Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 259-266 recibido 04-05-2005; aceptado 01-06-2005

Diccionarios basados en taxonomías con estructura de grafo orientado acíclico

Antonio Vaquero Sánchez
Departamento de Sistemas Informáticos y Programación
Facultad de Informática
Universidad Complutense de Madrid
C/ Prof. José García Santesmases, s/n, 28040, Madrid
vaquero@sip.ucm.es

Francisco Alvarez Montero
Departamento de Sistemas Informáticos y Programación
Facultad de Informática
Universidad Complutense de Madrid
C/ Prof. José García Santesmases, s/n, 28040, Madrid
francisco_alvarezm@fdi.ucm.es

Fernando Sáenz Pérez
Departamento de Sistemas Informáticos y Programación
Facultad de Informática
Universidad Complutense de Madrid
C/ Prof. José García Santesmases, s/n, 28040, Madrid
fernan@sip.ucm.es

Resumen: Al irse multiplicando las bases de datos léxicas en diferentes formatos, ha crecido también la preocupación acerca del nivel de reusabilidad de los recursos léxicos. El intercambio e integración de datos, así como el desarrollo de software común, son importantes para evitar duplicaciones de esfuerzos y posibilitar el desarrollo de bases de datos de información lingüística a gran escala. Es necesario un enfoque unificado en el desarrollo de herramientas para la creación y gestión de bases de datos léxicas. Su estructura debe seguir unas pautas basadas en unos niveles de abstracción claramente delimitados y su implementación debe estar controlada por métodos bien fundados de ingeniería de software. La tecnología de bases de datos relacionales es adecuada para soportar los datos y lograr una mayor flexibilidad, reusabilidad y expandibilidad. Con esa metodología se han desarrollado herramientas para la creación y gestión de diccionarios monolingües basados en una taxonomía conceptual con una única relación en forma de grafo orientado acíclico, donde una categoría puede tener más de un padre. Es éste un paso más en el camino que se está recorriendo hasta llegar a herramientas que gestionen bases de datos léxicas bien estructuradas y con múltiples relaciones semánticas.

Palabras clave: Bases de datos léxicas, diccionarios, modelo E-R, ontologías, taxonomías

Abstract: As lexical databases have proliferated in multiple formats, there has been a growing concern over the reusability of lexical resources. The interchange and integration of data, as well as the development of common software is increasingly important to avoid duplications of effort and enable the development of large-scale databases of linguistic information. A unified approach for the development of tools to create and manage lexical databases is necessary. Their structure must follow some patterns based on some clearly delimited abstraction levels and the implementation must be controlled by sound and well-founded software engineering methods. Relational database technology is suitable for data support and to attain a greater flexibility, reusability and expandability. With such a methodology we have developed tools for the creation and management of monolingual dictionaries based on a Directed Acyclic Graph shaped conceptual taxonomy with a single relation where a category can have more than one parent. This is a new step in the way to develop tools for managing well structured lexical databases and with multiple semantic relations.

Keywords: Lexical Databases, Dictionaries, E-R Model, Ontologies, Taxonomies

ISSN: 1135-5948

1 Introducción

El léxico se ha convertido en el recurso lingüístico ineludible y el componente más caro de construir en cualquier aplicación de PLN (Douglas et al., 1994).

Mucha de la investigación sobre léxicos se ha centrado en la creación de bases de datos léxicas (BDL) a partir de la información contenida en diccionarios convencionales para el consumo humano, produciendo diccionarios informatizados para utilizarlos en tareas de PLN. Sin embargo, este enfoque tiene como resultado la creación de BDL en formatos ad-hoc al tratar de transformar directamente lo que está en papel a un formato informáticamente aprovechable.

Aunque siguiendo este enfoque se obtienen BDL útiles para el PLN, estas BDL proveen principalmente información lingüística acerca de los términos: su equivalente en otra lengua, información morfológica, etc. La información conceptual es escasa, poco organizada, inconsistente e implícita (Meyer et al., 1996).

Además de información puramente lingüística, los diccionarios deben poseer algún tipo de conocimiento conceptual que permita comprender el significado de un término, tanto a personas como a las aplicaciones de PLN.

Existen esfuerzos de construcción de recursos léxicos que no están ligados a la organización lineal y alfabética de sus predecesoras, por ejemplo, WordNet (Miller et al., 1993) o el enfoque seguido en (Véronis e Ide, 1992) para la construcción de léxicos en dominios cerrados.

Sin embargo, aunque estas BDL hacen referencia a conceptos, la estructura y relaciones entre éstos no está explícitamente representada. El conocimiento lingüístico y el conocimiento del dominio están enlazados de tal manera que forman un solo bloque monolítico. Esto impide su reutilización (Mahesh y Nirenburg, 1996), así como también la integración de diferentes BDL.

La construcción de BDL para PLN, independiente del tipo de aplicación que las utilice, debe contemplar, no una, sino dos entidades (Nirenburg y Raskin, 1987). En la primera se encuentra representado de manera lógica, consistente, estructurada y explícita el conocimiento del dominio, en forma de una ontología compartida e independiente del lenguaje. En la segunda debe estar plasmado el conocimiento lingüístico de las lenguas objeto de tratamiento. Cada léxico, correspondiente a un idioma determinado, ha de estar convenientemente enlazado a la ontología común.

Aunque no existe duda acerca de lo que es un léxico, el término ontología puede tener varias interpretaciones. Por ejemplo, Wordnet es mencionada como ontología en (Guarino, 1998) y en (Uschold y Gruninger, 1996). En el caso más simple, una ontología describe una jerarquía de conceptos organizados e indexados en base a un principio de clasificación, es decir, una taxonomía conceptual.

En el área de PLN, las taxonomías, a pesar de que sólo cubren una parte del problema de representación de conocimiento, proporcionan una base sólida para éste (Jacobs, 1991).

Sin embargo, tanto la construcción de ontologías (Gómez Pérez, 1999) como la construcción de recursos lingüísticos (Sáenz y Vaquero, 2005a) y terminológicos (Kurshid et al, 1995) adolecen de una falta de normalización y rigor metodológico.

Existe una metodología (Sáenz y Vaquero, 2005b) para el desarrollo de BDL basada en principios de ingeniería de software, que permite crear diccionarios fáciles de utilizar y, además, estructuralmente preparados para ser utilizados por cualquier aplicación que los necesite.

La metodología se ha aplicado para el desarrollo de diccionarios monolingües (Vaquero, Sáenz y López, 2003b), bilingües (Vaquero y Sáenz, 2002) y multilingües (Vaquero y Sáenz, 2003a) y ahora, basándonos en ella y en modelos precedentes, hemos desarrollado un modelo conceptual de datos bin fundado y sencillo para diccionarios monolingües.

Los modelos precedentes sólo permiten crear diccionarios limitados a representar taxonomías en forma de árbol. En el nuevo modelo la taxonomía puede adoptar la forma de un grafo orientado acíclico, permitiendo representar clasificaciones complejas, en donde una categoría puede tener más de un padre.

El resto del artículo está estructurado de la siguiente manera: en la sección 2 se hace mención a los diversos conceptos lingüísticos representados en el modelo conceptual. En la sección 3 hablamos sobre modelos conceptuales para BDL y se describe el modelo propuesto En la sección 4 se describe brevemente el estado de desarrollo de la herramienta de creación de diccionarios monolingües. En la sección 5 se

enumeran las conclusiones obtenidas y en la sección 6 se plantea el trabajo futuro.

2 Conceptos lingüísticos

En esta sección se señalan los distintos conceptos que se encuentran representados en los modelos conceptuales de los distintos diccionarios desarrollados con la metodología. En todos los modelos se representan una serie de conceptos lingüísticos que son importantes, tanto para el PLN como para la comprensión humana.

Aquí solo se mencionarán, pues ya han sido explicados en detalle en (Sáenz y Vaquero, 2005a) anteriormente. Dichos conceptos son: el significado, las categorías semánticas, la definición, la taxonomía conceptual, los términos y las relaciones léxicas.

3 Modelos conceptuales para bases de datos léxicas

Cuando se pretende usar un diccionario como referencia o para aplicaciones de PLN, se vuelve crucial la necesidad de almacenarlo en una estructura regular y consistente. Esto permite que los datos de la base de datos léxica puedan ser identificados, buscados o modificados fácil y consistentemente.

El modelo relacional clásico ha sido propuesto para representar diccionarios en (Nakamura y Nagao, 1988) y en (Tidemann, 2002). Sin embargo ninguno de estos proyectos parte de un modelo conceptual de datos apropiado para bases de datos léxicas, ni sigue un enfoque de ingeniería de software para la construcción de la base de datos ni de sus herramientas de creación y gestión. Por lo tanto, proponemos un modelo conceptual de datos sólido y sencillo para diccionarios monolingües, donde el conocimiento del dominio esté plasmado de manera explícita y estructurada.

El modelo propuesto representa un refinamiento sobre modelos anteriores (Vaquero y Sáenz, 2002) (Vaquero y Sáenz, 2003a) (Vaquero, Sáenz y López, 2003b) y es el punto de partida para la aplicación de una metodología que nos permitirá construir, no sólo la BDL, sino también las interfaces para su manipulación.

3.1 Consideraciones preliminares

Como estamos interesados en la creación de modelos conceptuales para la creación de bases de datos léxicas basadas en ontologías, antes de pasar a la descripción del modelo conceptual propuesto, es necesario hacer algunas precisiones.

Los modelos precedentes sólo permiten crear taxonomías en forma de árbol. Esto representa una limitación, pues se pueden obtener clasificaciones complejas en donde una categoría puede tener más de un padre.

Puede pensarse que esto refleja un error de criterio cometido a la hora de construir la taxonomía, pero en realidad estas clasificaciones existen en la vida real, se ha visto que son útiles y, por lo tanto, no pueden ser ignoradas (Raguenaud y Kennedy, 2002).

El modelo propuesto refleja este tipo de clasificaciones y permite representar taxonomías en forma de grafo orientado acíclico.

El significado está representado en la taxonomía como una entidad independiente del lenguaje que pertenece a una categoría. Éste está directamente relacionado con las categorías, no con los términos, lo que permite una clasificación jerárquica de significados. La sinonimia es una propiedad del conjunto de términos y es el conjunto mismo el que está relacionado con un significado. Ello es debido a usar definiciones intensionales para las categorías.

3.2 Modelo conceptual del diccionario monolingüe

A partir del modelo taxonomía-léxico, la siguiente etapa consiste en aplicar la metodología propuesta en (Sáenz y Vaquero, 2002), para obtener un modelo E-R que represente la estructura de la base de datos léxica.

Los modelos E-R están basados en cuatro elementos semánticos básicos: entidades, relaciones, atributos y valores. Siguiendo algunas recomendaciones de (Silberschatz, 2002), las entidades están representadas por rectángulos, los atributos por elipses y las relaciones por rombos.

Una entidad representa un objeto de interés en el dominio. Los atributos de una entidad describen sus propiedades utilizando valores apropiados. Las relaciones se conectan con las entidades por medio de líneas.

Si a una entidad le llega una línea con flecha (dirigida), significa que la relación tiene cardinalidad de uno a varios (1:N). Una línea no

direccional (sin flecha) indica que la cardinalidad es de varios a varios (N:N). Las líneas también se usan para unir los atributos a las entidades o relaciones

3.2.1 Entidades

En nuestro modelo conceptual (Figura 1) se destaca la entidad Significados, que representa los significados de la base de conocimiento. La entidad Términos representa todos los términos de la base de datos terminológica. La entidad Categorías denota la categoría a la que pertenece cada significado. La entidad Comentarios representa todos los posibles comentarios que se pueden asociar a los términos.

3.2.2 Relaciones

La relación Cosin entre Significados y Términos denota el conjunto de sinónimos (synset) bajo una acepción y es N:N porque un conjunto de sinónimos puede contener varios términos (sinonimia) y un mismo término puede estar en diferentes conjuntos de sinónimos (polisemia).

La relación Véase denota el conjunto de términos que, sin ser sinónimos, están relacionados con un synset y es N:N porque un mismo término puede referirse a otros y puede aparecer referenciado por otros términos.

La relación PerteneceA denota la categoría a la que pertenece un significado y es N:N porque hay varios significados correspondientes a una categoría y un mismo significado puede estar en varias categorías. Esta relación implica que nuestra clasificación no es léxica (no hay una relación directa entre categoría y término) sino semántica: se relacionan significados con categorías.

La relación ComentarioTérmino denota los comentarios asociados a cada término y es N:N porque un mismo término puede tener varios comentarios y el mismo comentario se puede referir a varios términos.

La relación PadreDe denota la taxonomía conceptual y es N:N porque una categoría puede tener más de una categoría padre (excepto el nodo raíz) y una categoría puede tener varias categorías hijo.

3.2.3 Atributos

La entidad Categorías tiene el atributo NombreCategoria, que denota el nombre textual de la categoría. La entidad Significados tiene el atributo Definición, que denota la definición textual del significado. La entidad Términos tiene el atributo NombreTérmino, que denota el nombre textual del término. Finalmente, la entidad Comentarios tiene el atributo TextoComentario, que denota el texto del comentario.

4 Herramientas de creación de diccionarios

A partir del modelo E-R de la base de datos léxica se ha proseguido con las siguientes fases de la metodología, obteniendo el diseño lógico y físico de la base de datos. También se han diseñado e implementado interfaces para visualización y control de diccionarios monolingües con una estructura taxonómica en forma de grafo acíclico, de acuerdo a la metodología de ingeniería de software descrita en (Sáenz y Vaquero, 2005b).

Nuestro trabajo no se centra únicamente en la representación de datos y sus relaciones, sino también en las restricciones entre ellos. Estas restricciones que pueden implementarse en el modelo relacional nos permiten imponer restricciones sobre datos y relaciones que debe verificar cualquier ejemplar de la base de datos. Aunque estas restricciones se pueden implementar en la capa de aplicaciones, es desaconsejable porque el propio gestor de bases de datos puede incorporarlas de forma general para todas las aplicaciones, al igual que sucede con los datos.

No obstante, dada la naturaleza del proceso de creación de BDL, no se pueden imponer todas las restricciones identificadas puesto que hay información ausente que puede conocerse a posteriori o por conveniencia del propio proceso de creación. Por ello, este tipo de restricciones se deben proporcionar como potencialmente violables, es decir, como restricciones débiles que el autor puede comprobar en cualquier momento del proceso de creación. En cambio, las restricciones denominadas fuertes nunca deben ser violadas y el sistema gestor de bases de datos las vigila automática y continuamente.

Nuestras herramientas incorporan ambos tipos de restricciones y proporcionan al autor, cuando éste lo requiere, un informe de las restricciones débiles violadas para que se tomen las medidas oportunas.

En la Figura 2 se muestra la herramienta de creación de BDL para diccionarios

monolingües. Dispone de varias fichas que catalogan diferentes acciones que el autor puede realizar sobre la BDL. En particular, con la ficha Informe de inconsistencias se permite obtener el informe de violación de restricciones débiles.

5 Conclusiones

A partir del trabajo desarrollado podemos concluir lo siguiente:

- Se ha refinado la metodología utilizada para construir BDL monolingües partiendo de una ontología común con estructura taxonómica en forma de grafo orientado acíclico.
- La metodología se ha aplicado a la construcción de herramientas para crear y gestionar diccionarios.
- Se han definido las interfaces correspondientes, con una visualización determinada por el usuario, lo más cómodas y amigables para éste.
- Se ha comprobado el buen funcionamiento de las herramientas y las interfaces.
- Se ha comprobado que el intercambio e integración de datos, así como el desarrollo de software con una metodología común es importante para evitar duplicaciones de esfuerzos y posibilitar el desarrollo de bases de datos de información lingüística a gran escala.
- La utilización de tecnología de bases de datos relaciones se ha mostrado muy eficaz para lograr una mayor flexibilidad, reusabilidad y expandibilidad de las BDL.
- El desarrollo de ontologías y de recursos léxicos es un campo de exploración importante en el que se acusa ausencia de normalización, sin la cual no puede haber un compartimiento efectivo del conocimiento ni una integración eficiente de recursos lingüísticos.

6 Trabajo futuro

Como líneas de trabajo futuro podemos destacar las siguientes:

- Crear diccionarios monolingües (y multilingües) con las herramientas según la metodología.
- Estudiar de manera empírica el fenómeno de creación de BDL con múltiples relaciones taxonómicas.
- Integrar las herramientas existentes y las nuevas en un sistema único para facilitar y controlar la creación y uso de estas BDL.
- Construir herramientas para crear vistas normalizadas de BDL no normalizadas.
- Crear las herramientas de migración para integrar BDL normalizadas.
- Dotar a las herramientas de la capacidad de control para asegurar el cumplimiento de las propiedades lógicas de las relaciones entre conceptos (Guarino y Welty, 2001).
- Utilizar todos estos recursos lingüísticos en las aplicaciones de PLN y comprobar su adecuación.

Bibliografía

Gómez Pérez, A. 1999. "Ontological Engineering: A State of the Art". Expert Update. British Computer Society. Autumn. Vol. 2. Nº 3.

Guarino, N. "Formal Ontology and Information Systems". 1998. Proceedings of the FOIS'98, IOS Press, pp.3-15.

Guarino N and Welty, C. 2001. "Ontological Analysis of Taxonomic Relationships", International Conference on Conceptual Modeling". The Entity Relationship Approach. Salt Lake City, USA.

Jacobs, P.S. 1991. "Integrating Language and Meaning in Structured Inheritance Networks", In John F. Sowa (Ed.), Principles of Semantic Networks: Explorations in the representation of knowledge. Chapter 18. San Mateo, California: Morgan Kaufman.

Khurshid, A. et al. 1995. "Aspects of terminology infrastructure in Europe", Volume 1 - Analysis of terminology management systems in Europe, Technical Report CS-95-12. Guildford: University of Surrey.

Mahesh, K. and Nirenburg S. 1996. "Meaning Representation for Knowledge Sharing in Practical Machine Translation", Proceedings of the Florida Artificial Intelligence Research Symposium, FLAIRS-96. Florida, USA.

Meyer, I. et al. 1992. "Towards a New Generation of Terminological Resources: An Experiment Developing a Terminological Knowledge Base", In Proc. 14th International Conference on

Computational Linguistics. Nantes, pp. 956-960.

Miller, G.A. et al. 1993. "Introduction to Wordnet: An On-line Lexical Database", Journal of Lexicography, 3(4):234--244, 1990. Revised August.

Nakamura J. and Nagao. 1988. "Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation", Proceedings of the 12th International Conference on Computational Linguistics, COLING'88. Budapest, Hungary. 459-464.

Nirenburg, S. and Raskin, V. 1987. "The subworld concept lexicon and the lexicon management system", Computational Linguistics 13(3-4): 276-289.

Raguenaud, C. and Kennedy, J. 2002. "Multiple Overlapping Classifications: Issues and Solutions", 14th International Conference on Scientific and Statistical Database Management (SSDBM'02). Edingburgh, Scotland.

Sáenz, F. and Vaquero, A. 2002. "Towards a Development Methodology for managing Linguistic Knowledge Bases", Proceedings ES'2002. Springer-Verlag. pp 453 – 466.

Sáenz, F. and Vaquero, A. 2005a. "Knowledge Representation Issues and Implementation of Lexical Databases", Second International workshop on UNL, other interlinguas and their applications in the framework of the conference on Intelligent Text Processing and Computational Linguistics CICLING-2005, February.

Sáenz, F. and Vaquero, A. 2005b. "Applying Relational Database Development Methodologies to the Design of Lexical Databases", Database Systems 2005, IADIS Virtual Multi Conference on Computer Science and Information Systems (MCCSIS 2005), April.

Silberschatz, A., Korth, H.F., and Sudarshan S. 2002. "Database System Concepts", WCB/McGraw-Hill.

Tiedemann, J. 2002. "MatsLex - a Multilingual Lexical Database for Machine Translation", Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, 29-31, May, pp 1909-1912.

Uschold, M. and Gruninger, M. 1996. "Ontologies: principles, methods, and applications", Knowledge Engineering Review, Vol. 11, 2, pp 93-155.

Vaquero, A. and Sáenz, F. 2002. "Creación de bases de conocimiento lingüístico para el aprendizaje de idiomas", Simposio internacional de informática educativa RIBIE'02, ISBN: 84-8158-228-X, Servicio de Publicacións da Universidade de Vigo, November.

Vaquero, A. and Sáenz, F. 2003a. "A Human-Learning Environment for Building and Querying Electronic Dictionaries", Asialex'03, Urayasu (Japan), August.

Vaquero, A., Sáenz, F. and López, C. 2003b. "Herramientas para la creación de diccionarios monolingües con objetivos pedagógicos", Challenges 2003 - 5º SIIE, Braga (Portugal), September.

Véronis, J. and Ide, N. 1992. "A Feature Based model for Lexical Databases", International Conference On Computational Linguistics. Proceedings of the 14th Conference on Computational Linguistics 2: 588-594.

Zock, M. and Carroll, J. 2003. "Les dictionaires électroniques". TAL, Vol. 44, 2.

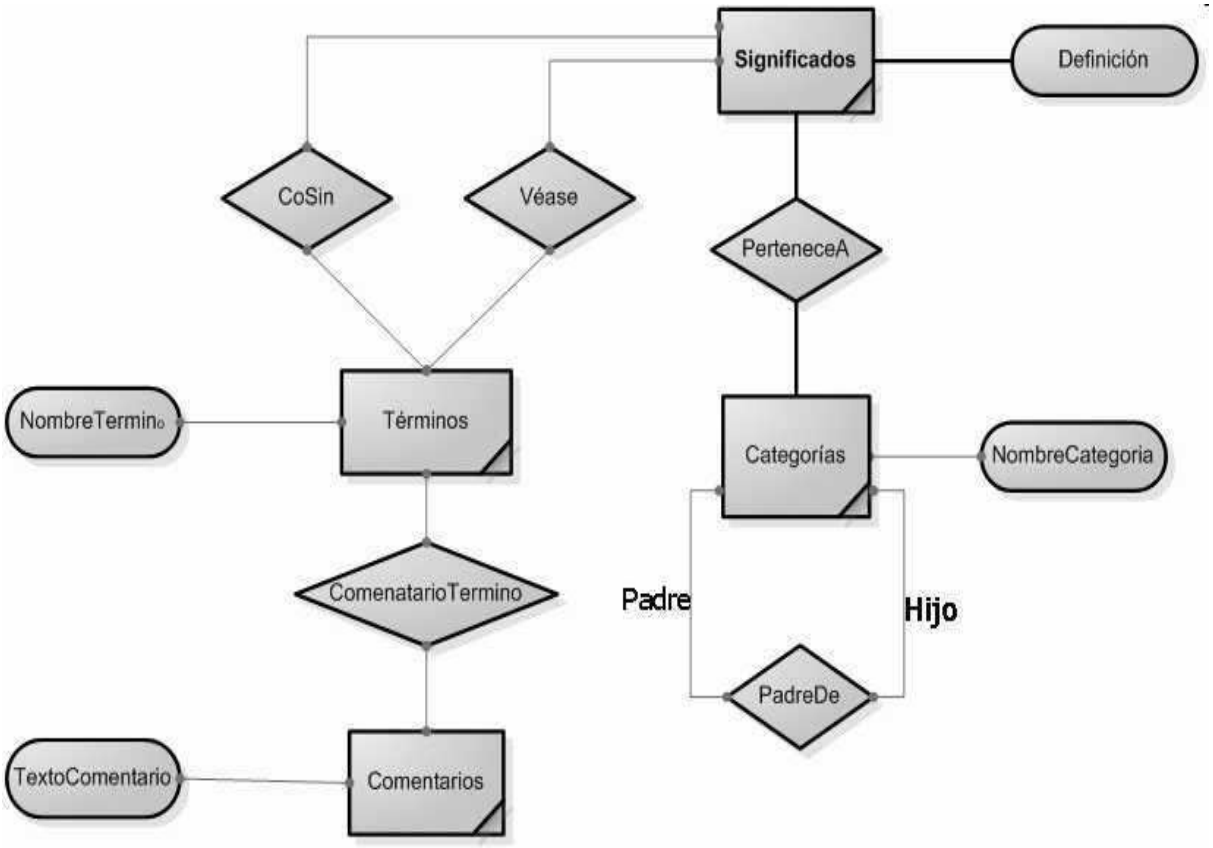

Figura 1. Modelo conceptual para diccionarios monolingües.

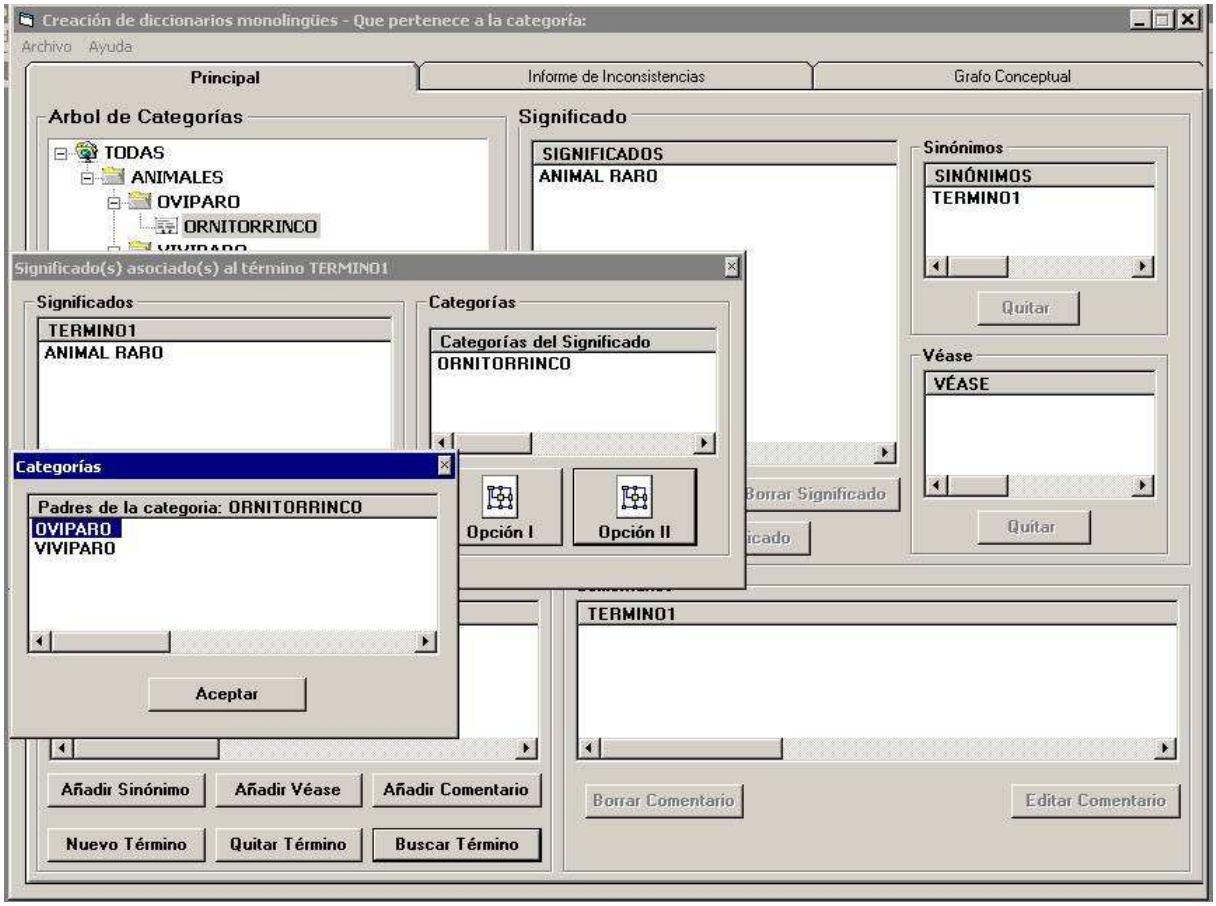

Figura 2. Herramienta de creación de BDL para diccionarios monolingües.