Establecimiento de equivalentes de traducción mediante ontologías en un lexicón computacional multilingüe

Antonio Moreno Ortiz
amo@uma.es

Chantal Pérez Hernández
mph@uma.es

Facultad Filosofía y Letras
Universidad de Málaga

1 INTRODUCCIÓN

En Moreno Ortiz (1998) se exponen las posibilidades de aplicación del marco lexicológico conocido como Modelo Lexemático-Funcional (MLF) en tareas de Traducción Automática (TA). En este trabajo se desarrolla una base de datos léxica capaz de contener la información léxica derivada del empleo de este modelo y se exploran las posibilidades de utilizar el lexicón computacional resultante en diversos sistemas de TA. De entre éstos, el enfoque basado en el conocimiento (*KBMT: Knowledge Based Machine Translation*) resulta especialmente atractivo por varias razones. En primer lugar, concibe una metodología que intenta simular los procesos cognitivos que se cree tienen lugar en la mente del traductor que, de un modo muy simplificado, vendrían a ser los siguientes: el traductor descodifica el mensaje de la lengua de origen, interpreta éste teniendo el cuenta el contexto en el que ocurre y, a continuación, codifica ese contenido comunicativo en la lengua meta, intentando conservar en la medida de lo posible todas las implicaciones contenidas en el origen. Aplicado al léxico, este proceso implica la existencia de un nivel conceptual, independiente de la lengua, que supone el enlace cognitivo entre las unidades léxicas de ambas lenguas (Wilss 1996). Los estudios léxicos en torno a fenómenos tales como las *lagunas léxicas* o la polisemia vienen a reforzar la suposición de que este nivel conceptual existe y, de hecho, es donde tiene lugar la asignación de equivalentes de traducción.

En la KBMT, este nivel conceptual no es sólo un constructo abstracto, sino que toma cuerpo en una base de conocimiento en la que se almacenan conceptos que servirán de nexo de unión entre las unidades léxicas de dos o más lenguas naturales y que, por definición, son independientes de la lengua. Esta base de conocimiento recibe el nombre de *ontología*, que se caracteriza como "una especificación de una conceptualización" (Gruber 1993).

Nuestro interés por este tipo de recursos proviene de la búsqueda de un modelado de datos léxicos lo más expresivo posible y de soluciones adecuadas a fenómenos tales como la prosodia semántica. En nuestra situación de recepción de análisis lexicográficos de otros componentes nuestro grupo de investigación, nuestra función es la de modelar la información obtenida tras la aplicación de los postulados del MLF (Martín Mingorance 1984, 1990; Faber & Mairal 1999). A menudo nos encontramos con que el lexicógrafo marca la prosodia semántica de predicados o argumentos recurriendo a descripciones entre paréntesis que acompañan a las restricciones de selección, para ofrecer un perfil más afinado del predicado o argumento en cuestión:

(1) **HOWL$_2$** to laugh loudly and repeatedly.
SV Adjunct of Manner (with)
S= +human (Ag)
Adjunct of Manner = *with* + [-concrete ∈ emotions < laughter > (Manner)]
Ex.: "They rolled around on the floor, clutching their stomachs and howling with laughter"

En esta expansión de marco predicativo el argumento 1 (Sujeto) tiene asignada la restricción de selección (SR) *+human* y la función semántica (SF) *Ag(ent)*, mientras que el argumento opcional 2 (satélite), cuya SF es la de *Manner*, ha sido caracterizado, además de con la correspondiente SR (-*concrete*), como [∈ *emotions* <*laughter*>], para conseguir un mayor grado de especificidad. Esto es perfectamente aceptable dentro del marco de trabajo en el que se desarrolla el análisis léxico, donde una RS puede ser tan general o específica como se desee (Dik 1989). Sin embargo, esta libertad también implica que no se puede ejercer ningún control sobre el número o tipo de las restricciones, lo

que está destinado a generar redundancia e inconsistencia, dos características que deben ser evitadas en toda base de datos. Por ejemplo, este lexicógrafo ha usado *emotions*, pero otro podría utilizar *feelings* para hacer referencia al mismo tipo de entidades del mundo real.

Un segundo problema es que el conjunto de RS no posee una estructura interna coherente, del mismo modo que lo tienen los predicados y argumentos propiamente. En este sentido, podemos visualizar el conjunto de restricciones de selección como una lista de elementos, en principio ilimitada, sin relación entre sí, que pretenden simbolizar entidades, propiedades o eventos del mundo real, pero cuyo estatus es el mismo que el de los objetos léxicos que pretenden describir.

Lo que resulta de esta situación es la nula utilidad que algo como <laughter> tiene para un sistema informático. En cuanto que la cadena de caracteres no tiene otro estatus dentro del sistema, no pasa de ser eso, una cadena de caracteres, sin valor alguno fuera del sistema de signos en el que se encuentra. Puesto que ese sistema de signos (el conjunto de las restricciones de selección consideradas) no posee ningún tipo de -valga la redundancia- restricción sobre el tipo o cantidad permitidos, ni posee estructura interna alguna, no se pueden efectuar operaciones con los mismos. Si el sistema de símbolos tuviese una estructura interna coherente, entonces sí sería posible, por ejemplo, medir distancias entre los mismos con objeto de obtener un índice de proximidad que nos pudiese ayudar en procesos de desambiguación semántica.

El problema de las RS es bien conocido dentro del ámbito de la lexicografía computacional, la TA y la inteligencia artificial. Una de las soluciones propuestas por equipos de investigación de KBMT consiste, como ya hemos apuntado, en la utilización de un sistema conceptual estructurado y autónomo, independiente de los diversos lexicones monolingües, es decir, una ontología.

2 LEXICONES Y ONTOLOGÍAS

Para nuestro estudio hemos empleado la ontología de uno de estos sistemas, Mikrokosmos (Mahesh 1996), a la que sus autores ofrecen acceso para usos académicos de forma desinteresada. El primer trabajo ha consistido en integrar dos tipos de recursos muy distintos: por un lado la ontología, originalmente un programa Lisp y, por otro, nuestros lexicones monolingües, implementados en bases de datos relacionales. Para conseguir esta integración hemos ideado un formato propietario para la ontología, que gestionamos desde una aplicación que en estos momentos se encuentra en fase de desarrollo, OntoWorks™, y que nos permite crear y gestionar ontologías e integrarlas con los lexicones. En un sistema de KBMT, la ontología supone el enlace, independiente de la lengua, entre las unidades léxicas. Cada uno de los lexemas de las bases de datos léxicas de las diversas lenguas es asignado a un concepto de la ontología, estableciéndose así, de un modo indirecto, las equivalencias de traducción.

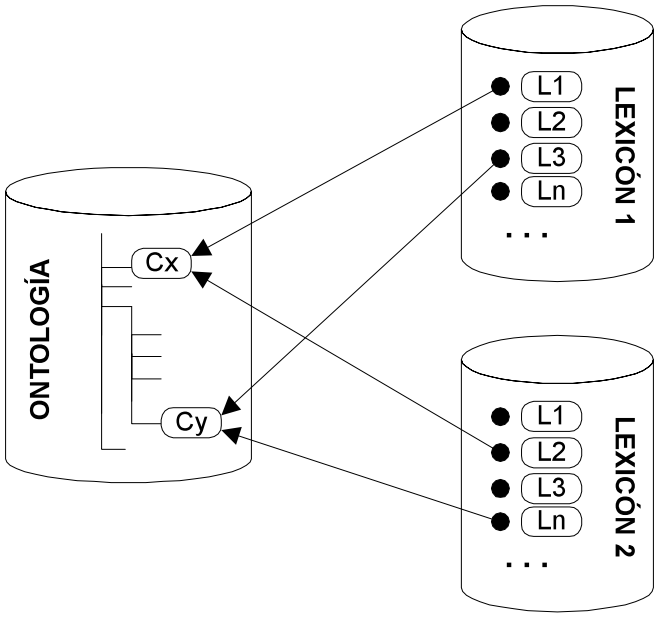

Figura 1. Esquema de integración

En un lexicón MLF los lexemas son caracterizados bien como predicados, que designan relaciones o propiedades, bien como términos, que designan entidades. Esto resulta totalmente compatible con la estructuración típica de un modelo superior de ontología, donde los tres nodos principales, hijos del nodo más alto en la jerarquía (T) son, precisamente OBJECT, PROPERTY y EVENT. En general, asignaremos los términos a OBJECTs, los predicados verbales a EVENTs y los adjetivales a PROPERTYs. A partir del momento de esta asignación, tanto los predicados como los términos de un determinado marco predicativo se convierten en instancias -léxicas- de los correspondientes conceptos de la ontología, heredando no sólo las propiedades y relaciones de éstos sino también las de sus antecesores.

En cuanto al eje paradigmático del lexicón MLF, la utilización de una ontología elimina la necesidad de representar la estructuración de los campos léxicos en los lexicones monolingües, ya que los lexemas quedan estructurados jerárquicamente según sus correspondientes conceptuales. Acorde con el espíritu de la GF, las dimensiones que estructuran el eje paradigmático de un campo léxico no son consideradas como abstracciones, sino como elementos de la lengua misma, con lo que volvemos a tener el mismo problema que mencionábamos en referencia a las restricciones de selección: la imposibilidad de efectuar operaciones abstractas sobre cadenas de caracteres sin otro estatus. En nuestros lexicones hemos mantenido la estructuración paradigmática de los predicados porque, además de su valor lexicográfico obvio, suponen un mecanismo de análisis de gran valor a la hora de construir la ontología.

3 EL CASO DE *HOWL*

Nos serviremos a continuación de un ejemplo, el del verbo inglés *howl*, para mostrar brevemente la forma en la que hemos usado la ontología para enriquecer la base de datos léxica. Este verbo pertenece al campo léxico SOUND y, dentro de éste, cuenta con cuatro acepciones pertenecientes a subdimensiones distintas. La estructuración básica del campo de sonido responde a dos parámetros fundamentales: el ente productor del sonido y la cualidad auditiva del sonido producido. No podemos detenernos aquí a explicar la compleja estructura de dicho campo, que incluye 151 lexemas para un total de 183 acepciones y 543 patrones de complementación. Es importante, sin embargo, señalar que existen cuatro grandes subdimensiones, organizadas en torno al primer parámetro antes mencionado: (i) *sounds produced by humans;* (ii) *sounds produced by animals;* (iii) *sounds produced by nature* y (iv) *sounds produced by objects.*

Howl aparece en las tres primeras de estas subdimensiones. En sus dos primeras acepciones el productor del sonido es identificado como *+Human*, en la tercera el productor del sonido es caracterizado como *+Animal [dog, wolf]* y en la cuarta, el productor del sonido es una fuerza de la naturaleza, descrita con la restricción de selección *+NForce (Natural Force)*:

Dimensiones de las 4 acepciones	Concepto	Eq. Tr. Esp.
howl$_1$: *to shout loud and continuously, because of pain or anger* ⇒ to make a loud sound by speaking ⇑ to make a sound by speaking ⇑ sounds produced by humans	SHOUT	gritar, aullar, berrear
howl$_2$: *to laugh loudly and repeatedly* ⇒ to make a sound expressing happiness ⇑ to make a sound indicating an emotion ⇑ sounds produced by humans	LAUGH	carcajearse, reir a carcajadas
howl$_3$: *to make a long crying sound like a dog or a wolf* ⇒ to make a sound like an angry or wild animal ⇑ sounds produced by animals	HOWL	aullar
howl$_4$: *to make a loud continuous crying noise (strong wind blowing)* ⇒ sounds produced by nature	HOWL	aullar

Tanto la definición en lenguaje natural como su posición en la jerarquía léxica delimitan las cuatro acepciones del verbo *howl*. Sin embargo, esta caracterización no deja de ser una simple cadena de caracteres, de poca utilidad en términos computacionales. Es aquí donde el uso de una ontología puede ser de gran ayuda en diversas aplicaciones de PLN. La información que la base de datos léxica

contiene sobre cada una de estas acepciones puede verse en la siguiente captura de pantalla que muestra la información fonética, morfológica, sintáctica y semántica relativa a una de las acepciones de *howl*:

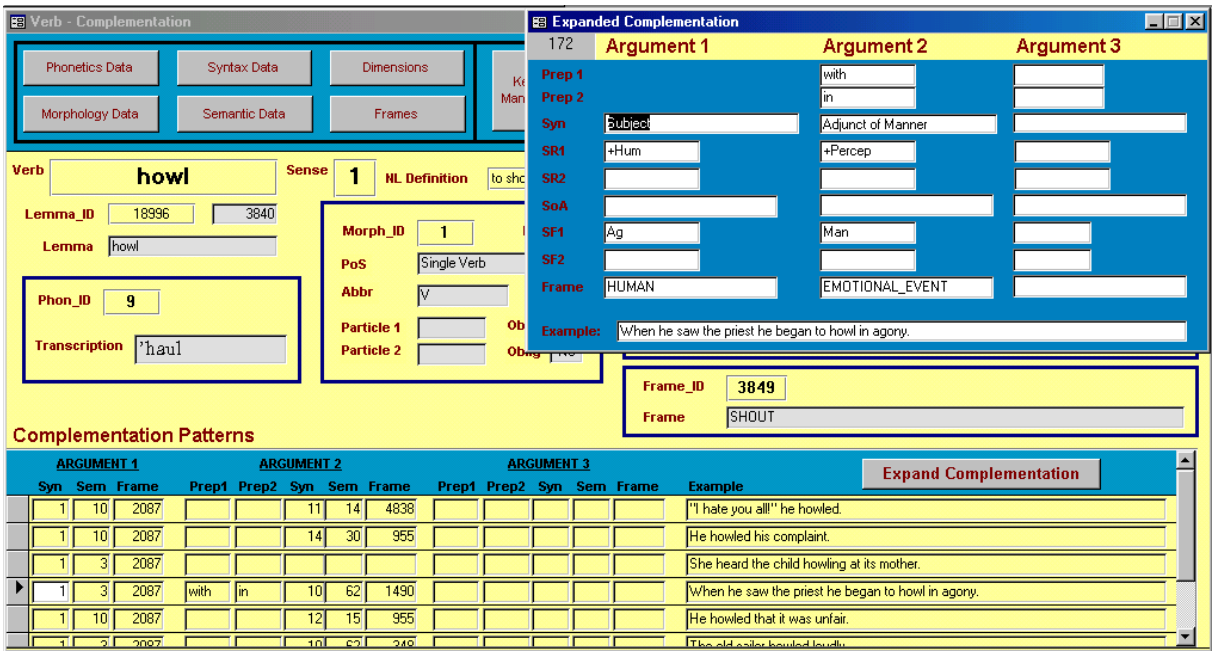

Figura 2. Entrada de la base de datos para $HOWL_1$

Cada una de las acepciones del verbo *howl* ha sido asignada a un concepto de la ontología: $howl_1$ al concepto (*frame)* SHOUT; $howl_2$ al concepto LAUGH y $howl_3$ y $howl_4$ al concepto HOWL. Las características que, mediante la inclusión en dimensiones, aportaban tanto la definición como la posición en la jerarquía léxica se manifiestan ahora en virtud de la asignación de las diferentes acepciones de *howl* a los conceptos de la ontología, además de mediante la información que la ontología ofrece sobre los mismos y sus conceptos padres:

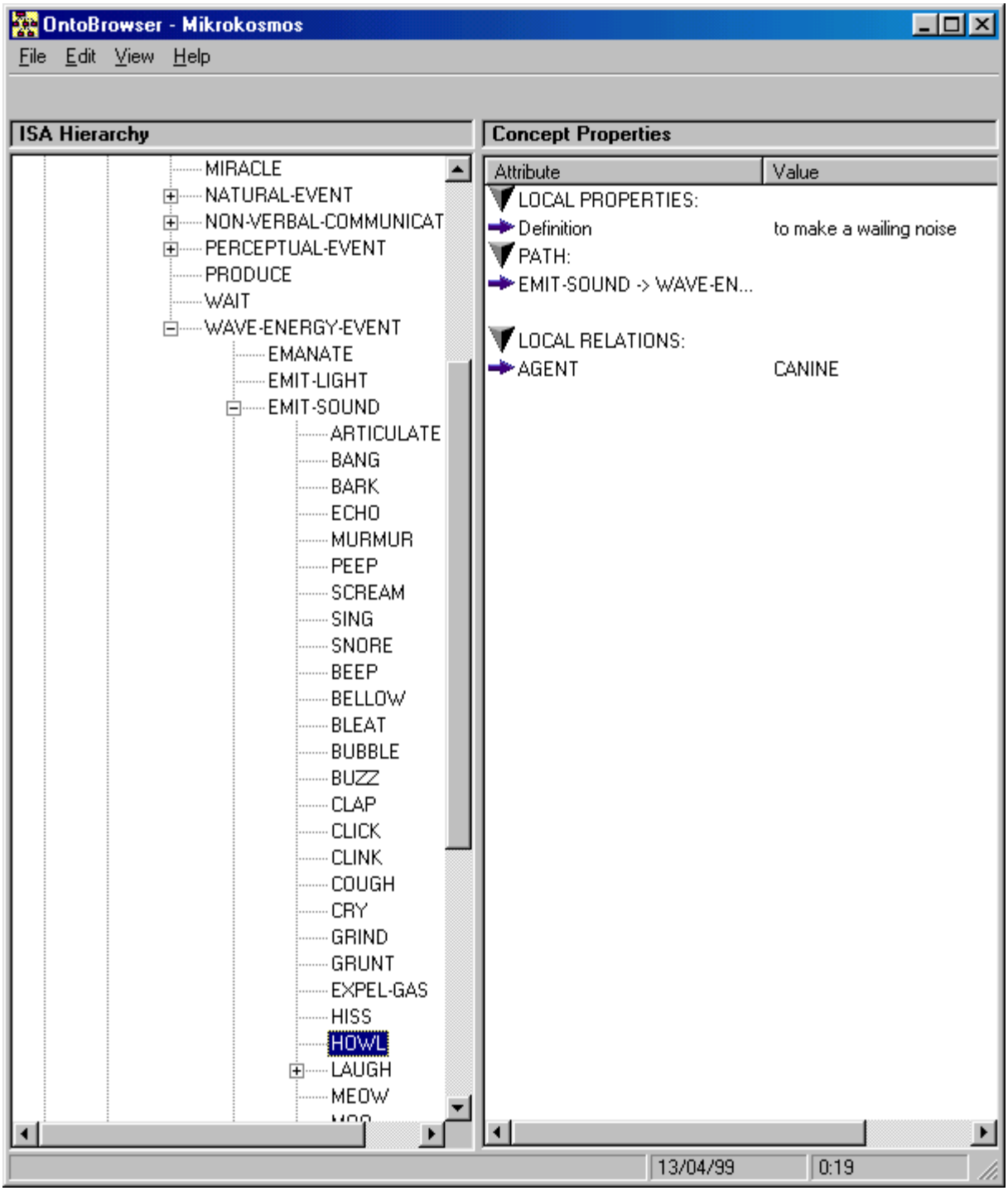

Figura 3. Concepto HOWL en la ontología

Como puede observarse en esta captura de pantalla, perteneciente al visualizador de jerarquías de OntoWorks™ (OntoBrowser), el concepto HOWL está descrito no sólo por medio de una definición, sino por las propiedades que este concepto hereda del superconcepto EVENT (ya que en el caso de *howl* estamos tratando con un predicado) y sus conceptos hijos en la jerarquía (EVENT ⇑ PHYSICAL EVENT ⇑ WAVE ENERGY EVENT ⇑ EMIT SOUND ⇑ [SHOUT | LAUGH | HOWL | CRY | ...]). Los conceptos también son caracterizados por diversas relaciones que se establecen entre los mismos, como es, en nuestro ejemplo, el tipo de agente prototípico que lleva a cabo la acción (y que son, a su vez, conceptos en la ontología): HUMAN en el caso de SHOUT y LAUGH y CANINE en el caso de HOWL. La siguiente captura de pantalla muestra las propiedades y las

relaciones locales del concepto MAMMAL, que serían heredadas a su vez por CANINE, el AGENT prototípico de HOWL.

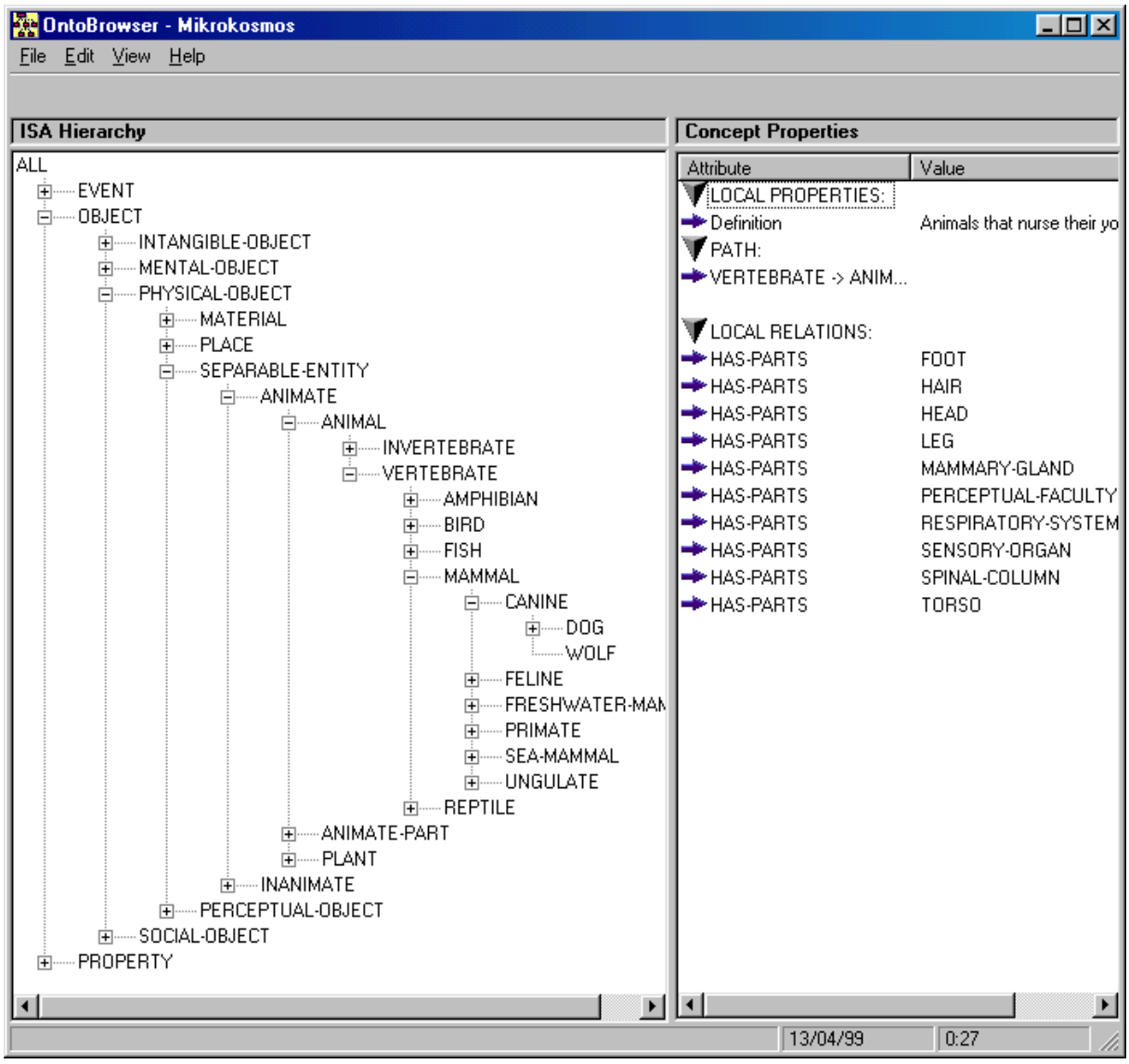

Figura 4. Concepto MAMMAL

Obviamente, todas las propiedades y relaciones especificadas para cada concepto, lo son por omisión. Un desvío del caso protípico va a marcar, en la mayoría de los casos, un uso metafórico del predicado. Por ello, para dar nombres a los conceptos, empleamos por sistema el verbo que tiene como primera acepción la denotación más aproximada del concepto.

4 CONCLUSIONES

Algo que hemos evitado comentar deliberadamente es el proceso de asignación de conceptos. Sin duda alguna éste es el punto más débil de lo que se ha dado en llamar *semántica ontológica* (Raskin & Nirenburg 1998) y se refiere a la arbitrariedad con la que los distintos objetos léxicos son asignados a los conceptos de la ontología, así como la construcción misma de ésta (Nirenburg, Raskin & Onyshkevych 1995). ¿Cuáles son los parámetros que guían la inclusión de un determinado concepto en un punto concreto de la ontología? En este contexto, los diseñadores de Mikrokosmos plantean el concepto de *ontología situada* (Mahesh & Nirenburg 1995), que se perfila como una aproximación absolutamente pragmática al problema: estos parámetros vienen dados por el objetivo concreto de la ontología, incluyéndose aquellos conceptos que sean relevantes para la tarea que se pretende llevar a

cabo. En el caso de Mikrokosmos, además de contener un *upper model* de conceptos generales, contiene conceptos específicos para facilitar la traducción automática de documentos sobre fusiones y adquisiciones empresariales en español e inglés. La construcción de una ontología con aspiraciones de convertirse en un modelo general del mundo es una empresa cuya viabilidad todavía está por demostrarse, pero mientras tanto las ontologías específicas están mostrando su utilidad en muy diversos ámbitos.

Sin embargo, sería interesante contar con parámetros objetivos para la construcción de ontologías enfocadas a la gestión de lenguas naturales. En este sentido, hemos podido comprobar cómo el tipo de análisis que se emplea para generar la macroestructura paradigmática de los campos léxicos, basado en la estructura definicional de los lexemas, arroja pistas importantes sobre las relaciones entre el léxico y el mundo real, por lo que podríamos utilizarlo para la construcción de nuestra ontología. De hecho, en la principal modificación de la sección de la ontología que hemos llevado a cabo para acomodar los conceptos corresopndientes a los campos léxicos SOUND y SONIDO, nos hemos guiado por las dimensiones que conforman estos campos y hemos podido comprobar su utilidad. Como apunta Guarino (1998), la principal peculiaridad de la construcción de ontologías, es la adopción de un enfoque interdisciplianrio donde la filosofía y la lingüística juegan un papel fundamental.

5 REFERENCIAS

FABER, P. & R. MAIRAL. 1999 *The Semantic Architecture of the English Verbal Lexicon*. Amsterda.: Mouton de Gruyter.

GUARINO, N. 1998. "Formal Ontology and Information Systems". Ed. N. Guarino. *Formal Ontology in Information Systems*. IOS Press.

GRUBER, T. R. 1993. "A Translation Approach to Portable Ontology Specifications", Knowledge Systems Laboratory. Technical Report KSL 92-71. Stanford University. Stanford, CA.

MAHESH, K. 1996. *Ontology Development for Machine Translation: Ideology and Methodology. Technical Report* MCCS 96-292, CRL, New Mexico State University, Las Cruces, NM.

MAHESH, K. & S. NIRENBURG 1995. "A Situated Ontology for Practical NLP". *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing.* Montreal, Canada.

MARTÍN MINGORANCE, L. 1984. "Lexical Fields and Stepwise Lexical Decomposition in a Contrastive English-Spanish Verb Valency Dictionary". Ed. R. R. K. Hartmann. *Proceedings from the I International Conference on Lexicography*, Tübingen: Niemeyer, 225-236.

MARTÍN MINGORANCE, L. 1990. "Functional Grammar and Lexematics". Ed. J. Tomaszczyk & B. Lewandowska. *Meaning and Lexicography*. Amsterdam/Philadelphia: John Benjamins.

MORENO ORTIZ, A. 1998. *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática*. Tesis doctoral. Universidad de Córdoba.

NIRENBURG, S., V. RASKIN & B. ONYSHKEVYCH 1995. "Apologiae Ontologiae". *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium.

RASKIN, V. & S. NIRENBURG 1998. "An Applied Ontological Semantic Microtheory of Adjective Meaning for Natural Language Processing". *Machine Translation* 13(2): 135-227.

WILSS, W. 1996. *Knowledge and Skills in Translation Behaviour*. Amsterdam. John Benjamins.