Una propuesta de estructuración del conocimiento para la adquisición de esquemas conceptuales de bases de datos a partir de textos

Paloma Martínez
pmf@inf.uc3m.es
Departamento de Informática
Universidad Carlos III de Madrid

Ana García-Serrano
agarcia@dia.fi.upm.es
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid

Resumen

La adquisición de conocimiento a partir de textos escritos mediante medios informáticos es un campo muy activo actualmente. El objetivo en la interpretación de textos es la producción de una estructura clara, con precisión medible y adaptable que incorpore la información del texto.

En este artículo se presenta una propuesta general de estructuración del conocimiento para la realización de modelos que incorporen gradualmente el conocimiento de un lingüista y que permitan posteriormente el desarrollo de un sistema con arquitectura cognitiva. Dicha arquitectura debe admitir fielmente la modularización del conocimiento planteada y el control no fijo para el análisis de textos, facilitando las labores de revisión y modificación del conocimiento por parte del experto en el dominio, aspecto clave en el caso de desarrollo de sistemas para tratamiento automático del lenguaje.

Finalmente se presenta un caso concreto de estudio, el diseño de bases de datos a partir de la interpretación de textos o enunciados en castellano.

Áreas Temáticas

Extracción y recuperación de Información, Sistemas Basados en el Conocimiento, Inteligencia Artificial

1 Introducción

El procesamiento automático del lenguaje natural ha sido abordado con formalismos y tecnología que provienen de la Inteligencia Artificial y más recientemente los sistemas informáticos incorporan además tratamiento estadístico o bien formalismos no simbólicos. Estas tecnologías han permitido el desarrollo con éxito de aplicaciones de lingüística computacional, generalmente para dominios concretos o con funcionalidad restringida. En estos sistemas la incorporación de conocimientos lingüísticos se realiza según un criterio minimalista para la obtención de la funcionalidad requerida y además la flexibilidad del sistema para ampliar o modificar su funcionalidad o el conocimiento incorporado es poco satisfactoria e incluso nula.

Sin embargo dado que éste es un campo en el que la interdisciplinaridad es ineludible, Dahl (1993), para el éxito de una aplicación de tratamiento automático de una lengua es necesario encontrar una solución conforme con los intereses tanto de los lingüistas como de los informáticos. Para ello es necesario disponer de una conceptualización adecuada así como de mecanismos eficientes para el tratamiento del gran volumen de información lingüística necesaria, Spark Jones (1996), para lo que se plantea el uso de una metodología que lleve a la especificación de una arquitectura con la funcionalidad requerida y con componentes que incorporen gradualmente el

conocimiento morfológico, sintáctico, semántico y pragmático de forma fácilmente revisable por el experto lingüista.

Desde mediados de los noventa en Inteligencia Artificial se ha investigado en la definición de metodologías para el de desarrollo de sistemas basados en el conocimiento. En Cuena y Molina (1996) se propone una metodología para diseño de este tipo de aplicaciones en la que la primera fase consiste en la formulación de modelos estructurados del conocimiento de forma independiente a su implementación final. Estos modelos abstractos del conocimiento, con escasas referencias a su codificación e implementación, son más comprensibles para el experto en el dominio. En este primer nivel el conocimiento se estructura y describe según dos perspectivas, la de las áreas de conocimiento y la de las tareas. En los niveles inferiores se detallan e incorporan los aspectos de simbolización e implementación del conocimiento identificado en la fase de descripción anterior, diseñándose una arquitectura de módulos que incorporan el conocimiento y definiendo la forma actuación conjunta que posibilita la realización de las tareas objetivo del sistema.

Esta metodología de desarrollo y el sistema final facilitan las labores de revisión y modificación del conocimiento por parte del experto en el dominio, aspecto clave en el caso de desarrollo de sistemas para tratamiento automático del lenguaje.

La adquisición de conocimiento a partir de textos escritos es un campo muy activo actualmente, debido a la utilización generalizada de la informática y el auge de las comunicaciones, lo que ha permitido la distribución masiva de información. En las aproximaciones existentes es posible identificar que el objetivo en la interpretación de textos es la producción a partir del texto, de una estructura clara, con precisión medible y adaptable utilizando métodos débiles o estadísticos, Jacobs y Rau (1993), durante el pre-proceso de los textos y técnicas basadas en el conocimiento con arquitecturas cognitivas complejas que controlen declarativamente la interacción de conocimiento lingüístico y conceptual. En línea con nuestra propuesta se encuentran los trabajos dirigidos por Jacobs y por Delisle.

La herramienta GENL, Jacobs y Rau, (1993), incluye conocimiento léxico con 15.000 entradas. Este conocimiento está estructurado según su significado y es dinámico, en el sentido de que es capaz de proponer nuevas formas léxicas y significados. El análisis no se realiza de forma secuencial y se beneficia de métodos estadísticos y heurísticos para obtener determinada información del texto. A continuación efectúa un análisis con control mixto, es decir dirigido por el objetivo para la creación de una estructura formal asociada al texto y dirigido por el lenguaje. Se ha evaluado su herramienta en diferentes dominios, consiguiendo interpretaciones correctas en un elevado número de casos.

En el proyecto TANKA, Delisle et al (1996), se define un modelo conceptual del dominio descrito en un grupo de textos a partir de escaso conocimiento a priori del dominio. El sistema elabora modelos del texto a partir del análisis léxico y sintáctico para identificar los constituyentes del mismo y a continuación realiza un análisis del texto obteniendo las relaciones semánticas entre los constituyentes. El control del análisis es fijo y secuencial. Finalmente transforma la interpretación semántica en una red semántica que integra en la descripción del dominio en curso. El usuario interactúa con el sistema supervisando y modificando la propuesta y el sistema aprende de la interacción con el usuario y mejora progresivamente sus sugerencias.

En este artículo se presenta una propuesta general de estructuración del conocimiento para la realización de modelos que incorporen gradualmente el conocimiento de un lingüista y que permitan posteriormente el desarrollo de un sistema con arquitectura cognitiva que incorpore fielmente la modularización del conocimiento

planteada y el control no fijo para el análisis del texto. En la propuesta de modularización del conocimiento lingüístico se refleja la naturaleza descriptiva de algunas expresiones para la definición de las perspectivas o fuentes de conocimiento que guían el proceso de análisis sin fijar una secuencialidad a priori.

Esta propuesta se muestra para un caso concreto, el diseño de bases de datos a partir de la interpretación de textos o enunciados en castellano objeto de estudio y desarrollo en el proyecto ENEAS/BD en el que trabaja la autora de la Universidad Carlos III, De Miguel et al. (1996). Este trabajo se encuentra en la línea de los trabajos de Burg y Van de Riet (1996), cuyo sistema COLOR-X posee una base lingüística combinada con técnicas gráficas para modelado conceptual. Sólo se centra en los verbos y utilizando bases de datos léxicas (Wordnet) propone al usuario los distintos significados verbales.

En los siguientes apartados se presenta la estructuración del conocimiento propuesta, los resultados obtenidos para ejemplo en el caso de definición de esquemas conceptuales para diseño de una base de datos, y por último algunas conclusiones.

2 Propuesta de estructuración del conocimiento

Para la presentación de la estructuración del conocimiento propuesta se ha utilizado la metodología KSM (Knowledge Structure Manager) descrita en Cuena y Molina (1996). Este entorno soporta una metodología de diseño y desarrollo de sistemas modular, incremental y multiforme a nivel de representación del conocimiento. El concepto básico en KSM es la Unidad Cognitiva (figura 1) que representa un ente inteligente que dispone de un determinado conocimiento a partir del que es posible la ejecución de diferentes tareas. Cada Unidad Cognitiva se compone de áreas de conocimiento que a su vez engloban a otras unidades cognitivas o bien componentes básicos denominados "primitivas" por incorporar conocimiento de base genérico o bien específico en un dominio. Las tareas o componentes operacionales asociados a las unidades cognitivas determinan la forma de combinación del conocimiento para la consecución de la funcionalidad.

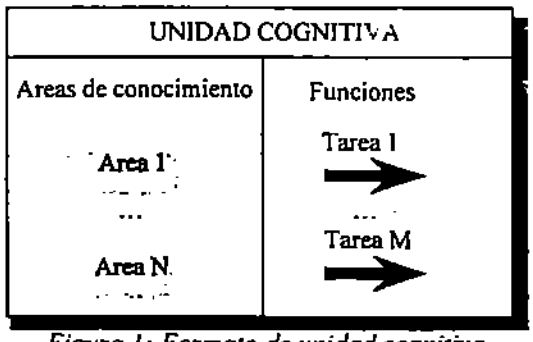

Figura 1: Formato de unidad cognitiva

Con esta metodología la descripción de una aplicación puede hacerse a tres niveles: El nivel del conocimiento, en el cual debe definirse *qué se sabe* y *qué se hace* con ese conocimiento; el nivel simbólico en el cual se describen las formas de representación del conocimiento de las unidades y los métodos o procedimientos con los que se ejecutan cada una de las tareas, y el nivel de implementación o soporte computacional.

En la figura 2 se muestra la estructuración a partir de la unidad cognitiva (UC) raíz que incorpora la definición de los mecanismos de análisis de textos y el conocimiento necesario en tres unidades que constituyen el segundo nivel en la

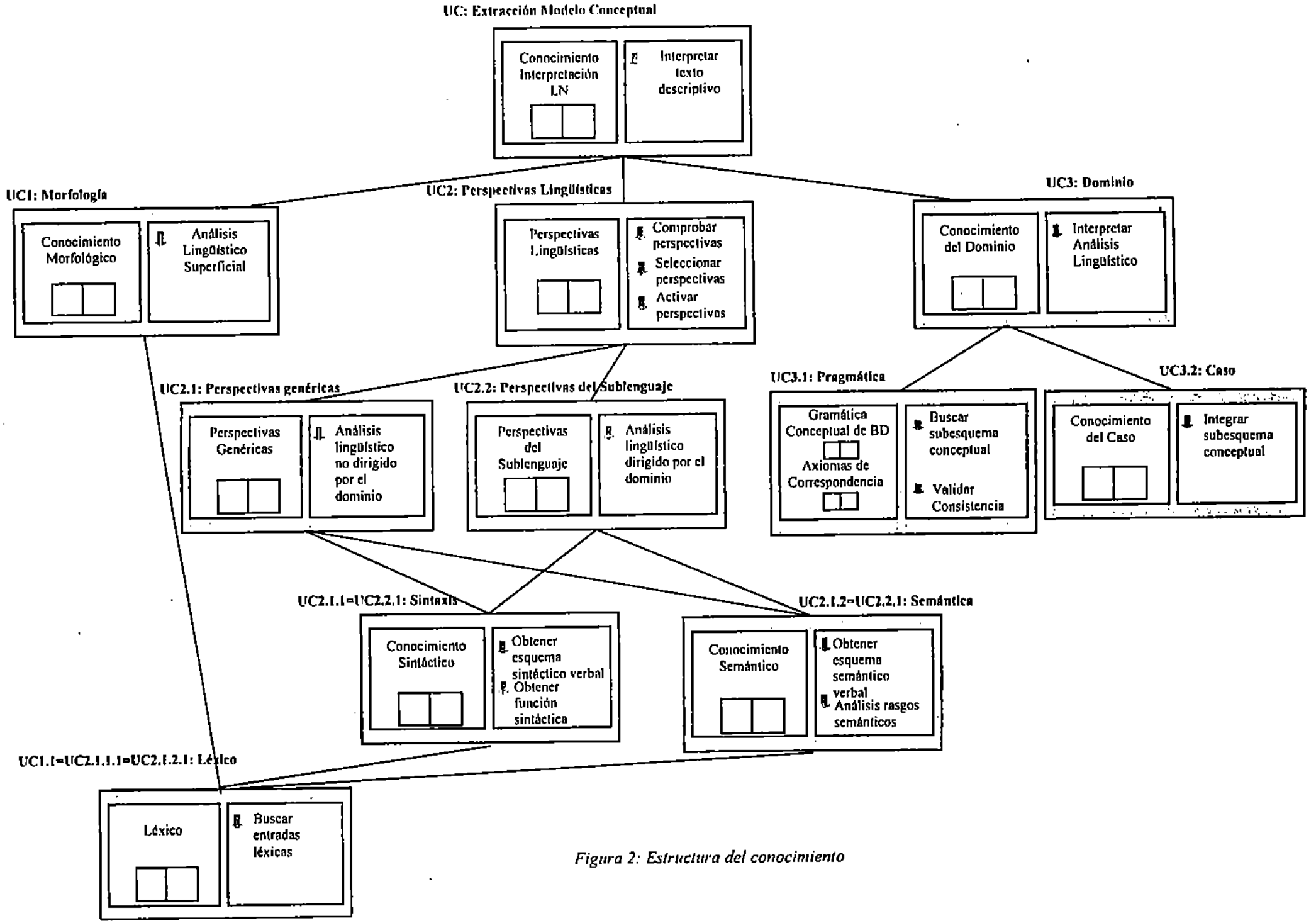

Figura 2: Estructura del conocimiento

jerarquía de unidades. Las unidades UC1, UC2 y UC3, incorporan el conocimiento sobre morfología, la segunda el conocimiento sintáctico y semántico y la tercera con conocimiento del dominio (en nuestro caso conocimiento sobre modelado conceptual de BD). En los siguientes apartados se describirán cada una de estas unidades cognitivas.

2.1 Unidad de Conocimiento Morfológica

Esta unidad cognitiva se subdivide en varias unidades de conocimiento. La primera está formada por un etiquetador morfológico, Sánchez-León y Nieto (1995), entrenado con textos en castellano, basado en un Modelo Oculto de Markov (HMM) utilizado para, en caso de ambigüedad, asignar a cada palabra su categoría morfológica más probable según el contexto que la rodea. Esta unidad es capaz de recibir textos en castellano y etiquetar las palabras y signos de puntuación que :o forman.

Cada etiqueta contiene información concerniente a la categoría y rasgos morfológicos como género, número, tiempo verbal, persona y otros.

La segunda unidad cognitiva contiene varios autómatas finitos sencillos que realizan un análisis superficial de las oraciones, Abney (1996), agrupando las etiquetas morfológicas de una oración en grupos nominales (gn), preposicionales (gp), verbales (gv), adverbiales (gadv), etc. No los denominamos sintagmas puesto que no siempre lo son en sentido gramatical. Este proceso facilitará análisis posteriores.

Oración: Cada curso tiene un código y un horario

Comp1:	cat: gn
	taglist: [QUDX. NCMS] (*lista de etiquetas*)
	lexlist: [cada, curso] (*lista de palabras*)
Comp2:	cat: gv
	taglist: [VLP13S]
	lexlist: [tiene]
Comp3:	cat: gn
	taglist: [ARCAMS, NCMS]
	lexlist: [un, código]
Comp4:	cat: conj
	taglist: [CC]
	lexlist: [y]
Comp5:	cat: gn
	taglist: [ARCAMS, NCMS]
	lexlist: [un, horario]

Figura 3

Un ejemplo de una oración etiquetada y segmentada se muestra en la figura 3.

Dentro de la UC Morfológica también se encuentran reglas para formación de singulares y plurales nominales, así como reglas para conjugación verbal, las cuales serán necesarias a la hora de visualizar información al usuario. Por último, también contiene una serie de heurísticas que ayudan a refinar la segmentación en cuanto a la unión de grupos preposicionales a grupos nominales si procede.

2.2 Unidad de Conocimiento de las perspectivas lingüísticas

Esta unidad trabaja sobre el texto etiquetado y segmentado producido por la UC Morfológica, encargándose de extraer las relaciones lingüísticas más relevantes (si las hubiera) de cada oración que serán utilizadas después por la Unidad de Conocimiento del Dominio. Al contrario que las propuestas centradas en análisis secuenciales con un orden prefijado (morfología → sintaxis → semántica → pragmática), el análisis lingüístico propuesto se basa en el empleo de las *perspectivas lingüísticas* para control del proceso de análisis.

Una perspectiva lingüística es un posible enfoque de análisis que puede realizarse sobre una oración (o parte de ella). Así, es posible centrarse en un enfoque

verbo-argumentos o en un enfoque nombre-modificadores. Además, cada enfoque puede abordarse según distintas combinaciones de aplicación del conocimiento disponible con el fin de mejorar la robustez del análisis.

Esta aproximación basada en perspectivas trata de aprovechar los aspectos que proporcionan más información o más útiles en un momento dado. Se distinguen dos tipos de perspectivas: *perspectivas genéricas* y *perspectivas del sublenguaje* (o dominio). Las primeras son independientes de cualquier dominio, es decir, conducen la extracción del significado de una frase u oración sin importar la clase de texto. Las segundas son propias del dominio y se basan en características particulares del sublenguaje que ayudan a dirigir el análisis.

El estudio del sublenguaje se ha realizado sobre una colección de textos procedentes de diversas fuentes de ámbito académico (exámenes y ejercicios de diseño de BD). Es importante destacar que todos los textos son descriptivos, es decir, cada uno explica una parte de un mundo real que se quiere modelar.

Las *perspectivas del sublenguaje* identificadas son:

Patrones específicos de estilo (P1): Está formada por un conjunto de patrones que describen diferentes estructuras sintácticas típicas de los textos del dominio que contemplan varios fenómenos de elipsis de verbos. Estos patrones se equiparan con partes de la oración. Esta perspectiva conduce la interpretación examinando en primer lugar la sintaxis (patrón) para, a continuación, examinar los atributos morfológicos y semánticos de los elementos que lo componen. Algunos patrones obtenidos del estudio del corpus disponible se muestran en la figura 4.

PERSPECTIVA 1

SSP1

gn (gr, gr, ... *y* gr)
gn (gr, gr, ... *o* gr)
gn (gr, gr, ..., *etc*)
gn (gr, gr, ..., gr)
gn (gr, gr, ...)
gn (gr, gr, ..., *etc*)

SSP2

gn ("gr", "gr",... "gr")
gn ("gr", "gr",... *y* "gr")
gn ("gr", "gr",... *o* "gr")
gn ("gr", "gr",... . *etc*)
gn ('gr', 'gr', ... 'gr')

donde gr= gn/gp

Figura 4

Patrones complejos (P2): Forman parte de esta perspectiva una serie de patrones que, al igual que los pertenecientes a P1, manejan fenómenos de elipsis y conjunción aunque de diferente naturaleza, pues en este caso tienen grado de oración. La activación de esta perspectiva hace que el análisis se dirija por la sintaxis, para examinar después características morfológicas y semánticas de los componentes que equiparan el patrón. Algunos de los patrones más usuales en el corpus se muestran en la figura 5.

PERSPECTIVA 2

CP1

gn gv gn, gn, ...*y* gn
gn gv gn, gn, ... etc
gn gv gn, ... *o* gn
gn gv gn, ... *o bien* gn
gn gv gn, ..., gn

CP3

tanto gn *como* gn gv gn
gn gv *de dos/tres/...* *tipos*: gn, gn, ...*y/o* gn
gv gn *de dos/tres/...* *tipos*: gn, gn, ...*y/o* gn
gv *dos/tres/...* *tipos de* gn: gn, gn, ...*y/o* gn

CP4

si gv gn *o no*
gn gv *o no* gn
gn gv gn *o no* gn

Figura 5

Palabras clave (P3): Esta perspectiva la componen un conjunto de palabras, o secuencias de palabras, entendidas como una unidad, que son propias del dominio con una clara correspondencia con algunos conceptos del mismo. En este caso se hace uso de las preferencias léxicas de las palabras según el dominio en que se utiliza. Por ello, el análisis mediante esta perspectiva está dirigido por la semántica. En este apartado no se incluyen los verbos que se verán en la perspectiva P4. La figura 6 muestra algunos ejemplos de palabras clave del dominio bajo estudio.

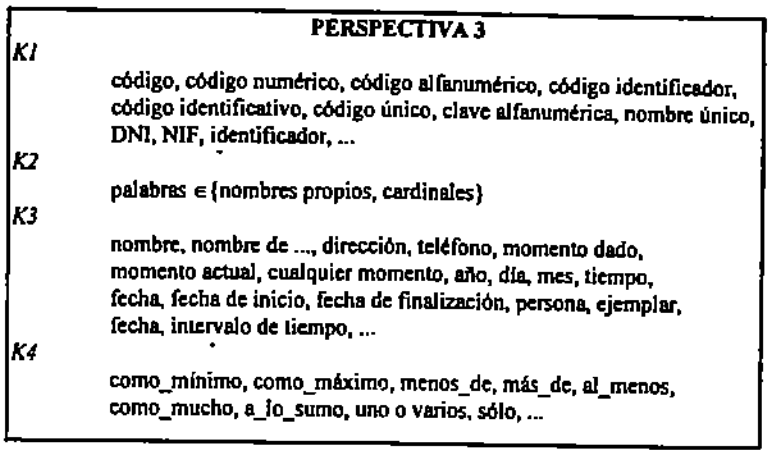

Figura 6

Verbos con preferencia semántica (P4): La última perspectiva del sublenguaje la constituyen los *verbos* que poseen un significado claro, es decir, que desarrollan una preferencia semántica en el dominio general que nos ocupa. Estos son susceptibles de aparecer en cualquier texto descriptivo y se agrupan según muestra la figura 7.

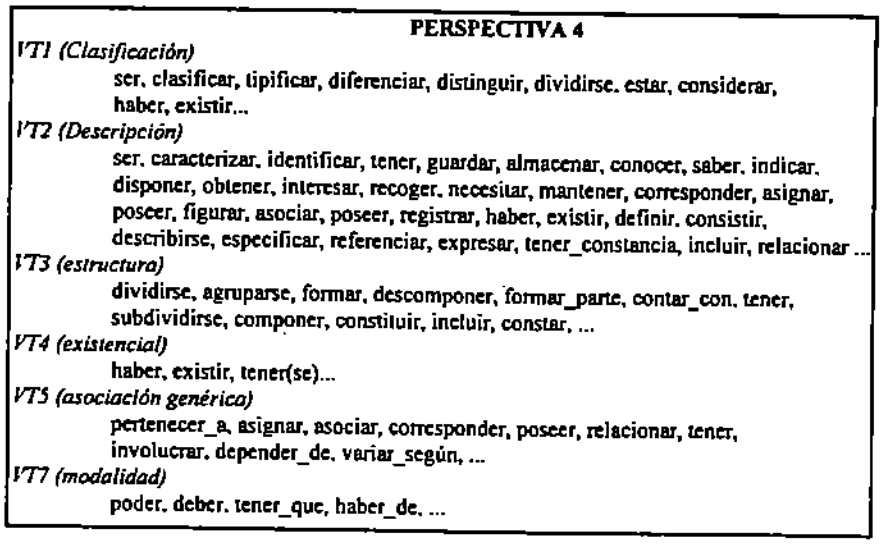

Figura 7

En cuanto a las *perspectivas genéricas* distinguimos dos tipos:

Verbos sin preferencia semántica (P5): Al igual que P4 también se centra en el verbo principal de la oración aunque está formada por aquellos verbos generales. En este caso no tenemos preferencias semánticas en cuanto a los argumentos que pueden regir. La figura 8 muestra algunos de los verbos extraídos del corpus.

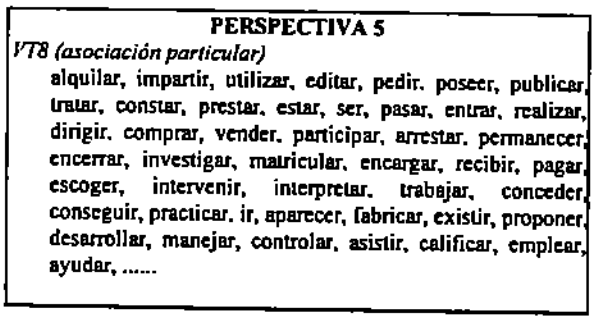

Figura 8

Interrelaciones sintagmáticas (P6): Esta perspectiva se utiliza para estudiar las interrelaciones que existen entre los componentes de un sintagma nominal de igual importancia a las que existen entre un verbo y sus argumentos. Con este fin P6 agrupa una serie de patrones sintácticos muy sencillos, como muestra la figura 9, a partir de los cuales se realiza el procesamiento semántico. Sería interesante incluir también los nombres deverbales que requieren los mismos complementos que los verbos de los que proceden.

PERSPECTIVA 6

PP1
nc

PP2
nc adjetivo

PP3
nc gp

PP4
cuantificador nc

PP5
artículo nc

PP6
ordinal nc
cardinal nc

PP7
una serie de nc
un conjunto de nc
un número de nc

donde
nc=nombre común
gp=grupo preposicional

Figura 9

La unidad cognitiva que incluye y gestiona estas perspectivas se encarga de su comprobación, selección y activación. El análisis lingüístico correspondiente a cada perspectiva es una combinación de morfología, sintaxis y semántica, como se muestra en la tabla 1, que relaciona las distintas perspectivas, con su tipo, el tipo de proceso asociado, la secuencia de análisis y su ámbito.

Posteriormente veremos cómo cada perspectiva conduce el análisis lingüístico con el fin de rellenar una estructura formal con la información extraída y que se utilizará durante la fase de análisis pragmático.

2.3 Unidad de Conocimiento Sintáctica

Esta unidad incluye información sintáctica y proporciona a cada una de las perspectivas la información que soliciten para realizar sus tareas de comprobación de patrones, obtención de funciones sintácticas, obtención de estructuras sintácticas verbales, y otras

Contiene cuatro tipos de conocimiento: Reglas gramaticales necesarias para la localización de segmentos de una oración que desempeñan una determinada función sintáctica (sujeto, complemento directo, atributo del sujeto, complemento indirecto, etc.); una jerarquía sintáctica de verbos que permite clasificar los esquemas verbales dependiendo de los rasgos del verbo y de los segmentos encontrados en la frase; un conjunto de reglas para distinguir los elementos de los grupos verbales (verbo principal, auxiliar, modal, voz, etc.); y por último, todos los patrones sintácticos que utilizan las perspectivas dirigidas por sintaxis (P1, P2 y P6).

Perspectiva	P1	P2	P3	P4	P5	P6
Tipo	sublenguaje	sublenguaje	sublenguaje	sublenguaje	genérica	genérica
Proceso	dirigido por la sintaxis	dirigido por la sintaxis	dirigido por la semántica	dirigido por la semántica	dirigido por la sintaxis	dirigido por la sintaxis
Secuencia[1]	MF-SX-SM-PR	MF-SX-SM-PR	MF-SM-PR	MF-SM-SX-PR	MF-SX-SM-PR	MF-SX-SM-PR
Ámbito	frase	oración	palabra	oración	oración	sintagma

Tabla 1

La figura 10 muestra un subconjunto muy simplificado de la jerarquía sintáctica verbal. En ella aparecen algunos esquemas verbales, es decir, las realizaciones sintácticas de los argumentos de los verbos (transitivo, bitransitivo, intransitivo, atributivo, etc.). Un verbo puede aparecer con más de un esquema sintáctico, dependiendo por ejemplo, de la voz de la construcción verbal (activa, pasiva o media). Cada entrada léxica verbal del diccionario debe contener los distintos esquemas sintácticos y semánticos asociados al verbo.

[1] Se refiere al procedimiento seguido en la aplicación de los distintos tipos de conocimiento: SX es sintaxis, SM es semántica, MF es morfología y PR es pragmática.

Los símbolos que aparecen señalados con "+" son los complementos sintácticos superficiales de los verbos. Algunos de los esquemas sintácticos son:

Sujeto (Oración con argumento Sujeto)
Sujeto-ObjDir (Oración con argumentos Sujeto y Complemento Directo)
Sujeto-ObjInd (Oración con argumentos Sujeto y Complemento Indirecto)
Sujeto-Supl (Oración con argumentos Sujeto y complemento preposicional con preposición fuertemente regida, por ejemplo en "*un empleado pertenece a un departamento*")
Sujeto-CompPrep (Oración con argumentos Sujeto y complemento preposicional con preposición débilmente regida)
Sujeto-Atrib (Oración con argumentos Sujeto y Atributo del sujeto)

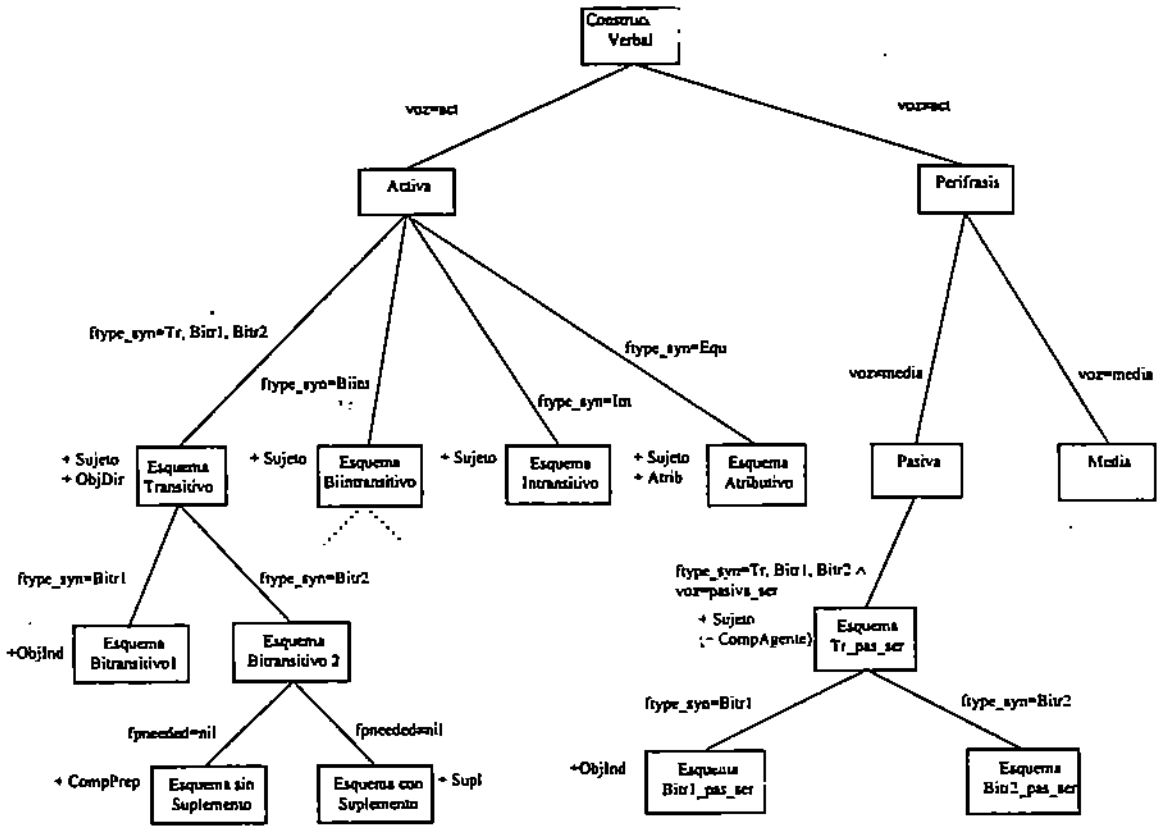

Figura 10

Para navegar por la jerarquía se utilizan los rasgos obtenidos del grupo verbal (gv) de la oración (voz, verbo auxiliar, verbo modal) así como los rasgos propios del verbo contenidos en el diccionario.

Cuando por equiparación de los rasgos anteriores se llega a un nodo de la jerarquía, se localizan los complementos que aparecen señalados con "+". La UC sintaxis contiene también las reglas gramaticales que encuentran en la oración los segmentos que realizan la función sintáctica de esos complementos. Por ejemplo, para localizar un Sujeto, hay que buscar los sintagmas nominales que concuerden con el verbo (puede haber varios que cumplan esta condición).

En el estado actual del trabajo las reglas son sencillas puesto que sólo se trabaja con oraciones simples (un verbo principal) sin frases subordinadas, extraídas del corpus. La figura 11 muestra las reglas sintácticas simplificadas que buscan el sujeto en una oración a partir de los segmentos obtenidos en la UC morfológica.

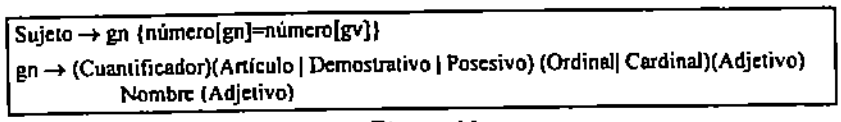

Figura 11

2.4 Unidad de Conocimiento Semántica

El conocimiento de esta unidad cognitiva es una estructura jerárquica que contiene los esquemas semánticos verbales. Esta jerarquía semántica se utiliza durante el análisis de las perspectivas P4 y P5 y proporciona los roles semánticos (Agente, Objeto, Beneficiario, Lugar, etc.) de los complementos sintácticos (Sujeto, Objeto Directo, Objeto Indirecto, etc.) de los verbos.

El estado del arte en clasificaciones verbales contiene diversos estudios que abarcan el análisis de oraciones centrado en el verbo. En Cook (1989) se describen varios enfoques a la gramática de casos (sintáctica y semántica) concluyendo que existe alguna posibilidad de descubrir correlaciones entre sintaxis y semántica que hagan corresponder el significado profundo de una oración con su expresión superficial. En esta línea se desarrollan los trabajos sobre clasificaciones descritos en Levin (1993).

Los casos o roles semánticos deben ser lo suficientemente generales para que puedan utilizarse en cualquier dominio, es decir, los roles específicos del dominio hacen que sean difíciles de migrar de un dominio a otro. Nuestra propuesta considera los siguientes roles semánticos requeridos por la valencia del verbo: Agente (Agt), Propiedad (Prop), Objeto (Obj), Beneficiario (Ben), Experimentador (Exp), Locativo (Loc) y Tiempo (Tm). En cuanto a los casos modales (no son esenciales para el significado del verbo) están Tiempo (Tm), Instrumento (Inst) y Modo (Mod) y no se tratarán en esta descripción.

Aunque tradicionalmente los verbos estativos no tienen roles semánticos asociados debido a que indican interrelaciones entre sujeto y atributos se manejaran utilizando Propiedad (Prop) y otros casos esenciales.

En una primera aproximación se propone la jerarquía semántica verbal basada en el modelo *case grammar matrix*, Cook (1989). En la figura 12 se muestra una parte simplificada de esta jerarquía. Se distinguen tres tipos de verbos según el tipo de evento que denotan (estado, proceso y acción). A su vez cada uno de estos tipos se descompone en varios dominios semánticos (básico, benefactivo, experimental, temporal y locativo). Estos dominios son excluyentes y un verbo sólo puede pertenecer a una de estas áreas semánticas. En la figura sólo aparece el dominio semántico *locativo* de los verbos de estado y de acción.

Cada hoja de la estructura jerárquica contiene los roles semánticos y sus funciones sintácticas asociadas cuando el verbo se utiliza en activa. Para estas funciones sintácticas se buscarán sus equivalentes en su esquema sintáctico correspondiente (por ejemplo, un ObjDir puede corresponderse con un CompAgente si la oración está en pasiva).

2.5 Unidad de Conocimiento del Léxico

Como muestra la figura 1 esta unidad cognitiva contiene conocimiento necesario para las UC Morfológica, Sintáctica y Semántica. Incluye los rasgos morfológicos, sintácticos y semánticos de las palabras.

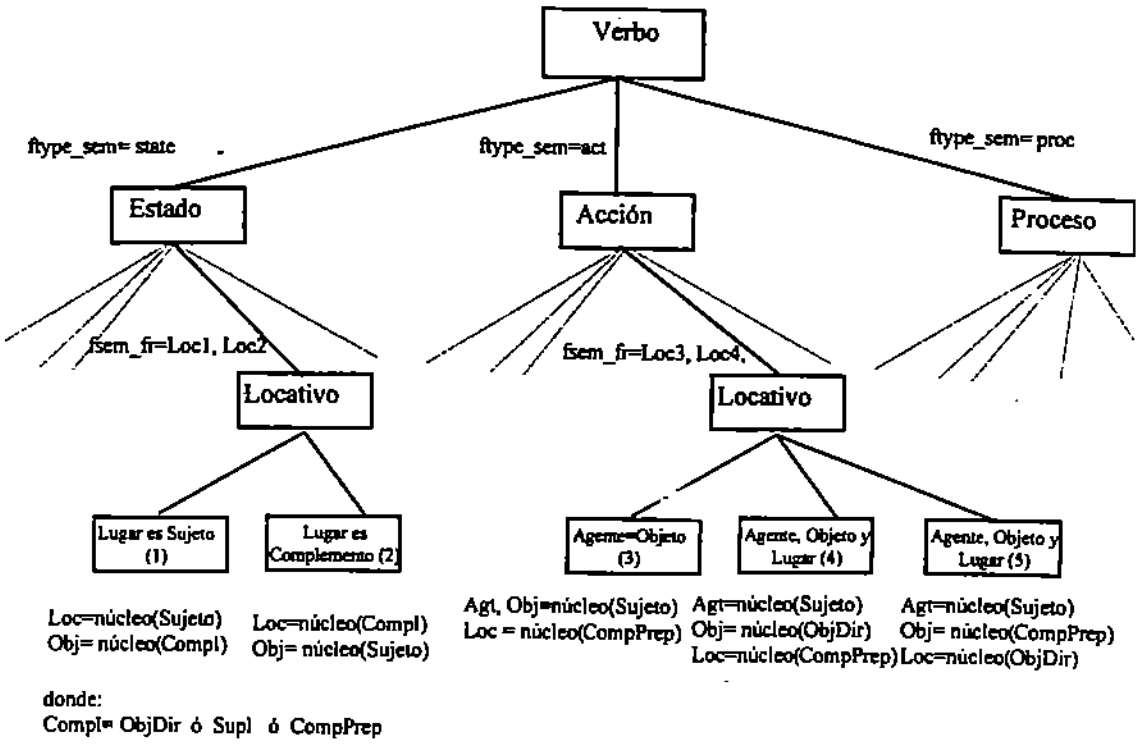

Figura 12

La información contenida en una entrada léxica verbal se muestra en la figura 13. Por ejemplo, la entrada léxica correspondiente al verbo incluir es la que se incluye en la figura 14.

```
entrada_verbal: nombre
   sx_features:
                  fpas: booleano                          (*capacidad para construcción pasiva*)
                  fexclpron: booleano                     (*verbo exclusivamente pronominal*)
                  fpron: booleano                         (*capacidad para uso pronominal*)
                  fmod: booleano                          (*capacidad para uso modal*)
                  faux: booleano                          (*capacidad para uso auxiliar*)
   list_sign:     {
                  sign:
                        frame_id:                         (*identificador significado*)
                        faux_ser: booleano
                        faux_estar: booleano
                        ftype_syn: valor                  (*tr. intr. bitr.....*)
                        fpneeded1: lista_valores          (*preposiciones para argumento 1*)
                        fpneeded2: lista_valores          (*preposiciones para argumento 2*)
                        ftype_sem: estado/acción/proceso  (*según tipo de evento*)
                        fsem_fr:valor   (*básico/experimental/temporal/benefactivo/ locativo *)   }*
                        f_pref: booleano
```

Figura 13

```
entrada_verbal: incluir
  sx_features:
        fpas: si
        fexclpron: no
        fpron: si
        fmod: no
        faux: no
  list_sign:     {
        sign:                                   sign:
                  frame_id: #1                            frame_id: #2
                  faux_ser: no                            faux_ser: si
                  faux_estar: si                          faux_estar: si
                  ftype_syn: tr                           ftype_syn: bitr2
                  fpneeded1: nil                          fpneeded1: nil
                  fpneeded2: nil                          fpneeded2: en
                  ftype_sem: estado                       ftype_sem: acción
                  fsem_fr: Loc1                           fsem_fr: Loc4
                  f_pref: si                              f_pref: no          }
```

Figura 14

Para navegar en las jerarquías verbales no sólo se utiliza el contenido de las entradas léxicas verbales sino también la información obtenida mediante la aplicación de reglas a los grupos verbales.

2.6 Unidad de Conocimiento del Dominio

Una vez realizado el análisis lingüístico de una oración del esquema descriptivo, la unidad cognitiva principal pasa el control a esta unidad para el análisis pragmático. Sobre una oración se pueden haber aplicado varios análisis de perspectivas (cada perspectiva cubre un segmento de la oración y puede haber intersecciones entre segmentos).

Como muestra la figura 2 el conocimiento del dominio se descompone en dos unidades: UC Pragmática y UC Caso. El conocimiento pragmático consiste en un conjunto de reglas (axiomas de correspondencia) para transformar el resultado del análisis lingüístico en conceptos de un modelo de datos. Se describe la tipología de conceptos utilizados en un modelo conceptual según la sintaxis del lenguaje DDL (Domain Description Language) utilizado en la metodología KADS, Schreiber (1993). La figura 15 muestra un subconjunto de la gramática conceptual definida utilizando notación de BNF. Esta gramática permite definir los conceptos de *clase*, *atributo*, *asociación*, *generalización*, etc. de un modelo orientado al objeto descrito en Marcos et al (1997).

La unidad de conocimiento del Caso contiene una estructura similar a una red semántica con los conceptos instanciados que se han ido adquiriendo a partir del texto.

Los axiomas pragmáticos de correspondencia utilizan el conocimiento lingüístico almacenado en el *frame oración* para obtener conceptos instanciados de la gramática conceptual anterior. En la figura 16 se muestra un axioma pragmático. La parte izquierda de la regla corresponde a condiciones sobre los rasgos lingüísticos (de cualquier tipo) de una perspectiva y la parte derecha a la instanciación de un concepto de la gramática conceptual. Puede ocurrir que los rasgos lingüísticos correspondientes a la instanciación de una perspectiva unifiquen con más de un axioma y consecuentemente produzcan más de un concepto instanciado o más de una instanciación de un concepto. Por ello, es necesario realizar una validación de consistencia con las interpretaciones de las otras perspectivas de la oración.

3 Un ejemplo

En apartados anteriores se ha mencionado que el resultado del análisis lingüístico de cada perspectiva se representaba en una estructura denominada *frame oración*. A continuación se describe esta estructura y como se completa para una oración: "Una titulación (nombre, plan) incluye varias asignaturas".

La tabla 2 muestra una parte del *frame oración* correspondiente a esta oración. Una vez que la UC morfológica etiqueta y segmenta el texto (slot *grupos de palabras*) y en la UC sintaxis se obtienen los atributos del GV (voz *activa* y verbo principal *incluir*), se procede a comprobar las perspectivas. En este caso, pueden ser aplicables las perspectivas del sublenguage P1 y P4. Centrándonos en P4, el verbo incluir es un verbo del sublenguaje y por ello posee una preferencia semántica en cuanto a los argumentos que rige. El slot *segmento input* muestra la secuencia de grupos de palabras que se pasan para el análisis de la perspectiva.

```
1 Clases
def-clase ⇒ clase : nombre-clase ;
                    [ atributos ; ]
                    [def-restricciones-clase ] .
2 Atributos
atributos ⇒ atributos : def-atributo ⟨ , def-atributo ⟩* .
def-atributo ⇒ atributo: nombre-atributo ;
                    [ def-valores ; ]
3 Dominio
def-valores ⇒ values : {valor-string ⟨ , valor-string ⟩*} )
4 Restricciones de clase y atributos
def-restricciones-clase ⇒ restricciones-clase :
                                  def-clave ;
def-clave ⇒ clave : nombre-atributo.
5 Asociaciones binarias
def-asociación-binaria ⇒ asociación-binaria : nombre-asociación ;

                                  participante 1 : def-participante ;
                                  participante 2 : def-participante ;
                                  [ atributos ;]
def-participante ⇒ nombre-clase ;
                    def-cardinalidad .
6 Restricciones de asociaciones
def-cardinalidad ⇒ cardinalidad : min nat max nat .
7 Asociaciones de grado superior
def-asociación-superior ⇒ asociación-superior : nombre-asociación ;
                                  participantes : def-participantes ;
                                  [ atributos ] .
def-participantes ⇒ def-participante , def-participante ⟨ , def-participante ⟩+ .
.....
```

Figura 15

```
unify(sem_att(main_v(v_type)), acción) ∧
unify(sem_att(main_(fr_class)), experimental) ∧
unify(sx_att(sx_schema), transitivo) ∧
→
asociación-binaria : mp_att(main_v(lex_inf)) ;
        participante 1 : mp_att(llsta_argumentos(sujeto(núcleo(lex))) :
        participante 2 : mp_att(lista_argumentos(do(núcleo(lex)))
```

Figura 16

El proceso para el tratamiento de la perspectiva P4 es SM-SX-PR. Con los rasgos semánticos de la entrada léxica *incluir*, en la jerarquía semántica de verbos se localiza el tipo *estado* (preferencia) y los roles semánticos asociados a Loc2: Loc y Obj (ver figura 12). Después se busca el esquema sintáctico con ayuda de la jerarquía sintáctica. Puesto que la oración es activa el esquema es Tr (Sujeto-ObjDir). Por último se realiza la correspondencia de estos complementos sintácticos con los que aparecen junto a los roles semánticos.

Los dos últimos slots del frame oración contienen los axiomas pragmáticos (sólo se muestra uno) y el(los) concepto(s) de la gramática conceptual (figura 15) instanciado(s) (no se muestran las cardinalidades).

La tabla 2 muestra el frame oración con la información lingüística contenida en una instancia de la perspectiva P4. Se han omitido los rasgos concernientes a los modificadores de los núcleos de los complementos verbales.

Frame Oración

Oración	*Una titulación (nombre. plan) incluye varias asignaturas.*
Grupos de Palabras	*GN1[arcafs(una), ncfs(titulación)] PM2[lp(( )] GN3[ncms(nombre)] PM4[cm(.)] GN5[ncms(plan)] PM6[rp( ))] GV7=[vlpi3s(incluye)] GN8[quxfp(varias), ncfp(asignaturas)]* *GN1 PM2 GN3 PM4 GN5 PM6 GV7 GN8*
Perspectiva	**P4**
Proceso	conducido por la semántica
Palabras	*Una titulación incluye varias asignaturas*
Segmento entrada	*GN1 GV7 GN8*

LINGUISTIC KNOWLEDGE

Morphologic Kw

```
mp_att:
  main_v:
    lex: incluye
    lex_inf: incluir
    tiempo: presente ind.
    número: sing
    persona: 3ª
  aux:
    –
  modal:
    –
  lista_argumentos: {
    sujeto:
        núcleo:
          núcleo_id: H1
          lex: titulación
          category: nombre
          number: sing
        lista_mods:
          ...
    do:
        núcleo:
          núcleo_id: H8
          lex: asignatura
          category: nombre
          number: plur
        lista_mods:
          ...
    io:
      –
    atrib:
      –
    supl:
      –
    cc:
      –
  }
```

Syntax Kw

```
sx_att:
  sx_esquema: transitivo
  voz: activa
  lista_argumentos: {
    sujeto:
      núcleo: H1
      seg: GN1
      lista_mods:
      .....
    do:
      núcleo: H8
      seg: GN8
      lista_mods:
      .....
    io:
      –
    atrib:
      –
    supl:
      –
    cc:
  }
```

Semantic Kw

```
sem_att:
  main_v:
    v_type: estado
    fr_class: Loc1
  lista_argumentos: {
    sujeto:
      role_sem: Loc
      núcleo:
        núcleo_id: H1
        sem_núcleo: común, abstracto
      lista_mods:
      .....
    do:
      role_sem: Obj
      núcleo:
        núcleo_id: H8
        sem_núcleo: común, abstracto
      lista_mods:
      .....
    io:
      --
    atrib:
      –
    supl:
      –
    cc:
      --
  }
```

Pragmatic Kw

```
unify(sem_att(main_v(v_type)), estado) ∧
unify(sem_att(main_(fr_class)), Loc1) ∧
unify(sx_att(sx_esquema), transitivo) ∧
.....
→
asociación-meronimia : mp_att(main_v(lex_inf)) ;
                       holónimo : mp_att(lista_argumentos(sujeto(núcleo(lex))) ;
                       merónimos :
                           merónimo : mp_att(list_argumentos(do(núcleo(lex))) ;
```

NON LG KW — **Conceptual Design**

```
asociación-meronimia : incluir;
                       holónimo : titulación ;
                       merónimos :
                           merónimo : asignatura ;
```

Tabla 2

4 Conclusiones

En este artículo se ha propuesto una estructuración del conocimiento lingüístico para el desarrollo de aplicaciones de tratamiento automático del lenguaje natural, con objeto de abordar de forma estructurada y metodológica el tratamiento del conocimiento multiforme y de gran complejidad que este tipo de sistemas requiere.

La arquitectura basada en unidades cognitivas permite definir áreas con tipos específicos de conocimiento, incluyendo sus funcionalidades asociadas y declarando explícitamente la interacción entre ellas

El enfoque propuesto basado en perspectivas lingüísticas permite el control dinámico de los distintos procesos lingüísticos durante el análisis del texto, con objeto de utilizar la información más prometedora en cada paso, empleando tanto técnicas de análisis superficial como técnicas de análisis más profundo. Por otro lado, este enfoque permite la posibilidad de una instrumentación paralela y la revisión del conocimiento por parte de los expertos.

En la actualidad se está trabajando en la operacionalización de las distintas unidades de conocimiento y en la inclusión de otros fenómenos lingüísticos, como es el caso de las ambigüedades que surgen en algunos procesos de análisis explicados, no planteados hasta este momento.

REFERENCIAS

Abney (1996), S. Abney. "Part-of-speech tagging and partial parsing." En Corpus-Based Methods in Language and Speech. An ELSNET book, Kluwer Academic Publishers, Dordrecht, 1996.

Burg y Van de Riet (1996), J.F.M. Burg y R.P. van de Riet. "Analyzing Informal Requirements Specifications: A First Step towards Conceptual Modeling". En Applications of Natural Language to Information Systems. Editado por R.P. van de Riet, J.F.M. Burg y A.J. van der Vos, IOS Press, 1996.

Cook (1989), Water A. Cook. Case Grammar Theory. Georgetown University Press, Washington, D.C, 1989.

Cuena y Molina (1996), J. Cuena y M. Molina. KSM: An environment for Design of Strucured Knowledge Models. To be published as chapter of the book "Knowledge-Based Systems-Advanced Concepts, Techniques and Applications". Edited by Spyros G. Tzafestas. Publisher: World Scientific Publishing Company, 1996.

Dalh (1993), Dalh, "What the study of language can contribute to AI". AI Communications, 6, 2, 1993.

De Miguel et al. (1996), A. De Miguel, M. Piattini, P. Martínez y E. Marcos. "ENEAS/BD: Un Entorno para la Enseñanza Avanzada de Sistemas de Bases de Datos". Primeras Jornadas de Investigación y Docencia en Basesde Datos, La Coruña, Junio 1996.

Delisle et al. (1996), S. Delisle, K. Barker, T. Copeck y S. Szpakowicz."Interactive Semantic Analysis of Techical Texts". Computational Intelligence, 12 (2), 273-306, 1996.

Jacobs y Rau (1993), P. S. Jacobs y L. F. Rau, "Innovations in text interpretation." Artificial Intelligence 63, 143-191, 1993.

Levin (1993),B. Levin, English Verb Classes and Alternations, The University of Chicago Press, 1993.

Marcos et al. (1997), "SQL3/ODMG-93 integration through MIMO". E. Marcos, A. de Miguel, M. Piattini, P. Martínez. Aceptado en BIWIT'97 (Third Basque International Workshop on Information Technology). Biarritz (France), 2-4 de julio, 1997.

Sánchez-León y Nieto (1995), F. Sánchez-León y A. Nieto. "CRATER - Corpus Resources and Terminology Extraction". WP6 - Public Domain POS Tagger for Spanish, November, 1995.

Schreiber (1993), G. Shreiber, "KADS: Domain Description Language". En KADS: Principled approach to knowledge based system development. Editado por G. Shreiber, B. Wielinga y J. Breuker. Academic Press, Cambridge, 1993.

Sparck Jones (1996), K. Sparck Jones, Evaluating Natural Language Processing Systems. NL AI Series, 1996.