Dos aproximaciones a la adquisición del conocimiento léxico y conceptual para sistemas de procesamiento del lenguaje natural

Alicia Ageno
M. Antonia Martí
Horacio Rodríguez
Felisa Verdejo
Irene Castellón
German Rigau
Mariona Taulé

Report LSI–91–45

Dos aproximaciones a la adquisición del conocimiento léxico y conceptual para sistemas de procesamiento del lenguaje natural

M. Felisa Verdejo
Departamento de Ingeniería Electrica, Electrónica y de Control
Escuela Técnica Superior de Ingenieros Industriales
UNED

A.Ageno, I. Castellón, G.Rigau, H.Rodriguez
Sección de Inteligencia Artificial
Departamento de Lenguajes y Sistemas Informáticos
Universidad Politécnica de Cataluña

M.A.Martí, M.Taulé
Departamento de Filología Románica
Universidad de Barcelona

1. Introducción.

La adquisición de conocimiento juega un papel primordial en cualquier sistema de comprensión del lenguaje natural. Dos de los aspectos relevantes conciernen a los niveles léxico y conceptual. El nivel conceptual constituye el dominio semántico de la aplicación, y por tanto los objetos que se incluyen en dicho nivel definen la cobertura conceptual del sistema. Por otra parte en el nivel léxico es donde se expresa información de diferente tipo, como por ejemplo, categorización morfológica y semántica, valencias sintácticas, etc. lo que determina en gran parte la cobertura lingüística del sistema.

Dos son las aproximaciones básicas para la adquisición de dicho conocimiento: la primera se basa en la reutilización de fuentes existentes, fundamentalmente diccionarios en soporte magnético, mientras que la segunda propone un tratamiento de elicitación, con expertos humanos, en base a un diálogo cooperativo.

Durante los últimos años estamos desarrollando dos proyectos, ACQUILEX (1) y GUAI (2) que se enmarcan en la primera y segunda linea respectivamente. El objetivo de la presente comunicación es ofrecer una visión detallada de ambos planteamientos, mostrar los entornos de adquisición que hemos implementado, así como discutir la complementariedad de las aproximaciones. La organización del artículo es la siguiente: en el apartado 2 presentamos el proyecto ACQUILEX, en el 3 el módulo de adquisición de GUAI y finalmente en el 4 establecemos las conclusiones más relevantes.

2. El Proyecto Acquilex.

El objetivo básico del proyecto **ACQUILEX** (1) es el desarrollo de técnicas y métodos que permitan la utilización de diccionarios en soporte magnético (M.R.D, Machine Readable Dictionaries) para la construcción de componentes léxicos para sistemas de procesamiento del lenguaje natural (P.L.N).

Los diccionarios automatizados constituyen una fuente de adquisición de Conocimiento Léxico y conceptual que, potencialmente, permite abordar algunos aspectos especialmente costosos de la construcción de una base de conocimiento para un sistema de P.L.N. de forma rápida y competitiva. Se trata de un campo relativamente poco explorado del área de la Adquisición del

Conocimiento, debido a la dificultad que supone el tratamiento complejo de grandes volúmenes de información y a las limitaciones de las teorías lingüísticas que abordan el tema del léxico.

No es nuestra intención presentar aquí una descripción, ni aún resumida, del proyecto (la referencia [Acquilex 89] cubre tal propósito) sino simplemente dar las líneas generales del mismo para enmarcar nuestra comunicación. A largo plazo el objetivo del proyecto es la construcción de una Base de Conocimiento léxico multilingüe con las siguientes caraterísticas:

- Contendrá Información Léxica general e independiente del dominio.
- La Representación del Conocimiento favorecerá al máximo su reutilización.
- Se utilizarán exclusivamente fuentes léxicas ya existentes.
- Los procesos de extracción de la información léxica y de utilización de la misma por los diferentes sistemas de tratamiento del Lenguaje Natural serán distintos e independientes.
- Se utilizará un formato estándar de intercambio de fuentes léxicas.
- Se definirá una estructura conceptual común, ligada a los significados individuales de las palabras en las diferentes lenguas cubiertas y capaz de soportar un procesamiento del lenguaje basado en el Conocimiento.
- Se incluirá un vocabulario general con información fonológica, morfológica, sintáctica y semántico-pragmática para las diversas lenguas que forman parte del proyecto.

Los objetivos primeros del proyecto se centran en el desarrollo de un prototipo de Base de Datos Léxica (L.D.B) y de Base de Conocimientos Léxica (L.K.B.) multilingües para un subconjunto manejable, pero significativo, del vocabulario y en el desarrollo de técnicas para la extracción semiautomática de información léxica del diccionario.

Las principales dificultades que plantea el acceso a la información léxica contenida en los M.R.D. son, por una parte, la aparición de la misma información en un formato poco estructurado, lo cual dificulta su utilización, y por otra parte, la ineficiencia que la propia organización de la fuente (normalmente las cintas de fotocomposición) supone para el acceso. Las figuras 1 y 2 nos muestran un ejemplo, tomado del Vox [Vox87], de la versión impresa de una definición y del formato en que la misma aparece dentro del M.R.D.

> **I) cacho** (*l. calculu , piedrecita*) m. *fam.* Pedazo pequeño de alguna cosa. *2* m. Cierto juego de naipes. *3* m. *Méj. y P.Rico.* Participación pequeña en un número de la lotería. SIN.1 v.Pedazo.

Fig. 1: Entrada editada del Diccionario Vox.

> [EP[j2]I) cacho [k1](l. [k2]calculu, [k1]piedrecita) [k2]m.
> [k1]fam. Pedazo pequeño de alguna cosa.[k2] 2 [k1]Cierto juego de naipes.[k2] 3 Méj. [k1]y[k2] P. Rico. [k1]Participación pequeña en un número de la lotería.[EP[j3] [j6]Sin.[j7] [k2]1 [k3][k1]v[k3].
> Pedazo.

Fig. 2: Entrada del Diccionario Vox en el M.R.D.

La principal motivación del proyecto Acquilex es la de integrar en un proyecto común buena parte de la investigación que se desarrolla en Europa sobre el tema de diccionarios automatizados. Ello implica una integración no sólo de equipos de investigación sino de herramientas y bases léxicas disponibles.

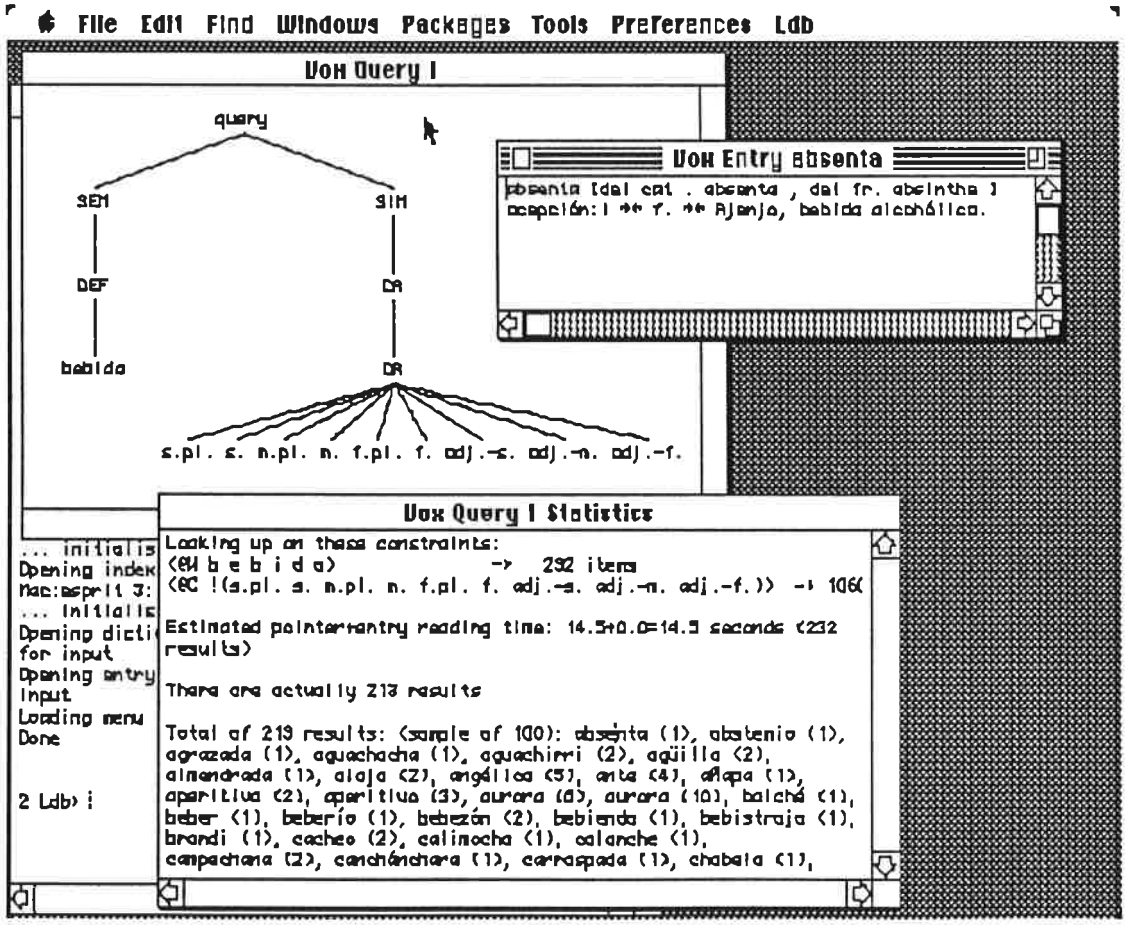
La primera característica, pues, de nuestra propuesta es la de trabajar en un contexto multilingüe. De hecho, el material con el que trabajamos incluye actualmente el LDOCE (Longman Dictionary of Contemporary English) inglés, el Garzanti italiano, el Van Dale holandés, el VOX español y los bilingües COLLINS inglés/italiano y Van Dale inglés/holandés. El empleo de diccionarios bilingües nos ha permitido explorar su utilización como vehículo potencial de transferencia de información léxica.

Las etapas principales en el desarrollo del proyecto, buena parte de las cuales ya han sido cubiertas, son las siguientes:

1. Elaboración de un modelo computacional de diccionario de forma que tengan expresión en él todas las características diferenciales de los diferentes diccionarios individuales [Calzolari90].

2. Descripción de los diccionarios individuales en términos del modelo.

3. Definición y desarrollo de software para la gestión de dicha L.D.B. [Carroll90]. La pantalla 1 nos muestra el aspecto de una sesión de trabajo en el entorno de la L.D.B. En ella se observa una "query" realizada de forma gráfica, en la cual se demandan aquellas entradas que contengan la palabra "bebida" en la definición, y cuya categoría sintáctica corresponda a un sustantivo. Además aparece otra ventana con los resultados de esta "query" (número total de entradas con estas características y una muestra de 100, tal como ha pedido el usuario), y una última ventana conteniendo la primera de éstas.



Pantalla 1.

4. Carga de la información de los diccionarios individuales en la L.D.B. [Castellón et al.90,91]. La figura 3 nos muestra el resultado del proceso de extracción de información

léxica a partir de la entrada que las figuras 1 y 2 nos mostraban. La estructura (que se suele conocer como "estructura lispificada") contiene la misma información presente en el diccionario fuente clasificada y segmentada, y constituye la entrada al proceso de carga de la L.D.B. La transformación de las entradas en formato M.R.D. a una estructura lispificada requiere un tratamiento específico dependiente del diccionario que se quiere cargar. En nuestro caso, desarrollamos un programa que transformaba, mediante una gramática, cada entrada del M.R.D. a una estructura lispificada.

5. Derivación de una estructura conceptual común. Relaciones entre esta estructura y las definiciones individuales de cada diccionario [Calzolari 90].

6. Desarrollo de técnicas para la extracción de información conceptual a partir de la información léxica contenida en la L.D.B.

7. Carga en la L.K.B. de un subconjunto significativo de la información léxica de los diversos diccionarios individuales.

8. Chequeo y evaluación del sistema a través de la actuación de un Sistema de P.L.N. cuyo componente léxico se haya extraído de la L.K.B. El Sistema chequeará las dos funciones de comprensión y generación [Cater90].

En el presente artículo vamos a centrarnos en el punto 6, la problemática que representa y la metodología que hemos seguido para su resolución.

```
((cacho )
(NH I)
(ETIM l. calculu , piedrecita )
(acepción 1)
(CA m.)
(REG fam.)
(DEF Pedazo pequeño de alguna cosa.)
(acepción 2)
(CA m.)
(DEF Cierto juego de naipes.)
(acepción 3)
(CA m.)
(GEO Méj. y P.Rico.)
(DEF Participación pequeña en un número de la lotería.)
(RELA 1)
(TIPOR Sin.)
(TXR 1  v.Pedazo.)
)
```

Fig. 3: Entrada del Diccionario Vox lispificada.

2.1 Esquema General del Sistema.

El contenido de la LDB es exclusivamente léxico. Para extraer información semántica debemos definir en primer lugar los elementos atómicos de la representación semántica, para, a continuación, establecer qué tipo de propiedades los definen y como se relacionan entre sí.

En nuestra aproximación, las unidades semánticas corresponden a las diferentes acepciones de las entradas del diccionario (se han dejado para más adelante los problemas que plantea la fusión de estructuras semánticas). La principal relación con la que trabajamos es la ES-UN que liga un concepto con su genérico. Esta relación es la base de la estructura taxonómica que extraemos y actúa como soporte del mecanismo de herencia de propiedades. Otras relaciones taxonómicas que extraemos igualmente son las que ligan un concepto con sus partes, un conjunto con sus miembros, etc... Existen, por otra parte, determinadas propiedades que pueden ser extraídas de la

LDB como color, forma, tamaño, etc... que son incorporadas a los nodos de la estructura conceptual.

Nuestro primer objetivo, por lo tanto, consistió en realizar un Sistema semiautomático de extracción de información semántica, básicamente taxonómica, del diccionario Vox cargado en la LDB. Para realizar nuestro diseño tomamos en consideración los siguientes criterios generales:

• La extracción de información semántica a partir de las entradas del diccionario supone un problema que no puede ser resuelto de forma completamente automática. La decisiones tomadas por el sistema deben ser validadas y confirmadas por un experto humano. Esto implica el uso de un entorno interactivo.

• Otra consideración importante es la reusabilidad de las estructuras de datos resultantes en otros entornos, especialmente el proceso de conversión a la LKB [Ageno et al. 91c,d].

• Es importante en cualquier proyecto de las características de **ACQUILEX** la reusabilidad del software producido. En este sentido, hemos utilizado al máximo tanto la metodología como las herramientas de otros participantes del proyecto adaptándolas a nuestras necesidades e integrándolas en nuestro propio software. Concretamente hemos usado elementos del software LDB de Cambridge [Caroll 90] incluyendo el analizador sintáctico-semántico FPar [Alshawi 89,90] y el analizador morfológico Seg-Word [SanFilippo 90a,90b] así como también hemos utilizado la aproximación para la extracción de información taxonómica a partir de las definiciones del diccionario Tax-Build [Copestake 90a] [Copestake 90b].

• Las tareas a realizar y el conocimiento asociado a ellas, supone la necesidad de un sistema flexible donde inicialmente será requerida una gran intervención humana que permita, de forma incremental, una mayor autonomía del sistema.

• Algunas de las tareas involucradas en la extracción de información semántica, como el análisis de las definiciones del diccionario, consumen un gran volumen de tiempo y no permiten, por tanto, su integración dentro de los procesos interactivos. Así, el sistema debe permitir la cooperación entre los procesos batch e interactivos.

2.2 Fuentes de conocimiento.

• La principal fuente de conocimiento de nuestro entorno es, por supuesto, la LDB del Vox. La LDB es una fuente de conocimiento estática que contiene la siguiente información actualmente accesible y susceptible de ser usada en el proceso de extracción:

- la entrada.
- la etimología.
- la categoria morfo-sintáctica.
- la definición o acepción.
- los usos particulares: figurado, no usual, informal, etc.
- el tema: biología, medicina, etc.
- la información geográfica: América, Aragón, etc.
- las relaciones semánticas: sinonimia, antonimia, etc.

Otras fuentes de conocimiento consultadas durante el proceso de extracción son:

• El conjunto de gramáticas para el análisis de las definiciones.
• El conjunto de reglas morfológicas.
• Un LEXICON para aquellas palabras que no aparecen en el diccionario o son consultadas con gran frecuencia por el analizador morfológico.
• El conjunto de heurísticos que son aplicados en distintos puntos para ayudar al usuario en su toma de decisiones.

Todas estas fuentes de información son dinámicas. Al principio, el sistema no cubre todos los

casos posibles. A medida que vamos construyendo taxonomías debemos incorporar, tanto al conjunto de reglas morfológicas como al LEXICON, aquellos nuevos casos que vayan presentándose y queramos tomar en consideración. Asimismo, debemos desarrollar distintas gramáticas sintactico-semánticas, según el ámbito temático al que pertenezca la taxonomía que vayamos a construir [Agenoetal.91b]. Estas gramáticas, al principio, tampoco captarán toda la información que deseamos extraer. Debemos mejorarlas paulatinamente hasta conseguir los resultados esperados.

2.3 El proceso.

Nuestro sistema realiza cuatro tareas básicas, tal como se muestra en la figura 5. La primera consiste sólamente en la extracción de la estructura taxonómica que subyace en las acepciones del Vox comenzando por una entrada inicial. Estas entradas iniciales, pueden ser localizadas fácilmente por su alta frecuencia de aparición como genéricos en las acepciones [Copestake90b]. La segunda tarea permite la extracción de las propiedades semánticas, no taxonómicas, que aparecen en las acepciones de la taxonomía creada anteriormente. En la tercera tarea, se validan los heurísticos aplicados en la construcción de la taxonomía. Finalmente, toda la información adquirida anteriormente deberá ser integrada en la LKB.

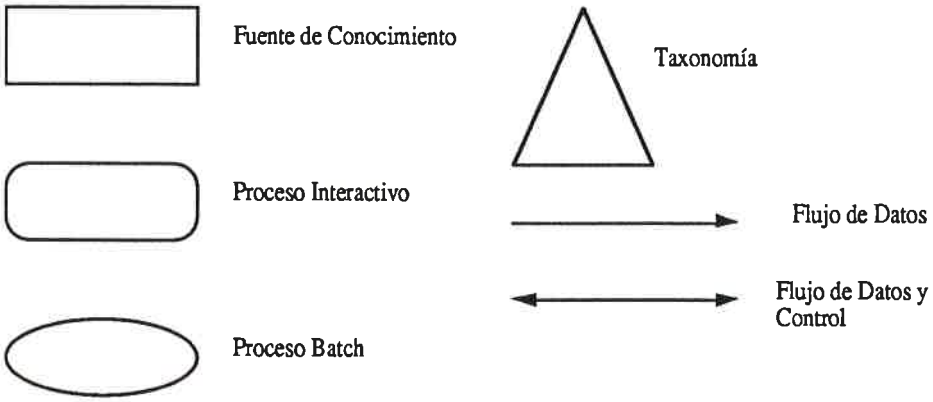


Fig. 4: Descripción de los símbolos.

Una parte del proceso de extracción incluye un conjunto de heurísticos que permiten automatizar el proceso de desambiguación de sentidos. Importantes características de nuestra aproximación, son por tanto, la definición de un conjunto de heurísticos parametrizados, su uso como guía en los procesos de selección y la existencia de mecanismos de evaluación de su actuación así como modificación de los mismos de acuerdo con dicha evaluación.

El sistema tiene dos modos de funcionamiento: el modo adquisición y el modo validación. En el modo adquisición, el sistema facilita la construcción interactiva de las taxonomías y el posterior análisis semántico de las definiciones (tareas 1 y 2). El análisis semántico de las taxonomías no necesita de la intervención humana y puede realizarse en un proceso batch. En modo validación, el análisis realizado anteriormente es validado y corregido mediante un proceso interactivo. El usuario entonces puede optar entre modificar la gramática con la que se ha hecho el análisis de la taxonomía y volver a lanzar el proceso batch o simplemente corregir el último análisis realizado. Este proceso puede realizarse tantas veces como el usuario considere necesario. Así, la gramática de cada taxonomía se modifica de forma incremental hasta conseguir un resultado óptimo. A continuación, en modo validación, podemos confrontar una estructura taxonómica existente (y que consideremos correcta) con diferentes conjuntos de heurísticos (tarea 3). Esta operación no necesita tampoco de la intervención del usuario.

Para facilitar la comprensión al lector, en la figura 5, el proceso de extracción de las relaciones semánticas no taxonómicas ha sido colocado antes que el proceso de validación de los heurísticos, pero ambos son independientes entre sí, pudiendo llevarse a cabo en cualquier orden y realizarse con taxonomías no completas.

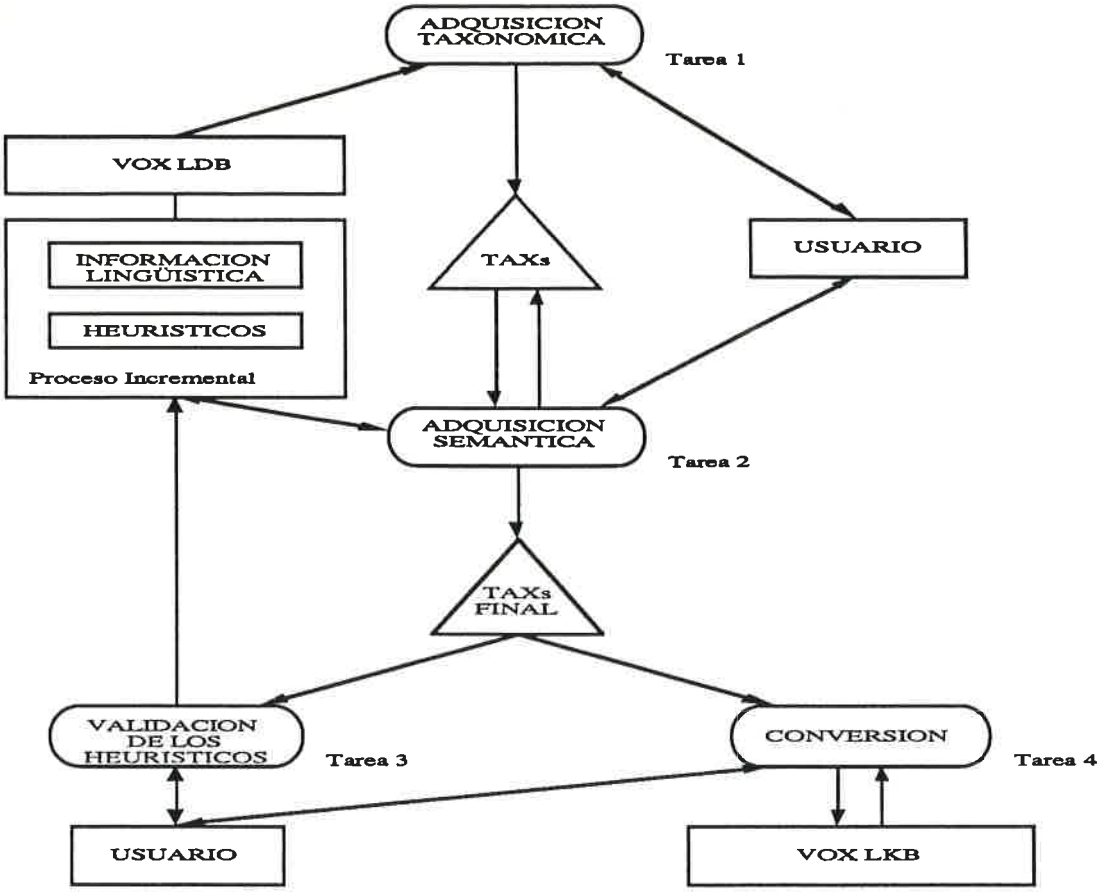


Fig. 5: Esquema General del Sistema.

Nuestro entorno es completamente compatible con el sistema LDB, permitiendo que durante la construcción de la taxonomía puedan formularse a la misma tantas y tan complejas preguntas como se considere necesario.

2.3.1 Adquisición de la información taxonómica.

El módulo de adquisición es el encargado de la construción de la estructura taxonómica (relación ES-UN) de un subconjunto de acepciones [Amsler 81]

El proceso de extracción de las relaciones taxonómicas, debe resolver dos problemas principales: la extracción y la desambiguación del término genérico [Copestake 90a].

En nuestro caso, el problema de la extracción del término genérico se resuelve mediante el análisis sintáctico parcial de las definiciones. Usamos para ello el analizador FPar [Alshawi89]. Dada una acepción, podemos detectar cuál es o cuáles son sus palabras hiperónimas [Ageno et al. 91a] junto con otras propiedadesemánticas.

A la hora de elegir un analizador para las definiciones, se ha considerado la necesidad de producir una interpretación **semántica** razonable de las mismas, aunque el análisis completo de éstas no sea posible.

El analizador sintáctico FPar utiliza para su proceso de análisis el texto de las definiciones aumentado con su categorización morfológica.

La alternativa escogida para solucionar el problema del análisis morfológico ha tratado de ser consecuente con la metodología que intentamos seguir, usando las mismas entradas del

diccionario como fuente básica de información para la categorización y tratando de reusar herramientas ya existentes. En este sentido hemos utilizado el analizador Seg-Word optimizado [SanFilippo 90 a,b] [Ageno et al.91a].

Este analizador ha sido diseñado específicamente para trabajar en conjunción con la LDB. Utiliza una típica aproximación basada en "string-unification", con listas tanto de afijos como de raíces, con la particularidad de que ésta última no se ha de construir previamente, pues a partir de la subcadena de entrada que se toma como raíz potencial se construye una lista de posibles entradas de diccionario. Esta lista se usa para acceder al diccionario con una simple y rápida consulta a la LDB. De esta consulta se extrae ya la información necesaria, es decir, se reconoce la subcadena de entrada como raíz correcta, en cuyo caso se recoge la información categorial de la entrada correspondiente; si ésta es compatible con la de los afijos usados, se devolverá como posible categoría asociada a la palabra de entrada.

Una vez efectuado el proceso básico de categorización por parte del Seg-Word, hemos introducido un paso adicional de optimización cuyo fin es facilitar la operación del analizador sintáctico. Esta fase trata de minimizar los efectos de la alta ambiguedad de categorías, siempre teniendo en cuenta las especiales características de las definiciones del diccionario. Así, se preprocesan casos de coordinación, detección de "patterns" arquetípicos (segun el subconjunto de acepciones), etc. Para mayor información sobre estos procesos adicionales al Seg-Word [Ageno et al. 91a].

El FPar es un analizador "pattern-based" que utiliza una gramática en forma de jerarquía de "patterns", y lleva a cabo un análisis descendente, aplicando patrones más específicos a medida que los de niveles superiores (más generales) van verificándose, proporcionando así un análisis parcial cuando no le sea posible aplicar patrones más detallados.

La estructura jerárquica permite también dar prioridad a la extracción de los componentes más importantes (en nuestro caso, el término genérico de las acepciones), y por otro lado restringir la aplicación de "patterns" a aquellos casos donde más probable sea su éxito (según la categoría sintáctica de la entrada, el ´smbito semántico a que pertenece,...). Se pretende así proporcionar suficiente información semántica para construir las taxonomías y propiedades asociadas [Ageno et al. 91b].

En cada regla de la gramática se especifica:

(<identificador> <pattern de frase> <identificadores hijos>)

Pudiendo aparecer en el pattern de frase literales o diversos tipos de variables (asociadas a categorías, arbitrarias, obligatorias, opcionales, ...) .

Además, cada regla de análisis o "context-free" tiene asociada una regla de construcción de estructura semántica, tal que las asociaciones de variables generadas por el proceso de "matching" se usan para generar los datos semánticos, etiquetados según especifique la regla.

Existe además la opción de especificar en determinados casos, transformaciones a efectuar automáticamente en la estructura resultante así como la posibilidad de categorizar ciertas palabras (en un lexicón adicional) y evitar de esta manera el uso del Seg-Word.

Por último, veamos el resultado del análisis en el ejemplo siguiente:

carbolíneo [de carbón + l. oleum , aceite]
acepción:1 ** m. ** Substancia líquida y grasa, obtenida de la destilación del alquitrán de la hulla, us. para hacer impermeable la madera.

(((CLASS SUBSTANCIA) (PROPERTIES (L\íQUIDA GRASA))
(SOURCE (DESTILACI\óN (PREP-MOD (DEL (OBJECT ALQUITR\áN)))))
(R-130)))

Esta estructura de salida nos permite establecer que la única acepción de *carbolíneo* corresponde a un hipónimo de una de las acepciones de substancia. Además, se señala la relación semántica FUENTE "destilación del alquitrán" y las propiedades "líquida" y "grasa".

Dada una entrada inicial, proporcionada por el usuario, el sistema busca automáticamente todas sus ocurrencias dentro de las definiciones del Vox, usando la LDB. A continuación, se analizan las definiciones selecionando únicamente aquellas en las que la entrada inicial aparece como término genérico. Si la palabra inicial es su término genérico pasaremos al proceso de desambiguación. En caso contrario, pasaremos a la siguiente ocurrencia.

Como una entrada puede tener más de un sentido, una vez se determina que una acepción se considera hipónima de una entrada sólo nos resta vincularla a alguna de sus acepciones. Este proceso está asistido por el sistema mediante un conjunto de heurísticos que determinan qué acepción de la entrada tiene más posiblidades de ser el auténtico hiperónimo que buscamos. El sistema sólo sugiere cual puede ser el hiperónimo, el usuario debe ratificar o rectificar el resultado. En caso de que la entrada tenga una sola definición la asignación será automática. El exito o fracaso de los heurísticos se registra para una posterior evaluación.

Para mejorar el rendimiento de los heurísticos, durante el proceso de construcción de la taxonomía, cierta información de las definiciones de los nodos superiores, como el tema y las palabras más significativas, son heredadas por el nodo actual que está siendo desambiguado. De esta forma, los heurísticos no sólo trabajan con la información de la acepción a desambiguar sino también con la información adquirida en los nodos superiores.

Cuando determinamos que una acepción es hipónima de otra, la entrada a la cual pertenece esa acepción se convierte en la siguiente entrada raíz. Cuando una acepción no genera ningún hipónimo, éste se convierte en un nodo terminal de la taxonomía.

Dado que la realización de una taxonomía completa es un proceso largo, debido al tiempo de análisis empleado, el sistema ofrece la posibilidad de construirla de forma incremental. El sistema también permite el tratamiento y posterior modificación de las taxonomías, eliminando o rehaciendo partes de éstas.

2.3.2 Adquisición semántica.

Una vez tenemos una taxonomía creada, disponemos de una estructura en forma de árbol donde todas las acepciones que pertenecen a ella están conectadas con su hiperónimo (excepto la raiz de la taxonomía) y con sus hipónimos (excepto las acepciones terminales).

El siguiente paso consiste en realizar el mismo proceso descrito anteriormente pero con una gramática distinta y sin la intervención del usuario. La gramática, por supuesto, debe ser más completa y compleja que la usada para la extracción del término genérico y debe permitir la extracción de la 'diferentia' [Calzolari 91] [Ageno et al. 91b] de las definiciones de la taxonomía. Cada taxonomía asociada a un ámbito semántico concreto tiene su própia gramática (por ejemplo, la información que podemos extraer de la taxonomía de "persona" es diferente a la que podemos extraer de la taxonomía de "substancia").

Dado que el proceso de extracción de la información semántica de las acepciones contenidas en una taxonomía puede dar resultados erroneos y/o incompletos, es necesario un proceso de validación de estos análisis. Este proceso es interactivo y permite observar el incremento de información capturada a medida que se emplean las distintas gramáticas.

Viendo el resultado de los múltiples análisis, el usuario puede determinar las modificaciones que deben realizarse en la gramática y/o en el módulo morfológico. Cuando el usuario está finalmente satisfecho con el resultado del análisis de las acepciones de la taxonomía, o bien considera que no pueden mejorarse, se lleva a cabo el proceso manual de corrección de los últimos análisis. Este proceso es también interactivo y permite al usuario validar, y si es necesario, corregirlos o modificarlos. Una vez terminado este proceso, la taxonomía está preparada para

comenzar la creación interactiva de las nuevas entradas léxicas en la LKB a partir de los análisis validados. Esto se realiza en el módulo de conversión [Ageno et al. 91c, d].

2.3.3.La Base de Conocimiento léxico

El lenguaje de representación [Copestake 91] puede verse como un formalismo basado en la unificación de grafos tipados, con herencia obligatoria, múltiple, permitiendo expresar de forma relacionada tanto información sintáctica como semántica a gran escala. El sistema de tipos se utiliza como soporte para el mecanismo de herencia así como para restringir otras operaciones. Además también se incluye la posibilidad de herencia por defecto formalizada en términos de una operación de unificación-por -defecto de un conjunto de características (llamadas psorts) asociadas a los tipos y ordenadas en una estructura jerárquica.

Las operaciones que soporta la BCL son herencia por defecto, unificación por defecto y aplicación de reglas léxicas.

El sistema de tipos

La jerarquía de tipos define un orden parcial, y especifica los tipos que son consistentes. Solo pueden unificarse las estructuras de características asociadas a tipos mutuamente consistentes. En la figura 6 puede verse un fragmento de una jerarquía de tipos (se utiliza el inglés como metalenguaje para la BCL multilingue), en las que aparecen los nodos *artefacto* y *físico*. Ambos son consistentes, pudiendo definirse un nuevo tipo como resultado de la unificación de los mismos: *artefactofísico*. (physart)

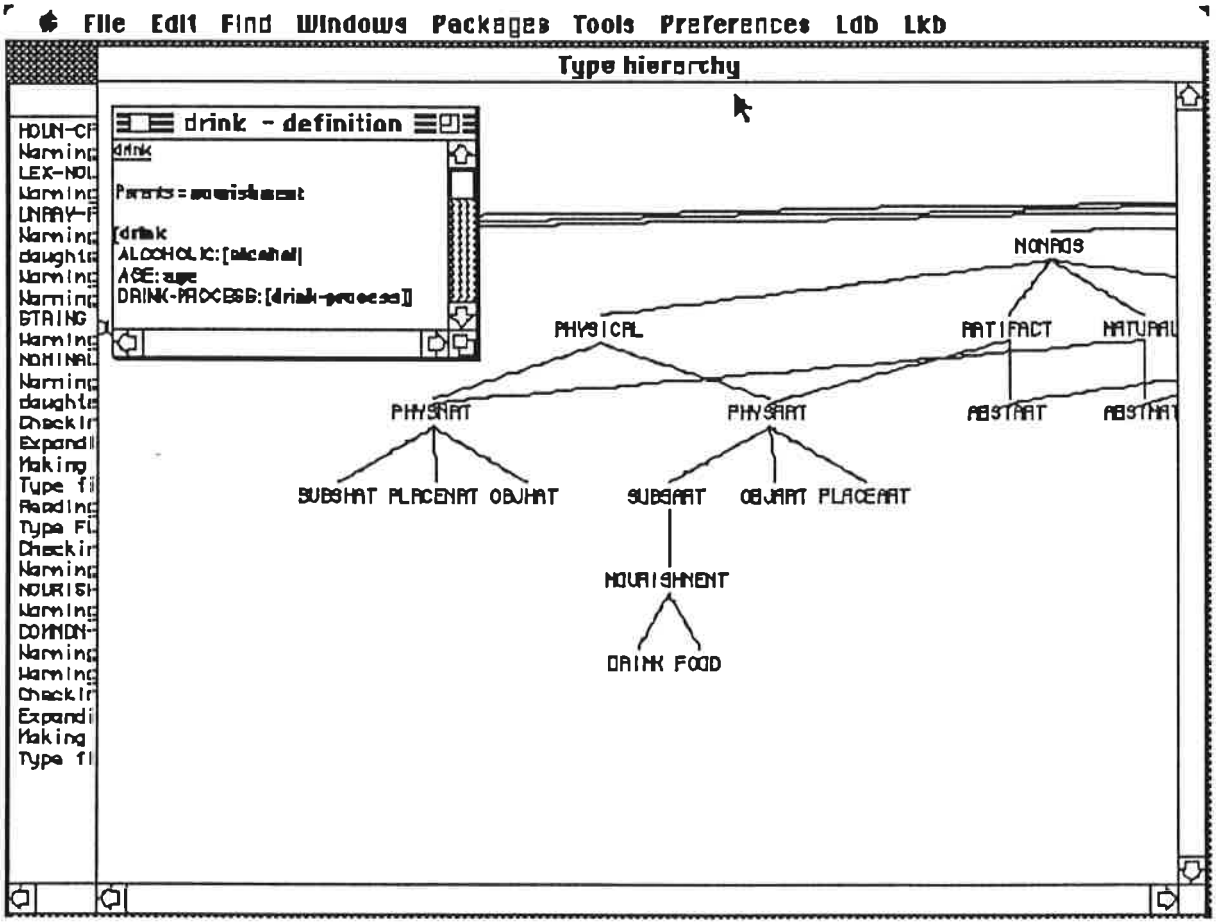


Figura 6: vista parcial del sistema de tipos.

La herencia de estructuras permite definir de forma concisa nuevos tipos. Como ejemplo puede citarse la definición del nodo *comida* (food) , donde solo es necesario especificar el tipo padre *alimento* (nourishment) y las propiedades directas que posee. Parte de su definición completa puede verse en la figura 7, en donde aparece explícitamente el resto de la información que automáticamente ha heredado a través del sistema de tipos.

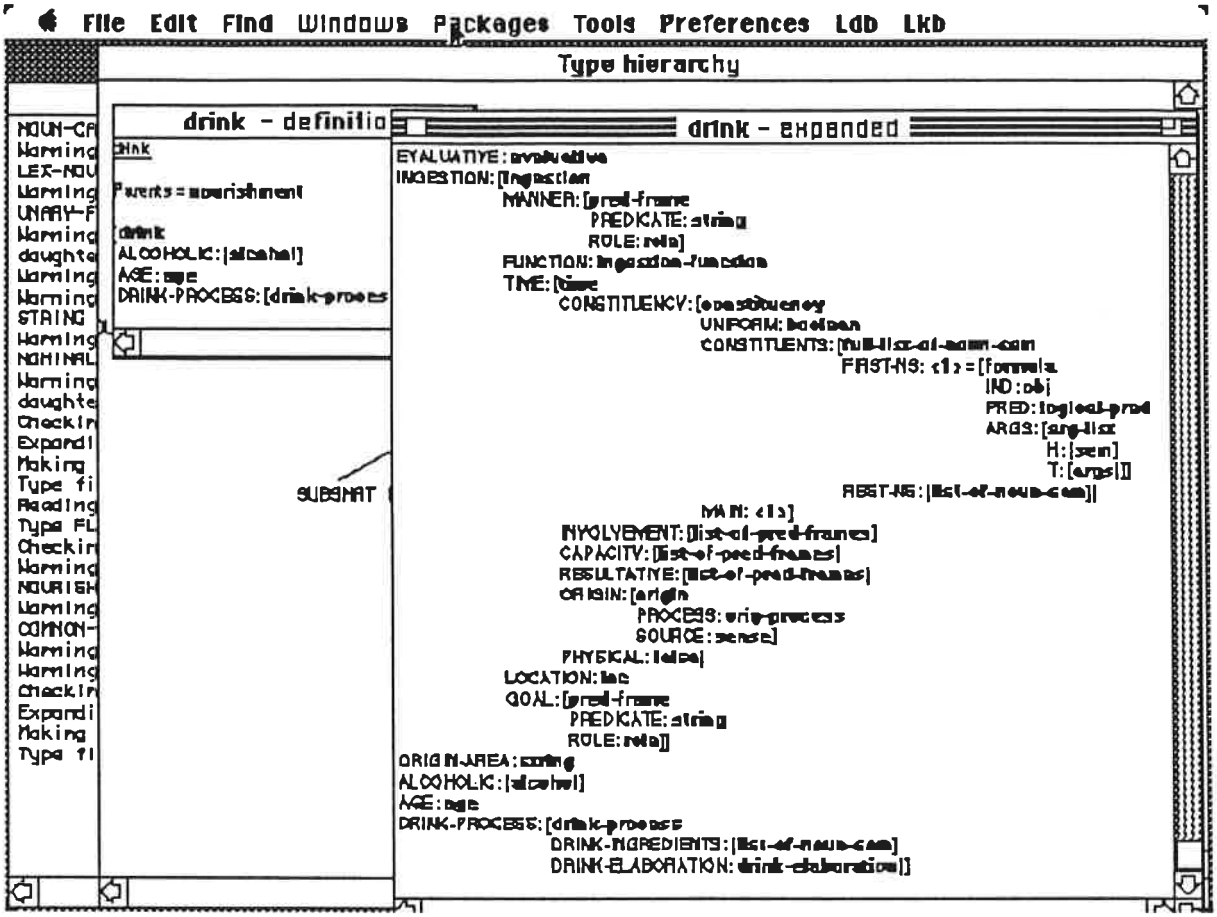


Figura 7: vista extendida del tipo *food.*

3. La adquisición conceptual y léxica en GUAI

GUAI (Generador Universal de Interfaces) [Rodriguez-89] es un generador de interfaces en lengua natural. Su misión consiste en adquirir el conocimiento ligado a las características específicas que una interfaz concreta debe poseer y generarla. La interfaz generada se incluye dentro de la aplicación para la que ha sido diseñada y asume, en forma autónoma e inteligente, la tarea comunicativa, en lengua natural, entre dicha aplicación y el interlocutor humano. El conocimiento que debe ser adquirido incluye elementos relativos al entorno operativo en que la interfaz debe insertarse, al dominio semántico de la aplicación, al usuario o interlocutor humano y a la propia interfaz.De las diversas adquisiciones de conocimiento que GUAI debe abordar, la más crítica, tanto por su volumen como por su frecuencia y la complejidad de las relaciones implicadas es la que se refiere al conocimiento léxico y conceptual.

El dominio semántico de cada aplicación prefigura los objetos y relaciones que constituyen el objeto del proceso comunicativo y determinan la cobertura conceptual de los diálogos que lo constituyen. GUAI organiza su dominio semántico en forma de estructura taxonómica en la cual las unidades son los conceptos básicos. Además de la relación ES_UN que articula la taxonomía, otras relaciones semánticamente significativas tienen cabida en el formalismo. Los niveles

superiores de la taxonomía son relativamente independientes del dominio y , por tanto, comunes a la mayoría de las aplicaciones. Los objetos específicos de cada aplicación, normalmente situados en los niveles inferiores de la taxonomía deberán definirse en cada caso.

La adquisición en Guai se realiza a través de la interacción del Sistema con un experto humano. Se escogió esta opción fundamentalmente por dos razones:

1. La adquisición basada en diccionarios como hemos visto en el apartado anterior es fundamentalmente léxica. Es decir se llega a los conceptos a través de las palabras. Para un marco como Guai en el que el proceso inferencial se lleva a cabo sobre todo en el nivel conceptual es más razonable el enfoque inverso, definir primero los conceptos y luego las palabras que lo realizan., en palabras de Niremburg, "world first, word later".

2. GUAI es un sistema que trabaja en dominios semánticos restringidos y por tanto no tiene demasiado sentido alimentarlo a través de fuentes forzosamente generales, y no inmediatamente disponibles.

3.1 La adquisición conceptual

Los conceptos en GUAI se especifican mediante una serie de *descriptores*. Cada descriptor queda caracterizado por la clase de objetos que lo soportan (su dominio) y por la clase de objetos que admite como valores (su rango). Las posibilidades de descripción del rango son: una clase, una lista de clases o una lista de ejemplares. Restricciones más complejas pueden incorporarse al concepto a través de su *delineación*. GUAI admite dos tipos de descriptores: los estructurales, encargados de expresar vínculos del concepto con otros conceptos, no siempre expresados a través de relaciones, y los restricitivos o de asignación de cualidades. Esta estructura describe clases de objetos. Los ejemplares concretos son instancias de dichas clases. GUAI permite que determinados descriptores se apliquen unicamente a las clases y no a sus ejemplares, es decir están expresamente excluidos del proceso de herencia.

Las operaciones que GUAI admite para la adquisición conceptual son obtención, búsqueda, modificación, alta y baja.

La operación de obtención permite acceder a un concepto a través de su nombre. El concepto es presentado graficamente en su entorno, es decir ligado, mediante arcos, a los conceptos relacionados con él. Se pueden seleccionar las relaciones a mostrar. Varios niveles de "zooming" son factibles. También es posible obtener la lista de todas las posibles realizaciones léxicas del concepto.

Las operaciones de BAJA y MODIFICACION implican en primer lugar la localización del concepto a dar de baja/modificar. Ello se puede lograr nombrándolo explícitamente o señalando un concepto presente en la pantalla. Toda una serie de comprobaciones se realizan, por ejemplo, un concepto no puede ser eliminado si existen referencias léxicas a él.

Para dar de ALTA un nuevo concepto el primer paso será situarlo dentro de la estructura taxonómica. Distinguiremos dos casos: que el nuevo concepto sea un terminal descendiente directo de un concepto ya existente en la estructura o que deba incluirse en el interior de la misma. En el primier caso, GUAI propondrá al usuario las propiedades, descriptores y restricciones sobre el rango de los mismos, asociadas al ascendiente en la jerarquía para que éste pueda, editando estos valores, seleccionar las del concepto a incluir. Los rangos o valores asociados a los descriptores deberán ser más restringidos que los del ascendiente. Es también posible añadir nuevos decritpores a los propuestos por el sistema. En este caso el nuevo descriptor deberá existir. El valor por defecto que GUAI propondría al ususrio sería, en este caso, el correspondiente al rango del descriptor. En cuanto a la inserción interna de un nuevo concepto, las direrencias afectan básicamente a la comunicación al usuario de los efectos laterales sobre los descendientes del concepto a través de cualquiera de sus relaciones. El usuario deberá simplemente darse por enterado de estos efectos y confirmar el alta o no hacerlo.

3.2 La adquisición léxica

La información léxica en GUAI está distribuída entre los diccionarios de raíces, sufijos, modelos de raíz, modelos de sufijo y semántico. Los módulos de adquisición de los distintos diccionarios son independientes. La modificación de sufijos y de modelos, así como la de otras estructuras de datos morfológicas requieren un interlocutor especialista y se tratan independientemente del resto de la información léxica.

En cuanto al diccionario semántico, las operaciones posibles son OBTENCION, ALTA, BAJA, y MODIFICACION. La OBTENCION se puede realizar a través de un lexema, a través de una raíz o a través de una palabra. El lexema constituye la entrada al diccionario semántico. Si se accede a través de una raíz se obtiene la información accesible a cada una de sus posibles interpretaciones. Si la entrada es a través de una palabra sólo las interpretaciones semánticas compatibles con la categoría morfológica serán recogidas. La información accesible incluye la categoría semántica , y si ello es posible, su interpretación. El ALTA y la MODIFICACION exigen la verificación de todas las restricciones asociadas a los objetos implicados en especial los conceptos asociados a los argumentos de la categoría semántica y que formarán parte de la interpretación.

Las operaciones posibles en el mantenimiento del diccionario de Raíces son también OBTENCION, ALTA, BAJA, y MODIFICACION. El mantenimiento del Diccionario de Modelos de Raíz es similar al de Raíces. La información que se le da al usuario incluye, en este caso, la lista de raíces asociadas al modelo. Se puede, a iniciativa del usuario, trasladar información (rasgos morfológicos), del modelo a sus raíces y viceversa. Es posible también obtener la lista de reglas en las que el modelo figura. La ausencia de elementos en una u otra lista es detectada y comunicada al usuario. Los diccionarios de sufijos y de modelos de sufijo son en todo similares a los de raíces y su adquisición sigue un proceso semejante.

4. Conclusiones

Existen dos enfoques básicos para el tratamiento de la adquisición de conocimiento léxico y conceptual con el fin de construir bases de conocimiento utilizables por sistemas automáticos de procesamiento del lenguaje natural:

- Adquisición a través de diccionarios convencionales
- Adquisición a través de la interacción con un experto humano

Ambos enfoques, en el estado actual del arte, no permiten procesos totalmente autónomos, es decir en cualquiera de los casos se requiere siempre de una interacción con un experto.

En nuestra opinión ambos enfoques son complementarios, la ontología creada a partir del primer enfoque, necesariamente general, puede ser aprovechada como una base de conocimiento global, común a diferentes aplicaciones, mientras que aquellos aspectos más específicos, ligados a un dominio concreto, pueden introducirse a través de un diálogo guiado tal y como se propugna en el segundo enfoque. Por tanto una linea abierta sería el diseño de un entorno que englobara las dos aproximaciones, permitiendo mediante operaciones sencillas aprovechar el subconjunto de interés de la base general, y refinándolo atendiendo a la ontología específica del dominio que se deea representar.

Referencias

[Acquilex89] Acquilex.: "Technical Annex". ESPRIT BRA-3030 ACQUILEX

[Ageno et al. 91a] Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F.: "An environment for management and extraction of taxonomies from on-line dictionaries". Universitat Politécnica de Catalunya, Barcelona.ESPRIT BRA-3030 ACQUILEX WP NO.020

[Ageno et al. 91b] Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F.:"The Extraction of Semantic Information from MRDs". Universitat Politécnica de Catalunya, Barcelona. ESPRIT BRA-3030 ACQUILEX WP NO.027

[Ageno et al. 91c] Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F., forthcoming.:"From LDB to LKB". Universitat Politécnica de Catalunya, Barcelona.ESPRIT BRA-3030 ACQUILEX WP NO.028

[Ageno et al. 91d] Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F., forthcoming.: "A Semi-automatic Process to create LKB entries". Universitat Politécnica de Catalunya, Barcelona. ESPRIT BRA-3030 ACQUILEX WP NO.029

[Alshawi 89] Alshawi H.: "Analysing the dictionary definitions". In Boguraev B., Briscoe T. (eds) *Computational Lexicography for NLP* , chapter 7. Longman, London.

[Alsawi 90] Alshawi H.: "Flexible Pattern Matching Parsing Tool (FPar). Technical Manual. Computer Laboratory, University of Cambridge. ESPRIT BRA-3030 ACQUILEX

[Amsler 81] Amsler R.: "A taxonomy for English nouns and verbs". *Proceedings of the 19th Anual Meeting of the ACL,* Stanford, California, pp 133-8.

[Calzolari 90] Calzolari N., Peters C., Roventini A.: "Computational model of the dictionary entry". Preliminary Repport. 6 Month Deliverable. Pisa. ESPRIT BRA-3030 ACQUILEX ICL-ACQ-1-90

[Calzolari 91] Calzolari N.: "Acquiring and Representing Semantic Information in a Lexical Knowledge Base". *Proceedings of the Workshop on Lexical Semantics,* Berkeley, USA.ESPRIT BRA-3030 ACQUILEX WP NO.016

[Carroll 90] Carroll J.: "Lexical Data Base System User Manual". Computer Laboratory, University of Cambridge.ESPRIT BRA-3030 ACQUILEX

[Castellón et al. 90] Castellón I., Martí M. A.: "Gramática del Diccionario Vox". *Proceedings of the 6th Anual Meeting of the SEPLN* . San Sebastian, Spain.

[Castellón et al. 91] Castellón I., Martí M. A., Rigau G., Rodríguez H., Verdejo M. F.: "Loading the MRD into the LDB. Characteristics of Vox Dictionary". Universitat Politécnica de Catalunya, Barcelona. ESPRIT BRA-3030 ACQUILEX WP NO.019

[Cater90] Cater, A.: "Tesbed Enghish Language Analyser". Deriverable #7. University College, Dublin.ESPRIT BRA-3030 ACQUILEX

[Copestake 90a] Copestake A.: "Building Taxonomies with disambiguated word senses". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO.008

[Copestake 90b] Copestake A.: "A System for building disambiguated taxonomies: draft version". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO.012

[Copestake 91] Copestake A: "The LKB: a system for representing lexical information extracted form machine-readable Dictionaries.In Proceedings of the ACQUILEX Workshop on Default Inheritence. Cambridge 1991.

[Rodriguez 89] Rodriguez Hontoria H: "GUAI un generador automático de interfaces en lengua natural". Tesis Doctoral. Universidad Politécnica de Cataluña.1989.

[Sanfilippo 90a] Sanfilippo A.: "A morphological Analyser for English & Italian". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO. 004

[Sanfilippo 90b] Sanfilippo A.: "Notes on Seg-Word". Computer Laboratory, University of Cambridge.ESPRIT BRA-3030 ACQUILEX

[Vox 87] "Diccionario General Ilustrado de la Lengua Española VOX". Ed. Biblograf S.A. Barcelona.

(1) **ACQUILEX** es un proyecto integrado, en el que participan las universidades de Amsterdam, Cambridge, Dublin y Politécnica de Catalunya y el Instituto de Lingüística Computacional de Pisa. El proyecto esta financiado por la C.E.E. a través del programa ESPRIT (Acción 3030).

(2) **GUAI** es un prototipo para desarrollar interfaces que permitan la comunicación en lenguaje natural entre una persona y un sistema informático.