ANALISIS DE DEFINICIONES DEL DICCIONARIO VOX

Alicia Ageno; Sivar Cardoce (*); Irene Castellón (*); Ma. Antònia Martí (**); German Rigau; Horacio Rodríguez (*); Mariona Taulé (**); Felisa Verdejo (*)

(*) Universitat Politècnica de Catalunya
(**) Universitat de Barcelona

1.- Introducción

El presente trabajo se enmarca en el área de la lexicografía computacional, en concreto, en el desarrollo de sistemas de extracción de información léxica a partir de diccionarios en soporte magnético (Machine Readable Dictionaries, MRDs). Dicho trabajo se ha desarrollado en el marco del proyecto Esprit Acquilex[1]. La labor lexicográfica se ha realizado sobre diversos diccionarios monolingües y bilingües[2].

En nuestro proyecto, la extracción de información léxica a partir de MRDs tiene como objetivo último la representación de dicha información en términos de una Base de Conocimiento Léxico (BCL). Este proceso es complejo y se realiza de manera interactiva y semiautomática. Para ello se ha diseñado un Sistema de Información para la Ayuda a la Toma de Decisiones Lexicográfica, SIATDL, (Agen-91a) que se aplica al material lexicográfico para la consecución de los objetivos trazados.

Las tareas más relevantes de este proceso de extracción de información léxica son las siguientes: a)- El volcado del MRD en una estructura de base de datos, la Base de Datos Léxica (BDL). b)- El análisis de la información contenida en los diferentes campos de la BDL. En este proceso se identifica el término genérico de cada acepción y sus modificadores. c)- La construcción de la taxonomía y la templeta, estructura de información que servirá de puente con la Base de Conocimiento, mediante la aplicación del Sistema de Extracción de Información Semántica de los Diccionarios (SEISD), que forma parte de SIATDL. d)- La traducción de la información extraída (el término genérico y sus modificadores) a su representación en términos de la Base de Conocimiento Léxico.

Para la realización de estos procesos hemos utilizado las herramientas desarrolladas por la Universidad de Cambridge (Sanf-90, Copes-90, Alsh-89) que hemos adaptado a nuestro diccionario y herramientas adicionales construidas por el equipo de la U.P.C.: el SIATDL (Agen-91a,b,c).

Desde un punto de vista lingüístico, la realización de este proceso de transformación de los datos contenidos en un diccionario ha implicado:

---

1 Aquilex es un proyecto Esprit (BRA 3030)en el que participan el Istituto di Linguistica Computazionale di Pisa y las universidades de Cambridge, Amsterdam, Dublín y la Universitat Politècnica de Catalunya. La labor lexicográfica se ha realizado sobre los diccionarios Longman (inglés), VanDale (holandés), Garzanti (italiano) y VOX (español) . Se utilizan también diversos diccionarios bilingües.

2 Queremos expresar nuestro agradecimiento a la Editorial Biblograf por la cesión de la versión en cinta magnética del Diccionario General Ilustrado de la Lengua Española Vox, sobre el cual estamos trabajando.

a)- La definición de la estructura del diccionario Vox y la elaboración de la gramática correspondiente, utilizada en el proceso de carga del diccionario a la Base de Datos Léxica (Cast-90 y Cast-91).

b)- La construcción de la gramática de la definición de cada una de las acepciones y la especificación de la estructura de la información resultante del proceso de análisis.

c)- La construcción de la taxonomía y la templeta, estructura de información que servirá de puente con la Base de Conocimiento, mediante la aplicación del Sistema de Extracción de Información Semántica de los Diccionarios (SEISD) (Agen-91a).

d)- La definición de la estructura de tipos de la base de conocimiento y la aplicación del programa Convert (Agen-91b,c) para proceder a la traducción de las estructuras de datos obtenidas mediante los procesos anteriores a la Base de Conocimiento.

En este artículo presentamos la metodología general del proceso de análisis y extracción de información de las definiciones. Se tratan con detalle las gramáticas que permiten el análisis de las definiciones del corpus seleccionado, correspondiente a subconjuntos de substantivos y verbos. A continuación se evalúa la metodología utilizada respecto de otras aproximaciones y, finalmente, se plantean las líneas de investigación futuras a partir del trabajo realizado.

2.- Metodología

El objetivo fundamental del análisis de las definiciones es la extracción de la información que contienen (el término genérico o nuclear y los modificadores) y la construcción de la estructura taxonómica. Entendemos por taxonomía la clasificación jerárquica de los sentidos del diccionario conectados entre sí por la relación ES-UN (Ams-81).

Dado el volumen del diccionario y la complejidad y diversidad de la información que contiene, se decidió centrar el proceso de extracción de información léxica en el análisis de las definiciones. El campo definición, a su vez, no es homogéneo ya que incluye, además de la definición, ejemplos, información contextual y restricciones de combinatoria léxica.

Por estas razones se optó por seguir una aproximación incremental y descendente en el proceso de extracción de información léxica, tanto en el diseño de las gramáticas como en la construcción de la taxonomía.

Así, se han diseñado gramáticas para la extracción del término genérico y otras para la extracción del resto de la información contenida en las definiciones. En función de los resultados obtenidos, dichas gramáticas se pueden ir mejorando hasta conseguir los objetivos previstos.

El análisis de las definiciones se realiza de manera incremental y, como ya se ha dicho, por subconjuntos temáticos: a partir de una primera gramática que tan sólo extrae el término genérico, se van construyendo sucesivas gramáticas, cada vez más refinadas hasta conseguir analizar la totalidad o la mayor parte de la definición.

Por otra parte, la construcción de la taxonomía se realiza mediante un proceso descendente a partir de un concepto inicial que funciona como cabeza de la taxonomía. Se trata de la misma estrategia propuesta por Amsler (Ams-81), Chodorow, Bird y Heidorn (Chod-85).

Se parte de la hipótesis de que todos aquellos conceptos definidos a partir del mismo término genérico y pertenecientes a una determinada categoría sintáctica comparten una serie de propiedades que los caracterizan y una estructura de información común. Esta estrategia conlleva que el diseño de las gramáticas se realice teniendo en cuenta el campo semántico que analizan ('substancia', 'persona', 'instrumento', etc.).

La elección de un proceso descendente (*top-down*) viene determinada, en parte, por las propias características del diccionario. La falta de un vocabulario limitado, el volumen de las acepciones, así como la gran variedad de estructuras que en ellas aparecen, puso de manifiesto

la imposibilidad de crear, a corto plazo, una gramática única para el análisis de este campo. En un proceso ascendente (*bottom-up*) resultaría imprescindible realizar un análisis previo de todas las definiciones o bien disponer de una gramática global que permitiera analizar cualquier definición, ya que, en este tipo de estrategia, no se pueden establecer ámbitos temáticos ni categoriales y se trabaja en todo momento con la totalidad del diccionario.

Por otra parte, el proceso ascendente no admitiría que la taxonomía se construyera de manera semiautomática ya que el lingüista no podría controlar la diversidad de niveles taxonómicos que se activarían durante el proceso; ello implicaría que el sistema tuviera incorporado todo el conocimiento para la desambiguación y la unificación de acepciones, lo cual no era factible en nuestro caso. Otros grupos del proyecto tiene previsto ha seguido esta estrategia por tener ya realizado el análisis completo del términos genéricos de las definiciones.

En esta primera fase de nuestro proyecto, de entre las diferentes relaciones semánticas que se pueden establecer entre las acepciones del diccionario, se ha acordado limitar el estudio a la relación ES-UN; más adelante está previsto tratar otras relaciones como Parte-de, Conjunto-de, etc. Existen trabajos ya realizados en este sentido como es el proyecto Links llevado a cabo por el equipo de la Universidad de Amsterdam (Voss-90). En el marco de esta investigación se realizó un análisis completo de las definiciones del diccionario Longman con el fin de hacer accesible su contenido de manera automática. El proyecto incluía un estudio completo de los diferentes tipos de relaciones que aparecen en las diferentes acepciones del diccionario.

2.1.- Definición de subconjuntos del diccionario

La definición del corpus es el primer requisito para llevar a cabo la metodología que hemos adoptado. A continuación exponemos el modo de extracción de los subconjuntos de diccionario con los que se trabaja.

Cada subconjunto del diccionario se define por una determinada categoría gramatical (de momento nos hemos limitado a nombres y verbos) y por tener una determinada forma en su definición. Un ejemplo de subconjunto sería el constituido por todas las acepciones de categoría nombre cuya definición contiene la forma 'substancia'. Sólo una vez aplicado el programa SEISD, se descartarán las acepciones en las que 'substancia' no aparece como genérico y las que hayan dado un resultado positivo se asociarán a la acepción de 'substancia' que les corresponda.

La selección del corpus se realiza de modo automático mediante preguntas a la Base de Datos Léxica (LDB) (Cast-90). Las preguntas que se formulan para conseguir aislar todas las acepciones que tengan como valor nombre o verbo en el campo categoría , son del siguiente tipo:

p.e.:

para obtener todos los substantivos:

```
[[SIN
  [CA (OR s.pl. s. m.pl. m. f.pl. f. adj.-s. adj.-m. adj.-f.)]]]
```

para obtener todos los verbos:

```
[[SIN
 [CA (OR tr.-prnl. tr.-intr.-prnl. tr. rec. prnl. pp.irreg. p.p. intr.-tr. intr.-prnl. intr. impers.)]]]
```

A estas condiciones se le aplican otras que conducirán a la delimitación del ámbito semántico, como es la presencia de una determinada forma en la definición:

p.e.:

para obtener todas las acepciones que contienen la forma 'substancia' en la definición se realiza la búsqueda:

[[**SEM**
[**DEF** substancia]]]

Así, la demanda completa sería:

[[**SEM**
[**DEF** substancia]]
[**SIN**
[**CA** (OR s.pl. s. m.pl. m. f.pl. f. adj.-s. adj.-m. adj.-f.)]]]

De igual modo, esta metodología se aplica a las definiciones verbales. De momento hemos seleccionado los verbos de movimiento (ir, andar, mover, etc.) y los verbos relacionados con 'comida' y 'bebida' (comer, beber, guisar, asar, etc.).

3.- Análisis de las definiciones

3.1.- Objetivo del análisis

El objetivo principal del análisis de las definiciones es la extracción de la información semántica contenida en las acepciones de una entrada. Una vez identificada y etiquetada esta información, el paso siguiente es la construcción de taxonomías y la definición de templetas o estructuras argumentales para su posterior proyección a la Base de Conocimiento Léxico (BCL).

La extracción de información se realiza en dos fases :

1.- En primer lugar, la obtención del término genérico , es decir, la palabra núcleo de la definición que mantiene una relación de inclusión (*del tipo es-un*) con la entrada definida.

p.e.:

(i)"aguardiente (1) : **bebida** *alcohólica* que por *destilación* se obtiene del *vino.*"

En esta definición el término genérico es el sustantivo 'bebida' que establece una relación de inclusión con aguardiente, es decir, "aguardiente (1) -es-una- bebida (3)" .

(ii) "circular (1): **andar o moverse** *en derredor.*"

En esta definición el término genérico es el verbo 'andar(1)' que establece una relación con 'circular(1)' del tipo 'es-un' (circular es una manera de andar).

2.- El siguiente paso es la obtención de las propiedades que modifican al término genérico de la definición (los modificadores o 'differentia') explicitando al máximo su naturaleza. Esta información rellenará las templetas en el caso de los substantivos y la estructura argumental en el caso de los verbos.

p.e.:
En el ejemplo (i), los modificadores de 'bebida' son el adjetivo 'alcohólica' y las expresiones 'se obtiene del vino' y 'por destilación'. El resultado del análisis es:

properties: *alcohólica*
source: *vino*

means: *por destilación*

En el ejemplo (ii), el modificador de 'andar' es la locución 'en derredor'. La salida del análisis es:

- locativo: *en derredor*.

3.2.- Analizadores

El proceso de extracción de información de las definiciones se ha realizado mediante la aplicación del analizador morfológico *Seg-Word* de A. Sanfilippo (Sanf-90) y del analizador sintáctico-semántico *FPar* de H. Alshawi (Alsh-89) desarrollados en la U. de Cambridge. El objetivo de la aplicación de estos analizadores es la extracción de la máxima información semántica posible contenida en las definiciones. Hasta el momento se ha trabajado sobre las definiciones de los substantivos correspondientes al ámbito semántico de 'substancia', 'persona', 'instrumento' y 'lugar' y de los verbos de movimiento.

Seg-Word es una analizador morfológico que tiene como característica fundamental usar el propio diccionario como fuente básica de información para el análisis de las formas flexivas y derivadas. Mediante un diccionario de sufijos y unas reglas morfológicas para la identificación del lema, consulta su categoría correspondiente en el diccionario y se la asigna. En caso de ambigüedad, se obtiene más de un resultado que se resolverá en el análisis posterior.

El FPar es un analizador "pattern-based" que utiliza una gramática en forma de jerarquía de esquemas ("patterns") y lleva a cabo un análisis descendente aplicando patrones de definición cada vez más específicos a medida que los niveles superiores (más generales) van verificándose. Este procedimiento permite proporcionar un análisis parcial cuando no es 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 esquemas a aquellos casos donde sea más probable un resultado positivo.

En la estructura de salida que proporciona este analizador la información viene segmentada y etiquetada sintáctica y/o semánticamente. El análisis está orientado a la extracción de toda la información semántica posible, sin embargo nuestro análisis puede también proporcionar resultados puramente sintácticos, sobre todo en casos difíciles de desambiguar como son los argumentos de un verbo.

Las unidades con las que trabaja el analizador son de diversos tipos: categorías sintácticas (nombre, verbo...), esquemas recurrentes ('que sirve para', 'compuesto de' ...) y diversos operadores que permiten declarar un nodo de la gramática como opcional u obligatorio, ya sea terminal (categoría sintáctica) o bien no terminal (regla de contexto libre):

+	indica que la categoría debe aparecer obligatoriamente una vez.
+0	indica que la categoría puede aparecer opcionalmente una vez o ninguna.
&	indica que la categoría debe aparecer obligatoriamente una vez o más de una.
&0	indica que la categoría puede aparecer opcionalmente una vez , más de una o ninguna.
*	indica que el nodo es una regla de contexto libre.
*0	indica que el nodo es una regla de contexto libre opcional.

&& permite tratar una cadena de texto que, en principio, no interesa analizar.

3.3.- Gramática para las definiciones de substantivos.

Como se ha indicado anteriormente, se han definido diversas gramáticas para el análisis de los substantivos. Por el momento, éstas son la gramática de *substancia*, de *lugar*, de *persona* y de *instrumento*.

Las entradas con categoría nombre se definen, en su mayoría, mediante una frase nominal, es decir un nombre introducido opcionalmente por un determinante y acompañado por sus complementos (grupos adjetivales, grupos preposicionales o frases de relativo).

Existen dos fases en la construcción de la gramática de las definiciones: la extracción del término genérico y la de los modificadores.

3.3.1.- Extracción del término genérico

El término genérico de las entradas de categoría sustantivo puede aparecer de diversos modos:

a.- El término genérico es el primer nombre que aparece en la definición precedido opcionalmente de un determinante.

p.e.:
espíritu: 5 ***alma*** individual esp. la de un muerto ...

Este término genérico puede ser simple o puede aparecer coordinado con otro genérico en cuyo caso la entrada afectada tendrá dos hiperónimos. Siguiendo con el ejemplo anterior, tenemos:

ánimo: 1 ***alma o espíritu*** en cuanto es principio de la actividad humana.

La relación establecida en todos estos casos es la de inclusión (es-un) por lo que estos ejemplos producirían una estructura del tipo:

espíritu **es-un** alma
ánimo **es-un** alma
ánimo **es-un** espíritu

También pueden aparecer dos términos genéricos yuxtapuestos en cuyo caso el primero suele mantener una relación de sinonimia con la entrada y el segundo una relación de hiperonimía.

p.e.:
matarratas (1) : ***raticida, substancia*** para matar ratas

b.- El término genérico viene introducido por un grupo nominal o preposicional que actúa de **especificador** o de **relator**.

i) Especificador se trata del grupo nominal o preposicional que antecede al elemento nuclear de la definición. Aporta información sobre el contexto temático o (al que se asigna el código Ilex como resultado del análisis) o léxico (Collocation) de la entrada.

i.1) "Ilex" agrupa toda aquella información referente al ámbito temático o geográfico al que pertenece la entrada, p.e.: "entre mineros" o " en el lenguaje de la droga" . Se trata de un tipo de información que suele aparecer codificada en el diccionario mediante marcadores como BIOL., DER., etc.

p.e.:
canuto (1.5) ***En el lenguaje de la droga*** , porro

En esta definición se especifica el uso de la entrada en un contexto determinado: *en el lenguaje de la droga.*

i.2) "Collocation" expresa restricciones de coocurrencia léxica regular, es decir describe conceptos que se expresan mediante dos o más palabras.

p.e.:
'nitrato(2) ~ ***de chile***, abono nitrogenado natural...,

En el ejemplo se indica no un sentido de 'nitrato' sino el sentido de la construcción 'nitrato de chile'.

Las estructuras que acompañan a la entrada en estos casos pueden ser de diversos tipos, con claro predominio de los adjetivos y grupos preposicionales.

ii) Un **relator** es aquel grupo nominal que expresa un tipo de relación entre la entrada y su término genérico,

p.e.:
cedreno : 1 ***parte*** líquida de la esencia del cedro,

esta definición establece la relación siguiente:

cedreno **part-of** esencia del cedro

Por ahora los diferentes relatores observados son:

ii.1) Relación 'es-un' expresada mediante los esquemas: 'variedad de' , 'tipo de', etc. que expresan la relación de hiponimia. Como resultado del análisis, las entradas con este tipo de relator se les asigna el código TYPE GENERIC.

ii.2) Relación 'conjunto-de' (TYPE COLLECTIVE) expresada mediante los esquemas: 'conjunto de', 'hato de', etc.

ii.3) relación 'parte-de' (TYPE PART) expresada mediante los esquema: ' 'parte de' ,'trozo de' , 'pedazo de', etc.

La gramática de substantivos detecta este tipo de relatores, sin embargo hasta el momento nuestro estudio se ha centrado únicamente en las relaciones del tipo 'es-un' (hiponimia) que se realiza léxicamente mediante un relator de tipo (ii.1) o bien por ausencia de relator .

3.3.1.1.- Las reglas

Las reglas principales de este analizador son dependientes del contexto y están apoyadas por otro tipo de reglas de contexto libre.

Las reglas que se ocupan de la extracción de término genérico son las siguientes:

(**n**- (n &&) n-000)

esta regla condiciona que la definición que se analiza tenga como categoría nombre.

La regla n-000 analiza un grupo nominal del que sólo se especifica el determinante y el nombre y la regla n-100 analiza un grupo nominal introducido por un especificador o relator:

```
(n-000 (n +0det && +noun  &&)  n-100)

(n-100
  (n *especi && +0det && +noun  &&))
```

Estos especificadores vienen introducidos por reglas independientes del contexto de las cuales exponemos algunos ejemplos.

```
(sub-pats (QUOTE (*especi))
    (QUOTE
        ((**especi1 (*esp1))
        (**especi2 (*esp2))
           (**especi3 (*esp3))
        (**especi4 (*esp4))
        (**especi5 (*esp5)))))
```

Esta regla asocia al nodo *especi diferentes tipos de relatores. El relator 'generic' que recoge todos los modos de expresar la relación 'es-un', viene introducido por la siguiente regla:

```
(sub-pats (QUOTE (*esp1))
    (QUOTE
        ((**e11 (cierto)
           (**e12 (cierta)
           (**e13 (+esp1 && &adj && prepo))
           (**e14 (+esp1 && prepo)))))
```

El segundo relator expresa la relación 'conjunto-de' y se analiza mediante la siguiente regla:

```
(sub-pats (QUOTE (*esp2))
    (QUOTE
        ((**e21 (+esp2 && &adj && prepo))
           (**e22 (+esp2 && prepo)))))
```

La regla *esp3 analiza las relaciones del tipo 'miembro-de':

```
(sub-pats (QUOTE (*esp3))
    (QUOTE
        ((**e31 (cada uno de))
           (**e32 (+esp3 && &adj && prepo)
           (**e33 (+esp3 && prepo)))))
```

Las relaciones del tipo 'parte de' se analizan a partir de la regla siguiente:

```
(sub-pats (QUOTE (*esp4))
    (QUOTE
        ((**e40 (+esp4 &0adj en forma prepo +noun && de))
           (**e41 (+esp4 && &adj && prepo))
           (**e42 (+esp4 && prepo)))))
```

Por último el especificador 'collocation', analiza las restricciones de coocurrencia léxica:

```
(sub-pats (QUOTE (*esp5))
```

```
(QUOTE
    ((**e50 (~ +adj +0coma))
       (**e51 (~ +noun +0coma))
       (**e52 (~ de +0det +noun +0coma))
       (**e53 (~ del +noun +0coma))
       (**e54 (+adj ~ +0coma)))))
```

3.3.2.- Extracción de modificadores

Llamamos modificadores o 'differentia' a los elementos que complementan y describen al término genérico. Estructuralmente se trata de adjetivos, sintagmas preposicionales y frases de relativo que suelen aparecer a la derecha del nombre.

El análisis de los modificadores se enfoca hacia un resultado semántico-sintáctico. Es por ello que se tratan de modo exhaustivo las construcciones sintácticas que son indicativas de un determinado campo semántico (p.e.: 'que sirve para', 'para', 'que se usa para', etc. expresan el campo semántico GOAL). Debido a que cada ámbito conceptual tiene asociados diferentes campos semánticos, se han construido diferentes gramáticas para los distintos ámbitos conceptuales que hemos tratado.

Las diferencias fundamentales de las gramáticas aparecen en el análisis de los modificadores que se expresan mediante reglas independientes del contexto. Así, a partir de reglas comunes para todos los substantivos como n-130

```
(n-130
        (n && +0det && +noun *modif && *post-mod11 *0post-mod2 &&))
```

se introducen las reglas independientes del contexto *modif, *post-mod11 y *0post-mod2. La regla *modif analiza sintagmas adjetivales, grupos preposicionales y frases de relativo, es un nodo común a todas las gramáticas. Las reglas *post-mod11 y *0post-mod2 introducen esquemas de definición.

Es en la descripción de los diversos esquemas de definición donde aparecen los cambios entre las diferentes gramáticas.

p.e.:
*post-mod11 en la gramática de substancia introduce siete tipos de modificadores:

```
(opt-pats (QUOTE (*post-mod11))
    (QUOTE
        ((**p-m1 (*rel1))
           (**p-m2 (*rel2)
           (**p-m3 (*rel3)
           (**p-m4 (*rel4))
           (**p-m5 (*rel5))
           (**p-m6 (*rel6))
           (**p-m7 (*rel7)))))
```

donde rel3 analiza esquemas del siguiente tipo:

```
(sub-pats (QUOTE (*rel3))
    (QUOTE
        ((**r31 (que sirve && para +v +0det +0noun *0modif1))
        (**r32 (que se usa && para +v +0det +0noun *0modif1))
        (**r33 (+0adv +pattern3 +v +0det +0noun *0modif1))
        (**r34 (+0adv +pattern3 +0det +noun *0modif1))
```

```
(**r35 (para +v +0noun))
(**r36 (que se emplea para +v +0noun))
(**r37 (con que && se +v +0noun))
(**r38 (que se toma && para +v +0noun))
(**r39 (que se usa && como +0det +noun *0modif1)))))
```

Estos esquemas indican la funcionalidad (GOAL) de una 'substancia'. Este tipo de 'pattern' no se presentará en la gramática de persona, que hasta el momento únicamente tiene un tipo de esquema que indica la actividad a la que se dedica una persona :

```
(opt-pats (QUOTE (*post-mod11))
    (QUOTE
        ((**p-m2 (*rel88)))))

(sub-pats (QUOTE (*rel88))
    (QUOTE
        ((**r821 (+0adv +pattern8 +0det +noun *0pp-mod))
        (**r822 (que por && se dedica a +0det +noun *0pp-mod))
        (**r823 (que se dedica a +v +0det +0noun))
        (**r824 (que practica +0det +noun *0pp-mod))
        (**r825 (que se dedica a +0det +0noun)))))
```

3.3.3.- Resultados del análisis

Cada regla de la gramática lleva asociada una estructura de salida donde la información obtenida del análisis se codifica sintáctica o semánticamente.

p.e.:

la regla n-130

```
(n-130
    (n && +0det && +noun *modif && *post-mod11 *0post-mod2 &&))
```

tiene asociada la estructura de salida:

```
(n-130
    ((compound-class +noun)
        (pm1 *modif)
        (pm1 *post-mod11)
        (pm1 *0post-mod2)
        (r-130)))
```

Veamos a continuación unos ejemplos del resultado del análisis:

```
(ACRIDINA) (1)
(substancia orgánica sintética que se usa en medicina como colorante antiséptico)

(((CLASS SUBSTANCIA)
(PROPERTIES ORGáNICA SINTéTICA)
(GOAL COLORANTE)))

(ADHERENTE) (1)
(adhesivo, substancia que sirve para unir otras)

(((CLASS ADHESIVO)
(RELATED-TO SUBSTANCIA)
(GOAL UNIR)))
```

(ACACIA)
(substancia medicinal que se extrae de la acacia de egipto o del endrino)

(((CLASS SUBSTANCIA)
(PROPERTIES MEDICINAL)
(SOURCE ACACIA (PREP-MOD (DE (OBJECT EGIPTO))))))

3.4.- Gramática para las definiciones de verbos

En un principio partíamos de la hipótesis de que la definición de una entrada verbal podía proporcionar información acerca de la subcategorización del verbo definido o acerca de su aridad, papeles temáticos, etc. Sin embargo nuestros estudios han confirmado que no se puede inferir de modo inmediato este tipo de información a partir de la definición. Otros campos de la entrada pueden proporcionar información pertinente para ello. De momento, la única información que se puede extraer es de tipo semántico, es decir, sobre el significado que expresa el verbo en cuestión. Veamos un ejemplo:

p.e.:
BEFAR (1):
mover los caballos el befo

En esta definición 'caballos' y 'befo' son argumentos del término genérico 'mover', sin embargo no reflejan la subcategorización sintáctica del verbo 'befar' sino sus argumentos semánticos implícitos, ya que 'befar' no subcategoriza ningún objeto directo por tratarse de un verbo intransitivo.

Antes de presentar la gramática que analiza el contenido de estas definiciones, es necesario describir como éstas se estructuran.

Se pueden contemplar tres partes básicas en la definición de una determinada entrada léxica verbal: el campo denominado "**Ilex**", el **término genérico** y a continuación las "**differentia**" o complementos del término genérico. Así como el campo Ilex y las "differentia" pueden o no encontrarse en la definición, son opcionales, la presencia del término genérico siempre es obligatoria. Veamos ahora de forma más detallada como se presentan cada una de estas partes de la definición.

3.4.1.- "Ilex"

Este código, que puede parafrasearse como 'información léxica', siempre se realiza como un sintagma preposicional, normalmente introducido por la preposición "*en*", a la izquierda del término genérico y encabezando la definición. La función de este sintagma preposicional es la de definir el contexto de la acepción que se define. En ningún momento debe ser interpretado como un complemento del término genérico. Este tipo de información es equivalente a la que se encuentra en otros campos de la entrada como Tema, Geo o Uso.

p.e.:
ENROCAR:
en el juego de ajedrez, mover en una misma jugada el rey y un roque bajo condiciones prescritas.

En este ejemplo el sintagma preposicional '*en el juego de ajedrez*' será tratado como Ilex porque aporta información sobre el uso del verbo en ese contexto determinado. El verbo adquiere un significado especial en ese contexto.

3.4.2.- El término genérico

En el caso de las entradas léxicas verbales el término genérico adquiere otra connotación, no se refiere a una relación de inclusión (relación ES-UN) con respecto a la entrada definida, sino que hace referencia al tipo de acción a la cual está asociado el verbo.

El término genérico es el elemento nuclear de la definición y siempre es un verbo en infinitivo. Este hecho condiciona absolutamente el tipo de oración que se analiza, es decir, la definición de una entrada léxica verbal se caracteriza por la presencia de una oración de infinitivo (o simplemente por un infinitivo).

El término genérico puede ser simple o bien compuesto por más de un verbo en infinitivo, ya sea por la presencia de una conjunción o por la presencia de una "coma". La coordinación (copulativa "*y*") de dos términos genéricos indica básicamente simultaneidad de acciones, mientras que la yuxtaposición (",") y coordinación distributiva de ellos ("*o*") indica sentidos aproximados

p.e.:

QUEBRANTAR (8):
persuadir, mover.

APALANCAR (1):
levantar, mover una cosa con palanca.

ZIGZAGUEAR (1):
andar, moverse o extenderse en zigzag.

DEAMBULAR (1):
andar o pasear sin objeto determinado, por pasatiempo.

ESCARABAJEAR (1):
andar y bullir desordenadamente.

BORNEAR (3):
disponer y mover los sillares hasta dejarlos en su debido lugar.

También puede darse yuxtaposición o coordinación de dos o más oraciones de infinitivo, es decir, de dos términos genéricos con sus respectivos complementos.

CODEAR (1):
mover los codos o **dar** golpes con ellos.

TRAFAGAR (1):
andar por varios países, **correr** mundo.

3.4.3.- Análisis de los modificadores

Los modificadores o complementos del término genérico aparecen situados siempre a su derecha y expresan las funciones de sujeto, de objeto, circunstanciales... Estos complementos se realizan básicamente en sintagmas nominales, sintagmas preposicionales, frases completivas y adverbios.

Como ya se ha señalado anteriormente la información que podemos inferir de la definición es de tipo semántico, la única información sintáctica que quizás pueda extraerse sea a partir de la subcategorización del término genérico, pero raramente de la entrada léxica que se define.

3.4.4.- Gramática

Actualmente se ha realizado la gramática de los verbos de movimiento, que consta de un total de 14 reglas dependientes del contexto y de 9 reglas independientes del contexto.

Las dos primeras reglas de la gramática tienen la función de extraer el término genérico de la definición:

```
(v- (v &&)  v-100)

(v-100
    (v && +v &&) v-110 v-200 v-300 v-400)
```

La primera regla selecciona para el análisis únicamente aquellas definiciones que están categorizadas como verbo y la regla v-100 es la responsable de la obtención del término genérico.

Las reglas restantes se ocupan del análisis de los modificadores y de las distintas combinaciones en que éstos aparecen.

Actualmente se analizan cinco tipos de modificadores del verbo (SN, SP, SGER, ADV y ADJ) y todas las combinaciones posibles entre ellos que se dan en las definiciones.

En el momento de determinar de forma automática el valor semántico de estos modificadores, lo que constituirá la estructura de información resultante del análisis, se plantea el problema de su desambiguación semántica. Los casos más problemáticos, por ser los más frecuentes y por lo tanto los que plantean una mayor ambigüedad, corresponden a los sintagmas nominales (SN) y a los sintagmas preposicionales (SP).

Un SN puede funcionar como un agente o bien como un tema. El orden no es un criterio suficiente para su desambiguación, porque el orden de los argumentos no es fijo. La determinación o indeterminación de los diferentes argumentos puede ser un criterio que posibilite su desambiguación de forma semiautomática. De momento, en la estructura de salida hemos introducido el código AG/TEM (Agente/Tema) que indica esta ambigüedad.

Los SP todavía presentan una diversidad semántica mayor, ya que pueden expresar complementos de manera, de finalidad, locativos (de origen, de destino, de dirección, de lugar), etc. Posiblemente, las estrategias para su desambiguación serán más efectivas y productivas: las distintas preposiciones en cada subconjunto verbal realizan funciones específicas. Por ejemplo, en las definiciones que contienen el verbo "andar" las preposiciones "por" y "sobre" siempre introducen locativos, la preposición "para" finalidad, etc. Este tipo de estrategias facilitan la desambiguación semiautomática de estos sintagmas preposicionales. De momento, en la estructura de salida sólo se dan códigos semánticos a aquellos SPs claramente desambiguables, los restantes se codifican como PP-MOD (Modificadores Preposicionales).

Las siguientes reglas muestran las diferentes combinaciones en que aparecen los argumentos del término genérico:

```
(v-111
    (v && +v *0sadv *pp-mod *0sn *0pp-mod1 &&))

(v-220
    (v && +v *sadv *sn1 *0sadv *0sn *0pp-mod1 &&))
```

```
(v-310
    (v && +v &adj *sn1))

(v-410
    (v && +v &0adj *0pp-mod1 *sger &&))
```

Un ejemplo de regla de contexto libre es la que analiza un verbo en gerundio (+ger) y sus complementos (adv, det noun adj, Opp-mod1, pp-mod, etc.):

```
(sub-pats (QUOTE (*sger))
    (QUOTE
        ((**ger1 (+ger &0adv +0det +noun &0adj *0pp-mod1))
        (**ger2 (+ger *0sadv *pp-mod))
        (**ger3 (+ger para +v *0sn))
        (**ger4 (+ger &adj))
        (**ger5 (+ger &adv))
        (**ger6 (+ger)))))
```

A continuación presentamos un ejemplo de estructura de salida, concretamente la correspondiente a la regla dependiente del contexto v-111

```
(v-111
    ((compound-class +v)
        (iter *0sadv)
        (ag-tem *0sn)
        (pm1 *pp-mod)
        (pm1 *0pp-mod1)
        (r111)))
```

La siguiente regla ilustra la estructura de salida de una regla de contexto libre, concretamente una de las correspondientes al gerundio.

(**ger1 (+ger (maner &0adv)(tema +noun (properties &0adj)) (pm1 *0pp-mod1)))

es decir, el gerundio puede tener un complemento de modo(*maner*), un *tema* que estará realizado como un nombre (noun) con una serie de complementos optativos (*properties*) realizados como adjetivos, seguido opcionalmente de complementos preposicionales, *pm1*.

Con estos ejemplos hemos querido mostrar el nivel de profundidad en el análisis de las definiciones verbales a que se llega actualmente.

En el ámbito de las definiciones verbales, en estos momentos los esfuerzos se centran básicamente en los temas siguientes:

- la búsqueda de criterios que permitan desambiguar semánticamente los argumentos verbales ;
- la subcategorización verbal a partir de la información contenida en el diccionario;
- la extensión del análisis a otros dominios semánticos.

5.- Evaluación

La gramática de nombres y de verbos en su estado actual de desarrollo permite analizar una amplia gama de definiciones del diccionario correspondientes a estas dos categorías. De momento se ha aplicado para la generación de diversas taxonomías de tamaño medio/alto: la taxonomía de *substancia* (1.200 entradas), de *alimento* (146 entradas), de *bebida* (260 entradas) y de cierta complejidad (la profundidad de la taxonomía cuya raíz es "bebida" tiene una profundidad de cinco niveles). En estos casos, el índice de aciertos en las extracciones automáticas del término genérico ha sido de un 95%.

Siguiendo el criterio incremental que hemos comentado al principio de nuestra exposición, la gramática se aplicará para la generación de sucesivas taxonomías y, a medida que se planteen problemas de análisis, se irá mejorando o adaptando según las necesidades que surjan. Como ya se ha apuntado más arriba, existe también la posibilidad de crear gramáticas específicas por conceptos y categorías.

6.- Conclusiones

Esta investigación representa una aportación importante en el marco de la lexicografía de la lengua castellana, ya que no existía ni una metodología lingüística ni unas herramientas informáticas que permitieran el tratamiento de la información contenida en los diccionarios: la extracción y codificación de la información, la construcción de la estructura taxonómica y su traducción en términos de un lenguaje de representación del conocimiento.

Un aspecto que cabe destacar es que los resultados que se han obtenido hasta el momento corresponden a las expectativas previstas según los objetivos que nos habíamos trazado (extracción del término genérico y del máximo de información complementaria)y la metodología que habíamos definido (proceso incremental y semiautomático).

BIBLIOGRAFIA

Agen-91a Ageno A., S. Cardoze, I. Castellón, M.A. Martí, G. Rigau, H. Rodríguez, M. Taulé, M.F. Verdejo (1991a) "An interactive environment fot the extraction and management of taxonomies from MRDs" Esprit BRA-3030 Acqilex WP No. 020.
Agen-91b A. Ageno - S. Cardoze - I. Castellón - M.A. Martí - G. Rigau - H. Rodríguez - M. Taulé -M.F. Verdejo (1991c) "From the LDB to the LKB" Esprit BRA-3030 Acquilex (en preparación).
Agen-91c A. Ageno - S. Cardoze - I. Castellón - M.A. Martí - G. Rigau - H. Rodríguez - M. Taulé -M.F. Verdejo (1991d) "A semiauthomatic process of creation LKB entries"Esprit BRA-3030 Acquilex (en preparación).
Alsh-89 H. Alshawi (1989) "Analysing the dictionary definitions" in Computational Lexicography for Natural Language Processing B. Bougarev & T. Briscoe Eds., LONGMAN Group. London
Ams-81 R.A.Amsler(1981) "A taxonomy for English nouns and verbs" Proceedings of the 19 Annual Meeting of the Association of Computational Linguistics.
Cast-90 Castellón I. , M.A. Martí (1990) "Gramática del diccionario Vox" .VI Congreso de la SEPLN. San Sebastián, 1990. Esprit BRA-3030 Acqilex WP No. 018.
Cast-91 Castellón I., G. Rigau, H. Rodríguez, M.A. Martí, M.F. Verdejo (1991) "Loading the MRD into the LDB. Characteristics of Vox Dictionary" Esprit BRA-3030 Acqilex WP No. 019.
Copes-90a A. Copestake (July 1990) "An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary" presentado en el International Workshop on Inheritance in Natural Language Processing. Esprit BRA-3030 Acqilex WP No. 008.
Copes-90b A. Copestake (September 1990) "A system for building disambiguated taxonomies" Esprit BRA-3030 Acqilex WP No. 012.
Cho-85 M. S. Chodorow, R. Byrd, G.E.Heidorn (1985) "Extracting Semantic Hierarchies from a large online Dictionary", Proceedings of the 23 Annual Meeting of the Association of Computational Linguistics.
Sanf-90 A. Sanfilippo (1990) " Morphological Analyzer for English and Italian" Esprit BRA-3030 Acqilex WP No. 004.
Voss-90 Vossen, Piek (1990) A parser-grammar for the Mening descriptions of the Longman Dictionary of Contemporary English. Technical Report. Amsterdam University, May 1990.

Diccionario General Ilustrado de la Lengua Española VOX Ed. Biblograf S.A. Barcelona 1987.