ESTRUCTURACIÓN DEL CONOCIMIENTO PARA LA INTERPRETACIÓN DE TEXTOS Y SU APLICACIÓN AL DISEÑO DE ESQUEMAS CONCEPTUALES DE BASES DE DATOS

Paloma Martínez

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

Ana Mª García-Serrano

Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid
agarcia@dia.fi.upm.es

Adoración de Miguel

Departamento de Informática,
Universidad Carlos III de Madrid
admiguel@inf.uc3m.es

Abstract

El desarrollo de sistemas genéricos para tratamiento automático del lenguaje está limitado por la imposibilidad de tener disponible todo el conocimiento requerido para cualquier dominio de aplicación. Por ello, la solución propuesta en este trabajo se basa en el desarrollo de un sistema modular y multiforme que permita la incorporación de los distintos tipos de conocimiento lingüístico y extralingüístico. Con este fin, se ha elaborado una propuesta general de estructuración del conocimiento para la interpretación de textos, que proporciona un mecanismo de control flexible de los distintos procesos lingüísticos identificados y cuya funcionalidad se lleva a cabo empleando distintos tipos de tecnología de procesamiento del lenguaje natural (PLN). Para mostrar la potencialidad de la arquitectura definida se ha utilizado el dominio de aplicación del modelado conceptual de bases de datos (BD) a partir de textos.

1. Apuntes históricos de la tecnología del lenguaje: enfoques racionalista, empírico e híbrido

La primera generación (1950-1975) y la segunda generación (1975-1985) de trabajos en LN estuvieron dominadas por un enfoque cognitivo en la tradición de la lingüística generativa. En la tercera generación (1985-actualidad) se une a esta corriente racionalista un enfoque motivado por la cobertura empírica, con colecciones de datos reales (corpus) como evidencia primaria. Cada enfoque tiene sus propios tipos de teoría, métodos y criterios para la consecución de sus objetivos.

La confrontación entre ambas aproximaciones impidió una interacción productiva, pues la lingüística computacional estaba dominada por la perspectiva teórica de la gramática generativa, hostil a los métodos cuantitativos, así como por la comunidad del procesamiento del habla, dominada por la teoría de la información estadística, discrepante de la lingüística teórica.

Los primeros contactos entre estas corrientes, hasta entonces contrapuestas, se produjeron a primeros de los noventa, debido sobre todo a un cambio en la política de financiación y a una proliferación de textos electrónicos. El resultado fue que se ampliaron los objetivos y se abrieron nuevas líneas de tratamiento del lenguaje con nuevos objetivos. Además, problemas clásicos en este campo, como la identificación de las categorías léxicas de las palabras de una oración o la ambigüedad de los modificadores preposicionales durante el análisis sintáctico, se abordaron con métodos estadísticos, procedentes de la comunidad del habla, a menudo con éxito.

Aunque las técnicas empíricas específicas pueden variar, todas ellas se basan en la idea de inducir el conocimiento necesario para resolver un problema específico mediante el análisis estadístico de grandes corpus de texto real en vez de construir tal conocimiento simbólicamente, como en la corriente racionalista.

Aparte de la corriente racionalista y empírica, la tercera generación dió origen a una aproximación híbrida en el estudio y tratamiento del LN. Los factores que han obligado a este planteamiento híbrido del PLN son, Abney (1996a):

- La constatación de que cada comunidad se beneficia de los trabajos de la otra. Los enfoques cuantitativos añaden robustez y cobertura a los sistemas PLN cualitativos, al permitir, por ejemplo, la adquisición (semi)automática de conocimiento léxico. Por otro lado, los métodos estadísticos inductivos funcionan mejor si están sesgados a priori con conocimiento preciso.
- La mejora de recursos hardware (ordenadores rápidos, abaratamiento de los dispositivos, distribución de datos en CDROM, etc.), la disponibilidad de grandes volúmenes de textos en formato electrónico debido al uso masivo de los ordenadores (publicaciones electrónicas, bases de datos, etc.), así como las iniciativas financiadas para recolección de datos.
- El desplazamiento de las aplicaciones para dominios limitados hacia aplicaciones que trabajan sobre textos reales. Es necesario satisfacer algunas demandas reales del mercado: grandes vocabularios para reconocimiento del habla, traducción automática de texto libre para servicios en línea, así como recuperación de información en WWW. Aplicaciones de este tipo deben enfrentarse con entradas imprevisibles de los usuarios que no están familiarizados con la tecnología. Todo ello exige sistemas con soluciones a gran escala.

- Por último, queda pendiente el análisis de la influencia sobre la investigación científica pura de la lingüística computacional, mediante la construcción de modelos que explícitamente tengan en cuenta la variabilidad y la incertidumbre y no sólo con el fin de salvar las lagunas de conocimiento simbólico, como se emplean en el sistema Japangloss, Knight et al. (1995).

Por lo tanto, no sólo no existe contradicción en la combinación de ambos enfoques sino que, además, esta combinación es productiva. Es un hecho que todo uso de la estadística se fundamenta en un modelo simbólico, no importa de qué aplicación se trate. Para el lenguaje en particular, las unidades naturales que se manipulan en cualquier modelo estadístico son constructores discretos tales como fonemas, morfemas y palabras, así como interrelaciones discretas entre estos constructores tales como adyacencia superficial o relaciones predicado-argumentos. Las valoraciones numéricas sólo son significativas en el contexto de soporte a un modelo simbólico y, a la inversa, los soportes simbólicos aislados no son suficientes para captar la variabilidad inherente que se da en los datos lingüísticos reales. Desde este punto de vista, no existen métodos “puramente” estadísticos.

Los factores anteriores se han concretado en la combinación de las aproximaciones racionalista y empírica en un enfoque híbrido con las siguientes características:

- Análisis Intensional y Análisis Extensional
- Modelos Probabilistas y Conocimiento basado en reglas
- Mejora de los tiempos de desarrollo
- Reutilización
- Amplia cobertura y Análisis profundo

Los sistemas basados en reglas se basaban mayormente en el tratamiento de oraciones individuales, para análisis completos en dominios muy restringidos (para un conjunto limitado de oraciones) o incluso artificiales y no soportaban mecanismos de recuperación ante entradas desconocidas. La demanda de nuevas aplicaciones que exigen tratamiento de grandes volúmenes de texto, como recuperación e indexación de documentos, extracción de información, etc. hace que los nuevos trabajos empleen técnicas superficiales y métodos cuantitativos. Se puede conseguir una comprensión suficiente del LN mediante análisis superficiales y parciales para ciertas tareas, intentando evitar así los problemas de cobertura. La clave para alcanzar los objetivos necesarios está en realizar análisis de textos que combinen tanto análisis superficiales, pero con una amplia cobertura de textos reales, junto con análisis profundos, cuando sea necesario

Observando las distintas tendencias a lo largo de las tres generaciones (racionalista, empírica e híbrida) de investigación y desarrollo en PLN, se pueden extraer las siguientes reflexiones concernientes a esta área.

En términos de lo que el tratamiento del lenguaje requiere, especialmente el tratamiento del lenguaje de objetivo general, la mayor parte del progreso se ha producido en el área de la sintaxis, donde se tienen medios efectivos para la caracterización de las gramáticas así como técnicas potentes de análisis, Sparck-Jones (1996).

Gran parte de los sistemas desarrollados para tratamiento de grandes corpus de texto, FASTUS, Appelt et al. (1993), PLUM, Weischedel et al. (1992), y otros están demasiado orientados a la funcionalidad requerida (extracción de información de dominios específicos) por lo que se impone una filosofía de sistemas basados en conocimiento que permita que el conocimiento lingüístico y no lingüístico

sea utilizable por distintas aplicaciones. Relacionado con este aspecto de reutilización, se han desarrollado grandes bases de datos léxicas de propósito general que fomentan la reutilización del conocimiento como son ACQUILEX, Calzolari et al. (1994), Wordnet, Miller (1995), EuroWordnet, Vossen et al. (1997), y otros. Incluso se han creado organizaciones como ELRA (European Language Resources Association) cuyo objetivo es promover la creación, verificación y distribución de recursos lingüísticos en Europa tales como léxicos, gramáticas, corpus de textos, datos terminológicos, etc. esenciales para el desarrollo de sistemas PLN de gran cobertura y cuyo coste en muchos casos es prohibitivo para las empresas.

El problema de la ampliabilidad todavía no ha sido resuelto, pues los sistemas se construyen a modo de prototipos y cualquier intento por aumentar el conocimiento de que disponen o por salvar las lagunas de conocimiento resulta de gran dificultad. Los trabajos expuestos en la corriente híbrida como los realizados en traducción automática por el sistema Japangloss, Knight et al. (1996), o en léxicos para aplicaciones PLN como el sistema ARIOSTO_LEX, Basili et al. (1996), demuestran la posibilidad de combinar conocimiento explícito simbólico y métodos puramente cuantitativos que permiten aprender modelos predictivos del lenguaje a partir de recursos on-line con el fin de mejorar la robustez y cobertura en determinados dominios.

Mientras que las arquitecturas altamente modulares al estilo de TANKA, Delisle et al. (1996), son ampliamente aceptadas, todavía existen problemas para determinar la distribución de la información y del esfuerzo entre los elementos lingüísticos y no lingüísticos de un sistema, así como entre los componentes de propósito general y los específicos del dominio. Una de las principales áreas de investigación sobre PLN en la actualidad es el desarrollo de modelos paralelos que utilicen una gran variedad de información lingüística como los sistemas CARAMEL, Sabah (1993), y PARSETALK, Hahn et al. (1994). Esto involucra el diseño de estructuras de datos apropiadas que representen la interconexión de los distintos tipos de conocimiento lingüístico y el diseño de algoritmos que busquen y apliquen el conocimiento apropiado en el momento adecuado durante el proceso evitando bloquear el sistema. Por ello, es necesario desarrollar modelos computacionales avanzados completos y más cercanos a la realidad psicolingüística que eliminen las deficiencias producidas por los enfoques convencionales con control fijo.

Los entornos software para investigación, definición y desarrollo de tecnología PLN constituyen otro aspecto al que se ha dedicado especial atención. Plataformas como GEPPETTO, Ciravegna et al. (1997), ALEP, Simpkins (1994) y GATE, Cunningham et al. (1997), ponen de manifiesto la necesidad de considerar el desarrollo tanto de recursos como de aplicaciones PLN como una nueva disciplina denominada *Ingeniería Lingüística*. Ello supone investigar en metodologías que cubran el ciclo de vida de las aplicaciones que involucran la tecnología PLN, en entornos de desarrollo que faciliten su aplicación así como en métodos de conceptualización adecuados.

En la revisión de los trabajos actuales en PLN, se muestra cómo la mayoría pueden incluirse en la aproximación híbrida de la tercera generación y de ello dan fe las series ACL 96, 97 y 98 así como las conferencias ANLP de los mismos años. Estos trabajos muestran que, en la actualidad, coexisten dos líneas de investigación; la corriente teórica busca formalismos híbridos que den

cabida tanto al conocimiento lingüístico racionalista o simbólico (enfoque cualitativo) como al empírico o basado en corpus (enfoque cuantitativo) siempre motivada lingüísticamente. La corriente en el marco de la Ingeniería Lingüística busca soluciones adecuadas a problemas reales, centrándose en la definición de arquitecturas para sistemas PLN lo suficientemente flexibles para alcanzar la robustez requerida y combinando tanto técnicas superficiales como profundas. Esta corriente también está interesada en la definición y construcción de plataformas a modo de herramientas CASE que faciliten las labores de análisis, diseño e implementación de aplicaciones que implican tratamiento automático del lenguaje.

En relación con la segunda vertiente y con el fin de comprender el término "Ingeniería Lingüística", se definirá en primer lugar el término Ingeniería. "*Ingeniería* es la aplicación de un enfoque sistemático, disciplinado y cuantificable a estructuras, máquinas, productos, sistemas o procesos. Se lleva a cabo en respuesta a las necesidades y deseos humanos percibidos y utiliza conocimiento, principios, técnicas y métodos derivados tanto de la ciencia como de la experiencia", Thomé (1993). Las características esenciales de la ingeniería son:

- *orientada al objetivo* (resultado final físico bien definido y previsto)
- *centrada en el hombre* (deriva sus objetivos de las necesidades y deseos del hombre y por ello debe considerar las condiciones humanas y su entorno)
- *sistemática* (se construye sobre principios, técnicas y métodos y utiliza procesos)
- *creativa* (cambia el entorno del hombre mediante la creación de nuevos sistemas)

Según la definición anterior, la Ingeniería Lingüística se centra en los métodos, técnicas y aplicaciones necesarias para conseguir resultados prácticos del modelado del uso del lenguaje. La Ingeniería Lingüística trata todos los aspectos relacionados con la aplicación del conocimiento sobre el lenguaje para proporcionar soluciones informáticas. Su objetivo es crear productos software que incorporen parte del conocimiento del lenguaje y que se requieren para mejorar la interacción hombre-máquina, la recuperación y extracción de información, etc. Estas necesidades crecientes hacen que se haya producido un desplazamiento desde los primeros sistemas PLN "artesanos" hacia un enfoque de ingeniería a nivel de producción no académica (comercial).

Estableciendo un símil con la disciplina de la Ingeniería del Software, la Ingeniería Lingüística también contempla ciclos de vida de desarrollo de sistemas así como metodologías y técnicas de análisis y diseño, en las que se fomente la participación de los usuarios y expertos, siempre teniendo en cuenta sus características propias y diferenciadoras.

Un producto de la Ingeniería Lingüística incluye el reconocimiento y validación automática del lenguaje en forma hablada o escrita; el análisis de la entrada para alcanzar niveles de comprensión adecuados (interpretación); la aplicación de esta interpretación para generar el resultado requerido (por ejemplo, convertir un texto hablado en escrito en un sistema de dictado o equiparar una solicitud de información con los contenidos de una BD); y, por último, la salida por ordenador en forma hablada, visualizada o impresa. En la actualidad, existen diversos productos comerciales que incluyen correctores ortográficos y gramaticales, traductores automáticos, sistemas de dictado y paquetes de recuperación de información que

muestran los resultados de la Ingeniería del Lenguaje.

De esta reflexión sobre el estado de la cuestión se puede derivar la propuesta planteada en este trabajo que aborda el tratamiento lingüístico de un corpus genérico justificándose el tipo de diseño y desarrollo con tecnología que integra aspectos tanto de la Ingeniería del Conocimiento como de la Ingeniería del Software.

2. Objetivos del trabajo

Es precisamente en la línea de la Ingeniería Lingüística en la que se encuadra este trabajo, Martínez (1998), mediante la propuesta de una arquitectura cognitiva que incluye conocimiento lingüístico y de control para el análisis de textos. Su funcionalidad se lleva a cabo empleando distintos tipos de tecnología PLN empleando un mecanismo de control flexible de los distintos procesos lingüísticos. Se plantea una aproximación híbrida que combina técnicas de análisis superficiales y parciales dirigidos por expectativas con el fin de mejorar la robustez y la cobertura, junto con técnicas de análisis más profundas.

En esta propuesta de modularización del conocimiento lingüístico se ha definido un conjunto de perspectivas lingüísticas que combinan las fuentes de conocimiento para guiar el proceso de análisis sin fijar una secuencialidad a priori. Se introduce así la interpretación parcial mediante una forma de actuación no determinista dirigida por una serie de “pistas” incluidas en el propio texto y que hemos denominado perspectivas lingüísticas.

Dicha arquitectura reflejará fielmente la estructuración del conocimiento identificado 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 en el que intervienen usuarios, informáticos y lingüistas en los distintos módulos del sistema.

El problema que se va a resolver en un dominio concreto es el modelado conceptual de BD a partir de especificaciones textuales. No se propone un modelo teórico para la combinación de conocimientos de distinto tipo sino un modelo cognitivo que hace uso de tecnología PLN existente y que busca cierta flexibilidad a la hora de utilizar las fuentes de conocimiento de que dispone.

La conceptualización del conocimiento lingüístico que se requiere para tratamiento automático del lenguaje con un objetivo específico aporta propuestas de solución a algunos de los aspectos mencionados anteriormente:

- aportación metodológica al desarrollo de sistemas que incorporan tecnología PLN
- combinación de técnicas de tratamiento del LN siguiendo la aproximación híbrida
- propuesta de control no fijo guiado por las fuentes de información lingüística
- fomento de la reutilización de conocimiento
- Localización de los aspectos relativos al dominio

3. Metodología para el desarrollo de aplicaciones basadas en el conocimiento

El objetivo principal es elaborar 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. Se ha seguido un enfoque metodológico para el desarrollo de aplicaciones basadas en

conocimiento soportado por el entorno KSM (Knowledge Structure Manager), Cuena y Molina (1996). KSM permite la operacionalización directa de un modelo cognitivo y se ha empleado con éxito en la implementación de otro tipo de aplicaciones.

Desde mediados de los ochenta se viene investigando en metodologías para sistemas basados en conocimiento. Estos sistemas son también software pero con algunas características específicas que no aparecen en el software tradicional de gestión; por ejemplo, los objetos son ideas y conceptos que se ajustan a una descripción de tipo simbólico y los procesos son más complejos que los procesos algorítmicos llevados a cabo en un sistema de gestión.

El modelo de representación basado en unidades cognitivas KSM propone una metodología de diseño orientado al conocimiento que descompone el desarrollo de un sistema en tres fases: Formulación del modelo a nivel de conocimiento, selección de la representación del conocimiento y, por último, la instanciación en un dominio.

Es posible establecer un paralelismo con las metodologías tradicionales de ingeniería del software en las que se descompone el ciclo de vida de un sistema en análisis, diseño e implementación. Por ejemplo, en el desarrollo de BD, en la fase de análisis se construye un esquema E/R que representa los datos que manejará el sistema, en la fase de diseño se decide si se empleará un gestor relacional, en red o jerárquico y, finalmente, en fase de implementación se selecciona el producto comercial que soportará la base de datos. En la metodología KSM, en la fase de formulación del modelo a nivel de conocimiento se definen las distintas unidades de conocimiento y las relaciones entre ellas; en la fase de selección de la representación del conocimiento se decide para cada unidad su representación simbólica (por ejemplo, marcos, redes semánticas, reglas de producción, etc.); por último, en la fase de instanciación del conocimiento del dominio se incorpora el conocimiento concreto de cada una de las bases de conocimiento con el fin de obtener un sistema operativo.

A partir de ahora se distinguirá entre la metodología orientada al conocimiento KSM y el entorno software KSM que proporciona soporte a la definición y operacionalización de los modelos de conocimiento.

La metodología KSM aborda la construcción de sistemas basados en el conocimiento desde una perspectiva distinta a, aunque no contradictoria con, la ingeniería del software tradicional. Por ejemplo, KSM persigue los mismos objetivos que otras metodologías de desarrollo software (SSADM, Merise, Métrica, etc.): desarrollo de sistemas modular, incremental y multiforme a nivel de representación de conocimiento, integrando tanto aspectos estáticos (o declarativos) como dinámicos.

La metodología KSM combina diferentes aspectos de metodologías de análisis y diseño estructurado (fases de análisis, diseño, implementación y mantenimiento), conceptos de la orientación al objeto (encapsulamiento de datos y procesos) y de desarrollo rápido de aplicaciones (el modelo definido en la primera fase es ejecutable una vez se ha instanciado el conocimiento necesario del dominio).

La pieza clave del modelado con la metodología KSM es la unidad cognitiva (UC), que representa a un ente inteligente que dispone de un determinado conocimiento, a partir del que es posible la ejecución de diferentes tareas. Cada unidad cognitiva (Figura 1) se compone de áreas de conocimiento que a su vez engloban a otras unidades cognitivas o bien componentes

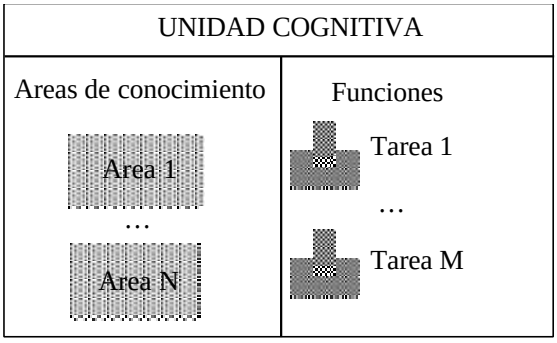


Figura 1: Formato de Unidad Cognitiva

básicos denominados unidades primarias por incorporar conocimiento de base genérico o bien específico de un dominio. En la descripción del modelo se ha hecho uso de la notación de la Figura 1 para reflejar la estructuración del conocimiento. Las tareas o componentes operacionales asociados a las unidades cognitivas determinan la forma de combinación del conocimiento para conseguir la funcionalidad perseguida por la tarea.

La metodología KSM permite describir una aplicación en varios niveles:

1. El **nivel de conocimiento**, en el cual debe definirse *qué se sabe* (conocimiento) y *cómo se usa* ese conocimiento (tareas), tanto de forma genérica como de un dominio en particular. Este modelo genérico posee tres vistas:
   - **Vista del conocimiento**: Estructura jerárquica de áreas de conocimiento (jerarquía de inclusión) que organiza el conocimiento.
   - **Vista de tareas**: Una tarea es una forma de resolución de problemas soportada por una unidad cognitiva (cada unidad cognitiva puede proporcionar diversas tareas o funcionalidades). Cada tarea es realizada por un método que representa el modo en que se resuelve una tarea (estrategia de control). KSM proporciona un lenguaje (LINK) para la definición de los métodos.
   - **Vista conceptual**: Formada por varios vocabularios conceptuales pertenecientes a las unidades cognitivas primarias con el fin de compartir conceptos entre bases de conocimiento distintas.
2. 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 definidas en el modelo de conocimiento.
3. **Nivel de implementación e instanciación**, en el que el modelo de conocimiento del nivel simbólico se instancia con conocimiento del dominio de aplicación.

Una contribución importante a la modularidad del conocimiento es que una vista del conocimiento se corresponde, generalmente, con varias vistas de tareas debido a que una unidad cognitiva puede proporcionar varias funcionalidades, es decir, un modelo que dispone de una vista de áreas de conocimiento admite varias vistas de tareas, una por cada tarea global que el sistema puede ofrecer.

3.1 Criterios metodológicos para la estructuración del conocimiento lingüístico

Cualquier sistema de comprensión de lenguaje natural tiene como objetivo genérico pasar de una representación origen a una representación final que se integrará en otro sistema con funcionalidad propia, Rich y Knight (1991). El desarrollo de estos sistemas debe contemplar los siguientes aspectos: arquitectura de control, fuentes de conocimiento y su representación, interrelaciones entre esas fuentes de conocimiento y técnicas que se utilizarán para llevar a cabo la funcionalidad

requerida. Todos estos aspectos responden a las siguientes preguntas:

- ¿cuáles son las características que debe poseer el sistema?
- ¿con qué tipo de textos funciona?
- ¿cuál es la funcionalidad requerida?
- ¿cómo se consigue la funcionalidad requerida?
    - ¿cuáles son los procesos que conducen a esta funcionalidad?
    - ¿cómo se encadenan estos procesos?
    - ¿cómo se realiza cada uno de ellos?
- ¿qué conoce el sistema para llevar a cabo su funcionalidad?
    - ¿cómo se relacionan los bloques de conocimiento entre sí?
    - ¿cómo se representa el conocimiento?

Todas estas cuestiones se van respondiendo a lo largo de las fases del ciclo de vida de un sistema PLN (análisis, diseño e implementación). Estos aspectos se detallan a continuación y su estudio ha ayudado al diseño del modelo de conocimiento.

Los sistemas que incorporan PLN son sistemas cuyo conocimiento es *heterogéneo* e *interdisplicinar* (distintas fuentes de conocimiento lingüístico y no lingüístico) y *complejo* (cada una de las fuentes de conocimiento lo es). Además, es necesario manejar la incertidumbre y la variabilidad del conocimiento. Todo ello requiere una conceptualización adecuada capaz de interactuar con los expertos tanto lingüistas como informáticos.

En cuanto al tipo de textos que se han utilizado se aclararán primero algunos conceptos. Se denomina *Lenguaje* a un sistema de comunicación o capacidad de los humanos para comunicarnos y *Lengua*[1] a un sistema organizado de signos orales o escritos, regulados por un código denominado *gramática*, propio de un pueblo o nación o común a varios. También se conoce como lengua al vocabulario y gramática peculiares de una época, de un escritor o de un grupo social.

Centrándonos en esta segunda definición de lengua, cuando se usa el lenguaje para hablar sobre un dominio restringido, y, en concreto, cuando lo utiliza una comunidad de hablantes que comparten un conocimiento especializado, la lengua frecuentemente adopta características distintas del lenguaje en su totalidad. Puede ser más restringida en cuanto a sus propiedades léxicas, sintácticas, semánticas o discursivas; puede incluir también expresiones que no estén presentes en la lengua común. La lengua resultante se denomina sublengua o sublenguaje. Esta distinción entre lenguaje común y sublenguaje es necesaria con el fin de definir el tipo de textos con los que se trabaja.

Por otro lado, el estudio del dominio permitirá descubrir el conocimiento concreto que servirá para particularizar el modelo de conocimiento definido. Determinadas aplicaciones PLN tratan con textos de un dominio que poseen, además, ciertas características superficiales según el tipo de texto (narrativa, descriptivo, diálogos, etc.), por lo que el conocimiento sintáctico, semántico, etc. se completa con información relativa al dominio concreto y al tipo de textos. Por otro lado, también es necesario incluir conocimiento que no sea dependiente del dominio sino más general, es decir, conocimiento sobre el lenguaje común.

Así, se plantea el desarrollo de un sistema basado en el conocimiento con las siguientes características:

[1] En la disciplina del PLN se considera a menudo ambos términos, lenguaje y lengua, como sinónimos.

- *Modularidad*: Se estructura el conocimiento de distinta naturaleza y complejidad (conocimiento declarativo, conocimiento de control y estrategias de inferencia) y razonamientos de distinto tipo, conviviendo con principios de software convencionales mediante la integración de aproximaciones procedentes de diversas tecnologías (simbólica, conexionista, empírica, etc.).
- *Flexibilidad en la consulta y mantenimiento del conocimiento*: Diseño que sea fácilmente revisable por expertos lingüistas e informáticos ya que se separan los aspectos de análisis, diseño e implementación del sistema, lo que permite, además, una mayor comprensión de la actuación del sistema para la evaluación de sus capacidades.
- *Portabilidad*: Se distinguen dos niveles de reutilización: de la configuración de sistema y de módulos software. Por un lado, la arquitectura final es reutilizable en distintos dominios intercambiando el conocimiento específico del dominio ya que, como se verá después, la definición de una estrategia de interpretación basada en perspectivas lingüísticas es fácilmente rediseñable. Por otro lado, los expertos pueden decidir la sustitución de determinados módulos de conocimiento e incluso la inserción o eliminación de otros.

En este trabajo se ha optado por una organización sobre la base de los niveles de descripción lingüística convencionales: morfología (conocimiento sobre flexión nominal, conjugación, categorías morfológicas de las palabras), sintaxis (conocimiento sobre las restricciones entre las categorías de las palabras que dan lugar a oraciones correctas), semántica (conocimiento sobre el significado de las palabras y la relación entre estos significados) y pragmática (conocimiento que relaciona el conocimiento del mundo y del lenguaje). Esta descomposición busca una agrupación del conocimiento del mismo tipo en áreas separadas. Esta separación no es del todo real, pues están integradas a un nivel superior mediante el concepto de las perspectivas lingüísticas. Cada una de estas áreas se descompone a su vez en áreas más específicas y, por tanto, más manejables.

En cuanto al dominio de modelado conceptual de BD seleccionado para desarrollar esta propuesta, los textos describen diferentes UoD, es decir, no se ajustan a un dominio específico por lo que hay que tener en cuenta los dos tipos de lenguaje: el sublenguaje empleado en los textos descriptivos sobre BD y el lenguaje de uso común.

Para el estudio de este dominio se ha recolectado un corpus de aproximadamente 60 textos descriptivos de enunciados para modelado de BD que han permitido extraer información morfológica, sintáctica, semántica y pragmática que se utilizará para completar los módulos de conocimiento de cada tipo.

3.2 Modelo estructurado para la interpretación de textos

Se distinguen dos fases bien diferenciadas en la extracción de conocimiento a partir de textos descriptivos: la primera realiza una interpretación lingüística del texto y la segunda emplea el resultado del análisis lingüístico para buscar la correspondencia con los conceptos del dominio que se pretende adquirir, en nuestro caso, el dominio de modelado conceptual de BD.

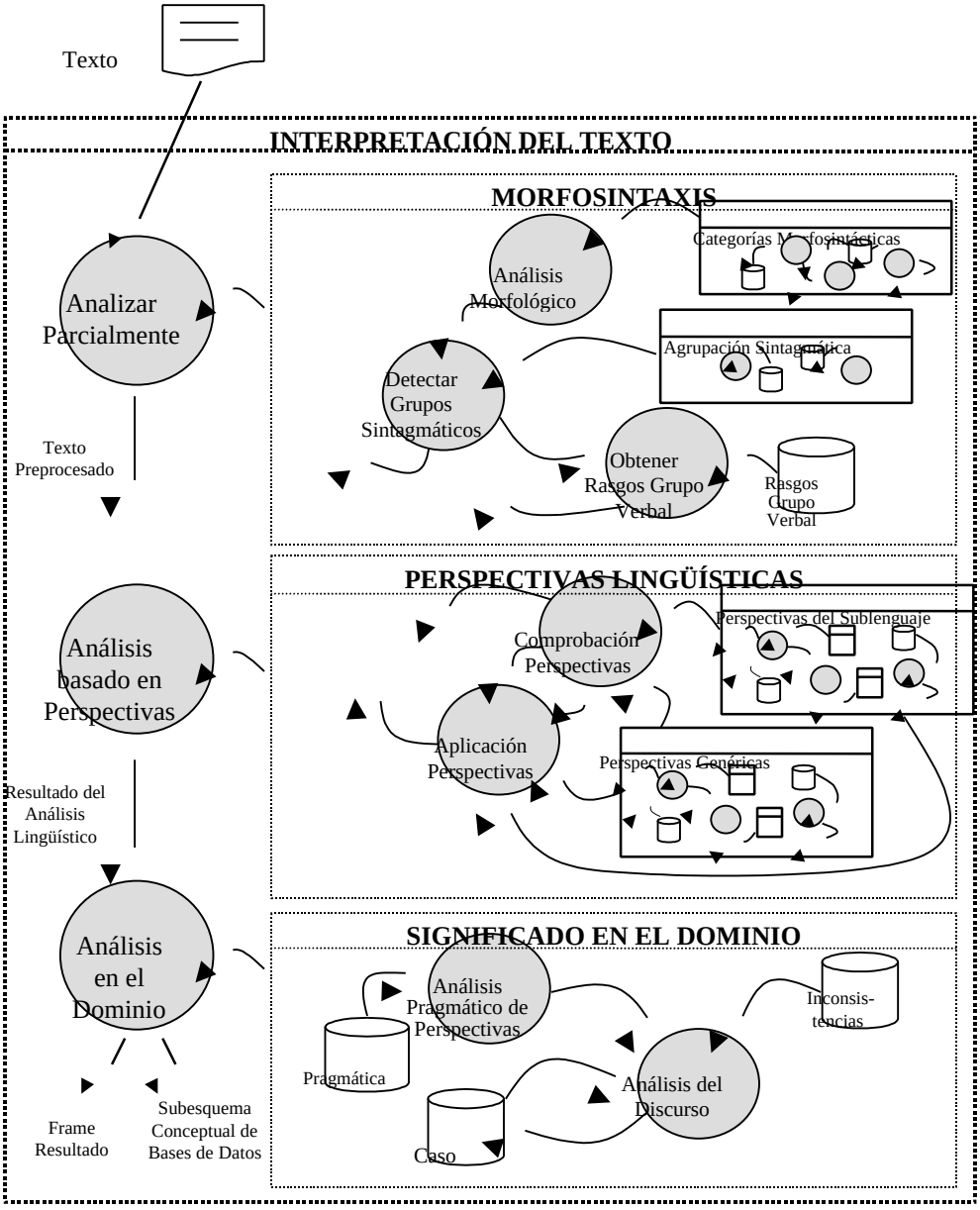


Figura 2: Diagrama de flujo en sus dos primeros niveles

La Figura 2 muestra cómo se estructura el conocimiento en un primer nivel:

- Conocimiento morfosintáctico para realizar un análisis parcial del texto
- Conocimiento sobre las perspectivas lingüísticas para llevar a cabo un análisis flexible
- Conocimiento del significado en el dominio para la interpretación del análisis lingüístico de acuerdo al fin al que se destina el sistema

El conocimiento para obtener el significado en el dominio es necesario en todo sistema PLN para traducir la información lingüística a una representación útil para el dominio de aplicación (pragmática) y para integrar la información adquirida con otra existente con anterioridad o extraída durante la interpretación (conocimiento del Caso).

En la primera parte de esta tesis se estudiaron dos tipos de trabajos en sistemas PLN para adquisición de conocimiento a partir de textos según el conocimiento del mundo de que disponen:

- Aquellos que tratan de construir una base de conocimiento con la información contenida en los textos, Delisle et al. (1996). Para ello parten de un conocimiento mínimo a priori sobre el dominio en que actúan ya que el objetivo es construir un modelo conceptual con el conocimiento contenido en el texto.

bases de datos se analiza una consulta, se realiza el análisis pragmático para obtener una consulta SQL, ésta se ejecuta y se procesa la siguiente. Sin embargo, si se procesan textos sobre noticias financieras, el resultado del análisis de una o varias oraciones puede integrarse en una estructura que represente el conocimiento que se va extrayendo del texto.

Se observa en la Figura 3 que el área de conocimiento raíz posee un único objetivo o tarea global denominada *Interpretar Texto*. Sin embargo, el conocimiento de la

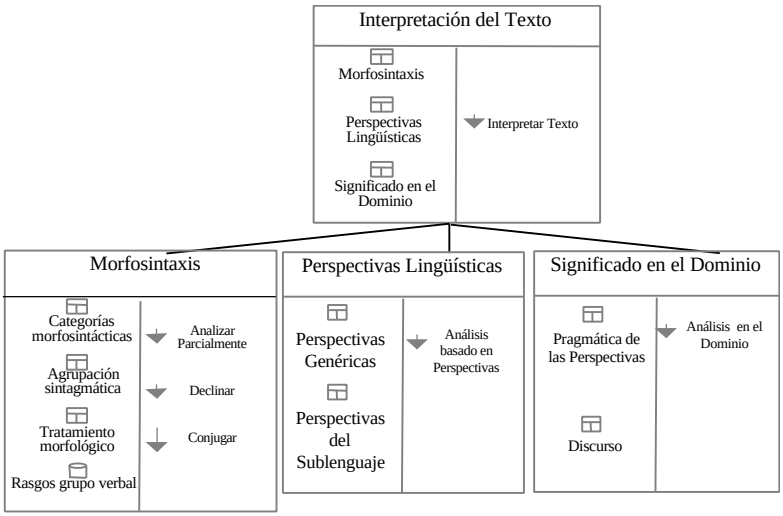


Figura 3: Modelo de Conocimiento en un primer nivel

- Aquellos que parten de un modelo conceptual del dominio que utilizan de ayuda en el análisis lingüístico de los textos, Hahn et al (1994), de tal forma que el conocimiento extraído completa el ya existente y permite realizar inferencias sobre él.

En este caso, el modelo propuesto se engloba en el primer tipo. El área de conocimiento del *Significado en el Dominio* debe ser definida completamente cuando se migra de un dominio a otro. Dependiendo del tipo de aplicación, el área de conocimiento del Caso puede ser de distinta naturaleza, incluso puede no existir. Por ejemplo en interfaces en LN de acceso a

Morfosintaxis proporciona tres tareas globales, a saber, *Analizar Parcialmente*, *Declinar* y *Conjugar*. Cada área de conocimiento se descompone en subáreas más especializadas, dando lugar a una jerarquía de varios niveles. Las tareas pertenecientes a cada área de conocimiento se descomponen en tareas de las áreas de los niveles inmediatamente inferiores.

En las siguientes secciones se estudiarán cada uno de los módulos de los que consta la arquitectura, examinando cómo las unidades inferiores contribuyen a que las superiores lleven a cabo su funcionalidad. Las unidades que ya no se descomponen en otras se denominan primarias y contienen el

conocimiento propiamente dicho, como la unidad del Léxico.

4. Construyendo los recursos lingüísticos

A continuación se describirán las unidades de conocimiento mostradas en la Figura 3 haciendo hincapié en las unidades primitivas que contienen el conocimiento propiamente dicho.

4.1 Conocimiento morfosintáctico

La Figura 4 muestra la estructuración completa del conocimiento morfosintáctico compuesto de:

Categorías Morfosintácticas: Conocimiento sobre la categorización de las palabras en clases denominadas *partes del discurso* (en inglés, part-of-speech) que suelen descomponerse en categorías cerradas (preposiciones, determinantes) y abiertas (nombres, adjetivos, etc.). También incluye conocimiento sobre la concordancia de las palabras y su relación con las etiquetas morfosintácticas de las palabras que los contienen. y un Modelo Oculto de Markov (HMM) entrenado que modela la dependencia del orden de las palabras en términos de etiquetas léxicas mediante las probabilidades de ocurrencia de pares de etiquetas, Sánchez-León y Nieto (1995).

Un ejemplo de la salida que produce el etiquetador es el siguiente (secuencia de pares de la forma *palabra_etiqueta*):

Los_ARTDMP departamentos_NCMP pueden_VMPI3P estar_VEINF en_PREP una_ARCAFS sola_ADJGFS facultad_NCFS o_CC ser_VSINF interfacultativos_ADJGMP ,_CM agrupando_VLGER en_PREP este_DMPXMS caso_NCMS cátedras_NCFP que_CQUE pertenecen_VLPI3P a_PREP facultades_NCFP distintas_ADJGFP ._FS

Agrupación Sintagmática: Conocimiento sobre las formas en que las categorías morfosintácticas se pueden agrupar para componer sintagmas básicos: grupos

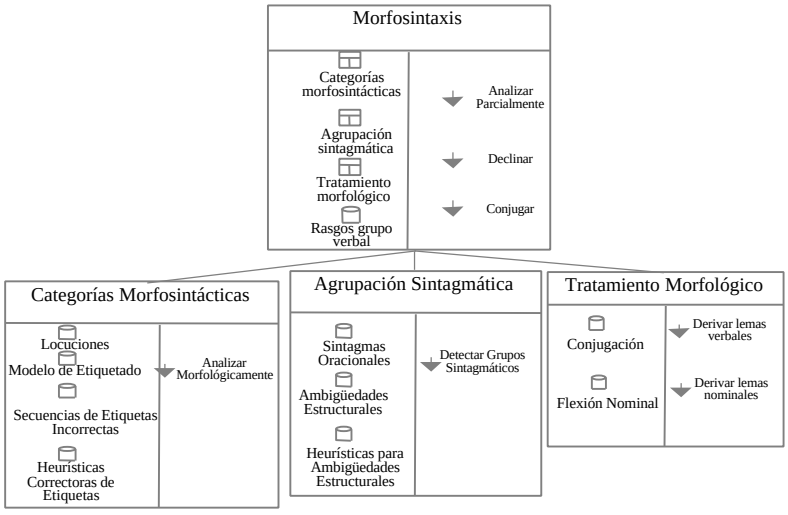


Figura 4: Estructuración del conocimiento Morfosintáctico

palabras (género, número, persona, caso). Sus componentes principales son un lexicón que contiene palabras con sus posibles partes del discurso (etiquetas morfosintácticas o léxicas) así como conocimiento sobre los sufijos de las

nominales, grupos preposicionales y grupos verbales, representados por un conjunto de autómatas en cascada, es decir, Redes de Transición Recursiva (RTN), Abney (1996b). La funcionalidad de este componente es *Segmentar Oración* que

detecta los sintagmas oracionales a partir de las *secuencias de pares palabra-etiqueta*, procedentes del análisis morfológico. Para la oración etiquetada en el ejemplo anterior se obtendría:

[f,[seg,[sn,[det,artdmp],[nom,ncmp],[]]],[seg,[gv,[modal,vmpi3s],[infinitivo,veinf],[]]],[seg,[sp,[prep,prep],[sn,[det,arcafs],[adj,adjgfs],[nom,ncfs],[]]]],[seg,[conjuncion,cc]],[seg,[gv,[infinitivo,vsinf]]],[seg,[sadj,[adj,adjgmp]]],[seg,[signo,cm]],[seg,[gv,[gerundio,vlger]]],[seg,[sp,[prep,prep],[sn,[demostrativonp,dmpxms],[nom,ncms],[]]]],[seg,[sn,[nom,ncfp],[]]],[seg,[sn,[relativo,cque],[]]],[seg,[gv,[conjugado,vlpi3p],[]]],[seg,[sp,[prep,prep],[sn,[nom,ncfp],[adj,adjgfp],[]]]]]

sintagmas=[sn,gv,sp,conjuncion,gv,sadj, signo,gv,sp,sn,sn,gv,sp]

Tratamiento Morfológico: Conocimiento sobre la forma en que las palabras pueden estructurarse mediante unidades más básicas, los morfemas.

Rasgos Grupo Verbal: Contiene un conjunto de reglas que se encargan de extraer determinados rasgos sintácticos de los grupos verbales y de la oración, por ejemplo, la voz (activa, pasiva, media), verbo principal, verbo auxiliar y verbo modal. Esta información se extrae de las etiquetas morfosintácticas que forman el grupo verbal (gv) procedente del análisis superficial llevado a cabo por los procesos lingüísticos *Analizar Morfológicamente* y *Detectar Grupos Sintagmáticos.* Por *ejemplo:*

'pueden estar'=[seg,[gv,[modal,vmpi3s], [infinitivo,veinf],[]]]

4.2 Perspectivas Lingüísticas

Una vez visto el conocimiento requerido para llevar a cabo un análisis parcial del texto, se describe el núcleo central del modelo propuesto: el conocimiento sobre las perspectivas lingüísticas, Martínez y García-Serrano (1997 y 1998).

Uno de los objetivos de la estructuración propuesta consiste en permitir flexibilidad en la combinación de distintos tipos de conocimiento lingüístico, de tal forma que el proceso de interpretación sea configurable según la información disponible en cada momento y el objetivo del sistema, ya que se constata que aunque el estilo es variable temáticamente, sigue unas pautas lingüísticas claras en cada campo técnico.

Las distintas combinaciones de utilización de las fuentes de conocimiento hacen que los procesos lingüísticos no tengan por qué aplicarse siempre en el mismo orden. Esta característica indica que esta propuesta puede enmarcarse entre dos tipos de control:

- *Control fijo secuencial*: Las fuentes de conocimiento se utilizan en una secuencia fija definida a priori, habitualmente morfología→ sintaxis→ semántica→ discurso→ pragmática. Una fuente de conocimiento se utiliza una sola vez y en un determinado paso del proceso.
- *Control variable con conocimiento distribuido*: Las fuentes de conocimiento se utilizan en un proceso con control dinámico. Una fuente de conocimiento se utiliza varias veces y en varios pasos del proceso. Esta arquitectura supone un primer intento de aproximación a la forma de modelar conocimiento lingüístico por un experto.

En una clase intermedia puede situarse una arquitectura que posibilita un control no fijo de los procesos lingüísticos en el que se manejan distintas posibilidades de combinación de las fuentes de conocimiento. Cada una de estas estrategias corresponde a lo que en este trabajo se ha denominado *perspectiva lingüística*.

Una perspectiva lingüística es un enfoque de análisis que se realiza sobre una oración o parte de ella utilizando determinadas

fuentes de conocimiento lingüístico. El uso de perspectivas lingüísticas permite definir estrategias que consideren la información más prometedora en cada momento del proceso de interpretación.

La definición de las perspectivas contempladas en la arquitectura propuesta se apoya en los siguientes parámetros:

1. *Conocimiento conceptual del dominio*, es decir, expectativas conceptuales sobre el texto o conocimiento sobre el dominio al que se refiere.
2. *Conocimiento tanto del lenguaje general como del lenguaje del dominio* o sublenguaje, es decir, conocimiento detallado sobre los fenómenos y restricciones lingüísticas específicas de cada dominio, así como sobre el lenguaje común.

Examinando las características propias del sublenguaje y relacionándolas con los distintos tipos de conocimiento habituales en un sistema de tratamiento automático de LN, se han identificado dos tipos de perspectivas lingüísticas: Genéricas y del Sublenguaje.

Esta distinción es necesaria para diferenciar entre aquellos fenómenos lingüísticos, tanto sintácticos como semánticos, propios del sublenguaje de los del uso general del lenguaje y que, consecuentemente, pueden darse en cualquier sublenguaje.

A continuación se expone de forma precisa cada una de las perspectivas diseñadas para este trabajo.

Patrones de Estilo

En lo sucesivo denominaremos P1 a esta perspectiva. Está formada por un conjunto de patrones que describen diferentes estructuras sintácticas típicas de los textos del dominio que responden a convenciones utilizadas para representar significados particulares. Generalmente 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 componen el patrón.

A continuación se muestran algunas oraciones extraídas del corpus con los patrones de estilo marcados:

"En una película pueden participar varios actores (nombre, nacionalidad, sexo)"
P1

"Un delincuente (DNI, nombre, teléfono) es
P1
arrestado por uno o varios policías"

Patrones Complejos

En lo sucesivo denominaremos P2 a esta perspectiva de la que forman parte una serie de patrones oracionales que manejan fenómenos de elipsis y conjunción, oraciones comparativas y condicionales, etc.

En principio sólo se contemplarán oraciones univerbales con sintagmas alargados por coordinación (yuxtaposición, conjunción y disyunción). Estos patrones complejos llevarán asociado un tipo de verbo que ayude a su posterior interpretación. Algunos ejemplos son:

"De cada aparato se quiere tener almacenado su código, descripción y estado de conservación"
P2

"Los puntos de ruta de un tour pueden ser ciudades, monumentos, zonas geográficas, etc."
P2

Palabras Clave

Esta perspectiva, que denominamos P3, 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 (terminología específica). En este caso se hace uso de las preferencias léxicas de las palabras según el

dominio en que se utilizan. Por ello, el análisis mediante esta perspectiva está dirigido por la semántica. No incluimos aquí los verbos, que se verán en la siguiente perspectiva. Algunos ejemplos son:

"Las asignaturas tienen un código identificador, un nombre y un curso" P3
P3

"Los documentos son de dos tipos: libros y artículos" P3

Verbos con preferencia semántica

Esta perspectiva del sublenguaje, que denominamos P4, la constituyen los *verbos* que desarrollan una preferencia semántica en el dominio que nos ocupa. Estos verbos son susceptibles de aparecer en cualquier texto descriptivo y, aunque pueden tener varios significados, sólo uno de ellos tiene preferencia en los textos que se tratan.

Por ejemplo, el verbo "disponer" puede denotar, entre otros, los siguientes significados:

1. DISPONER: Significado "mandar"
   verbo de acción que requiere un Agente(Agt) y un Objeto (Obj):
   esquema 1: Alguien (Agt) *dispone* Algo (Obj)
2. DISPONER: Significado "poseer"
   verbo de estado que requiere un Beneficiario (Ben) y un Objeto (Obj):
   esquema 1:Algo/Alguien (Ben) dispone de Algo/Alguien (Obj)
   esquema 2: Se dispone de Algo/Alguien (Obj) en Algo (Ben)
   esquema 3: Se dispone *de* Algo (Obj) *de* Algo (Ben)

De estos dos significados, el preferente en los textos descriptivos de bases de datos es el segundo (poseer).

Otros verbos sólo poseen un significado, y como tal, es también el preferente, como ocurre con el verbo "pertenecer":

1. PERTENECER: Significado "poseer"
   Verbo de estado que requiere un Beneficiario (Ben) y un Objeto (Obj)
   esquema 1: Algo/Alguien (Obj) *pertenece a* Algo/Alguien (Ben)

El manejo de las preferencias semánticas de los verbos permitirá iniciar un análisis basado en la semántica: cuando se conoce la preferencia de un verbo se conocen también los esquemas semánticos en los que aparece y, consecuentemente, los casos semánticos que hay que completar. Si no es posible completar los roles semánticos con los complementos sintácticos del verbo, entonces la preferencia no es válida y habría que probar con otro significado del verbo.

Verbos con preferencia semántica indeterminada

Denominaremos P5 a esta perspectiva que, al igual que P4, también se centra en el verbo principal de la oración, aunque en este caso los verbos no poseen preferencia semántica. Por ello, esta perspectiva no pertenece al sublenguaje.

Un verbo de este tipo es "dirigir":

1. DIRIGIR: significado "comunicar"
   Verbo de acción que requiere un agente(Agt) y un experimentador (Exp):
   Alguien (Agt) *se dirige* a Alguien (Exp)
2. DIRIGIR: significado "gobernar"
   Verbo de acción que requiere un Agente(Agt) y un Objeto (Obj):
   Alguien (Agt) *dirige* Algo (Obj)
3. DIRIGIR: significado "conjuntar un espectáculo"
   Verbo de acción que requiere un Agente (Agt) y un Objeto (Obj):
   Alguien (Agt) *dirige* Algo (Obj)
4. DIRIGIR: significado "movimiento"
   Verbo que requiere un Agente que sea a la vez Objeto (Agt=Obj) y un Locativo (Loc):
   Alguien/Algo (Agt=Obj) *se dirige* a Algo (Loc)

En este caso no es posible determinar un significado preferente, lo que supone tener que iniciar un análisis sintáctico que, a partir de todos los esquemas sintácticos correspondientes a un verbo, seleccione el que mejor se ajuste a la oración para iniciar después el análisis semántico.

Patrones de sintagmas

Esta perspectiva, que denominaremos P6, pertenece al grupo de las perspectivas

genéricas. Se utiliza para estudiar las interrelaciones que existen entre los componentes de un sintagma nominal o preposicional de igual importancia a las que existen entre un verbo y sus argumentos. Con este fin, P6 agrupa una serie de patrones sintácticos del lenguaje general, a nivel de sintagma, muy sencillos pero que en caso de no haber sido tratados en alguna de las otras perspectivas, pueden ser de utilidad.

Por ejemplo, en las dos oraciones siguientes existe un patrón sintáctico *nombre-adjetivo*:

"El personal académico imparte cursos y puede
P6
realizar trabajos de investigación"

"Los estudiantes graduados investigan"
P6

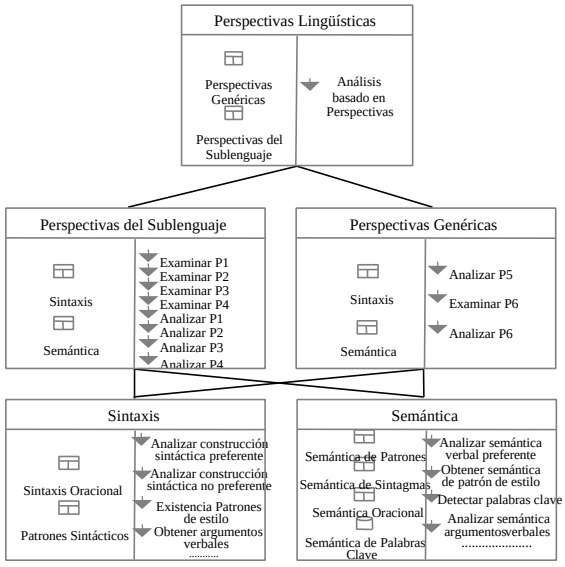


Figura 5: Estructuración del conocimiento de las perspectivas lingüísticas

No hay que olvidar que existen sintagmas nominales cuyo núcleo tiene una estructura argumental y sintagmas cuyo núcleo no subcategoriza a los argumentos que lo acompañan. Veamos con algunos ejemplos en qué se diferencian. El sintagma nominal "la selección de los empleados por el departamento de personal" tiene como núcleo la palabra *selección* que procede de un verbo (seleccionar) por lo que hereda de él su estructura argumental, Escandell (1995). Así, sabemos que el argumento *de los empleados* es en realidad el objeto del verbo seleccionar y que el argumento *por el departamento de personal* es el agente de la acción. El nombre *selección* pertenece a los nombres deverbales de proceso y resultado. Existen otros tipos de nombres con estructura argumental como las nominalizaciones de agente (por ejemplo, *constructor*, *traductor*, etc.) y las nominalizaciones deadjetivales (por ejemplo, *inteligencia*, *velocidad*, etc.)

En cuanto a los nombres que carecen de estructura argumental (por ejemplo, *casa*, *edificio*) no pueden imponer a sus modificadores una interpretación unívoca, es decir, los modificadores tienen un carácter opcional o adjunto.

La Figura 5 muestra cómo las distintas perspectivas hacen uso del conocimiento sintáctico y semántico como se expondrá en las siguientes secciones.

4.3 Léxico

El léxico contiene toda la información lingüística sobre las palabras, es decir, los rasgos sintácticos y semánticos. La aproximación seguida en este trabajo emplea un léxico con conocimiento sintáctico y semántico general. En el diseño del léxico se han utilizado ideas de WordNet para la semántica de nombres y adjetivos, así como de algunos trabajos realizados en EUROTRA (1991) para el diseño de las entradas léxicas verbales.

Nos centraremos en primer lugar en la descripción de las entradas léxicas verbales, puesto que los verbos son en la mayor parte de los casos el pivote de la interpretación. Representaremos los rasgos sintácticos y semánticos de un verbo en forma de marcos. Cada verbo constará de dos partes: Rasgos sintácticos generales

(*rasgos_sx_gen*) y una lista que representa uno o más significados (*list_sign*).

roles (Agt, Obj, Loc) de los que el primero es *persona* y el tercero *entidad*.

```
entrada_verbal: incluir
   rasgos_sx_gen:
        f_pas: si              (% capacidad para construcción pasiva %)
        f_exclpron: no         (%uso exclusivamente pronominal %)
        f_pron: si             (%capacidad para uso pronominal %)
        f_mod: no              (% capacidad para uso modal %)
        f_aux: no              (% capacidad para uso auxiliar %)
   list_sign:        {
        sign:                                   sign:
             frame_id: #1                            frame_id: #2
             f_coment: 'formar parte de algo'        f_coment: 'poner algo dentro de algo'
             f_aux_ser: no                           f_aux_ser: si
             f_aux_estar: si                         f_aux_estar: si
             f_esquema_sx: tr                        f_esquema_sx: bitr2-2
             f_prep_req1: nil                        f_prep_req1: en, dentro de
             f_prep_req2: nil                        f_prep_req2: nil
             f_tipo_evento: estado                   f_tipo_evento: acción
             f_esquema_sm: Loc1                      f_esquema_sm: Loc4
             f_sem_arg: nil, nil                     f_sem_arg: persona, nil, entidad
             f_pref: si                              f_pref: no                  }
```

Figura 6: Entrada léxica verbal de "incluir"

La Figura 6 muestra dos significados del verbo "incluir" (#1 y #2). El primero representa 'formar parte de' (significado preferente), su construcción sintáctica es transitiva (sujeto y complemento directo) y su esquema semántico es locativo con dos roles (Loc y Obj) sin restricciones de selección. El segundo representa 'poner algo dentro de algo', su construcción

Para el resto de las categorías léxicas (nombres, adjetivos, preposiciones, etc.) el conocimiento es más débil. Por ejemplo, una entrada léxica que representa un nombre incluye rasgos semánticos generales (nombre común, propio, animado, inanimado, humano, organización, abstracto, tiempo y lugar). Se ha construido una clasificación de rasgos semánticos

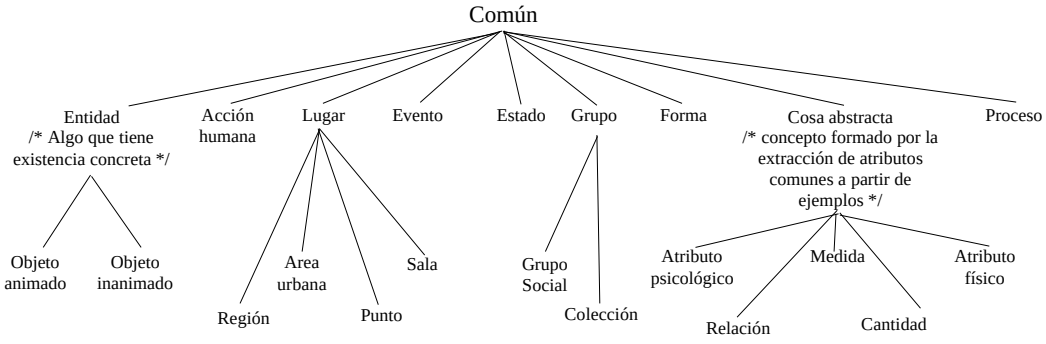


Figura 7: Una vista parcial de la jerarquía de rasgos semánticos nominales

sintáctica es bitransitiva sin suplemento (sujeto, complemento directo y complemento preposicional) y su esquema semántico es también locativo pero con tres

basada en la utilizada en Wordnet, Miller(1995), una vista de la cual aparece en la Figura 7.

4.4 Conocimiento Sintáctico

Existen dos tipos de conocimiento sintáctico; el componente de la *Sintaxis Oracional* (Figura 8) contiene el conocimiento relacionado con los distintos esquemas sintácticos de los verbos, así como el conocimiento sobre los complementos sintácticos; los *Patrones Sintácticos* poseen el conocimiento relativo a las secuencias sintácticas definidas como patrones de estilo y complejos así como diversas relaciones intrasintagmáticas.

En la descripción del Léxico se expuso que cada entrada verbal contiene varios marcos o esquemas sintáctico-semánticos, siendo uno de ellos el preferente en el dominio bajo estudio. Los esquemas sintáctico-semánticos contienen los argumentos

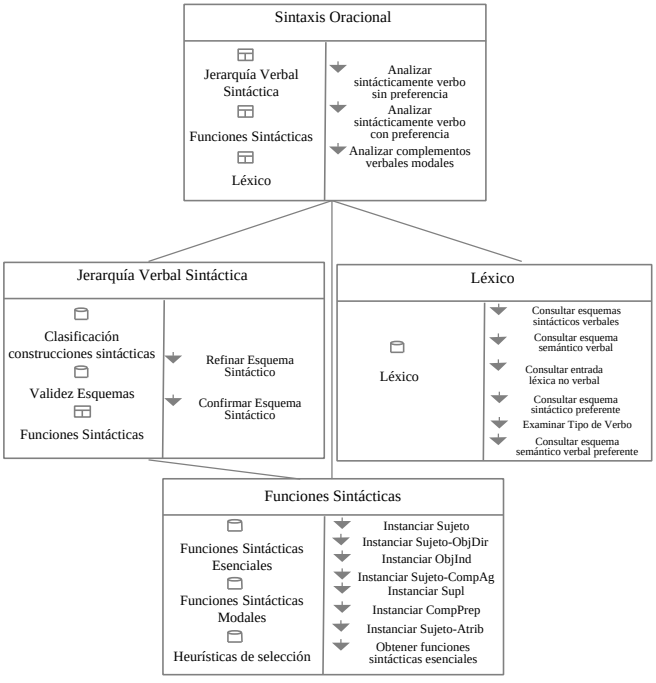


Figura 8: Estructuración de la Sintaxis

requeridos por el verbo y no los argumentos modales correspondientes con complementos circunstanciales (modo, tiempo, lugar, etc.). El conocimiento sobre la sintaxis oracional se ha obtenido de la gramática de la lengua española, RAE (1996), de textos de lingüística general, Moreno (1991) y otros trabajos más específicos sobre sintaxis. Adicionalmente, para el estudio de las funciones sintácticas verbales se han consultado los trabajos sobre complementos argumentales del verbo de Porto (1994).

La Figura 9 muestra una visión parcial de la *Clasificación de Construcciones Sintácticas*, junto con las condiciones que conforman su validez (etiquetas en las ramas y los argumentos que cada frame requiere indicados con el signo‘+’). En el primer nivel se distingue entre construcción activa (voz=activa) y construcción perifrástica (voz=pasiva_ser, pasiva_estar, media). Este nivel es necesario porque los esquemas de los verbos contenidos en el Léxico se especifican para la voz activa.

4.5 Conocimiento Semántico

Como muestra la Figura 5, la semántica se estructura en la *Semántica Oracional* contiene el conocimiento necesario para obtener el significado de una oración utilizando el verbo como pivote; la *Semántica de Palabras Clave* se encarga de detectar las palabras usuales en el dominio concreto; la *Semántica de Patrones* posee el conocimiento requerido para asociar significado a los patrones sintácticos y, por último, la *Semántica de los Sintagmas Simples* lleva a cabo el análisis semántico de los sintagmas simples. Nos centraremos en la semántica oracional (una descripción detallada de la semántica de los patrones se encuentra en Martínez (1998)).

El conocimiento de la *Semántica Oracional* se estructura en una *Jerarquía Verbal Semántica* que contiene una clasificación semántica de verbos basada en roles o casos semánticos, que permite construir la semántica oracional alrededor del verbo central mediante los argumentos esenciales y argumentos modales y el *Léxico*.

El análisis semántico utilizando el verbo como pivote se puede activar de dos formas:

a) Si el verbo **tiene una preferencia semántica determinada** en el dominio, se extrae el esquema semántico preferente de su entrada léxica del diccionario. Esto significa que directamente se tiene el significado del verbo y consecuentemente sus roles semánticos asociados dependiendo de su tipo. En este caso el análisis está conducido por expectativas y sigue la secuencia semántica-sintaxis.

b) Para los verbos **con preferencia indeterminada (o sin preferencia)**, mediante el análisis de los rasgos del grupo verbal, se intenta primero un análisis sintáctico (teniendo en cuenta que se conocen todos los esquemas sintácticos que pueden corresponder a ese verbo). Si se encuentra uno para el que se haya podido localizar sus funciones sintácticas, entonces, este esquema sintáctico tendrá uno o varios esquemas semánticos asociados. Buscando todos ellos en la jerarquía semántica verbal, se realiza la correspondencia entre roles semánticos y funciones sintácticas, comprobando, además, las restricciones semánticas que impone el verbo a sus argumentos (animado, humano, cosa, etc.). En este caso podría surgir más de un análisis, es decir, un verbo con varios significados, todos ellos con el mismo esquema sintáctico, mismo esquema semántico y mismas restricciones sobre los argumentos.

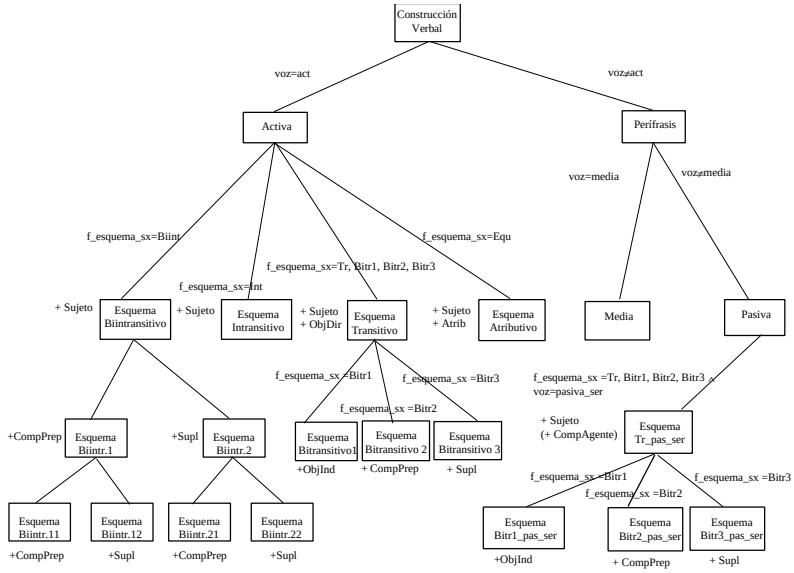


Figura 9: Clasificación y validez de construcciones sintácticas

La *Jerarquía Verbal Semántica* contiene una clasificación de esquemas semánticos (distintos esquemas semánticos asociados a los verbos) y el conocimiento sobre su validez. La clasificación semántica verbal se estructura en tres niveles (el primero para distinguir el tipo de evento, el segundo con los dominios semánticos y el tercero con las posibles combinaciones de roles sintácticos y semánticos). La validez de esquemas semánticos se realiza comprobando las condiciones necesarias que confirman un determinado esquema semántico verbal según los siguientes atributos:

- características semánticas del verbo (tipo de evento y dominio semántico)
- correspondencia de roles semánticos y sintácticos

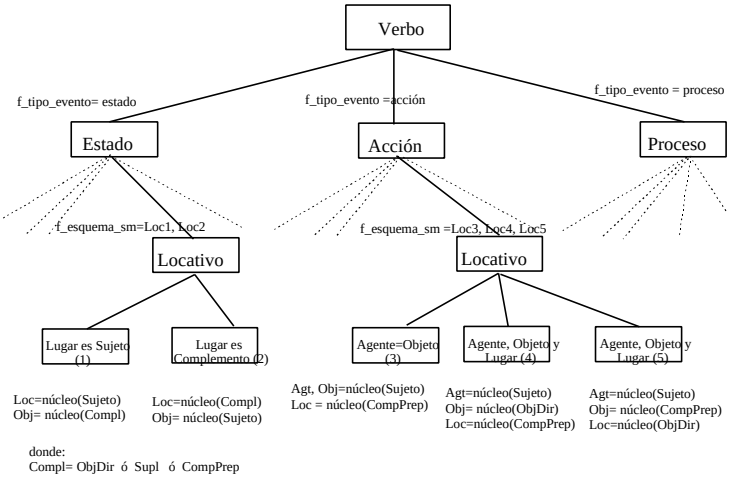


Figura 10: Una vista parcial de la jerarquía verbal semántica

- restricciones de selección del verbo respecto a los argumentos contenidas en la entrada léxica verbal.

La clasificación semántica verbal (Figura 10) se basa en el modelo *Case Grammar Matrix*, Cook (1989), que distingue entre casos esenciales o proposicionales (requeridos por la valencia semántica del verbo) y modales (generalmente adverbios que representan casos opcionales no requeridos por el verbo). Esta clasificación se combina con jerarquías similares de nombre, adjetivos y adverbios necesarias para obtener la semántica de los argumentos verbales.

Para cada dominio semántico (básico, experimental, temporal, locativo y benefactivo) según el tipo de evento (estado, acción y proceso) es posible distinguir distintas asociaciones de casos con funciones sintácticas, lo que da lugar a nuevos esquemas semánticos según se

Dominio Semántico	Correspondencia de roles sintácticos y semánticos	Verbos en el corpus
Básico	Obj es Complemento (bas1)	Existir, Haber
	Obj es Sujeto Prop es Atrib (bas2)	Ser
	Obj1 es Sujeto Obj2 es Complemento (bas3)	Costar, Medir, Valer, Pesar
Benefactivo	Ben es Sujeto Obj es Complemento (ben1)	Disponer, Poseer, Requerir, Necesitar, Tener, Tratar, Contar, Implicar, Obtener, Participar, Tratar
	Ben es Complemento Obj es Sujeto (ben2)	Componer, Caracterizar, Corresponder, Identificar, Pertenecer, Ser
Locativo	Loc es Sujeto Obj es Complemento (loc1)	Agrupar, Alojar, Constar, Encerrar, Involucrar, Incluir, Presentar
	Loc es complemento Obj es Sujeto (loc2)	Encontrar(se), Estar, Constituir, Formar, Formar_parte, Tener_lugar, Pasar, Ser, Haber
Experimental	Exp es Sujeto Obj es Complemento (exp1)	Conocer, Saber
	Exp=Obj es Sujeto, Prop es Complemento (exp2)	Encontrar(se), Estar
Temporal	Tm es Atributo Obj es Sujeto (Tem1)	Ser

Figura 11: Verbos de estado localizados en el corpus

realice esta correspondencia[2]. Se ha llevado a cabo un estudio de aproximadamente 100 verbos, con el fin de analizar los distintos ordenamientos sintácticos de las oraciones del corpus que los contienen, teniendo en cuenta también otros significados que puede tener un verbo y que no aparecen en el corpus. Para ello se ha seguido una descripción similar a la empleada en Wordnet mediante esquemas sintáctico-semánticos que muestran los roles semánticos superficiales (*Algo* <verbo> *Alguien*/*Algo*) que requiere un verbo, por ejemplo,

ACTUAR(Agt, Obj): Alguien/Algo actúa como Alguien/Algo /*interpretar un papel o función */

ACTUAR(Agt=Obj, Loc): Alguien/Algo actúa sobre/en Alguien/Algo /*producir un efecto sobre algo */

La Figura 11 muestra las agrupaciones de los verbos de *estado* según la clase semántica a la que pertenecen, considerando las realizaciones sintácticas de sus roles semánticos.

5. Conocimiento del significado en el dominio: modelado conceptual de bases de datos

Una vez realizado el análisis lingüístico de una oración según las distintas perspectivas, se procede a realizar el análisis pragmático de la oración así como su integración en el discurso, Martínez et al. (1998). El conocimiento del dominio (Figura 3) se descompone en *Pragmática de las Perspectivas*, formada por un conjunto de reglas que llevan a cabo la interpretación del resultado del análisis lingüístico de las perspectivas, y conocimiento del *Discurso*, para la integración de la pragmática de una perspectiva con el conocimiento extraído del texto hasta el momento mediante la resolución de inconsistencias

El resultado del análisis semántico de cada una de las perspectivas lingüísticas estudiadas se traduce a una estructura con los roles específicos del dominio. Para ello, se utilizará un conjunto de reglas que contendrán en su parte izquierda condiciones sobre atributos lingüísticos y en su parte derecha constructores propios de un modelo conceptual de datos orientado a objetos (tipos de objeto, interrelaciones, atributos, generalizaciones, agregaciones, etc.). Estos axiomas actúan sobre las instancias de las perspectivas lingüísticas sin tener en cuenta interrelaciones entre ellas. Será después, en el componente del *Discurso*, donde se estudiarán estas interrelaciones, lo que permitirá a su vez eliminar interpretaciones inconsistentes de las perspectivas.

En el caso de los verbos, el análisis pragmático de las perspectivas P4 y P5 relaciona los distintos tipos de evento y dominios de la clasificación semántica verbal (vista en la descripción del conocimiento semántico) con sus correspondientes conceptos del dominio, generalmente interrelaciones, generalizaciones y agregaciones. Los casos o roles del dominio se corresponden con los conceptos del dominio que queremos adquirir (Tipo de objeto, atributo, atributo clave, subtipo, supertipo, valor de atributo, instancia de tipo de objeto,...).

El análisis pragmático no es más que un refinamiento del análisis semántico. Por ejemplo, la Figura 12 muestra cómo los roles semánticos de los verbos de estado, según los dominios semánticos benefactivo y locativo, se corresponden con los roles del dominio o pragmáticos[3].

[2] Teniendo en cuenta la restricciones de selección semántica del verbo y la léxico-semántica de las palabras.

[3] Correspondencias similares se han definido para los verbos de acción y proceso, Martínez (1998).

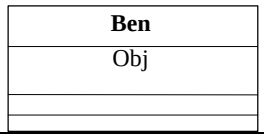
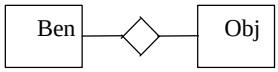
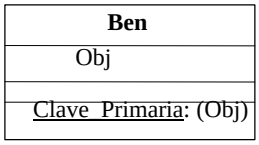
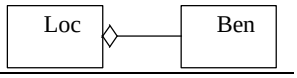
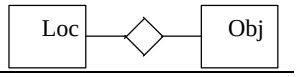
Verbos de Estado				
Clase Semántica del Verbo	**Roles Semánticos**	**Pragmática del Verbo**	**Roles Pragmáticos**	**Representación Gráfica (esquema conceptual de BD)**
Benefactivo		Descripción	Ben= TO Obj= Atributo	
	Ben, Obj	Asociación Genérica	Ben=TO Obj=TO	
		Identificación	Ben=TO Obj= Clave_Primaria	
Locativo		Agregación	Loc=Colección Obj= Miembro	
	Loc, Obj	Asociación Genérica	Loc= TO Obj=TO	
		Generalización	Obj= TO Supertipo Loc= TO Subtipo	

Figura 12: Correspondencia semántico-pragmática de los verbos de estado

Otros trabajos en el ámbito del modelado conceptual de BD son COLOR-X, Burg (1997), y NL-OOPS, Mich (1996). COLOR-X es el más avanzado en cuanto a que posee un formalismo de modelado conceptual apoyado en el lenguaje natural pero no profundiza en la forma de extraer el conocimiento necesario del texto para diseñar un esquema conceptual. NL-OOPS sí aborda el modelado conceptual de forma automática y emplea para ello el sistema PLN genérico LOLITA aunque en una forma secuencial fija. Sin embargo, se carece de marcos que aprovechen de forma (semi)automática y con control no fijo el conocimiento lingüístico contenido en los esquemas descriptivos que representan un UoD.

5. Un caso de estudio

En esta sección se expondrá un ejemplo de interpretación utilizando un esquema descriptivo extraído del corpus de textos sobre el que se ha realizado el estudio. El esquema descriptivo se muestra a continuación:

Oración 1: Durante un campeonato se desarrollan varias eliminatorias.

Oración 2: De cada eliminatoria interesa su tipo (octavos de final, cuartos de final, semifinal y final), su fecha de inicio y su fecha de finalización.

Oración 3: Un determinado número de selecciones participa en el campeonato.

Oración 4: De cada selección interesa su país y su continente.

Oración 5: Se asigna a cada selección un determinado grupo.

Oración 6: Cada selección ocupa una determinada posición (1, 2, 3 ó 4) en un grupo.

Oración 7: Cada grupo es identificado por un código (A, B, C, D, E o F).

Oración 8: Varias selecciones componen cada eliminatoria.

Oración 9: Cada encuentro está caracterizado por el sitio, fecha, hora y resultado.

Oración 10: Las selecciones participan en diversos encuentros.

Oración 11: Cada selección está compuesta de un determinado número de participantes.

Oración 12: De los participantes interesa el nombre, número de pasaporte y fecha de nacimiento.

Oración 13: Entre los participantes distinguimos: jugadores, entrenador y técnicos.

Oración 14: Cada jugador ocupa un determinado puesto (‘Defensa’, ‘Centrocampista’, etc.) en el equipo de la selección.

Oración 15: Cada entrenador tiene varios años de experiencia.

Oración 16: Cada técnico desempeña una función (por ejemplo, ‘Médico’, ‘Masajista’, etc.).

Oración 17: Los equipos de liga proporcionan jugadores a las selecciones.

Oración 18: De cada equipo de liga interesa el nombre y el país.

Oración 19: Los entrenadores preparan a los equipos de liga durante un intervalo de tiempo.

Se mostrará el análisis de la oración 14 por ser la que activa más perspectivas activa.: P1, P5 y P6. Una vez realizado el etiquetado morfosíntactico y la segmentación de la oración, se comprueba que la perspectiva P1 equipara el segmento ‘puesto (‘Defensa’, ‘Centrocampista’, etc.)’ con un patrón de estilo cuyo análisis semántico devuelve:

enumeración_valores: relación_orden→ (persona, persona)

Su correspondiente interpretación pragmática supone dos posibilidades[4]:

E1 = Tipo_Objeto : puesto ;
Tipo_Objeto : centrocampista ;
Tipo_Objeto : defensa ;
Generalización : es1 ,
Supertipo : puesto ,
Subtipos : defensa, centrocampista ;

E2 = Atributos : puesto :
Dominio : dom1 (defensa, centrocampista)

es decir, E1 es una generalización en la que *puesto* es el supertipo y *defensa* y *centrocampista* son los subtipos mientras que E2 contiene *puesto* como un atributo que puede tomar los valores *defensa* o *centrocampista*.

La perspectiva P5 se activa porque el verbo ‘ocupar’ es un verbo sin preferencia semántica que tiene dos significados[5]:

:-instanciar_sign_verbo([ocupar2,ocupar,no,'relacion de orden en algo', si,si,no,si,bitr3,[], [en],loc5, [persona, relacion_orden, [evento, lugar, grupo]]]).

:-instanciar_sign_verbo([ocupar1,ocupar,no,'tomar posesion o instalarse',si,si,no,si,tr,[],[],loc3, [persona, [evento, lugar, grupo]]]).

El único esquema sintáctico que encaja es el bitransitivo3 que requiere Sujeto, ObjDir y Supl. Este análisis tiene éxito con la siguiente instanciación de funciones sintácticas:

Sujeto= Cada jugador ,
ObjDir= un determinado puesto
Supl = en el equipo de la selección

Dentro de esta perspectiva P5 se encuentra el segmento ‘el equipo de la selección’ como un argumento del verbo. Si este segmento no hubiera sido utilizado durante el análisis del verbo y sus argumentos, entonces se habría activado la perspectiva P6, con el fin de extraer información de este segmento oracional. Además, las reglas utilizadas para el análisis sintáctico y semántico de los segmentos que realizan los argumentos verbales son las mismas que las utilizadas para el análisis de la perspectiva P6.

Comenzando por el segmento ‘el equipo de la selección’, las entradas léxicas de ‘equipo’ y ‘selección’ nos indican que la palabra ‘equipo’ puede tener los rasgos semánticos *instrumento* y *organización* y la palabra ‘selección’ los rasgos *acción_humana* y *organización* Esto significa que las posibles combinaciones

[4] En lo sucesivo Ei se corresponderá con los subesquemas de BD que se obteniendo en los sucesivos pasos de la interpretación; además, no se obtiene una única propuesta de esquema.

[5] Implementación PROLOG de la entrada léxica verbal mostrada en la sección 4.3.

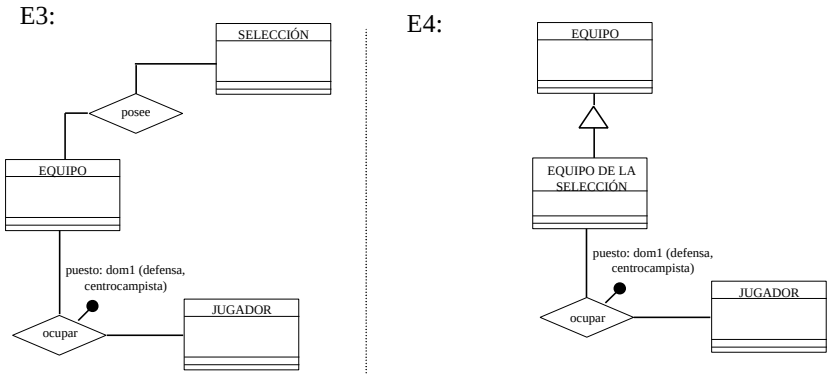


Figura 13: Dos subesquemas conceptuales obtenidos a partir de la Oración 14

que se ofrecen para el análisis semántico son:

instrumento + de + *acción_humana* ➔ USO
instrumento + de + *organización* ➔ POSESIÓN
organización + de + *acción_humana* ➔ USO
organización + de + *organización* ➔ POSESIÓN

Analizando el segmento aisladamente, todas ellas son válidas puesto que esta es la primera oración que se analiza. Sin embargo, al realizar el análisis semántico centrado en el verbo de la oración se obtiene:

Verbo: ocupar
Tipo de evento: acción
Dominio semántico: locativo
Roles esenciales:
Agt= (jugador=persona)
Obj= (puesto= relación_orden)
Loc= (equipo=organización,
((selección=organización ←posesión→ equipo= organización) OR (equipo=organización ←uso→ selección=acción_humana)))

Se observa que se ha eliminado la interpretación *instrumento* + de + *organización* ➔ POSESIÓN e *instrumento* + de + *acción_humana* ➔ USO para el segmento 'equipo de la selección' porque el rasgo *instrumento* no es compatible con las restricciones de selección del verbo 'ocupar' sobre el rol locativo (*lugar*, *grupo*, *evento*).

La interpretación pragmática para esta perspectiva P5 da lugar a:

E3 = E2 ∪ Tipo_objeto : equipo ;
Tipo_Interrelación : ocupar ;
Participantes : jugador, equipo ;
Atributos : puesto
Tipo_Interrelacion : posee ;
Participantes : selección, equipo

E4 = E2 ∪ Tipo_objeto : equipo de la selección ;
Tipo_objeto : equipo;
Generalización : es2 ,
Supertipo : equipo,
Subtipos : equipo de la selección ;
Tipo_Interrelación : ocupar ;
Participantes : jugador, equipo de la selección ;
Atributos : puesto

Se ha eliminado la interpretación E1 obtenida para la perspectiva P1 por la incompatibilidad:

Tipo_Objeto(puesto) incompatible con Atributo(puesto)

La Figura 13 muestra la representación gráfica de los dos subesquemas conceptuales de BD obtenidos para la oración 14.

Para el resto de las oraciones se procede del mismo modo y se van añadiendo tantos subesquemas de BD como posibles interpretaciones haya. Así, una de las posibles soluciones al texto propuesto es la mostrada en la Figura 14.

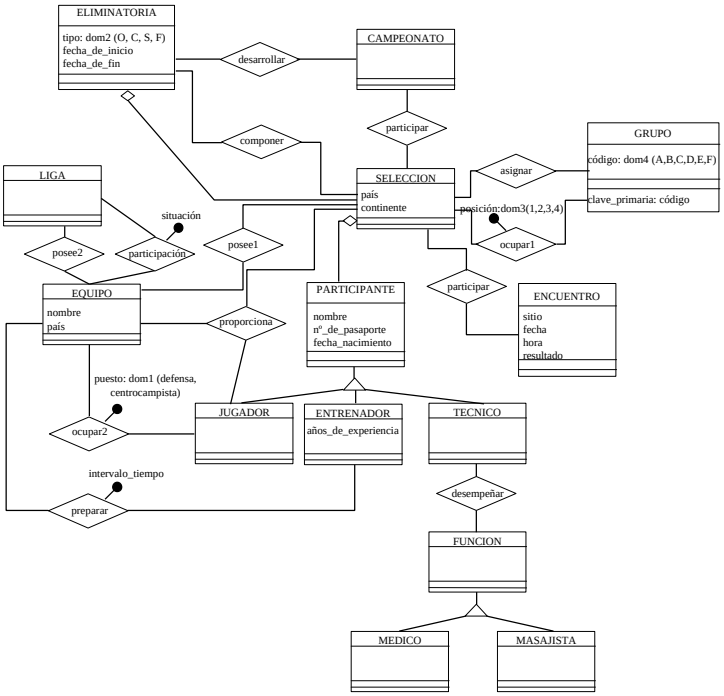


Figura 13: Una de las posibles interpretaciones del Caso

6. Conclusiones y Trabajos futuros

La línea de trabajo planteada ha sido desarrollar un sistema modular y multiforme que permita la incorporación y gestión automática de los distintos tipos de conocimiento lingüístico. Una idea clave en este trabajo es que la disciplina del tratamiento automático del lenguaje requiere un *enfoque de ingeniería* similar a los existentes en otras disciplinas tanto desde el punto de vista del proceso como de los productos obtenidos.

Se ha propuesto una estructuración adecuada del conocimiento que se requiere para el tratamiento automático del LN con un objetivo específico, de modo que:

- se apliquen criterios metodológicos (durante el análisis y diseño del modelo propuesto se han empleado principios procedentes de una metodología para sistemas basados en conocimiento y de la ingeniería del software) al desarrollo de sistemas que incorporan tecnología PLN.
- se fomente la utilización de técnicas de tratamiento del LN siguiendo la aproximación híbrida en busca de soluciones a problemas reales.
- se permita un control no fijo guiado por las fuentes de información lingüística para lo que se ha diseñado una estrategia de análisis basada en perspectivas lingüísticas que permite que la conceptualización propuesta sea configurable según el objetivo hacia el que se destina. Esta estrategia es

dependiente del dominio en una opción no habitual pues no es dependiente del contenido del texto pero sí de las expresiones que aparecen en él.

- se facilite tanto la reutilización del conocimiento como la validación por parte de los expertos tanto lingüistas como informáticos.

Además, se posibilita un tratamiento del lenguaje en un dominio complejo y real como el del diseño de BD. En este campo no hay resultados definitivos, y existen propuestas que extraen tanto esquemas conceptuales como relacionales de BD de forma (semi)automática que carecen de un análisis de la influencia del conocimiento lingüístico en las tareas de diseño de BD.

Desde el punto de vista lingüístico, el dominio no está claramente delimitado en cuanto al vocabulario aunque sí lo está en relación con el sublenguaje de los textos técnicos, por lo que se ha reutilizado un lexico de uso general y se han definido las características del sublenguaje.

En la actualidad se está investigando en un planteamiento distribuido y concurrente del modelo que surge de forma natural a partir de la estructuración propuesta. Así mismo, se está estudiando si la aproximación basada en perspectivas lingüísticas es adecuada para otro tipo de aplicaciones (por ejemplo, en extracción de información en Web) mediante nuevas configuraciones y refinamientos. Esto justificaría una propuesta de un marco genérico que permita construir distintas aplicaciones de la ingeniería lingüística.

Desde el punto de vista del dominio de la aplicación, puesto que la aplicación de modelado de BD se enmarca en un entorno para la enseñanza y desarrollo de BD, en una investigación futura se puede extraer del trabajo realizado una metodología completa para el desarrollo de BD con el prototipo implementado de base y que podría integrarse con un tutor inteligente para aprendizaje de BD.

Referencias

Abney (1996a), Abney, S. Statistical Methods and Linguistics. En Judith Klavans and Philip Resnik, Eds., *The Balancing Act*. MIT Press, Cambridge, MA, 1996.

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

Appelt et al. (1993), Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D. y Tyson M. FASTUS: A Finite-state Processor for Information Extraction from Real-word Text. *Proceedings IJCAI 93*, Chambéry, France, 1993.

Basili et al. (1996), Basili, R., Pazienza, M. T. y Velardi, P. An empirical symbolic approach to natural Language processing. *Artificial Intelligence*, 85, pp. 59-99,1996.

Burg (1997), Burg, J.F.M. Linguistic Instruments in Requirements Engineering. Tesis Doctoral, IOS Press, 1997.

Calzolari et al. (1994), Calzolari, N., Federici, S., Montemagni, S. y Peters, C. Extracting, Representing and Using Syntactic-Semantic Information from Cobuild Definitions. En J. Sinclair, M. Hoelter y C. Peters, editors, *The Languages of Definition: The Formalization of Dictionary Definitions for Natural Language Processing. Studies in Machine Translation and Natural Language Processing*, 7, pp. 59-148. Office for Official Publications of the European Communities, 1994.

Ciravegna et al. (1997), Civaregna, F., Lavelli, A., Petrelli, D. y Piaseni, F. Participatory Design for Linguistic Engineering: the Case of the GEPPETTO Development Environment. *Proceedings of the Workshop Computational Environments for Grammar Development and Linguistic Engineering*. Sponsored by the Association for Computational Linguistics, pp. 16-23, Madrid, July 1997.

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

Cuena y Molina (1996), Cuena, J. y Molina, M. KSM: An environment for Design of Structured Knowledge Models. En Spyros G. Tzafestas Ed, *Knowledge-Based Systems-Advanced Concepts, Techniques and Applications*, World Scientific Publishing Company, 1996.

Cunningham et al. (1997), Cunningham, H., Humphreys, K., Gaizauskas, R. y Wilks, Y.

Software Infrastructure for Natural Language Processing. *Proceedings of 5th Conference on Applied Natural Language Processing*, 1997.

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

Escandel (1995), Escandell, M. V. *Los Complementos del Nombre*. Cuadernos de Lengua Española. Editorial Arco Libros, 1995.

EUROTRA (1991), Argument Structure. En *Spanish Final Implementation Report of the EUROTRA Project*, Febrero 1991.

Hahn et al. (1994), Hahn U., Schacht S. y Broker N. Concurrent, Object-Oriented Natural Language Parsing: The ParseTalk Model. *International Journal of Human-Computer Studies*, 41, pp. 179-222, 1994.

Knight et al. (1995), Knight, K, Chander, I., Haines, M., Hatzivassiloglou, V., Hovy, E., Iida, M., Luk, S. K., Whitney, R. y Yamada, K. Filling Knowledge Gaps in a Broad-Coverage Machine Translation System, *Proceedings IJCAI 95*, pp. 1390-1396, 1995

Martínez (1998), Martínez, P. Propuesta de estructuración del conocimiento lingüístico para interpretación de textos: Aplicación al Diseño de BD. Tesis doctoral, Facultad de Informática, UPM, 1998.

Martínez et al. (1998), Martínez, P., De Miguel, A. y Marcos, E. A Knowledge-based Approach to Database Conceptual Modelling through Natural Language. *International Workshop on Issues and Applications of Database Technology*, (IADT'98). Berlin, Germany, July, 1998.

Martínez y García-Serrano (1997), Martínez, P. y García-Serrano, A. Una propuesta de estructuración del conocimiento para la adquisición de esquemas conceptuales de bases de datos a partir de textos. *Conferencia paralela al ACL-EACL'97* y *Procesamiento del Lenguaje Natural,Revista nº* 21, pp. 91-105, Julio 1997.

Martínez y García-Serrano (1998), Martínez, P. y García-Serrano, A. A Knowledge-based Methodology applied to Linguistic Engineering. En R. Nigel Horspool Ed., *Systems Implementation 2000: Languages, Methods and Tools*. London: Chapman & Hall, pp. 166-179, 1998.

Mich (1996), Mich, L. NL-OOPS: from natural language to object oriented requirements using the natural language processing system LOLITA. *Natural Language Engineering*, 2, 2, pp.161-187, 1996.

Miller (1995), Miller, G. A. WordNet: A lexical Database for English. *Communications of the ACM*, 38, 11, pp. 39-41, Noviembre 1995.

Moreno (1991), Moreno Cabrera, J.C. *Curso universitario de lingüística general, vol. 1: Teoría de la gramática y sintaxis general*. Madrid:Síntesis, D. L. 1991.

Porto (1994), Porto Dapena, J. *Complementos argumentales del verbo: directo, indirecto, suplemento y agente*. Cuadernos de Lengua Española. Arco/Libros, S.L. 1994.

RAE (1996), Real Academia de la Lengua Española, *Comisión de Gramática. Esbozo de una nueva gramática de la lengua española*. 1ª edición, 16ª reimpresión, Madrid: Espasa-Calpe, 1996.

Rich y Knight (1991) Rich, E. y Knight, K. *Artificial Intelligence*. New York: McGarw-Hill, 1991.

Sabah (1993), Sabah, G. Knowledge Representation and Natural Language Processing. *AICOM*, 6, 3-4, pp. 155-186, Septiembre-Diciembre 1993.

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

Simpkins (1994), Simkins, N. An Open Architecture for Language Engineering. *First Language Engineering Convention*, Paris, 1994.

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

Thomé (1993), Thomé, B. Definition and Scope of System Engineering. En B. Thomé Ed., *Systems Engineering. Principles and Practice of Computer-Based Systems Engineering*. John Wiley & Sons Ltd, 1993.

Vossen et al. (1997), Vossen, P., Diez-Orzas, P. y Peters, W. Multilingual design of EuroWordNet. *Proceedings of the Automatic Information Extraction and Building of Lexical Semantic Resources Workshop*, pp. 1-8, Madrid, Julio 1997.

Weischedel et al. (1992), Weischedel, R., Ayuso, D., Boisen, S., Fox, H. and Ingria R. A new approach to text understanding. *Proceedings of the 5th DARPA Workshop on Speech and Natural Language*, 1992.