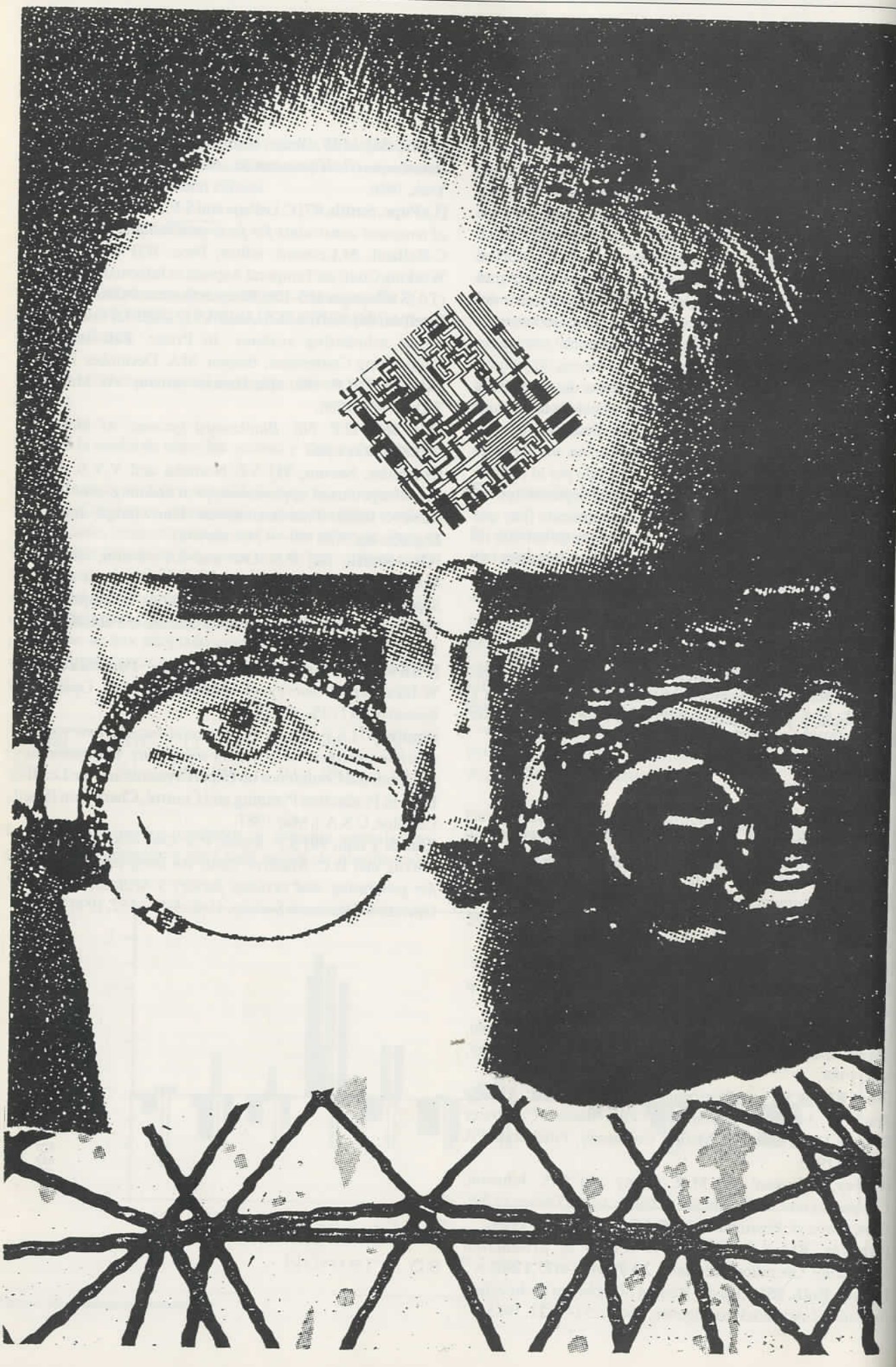


Inteligencia Artificial

Un entorno para la extracción de información semántica del diccionario VOX

Alicia Ageno, Irene Castellón, Francesc Ribas, Germán Rigau, Horacio Rodriguez;
Departament de LSI, UPC, Barcelona.
M.A. Martí, Mariona Taulé;
Departament de Filología Romànica, Univ.Barcelona.
Felisa Verdejo;
Dp.Ing.Eléctrica, Electrónica y Control; UNED Madrid

1. Introducción

Presentamos aquí un entorno interactivo para la obtención y representación de información semántica de una base de datos léxica construida a partir de diccionarios en soporte magnético. El entorno posibilita el tratamiento de la información léxica, realizando de forma automática los procesos que lo permitan y dotando al lexicógrafo de la ayudas precisas para completar, en forma interactiva, el proceso de extracción. El presente artículo se organiza del siguiente modo: la sección 2 presenta una breve exposición del marco dónde se desarrolla nuestro sistema, el proyecto integrado ACQUILEX, en el que participan el Instituto de Lingüística Computacional de Pisa y las universidades de Amsterdam, Cambridge, Dublin y Politécnica de Catalunya: proyecto esta financiado por la C.E.E. a través del programa ESPRIT (Acción BRA 3030). En la sección 3 se expone el sistema de forma general. La sección 4 se centra en la extracción de taxonomías, a partir del diccionario Vox. Por último, la sección 5 establece una serie de conclusiones y perspectivas de futuro.

2. El Proyecto Acquilex

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

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 PLN 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 no siempre sistemática 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 el entorno de extracción de información semántica que constituye nuestro artículo. 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 o sistema de tipos, 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.

2.1. Objetivos primeros del proyecto

Se centran en el desarrollo de un prototipo de Base de Datos Léxica (LDB) y de Base de Conocimientos Léxica (LKB) 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 MRDs. 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 MRD.

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 MRD.

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 de bases léxicas disponibles.

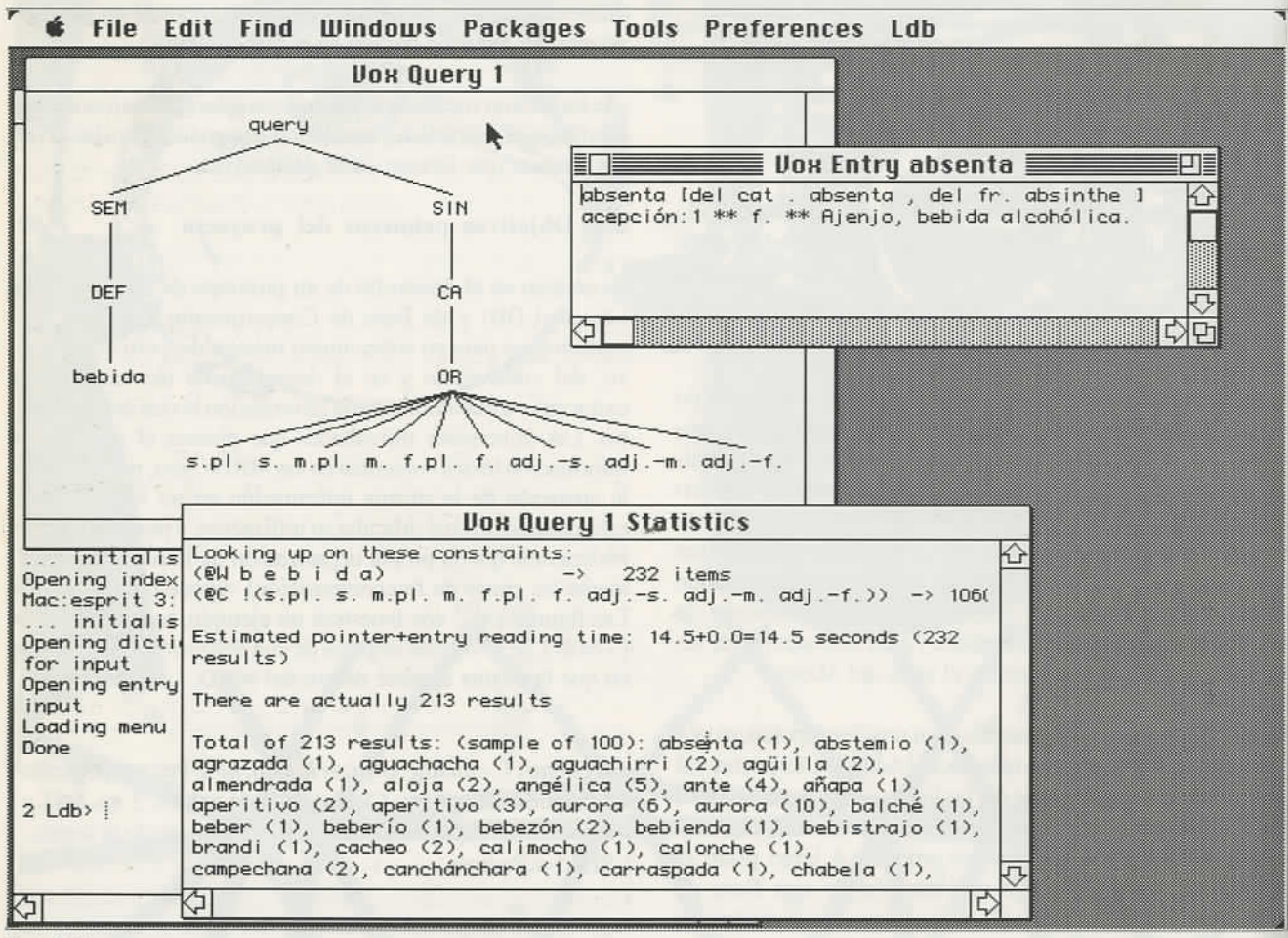
La primera característica de nuestra propuesta es la de trabajar en un contexto multilingüe. De hecho, el material con el que trabajamos incluye actualmente el LDOCE (Longman Dictionary of Contemporary English) inglés, el Garzanti italiano, el Van Dale holandés, el VOX español y los bilingües COLLINS 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.

2.2. Etapas principales en el desarrollo del proyecto

1. Elaboración de un modelo computacional de diccionario de forma que tengan expresión en él todas las características diferenciales de los diferentes diccionarios individuales [Calzolari 90].

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

3. Definición y desarrollo de software para la gestión de la BD Léxica, LDB [Carroll 90]. La **pantalla 1** muestra el aspecto de una sesión de trabajo en el entorno de la LDB: 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 LDB [Castellón et al. 90,91]. La **figura 3** muestra el resultado del proceso de formalización de información léxica a partir de la entrada que las **figuras 1 y 2** mostraban. La estructura (que se suele conocer como **"lispificada"**) contiene la misma información presente en el diccionario fuente clasificada y segmentada, y constituye la entrada al proceso de carga de la LDB. La transformación de las entradas en formato MRD 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 MRD 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. Extracción de información semántica de las definiciones a partir de la información contenida en la LDB.

7. Carga en la LKB 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 PLN cuyo componente léxico se haya extraído de la LKB. El Sistema chequeará las dos funciones de comprensión y generación [Cater 90].

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.

3. 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 o sentidos de las entradas del diccionario. La principal relación con la que trabajamos es la clase-subclase (**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, consistióen 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], [Carroll 90] y el analizador morfológico Seg-Word [Sanfilippo 90a,90b] así como también hemos usado la aproximación para la extracción de información taxonómica a partir de las definiciones del diccionario Tax-Build [Copestake 90a] [Copestake 90b] y el lenguaje de representación del conocimiento léxico LKB.

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.

3.1. Fuentes de conocimiento

La principal fuente de conocimiento de nuestro entorno es por supuesto la LDB, que 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 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 gramáticas para el análisis de las definiciones.

- El conjunto de heurísticos que son aplicados en distintos puntos para ayudar al usuario en su toma de decisiones.

- Reglas de conversión que nos permitan transportar la información resultante de la adquisición de semántica a entradas léxicas susceptibles de ser cargadas en la LKB.

- *Tlinks*, enlaces de traducción que permitan vincular las entradas léxicas correspondientes a los distintos lexicones monolingües.

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 [Ageno et al. 91b]. Estas gramáticas, al principio, tampoco captan toda la información que deseamos extraer. Debemos mejorarlas paulatinamente hasta conseguir los resultados esperados.

3.2. El proceso

El 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 términos genéricos en las acepciones [Copestake 90b]. La segunda fase permite la extracción de las propiedades semánticas, no taxonómicas, que aparecen en las acepciones de la taxonomía creada anteriormente. En la tercera fase, se validan los heurísticos aplicados en la construcción de la taxonomía. Toda la información adquirida anteriormente se integra en la LKB mediante un proceso de conversión. Finalmente, el proceso de *mapping* vincula entradas léxicas de distintas fuentes.

Una parte del proceso de extracción incluye un conjunto de heurísticos que permiten automatizar el proceso de *desambigüación* 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: *adquisición* y *validación*. En modo adquisición, el sistema facilita la construcción interactiva de las taxonomías y el posterior análisis semán-tico de las definiciones. El análisis semántico de las taxonomías así como su conversión no necesitan intervención humana y pueden realizarse en un proceso batch. En modo validación, el análisis realizado anteriormente se valida y corrige mediante un proceso interactivo. El usuario 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 y reglas de conversión de cada taxonomía se modifican de forma incremental hasta conseguir un resultado óptimo. A continuación, en modo validación, se puede confrontar una estructura taxonómica existente considerada correcta con diferentes conjuntos de heurísticos, lo que tampoco necesita la intervención del usuario. Para facilitar la comprensión al lector, en la figura 5 el proceso de extracción de las relaciones semánticas no taxonómicas se ha colocado antes que el proceso de validación de los heurísticos, pero ambos son independientes

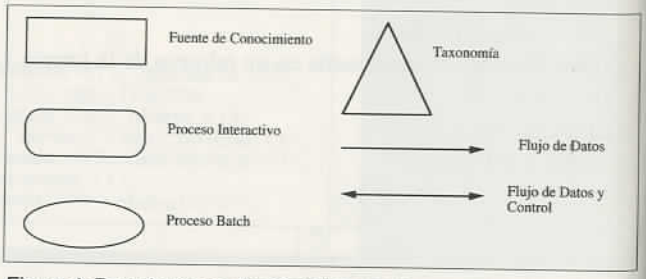


Figura 4: Descripción de los símbolos usados

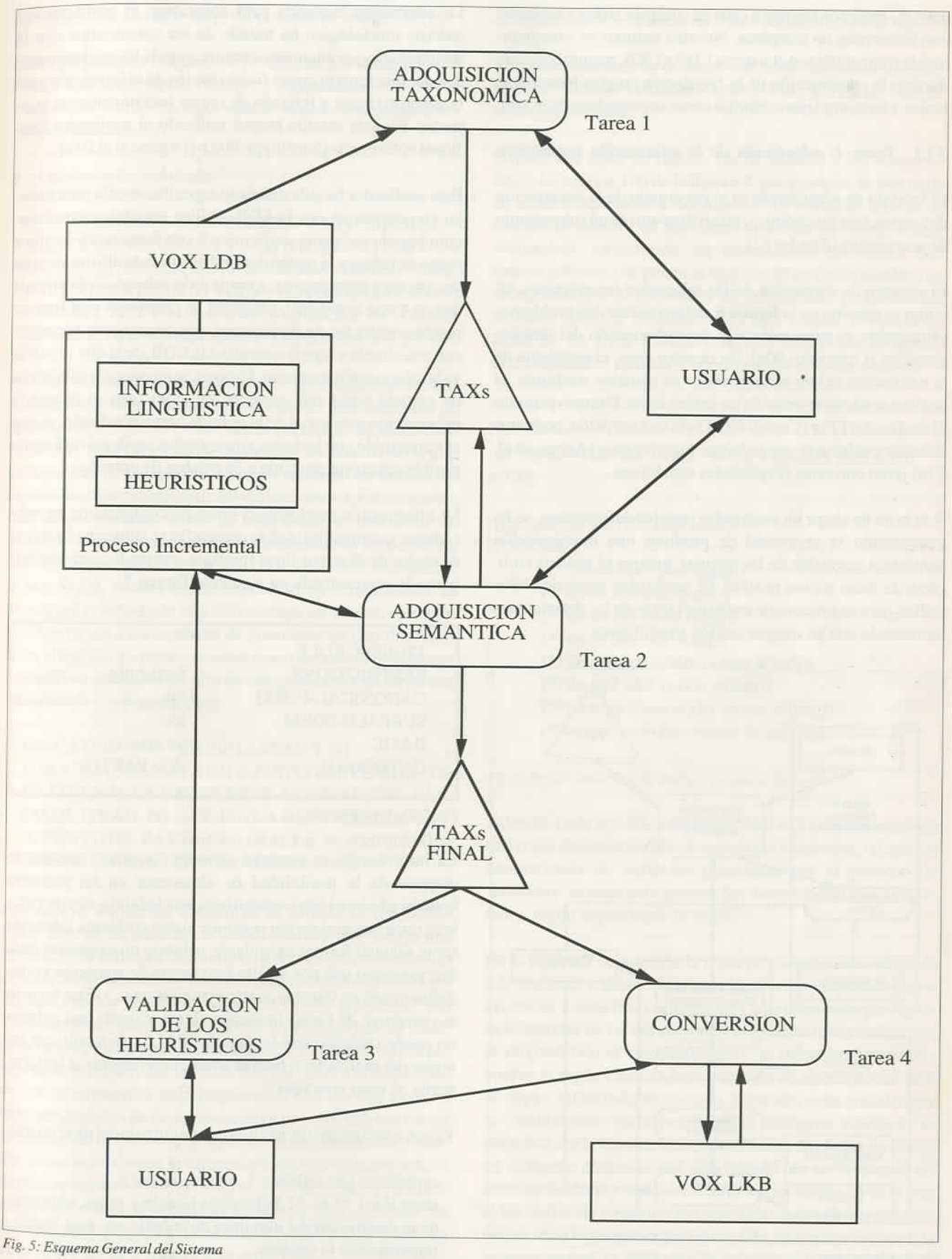


Fig. 5: Esquema General del Sistema

entre sí, pudiendo llevarse a cabo en cualquier orden y realizarse con taxonomías no completas. Nuestro entorno es completamente compatible con el sistema LDB y LKB, permitiendo que durante la construcción de la taxonomía puedan formularse tantas y tan complejas consultas como se considere necesario.

3.2.1 Tarea 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, tal como se muestra en la **figura 6**, debe resolver dos problemas principales: la extracción y la desambigüación 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 [Carroll 89]. Dada una acepción, podemos detectar cuáles son sus palabras hiperónimas [Ageno et al. 91a] junto con otras propiedades semánticas.

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.

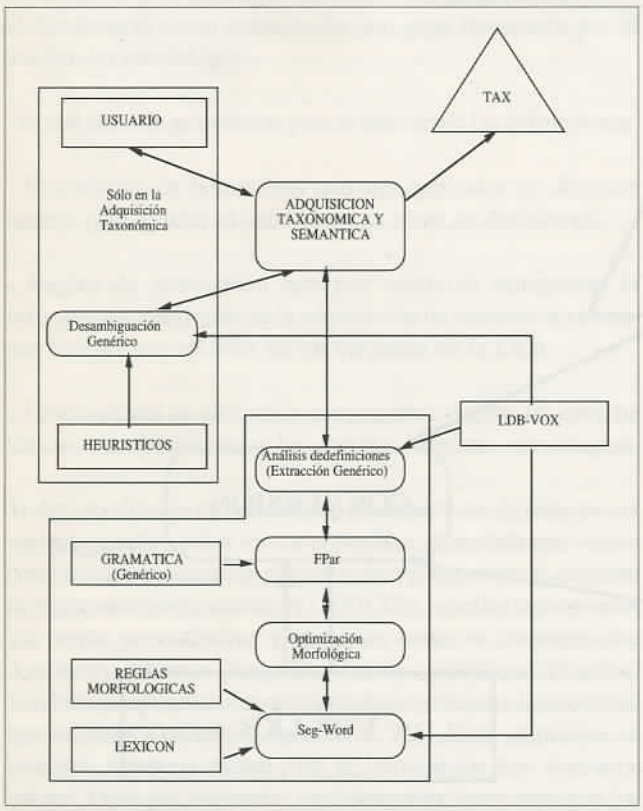


Fig. 6: Adquisición

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 90a,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 y la particularidad de que ésta última no se ha de construir previamente: a partir de la subcadena de entrada que se toma como raíz potencial se construye una lista de posibles entradas de diccionario, que se usa para accederle con una simple y rápida consulta a la LDB, de la que 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.

La información morfológica sobre descomposición en subcadenas y compatibilidad de categorías se almacena en forma de reglas de diversos tipos (prefijos, sufijos y composición) como la representada en siguiente **figura 7**:

```
(SUFFIX-RULE
RESTRICTIONS          final-only
CANONICAL-FORM        ido
SURFACE-FORM          ido
BASIC                 (er ir)
CATEGORY              (V > PARTI))
```

Fig. 7: Regla morfológica.

La información se consulta siempre "on-line", aunque se dispone de la posibilidad de almacenar en un pequeño lexicón adicional la(s) categoría(s) asociada(s) a ciertas palabras cuya categorización ocasione algún problema (abreviaturas, algunas formas irregulares, palabras no existentes en el diccionario o que por su alta frequencia de aparición en las definiciones no interesa analizarlas cada vez...). Este lexicón se construye de forma incremental; así, cuando una palabra no puede categorizarse (categoría "PAL"), se modifican las reglas del analizador o bien se añade directamente al lexicón, según el caso concreto.

Veamos un ejemplo de análisis de una entrada del diccionario:

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

La entrada (en formato Lisp) al proceso de análisis sería:

(N «Substancia» «líquida» «y» «grasa» «,» «obtenida» «de» «la» «destilación» «del» «alquitrán» «de» «la» «hulla» «,» «us.» «para» «hacer» «impermeable» «la» «madera» «.»)

Y el subsecuente resultado:

((N CATEGORY) (SUBSTANCIA V N) (L\íQUIDA ADJ) (Y CONJ) (GRASA V ADJ N) (\,PUNT) (OBTENIDA PAL) (DEP) (LA DET) (DESTILACI\óN N) (DEL PAL) (ALQUITR\áN N) (DE P) (LA DET) (HULLA N) (\, PUNT) (US. PAL) (PARA P) (HACER V) (IMPERMEABLE ADJ N) (LA DET) (MADERA V N))

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 (según el subconjunto de acepciones), etc. Para mayor información sobre estos procesos adicionales al Seg-Word, véase [Ageno et al. 91a]. A continuación, se muestra el resultado de esta última etapa en nuestro ejemplo: se observa un caso concreto de coordinación (los dos adjetivos «líquida» y «grasa» se unen como un solo *item* a efectos del análisis sintáctico), y otro de categorización de un patrón predefinido («obtenida de»):

((N CATEGORY) (SUBSTANCIA V N) (L\íQUIDA_GRASA ADJ) (\,PUNT) (OBTENIDA_DE PATTERN4) (LA DET) (DESTILACION N) (DEL PAL) (ALQUITRAN N) (DE P) (LA DET) (HULLA N) (\, PUNT) (US. PATTERN3) (HACER V) (IMPERMEABLE ADJ N) (LA DET) (MADERA V N))

Esta última expresión constituye la entrada al analizador FPar. Este es un analizador *'pattern-based'* que utiliza una gramática en forma de jerarquía de patrones y lleva a cabo un análisis descendente, aplicando patrones más específicos a medida que los de niveles superiores (más generales) van verificándose y 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 ámbito 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 ...).

Existen dos tipos de reglas, las de análisis y las «context-free». La regla n-100 de la **figura 8**, por ejemplo, es una regla de *análisis* aplicable a definiciones de entradas léxicas de categorías *nombre* en las cuales aparezca un especificador obligatorio ya definido, un determinante opcional y una cadena arbitraria de palabras seguida de un único nombre, un sintagma adjetival opcional y un modificador preposicional también opcional terminando en otra cadena arbitraria de palabras. La regla *'context-free'* que define los sintagmas preposicionales aparece a continuación de la de análisis.

```
(n-100
(n *especi +0det && +noun *0s-adj *0pp-mod &&) n-110
n-120)

(opt-pats (QUOTE (*0pp-mod))
(QUOTE
      ((**de-co (de color &adj))
      (**de-to (de tono &adj))
      (**de-sa (de sabor &adj))
      (**de-ol (de olor &adj))
      (**de-pp1 (de +0det +noun &0adj))
      (**de-pp2 (del +noun &0adj))
      (**para-pp (para +0det +noun &0adj))
      (**en-pp (en +0det +noun &0adj)))))
```

Fig. 8: Reglas sintácticas de análisis y «context-free».

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.

En la **figura 9** se muestra la regla de construcción asociada a la regla de análisis mencionada, con las etiquetas que se asignarán a aquellos componentes significativos que logremos capturar de las definiciones: el genérico de la definición se etiqueta con «compound-class», la información que devuelve la regla *'context-free'* que trata el especificador será el «type», el(los) adjetivo(s) que, de existir, serán tratados por la *'context-free'* correspondiente al sintagma adjetival, se etiquetan como «properties», y por último, en el caso de existir un sintagma preposicional será tratado por su *'context-free'* correspondiente y marcado como «prep-mod». Véase por ej.las reglas de construcción asociadas a dos de los componentes de la regla que define el modificador preposicional: la primera extrae la información correspondiente a aquellos

modificadores que indican un color (señalando que se trata de esta propiedad así como el color en concreto), y la segunda a aquellos modificadores que presentan la preposición «de» seguida de un nombre (que se etiqueta como «object») y posiblemente de uno o más adjetivos (también marcados como «properties» pero esta vez del objeto en cuestión).

```
(n-100
        ((compound-class +noun)
        (type *especi)
        (properties *0s-adj)
        (prep-mod  *0pp-mod))

(**de-co (color &adj))

(**de-pp1 (de (object +noun) (properties  &0adj)))
```

Fig. 9: Reglas sintácticas de construcción.

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 del apartado anterior:

(((CLASS SUBSTANCIA) (PROPERTIES (LIQUIDA GRASA)) (SOURCE (DESTILACION (PREP-MOD (DEL (OBJECT ALQUITRAN))))) (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 y 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 desambigüado. De esta forma, los heurísticos no sólo trabajan con la información de la acepción a desambigüar, 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.

3.2.2 Tarea 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 en la **figura 6**, pero con una gramática distinta y sin la intervención del usuario (véase la **figura 10**). 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, que es interactivo, 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.

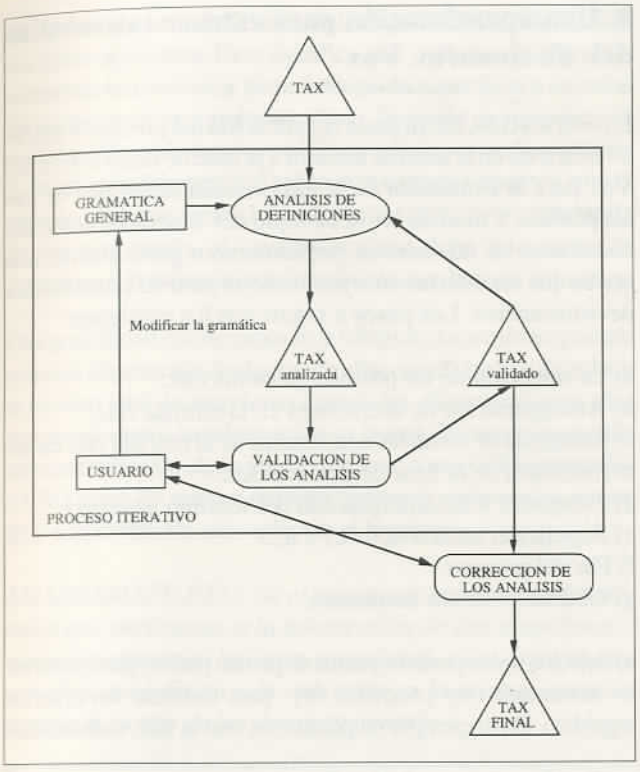


Fig. 10: Adquisición semántica

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 validarlos, y si es necesario, corregirlos o modificarlos; una vez terminado, 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].

3.2.3 Tarea 3: Validación de los heurísticos

Una vez creada la taxonomía y por lo tanto sin ambigüedad (de la acepción de taxonomía sabemos cuál es su hiperhónimo y cuáles sus hipónimos) podemos poner a punto los heurísticos empleados en su construcción. Este proceso se lleva a cabo también sin la supervisión del usuario. Sabiendo el éxito o fracaso de los heurísticos, pueden cambiarse sus parámetros para conseguir un mejor rendimiento del sistema. Los heurísticos implementan estrategias para la toma de decisiones allí donde no está definida ninguna solución algorítmica.

Básicamente un heurístico es un procedimiento que asigna una puntuación a cada una de las diferentes opciones que se le plantean. Para cada decisión a tomar, se puede aplicar un conjunto de heurísticos. Tras su actuación, se calcula una puntuación global, partiendo de las asignadas por los diferentesheurísticosa partir de la cual se toma una decisión.

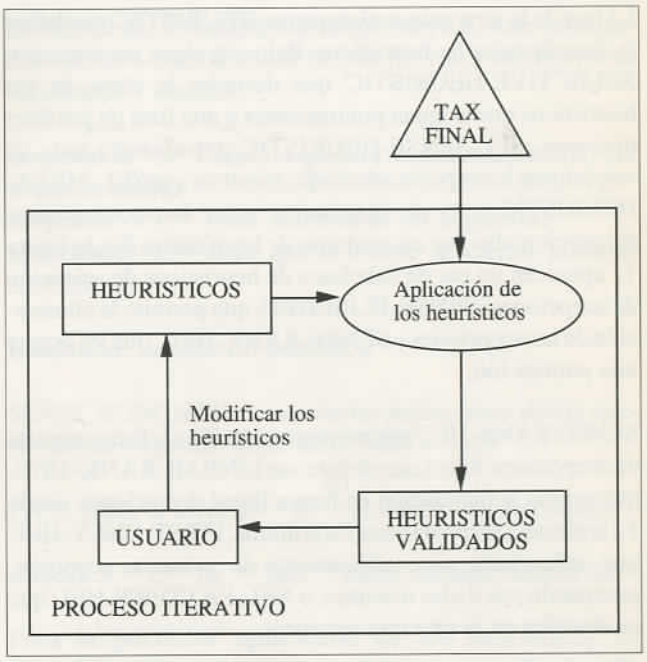


Fig. 12: Vista parcial de la estructura de heurísticos

Los heurísticos se representan en forma declarativa, mediante un formalismo orientado a objetos. Los diferentes heurísticos que se aplicarán en diferentes puntos de decisión, se organizan en una estructura jerárquica que permite la herencia de propiedades. Una muestra de la estructura aparece en la **figura 12**.

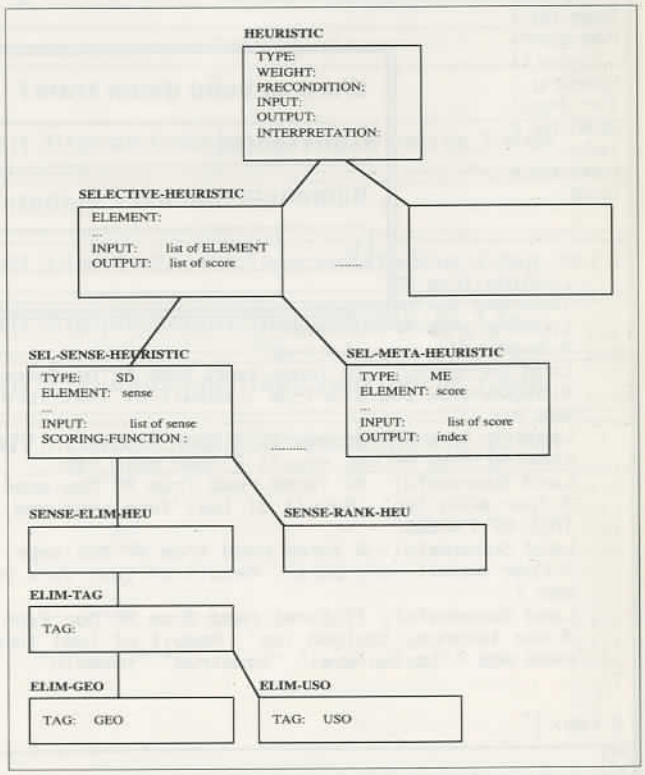


Fig. 12: Vista parcial de la estructura de heurísticos.

La raiz de la jerarquía es el esquema HEURISTIC que define la clase de todos los heurísticos. Bajo esta clase, encontramos SELECTIVE-HEURISTIC que describe la clase de los heurísticos que asignan puntuaciones a una lista de posibles opciones. SEL-SENSE-HEURISTIC por ejemplo trata de seleccionar la acepción adecuada, mientras que SEL-META-HEURISTIC trata de discriminar entre las puntuaciones proporcionadas por un conjunto de heurísticos. En la figura 12 aparecen un par de subclases de heurísticos de selección de acepciones: SENSE-ELIM-HEU, que permite la eliminación de las acepciones y SENSE-RANK-HEU, que les asigna una puntuación.

SENSE-RANK-HEU asigna una puntuación a la lista completa de acepciones. Subclases de éste son LINEAR-RANK- HEU, que asigna la puntuación en forma lineal decreciente, desde 1 a la primera acepción hasta 0 a la última; FIRST-ONLY-HEU que selecciona sistemáticamente la primera acepción, puntuando con 0 a las restantes, o SEL-PATTERN-HEU que se describe en la próxima sección.

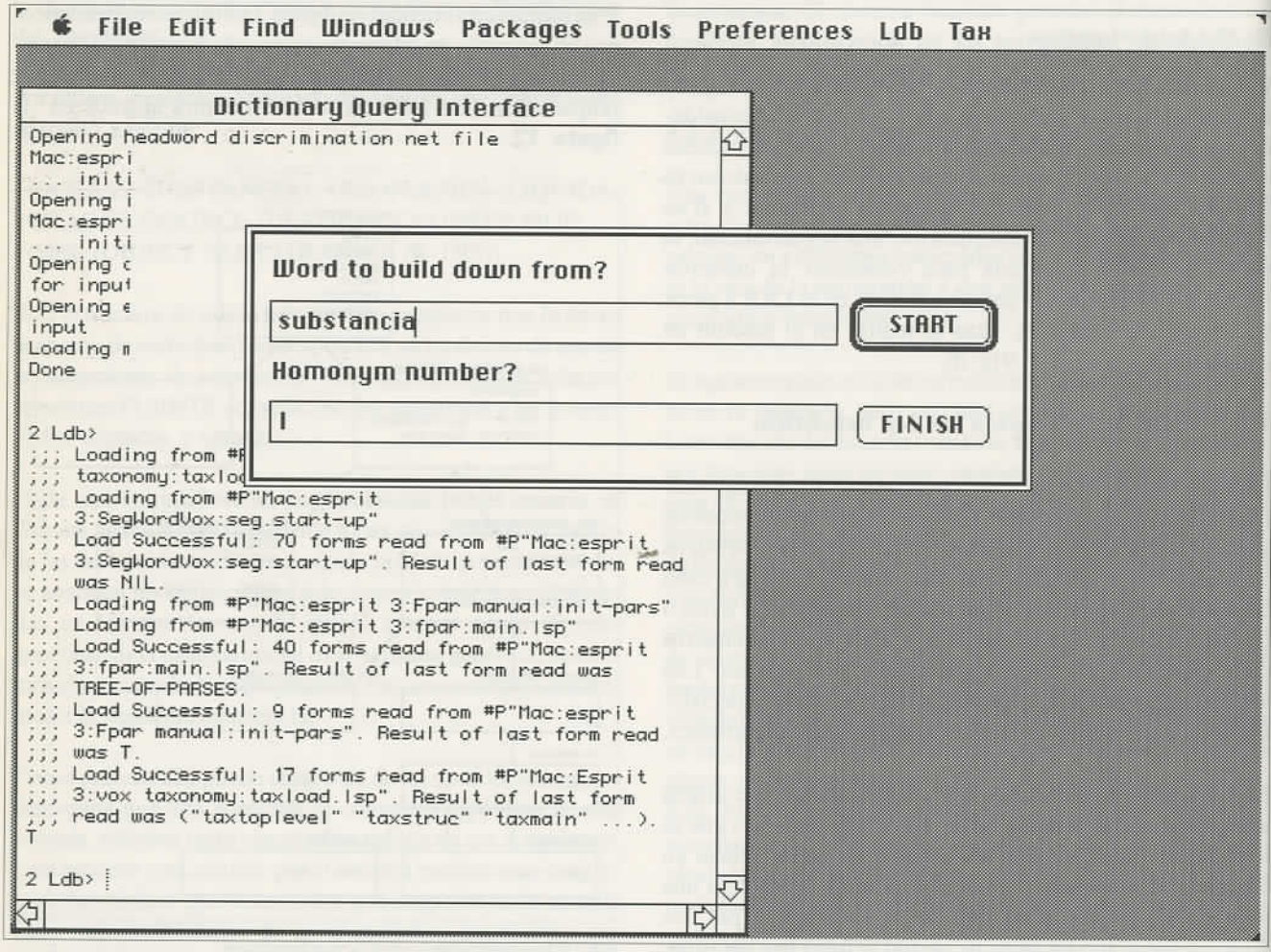
Se pueden construir colecciones de instancias de heurísticos, para su utilización durante el proceso de extracción.

4. Una aproximación para extraer taxonomías del diccionario Vox

En esta sección, mostramos la aplicación del proceso y entorno descritos en la sección anterior a la base de datos léxica del Vox para la extracción de la información semántica de sus acepciones. Como un breve ejemplo del funcionamiento del constructor de taxonomías, presentaremos gráficamente una sesión que muestra la construcción de un parte de la taxonomía de«substancia». Los pasos a seguir son los siguientes:

a) La selección de las posibles palabras raíz.
b) Amalgamación de acepciones en la entrada raíz.
c) Búsqueda de todas las ocurrencias de la entrada raíz en las definiciones de la base de datos léxica.
d) Búsqueda y desambigüación del término genérico.
e) Repetición recursiva de c) y d).
f) Fin del proceso.
g) Obtención de los resultados.

a) Seleción de los posibles puntos de partida (raíces) para construir la taxonomía (ver [Acquilex 90] para clarificar los criterios seguidos). En el ejemplo empezaremos con la raíz «substancia»



Pantalla2

b) El sistema permite al usuario trabajar a nivel de definición o a nivel de palabra. Esto significa que podemos construir la estructura taxonómica de una acepción específica o de todas las acepciones (o algunas de ellas). Después de introducir la entrada raíz «substancia», como se puede ver en la **pantalla 2**, la información general sobre la taxonomía aparece ante el usuario para que valide los valores por defecto (ver **pantalla 3**). En este punto, el usuario también debe seleccionar la gramática y la colección de heurísticos que han ser aplicados.

Otra posibilidad en este punto es la fusión de dos o más acepciones o bien la eliminación de algunas de ellas con el objetivo de reducir el número total de acepciones cuando las diferencias entre ellas no son relevantes (amalgamación) o cuando las acepciones están caracterizadas con algún tipo de etiqueta, como GEO (geografía) o USO (uso). En nuestro ejemplo, la entrada «substancia» consta de 9 acepciones.Ha sido aplicado el heurístico

AMALGAMATE-HEU: "si el número de palabras no funcionales que pertenecen a la intersección de dos acepciones y que no aparecen en las otras acepciones, es mayor que una cantidad predefinida y la categoría es la misma, entonces propón la fusión de las dos acepciones".

De este modo, fusionamos las acepciones 4, 5, 6 de «substancia» debido a la presencia de formas flexivas y derivativas de «alimento» y «nutrir».

acepción:4**f.**Cosa con que otra se alimenta y nutre y sin la cual se acaba.
acepción:5**f.** Parte nutritiva de los alimentos.
acepción:6**f.** Jugo que se extrae de ciertas materias alimenticias.

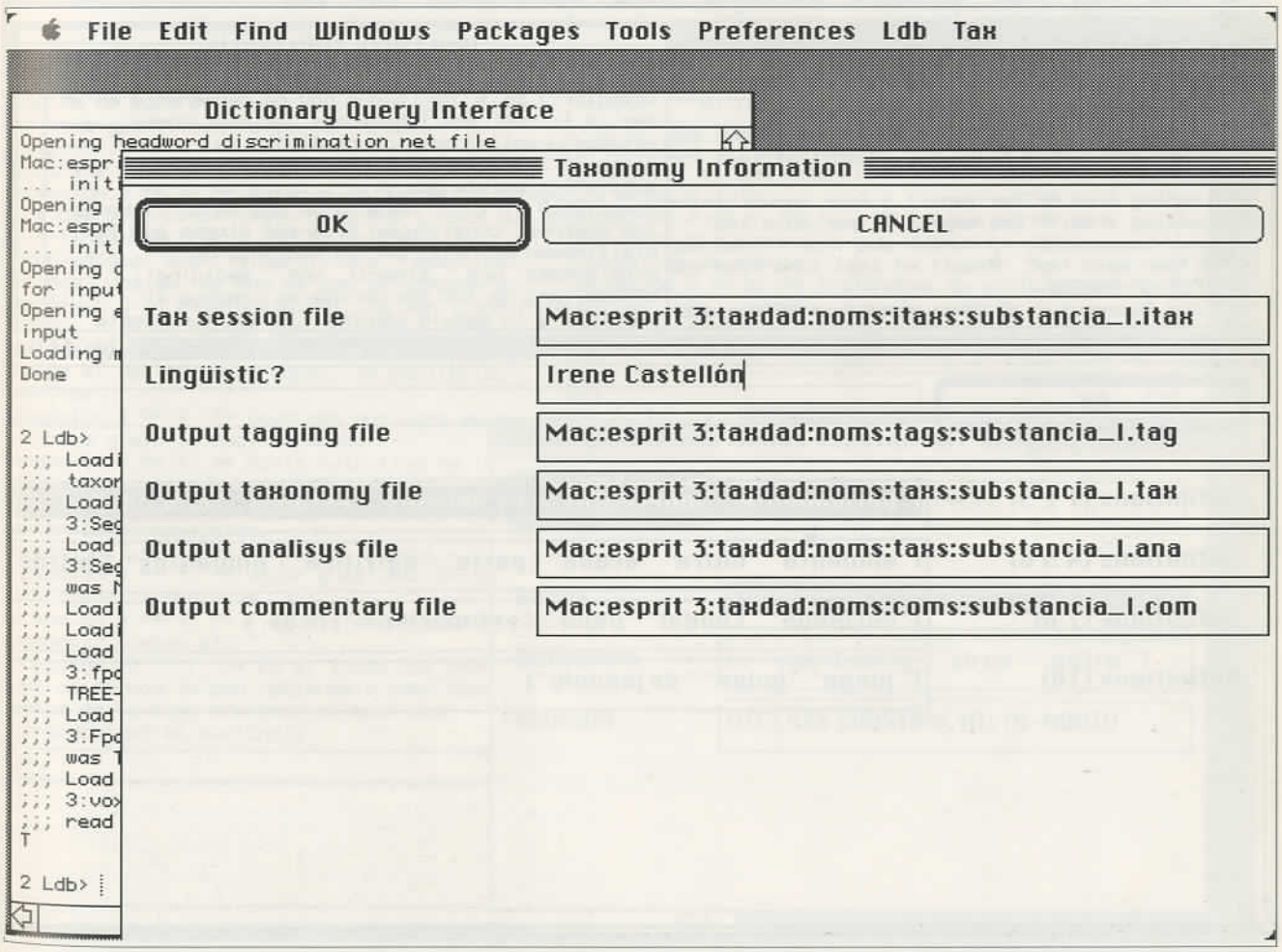
Igualmente, se aplica el heurístico

SENSE_ELIM_HEU: "descarta las definiciones donde aparezcan las etiquetas SEM, REG, USO o GEO".

Por lo tanto, la acepción 9 de «substancia» es descartada.

acepción:9 ** f. ** fig. ** fam. ** Juicio, madurez: hombre sin ~.

Para la posterior aplicación de los heurísticos de desambiguación, extraemos las palabras más significativas de cada acepción pudiendo el usuario validarlas y/o modificarlas (ver **pantalla 4**).



Pantalla 3

c) El siguiente paso consiste en buscar en las definiciones del diccionario (mediante la LDB) todas las ocurrencias de la palabra raíz (y quizá, de algunas formas flexivas o derivativas). Se obtendrá entonces una lista de posibles acepciones hipónimas.

d) Para cada acepción hay dos problemas a resolver: ¿Está actuando la palabra raíz en esta acepción como genérico?; y el problema de desambiguación de las acepciones.

Como resultado del analizador FPar pueden darse varios casos:

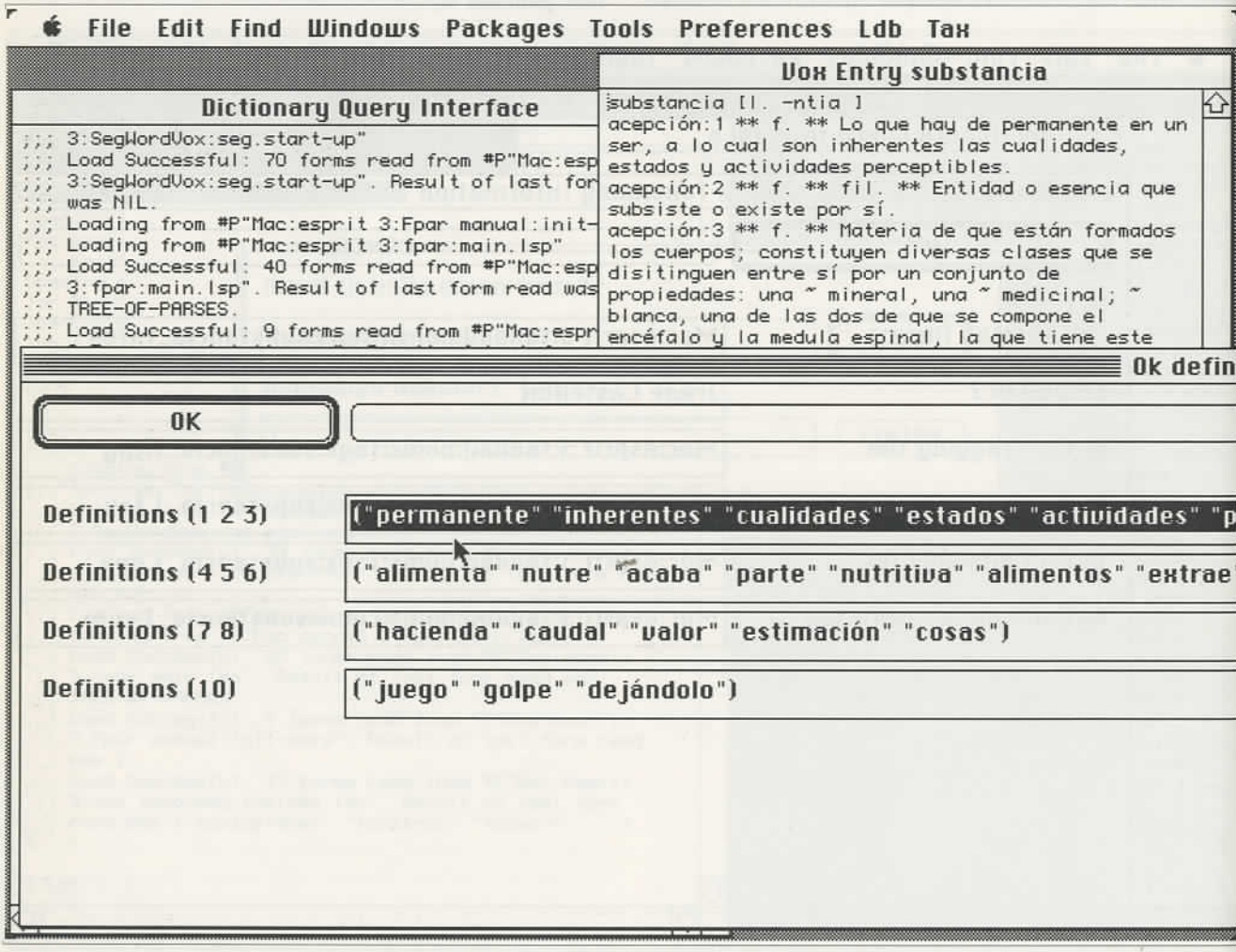
- la palabra ha sido seleccionada correctamente como genérico;
- la palabra ha sido descartada correctamente como genérico;
- la actuación del analizador no ha sido correcta debido a: la falta de categorización de algunas palabras que aparecen en la acepción; o a incompletitud de la gramática usada por el analizador.

La última posibilidad de funcionamento defectuoso del analizador implica algún tipo de modificación interactiva en el léxico, en las reglas morfológicas o en la gramática. En la **pantalla 5**, podemos ver una de las acepciones de alimento (1), con el resultado de la aplicación de la extracción de su genérico, así como las palabras mas significativas de la acepción. En este punto, si el usuario considera que alimento (1) es verdaderamente un hipónimo de "substancia", debe confirmar el juicio del sistema, modificando, si es necesario, la lista de palabras más significativas y el resultado del análisis del FPar.

Una vez que el término genérico ha sido seleccionado, se aplican diferentes tipos de heurísticos para realizar la desambigüación de las acepciones. En nuestro ejemplo, se aplica el siguiente heurístico:

SEL-PATTERN-HEU: «favorece la mayor intersección entre las palabras no funcionales que aparecen en la acepción actual y las palabras que aparecen en las acepciones de la entrada hiperónima».

Por esto, alimento (1) es propuesto como el hipónimo de substancia (4 5 6) debido a la ocurriencia de «nutrir» en ambas acepciones (vease **pantalla 6**).



Pantalla 4

En este punto, el usuario debe elegir entre el conjunto de posibles acepciones hiperónimas de alimento (1). Estas aparecen ordenadas según los valores que hayan obtenido de los heurísticos aplicados. Así aparece en primer lugar la acepción de «substancia» que tiene mayor posibilidad de ser el auténtico hiperónimo de alimento(1): **substancia (4 5 6).**

e) El proceso entonces repite recursivamente los pasos c) y d) en un recorrido en profundidad, tomando, en nuestro ejemplo, la palabra «alimento» como entrada raíz. Siempre que una entrada raíz tenga posibles hipónimos, se dan al usuario cuatro opciones: continuar la construcción de la taxonomía; podar la rama del árbol que iba a realizarse (dejando la entrada raíz en la estructura taxonómica para una realización posterior); parar el proceso (salvando previamente el estado actual); o simplemente salvar la situación actual (volviendo al mismo punto).

f) El proceso para automáticamente cuando no quedan ocurrencias por tratar. En general otros heurísticos podrían decidir cuando debe pararse la construcción de una taxonomía.

g) El proceso genera cinco ficheros de salida diferentes que contienen:

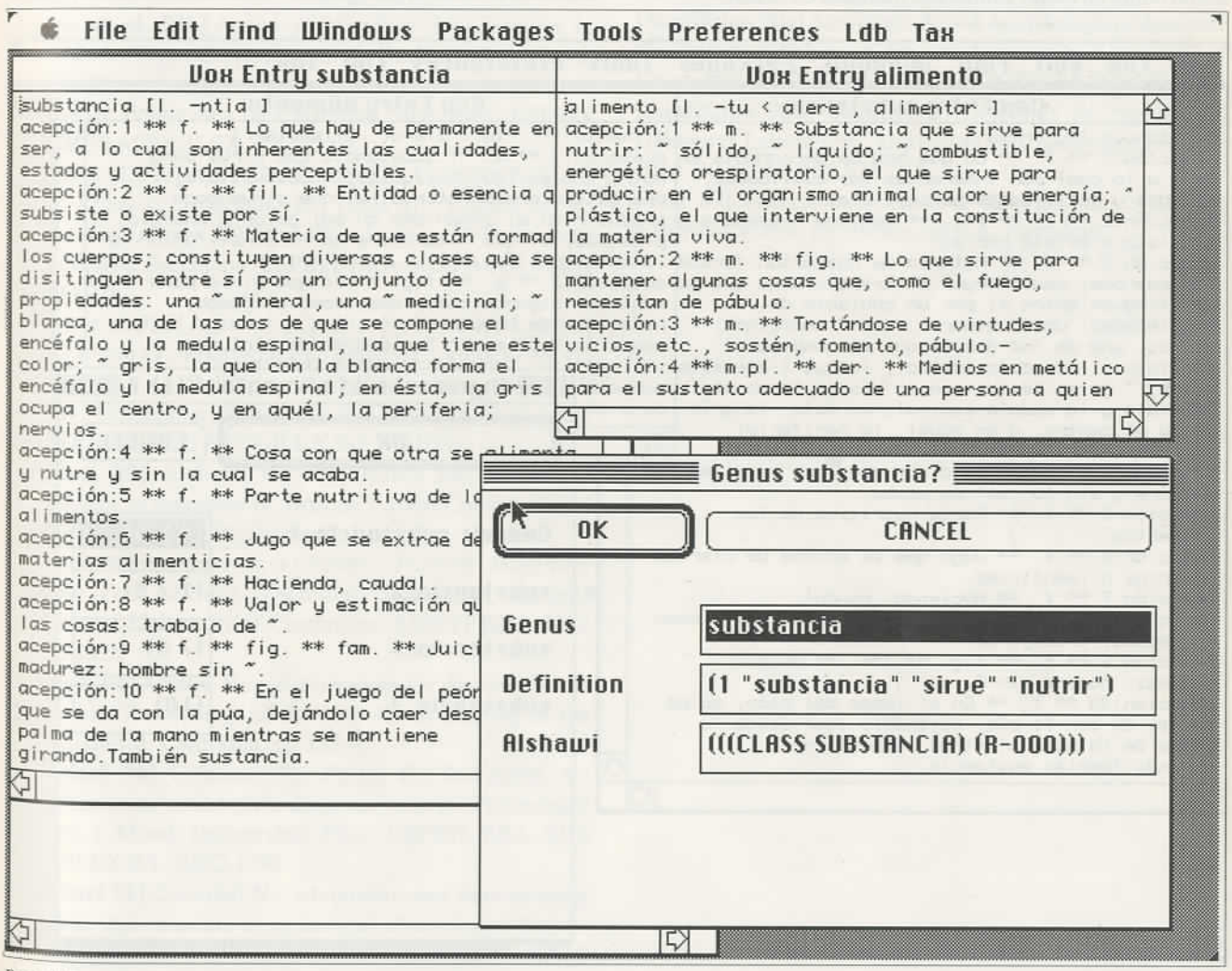
- *Información general* (para uso interno).

- *La estructura de la taxonomía* (también para uso interno).

- *El último análisis realizado* (en texto):

```
substancia I (1 2 3)
    ...
substancia I (4 5 6)
    ...
    alimento I (1)
    (((CLASS SUBSTANCIA) (R-000))) T

        ahumado I (3)
        (((CLASS ALIMENTO) (R-000))) T
        ...
    ...
substancia I (7 8)
```



Pantalla 5

...
substancia I (10)
...

- *El esqueleto de la taxonomía* (también en texto):

substancia I (1 2 3)
...
substancia I (4 5 6)
...
alimento I (1)
ahumado I (3)
...
...
substancia I (7 8)
...
substancia I (10)
...

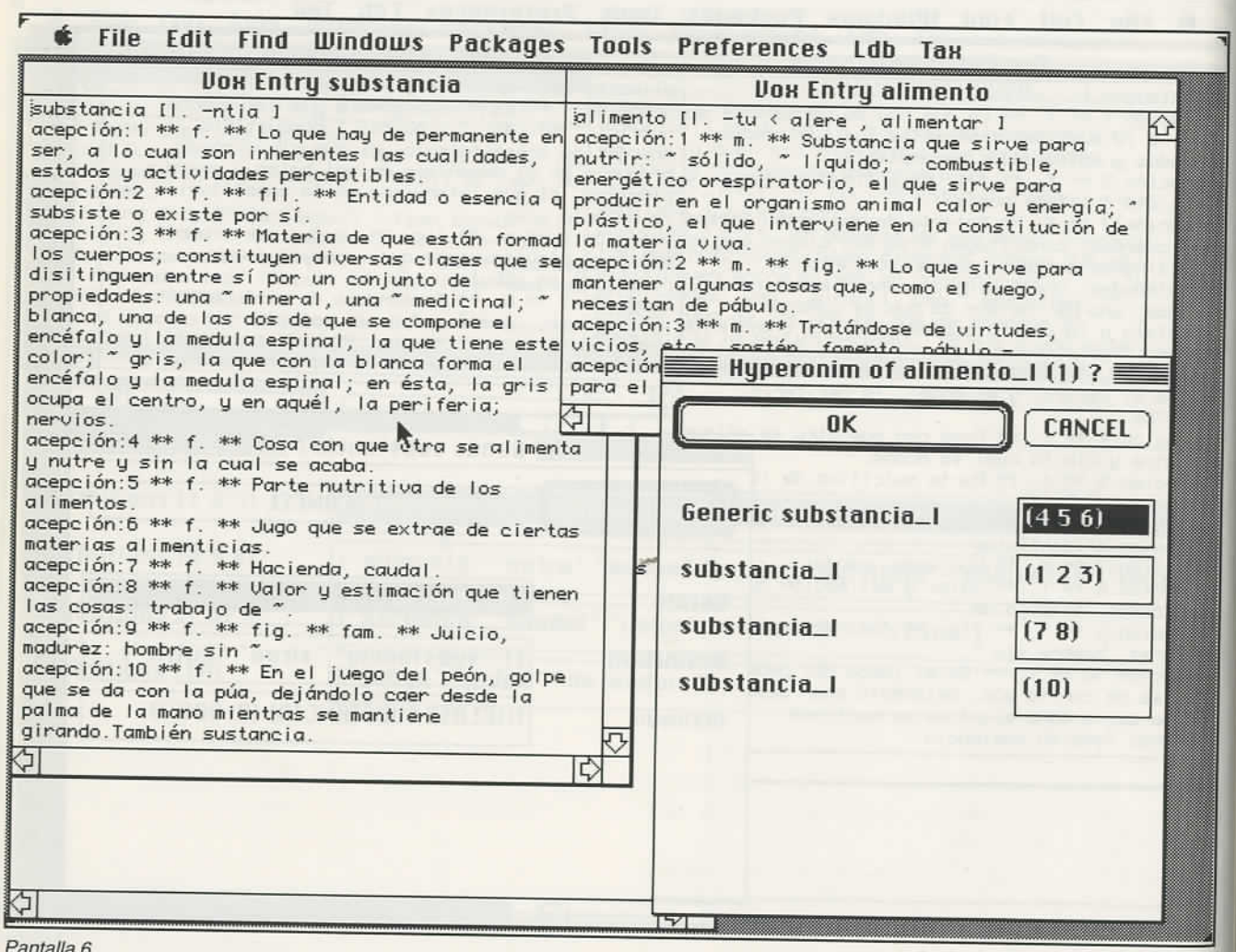
La información que ha sido recogida tanto en modo adquisición como en modo validación (también en texto).

5. Conclusiones

El Sistema ha sido implementado sobre un Mac-IIci utilizando Procyon Common Lisp. Combina modos de ejecución interactivos y batch, garantiza un alto grado de automatismo en su proceso y proporciona al usuario un marco amigable de desarrollo en el que toda la información necesaria para tomar decisiones es fácilmente accesible.

En paralelo se han desarrollado las fuentes de conocimiento lingüístico necesarias (Lexicón, conjunto de reglas para el Analizador Seg-Word, gramáticas para el analizador Fpar). Se han definido conjuntos iniciales de heurísticos para las tareas de desambigüación y fusión de acepciones. Asimismo se ha creado un sistema de conversión (CRS Convertion Rule System) que mediante un conjunto de reglas de conversión mapea la información extraida de las definiciones a la LKB (tarea 4). Actualmente, estamos trabajando en la tarea 5, enlazando entradas léxicas de distintas lenguas de forma semiautomática.

Los resultados son, hasta el momento, alentadores. Se ha utilizado el Sistema para generar varias taxonomías de tamaño



Pantalla 6

medio/alto tanto nominales como verbales («substancia», con 1200 entradas, «alimento» con 146 o «bebida» con 260) y de cierta complejidad (la profundidad de la taxonomía cuya raiz es «bebida» es de 5 niveles). El índice de acierto en las extracciones automáticas ha sido de un 95%. Ello supone un ahorro indudable frente a la alternativa que supondría un proceso de extracción manual.

Referencias

[Acquilex89] Acquilex.: *«Technical Annex»*. ESPRIT BRA-3030 ACQUILEX

[Acquilex90] Acquilex.: *«Initial definition of the vocabulary subset»*. Preliminary Report. 12 MonthDeliverable. Amsterdam. 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.

[Carroll 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

[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.