Propuesta de incorporación de información semántica desde WordNet al análisis sintáctico parcial orientado a la resolución de la anáfora.

Maximiliano Saiz-Noeda, Armando Suárez, Jesús Peral
{max, armando, jperal}@dlsi.ua.es

Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información
Dpto. Lenguajes y Sistemas Informáticos. Universidad de Alicante
Apartado 99. 03080 Alicante, España

Resumen.

En este trabajo se presenta una propuesta para incorporar información semántica en el proceso de análisis sintáctico parcial. La propuesta está basada en el método IRSAS para textos restringidos y extiende su acción a textos no restringidos con el uso de un recurso léxico de propósito general (WordNet). Tras la generación de una serie de patrones semánticos extraídos de un corpus de entrenamiento, el mecanismo propuesto pretende definir combinaciones sustantivo-verbo semánticamente correctas. Esto permite definir la compatibilidad semántica existente entre un sujeto y un verbo en una oración.

El comportamiento de este mecanismo será evaluado en la resolución de la anáfora pronominal de tipo sujeto con el objetivo de resolver aquellos casos que, debido a la ausencia de información semántica, son tratados de manera incorrecta.

1. Introducción.

En este trabajo se presenta una propuesta de incorporación de información semántica en el análisis sintáctico parcial. Esta propuesta está basada en el método IRSAS para textos restringidos planteado en Moreno *et al.* (1992) y utiliza WordNet como recurso léxico.

Para llevar a cabo esta propuesta, se ha tomado como punto de partida un corpus etiquetado con el sentido correcto de cada palabra, para posteriormente realizar un estudio, aprendizaje y extracción de patrones semánticos de relación sujeto-verbo existentes en un fragmento del corpus. En último lugar, se ha propuesto la evaluación del comportamiento de estos patrones en otro fragmento del corpus para la resolución de la anáfora pronominal de tipo sujeto, proceso que puede requerir información semántica para la elección del antecedente correcto.

De WordNet se ha extraído la ontología de rasgos semánticos utilizada en la obtención de patrones. Dado que WordNet es un recurso disponible actualmente sólo en inglés, se ha definido una gramática inglesa para el análisis.

WordNet plantea diferencias claras frente al mencionado método IRSAS. El primero es un recurso de ámbito global, que proporciona una ontología de rasgos semánticos general aplicable a cualquier dominio y a cualquier texto, mientras que la ontología que facilita IRSAS es muy restringida y adecuada al corpus usado. Esto hace que plantee una mayor eficacia por su adecuación al corpus. Sin embargo su ámbito de acción resulta claramente limitado.

El artículo presentado comienza con una definición de los elementos involucrados en el ámbito de aplicación de la propuesta. En la siguiente sección se describe el mecanismo de incorporación de información semántica al análisis sintáctico parcial. A continuación se discuten los puntos principales de la propuesta para finalizar con las conclusiones.

2. Ámbito de aplicación de la propuesta.

En esta sección se presenta, en primer lugar, la *Gramática de Unificación de Huecos* (*Slot Unification Grammar, SUG*) como base del analizador parcial *SUPP*, que se describe a continuación. Posteriormente se analiza el método IRSAS como base del trabajo presentado. Seguidamente se detalla la resolución de la anáfora y el algoritmo planteado. Por último se presentan las características del recurso léxico WordNet que se han considerado más relevantes para la propuesta presentada.

2.1. Gramática *SUG*.

Las *SUG* fueron presentadas en Ferrández *et al.* (1997) como una extensión de las *Gramáticas de Cláusulas Definidas (Definite*

Clause Grammars, DCG), introducidas por Pereira y Warren (1980), con el objetivo de ampliar las capacidades de las *DCG* para facilitar la resolución de manera modular de diversos problemas lingüísticos.

Las *SUG* se definen como una cuádrupla *(NT,T,P,H)*, donde *NT* es un conjunto finito de símbolos no terminales y *T* es un conjunto finito de símbolos terminales disjunto con *NT*. *P* son las *reglas de producción* de la gramática: un conjunto finito de pares $\alpha \rightarrow \beta$ donde $\alpha \in NT$, $\beta \in (T \cup NT)^* \cup$ *{llamadas a procedimientos}*. Por último, H son *hechos SUG* que sólo tienen el primer miembro de la regla, donde α puede ser *coordination*, *juxtaposition*, *fusion*, *basicWord* o *isWord*.

Al ser *SUG* una extensión de *DCG*, heredará muchas de sus características. La principal diferencia es que las reglas de producción son de la forma $\alpha \rightarrow \beta$ (frente a $\alpha \rightarrow \beta$), y cada subconstituyente de β puede omitirse en la oración si se escribe entre el operador << >> (constituyente opcional).

2.2. Sistema *SUPP*.

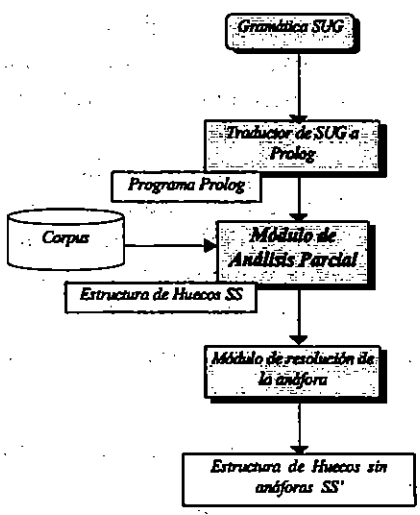

Figura 2.1 Análisis parcial SUPP y módulo de resolución de la anáfora

En la *Figura 2.1* se muestra el esquema general del funcionamiento del analizador *SUPP*, presentado en Martínez-Barco *et al.* (1998), y del módulo de resolución de la anáfora.

En primer lugar, se define una gramática *SUG* capaz de reconocer sintagmas nominales, sintagmas preposicionales y *chunks*[1] verbales. Esta gramática se traduce automáticamente a cláusulas Prolog (mediante el traductor *SUG*), dando como resultado el analizador que, con el uso de una técnica de análisis descendente, proporciona como salida los constituyentes obtenidos y los almacena en la *Estructura de Huecos* (*Slot Structure*, *SS*). Esta *SS* contiene la información necesaria para posteriormente utilizarla en el módulo de resolución de la anáfora (en este caso se usa información léxica, morfológica y sintáctica). Tras el módulo de resolución de la anáfora, la estructura de huecos resultante (*SS'*) es idónea para su posterior aplicación a tareas de recuperación y extracción de información, traducción automática, etc.

2.3. Método IRSAS.

Una aproximación a la incorporación de información semántica en el análisis sintáctico es el método IRSAS (Incorporar Restricciones Semánticas en el Análisis Sintáctico) presentado en Moreno *et al.* (1992).

Para ello, cada objeto del universo del discurso se asocia por sus propiedades a un rasgo semántico. Esta clasificación (*ontología de los rasgos semánticos*) se define mediante la aplicación de dos relaciones básicas:

a) Las relaciones de división y herencia, por las que un concepto se puede dividir en subconceptos que heredan del tipo padre sus rasgos semánticos, añadiendo nuevas características especificadas. Por ejemplo, un *ser vivo* puede ser *animal* o *vegetal*. Un *animal* puede ser *macho* o *hembra*, etc. Esta relación se denota como

$$rasgo \Leftrightarrow rasgo1 \vee rasgo2 \vee ...rasgoN$$

donde $\vee$ es una disyunción exclusiva.

b) Las relaciones de implicación, por las que un rasgo específico puede implicar a otro rasgo más genérico. Así se pueden definir, por ejemplo, relaciones del tipo: un *perro* es un *cazador*. Se denota como:

$$rasgoA \Rightarrow rasgoB$$

Una vez definida correctamente la ontología de rasgos siguiendo las relaciones anteriores, el método IRSAS las almacena en tres grafos:

- El grafo de herencias, donde se relaciona cada rasgo con sus herederos.
- El grafo de incompatibilidades, donde quedan relacionados aquellos rasgos que por formar parte de la disyunción exclusiva de una herencia no podrían

[1] Un chunk se define, según Abney (1997), como una secuencia de elementos con cierto sentido sintáctico alrededor de un núcleo o cabecera.

darse simultáneamente en un mismo tipo semántico (ningún tipo semántico podría contener a la vez los rasgos *animal* y *vegetal*).

- El grafo de implicaciones, que mantiene las relaciones de implicación definidas.

Finalmente, partiendo de estos grafos indicados y mediante la aplicación de una serie de condiciones de consistencia entre listas de rasgos, el método IRSAS es capaz de determinar la compatibilidad semántica entre dos términos.

En la *Figura 2.2* se muestra un fragmento de la ontología IRSAS con relaciones de división y herencia y relaciones de implicación.

```
Relaciones de división y herencia:
ontolog(entity,[concrete,abstract]).
ontolog(entity,[individual,collective]).
ontolog(abstract,[temporal,espatial,mental]).
ontolog(mental,[state,action]).
ontolog(concrete,[organic,inorganic]).
ontolog(organic,[vegetal,animal]).
ontolog(animal,[human,nonhuman]).
ontolog(animal,[male,female]).
ontolog(inorganic,[geography,physics]).
ontolog(geography,[town,mountain]).
ontolog(physics,[solid,liquid,gas]).
ontolog(liquid,[drinkable,undrinkable]).
ontolog(solid,[eatable,uneatable]).
Relaciones de implicación:
ontolog2(animal,[solid]).
ontolog2(town,[solid]).
```

Figura 2.2 Fragmento de la ontología de rasgos semánticos de IRSAS

El método IRSAS se aplicó en sus orígenes al análisis de oraciones, posteriormente a la resolución de la elipsis intraoracional y a la resolución de la anáfora pronominal en textos restringidos.

2.4. Resolución de la anáfora.

```
Análisis sintáctico de una oración. Obtención de su estructura de huecos (SS).
Para cada anáfora en SS:
  Selección de antecedentes de las X oraciones anteriores
  dependiendo de la clase de anáfora en L0
  Aplicar restricciones (según el tipo de anáfora) a L0 e introducir el
  resultado en L1:
  Caso de que:
    |L1| = 1 Entonces:
      Éste será el antecedente de la anáfora
    |L1| > 1 Entonces:
      Aplicar preferencias (según el tipo de anáfora) a L1 e introducir el
      resultado ordenado en L2:
      El primero de L2 será el antecedente seleccionado
Actualizar SS con el antecedente de cada anáfora con el resultado de SS'.
```

Figura 2.3 Algoritmo de resolución de la anáfora según Ferrández et al. (1998)

La *Figura 2.3* describe el algoritmo que resuelve la anáfora discursiva en textos no restringidos usando análisis parcial (sin utilización de información semántica) y sobre el que se pretende aplicar el método propuesto en este trabajo.

El algoritmo aplicará un conjunto de restricciones (concordancia morfosintáctica y restricciones *c-command*) a la lista de posibles antecedentes con el objetivo de descartar candidatos. Si sólo hay un candidato, éste será el antecedente de la anáfora. De otro modo, si quedan más de un candidato, se aplican un conjunto de preferencias (paralelismo sintáctico, información léxica, reiteración de un antecedente en el texto, etc.). Estas preferencias ordenarán la lista de antecedentes restantes. Como resultado, el primero de la lista será el antecedente seleccionado.

Este algoritmo ha sido evaluado sobre textos no restringidos que no disponían de información semántica. En concreto, se ha utilizado el corpus *The Blue Book*[2] (manual técnico de telecomunicaciones, *International Telecommunications Union CCITT handbook*) en sus versiones en español e inglés.

Para la versión española del corpus *The Blue Book* se ha detectado el 100% de las anáforas pronominales, la longitud media de las oraciones con anáforas es de 48 palabras y se ha obtenido un 83% de éxito (pronombres correctamente resueltos dividido por número total de pronombres) para las anáforas pronominales.

Tomando como base el sistema en español se han realizado cambios para adaptarlo al idioma inglés.

Para el corpus *The Blue Book* en inglés, se ha detectado el 100% de las anáforas pronominales, la longitud media de las oraciones de 22 palabras y se ha obtenido un 87'3% de éxito.

2.5. El recurso léxico: WordNet.

WordNet, introducido en Miller (1990). es una base de datos léxica de términos en inglés organizados en unas estructuras denominadas synsets. Estas estructuras reflejan la propiedades de sinonimia entre las palabras.

En lo que respecta a este trabajo, tal y como se puede ver en Fellbaum (1998), WordNet organiza sus *synsets* en los llamados ficheros

[2] Corpus incluido en el Proyecto CRATER. Corpus Resources and Terminology Extraction. Proyecto financiado por la Comisión de las Comunidades Europeas (DG-XIII). Investigadores principales Marcos, F. y Sánchez, F. Laboratorio de Lingüística Informática. Facultad de Filosofía y Letras. Universidad Autónoma de Madrid.

lexicográficos, que definen una lista de conceptos clave basados en categorías sintácticas y agrupamientos lógicos. Los sustantivos quedan agrupados en 25 conceptos y los verbos en 15. La *Figura 2.4* muestra ambos grupos.

Sustantivos	act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time
Verbos	body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather

Figura 2.4 Clasificación conceptual de WordNet para sustantivos y verbos

De aquí en adelante, denominaremos ontología principal a este conjunto de elementos conceptuales.

3. Propuesta de incorporación de información semántica al análisis.

Como ya se ha comentado, el principal objetivo de este trabajo es la incorporación de información semántica al análisis parcial aplicado a la resolución de la anáfora pronominal de tipo sujeto. Para llevar a cabo con garantías este objetivo es necesario contar con un recurso léxico adecuado.

El método tomado como base, IRSAS, define su propia ontología de rasgos para textos restringidos. Ante el tratamiento de textos no restringidos se ha decidido utilizar WordNet, por lo que el trabajo se circunscribe al estudio y propuesta de un mecanismo que obtenga información semántica de WordNet. Este mecanismo seleccionará información en forma de patrones semánticos del tipo sustantivo-verbo, información que WordNet no suministra directamente.

El proceso de incorporación de información semántica se ha desarrollado en varias etapas.

Una primera fase es la obtención de los patrones semánticos. Para ello se ha usado la ontología principal, conjunto de conceptos base descritos con detalle en la sección anterior.

Para la obtención de los patrones se extraen de un corpus de entrenamiento los pares formados por un sustantivo y una raíz verbal. El sustantivo es el núcleo de un sintagma nominal con función de sujeto y el verbo es el núcleo del sintagma verbal que acompaña a dicho sujeto.

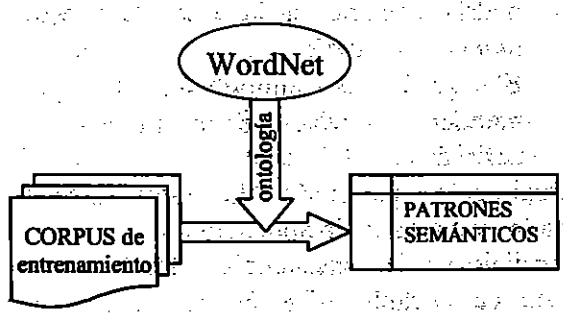

Figura 3.1 Obtención de patrones semánticos

Con el fin de establecer patrones generales, se consulta en la ontología principal de WordNet ambos elementos de los pares sustantivo-verbo extraídos. De cada palabra se obtiene el elemento de la ontología. El esquema general de este proceso se describe gráficamente en la *Figura 3.1.*

El conjunto de patrones genéricos así definidos describe el comportamiento semántico de los verbos y de los sustantivos como sujetos de esos verbos.

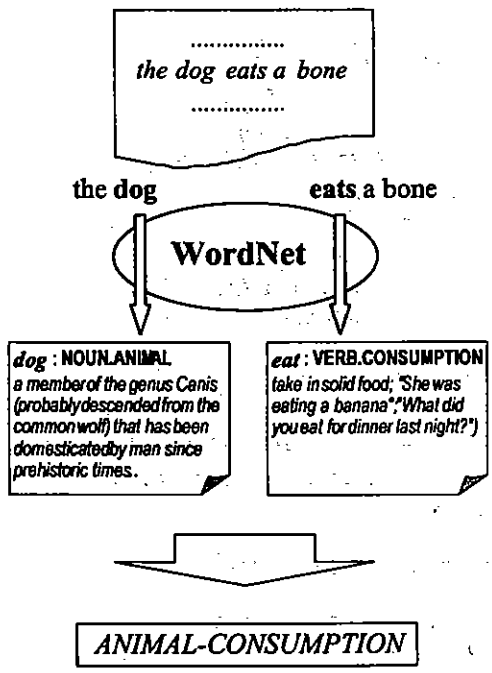

Figura 3.2 Ejemplo de obtención de un patrón en el aprendizaje

La *Figura 3.2* ilustra un ejemplo sencillo del mecanismo de extracción de patrones. En la oración *"The dog eats a bone"* se extraería el par *dog-eat* donde *"dog"* es el núcleo del sintagma nominal sujeto *"the dog"* y *"eat"* es la raíz del sintagma verbal *"eats a bone"*. A partir de este par, la consulta en WordNet daría como resultado que *"dog"* tiene como elemento de ontología básico *animal* y *"eat"* tiene como elemento de ontología verbal básico

consumption[3]. Esto genera un patrón de comportamiento *animal-consumption* que se añadirá a la lista de patrones genéricos de aprendizaje.

En la siguiente fase, una vez obtenidos los patrones semánticos, se realiza el análisis. Dado que las estructuras de huecos resultantes de dicho análisis cuentan con la información referente al sentido de cada palabra, se puede comprobar la compatibilidad semántica entre el sintagma nominal sujeto y su verbo, de manera que, aunque el análisis haya resultado correcto sintácticamente, la oración puede no estar semánticamente bien construida. Este mecanismo queda patente en la *Figura 3.3*.

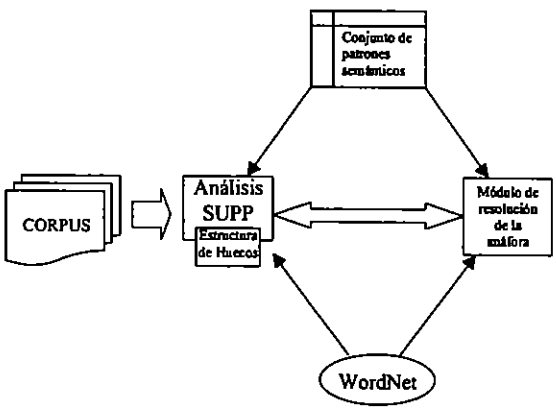

Figura 3.3 Mecanismo de incorporación de información semántica

Por otro lado, para el caso particular de la resolución de la anáfora, este sistema proporciona un mecanismo adicional de evaluación del antecedente semánticamente más compatible.

Este proceso, definido gráficamente en la *Figura 3.3*, parte de la lista de antecedentes (sintagmas nominales) y del verbo de la expresión anafórica. A partir de las etiquetas semánticas, se realiza una consulta para extraer de WordNet el elemento de la ontología principal correspondiente al sustantivo núcleo de cada sintagma nominal (antecedente) así como el de la raíz del verbo de expresión anafórica. Los pares de elementos de ontología así obtenidos se contrastan con la lista de patrones extraída del corpus de aprendizaje. Esta comparación establece el antecedente semánticamente más compatible según los patrones aprendidos.

La *Figura 3.4* muestra la imagen de un prototipo utilizado para ponderar estos antecedentes con el verbo a partir de la lista de patrones de comportamiento.

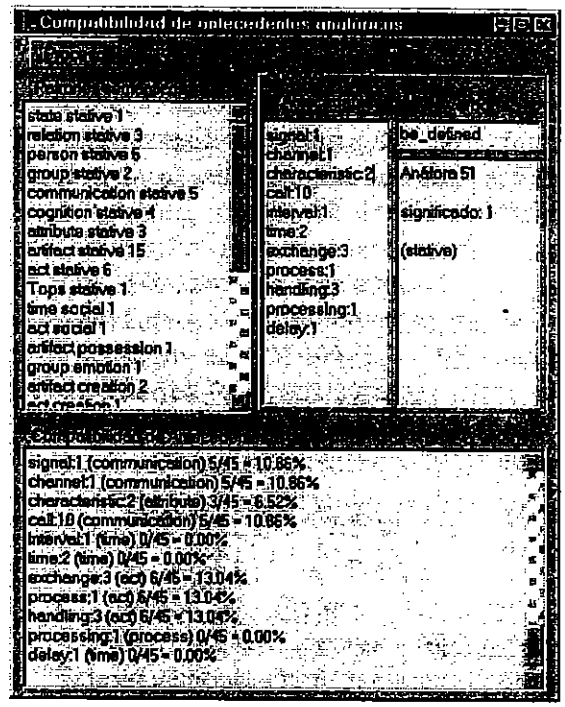

Figura 3.4 Prototipo para la elección del antecedente más compatible

4. Discusión.

A continuación se presenta un análisis del problema desde la perspectiva de esta propuesta, mostrando un ejemplo en el que la generación de patrones semánticos resuelve una anáfora que el sistema original (sin información semántica) es incapaz de resolver.

The Hold function is used to put existing calls which are in the establishment or in the active phase in the Call Held auxiliary state. By default, it reserves the B-channel in use.

Este fragmento de texto contiene una anáfora pronominal (*it*) de tipo sujeto. Tras aplicar las restricciones que se encargan de descartar aquellos antecedentes que no son compatibles morfológicamente (género y número), el módulo de resolución de la anáfora proporciona una lista de posibles antecedentes (*Tabla 4.1*).

antecedente	*núcleo*
default	default
the Call Held auxiliary state	state
the active phase	phase
the establishment	establishment
the Hold function	function

Tabla 4.1. Antecedentes de la anáfora ejemplo

En la aplicación de preferencias, el sistema original determina *default* como el antecedente más probable (por su proximidad a la expresión

[3] Es conveniente recordar que se parte de un corpus etiquetado con el significado correcto de cada palabra en el texto. Las etiquetas siguen la codificación de WordNet.

anafórica una vez agotado el resto de los criterios). Sin embargo, analizando el texto, se comprueba que el antecedente correcto es *The Hold function*.

En nuestra propuesta, en el aprendizaje se genera una lista de patrones con el índice de aparición de los mismos. Estos patrones contienen los conceptos de la ontología principal extraída de WordNet

A continuación se extrae el núcleo de los sintagmas nominales antecedentes y se consulta su correspondencia en la ontología principal. En la *Tabla 4.2* se muestra cada antecedente representado por su núcleo, su sentido de la palabra en el texto (etiqueta de WordNet) y el concepto ontológico asociado.

núcleo	*sentido WN*	*concepto*
default	1	act
state	4	Tops
phase	1	time
establishment	1	act
function	6	communication

Tabla 4.2. Resultado de la consulta de sustantivos en WordNet

Por otra parte, el verbo de la anáfora es *reserve* y su etiqueta semántica corresponde a su tercer sentido de WordNet. La consulta en la ontología principal de verbos devuelve el concepto *possession*.

sustantivo	*verbo*	*apariciones*	
artifact -	possession	150	(38%)
communication -	possession	60	(15%)
person -	possession	50	(13%)
act -	possession	40	(10%)
relation -	possession	30	(8%)
attribute -	possession	30	(8%)
group -	possession	20	(5%)
state -	possession	10	(3%)
Tops -	possession	10	(3%)

Tabla 4.3. Patrones correspondientes al concepto possession

Tras este proceso, contamos con los conceptos ontológicos del verbo (*possession*) y de los núcleos de los antecedentes (*Tabla 4.2*), así como con los patrones extraídos en el aprendizaje. Con estos elementos, el siguiente paso es consultar en la lista de patrones los correspondientes al elemento conceptual del verbo. Tal y como se muestra en la *Tabla 4.3*, esta consulta nos proporciona todos los patrones que incluyen el concepto *possession* como verbo. A cada patrón se le asocia un índice que representa el número de apariciones que tiene en el corpus.

Por último, la elección del antecedente se basa en el emparejamiento de cada uno de estos patrones con las posibles combinaciones de los conceptos ontológicos de los núcleos de los antecedentes y el verbo.

Esta combinación da como resultado el índice de compatibilidad existente entre antecedente y verbo anafórico (*Tabla 4.4*).

núcleos	*concepto*	*compatibilidad con possession*
default	cognition	0%
state	Tops	3%
phase	time	0%
establishment	act	10%
function	communication	15%

Tabla 4.4. Compatibilidad antecedente-verbo

Tal y como se muestra en la *Tabla 4.4*, el resultado define a *function* como el candidato más compatible de todos, y por tanto, el sistema ampliado con información semántica resolvería correctamente la anáfora.

Cabe la posibilidad de que sean varios los antecedentes que queden cubiertos por el mismo patrón, por lo que el resultado daría una equiprobabilidad entre ellos. Esta información sería útil, no tanto para seleccionar el antecedente correcto, pero sí para descartar aquellos que no lo son.

Como resultado de una evaluación manual inicial, y partiendo del índice de éxito que proporciona el sistema sin información semántica, estimamos un incremento en este índice de un 6% (del 87,3 al 93% aproximadamente)

5. *Conclusiones.*

En este trabajo se ha propuesto un mecanismo de incorporación semántica en el análisis sintáctico parcial. La propuesta presentada ha sido extendida a la resolución de la anáfora pronominal de tipo sujeto y plantea la posibilidad de incrementar su índice de éxito.

Consideramos que la combinación de determinado tipo de conceptos intuitivamente incompatibles, como por ejemplo *artefacto-sentimiento*, parece incoherente desde el punto de vista semántico. Así mismo, parece lógico pensar que en un texto referente a un tema concreto, la aparición de ciertos patrones semánticos sea más probable que otros.

Esta reflexión nos ha llevado a la extracción de patrones que definan el comportamiento de

pares sustantivo-verbo como sujeto y núcleo verbal de una oración.

Se ha pretendido con esta propuesta encontrar la compatibilidad, en el caso del análisis, y el grado de aceptabilidad de un antecedente en el caso de la resolución de la anáfora entre estos pares de términos.

En la línea de trabajos futuros, se pretende la implantación de un sistema que permita la incorporación de información semántica en el análisis sintáctico, orientado a la resolución de la anáfora. Para ello se contrastarán sistemas basados en el conocimiento y sistemas estocásticos con el fin de estudiar qué tipo de estrategias ofrecen mejores resultados.

6. Agradecimientos.

Este artículo ha sido subvencionado por la Comisión Interministerial de Ciencia y Tecnología con el proyecto número TIC97-0671-C02-02.

7. Referencias.

Abney, S. (1997). Part-of-Speech Tagging and Partial Parsing. En *Corpus-based Methods in Language and Speech Processing*. S. Young and G. Bloothooft, Eds., Kluwer Academic publishers. Holanda. 1997. pp. 119-136.

Fellbaum, C. (1998) *WordNet, an electronic lexical database*. Fellbaum, C. eds. MIT Press. ISBN 0-262-06197-X. 1998.

Ferrández, A.; Palomar, M.; Moreno, L. (1997). Slot Unification Grammar. En *Actas de APPIA-GULP-PRODE* (Grado, Italia, 1997). pp. 523-532.

Ferrández, A.; Moreno, L.; Palomar, M. (1998). Anaphor resolution in unrestricted texts with partial parsing. En *Actas del 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL'98 - COLING'98* (Montreal, Quebec, Canada, 1998). pp. 385-391.

Martínez-Barco, P.; Peral, J.; Ferrández, A.; Moreno, L.; Palomar, M. (1998). Analizador Parcial *SUPP*. En *Actas de VI biennial Iberoamerican Conference on Artificial Intelligence, IBERAMIA'98* (Lisboa, Portugal, octubre 1998). pp. 329-341.

Miller, G.A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. (1990). *Five Papers on WordNet*. CSL Report 43, Cognitive Science Laboratory, Princeton University. (1990).

Moreno, L.; Andrés, F.; Palomar, M. (1992). Incorporar Restricciones Semánticas en el Análisis Sintáctico: IRSAS. *Procesamiento del Lenguaje Natural, 12* (1992).

Pereira, F.; Warren, D. (1980). Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks. *Artificial Intelligence, 13* (1980).