building, Form, Function, Group, Object, Origin]

que contiene el rasgo de 'grupo' y que permanecerá por ello en el conjunto de candidatos posibles. Este caso provoca un fallo del sistema al aplicar la morfosemántica de forma independiente y elegir *casa* como antecedente por ser el candidato más cercano al pronombre. Este error puede ser atribuido más al uso que WordNet hace del concepto de grupo para determinados nombres[9] que

[9] Si bien el rasgo de 'grupo' está asociado en WordNet a sustantivos que potencialmente pueden formar grupos (*casa*, *plátano*, *árbol*, ...) también lo está a sustantivos que lo son en sí mismos (*pueblo*, *compañía*, *policía*, ...), con lo que

a un fallo del propio proceso de resolución.

Restricciones sintáctico-semánticas. Uno de los aspectos más destacables de estas restricciones es la eficacia de su aplicación. Dado que aplican reglas de restricción muy concretas que actúan sobre pronombres específicos, generan, tal y como puede verse en los resultados de su adición (cuadro 5.20), un incremento directo de un 9,44 % sobre la resolución de algunos pronombres personales sin alterar el comportamiento en el resto. Esto produce una mejora en los resultados globales (+4,13 %) cuando se incorpora de forma aislada el caso base de selección del candidato más cercano.

BASE de adición: el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Adición restricciones Sintáctico-semánticas

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	17	32,08%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	37	30,58%

BASE de supresión: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	43,64%
Personales	53	34	64,15%
Demostr.	3	3	10000%
Reflexivos	10	10	100,00%
	121	69	58,68%

Supresión restricciones Sintáctico-semánticas

TOTAL	Anaf	OK	
Omitidos	55	24	43,64%
Personales	53	32	60,38%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	69	57,02%

Cuadro 5.20. Adición y supresión de restricciones sintáctico-semánticas en la evaluación

Un ejemplo de mejora en la adición se muestra en (111), extraído del bloque L009, donde los antecedentes *colmo*, *claridad* y *coherencia* son eliminados por no poseer el rasgo de 'animado' necesario para correferir con el pronombre personal de sujeto *ellos*.

el mismo rasgo ontológico se usa para conceptos semánticos algo diferentes lo que contribuye a este tipo de errores.

(111) ...mientras que los *hombres*$_i$ aparecían como el más luminoso colmo de la claridad y la coherencia. Pues bien, de eso nada: *ellos*$_i$ son desconcertantes calamitosos y rarísimos.

Al igual que lo que ocurría en el caso de la información sintáctica, la supresión de las restricciones sintáctico-semánticas del conjunto total de restricciones ofrece una influencia algo más débil (−1,66 %). Esto es debido a que el resto de las restricciones cubren la mayoría de los casos que las sintáctico-semánticas resuelven correctamente. Algunas excepciones en este sentido son ejemplos como (112) y (113) que sólo pueden ser resueltas correctamente por este tipo de condiciones de no correferencia dentro del conjunto de restricciones del método.

(112) La televisión está encendida cuando *Luisa*$_i$ llega a la cocina. *Ella*$_i$ la apaga cuando se acuesta.

(113) *Luis*$_i$ ganó el premio al mejor cortometraje. *Le*$_i$ vi muy contento.

En el primer caso, el pronombre personal de sujeto obliga a su antecedente a tener un rasgo de 'animado', mientras que en el segundo, el pronombre de objeto directo *le* obliga a su antecedente a ser 'humano'. En ambos casos, todos los candidatos, excepto el antecedente, serán eliminados a través de las restricciones sintáctico-semánticas.

Preferencias semánticas. Las preferencias semánticas tienen una doble función. Por un lado valoran positivamente aquellos antecedentes que no son de tiempo, dirección, cantidad ni tipo abstracto y, por otro lado, establecen un grado de compatibilidad semántica entre el antecedente y el pronombre a través de su verbo.

Los resultados de la evaluación de la adición y supresión de las preferencias semánticas en el proceso de resolución (cuadro 5.21) revelan dos datos muy interesantes.

BASE de adición: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**71**	**58,68%**

Adición preferencias Semánticas

TOTAL	Anaf	OK	
Omitidos	55	27	**49,09%**
Personales	53	38	**71,70%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**78**	**64,46%**

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión preferencias Semánticas

TOTAL	Anaf	OK	
Omitidos	55	53	**96,36%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**112**	**92,56%**

Cuadro 5.21. Adición y supresión de preferencias semánticas en la evaluación

Por un lado, su adición individual proporciona un incremento (+5,78 %) sobre la aplicación base de todas las restricciones. Este dato, aunque inferior al de la adición de las preferencias sintácticas, es algo superior al de las estructurales o las morfológicas y hace patente la positiva influencia del uso de los patrones de compatibilidad semántica en el proceso de resolución.

A pesar de este incremento observado, se puede hablar una vez más de un caso aislado que, al aplicar las preferencias semánticas, hace descender el índice de éxito del sistema. Este caso se muestra en el ejemplo (114), extraído del bloque L009 del corpus de evaluación.

(114) Siempre creí que a lo que yo aspiraba era a la comunicación perfecta con un hombre, o, mejor dicho, con el hombre, con ese príncipe azul de los sueños de infancia, un ser que sabría adivinarme hasta en los más menudos pliegues interiores. Ahora he aprendido no sólo que esa *fusión*$_i$ es imposible, sino además que $\emptyset_i$ es probablemente indeseable.

Al aplicar las preferencias semánticas sobre los candidatos de la anáfora generada por el pronombre omitido, los SN con núcleos *hombre* y *príncipe* reciben una mayor ponderación a través del

patrón semántico que les asocia con el verbo *ser* mientras que, cuando no se aplican estas preferencias semánticas, se escoge *fusión* por razones de cercanía y compatibilidad de papel sintáctico.

Este ejemplo aislado, sirve para reflexionar sobre el uso especial que tienen determinados verbos, como es el caso del *ser* copulativo. Es evidente que este verbo no proporciona ninguna clase de información semántica y es precisamente el atributo el que añade dicha información[10].

Por otro lado, la supresión de la información semántica no proporciona resultados satisfactorios, debido probablemente al elevado índice de resolución que proporciona en el corpus el conjunto global de restricciones y preferencias. De hecho, el error que el método comete en el caso anterior consigue incluso un leve incremento en los resultados.

Preferencias semánticas combinadas. Para comprobar la influencia de la semántica en términos generales, se han realizado experimentos que aunan la información semántica procedente de las preferencias definidas como puramente semánticas y la procedente de las preferencias de carácter estructural que usan la semántica para su aplicación (preferencias semántico-estructurales). Los resultados de la adición y la supresión de estas preferencias combinadas se muestran en el cuadro 5.22.

Estos resultados (+19,83 %,−4,14 %) revelan una clara influencia de la aplicación conjunta de las preferencias semánticas combinadas en el proceso de resolución. Se observa que la combinación de ambos conjuntos de preferencias mejoran los resultados en un porcentaje mayor que la suma de las mejoras parciales de cada conjunto, siendo de nuevo más relevante la adición que la supresión de la semántica combinada en la evaluación global.

Combinación de restricciones y preferencias semánticas. Haciendo un balance global de la influencia de la semántica en la aplicación del método ERA, en el cuadro 5.23 puede comprobarse

[10] En la sección 7.2 se aborda ésta y otras líneas de mejora en la ampliación del método ERA.

BASE de adición:
todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**71**	**58,68%**

Adición preferencias
Semánticas combinadas

TOTAL	Anaf	OK	
Omitidos	55	39	**70,91%**
Personales	53	43	**81,13%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**95**	**78,51%**

BASE de supresión:
todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión preferencias
Semánticas combinadas

TOTAL	Anaf	OK	
Omitidos	55	50	**90,91%**
Personales	53	43	**81,13%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**106**	**87,60%**

Cuadro 5.22. Adición y supresión de preferencias semánticas combinadas en la evaluación

el resultado de la adición y la supresión del conjunto global de restricciones y preferencias de carácter semántico, basado fundamentalmente en los patrones de compatibilidad y de incompatibilidad semántica.

BASE de adición:
el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	**27,27%**
Personales	53	12	**22,64%**
Demostr.	3	0	**0,00%**
Reflexivos	10	5	**50,00%**
	121	**32**	**26,45%**

Adición restr. y pref..
Semánticas

TOTAL	Anaf	OK	
Omitidos	55	17	**30,91%**
Personales	53	15	**28,30%**
Demostr.	3	1	**33,33%**
Reflexivos	10	5	**50,00%**
	121	**38**	**31,40%**

BASE de supresión:
todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión restr. y pref.
Semánticas

TOTAL	Anaf	OK	
Omitidos	55	51	**92,73%**
Personales	53	45	**84,91%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**109**	**90,08%**

Cuadro 5.23. Adición de restricciones y preferencias semánticas en la evaluación

La observación de estos datos muestra una débil pero positiva influencia de esta información en la selección de antecedentes correctos que, como en otros muchos casos, se asocia más a la adición de la restricciones y preferencias (+4,95 %) que a su supresión (−1,66 %). El grado de influencia de la fuente semántica aislada se corresponde con los datos obtenidos de la aplicación parcial de restricciones y preferencias.

Si combinamos además la semántica basada en patrones con el resto de fuentes de información que hacen uso de ella (morfo-semánticas, sintáctico-semánticas, semántico-estructurales) la influencia demostrada (ver cuadro 5.24) es mucho más que satisfactoria.

BASE de adición:
el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Adición restr. y pref..
Semánticas combinadas

TOTAL	Anaf	OK	
Omitidos	55	39	70,91%
Personales	53	39	73,58%
Demostr.	3	3	100,00%
Reflexivos	10	4	40,00%
	121	85	70,25%

BASE de supresión:
todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Supresión restr. y pref.
Semánticas combinadas

TOTAL	Anaf	OK	
Omitidos	55	48	87,27%
Personales	53	37	69,81%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	98	80,99%

Cuadro 5.24. Adición y supresión de restricciones y preferencias semánticas combinadas en la evaluación

Como puede verse, el índice de resolución proporcionado por la adición de estas fuentes de información semántica combinada es del +43,8 % sobre el método de selección del más cercano, mientras que su supresión supone un decremento en la resolución de un −10,75 %.

La semántica es, por tanto, una fuente de información que incorpora criterios adicionales y que mejora los resultados de reso-

lución anafórica, especialmente cuando se combina con fuentes de información adicionales como la sintáctica o la estructural.

5.3.7 Influencia de la información estructural

Preferencias estructurales. Las preferencias estructurales tienen una interesante relevancia según los resultados de la evaluación (ver cuadro 5.25).

BASE de adición: todas las restricciones

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**71**	**58,68%**

Adición preferencias Estructurales

TOTAL	Anaf	OK	
Omitidos	55	28	**50,91%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**75**	**61,98%**

BASE de supresión: todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión preferencias Estructurales

TOTAL	Anaf	OK	
Omitidos	55	46	**83,64%**
Personales	53	45	**84,91%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**104**	**85,95%**

Cuadro 5.25. Adición y supresión de preferencias estructurales en la evaluación

El carácter positivo de la incorporación y la supresión de la información estructural (+3,3 %,−5,79 %) refuerza las teorías basadas en el reducido espacio de búsqueda de la solución anafórica, que para esta evaluación ha estado compuesto de la oración en la que aparece el pronombre y la oración anterior. Este espacio de búsqueda del antecedente cubre más del 99 % de las anáforas del corpus.

Un ejemplo de aplicación correcta de estas preferencias es el mostrado en (115), extraído del bloque **L065**, donde, al no aplicar preferencias estructurales, el método selecciona el candidato *hombros* por criterios eminentemente sintácticos (concordancia en

papel sintáctico). Sin embargo, la aplicación de preferencias estructurales señalarían a *relojes* como el candidato más adecuado.

(115) Había prescindido de la pieza superior del bañador, porque le fastidiaba la marca blanca que dejaban los tirantes sobre la piel, y que luego le impedía lucir los trajes de noche que dejaban al descubierto los hombros desnudos. Los *relojes*$_i$ de los navegantes marcaban todavía las diez y media, y Frans recomendó a Lisbeth adelantar*los*$_i$ dos horas...

La estructural es, por tanto, una fuente de información fundamental para establecer, no sólo las preferencias asociadas a los candidatos más próximos estructuralmente hablando, sino para determinar el espacio de búsqueda de la solución de un pronombre.

5.3.8 La semántica y los papeles sintácticos

A lo largo de los apartados anteriores se han tratado de manera independiente los resultados del uso de información semántica y sintáctica en la evaluación. Sin embargo, se han hecho varias referencias a las ventajas que proporciona la combinación de ambas fuentes de conocimiento. Siguiendo con el objetivo de este trabajo en lo referente a la valoración de la influencia de la información semántica y la información basada en papeles sintácticos sobre el proceso de resolución de la anáfora, se ha llevado a cabo una prueba relativa a la eliminación de toda restricción o preferencia que integre cualquiera de las dos fuentes mencionadas (ver cuadro 5.26).

Como se puede observar, la relevancia de estas fuentes de conocimiento no sólo es muy elevada, sino que, conjuntamente, proporcionan resultados mejores (+59,5 %,−42,98 %) que la suma de las mejoras obtenidas de forma individual. Esto abunda en la importancia de ambas fuentes de información, especialmente cuando se combinan entre sí.

BASE de adición:
el más cercano

TOTAL	Anaf	OK	
Omitidos	55	15	**27,27%**
Personales	53	12	**22,64%**
Demostr.	3	0	**0,00%**
Reflexivos	10	5	**50,00%**
	121	**32**	**26,45%**

Adición restr. y pref..
Sint. y Sem. combinadas

TOTAL	Anaf	OK	
Omitidos	55	46	**83,64%**
Personales	53	45	**84,91%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**104**	**85,95%**

BASE de supresión:
todas las restr. y pref..

TOTAL	Anaf	OK	
Omitidos	55	52	**94,55%**
Personales	53	46	**86,79%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**111**	**91,74%**

Supresión restr. y pref.
Sint. y Sem. combinadas

TOTAL	Anaf	OK	
Omitidos	55	25	**45,45%**
Personales	53	25	**47,17%**
Demostr.	3	2	**66,67%**
Reflexivos	10	7	**70,00%**
	121	**59**	**48,76%**

Cuadro 5.26. Adición y supresión de restricciones y preferencias sintácticas y semánticas combinadas en la evaluación

5.3.9 Influencia de la adquisición de patrones de compatibilidad

En relación al cuarto y último de los experimentos realizados con el método ERA es conveniente destacar algunos de los aspectos que se derivan de estos resultados (ver cuadro 5.27).

Por un lado, puede verse cómo el sistema prácticamente no varía su comportamiento por la adquisición previa de patrones de compatibilidad semántica cuando se aplican todas las restricciones y preferencias (+0,82 %). Esto parece deberse al hecho de que los resultados obtenidos de la aplicación conjunta de restricciones y preferencias son muy elevados y hacen muy difícil la mejora global del sistema. Por otro lado, esta idea se refuerza por el hecho de que, al aplicar la resolución basada únicamente en preferencias semánticas, el incremento en la resolución es notable, especialmente cuando la adquisición se ha realizado sobre todos los bloques del corpus (+20,66 %), algo que si bien es de esperar ya que se están adquiriendo los patrones que después intervendrán en la resolución, demuestra que un contexto más amplio y un corpus más extenso contribuirían a una mejora en los resultados de aplicación de estos patrones.

BASE 1: todas las restr. y pref.

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Adquisición: dos bloques
Resolución: el tercero

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Adquisición: todos
Resolución: todos

TOTAL	Anaf	OK	
Omitidos	55	51	92,73%
Personales	53	48	90,57%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	112	92,56%

BASE 2: todas las restr. y sólo pref. semánticas

TOTAL	Anaf	OK	
Omitidos	55	27	49,09%
Personales	53	36	67,92%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	76	62,81%

Adquisición: dos bloques
Resolución: el tercero

TOTAL	Anaf	OK	
Omitidos	55	28	50,91%
Personales	53	37	69,81%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	78	64,46%

Adquisición: todos
Resolución: todos

TOTAL	Anaf	OK	
Omitidos	55	43	78,18%
Personales	53	45	84,91%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	101	83,47%

Cuadro 5.27. Experimento de adquisición previa de patrones en la evaluación

5.4 Conclusiones

Tras el estudio exhaustivo y la interpretación de los resultados proporcionados por la aplicación del método ERA realizada sobre el corpus de evaluación, la relevancia de la incorporación de conocimiento basado en papeles sintácticos y de conocimiento semántico en el proceso de la resolución de la anáfora parece evidente. El cuadro 5.28 muestra un resumen de la influencia de cada fuente de conocimiento tal y como se ha expuesto en este capítulo y referida a la aplicación conjunta de restricciones y preferencias.

Fuente	Adición	Supresión
Morfológica	+18,8 %	−9,1 %
Sintáctica	+43,8 %	−15,71 %
Semántica	+4,95 %	−1,66 %
Estructural	+3,3 %	−5,79 %
Sintáctica combinada	+45,45 %	−16,53 %
Semántica combinada	+43,8 %	−10,75 %
Semántica + Sintáctica	+59,5 %	−42,98 %

Cuadro 5.28. Resumen de resultados sobre la influencia de cada fuente de información en el método ERA

En lo referente a la información sintáctica enriquecida, ésta proporciona por sí sola una tasa de resolución del 43,8 %, porcentaje que asciende al 45,45 % con la incorporación de restricciones sintáctico-semánticas (información sintáctica combinada).

Por otra parte, la información basada en patrones semánticos tiene una influencia positiva en el proceso de resolución (+4,95 %), relevancia que, sin embargo, es equiparable a la de la información sintáctica (+43,8 %) cuando se aplican todas las fuentes de carácter semántico (información semántica combinada).

Por tanto, la relevancia de la semántica queda reforzada en su combinación con otras fuentes de información como la sintáctica o la estructural (+59,5 %). En este sentido, resulta complejo trazar la línea que separa la sintaxis de la semántica en este trabajo. Es evidente que ambas fuentes de información van muy unidas y cooperan en la mejora de los resultados del proceso de resolución de la anáfora.

Se ha comprobado cómo, en términos generales, la incorporación aislada de fuentes de información proporciona mejores resultados que su supresión del conjunto general. Una de las razones de ello podría encontrarse en el elevado índice de resolución que todas las fuentes de conocimiento consiguen de forma global. No obstante, a través de los datos obtenidos de su supresión, tanto la información sintáctica (−16,53 %) como la semántica (−10.75) influyen muy positivamente en el proceso global, especialmente al combinarlas (−42,98 %).

En relación con las ventajas que el uso de enriquecimientos supone sobre métodos de conocimiento limitado, algunos resultados comparativos preliminares entre el método ERA y el método de conocimiento limitado corroboran la importancia de la incorporación de estas nuevas fuentes de información. En concreto, y para

el bloque L009[11], se ha comprobado un incremento del 17,69 % según los cálculos realizados en función de la medida-F[12].

Es importante destacar, tal y como se ha podido comprobar, que la incorporación de la semántica no resulta trivial desde un punto de vista puramente computacional y las posibilidades de enriquecer los módulos semánticos son todavía muchas. Este trabajo deja una línea de investigación abierta a la búsqueda de técnicas de enriquecimiento del método propuesto que mejoren el comportamiento de la semántica en los procesos de resolución de la anáfora. Algunas de estas líneas futuras de trabajo se tratarán en la sección 7.2 (pág. 226).

[11] El resultado de la aplicación del método de conocimiento limitado sobre el bloque L009 representa el comportamiento que este método tiene en la evaluación global sobre el corpus Lexesp, y que está en la línea de otros métodos de conocimiento limitado (Palomar et al., 2001a; Mitkov, 1998), proporcionando índices de éxito cercanos al 80 %.

[12] La medida-F pondera conjuntamente la precisión y la cobertura, teniendo en cuenta las diferencias existentes entre los datos de evaluación de cada uno de los métodos.

Los parámetros de cálculo de esta medida son:

$$\text{precisión} = \frac{A_c}{A_t} \qquad \text{cobertura} = \frac{A_c}{A_e} \qquad F = \frac{(\beta^2 + 1{,}0) \times P \times R}{\beta^2 \times P + R}$$

donde A_e es el número de anáforas existentes, A_c es el número de anáforas resueltas correctamente, A_t es el número de anáforas tratadas, P es la precisión, R es la cobertura y β es el índice de importancia dado a la cobertura sobre la precisión (en este caso $\beta = 1$).

6. Marco de aplicación del método ERA

6.1 Introducción

El método propuesto en esta Tesis requiere del uso de un corpus en el que cada palabra se acompaña, entre otras etiquetas, de su sentido correcto en el texto. Para ello se ha realizado un etiquetado manual del corpus de entrada usando como herramienta de referencia el recurso léxico WordNet. Estas etiquetas son, por tanto, los sentidos que las palabras toman en el texto.

Esta desambiguación realizada a partir de los sentidos proporcionados por WordNet y llevada a cabo por procedimientos manuales pretende simular el comportamiento de herramientas de desambiguación que se encarguen de realizar la anotación semántica por procedimientos automáticos. El campo de investigación en la desambiguación del sentido de las palabras (más conocida por el término en inglés *Word Sense Disambiguation*, en adelante WSD) ha sido uno de los más prolíficos durante los últimos años.

La integración de técnicas de WSD en tareas de Procesamiento de Lenguaje Natural, como la resolución de la anáfora, conducirán a una automatización de los procesos y a una mejora sustancial de los enfoques basados en semántica debido a la posibilidad de realizar etiquetados automáticos (Suárez et al., 1999; Saiz-Noeda et al., 2001b; Muñoz et al., 2002b).

Otro de los puntos relevantes tratados en esta Tesis es el uso de una ontología de rasgos semánticos en la que se fundamenta la generación de patrones de compatibilidad semántica. Cualquier tarea de PLN basada en el uso de ontologías se verá claramente beneficiada por una mayor riqueza de conceptos y niveles. Asimismo, la mayor diversificación y especialización de esta ontología posibilitará una mayor precisión en el tratamiento de corpus de

dominio restringido. En este sentido, es fundamental contar con ontologías lo más ricas posibles. La utilizada en este trabajo es la que EuroWordNet proporciona y está formada por 64 conceptos ontológicos organizados en cuatro niveles distintos. Una de las ventajas principales del uso de esta ontología es que está perfectamente integrada en la red semántica de EuroWordNet y, por tanto, favorece la combinación de técnicas que hacen uso de la información semántica proporcionada tanto por la ontología como por el resto de los elementos de dicha red.

A pesar de que WordNet es un recurso extremadamente útil y rico en conocimiento, todavía requiere la corrección de carencias terminológicas y semánticas importantes. Uno de los problemas que plantea es la enorme semejanza semántica que guardan muchos de los *synsets* de una misma palabra (este hecho ha sido definido por algunos autores como granulado fina). Esta semejanza entre sentidos dificulta enormemente las tareas de desambiguación léxica que tienen que seleccionar el correcto de entre un grupo de sentidos muy similares. Otro problema es la ya mencionada ausencia de terminología común[1], así como de terminología específica de dominios concretos.

Estas carencias han propiciado el surgimiento de tendencias orientadas al enriquecimiento de WordNet con el fin de hacerlo más adecuado para su uso en tareas de PLN. En esta línea, se han realizado trabajos para el agrupamiento de los sentidos de WordNet en campos temáticos terminológicos (Magnini y Cavaglia, 2000) que corrijan el granulado fino comentado previamente. También se han llevado a cabo propuestas para la extensión de WordNet con terminología restringida a un dominio semántico concreto, como por ejemplo el dominio médico (Buitelaar y Sacaleanu, 2002) o el dominio medioambiental (Stamou et al., 2002a).

La mejora, a través de estas propuestas de enriquecimiento y extensión de los recursos semánticos, de los resultados en desambiguación léxica redundan en una evidente mejora de la eficiencia de tareas que, como la resolución de la anáfora del método ERA,

[1] En el capítulo anterior, dedicado a la evaluación, se han comentado algunos casos de nombres y verbos muy comunes que no aparecían en la base de datos de WordNet español.

requieren el menor índice de error posible en la anotación de los sentidos correctos.

Además de los requisitos semánticos que en lo referente a la desambiguación léxica se proponen para mejorar el proceso de resolución de la anáfora, esta tarea es fundamental en aplicaciones de diversos campos del PLN como la extracción de información (EI), la recuperación de información (RI) o la búsqueda de respuestas (BR).

Estas dos ideas, requisitos semánticos y aplicaciones, definen a la resolución de la anáfora, y en particular al método ERA, como una tarea que requiere de un preproceso que incluya la máxima cantidad –y naturalmente la máxima calidad– de información posible referente a la representación lingüística –y semántica en particular– del texto de entrada para poder realizar una correcta aportación en la mejora de los sistemas de PLN en los que se aplique. Este marco de aplicación queda representado en la figura 6.1.

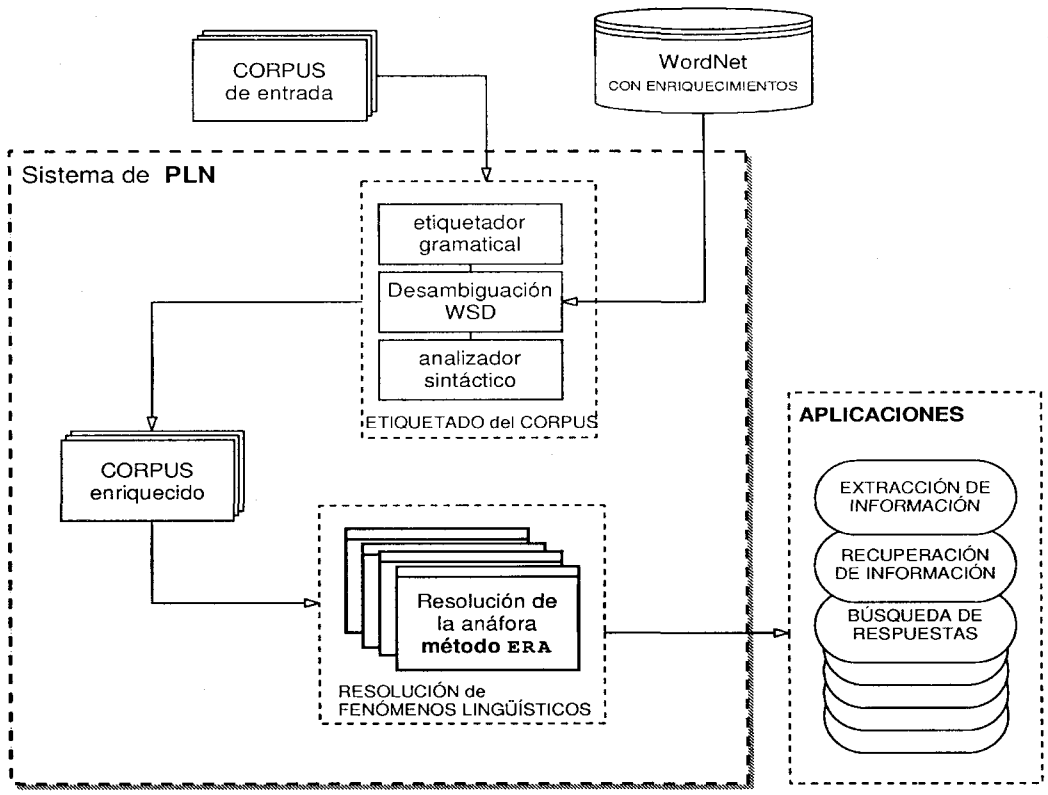


Figura 6.1. Marco de aplicación de la resolución de la anáfora en el PLN

De este modo, una vez presentado el marco de aplicación del método ERA, en este capítulo se presentarán las investigaciones realizadas en estas áreas, tanto en lo referente a requisitos de carácter semántico como a las aplicaciones de PLN. En primer lugar, en lo referente a los requisitos semánticos del método ERA, se tratarán a continuación dos propuestas de mejora del recurso léxico WordNet en la línea de lo comentado. Por un lado, se propondrá la combinación del método enriquecido de resolución de la anáfora con un mecanismo de desambiguación léxica basado en marcas de especificación de dominios. Por otro lado, se planteará la propuesta del proyecto EuroTerm, proyecto en el que el autor de esta Tesis ha participado activamente, y cuyo objetivo es el de extender el recurso EuroWordNet con terminología del sector público.

Asimismo, en segundo lugar y con el fin de demostrar la importancia del tratamiento del problema de la anáfora en el campo de las aplicaciones de PLN, este capítulo se encargara de encuadrar el proceso de resolución de la anáfora y en concreto el método ERA en el proyecto TUSIR, un proyecto cuyo objetivo es el de desarrollar técnicas de comprensión de textos en la recuperación de información.

6.2 El método ERA: Requisitos semánticos

Las tendencias actuales en el uso de recursos léxicos como WordNet, apuntan a una mejora de estos recursos en dos áreas diferentes.

Por un lado, y con el objetivo de reducir el problema de similitud entre los sentidos proporcionados por WordNet, se plantean agrupamientos de estos sentidos en función de su proximidad semántica. Esto conduce además a la definición de campos temáticos que agrupen también aquellos sentidos que compartan un dominio concreto. Además de contribuir a una mejor desambiguación del sentido de las palabras, el agrupamiento en campos temáticos aporta nuevas fuentes de información semántica a la propia resolución de la anáfora.

Por otro lado, la necesidad de recursos léxicos como WordNet en aplicaciones de dominios concretos obliga a extender este recurso con terminología propia de cada dominio.

A continuación se presentan dos propuestas que abarcan ambos enfoques, el agrupamiento en campos temáticos orientado a la desambiguación y la extensión de WordNet con terminología medioambiental.

6.2.1 Los campos temáticos en WordNet y la desambiguación de sentidos

El desarrollo de un sistema de Procesamiento del Lenguaje Natural debe contar con un módulo que resuelva la correferencia lingüística y, por tanto, que resuelva la anáfora. El problema de la anáfora ha sido definido como un fenómeno semántico y, tal y como se ha venido tratando a lo largo de este trabajo, la información semántica debe integrarse en el proceso de resolución junto con otras fuentes de conocimiento.

Para facilitar la incorporación de esta semántica se ha de contar con un módulo de desambiguación del sentido de las palabras (WSD) que proporcione para cada término un sentido correcto de entre los posibles. La mejora de estos sistemas de desambiguación redundan en una mayor precisión de los métodos de resolución de la anáfora que, como el método ERA, plantean la incorporación de la semántica en el conjunto de fuentes de información.

En (Muñoz et al., 2002b) se propone un sistema completo de PLN compuesto por un módulo de resolución de la anáfora pronominal, un módulo de resolución de descripciones definidas y un módulo de WSD. La propuesta de este módulo de WSD está basada en el método de marcas de especificación (Montoyo y Palomar, 2000) y, además de seleccionar un sentido de cada palabra, extrae una etiqueta de campo temático o dominio. Este método se ha denominado método de marcas de especificación de dominio.

Los campos temáticos[2] de WordNet (Magnini y Cavaglia, 2000) son una extensión de la versión 1.6 de este recurso en la que la

[2] *Campo temático* es una traducción directa del término *Subject Field* utilizado por los autores, si bien este término y el de *dominio* se usarán indistintamente a lo largo de este capítulo para referir al mismo concepto.

práctica totalidad de los *synsets* contenidos han sido anotados con una etiqueta de campo temático o dominio. Así, tanto nombres como verbos quedan agrupados dentro de WordNet en función de un dominio concreto perteneciente a una jerarquía de 250 códigos de campos temáticos (por ejemplo, tanto los nombres *hospital* y *doctor* como el verbo *operar*, en sus sentidos adecuados, estarían incluidos en el campo temático de 'Medicina'). En este sentido, en (Montoyo y Palomar, 2001) se demuestra cómo la tarea de WSD basada en marcas de especificación obtiene mejores resultados cuando se aplica a dominios. En concreto, este trabajo revela resultados cercanos al 95 % de éxito en la desambiguación léxica cuando el método de marcas de especificación se aplica a sistemas de clasificación como el IPTC[3] frente a un 68 % cuando se aplica a textos no restringidos usando WordNet como base de datos léxica.

Por tanto, teniendo en cuenta estos resultados, para la desambiguación del sentido de las palabras en el texto se ha utilizado el método de marcas de especificación[4]. En términos generales, una marca de especificación es un elemento jerárquico raíz que agrupa un conjunto de términos de manera similar a como lo hace una clase semántica en WordNet a través de las relaciones de hiperonimia/hiponimia. Este agrupamiento indica una proximidad en los términos incluidos en cada marca. Para realizar la desambiguación, se toma un contexto formado por las palabras que acompañan a la que se desea desambiguar. Para cada uno de los synsets asociados a las palabras del contexto se recorren las ramas de la jerarquía semántica definida por cada marca de especificación. Aquella marca que contenga al mayor número de sentidos de las palabras del contexto será la elegida para la desambiguación del sentido.

[3] El sistema de referencia temática IPTC ha sido desarrollado para permitir a los proveedores de información el acceso a un sistema universal de codificación independiente del lenguaje para indicar el contenido temático de nuevos elementos. Ver información detallada en `http://www.iptc.org`.

[4] La propuesta basada en información lingüística para resolver el problema de la ambigüedad léxica con el uso de marcas de especificación constituye una de las tesis doctorales más recientes desarrolladas en el seno del GPLSI (Montoyo, 2002).

A partir del sentido obtenido por el mecanismo de desambiguación de marcas de especificación, el método de marcas de especificación de dominio propuesto asignaría la etiqueta de dominio siguiendo tres pasos.

1. Obtención del *synset* en WordNet1.5. A partir del sentido desambiguado usando WordNet español, se obtiene el *synset* correspondiente en el WordNet 1.5. a través de un identificador de *synset* que ambos comparten. Por ejemplo, el *synset* de *teléfono#2* en el WordNet español se corresponde con el *synset* de *phone#1* en el WordNet 1.5.
2. Emparejamiento de WordNet 1.5 y WordNet 1.6. Dado que la versión de WordNet usada para esta investigación es la 1.5, es necesario establecer una correspondencia entre los synsets de una versión y de otra, ya que existen algunos cambios de estructura entre ambas. Para establecer esta correspondencia se propone el uso del emparejado de WordNets 1.5 y 1.6 (Daudé et al., 2001).
3. Obtención de la etiqueta de dominio. Por último, se consultan los dominios de WordNet y se extrae la etiqueta de campo temático para el *synset* de WordNet 1.6 obtenido en el paso anterior.

De esta manera, además del sentido de cada palabra en el texto, es posible contar con información del dominio del concepto asociado, algo que será muy útil para cualquier tarea de PLN en un dominio concreto, incluida, naturalmente, la resolución de la anáfora.

La resolución de la anáfora con campos temáticos. Además de favorecer los procesos de desambiguación que evidentemente mejoran el etiquetado semántico del corpus, el método ERA puede enriquecerse notablemente al contar con información semántica relativa al dominio.

Además de utilizar las ya detalladas preferencias semánticas basadas en el concepto de compatibilidad entre un nombre y un verbo, es posible instrumentar un mecanismo de preferencia basado en la afinidad de dominio existente entre el antecedente y el verbo de la anáfora.

Aunque las etiquetas de dominio están definidas tanto para nombres como para verbos, el método de desambiguación de las palabras trabaja, por el momento, únicamente con nombres. Por ello, es necesario recurrir a alguna estrategia que determine el dominio asociado al verbo.

Una propuesta para determinar este dominio es procesando la glosa que acompaña a cada término en el ILI. A partir de esta glosa, pueden extraerse un conjunto de palabras relevantes que ayuden a determinar el dominio de una palabra no etiquetada, en este caso, de un verbo. En realidad es una técnica similar a la usada para la desambiguación con marcas de especificación pero en la que el contexto de la palabra a desambiguar está formada no sólo por las palabras que aparecen en el entorno, sino también por las incluidas en su definición. Una vez desambiguadas las palabras contenidas en la glosa del verbo, se puede determinar que el dominio asociado al verbo es el dominio asociado a las palabras de su contexto.

Una vez que se ha determinado el dominio del verbo, puede establecerse una relación de dominio entre él y un candidato anafórico. Supongamos el ejemplo (116) perteneciente al dominio botánico.

(116) Los *hongos*$_i$ que nacen en las laderas de los montes contienen un veneno muy peligroso. $\emptyset_i$ Crecen muy rápidamente durante la primavera...

El verbo *crecer* tiene en español varios sentidos (aumentar de tamaño tanto en cosas concretas como abstractas, hacerse mayor referido a personas y otros seres vivos, prosperar, ...), y en WordNet español se recogen seis de ellos. El sentido del verbo *crecer* más afín al dominio botánico es el que se refiere al crecimiento de plantas representado en WordNet español por *crecer#2* y correspondiente con el synset de inglés *grow#4*. La glosa que acompaña en a este verbo es:

crecer#2: "of living matter, such as **plants** and animals". 2ndOrderEntity 30 Dynamic Quantity SituationType

Supongamos que los antecedentes de la anáfora están etiquetados con sus sentidos asociados al dominio concreto con los siguientes sentidos y glosas:

- *hongo#1*: "a parasitic **plant** lacking chlorophyll and leaves and true stems and roots and reproducing by spores". 1stOrderEntity 20 Group Living Natural Origin Plant
- *ladera#1*: "the side or slope of a hill". 1stOrderEntity Form Function Natural Object Origin Place Substance
- *monte#2*: "a land mass that projects well above its surroundings; higher than a hill". 1stOrderEntity Form Function Natural Object Origin Place Solid Substance
- *veneno#2*: "a substance that causes injury, illness, or death". 03 1stOrderEntity 27 Form Object Origin Substance

Además, consultando las relaciones semánticas, el término *hongo#2* tiene como hiperónimo directo al término *planta#1*. Así, a través de los conceptos de las glosas de *hongo#2* y de *crecer#2* se podría establecer un vínculo más fuerte que el existente entre el resto de los candidatos y el verbo, concluyendo que el verbo *crecer#2* pertenece al dominio de la botánica[5].

De esta manera, se enriquecería el método de desambiguación con una técnica de anotación de verbos y se incorporaría información adicional sobre el campo temático o dominio del verbo al método de resolución de la anáfora. La figura 6.2 muestra el esquema de integración de los dominios al método enriquecido de resolución de la anáfora.

Es evidente que esta estrategia, al igual que cualquier otra basada en la semántica, tendrá un comportamiento más satisfactorio cuando se aplica sobre dominios restringidos. Asimismo, debido a las carencias terminológicas del WordNet español sobre dominios concretos, los enriquecimientos y extensiones de WordNet que garanticen una mayor cobertura terminológica son esenciales en este tipo de estrategias de PLN. En la siguiente sección se tratará de una estrategia para extender EuroWordNet con terminología de

[5] Este ejemplo pretende ser únicamente ilustrativo del enfoque propuesto y no representa necesariamente un caso incorrectamente resuelto por el método ERA. De hecho, el método ERA, tal y como ha sido planteado, resolvería esta anáfora bien con el uso de un patrón de incompatibilidad del tipo $DEBE$(crecer#2,Living,S) o bien mediante la adquisición de patrones de compatibilidad que preferirían un antecedente de tipo 'viviente' para el verbo *crecer#2*.

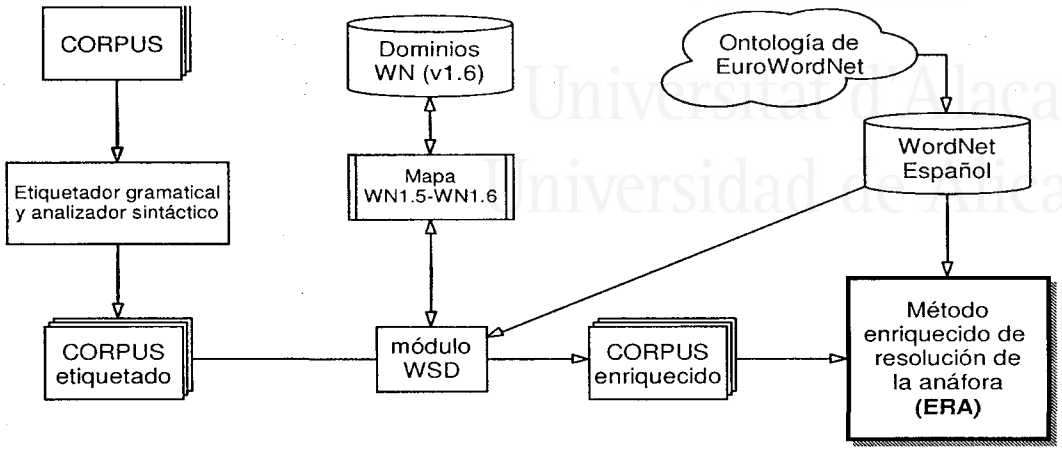


Figura 6.2. Integración del módulo de WSD y las etiquetas de dominio en el sistema ERA

dominio que podrá aplicarse a tareas de PLN en el ámbito del sector público y, en particular, en el dominio medioambiental.

6.2.2 Extensión de EuroWordNet con terminología del sector público: el proyecto EuroTerm

Además de la definición de dominios o campos temáticos, una tarea esencial en el enriquecimiento de WordNet es precisamente la incorporación de nuevos conceptos a los ya incluidos en su base de datos. Así, se han planteado extensiones de WordNet que permitan definir terminología específica de dominios restringidos. Una de estas extensiones ha sido desarrollada en el proyecto EuroTerm[6].

El objetivo de EuroTerm es la ampliación de EuroWordNet con terminología medioambiental en los idiomas griego, holandés y español. Tal y como se ha comentado en capítulos previos, EuroWordNet es una base de datos multilingual formada por WordNets genéricos en ocho idiomas europeos (Vossen, 1998, 2000). Los

[6] EuroTerm es un proyecto financiado por la Comisión Europea (EDC-2214) con una duración total de dieciocho meses (del 01/01/01 al 30/06/02) e incluido en las acciones preparatorias del programa *e-content*. El consorcio está formado por investigadores de las universidades de Patras (Grecia), de Tilburg (Holanda) y de Alicante (España). Los participantes españoles son miembros del Grupo de Procesamiento del Lenguaje y Sistemas de Información del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante, grupo y proyecto en el que el autor de esta Tesis mantiene una activa participación. La información relativa a este proyecto puede encontrarse en `http://dblab.upatras.gr`.

WordNets individuales incorporados en la base de datos central forman redes semánticas autónomas conectadas entre sí a través del índice inter-lenguas (*Inter-Lingual-Index* o ILI).

El objetivo de EuroTerm es enriquecer EuroWordNet con terminología específica de dominio medioambiental para los lenguajes antes mencionados. Cada uno de los WordNets monolingüe incorporará aproximadamente 1000 nuevos conjuntos de sinónimos (*synsets*) y será almacenado en una base de datos común, que será enlazada a la base de datos central de EuroWordNet bajo la etiqueta de dominio 'medio ambiente' (*Environment*).

Enfoque mixto: modelo de expansión y modelo de fusión. Existen dos aproximaciones para la construcción de una red semántica: el modelo de expansión y el modelo de fusión[7]. El modelo de expansión consiste en la traducción de conceptos en inglés a los idiomas respectivos y en el desarrollo de los *synsets* monolingües. El modelo de fusión implica el desarrollo independiente de *synsets* monolingües y su enlace al *synset* más equivalente del ILI (Vossen, 1996).

Para conseguir el solapamiento suficiente entre lenguas y que el vocabulario alcance una completitud y cobertura aceptables, se debe prestar mucha atención a la selección de los términos a incorporar en la red semántica. Así, tras una investigación exhaustiva de los dos modelos (expansión y fusión) usados con anterioridad para la construcción de redes semánticas, y teniendo en cuenta la aplicación de EuroTerm, se concluyó que la combinación de ambos modelos daría lugar a resultados más consistentes y fiables. De este modo, a diferencia de la extensión de EuroWordNet con terminología informática llevada a cabo siguiendo el modelo de expansión (Vossen et al., 1999), este enfoque marca unas pequeñas diferencias debidas al uso combinado de los modelos de expansión y fusión (Stamou et al., 2002a).

La razón de aplicar esta combinación de modelos es la de asegurar el suficiente solapamiento en la cobertura de los WordNets monolingües manteniendo las características, particularidades y diferencias específicas de cada lengua.

[7] Conocidos por los términos en inglés *expand model* y *merge model*.

Además de asegurar el solapamiento y las particularidades de cada idioma, se plantea como objetivo la interpretación de las diferencias encontradas entre los WordNets monolingües una vez que se han incorporado en sistemas de recuperación de información (RI), que es una de las aplicaciones principales consideradas en este proyecto.

Metodología de adquisición terminológica. Siguiendo el modelo de expansión, en la extracción de la terminología del dominio medioambiental se usaron un conjunto común de recursos léxicos en inglés. Para este proceso, se comenzó con un corpus medioambiental recopilado a partir de 429 documentos y glosarios en inglés que incluyen un total de 4972 términos junto con sus glosas. A este corpus se le aplicó un etiquetador gramatical (Zavrel y Daelemans, 1999) y un lematizador. A continuación se usó la métrica TF*IDF (Salton y Buckley, 1988) para contabilizar las frecuencias de los lemas que fueron clasificados en función de esta frecuencia. El proceso de la extracción se realizó automáticamente. Los 4500 términos más frecuentes fueron contrastados semiautomáticamente con los glosarios medioambientales en inglés y los que se encontraron en ellos se consideraron candidatos al WordNet medioambiental.

Los términos candidatos fueron contrastados de nuevo con el ILI en un proceso semiautomático para determinar los que estaban presentes con un sentido medioambiental asociado. Los términos medioambientales que no estaban en el ILI tenían también que ser contrastados con recursos monolingües antes de su incorporación a la base de datos.

Una vez que el primer conjunto de términos había sido extraído, se adoptó el modelo de fusión para contrastar estos términos con los existentes en varios recursos léxicos monolingües y enriquecerlos con términos monolingües ausentes. En esta etapa, se aplicó el modelo de fusión. En particular, los términos ausentes fueron contrastados manualmente con lexicones y corpus monolingües del dominio específico y sus glosas, encontradas en diccionarios, se investigaron para determinar su importancia en los idiomas correspondientes. Además, los términos medioambientales encontrados en los recursos monolingües pero no en la lista de

términos candidatos en inglés también fueron contrastados con los glosarios y diccionarios monolingües y, en el caso de que fueran relevantes, se incluyeron en la lista de términos candidatos.

La relevancia fue determinada en esta etapa por la frecuencia de un término en el corpus y por su presencia en los glosarios. Una vez que el proceso de selección había finalizado, los términos en inglés que no se encontraban en el ILI fueron manualmente traducidos a los idiomas correspondientes y etiquetados como medioambientales. Por otra parte, los términos encontrados en los recursos monolingües también fueron traducidos manualmente al inglés e incorporados al ILI. El desarrollo de los *synsets* en EuroTerm sigue la estructura de definición de relaciones internas al lenguaje usadas en EuroWordNet.

Si bien el proyecto EuroTerm está a punto de finalizar, no es posible dar cuenta de resultados concretos asociados a la fase de prueba y relativos a errores terminológicos o a la alineación de idiomas ya que esta fase está contenida en los últimos estadios de desarrollo del proyecto. No obstante, se ha estimado que los problemas que podrían aparecer están vinculados más a particularidades de cada idioma que a la propia metodología seguida para la adquisición terminológica.

La incorporación del conjunto final de términos a la base de datos se ha llevado a cabo con el uso de un sistema de alineación de terminología (*Terminology Aligment System*, TAS) que posibilita la conexión entre los tres lenguajes implicados.

El Sistema de Alineación de Terminología. El Sistema de Alineación de Terminología (TAS) es una parte clave de la infraestructura subyacente al proyecto EuroTerm (Hoppenbrouwers, 2001). A través del TAS, los miembros integrantes del proyecto pueden comunicar y coordinar su trabajo sobre los WordNets individuales en español, griego, y holandés.

El TAS se define como un sistema de alineación porque ayuda a los terminólogos a alinear su trabajo sobre los WordNets locales. El TAS no es una base de datos central unificada en la cual se combinan los WordNets locales, sino que es una base de datos de enlaces, diseñada para facilitar el trabajo cooperativo sobre los WordNets locales. Si bien el proyecto EuroTerm cuenta con una

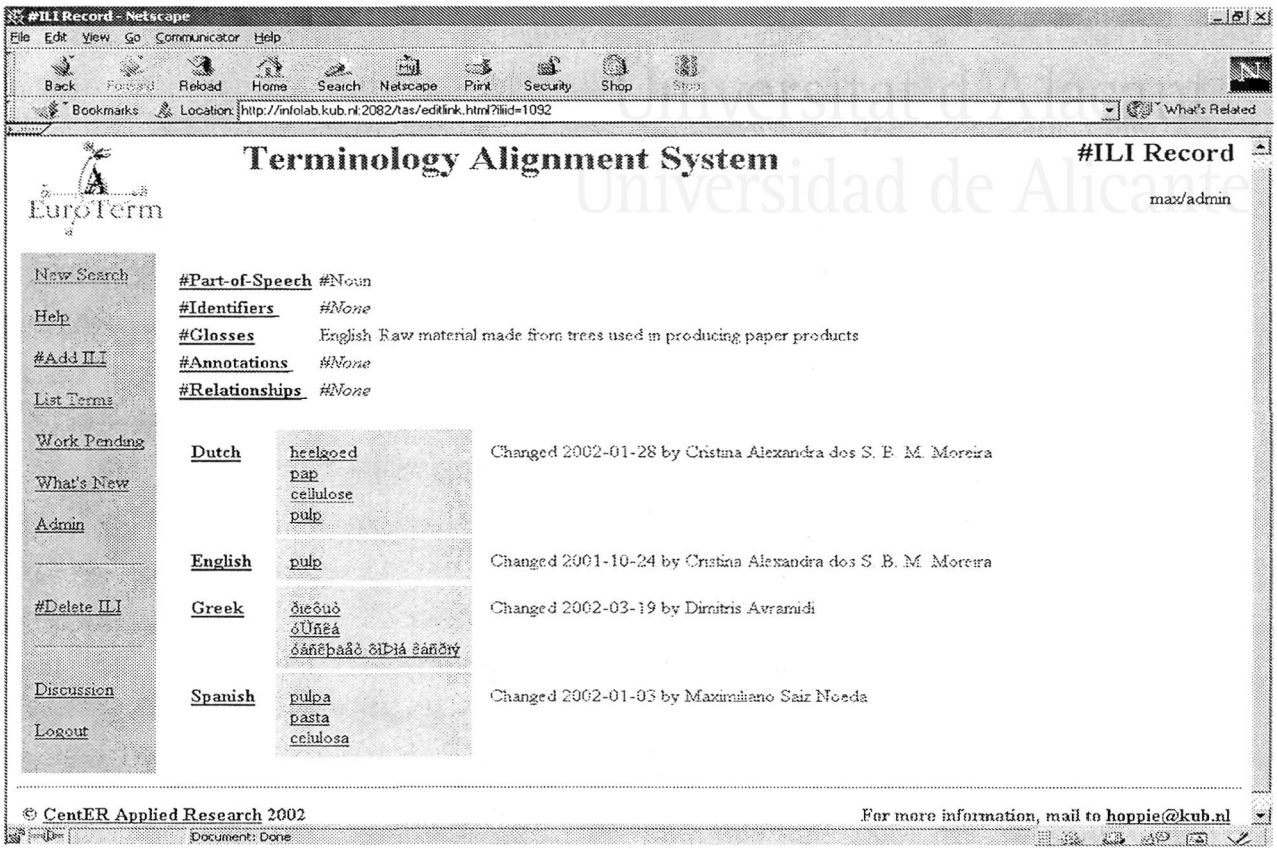


Figura 6.3. El Sistema de Alineación de Terminología (TAS) en el proyecto EuroTerm

base inicial común relativamente extensa que está formada por unos 1000 términos, aplicaciones futuras que hagan uso del TAS podrían contar con una cantidad inferior de datos base. En estos casos, el concepto de federación, donde todos los socios cooperan bajo una estandarización muy débil se explota al máximo. Si bien siempre es necesario seguir algunos estándares, en este caso se asume que EuroWordNet y su estructura proporcionan este mínimo necesario.

Cualquier herramienta utilizada para la gestión de WordNet, como puede ser Polaris o cualquier sistema abierto, puede seguir siendo utilizada para el mantenimiento del WordNet local. Por ello, el TAS puede considerarse como una herramienta federada en red que enlaza las herramientas locales y que ha sido construida como una base de datos Web. La figura 6.3 muestra una captura del sistema de alineación de terminología.

Las aplicaciones de EuroTerm. La aplicación más inmediata del WordNet multilingual de dominio específico es la incorporación del dominio medioambiental en un sistema de recuperación de información. Así, los documentos estarían representados semánticamente en vez de léxicamente en el índice del sistema. De esta manera, los términos que componen la pregunta (*query*) que se le hace al sistema se compararán con los documentos no sólo usando una medida de co-ocurrencia sino también a través de la similitud semántica de la pregunta y los conjuntos de índices documentales.

Para ir en esta dirección, es necesario realizar algunas modificaciones en el motor de búsqueda en el que se incorpore EuroTerm. En concreto, se requiere la incorporación de un directorio medioambiental en la interfaz del buscador que permita a los usuarios especificar si están o no interesados en realizar la consulta en este dominio. Además, el motor de búsqueda debe mantener dos índices por separado, uno que agrupe los documentos medioambientales y otro que contenga el resto de los documentos. El objetivo de EuroTerm es el de conseguir mejores resultados de precisión que los obtenidos si la búsqueda se realiza directamente con la co-ocurrencia de las palabras de la pregunta. Además, se espera que emparejando los términos de la pregunta con los *synsets* medioambientales correctos, los resultados de la recuperación mejoren en precisión y cobertura, dotándoles de más sentido para el usuario final.

La resolución de la anáfora como tarea integrada en sistemas de PLN orientados a la recuperación de información será tratada en la siguiente sección.

6.3 Aplicaciones: el proyecto TUSIR

Una vez que el método ERA pueda contar con un sistema de desambiguación léxica que proporcione de manera automática los sentidos de las palabras en el texto, el sistema de PLN definido en la figura 6.1, y en concreto el método ERA, puede ser usado en aplicaciones de Extracción de Información, Recuperación de In-

formación o Búsqueda de Respuestas enriquecidas con extensiones de WordNet para su uso en dominios concretos.

Relacionado con la recuperación de información, a continuación se expondrán las investigaciones que se están llevado a cabo en el seno del Grupo de Procesamiento del Lenguaje y Sistemas de Información relacionada con la aplicación de la resolución de la anáfora a la comprensión de textos y que constituye el marco de aplicación del método ERA.

El proyecto "TUSIR: Desarrollo de un sistema de comprensión de textos aplicado a la recuperación de información", subvencionado por la Comisión Interministerial de Ciencia y Tecnología (CICyT) con número de referencia TIC2000-0664-C02-01/02, es un proyecto coordinado entre la Universidad Politécnica de Valencia y la Universidad de Alicante y en el que el autor de esta Tesis participa como investigador de la segunda universidad. El objetivo de este proyecto consiste en el desarrollo de técnicas de análisis de textos para su incorporación en sistemas de procesamiento de lenguaje natural aplicables a la resolución de problemas de recuperación de la información.

Este proyecto sigue la línea de colaboración iniciada con el proyecto de investigación "Construcción de Analizadores Híbridos de Lenguajes Naturales" subvencionado por la CICyT de referencia TIC97-0671-C02-01/02 entre los mismos grupos de investigación.

Un sistema como el propuesto en el proyecto TUSIR debe tomar como entrada frases de consulta a un sistema de información documental, escritas en lenguaje natural, sin otras restricciones que las que marca la propia aplicación y debe proporcionar como salida la relación de documentos con información relevante sobre la consulta solicitada.

Para llevar a cabo este proceso global es necesario desarrollar una estructura semántica que represente los conceptos significativos de los documentos almacenados así como un conjunto de estrategias de búsqueda conceptual en esa estructura semántica que sean compatibles con el significado de la consulta realizada.

La consecución de estos objetivos pasa por la construcción de una plataforma de integración de todas las herramientas desarrolladas, mediante un entorno gráfico, para facilitar las tareas de

construcción y validación de corpus etiquetados léxica, sintáctica, y semánticamente, así como con anotación correferencial. Este etiquetado[8] sigue la línea definida en esta Tesis al respecto de los requisitos del método ERA en lo referente a la anotación del corpus (ver apartado 4.3.2).

En lo referente a los logros científicos, el proyecto TUSIR propone el desarrollo de analizadores sintácticos parciales utilizando aproximaciones basadas en reglas y en modelos estadísticos, así como el de nuevos métodos de resolución de la correferencia lingüística. Dado que TUSIR plantea una metodología basada en información lingüística de varios niveles, incluido el semántico, las propuestas de métodos de resolución de la correferencia han de integrar esta fuente de información tal y como lo hace el método ERA.

Además, se plantea el desarrollo de estrategias de comprensión de texto y de técnicas de desambiguación del significado de las palabras mediante el uso de conocimiento lingüístico, estadístico y aprendizaje automático, algo que enlaza directamente con las técnicas antes mencionadas de WSD (Suárez y Palomar, 2002; Molina et al., 2002; Montoyo, 2002).

Por último, TUSIR propone un estudio de la aplicabilidad de las técnicas desarrolladas a la recuperación de información. Dentro del desarrollo de nuevos métodos de resolución de la anáfora, se procederá a resolver las posibles relaciones de correferencia existentes entre los distintos sintagmas analizados. Con ello se pretende reducir el número de entidades existentes en el texto y agrupar toda la información disponible de cada una de ellas. Por ejemplo, inicialmente puede aparecer la descripción general de una entidad y, posteriormente, en el texto se pueden hacer referencias a ésta para introducir nueva información. Dichas referencias han de ser identificadas para así completar toda la información de cada entidad. Se resolverán las anáforas pronominales, alias, sintagmas nominales definidos y expresiones temporales de referencia. Aquí se

[8] El etiquetado manual del corpus Lexesp está siendo llevado a cabo por participantes en este proyecto: Manuel Pruñonosa (Universidad de Valencia), Borja Navarro (Universidad de Alicante) y Eugenia Ferrer (Universidad Politécnica de Valencia).

pretende construir un mecanismo que resuelva las correferencias detectadas en el texto. Para la construcción de estos mecanismos se necesita información léxica, morfológica, sintáctica, semántica y contextual.

Como puede verse, a través de sus objetivos, este proyecto desarrollará estrategias de resolución de la anáfora que se incorporarán en el proceso de comprensión (total o parcial) de frases y textos, definiendo sus requisitos lingüísticos y, en concreto, los semánticos y proponiendo su aplicación a tareas de PLN. La resolución de los pronombres es fundamental dentro de esta tarea, especialmente la aportación de un método de resolución de pronombres basada en información semántica. La información semántica es, por tanto, pilar fundamental de este proyecto y enlaza directamente con las propuestas realizadas en esta Tesis.

Trabajos realizados por miembros del GPLSI han mostrado la positiva influencia de la resolución de la anáfora pronominal en tareas de recuperación de información así como en sistemas de búsqueda de respuestas (Vicedo y Ferrández, 2000).

6.4 Conclusiones

La resolución de la anáfora es, sin lugar a dudas, una de las tareas clave dentro de un sistema global de PLN. El método ERA, propuesto en esta Tesis, puede ser incorporado a tareas de PLN que requieran de la resolución de pronombres para mejorar sus resultados. Esta resolución de la anáfora, basada en semántica ontológica extraída de WordNet, puede ser comprensiblemente mejorada con el uso de módulos de desambiguación basados en los enriquecimientos de WordNet a partir del agrupamiento de conceptos en campos temáticos y de la extensión terminológica de dicho recurso.

En este capítulo se ha descrito la integración del método ERA en un sistema de PLN con el uso de un módulo de desambiguación léxica. Para el método de desambiguación, basado en marcas de especificación, se ha propuesto un enriquecimiento a partir de la definición de un conjunto de etiquetas de campos temáticos

que agrupan los synsets de WordNet en función de su significado. Además, se ha descrito la propuesta de extensión de EuroWordNet con terminología del sector público llevada a cabo en el proyecto EuroTerm.

Con la presentación del proyecto TUSIR, se ha mostrado cómo la resolución de la anáfora puede combinarse con los sistemas de recuperación de información para mejorar sus resultados. La sustitución de los pronombres por sus antecedentes modifican los índices de frecuencia usados en la recuperación de información para determinar la relevancia de los documentos. Si, además, la estrategia de resolución anafórica está basada en información semántica, el uso de las etiquetas de dominios extraídas de WordNet y aplicadas a textos de dominio restringido afinarán aún más los criterios de selección de los documentos relevantes.

Por último, es conveniente mencionar el hecho de que, al igual que la recuperación de información, otras aplicaciones como la búsqueda de respuestas se han visto beneficiadas de los procesos previos de resolución de la anáfora, mejorando los resultados obtenidos sin la aplicación de esta tarea.

7. Conclusiones finales

7.1 Conclusiones sobre el trabajo presentado

En este trabajo se ha realizado un riguroso estudio de la influencia y el uso tanto de la información basada en los papeles sintácticos de los elementos oracionales como de la información semántica extraída de una conjunto de conceptos ontológicos definidos en EuroWordNet.

Este estudio se ha enfocado a la resolución de la anáfora pronominal de tercera persona en español, cubriendo los casos de anáfora provocados por pronombres personales, demostrativos, reflexivos y omitidos.

Tradicionalmente, y por razones de eficiencia y cobertura, los investigadores del campo de la resolución de la anáfora han centrado sus esfuerzos en la aplicación de información de origen morfológico y sintáctico. La mayoría de estos autores han coincidido en destacar la semántica como el complemento necesario a integrar en cualquier sistema de resolución de la anáfora con el fin de cubrir casos que el resto de las fuentes eran incapaces de resolver correctamente. Asimismo, muchos de estos trabajos basados en conocimiento limitado, y en particular los realizados para el español, han hecho uso de análisis sintáctico parcial, con lo que el etiquetado carecía de información al respecto de la función sintáctica que un sintagma nominal tenía con respecto al verbo al que acompañaba.

En esta Tesis, la información de papeles sintácticos y la semántica basada en ontologías han sido combinadas para proponer una metodología de resolución basada en información enriquecida que incorpora la morfología, la sintaxis y la semántica, así como información de carácter estructural. Se ha realizado un estudio ex-

haustivo de las fuentes que intervienen en el proceso de resolución así como de las necesidades que un sistema que incorpora este tipo de información adicional, requiere para su correcto funcionamiento.

En el estudio de la influencia que cada fuente de información tiene en el proceso de resolución de la anáfora se ha comprobado la enorme relevancia que tiene tanto la información de papeles sintácticos como la semántica. Esta relevancia queda reforzada en la combinación de la semántica con otras fuentes de información como la sintáctica o la estructural y, en general, se ha demostrado cómo la combinación de todas las fuentes de información es la que proporciona los mejores resultados.

De este modo, podríamos resumir las aportaciones que plantea esta Tesis en los siguientes puntos:

- Contextualización de la anáfora, en la que este fenómeno lingüístico se relaciona con otros fenómenos como la elipsis o la deixis, y clasificación de la anáfora en función de distintos criterios. El primero de ellos ha sido la relación existente entre la anáfora y su antecedente. La segunda clasificación ha tenido en cuenta la categoría sintáctica del antecedente. La tercera de las clasificaciones, la más extensa, ha usado como criterio la naturaleza sintáctica del elemento anafórico. Con el fin de mantener la dinámica de la propuesta, cada uno de los tipos de anáfora contenidos en la última clasificación ha contado con ejemplos de resolución basados en información morfosintáctica, por un lado, y ejemplos relativos a la necesidad de aplicación de la semántica por otro.
- Revisión del estado del arte, basado en los mismos criterios propuestos en el trabajo, bajo los cuales se han definido tres grupos principales: los trabajos denominados de conocimiento limitado, que fundamentan la resolución en información morfológica y sintáctica; los trabajos denominados enriquecidos, que incorporan información semántica y de discurso y, por último, un grupo de aproximaciones alternativas que resuelven la anáfora por mecanismos extra-lingüísticos.

- Estudio de las diferentes fuentes de conocimiento que intervienen en el proceso de resolución de la anáfora y repaso de algunos recursos que las proporcionan.
- Propuesta del método de conocimiento limitado basado en un conjunto de restricciones y preferencias de carácter morfológico y sintáctico. Evaluación de los índices de éxito de este método en la resolución de la anáfora y comparación de estos resultados con los obtenidos por otros métodos que han sido implementados y adaptados al español.
- Propuesta de etiquetado sintáctico-semántico enriquecido a partir de un análisis parcial del corpus de entrada. En este etiquetado se han incluido las necesidades de anotación adicionales que plantea el método enriquecido de resolución de la anáfora, entre las que se incluye el etiquetado de los papeles sintácticos de los elementos oracionales así como los sentidos correctos de las palabras a partir del recurso léxico WordNet.
- Propuesta del método enriquecido de resolución de la anáfora pronominal en español (ERA). Se ha propuesto un método que incorpora a las fuentes de conocimiento limitado las provenientes de los papeles sintácticos y la información semántica. Basado también en un conjunto de restricciones y preferencias, el método ERA aporta criterios adicionales a la resolución de la anáfora, criterios cuya eficacia se ha puesto de manifiesto en diferentes ejemplos y en la propia evaluación.
- Construcción de un banco de pruebas para la evaluación del método ERA. El banco de pruebas ha sido diseñado específicamente para determinar la influencia de las distintas restricciones y preferencias y, por tanto, de las diferentes fuentes de conocimiento, en la aplicación del método sobre un corpus de evaluación.
- Análisis de la influencia de las distintas fuentes de información en la resolución de la anáfora con el método ERA. Usando el banco de pruebas, se han realizado evaluaciones del comportamiento de diferentes grupos de restricciones y preferencias sobre un corpus de entrada. A partir de la adición y la eliminación de estos conjuntos de restricciones y preferencias se ha reflexionado sobre la importancia que cada fuente de información tiene de

forma individual y cooperativa sobre el proceso de resolución de la anáfora.

Con este trabajo se ha pretendido llenar un espacio hasta ahora vacío en la resolución de la anáfora pronominal en español, proporcionando una base lingüística, científica y metodológica de la aplicación de los papeles sintácticos y la semántica de ontologías en el proceso de resolución.

7.2 Trabajos en progreso y líneas futuras

Como ya se ha comentado, la aplicación de la semántica en la resolución de la anáfora deja abierta una gran cantidad de líneas de investigación y desarrollo.

Algunos aspectos concretos a tratar a corto plazo para la mejora del método ERA son:

- Incorporación de preferencias semánticas relativas a los adjuntos de los verbos copulativos. Es evidente que la carga semántica de un verbo copulativo es nula, proporcionando el adjunto de dicho verbo todos los matices semánticos asociados. La estrategia de construcción de patrones semánticos está basada en los sustantivos y sus verbos. Si se incorpora la posibilidad de tomar el sentido de un verbo copulativo a través de su adjunto, se podrían construir patrones específicos. Así, el tratamiento semántico de "*Los tomates maduran*" y de "*Los tomates están maduros*" tendría características similares.
  Además, el tratamiento de los adjuntos del verbos *ser* copulativo permitirían aplicar criterios de paralelismo semántico como en el ejemplo (117) extraído del corpus de evaluación, donde el pronombre relativo puede ser resuelto aplicando este tipo de paralelismo ("*el pensamiento freudiano era el pensamiento de nuestro siglo*").

(117) Es fácil sentir aún en Bergasse las huellas del *pensamiento*$_i$ freudiano, *que*$_i$ era el pensamiento de su siglo...

Estos criterio, además, pueden enriquecerse con la información semántica procedente de relaciones como la sinonimia o la hiperonimia para resolver casos como los de (118) o (119).

(118) Pedro se ha comprado un *coche*$_i$ nuevo. $\emptyset_i$ Es un *automóvil*$_i$ muy seguro.

(119) Pedro se ha comprado un *coche*$_i$ nuevo. $\emptyset_i$ Es un *vehículo*$_i$ muy seguro.

- Eliminación de candidatos semánticamente semejantes a candidatos incompatibles. Cuando durante el proceso de resolución un candidato se elimina por razones semánticas, en realidad se está estableciendo una incompatibilidad entre el pronombre y lo que el candidato eliminado representa. Según esto, también podría quitarse de la lista de candidatos todo aquel que correfiera con el eliminado. Esta correferencia puede ser determinada por relaciones de sinonimia existentes entre el candidato eliminado y otros incluidos en su lista. Esto supondría la incorporación al método de un nuevo filtro semántico.
- Extensión de los patrones semánticos con relaciones de hiperonimia (Saiz-Noeda y Palomar, 2000). Si bien esta extensión podría complicar notablemente la gestión computacional de la semántica incorporada, las relaciones de hiperonimia pueden hacer más útiles los patrones semánticos, considerando la compatibilidad no sólo como una ecuación resultante de un conjunto de elementos ontológicos, sino como una ponderación de la "distancia" semántica existente entre el candidato y los patrones asociados al verbo. Estas técnicas han sido aplicadas satisfactoriamente para la resolución de descripciones definidas (Muñoz et al., 2000).
- Extensión de los patrones de incompatibilidad con combinaciones de conceptos. Si bien esta mejora pertenece más a la parte de implementación del método, afecta a la propia definición teórica de las reglas de incompatibilidad. La creación de reglas

más complejas facilitaría la definición de casos que afectan a un mayor conjunto de elementos ontológicos e incluso posibilitarían la definición de elementos ontológicos más complejos con la combinación de los contenidos en la ontología principal de EuroWordNet[1]: Comestible + Líquido = Bebida

- Ampliación de patrones de incompatibilidad con los términos contenidos en el *synset* del verbo. Así los patrones para el verbo *madurar#1* serían los mismos que los del verbo *envejecer#2* aunque éste no haya aparecido en el texto. Esta extensión responde al mismo criterio usado con los sinónimos de los nombres para las preferencias estructurales. Si bien esta característica no está implementada en el sistema, su uso expandiría considerablemente el conjunto de patrones, algo que podría mejorar su rendimiento.
- Ajuste de pesos óptimo para la tarea de resolución de la anáfora. Tan pronto como se disponga de un corpus lo suficientemente grande para ello, y gracias a las posibilidades de configuración del banco de pruebas, se realizará un proceso de entrenamiento del sistema para determinar la configuración óptima de las preferencias[2].

Además, en lo referente al ámbito de aplicación del método ERA, se plantea en un futuro inmediato su ampliación a otros tipos de anáfora pronominal como la generada por pronombres posesivos o pronombres de relativo. Para ello, es necesario realizar un estudio de las características propias de este tipo de pronombres, así como de las que comparte con el resto.

Uno de los objetivos planteados con el uso de EuroWordNet es la extensión de la definición del método a cualquier idioma recogido en dicho recurso. En este sentido, la estrategia de incorporación de la semántica definida para el método ERA puede ser fácilmente adaptada al resto de los idiomas de EuroWordNet, ya que hace

[1] EuroWordNet define en su documentación (Vossen et al., 1998) una clasificación de *conceptos base* a partir de los conceptos ontológicos principales que siguen esta filosofía.

[2] En la evaluación realizada para este trabajo no se ha llevado a cabo una fase de entrenamiento previa, estableciendo, desde el comienzo de las pruebas, unos pesos que se han mantenido inalterados durante toda la evaluación.

uso de conceptos ontológicos asociados al módulo inter-lenguas (*ILI*) que comparten todos ellos.

Por otra parte, tal y como se ha visto en las clasificaciones realizadas, el fenómeno de la anáfora es de una casuística muy variada y son todavía muchos los tipos de anáfora que quedan sin tratar, especialmente aquellos que requieren un mayor conocimiento semántico y pragmático para su resolución (anáforas verbales, anáforas adverbiales, ...). Técnicas como la presentada en esta Tesis, basadas en la incorporación de información semántica, ontológica y conocimiento del mundo, permitirán afrontar en mayor medida esos tipos de anáforas.

Por último, es importante destacar los esfuerzos que en el seno del Grupo del Procesamiento del Lenguaje y Sistemas de Información del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante se están realizando para la construcción de un sistema completo de PLN que integre diferentes módulos para distintas tareas, entre las que naturalmente se encuentra la resolución de la anáfora. Líneas futuras de actuación incluirán la incorporación de estas técnicas basadas en la semántica en el mencionado sistema global.

7.3 Producción científica

Se exponen a continuación las publicaciones en las que el autor de esta Tesis ha participado. La gran mayoría de ellas están relacionadas directamente con este trabajo, bien en aproximaciones sobre resolución de la anáfora, bien en técnicas de incorporación o extracción de semántica para tareas de PLN. Cada referencia viene acompañada de una breve descripción para que el lector pueda conocer mejor la vinculación de cada publicación con el trabajo expuesto en esta Tesis.

La publicaciones se han agrupado en función de su naturaleza, tanto por el tipo de publicación en revista o en congreso como por su carácter nacional o internacional.

7.3.1 Revistas internacionales

PALOMAR, MANUEL, ANTONIO FERRÁNDEZ, LIDIA MORENO, PATRICIO MARTÍNEZ-BARCO, JESÚS PERAL, MAXIMILIANO SAIZ-NOEDA y RAFAEL MUÑOZ (2001). «An algorithm for Anaphora Resolution in Spanish Texts», *Computational Linguistics*, **27**(4), 545–567.

- El contenido de este artículo es la culminación del trabajo realizado por miembros del grupo de investigación interuniversitario de Procesamiento del Lenguaje, formado por miembros del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante y miembros del Departamento de Sistemas Informáticos y Computación de la Universidad Politécnica de Valencia. El contenido de este artículo se corresponde con la propuesta base de esta Tesis sobre el método de conocimiento limitado y supone, sin lugar a dudas, uno de los sistemas con mejores resultados de la bibliografía de este área de investigación.

7.3.2 Revistas nacionales

SAIZ-NOEDA, MAXIMILIANO, ARMANDO SUÁREZ y JESÚS PERAL (1999). «Propuesta de incorporación de información semántica desde WordNet al análisis sintáctico parcial orientado a la resolución de la anáfora», *Procesamiento del Lenguaje Natural*, **25**, 167–173.

- Este artículo es una de las primeras propuestas del trabajo presentado en esta Tesis. Supone una primera aproximación al concepto de incompatibilidad entre el sujeto y el verbo y presenta algunas ideas que, por su interés, se han mantenido prácticamente sin cambios.

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR, PATRICIO MARTÍNEZ-BARCO, JESÚS PERAL, RAFAEL MUÑOZ y MAXIMILIANO SAIZ-NOEDA (1999). «Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística.», *Procesamiento del Lenguaje Natural*, **25**, 217–218.

- Este trabajo fue presentado como una demostración del sistema de resolución de la anáfora. Suponía uno de los primeros prototipos que integraba una interfaz visual para la configuración de los parámetros principales de resolución.

Ferrández, Antonio, Jesús Peral, Patricio Martínez-Barco, Maximiliano Saiz-Noeda y Rafael Romero (1997). «Resolución de la extraposición a izquierdas con las gramáticas de unificación de huecos.», *Procesamiento del Lenguaje Natural*, **21**, 167–182.

- Esta publicación es la primera de las realizadas por el autor de esta Tesis en el campo del PLN y si bien no tiene una relación directa con el trabajo de esta Tesis, encuentra su afinidad en dos aspectos fundamentales: la resolución de un problema lingüístico como es la extraposición a izquierdas y el formalismo gramatical SUG, usado posteriormente para el análisis parcial de los métodos desarrollados con conocimiento limitado.

7.3.3 Series incluidas en *Journal Citation Report* (*JCR*)

Muñoz, Rafael, Maximiliano Saiz-Noeda y Andrés Montoyo (2002). «Semantic Information in Anaphora Resolution», en *Proceedings of the Portugal for Natural Language Processing (PorTAL'2002)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, págs. 63–70, Algarve, Faro, Portugal.

- Ésta es una de las publicaciones más recientes y se centra en la combinación del método ERA con técnicas de WSD basadas en marcas de especificación. La sección 6.2 trata con detenimiento el contenido de esta propuesta.

Saiz-Noeda, Maximiliano, Manuel Palomar y Lidia Moreno (2001). «Pronoun Resolution in Spanish from Full Parsing», en *Proceedings of the International Conference on Text, Speech and Dialogue (TSD'2001)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, págs. 84–91, Zelezna Ruda, República Checa.

- En este artículo se introduce el concepto de resolución basada en los papeles sintácticos obtenidos a partir del análisis completo del texto. Presenta un conjunto de restricciones y preferencias que se aproximan a los planteados en esta Tesis, incluyendo algunas ideas preliminares sobre la evaluación.

PALOMAR, MANUEL, MAXIMILIANO SAIZ-NOEDA, RAFAEL MUÑOZ, ARMANDO SUÁREZ, PATRICIO MARTÍNEZ-BARCO y ANDRÉS MONTOYO (2001). «PHORA: A NLP system for Spanish», en Alexander Gelbukh, editor, *Proceedings of the Second International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2001)*, Lectures Notes In Computer Science. Springer-Verlag, págs. 128–139, Springer Verlag, Mexico City, Mexico.

- Este artículo parte de la propuesta de Palomar et al. (2000) y hace uso de un conjunto de estrategias y teorías acerca de la resolución de la anáfora incluyendo además métodos de desambiguación léxica de distinta naturaleza. Él objetivo común es el de proporcionar un sistema global de PLN que integre todos los módulos necesarios para una resolución completamente automática.

SAIZ-NOEDA, MAXIMILIANO y MANUEL PALOMAR (2000). «Semantic Knowledge-driven Method to Solve Pronominal Anaphora in Spanish», en *NLP'2000 Filling the gap between theory and practice*, Lecture Notes In Artificial Intelligence. Springer-Verlag, págs. 204–211, Patras, Greece.

- Este artículo plantea una de las primeras teorías acerca de la incorporación de la semántica basada en ontologías. Hace uso de una ontología *ad-hoc* para exponer los mecanismos básicos de determinación de compatibilidad entre un nombre y un verbo.

7.3.4 Congresos internacionales

MUÑOZ, RAFAEL, RUSLAN MITKOV, MANUEL PALOMAR, JESÚS PERAL, RICHARD EVANS, LIDIA MORENO, CONSTANTIN ORASAN, MAXIMILIANO SAIZ-NOEDA,

ANTONIO FERRÁNDEZ, CATALINA BARBÚ, PATRICIO MARTÍNEZ-BARCO y ARMANDO SUÁREZ (2002). «Bilingual Alignment of Anaphoric Expressions», en *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Canary Islands, Spain.

- Este trabajo está basado en la experiencia de dos de los grupos de investigación más conocedores del fenómeno de la anáfora, el grupo interuniversitario de Procesamiento del Lenguaje de la Universidad de Alicante y de la Universidad Politécnica de Valencia y el Grupo de Lingüística Computacional de la Universidad de WolverHampton. A partir de esta experiencia el artículo propone un mecanismo de alineación de expresiones anafóricas en textos bilingües español-inglés orientado a tareas como la traducción automática o la generación de anáfora multilingüe.

SAIZ-NOEDA, MAXIMILIANO, ARMANDO SUAREZ y MANUEL PALOMAR (2001). «Semantic pattern learning through Maximum Entropy-based WSD technique», en *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, págs. 23–29, Toulouse, France.

- Se combina en este artículo una propuesta de extracción de patrones semánticos orientada a la resolución de pronombres en inglés con un método de desambiguación léxica basado en Máxima Entropía. El tipo de patrones extraídos son el fundamento de los patrones semánticos presentados en esta Tesis.

PALOMAR, MANUEL, MAXIMILIANO SAIZ-NOEDA, RAFAEL MUÑOZ, ARMANDO SUÁREZ y PATRICIO MARTÍNEZ-BARCO (2000). «PHORA: A system to solve the Anaphora in Spanish», en *Proceedings of Third Colloquium on Discurse Anaphora and Anaphor Resolution (DAARC'2000)*, págs. 206–211, Lancaster, UK.

- Este artículo es la definición inicial del sistema resolución de la anáfora del Grupo de Procesamiento del Lenguaje y Sistemas de Información. Este trabajo plantea las bases del sistema basado en información morfológica, sintáctica y semántica que se enriquece con nuevos módulos en (Palomar et al., 2001b).

Muñoz, Rafael, Maximiliano Saiz-Noeda, Armando Suárez y Manuel Palomar (2000). «Semantic Approach to Bridging Reference Resolution», en *Proceedings of the International Conference Machine Translation and Multilingual Applications in the New Millennium (MT'2000)*, págs. 17.1–17.8, Exeter, UK.

- Ésta es una aportación de los mecanismos basados en la semántica extraída de WordNet a la resolución de descripciones definidas en español. Se usan para ello las relaciones de sinonimia, hiperonimia, rol temático y antonimia proporcionadas por este recurso léxico.

Saiz-Noeda, Maximiliano, Manuel Palomar y David Farwell (2000). «NLP system oriented to anaphora resolution», en *Proceedings of the International Conference Machine Translation and Multilingual Applications in the New Millennium (MT2000)*, págs. 19.1–19.7, Exeter, UK.

- Este artículo presenta un sistema de resolución de la anáfora basado en gramáticas léxico-funcionales (LFG) que generan en el análisis sintáctico un conjunto de características sintáctico-semánticas enriquecidas similares a las enunciadas en esta Tesis, por lo que la estrategia enunciada en el artículo tiene algunos puntos en común con la del método ERA.

Saiz-Noeda, Maximiliano, Jesús Peral y Armando Suárez (2000). «Semantic Compatibility Techniques for Anaphora Resolution», en *Proceedings of International Conference on Artificial and Computational Intelligence For Decision, Control and Automation In Engineering and Industrial Applications (ACIDCA'2000)*, págs. 43–48, Monastir, Tunisia.

- Se presentan en este artículo un conjunto de técnicas muy similares a las presentadas en esta Tesis para la incorporación de información semántica basada en elementos ontológicos de WordNet 1.5 en la resolución de la anáfora en inglés. En la evaluación del método se obtienen índices de éxito cercanos al 81 %.

Peral, Jesús, Maximiliano Saiz-Noeda, Antonio Ferrández y Manuel Palomar (1999). «Anaphora resolution and generation in a multilingual system.

An interlingua mechanism», en *Proceedings of the Venezia per il Trattamento Automatico delle Lingue (VEXTAL'99)*, págs. 315–324, Venice, Italy.

- Planteamiento de resolución y generación bilingüe de la anáfora. Uno de los aspectos tratados en este artículo es el fundamento de las reglas morfosemánticas definidas en esta tesis y usadas, tanto en español como en inglés, para la resolución y generación correctas.

Suárez, Armando, Maximiliano Saiz-Noeda y Manuel Palomar (1999). «A method of restricted knowledge acquisition from WordNet», en *Proceedings of the Third International Conference on Knowledge-based Intelligent Information Engineering Systems (KES'99)*, págs. 38–41, Adelaide, Australia.

- Obtención de una subred de sentidos relacionados con un dominio asociado a un texto de entrada. Básicamente se trata de la aplicación de técnicas de desambigüación para la obtención de synsets de WordNet en inglés asociados al dominio restringido del texto.

Palomar, Manuel, Antonio Ferrández, Lidia Moreno, Maximiliano Saiz-Noeda, Rafael Muñoz, Patricio Martínez-Barco, Jesús Peral y Borja Navarro (1999). «A Robust Partial Parsing Strategy based on the Slot Unification Grammars», en *Proceeding of the Sixth Conference on Natural Language Processing, TALN'99*, págs. 263–272, Corsica, France.

- Trabajo sobre el analizador parcial usado para procesar el corpus de evaluación que se ha utilizado en el método de conocimiento limitado propuesto en (Palomar et al., 2001a). Este analizador genera su análisis a partir del formalismo gramatical SUG (*Slot Unification Grammar*. Gramáticas de Unificación de Huecos).

7.3.5 Congresos nacionales

Saiz-Noeda, Maximiliano, Patricio Martínez-Barco y Manuel Palomar (1997). «Paralelismo sintáctico-semántico para el tratamiento de elementos

extrapuestos en textos no restringidos», en *Proceedings of the VII Congreso de la Asociación Española para la Inteligencia Artificial CAEPIA-TTIA'97*, págs. 797–804, Málaga, Spain.

- A partir del trabajo presentado en (Ferrández et al., 1997), este artículo profundiza en los aspectos sintácticos y semánticos que relacionan los elementos extrapuestos en una oración. Este tema sirvió como introducción del autor de esta Tesis al área del Procesamiento del Lenguaje Natural y a la resolución de problemas lingüísticos. Sin poder establecer vínculos directos con los contenidos de esta Tesis, este artículo guarda relaciones con los aspectos computacionales de la resolución de fenómenos lingüísticos.

7.3.6 Informes internos

Llopis, Fernando, Rafael Muñoz, Armando Suárez, Andrés Montoyo, Manuel Palomar, Antonio Ferrández, Jesús Peral, Patricio Martínez-Barco, Rafael Romero y Maximiliano Saiz-Noeda (1998). «Sistema EXIT», *Informe interno*, DLSI. Universidad de Alicante. Alicante, Spain.

- Informe sobre el sistema de extracción de información EXIT del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante.

Bibliografía

ABAD, ANTONIO (1980). *Introducción a la lingüística*, Alhambra, Madrid, España.

ALARCOS, EMILIO (1994). *Gramática de la Lengua Española*, RAE. Espasa Calpe, Madrid, España.

AONE, CHINATSU y SCOTT WILLIAM BENNETT (1994). «Discourse tagging tool and discourse-tagged multilingual corpora», en *Proceedings of the International Workshop on Sharable Natural Language Resources (SNRL)*, págs. 71–77, Ikoma, Nara, Japan.

AONE, CHINATSU y SCOTT WILLIAM BENNETT (1995). «Evaluating automated and manual acquisition of anaphora resolution strategies», en Morgan Kaufmann Publishers, editor, *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, págs. 122–129, Cambridge, Massachusetts.

AONE, CHINATSU y SCOTT WILLIAM BENNETT (1996). «Applying machine learning to anaphora resolution», en Stefan Wermter, Ellen Riloff y Gabriele Scheler, editores, *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing. IJCAI '95 Workshop. 1995 Proceedings*, vol. 1040 de *Lecture Notes in Computer Science*, cap. Symbolic Approaches, págs. 302–314, Springer Verlag, Berlin, Germany.

AONE, CHINATSU y DOUGLAS MCKEE (1993). «A language-independent anaphora resolution system for understanding multilingual texts», en Association for Computational Linguistics, editor, *Proceedings of the 31nd Annual Meeting of the Association for Computational Linguistics (ACL'93)*, págs. 156–

163, Ohio State University, Columbus, Ohio, USA.

Aoun, Joseph E. (1981). *The Formal Nature of Anaphoric Relations*, Tesis Doctoral, Massachusetts Institute of Technology, Massachussetts, USA.

Atserias, Jordi, Josep Carmona, Irene Castellón, Sergi Cervell, Montse Civit, Lluis Màrquez, M. Antonia Martí, Lluis Padró, Roberto Placer, Horacio Rodríguez, Mariona Taulé y Jordi Turmo (1998). «Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text», en *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Azzam, Saliha, Kevin Humphreys y Robert Gaizauskas (1998a). «Coreference Resolution in a Multilingual Information Extraction System», en *Proceedings of the Workshop on Linguistic Coreference. First Language Resources and Evaluation Conference (LREC'98).*, págs. 74–78, Granada, Spain.

Azzam, Saliha, Kevin Humphreys y Robert Gaizauskas (1998b). «Evaluating a Focus-Based Approach to Anaphora Resolution», en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, págs. 74–78, Montreal, Canada.

Baldwin, Breck (1997). «CogNIAC: high precision coreference with limited knowledge and linguistic resources», en *Proceedings of the ACL'97/EACL'97 workshop on Operational Factors in Practical, Robust Anaphor Resolution*, págs. 38–45, Madrid, Spain.

Barbu, Catalina y Ruslan Mitkov (2000). «Evaluation environment for anaphora resolution», en *Proceedings of the International Conference Machine Translation and Multilingual Applications in the New Millennium. (MT'2000)*, págs. 18.1–18.8, Exeter, UK.

Barbu, Catalina y Ruslan Mitkov (2001). «Evaluation tool for rule-based anaphora resolution methods», en *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001)*, págs. 34–41, Toulouse, France.

BENÍTEZ, LAURA, SERGI CERVELL, GERARD ESCUDERO, MÒNICA LÓPEZ, GERMAN RIGAU y MARIONA TAULÉ (1998). «Methods and tools for building the Catalan WordNet», en *Proceedings of the Workshop on Language Resources for European Minority Languages. First Language Resources and Evaluation Conference (LREC'98).*, Granada, Spain.

BOGURAEV, BRANIMIR (1979). «Automatic resolution of linguistic ambiguities», *Technical report TR-11*, University of Cambridge Computer Laboratory, Cambridge, Massachusetts, USA.

BRENNAN, S.E., M.W. FRIEDMAN y C.J. POLLARD (1987). «A centering approach to pronouns», en *Proceedings of the 25st Annual Meeting of the Association for Computational Linguistics (ACL'87)*, págs. 155–162, Stanford, California. USA.

BRUCART, JOSÉ M. (1999). «La elipsis», *Gramática descriptiva de la lengua española*, **2**, 2787–2863.

BUITELAAR, PAUL y BOGDAN SACALEANU (2002). «Extending Synsets with Medical Terms», en *Proceedings of the First International Conference on WordNets*, Mysore, India.

BYRON, DONNA K. y JAMES F. ALLEN (1999). «Applying Genetic Algorithms to Pronoun Resolution», en *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI'99)*, pág. 957, Orlando, Florida.

BYRON, DONNA K. y AMANDA STENT (1998). «A Preliminary Model of Centering in Dialog», en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, págs. 1475–1477, Montreal, Canada.

CARBONELL, JAIME G. y RALPH D. BROWN (1988). «Anaphora resolution: a multi-strategy approach», en *Proceedings of 12th International Conference on Computational Linguistics (COLING'88)*, págs. 96–101, Budapest, Hungary.

CARDIE, CLAIRE y DAVID PIERCE (1998). «Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification», en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, págs. 218–224, Montreal, Canada.

CARDIE, CLAIRE y KIRI WAGSTAFF (1999). «Noun Phrase Coreference as Clustering», en *Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, págs. 82–89, Maryland, USA.

CARLETTA, JEAN (1996). «Assessing agreement on classification task: the kappa statistic», *Computational Linguistics*, **22**(2), 249–254.

CARLETTA, JEAN, AMY ISARD, STEPHEN ISARD, JACQUELINE C. KOWTKO, GWYNETH DOHERTY-SNEDDON y ANNE H. ANDERSON (1997). «The Reliability of a Dialogue Structure Coding Scheme», *Computational Linguistics*, **23**(1), 13–32.

CARTER, DAVID M. (1986). *A shallow processing approach to anaphor resolution*, Tesis Doctoral, University of Cambridge, Cambridge, Massachusetts, USA.

CARTER, DAVID M. (1987a). «Common sense inference in a focus-guided anaphor resolver», *Journal of Semantics*, **4**, 237–246.

CARTER, DAVID M. (1987b). *Interpreting anaphora in natural language texts*, Chichester: Ellis Horwood.

CHOMSKY, NOAM (1965). *Aspects of a Theory of Sintax*, MIT Press, Cambridge, Massachussetts, USA.

CHOMSKY, NOAM (1981). *Lectures on Government and Binding*, Foris Publications, Dordrecht, Holland.

CUTTING, DOUG, JULIAN KUPIEC, JAN PEDERSEN y PENELOPE SIBUN (1998). «A Practical Part-of-Speech Tagger», en *Proceedings of the Third Conference on Applied Natural Language Processing*, págs. 133–140, Trento, Italia.

DAGAN, IDO (1992). *Multilingual statistical approaches for natural language disambiguation*, Tesis Doctoral, Israel Institute of Technology, Haifa, Israel.

DAGAN, IDO y ALON ITAI (1990). «Automatic processing of large corpora for the resolution of anaphora references», en *Proceedings of 13th International Conference on Computational Linguistics (COLING'90)*, págs. 330–332, Helsinki, Finland.

DAGAN, IDO y ALON ITAI (1991). «A statistical filter for resolving pronoun references», *Artificial Intelligence and Computer Vision*, págs. 125–135.

DAGAN, IDO, JOHN JUSTESON, SHALOM LAPPIN, HERBERT LEASS y AMNON RIBAK (1995). «Syntax and lexical statistics in anaphora resolution», *Applied Artificial Intelligence*, **9**, 633–644.

DAUDÉ, JORDI, LLUIS PADRÓ y GERMAN RIGAU (2001). «A Complete WN1.5 to WN1.6 Mapping», en *Proceedings of the NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customisations*, págs. 83–88, Carnegie Mellon University, Pittsburgh, USA.

DUCROT, OSWALD y JEAN-MARIE SCHAFFER (1998). *Nuevo diccionario enciclopédico de las ciencias del lenguaje*, Arrecife, Madrid, España.

ECKERT, MIRIAM y MICHAEL STRUBE (2001). «Dialogue acts, synchronising units and anaphora resolution», *Journal of Semantics*, **17**(1), 51–89.

FELLBAUM, CHRISTIANE (1998). *WordNet, an electronic lexical database*, MIT Press.

FERNÁNDEZ, OLGA (1999). «El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos», *Gramática descriptiva de la lengua española*, **1**, 1209–1273.

FERRÁNDEZ, ANTONIO (1998). *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos*, Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, España.

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR, PATRICIO MARTÍNEZ-BARCO, JESÚS PERAL, RAFAEL MUÑOZ y MAXIMILIANO SAIZ-NOEDA (1999). «Sistema de procesamiento del lenguaje natural orientado a la resolución de la correferencia lingüística», *Procesamiento del Lenguaje Natural*, **25**, 217–218.

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR y LIDIA MORENO (1997). «Slot Unification Grammar», en *Proceedings of the Joint Conference on Declarative Programming. APPIA-GULP-PRODE*, págs. 523–532, Grado, Italy.

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR y LIDIA MORENO (1998). «Anaphora resolution in unrestricted texts with partial parsing», en *Proceedings of the 36th Annual Meeting of the As-*

sociation for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98), págs. 385–391, Montreal, Canada.

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR y LIDIA MORENO (1999). «An empirical approach to Spanish anaphora resolution», *Machine Traslation*, **14**(3/4), 191–216.

FERRÁNDEZ, ANTONIO y JESÚS PERAL (2000). «A computational approach to zero-pronouns in Spanish», en *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, págs. 166–172, Hong Kong, China.

FERRÁNDEZ, ANTONIO, JESÚS PERAL, PATRICIO MARTÍNEZ-BARCO, MAXIMILIANO SAIZ-NOEDA y RAFAEL ROMERO (1997). «Resolución de la extraposición a izquierdas con las gramáticas de unificación de huecos», *Procesamiento del Lenguaje Natural*, **21**, 167–182.

FOX, BARBARA (1987). *Discourse Structure and Anaphora. Written and conversational English*, Cambridge Studies in Linguistics, Cambridge University Press, Cambridge, Massachusetts, USA.

GAIZAUSKAS, ROBERT J. y KEVIN HUMPHREYS (1996). «Quantitative Evaluation of Coreference Algorithms in an Information Extraction System», en *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC'96)*, Lancaster, UK.

GE, NIYU (2000). *An approach to anaphoric pronouns*, Tesis Doctoral, Department of Computer Sicence. Brown University, Providence. Rhode Island. USA.

GE, NIYU, JOHN HALE y EUGENE CHARNIAK (1998). «A statistical approach to anaphora resolution», en Eugene Charniak, editor, *Proceedings of Sixth WorkShop on Very Large Corpora*, págs. 161–170, Montreal, Canada.

GILI-GAYA, SAMUEL (1961). *Curso superior de sintaxis española*, Vox, Barcelona, España.

GROSZ, BARBARA, ARAVIND JOSHI y SCOTT WEINSTEIN (1983). «Providing a unified account of definite noun phrases in discourse», en *Proceedings of the 21st Annual Meeting of the*

Association for Computational Linguistics (ACL'83), págs. 44–50, Cambridge, Massachusetts. USA.

GROSZ, BARBARA, ARAVIND JOSHI y SCOTT WEINSTEIN (1995). «Centering: a framework for modeling the local coherence of discourse», *Computational Linguistics*, **21**(2), 203–225.

GUENTHNER, FRANZ y HUBERT LEHMANN (1983). «Rules for pronominalization», en *Proceedings of the First Conference of the European Chapter of the Association for Computational Linguistics (EACL'83)*, págs. 144–151, Pisa, Italy.

HAEGEMAN, LILIANE (1994). *Introduction to Government and Binding Theory*, cap. Anaphoric Relations and Overt NPs, págs. 201–249, Blackwell Publishers, Oxford, UK.

HALLIDAY, MICHAEL A.K. y RUQAIYA HASSAN (1976). *Cohesion in English*, Longman, London, UK.

HARABAGIU, SANDA y STEPHEN MAIORANO (1999). «Knowledge-lean coreference resolution and its relation to textual cohesion and coreference», en Dan Cristea, Nancy Ide y Daniel Marcu, editores, *The Relation of Discourse/Dialogue Structure and Reference*, págs. 29–38, Association for Computational Linguistics, New Brunswick, New Jersey.

HARABAGIU, SANDA M. y STEVEN J. MAIORANO (2000). «Multilingual Coreference Resolution», en *Proceedings of the Language Technology Joint Conference on Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL'2000)*, págs. 142–149, Seattle, WA.

HEARST, MARTI A. (1994). «Multi-Paragraph segmentation of expository text», en Association for Computational Linguistics, editor, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, págs. 9–16, Las Cruces, New Mexico.

HERNANZ, Mª. LLUÏSA y JOSÉ Mª. BRUCART (1987). *La sintaxis. Principios teóricos. La oración simple*, Enseñanza/crítica. "Textos", Editorial Crítica, Barcelona, España.

HOBBS, JERRY R. (1976). «Pronoun resolution», *Research report # 76-1*, Department of Computer Sciences. City College. City University of New York, New York, USA.

HOBBS, JERRY R. (1978). «Resolving pronoun references», *Lingua*, **44**, 311–338.

HOBBS, JERRY R. (1986). «Resolving pronoun references», en Barbara J. Grosz, Karen Sparck Jones y Bonnie Lynn Webber, editores, *Readings in Natural Language Processing*, págs. 339–352, Morgan Kaufmann Publishers, Inc., Los Altos, California, 1978 paper reprint.

HOCKETT, CHARLES F. (1971). *Curso de lingüística general*, Eudeba, Buenos Aires, Argentina.

HOPPENBROUWERS, JEROEN (2001). «Requirements of the Terminology Alignment System», *EuroTerm EDC-2214 Technical Report D.3.2*, Infolab, CentER Applied Research, Tilburg, Netherlands.

KAMEYAMA, MEGUMI (1997a). «Intrasentential Centering: A case study», en *Centering Theory in Discourse*, págs. 89–112, Oxford University Press, Oxford, UK.

KAMEYAMA, MEGUMI (1997b). «Recognizing Referential Links: An Information Extraction Perspective», en *Proceedings of the ACL'97/EACL'97 workshop on Operational Factors in Practical, Robust Anaphor Resolution*, págs. 46–53, Madrid, Spain.

KAMP, HANS (1981). «A theory of truth and semantic representation», en *Formal methods in the study of language*, págs. 277–322, Mathematical centre. Tracts, Amsterdam, Netherlands.

KENNEDY, CHRISTOPHER y BRANIMIR BOGURAEV (1996). «Anaphora for everyone: pronominal anaphora resolution without a parser», en *Proceedings of 16th International Conference on Computational Linguistics*, vol. I, págs. 113–118, Copenhagen, Denmark.

LANGACKER, R. (1969). «On pronominalisation and the chain of command», en *Modern studies in English*, págs. 160–186, Prentice-Hall, Englewood Cliffs, New Jersey, USA.

LAPPIN, SHALOM y HERBERT LEASS (1994). «An algorithm for pronominal anaphora resolution», *Computational Linguistics*, **20**(4), 535–561.

LYONS, JOHN (1971). *Introducción en la lingüística teórica*, Teide, Barcelona. España.

MAGNINI, BERNARDO y GABRIELA CAVAGLIA (2000). «Integrating subject field codes into WordNet», en Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. y Stainhaouer G., editores, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, págs. 1413–1418, Athens, Greece.

MAHESH, KAVI y SERGEI NIRENBURG (1995). «A Situated Ontology for Practical NLP», en *Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing. International Joint Conference on Artificial Intelligence (IJCAI'95)*, Montreal, Canada.

MARCUS, MITCHELL P., BEATRICE SANTORINI y MARY ANN MARCINKIEWICZ (1993). «Building a large annotated corpus of English: the Penn Treebank», *Computational Linguistics*, **19**(2), 313–330.

MARTÍ, M.ANTONIA, HORACIO RODRÍGUEZ y J. SERRANO (1998). «Declaración de categorías morfosintácticas», Proyecto ITEM. Doc. núm. 2. http://sensei.ieec.uned.es/item (página visitada el 17/04/01).

MARTÍNEZ-BARCO, PATRICIO (2001). *Resolución computacional de la anáfora en diálogos: estructura del discurso y conocimiento lingüístico*, Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante, Alicante, España.

MCCORD, MICHAEL (1990). «Slot grammar: a system for simpler construction of practical natural language grammars», en Rudi Studer, editor, *Natural Language and Logic: International Scientific Symposium*, vol. 459 de *Lecture Notes in Computer Science*, págs. 118–145, Springer Verlag, Hamburg, Germany.

MCCORD, MICHAEL (1993). «Heuristics for broad-coverage natural language parsing», en Morgan Kaufmann, editor, *Proceedings of the ARPA Workshop on Human Language Technology*, Princeton, New Jersey.

MILLER, GEORGE A., RICHARD BECKWITH, CHRISTIANE FELLBAUM, DEREK GROSS y KATHERINE J. MILLER (1993). «Five Papers on WordNet», *Special Issue of the International Journal of Lexicography*, **3**(4), 235–312.

MITKOV, RUSLAN (1994). «An integrated model for anaphora resolution», en *Proceedings of 15th International Conference on Computational Linguistics (COLING'94)*, vol. III, págs. 1170–1176, Kioto, Japan.

MITKOV, RUSLAN (1996). «Anaphora resolution: a combination of linguistic and statistical approaches», en *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC'96)*, Lancaster, UK.

MITKOV, RUSLAN (1998). «Robust pronoun resolution with limited knowledge», en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, págs. 869–875, Montreal, Canada.

MITKOV, RUSLAN (2001). «Outstanding issues in anaphora resolution», en Alexander Gelbukh, editor, *Proceedings of the Second International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2001).*, Lectures Notes In Computer Science, págs. 110–125, Springer Verlag, Mexico City, Mexico.

MITKOV, RUSLAN (2002). *Anaphora resolution*, Longman, London. UK.

MOESHLER, JACKES y ANNE REBOUL (1991). *Dictionnaire Encyclopédique de Pragmatique*, Éditions du Seuil, París, France.

MOLINA, ANTONIO, FERRAN PLA, ENCARNA SEGARRA y LIDIA MORENO (2002). «Word Sense Disambiguation using Statistical Models and WordNet», en *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Canary Islands, Spain.

MONTOYO, A. y M. PALOMAR (2001). «Specification Marks for Word Sense Disambiguation: New Development», en Alexander Gelbukh, editor, *Proceedings of the Second International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2001).*, Lectures Notes In Computer Science, págs. 182–191, Springer Verlag, Mexico City, Mexico.

MONTOYO, ANDRÉS y MANUEL PALOMAR (2000). «Word Sense Disambiguation with Specification Marks in Unrestricted

Texts», en *Proceedings 11th International Conference on Database and Expert Systems Applications (DEXA'2000)*, págs. 103–107, Greenwich, London, UK.

MONTOYO, ANDRÉS (2002). *Desambiguación Léxica mediante Marcas de Especificidad*, Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante.

MORENO, LIDIA (1993). *Formalismos Lógicos para el Análisis e Interpretación oracional del Lenguaje Natural*, Tesis Doctoral, Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia, Valencia, España.

MORENO, LIDIA, FRANCISCO ANDRÉS y MANUEL PALOMAR (1991). «Incorporar Restricciones Semánticas en el Análisis Sintáctico: IRSAS», *Procesamiento del Lenguaje Natural*, **11**, 75–88.

MORENO, LIDIA, MANUEL PALOMAR, ANTONIO MOLINA y ANTONIO FERRÁNDEZ (1999). *Introducción al Procesamiento del Lenguaje Natural*, Servicio de Publicaciones de la Universidad de Alicante, Alicante, España.

MORENO-CABRERA, JUAN C. (1991). *Curso universitario de lingüística general*, vol. 1, Síntesis, Madrid, España.

MUC-6 (1995). *Sixth Message Understanding Conference*, Columbia, Maryland, USA.

MUÑOZ, RAFAEL (2001). *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, Alicante, España.

MUÑOZ, RAFAEL, RUSLAN MITKOV, MANUEL PALOMAR, JESÚS PERAL, RICHARD EVANS, LIDIA MORENO, CONSTANTIN ORASAN, MAXIMILIANO SAIZ-NOEDA, ANTONIO FERRÁNDEZ, CATALINA BARBÚ, PATRICIO MARTÍNEZ-BARCO y ARMANDO SUÁREZ (2002a). «Bilingual Alignment of Anaphoric Expressions», en *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Canary Islands, Spain.

MUÑOZ, RAFAEL y MANUEL PALOMAR (2001). «Semantic-driven Algorithm for Definite Description Resolution », en *Pro-*

ceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'2001), págs. 180–186, Tzigov Chark, Bulgaria.

MUÑOZ, RAFAEL, MAXIMILIANO SAIZ-NOEDA y ANDRÉS MONTOYO (2002b). «Semantic Information in Anaphora Resolution», en *Proceedings of the Portugal for Natural Language Processing (PorTAL'2002)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, págs. 63–70, Algarve, Faro, Portugal.

MUÑOZ, RAFAEL, MAXIMILIANO SAIZ-NOEDA, ARMANDO SUÁREZ y MANUEL PALOMAR (2000). «Semantic Approach to Bridging Reference Resolution», en *Proceedings of the International Conference Machine Translation and Multilingual Applications in the New Millennium. (MT'2000)*, págs. 17.1–17.8, Exeter, UK.

NASUKAWA, TETSUYA (1994). «Robust method of pronoun resolution using full-text information», en *Proceedings of 15th International Conference on Computational Linguistics (COLING'94)*, vol. III, págs. 1157–1163, Kioto, Japan.

PADRÓ, LLUIS (1997). *A Hybrid Environment for Syntax-Semantic Tagging*, Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos. Universidad Politécnica de Cataluña, Barcelona, Spain.

PALOMAR, MANUEL (1996). *Aportaciones a la resolución de la elipsis en lenguaje natural utilizando técnicas incrementales*, Tesis Doctoral, Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia, Valencia, España.

PALOMAR, MANUEL, ANTONIO FERRÁNDEZ, LIDIA MORENO, PATRICIO MARTÍNEZ-BARCO, JESÚS PERAL, MAXIMILIANO SAIZ-NOEDA y RAFAEL MUÑOZ (2001a). «An algorithm for Anaphora Resolution in Spanish Texts», *Computational Linguistics*, **27**(4), 545–567.

PALOMAR, MANUEL, ANTONIO FERRÁNDEZ, LIDIA MORENO, MAXIMILIANO SAIZ-NOEDA, RAFAEL MUÑOZ, PATRICIO MARTÍNEZ-BARCO, JESÚS PERAL y BORJA NAVARRO (1999). «A Robust Partial Parsing Strategy based on the Slot Unification Grammars», en *Proceeding of the Sixth Conference on Na-*

tural Language Processing (TALN'99), págs. 263–272, Corsica, France.

PALOMAR, MANUEL y PATRICIO MARTÍNEZ-BARCO (2001). «Computational approach to anaphora resolution in Spanish dialogues», *Journal of Artificial Intelligence Research*, **15**, 263–287.

PALOMAR, MANUEL, MAXIMILIANO SAIZ-NOEDA, RAFAEL MUÑOZ, ARMANDO SUÁREZ y PATRICIO MARTÍNEZ-BARCO (2000). «PHORA: A system to solve the Anaphora in Spanish», en *Proceedings of Third Colloquium on Discurse Anaphora and Anaphor Resolution (DAARC'2000)*, págs. 206–211, Lancaster, UK.

PALOMAR, MANUEL, MAXIMILIANO SAIZ-NOEDA, RAFAEL MUÑOZ, ARMANDO SUÁREZ, PATRICIO MARTÍNEZ-BARCO y ANDRÉS MONTOYO (2001b). «PHORA: A NLP system for Spanish», en Alexander Gelbukh, editor, *Proceedings of the Second International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2001).*, Lectures Notes In Computer Science, págs. 128–139, Springer Verlag, Mexico City, Mexico.

PERAL, JESÚS (2001). *Resolución y generación de la anáfora pronominal en español e inglés en un sistema interlingua de traducción automática*, Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante, Alicante, España.

PERAL, JESÚS, MAXIMILIANO SAIZ-NOEDA, ANTONIO FERRÁNDEZ y MANUEL PALOMAR (1999). «Anaphora resolution and generation in a multilingual system. An interlingua mechanism», en *Proceedings of the Venezia per il Trattamento Automatico delle Lingue (VEXTAL'99)*, págs. 315–324, Venice, Italy.

PLA, FERRAN (2000). *Etiquetado léxico y análisis sintáctico superficial basado en modelos estadísticos*, Tesis Doctoral, Departamento de Sistemas Informáticos y Computación. Universidad de Politécnica de Valencia, Valencia, España.

PLA, FERRAN y ANTONIO MOLINA (2001). «Part-of Speech Tagging with Lexicalized HMM», en *Proceedings of the Internatio-*

nal Conference on Recent Advances in Natural Language Processing (RANLP'2001), Tzigov Chark, Bulgaria.

REINHART, TANYA (1983). *Anaphora and Semantic Interpretation*, Croom Helm linguistics series, Croom Helm Ltd, London, UK.

RICH, ELAINE y SUSAN LUPERFOY (1998). «An Architecture for Anaphora Resolution», en *Proceedings of the Second Conference on Applied Natural Language Processing*, págs. 18–24, Austin, Texas.

RICO, CELIA (1994). *Aproximación estadístico-algebráica al problema de la resolución de la anáfora en el discurso*, Tesis Doctoral, Departamento de Filología Inglesa. Universidad de Alicante, Alicante, España.

RIGAU, GEMMA (1981). *Gramàtica del discurs*, Bellaterra : Universitat Autònoma de Barcelona, Barcelona, España.

SAIZ-NOEDA, MAXIMILIANO y MANUEL PALOMAR (2000). «Semantic Knowledge-driven Method to Solve Pronominal Anaphora in Spanish», en *NLP'2000 Filling the gap between theory and practice*, Lecture Notes In Artificial Intelligence. Springer-Verlag, págs. 204–211, Patras, Greece.

SAIZ-NOEDA, MAXIMILIANO, MANUEL PALOMAR y DAVID FARWELL (2000a). «NLP system oriented to anaphora resolution», en *Proceedings of the International Conference Machine Translation and Multilingual Applications in the New Millennium. (MT'2000)*, págs. 19.1–19.7, Exeter, UK.

SAIZ-NOEDA, MAXIMILIANO, MANUEL PALOMAR y LIDIA MORENO (2001a). «Pronoun Resolution in Spanish from Full Parsing», en *Proceedings of the International Conference on Text, Speech and Dialogue (TSD'2001)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, págs. 84–91, Zelezna Ruda, República Checa.

SAIZ-NOEDA, MAXIMILIANO, JESÚS PERAL y ARMANDO SUÁREZ (2000b). «Semantic Compatibility Techniques for Anaphora Resolution», en *Proceedings of International Conference on Artificial and Computational Intelligence For Decision, Control and Automation In Engineering and Industrial Applications (ACIDCA'2000)*, págs. 43–48, Monastir, Tunisia.

SAIZ-NOEDA, MAXIMILIANO, ARMANDO SUAREZ y MANUEL PALOMAR (2001b). «Semantic pattern learning through Maximum Entropy-based WSD technique», en *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL'2001)*, págs. 23–29, Toulouse, France.

SAIZ-NOEDA, MAXIMILIANO, ARMANDO SUÁREZ y JESÚS PERAL (1999). «Propuesta de incorporación de información semántica desde WordNet al análisis sintáctico parcial orientado a la resolución de la anáfora», *Procesamiento del Lenguaje Natural*, **25**, 167–173.

SALTON, GERARD y CHRIS BUCKLEY (1988). «Term Weighting Approaches in Automatic Text Retrieval», *Information Processing and Management*, **24**(5), 513–523.

SIDNER, CANDACE (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*, Tesis Doctoral, Massachusetts Institute of Technology, Cambridge, Massachussetts, USA.

SIDNER, CANDACE (1983). *Focusing in the comprehension of definite anaphora*, págs. 267–330, MIT Press, Cambridge, Massachussetts, USA, publicado también en Grosz, B., Jones, K.S. and Webber, B. (Eds), Readings in Natural Language Processing. Morgan Kaufmann Publishers, Inc. (1986).

STAMOU, SOFIA, ALEXANDROS NTOULAS, JEROEN HOPPENBROUWERS, MAXIMILIANO SAIZ-NOEDA y DIMITRIS CHRISTODOULAKIS (2002a). «EUROTERM: Extending EuroWordNet using both the expand and merge model», en *Proceedings of the First International Conference on WordNets*, Mysore, India.

STAMOU, SOFIA, KEMAL OFLAZER, PALA KAREL, DIMITRIS CHRISTODOULAKIS, DAN CRISTEA, DAN TUFIS, SVETLA KOEVA, GEORGE TOTKOV, DOMINIQUE DUTOIT y MARIA GRIGORIADOU (2002b). «BALKANET: A Multilingual Semantic Network for Balkan Languages», en *Proceedings of the First International Conference on WordNets*, Mysore, India.

STRUBE, MICHAEL (1998). «Never Look Back: An Alternative to Centering», en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Inter-*

national Conference on Computational Linguistics (COLING-ACL'98), págs. 1251–1257, Montreal, Canada.

STRUBE, MICHAEL y UDO HAHN (1999). «Functional Centering - Grounding Referential Coherence in Information Structure», *Computational Linguistics*, **25**(5), 309–344.

SUÁREZ, ARMANDO y MANUEL PALOMAR (2002). «Feature Selection Analysis for Maximum Entropy-based WSD», en Alexander Gelbukh, editor, *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2002).*, Lectures Notes In Computer Science, págs. 146–155, Springer Verlag, Mexico City, Mexico.

SUÁREZ, ARMANDO, MAXIMILIANO SAIZ-NOEDA y MANUEL PALOMAR (1999). «A method of restricted knowledge acquisition from WordNet», en *Proceedings of the Third International Conference on Knowledge-based Intelligent Information Engineering Systems (KES'99)*, págs. 38–41, Adelaide, Australia.

TAPANAINEN, PASI y TIMO JÄRVINEN (1997). «A non-projective dependency parser», en *Proceedings of the Fifth Conference on Applied Natural Language Processing*, págs. 64–71, Washington DC, USA.

TETREAULT, JOEL R. (1999). «Analysis of Syntax-Based Pronoun Resolution Methods», en *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, págs. 602–605, Maryland, USA.

VICEDO, JOSÉ LUIS y ANTONIO FERRÁNDEZ (2000). «Importance of Pronominal Anaphora resolution in Question Answering systems», en *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, págs. 555–562, Hong Kong, China.

VIEIRA, RENATA y MASSIMO POESIO (2000). «An Empirically-Based System for Processing Definite Descriptions», *Computational Linguistics*, **26**(4), 539–593.

VOSSEN, PIEK (1996). «Right or Wrong: Combining Lexical Resources in the EuroWordNet Project», en *Proceedings of the 7th Euralex International Congress on Lexicography*, págs. 715–728, Göteborg, Sweden.

VOSSEN, PIEK (1998). «EuroWordNet: Building a Multilingual Database with WordNets for European Languages», *The ELRA Newsletter*, **3**(1).

VOSSEN, PIEK (2000). «EuroWordNet: a Multilingual Database with WordNets in 8 languages», *The ELRA Newsletter*, **5**(1), 9–10.

VOSSEN, PIEK, LAURA BLOKSMA, WIM PETERS, CLAUDIA KUNZE, ANDREAS WAGNER, KAREL PALA, KADRI VIDER y FRANCESCA BERTAGNA (1999). «Extending the Inter-Lingual-Index with new Concepts», *Deliverable 2D010*, EuroWordNet, LE2-4003 TR-11.

VOSSEN, PIEK, LAURA BLOKSMA, HORACIO RODRÍGUEZ, SALVADOR CLIMENT, NICOLETTA CALZOLARI, ADRIANA ROVENTINI, FRANCESCA BERTAGNA, ANTONIETTA ALONGE y WIM PETERS (1998). «The EuroWordNet Base Concepts and Top Ontology», *Deliverable D017, D034, D036, WP5*, EuroWordNet, LE2-4003 TR-11.

WALKER, MARILYN A. (1998). *Centering, anaphora resolution and discourse structure*, cap. 4, Oxford University Press, Oxford, UK.

WILKS, YORICK (1975). *Preference semantics*, págs. 329–348, Cambridge University Press, Cambridge.

ZAVREL, JAKUB y WALTER DAELEMANS (1999). «Recent Advances in Memory-Based Part-Of-Speech Tagging», en *Actas del VI Simposio Internacional de Comunicación Social*, págs. 590–597, Centro de Lingüística Aplicada, Santiago de Cuba.

A. Resultados de la evaluación

Las siguientes páginas presentan los datos obtenidos en la evaluación del método ERA según los diferentes experimentos realizados.

Para cada uno de los experimentos, comentados en el apartado 5.3.3 (pág. 167), se detalla en cada cuadro el conjunto de pruebas realizadas, así como los resultados parciales para cada bloque del corpus y los resultados totales obtenidos. Estos resultados han servido como base fundamental de interpretación sobre la influencia de cada fuente de información realizada en los apartados del 5.3.4 al 5.3.9 (págs. 176–199).

A.1 Experimento 1. Estudio de las restricciones

A.1.1 Adición de restricciones

BASE de ADICIÓN: el más cercano

L009	Anaf	OK	
Omitidos	13	7	53,85%
Personales	14	8	57,14%
Demostr.			
Reflexivos	4	1	25,00%
	31	16	51,61%

L065	Anaf	OK	
Omitidos	39	7	17,95%
Personales	29	3	10,34%
Demostr.			
Reflexivos	4	3	75,00%
	72	13	18,06%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	1	10,00%
Demostr.	3	0	0,00%
Reflexivos	2	1	50,00%
	18	3	16,67%

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Sólo Morfológicas

L009	Anaf	OK	
Omitidos	13	9	69,23%
Personales	14	12	85,71%
Demostr.			
Reflexivos	4	2	50,00%
	31	23	74,19%

L065	Anaf	OK	
Omitidos	39	10	25,64%
Personales	29	11	37,93%
Demostr.			
Reflexivos	4	4	100,00%
	72	25	34,72%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	2	20,00%
Demostr.	3	2	66,67%
Reflexivos	2	1	50,00%
	18	6	33,33%

TOTAL	Anaf	OK	
Omitidos	55	20	36,36%
Personales	53	25	47,17%
Demostr.	3	2	66,67%
Reflexivos	10	7	70,00%
	121	54	44,63%

Sólo Morfosemánticas

L009	Anaf	OK	
Omitidos	13	9	69,23%
Personales	14	12	85,71%
Demostr.			
Reflexivos	4	2	50,00%
	31	23	74,19%

L065	Anaf	OK	
Omitidos	39	10	25,64%
Personales	29	10	34,48%
Demostr.			
Reflexivos	4	4	100,00%
	72	24	33,33%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	2	20,00%
Demostr.	3	2	66,67%
Reflexivos	2	1	50,00%
	18	6	33,33%

TOTAL	Anaf	OK	
Omitidos	55	20	36,36%
Personales	53	24	45,28%
Demostr.	3	2	66,67%
Reflexivos	10	7	70,00%
	121	53	43,80%

Sólo Sintáctico-Semánticas

L009	Anaf	OK	
Omitidos	13	7	53,85%
Personales	14	10	71,43%
Demostr.			
Reflexivos	4	1	25,00%
	31	18	58,06%

L065	Anaf	OK	
Omitidos	39	7	17,95%
Personales	29	4	13,79%
Demostr.			
Reflexivos	4	3	75,00%
	72	14	19,44%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	3	30,00%
Demostr.	3	0	0,00%
Reflexivos	2	1	50,00%
	18	5	27,78%

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	17	32,08%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	37	30,58%

Sólo Sintácticas

L009	Anaf	OK	
Omitidos	13	7	53,85%
Personales	14	8	57,14%
Demostr.			
Reflexivos	4	4	100,00%
	31	19	61,29%

L065	Anaf	OK	
Omitidos	39	8	20,51%
Personales	29	6	20,69%
Demostr.			
Reflexivos	4	4	100,00%
	72	18	25,00%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	1	10,00%
Demostr.	3	0	0,00%
Reflexivos	2	2	100,00%
	18	4	22,22%

TOTAL	Anaf	OK	
Omitidos	55	16	29,09%
Personales	53	15	28,30%
Demostr.	3	0	0,00%
Reflexivos	10	10	100,00%
	121	41	33,88%

Sólo Semánticas

L009	Anaf	OK	
Omitidos	13	7	53,85%
Personales	14	9	64,29%
Demostr.			
Reflexivos	4	1	25,00%
	31	17	54,84%

L065	Anaf	OK	
Omitidos	39	8	20,51%
Personales	29	3	10,34%
Demostr.			
Reflexivos	4	3	75,00%
	72	14	19,44%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	2	20,00%
Demostr.	3	1	33,33%
Reflexivos	2	1	50,00%
	18	5	27,78%

TOTAL	Anaf	OK	
Omitidos	55	16	29,09%
Personales	53	14	26,42%
Demostr.	3	1	33,33%
Reflexivos	10	5	50,00%
	121	36	29,75%

A.1.2 Supresión de restricciones

BASE de SUPRESIÓN: todas las restricciones

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	26	**83,87%**

L065	Anaf	OK	
Omitidos	39	14	**35,90%**
Personales	29	16	**55,17%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	34	**47,22%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	5	**50,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	11	**61,11%**

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**71**	**58,68%**

TODAS sin Morfológicas ni Morfosemánticas

L009	Anaf	OK	
Omitidos	13	7	**53,85%**
Personales	14	10	**71,43%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	21	**67,74%**

L065	Anaf	OK	
Omitidos	39	9	**23,08%**
Personales	29	7	**24,14%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	20	**27,78%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	4	**40,00%**
Demostr.	3	1	**33,33%**
Reflexivos	2	2	**100,00%**
	18	8	**44,44%**

TOTAL	Anaf	OK	
Omitidos	55	17	**30,91%**
Personales	53	21	**39,62%**
Demostr.	3	1	**33,33%**
Reflexivos	10	10	**100,00%**
	121	**49**	**40,50%**

TODAS sin SintacticoSemánticas

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	26	**83,87%**

L065	Anaf	OK	
Omitidos	39	14	**35,90%**
Personales	29	16	**55,17%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	34	**47,22%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	3	**30,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	9	**50,00%**

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	32	**60,38%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**69**	**57,02%**

TODAS sin Sintacticas

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	2	**50,00%**
	31	24	**77,42%**

L065	Anaf	OK	
Omitidos	39	13	**33,33%**
Personales	29	13	**44,83%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	30	**41,67%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	5	**50,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	1	**50,00%**
	18	10	**55,56%**

TOTAL	Anaf	OK	
Omitidos	55	23	**41,82%**
Personales	53	31	**58,49%**
Demostr.	3	3	**100,00%**
Reflexivos	10	7	**70,00%**
	121	**64**	**52,89%**

TODAS sin Semánticas

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	26	**83,87%**

L065	Anaf	OK	
Omitidos	39	11	**28,21%**
Personales	29	15	**51,72%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	30	**41,67%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	4	**40,00%**
Demostr.	3	2	**66,67%**
Reflexivos	2	2	**100,00%**
	18	9	**50,00%**

TOTAL	Anaf	OK	
Omitidos	55	21	**38,18%**
Personales	53	32	**60,38%**
Demostr.	3	2	**66,67%**
Reflexivos	10	10	**100,00%**
	121	**65**	**53,72%**

A.2 Experimento 2. Estudio de las preferencias

A.2.1 Adición de preferencias

BASE de ADICIÓN: todas las restricciones

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	26	**83,87%**

L065	Anaf	OK	
Omitidos	39	14	**35,90%**
Personales	29	16	**55,17%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	34	**47,22%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	5	**50,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	11	**61,11%**

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**71**	**58,68%**

Sólo morfológicas

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	26	**83,87%**

L065	Anaf	OK	
Omitidos	39	14	**35,90%**
Personales	29	17	**58,62%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	35	**48,61%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	5	**50,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	11	**61,11%**

TOTAL	Anaf	OK	
Omitidos	55	24	**43,64%**
Personales	53	35	**66,04%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**72**	**59,50%**

Sólo sintácticas

L009	Anaf	OK	
Omitidos	13	9	**69,23%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	26	**83,87%**

L065	Anaf	OK	
Omitidos	39	36	**92,31%**
Personales	29	24	**82,76%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	64	**88,89%**

E001	Anaf	OK	
Omitidos	3	3	**100,00%**
Personales	10	7	**70,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	15	**83,33%**

TOTAL	Anaf	OK	
Omitidos	55	48	**87,27%**
Personales	53	44	**83,02%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**105**	**86,78%**

Sólo semánticas

L009	Anaf	OK	
Omitidos	13	8	**61,54%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	25	**80,65%**

L065	Anaf	OK	
Omitidos	39	18	**46,15%**
Personales	29	18	**62,07%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	40	**55,56%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	7	**70,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	13	**72,22%**

TOTAL	Anaf	OK	
Omitidos	55	27	**49,09%**
Personales	53	38	**71,70%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**78**	**64,46%**

Sólo estructurales

L009	Anaf	OK	
Omitidos	13	10	**76,92%**
Personales	14	13	**92,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	27	**87,10%**

L065	Anaf	OK	
Omitidos	39	17	**43,59%**
Personales	29	16	**55,17%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	37	**51,39%**

E001	Anaf	OK	
Omitidos	3	1	**33,33%**
Personales	10	5	**50,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	11	**61,11%**

TOTAL	Anaf	OK	
Omitidos	55	28	**50,91%**
Personales	53	34	**64,15%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**75**	**61,98%**

Sólo semánticas y semántico-estructurales

L009	Anaf	OK	
Omitidos	13	10	**76,92%**
Personales	14	14	**100,00%**
Demostr.			
Reflexivos	4	4	**100,00%**
	31	28	**90,32%**

L065	Anaf	OK	
Omitidos	39	27	**69,23%**
Personales	29	22	**75,86%**
Demostr.			
Reflexivos	4	4	**100,00%**
	72	53	**73,61%**

E001	Anaf	OK	
Omitidos	3	2	**66,67%**
Personales	10	7	**70,00%**
Demostr.	3	3	**100,00%**
Reflexivos	2	2	**100,00%**
	18	14	**77,78%**

TOTAL	Anaf	OK	
Omitidos	55	39	**70,91%**
Personales	53	43	**81,13%**
Demostr.	3	3	**100,00%**
Reflexivos	10	10	**100,00%**
	121	**95**	**78,51%**

A.2.2 Supresión de preferencias

BASE de SUPRESIÓN: todas las restricciones y preferencias

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

TODAS sin morfológicas

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

TODAS sin sintácticas

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	33	84,62%
Personales	29	21	72,41%
Demostr.			
Reflexivos	4	4	100,00%
	72	58	80,56%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	14	77,78%

TOTAL	Anaf	OK	
Omitidos	55	46	83,64%
Personales	53	42	79,25%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	101	83,47%

TODAS sin semánticas

L009	Anaf	OK	
Omitidos	13	12	92,31%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	30	96,77%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	53	96,36%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	112	92,56%

TODAS sin estructurales

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	32	82,05%
Personales	29	24	82,76%
Demostr.			
Reflexivos	4	4	100,00%
	72	60	83,33%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	46	83,64%
Personales	53	45	84,91%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	104	85,95%

TODAS sin semánticas ni semántico-estructurales

L009	Anaf	OK	
Omitidos	13	9	69,23%
Personales	14	13	92,86%
Demostr.			
Reflexivos	4	4	100,00%
	31	26	83,87%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	24	82,76%
Demostr.			
Reflexivos	4	4	100,00%
	72	66	91,67%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	6	60,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	14	77,78%

TOTAL	Anaf	OK	
Omitidos	55	50	90,91%
Personales	53	43	81,13%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	106	87,60%

A.3 Experimento 3. Estudio conjunto de restricciones y preferencias

A.3.1 Adición de restricciones y preferencias

BASE de ADICIÓN: el más cercano

L009	Anaf	OK	
Omitidos	13	7	53,85%
Personales	14	8	57,14%
Demostr.			
Reflexivos	4	1	25,00%
	31	16	51,61%

L065	Anaf	OK	
Omitidos	39	7	17,95%
Personales	29	3	10,34%
Demostr.			
Reflexivos	4	3	75,00%
	72	13	18,06%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	1	10,00%
Demostr.	3	0	0,00%
Reflexivos	2	1	50,00%
	18	3	16,67%

TOTAL	Anaf	OK	
Omitidos	55	15	27,27%
Personales	53	12	22,64%
Demostr.	3	0	0,00%
Reflexivos	10	5	50,00%
	121	32	26,45%

Sólo Restricciones y Preferencias Morfológicas

L009	Anaf	OK	
Omitidos	13	9	69,23%
Personales	14	12	85,71%
Demostr.			
Reflexivos	4	2	50,00%
	31	23	74,19%

L065	Anaf	OK	
Omitidos	39	10	25,64%
Personales	29	11	37,93%
Demostr.			
Reflexivos	4	4	100,00%
	72	25	34,72%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	2	20,00%
Demostr.	3	2	66,67%
Reflexivos	2	1	50,00%
	18	6	33,33%

TOTAL	Anaf	OK	
Omitidos	55	20	36,36%
Personales	53	25	47,17%
Demostr.	3	2	66,67%
Reflexivos	10	7	70,00%
	121	54	44,63%

Sólo Restricciones y Preferencias Sintácticas

L009	Anaf	OK	
Omitidos	13	8	61,54%
Personales	14	11	78,57%
Demostr.			
Reflexivos	4	4	100,00%
	31	23	74,19%

L065	Anaf	OK	
Omitidos	39	31	79,49%
Personales	29	16	55,17%
Demostr.			
Reflexivos	4	4	100,00%
	72	51	70,83%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	4	40,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	11	61,11%

TOTAL	Anaf	OK	
Omitidos	55	41	74,55%
Personales	53	31	58,49%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	85	70,25%

Sólo Restricciones y Preferencias Semánticas

L009	Anaf	OK	
Omitidos	13	6	46,15%
Personales	14	10	71,43%
Demostr.			
Reflexivos	4	1	25,00%
	31	17	54,84%

L065	Anaf	OK	
Omitidos	39	10	25,64%
Personales	29	3	10,34%
Demostr.			
Reflexivos	4	3	75,00%
	72	16	22,22%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	2	20,00%
Demostr.	3	1	33,33%
Reflexivos	2	1	50,00%
	18	5	27,78%

TOTAL	Anaf	OK	
Omitidos	55	17	30,91%
Personales	53	15	28,30%
Demostr.	3	1	33,33%
Reflexivos	10	5	50,00%
	121	38	31,40%

Sólo Restricciones y Preferencias Semánticas combinadas

L009	Anaf	OK	
Omitidos	13	10	76,92%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	0	0,00%
	31	24	77,42%

L065	Anaf	OK	
Omitidos	39	27	69,23%
Personales	29	18	62,07%
Demostr.			
Reflexivos	4	3	75,00%
	72	48	66,67%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	1	50,00%
	18	13	72,22%

TOTAL	Anaf	OK	
Omitidos	55	39	70,91%
Personales	53	39	73,58%
Demostr.	3	3	100,00%
Reflexivos	10	4	40,00%
	121	85	70,25%

Sólo Restricciones y Preferencias Sintácticas combinadas

L009	Anaf	OK	
Omitidos	13	8	61,54%
Personales	14	11	78,57%
Demostr.			
Reflexivos	4	4	100,00%
	31	23	74,19%

L065	Anaf	OK	
Omitidos	39	31	79,49%
Personales	29	17	58,62%
Demostr.			
Reflexivos	4	4	100,00%
	72	52	72,22%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	5	50,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	12	66,67%

TOTAL	Anaf	OK	
Omitidos	55	41	74,55%
Personales	53	33	62,26%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	87	71,90%

Sólo Restricciones y Preferencias Sintácticas y Semánticas Combinadas

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	32	82,05%
Personales	29	24	82,76%
Demostr.			
Reflexivos	4	4	100,00%
	72	60	83,33%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	46	83,64%
Personales	53	45	84,91%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	104	85,95%

A.3.2 Supresión de restricciones y preferencias

BASE de SUPRESIÓN: todas las Restricciones y Preferencias

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

TODAS sin Restricciones y Preferencias Morfológicas

L009	Anaf	OK	
Omitidos	13	9	69,23%
Personales	14	13	92,86%
Demostr.			
Reflexivos	4	4	100,00%
	31	26	83,87%

L065	Anaf	OK	
Omitidos	39	34	87,18%
Personales	29	21	72,41%
Demostr.			
Reflexivos	4	4	100,00%
	72	59	81,94%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	46	83,64%
Personales	53	41	77,36%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	100	82,64%

TODAS sin Restricciones y Preferencias Sintácticas

L009	Anaf	OK	
Omitidos	13	12	92,31%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	0	0,00%
	31	26	83,87%

L065	Anaf	OK	
Omitidos	39	30	76,92%
Personales	29	19	65,52%
Demostr.			
Reflexivos	4	4	100,00%
	72	53	73,61%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	1	50,00%
	18	13	72,22%

TOTAL	Anaf	OK	
Omitidos	55	44	80,00%
Personales	53	40	75,47%
Demostr.	3	3	100,00%
Reflexivos	10	5	50,00%
	121	92	76,03%

TODAS sin Restricciones y Preferencias Semánticas

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	6	60,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	13	72,22%

TOTAL	Anaf	OK	
Omitidos	55	51	92,73%
Personales	53	45	84,91%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	109	90,08%

TODAS sin Restricciones y Preferencias Semánticas Combinadas

L009	Anaf	OK	
Omitidos	13	9	69,23%
Personales	14	10	71,43%
Demostr.			
Reflexivos	4	4	100,00%
	31	23	74,19%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	22	75,86%
Demostr.			
Reflexivos	4	4	100,00%
	72	64	88,89%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	5	50,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	11	61,11%

TOTAL	Anaf	OK	
Omitidos	55	48	87,27%
Personales	53	37	69,81%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	98	80,99%

TODAS sin Restricciones y Preferencias Sintácticas combinadas

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	0	0,00%
	31	25	80,65%

L065	Anaf	OK	
Omitidos	39	30	76,92%
Personales	29	19	65,52%
Demostr.			
Reflexivos	4	4	100,00%
	72	53	73,61%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	1	50,00%
	18	13	72,22%

TOTAL	Anaf	OK	
Omitidos	55	43	78,18%
Personales	53	40	75,47%
Demostr.	3	3	100,00%
Reflexivos	10	5	50,00%
	121	91	75,21%

TODAS sin Restricciones y Preferencias Sintácticas y Semánticas combinadas

L009	Anaf	OK	
Omitidos	13	10	76,92%
Personales	14	12	85,71%
Demostr.			
Reflexivos	4	2	50,00%
	31	24	77,42%

L065	Anaf	OK	
Omitidos	39	14	35,90%
Personales	29	11	37,93%
Demostr.			
Reflexivos	4	4	100,00%
	72	29	40,28%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	2	20,00%
Demostr.	3	2	66,67%
Reflexivos	2	1	50,00%
	18	6	33,33%

TOTAL	Anaf	OK	
Omitidos	55	25	45,45%
Personales	53	25	47,17%
Demostr.	3	2	66,67%
Reflexivos	10	7	70,00%
	121	59	48,76%

A.4 Experimento 4. Estudio de la adquisición de patrones de compatibilidad

BASE 1: todas las Restricciones y Preferencias

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Aprendizaje: dos bloques. Resolución: el tercero

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	38	97,44%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	67	93,06%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	52	94,55%
Personales	53	46	86,79%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	111	91,74%

Adquisición: TODOS. Resolución TODOS.

L009	Anaf	OK	
Omitidos	13	11	84,62%
Personales	14	14	100,00%
Demostr.			
Reflexivos	4	4	100,00%
	31	29	93,55%

L065	Anaf	OK	
Omitidos	39	37	94,87%
Personales	29	27	93,10%
Demostr.			
Reflexivos	4	4	100,00%
	72	68	94,44%

E001	Anaf	OK	
Omitidos	3	3	100,00%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	15	83,33%

TOTAL	Anaf	OK	
Omitidos	55	51	92,73%
Personales	53	48	90,57%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	112	92,56%

BASE 2. Sólo preferencias semánticas

L009	Anaf	OK	
Omitidos	13	8	61,54%
Personales	14	13	92,86%
Demostr.			
Reflexivos	4	4	100,00%
	31	25	80,65%

L065	Anaf	OK	
Omitidos	39	18	46,15%
Personales	29	18	62,07%
Demostr.			
Reflexivos	4	4	100,00%
	72	40	55,56%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	5	50,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	11	61,11%

TOTAL	Anaf	OK	
Omitidos	55	27	49,09%
Personales	53	36	67,92%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	76	62,81%

Adquisición: dos bloques. Resolución: el tercero

L009	Anaf	OK	
Omitidos	13	8	61,54%
Personales	14	13	92,86%
Demostr.			
Reflexivos	4	4	100,00%
	31	25	80,65%

L065	Anaf	OK	
Omitidos	39	19	48,72%
Personales	29	18	62,07%
Demostr.			
Reflexivos	4	4	100,00%
	72	41	56,94%

E001	Anaf	OK	
Omitidos	3	1	33,33%
Personales	10	6	60,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	12	66,67%

TOTAL	Anaf	OK	
Omitidos	55	28	50,91%
Personales	53	37	69,81%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	78	64,46%

Aprendizaje: TODOS. Resolución TODOS.

L009	Anaf	OK	
Omitidos	13	10	76,92%
Personales	14	13	92,86%
Demostr.			
Reflexivos	4	4	100,00%
	31	27	87,10%

L065	Anaf	OK	
Omitidos	39	31	79,49%
Personales	29	25	86,21%
Demostr.			
Reflexivos	4	4	100,00%
	72	60	83,33%

E001	Anaf	OK	
Omitidos	3	2	66,67%
Personales	10	7	70,00%
Demostr.	3	3	100,00%
Reflexivos	2	2	100,00%
	18	14	77,78%

TOTAL	Anaf	OK	
Omitidos	55	43	78,18%
Personales	53	45	84,91%
Demostr.	3	3	100,00%
Reflexivos	10	10	100,00%
	121	101	83,47%