

**Utilización de estructuras verbales en la identificación de relaciones y descriptores en
tesauros II***

J. A. Moreiro González, J. Lloréns Morillo, M. A. Marzal García-Quismondo, J. Morato Lara, S.
Sánchez Cuadrado y P. Beltrán Orenes.

Departamento de Biblioteconomía y Documentación y Departamento de Informática
Universidad Carlos III de Madrid

Dirección postal:

Facultad de Ciencias Sociales y Jurídicas Facultad de Humanidades, Comunicación y

Documentación

C/ Madrid, 126 – 28903

Ciudad y País: GETAFE (MADRID) SPAIN

Teléfono/Fax: 34916249238 Fax: 34-91-6249757

Correo electrónico: jamore@bib.uc3m.es

RESUMEN

Este documento trata de dar una visión de diferentes clasificaciones de relaciones. Se ofrece una perspectiva de diversos estudios. Al mismo tiempo establece similitudes y diferencias con otras propuestas. Se propone una nueva iniciativa aplicando nuevas tecnologías. Igualmente presentan nuevos conceptos y la identificación de los mismos para establecer relaciones tesaurales. Por último se plantea una nueva clasificación de relaciones.

I. INTRODUCCIÓN

En la primera parte de este trabajo¹ partíamos de la definición de tesoro dada por la Norma ISO 2788-1986² como un *vocabulario de un lenguaje de indización controlado organizado*

* Este trabajo ha sido realizado dentro del marco del Proyecto financiado por la CICYT (Comisión Interministerial de Ciencia y Tecnología), titulado "Desarrollo de un tesoro de verbos para entornos de

formalmente con objeto de hacer explícitas las relaciones a priori entre conceptos. Esta estructura mencionada anteriormente se basa, fundamentalmente, en la definición de una serie de relaciones entre descriptores³. Es decir, por un lado se identifican los términos que han de pertenecer a la lista de vocabulario a tener en cuenta –los descriptores–, mientras que por otro se analizan las posibles relaciones entre los mismos.

La idea que subyace bajo el término descriptor es la de designar de manera unívoca cada concepto del tesoro mediante un término normalizado (*controlado*). La normalización de los descriptores atañe también a su categoría gramatical, ya que la norma desaconseja el uso de adjetivos, adverbios y verbos, permitiendo casi exclusivamente sustantivos y frases nominales.

En cuanto a las relaciones, las tres que describe la ISO 2788-1986 son:

1. Equivalencia: para los sinónimos y cuasi-sinónimos (sinónimo sólo en determinado contexto).
2. Jerarquía: para describir las relaciones de super y subordinación. Existen tres subtipos: genérica, parte-todo y enumerativa
3. Asociativa: que son básicamente, los términos relacionados que no pueden ser agrupados bajo las relaciones de equivalencia y jerarquía. Existen varios subtipos: disciplinar, instrumentales, causalidad, atributivas, de medición, etc.

Por su parte, la norma UNE 50125 añadió una relación idiomática para permitir la implementación de varios idiomas en un mismo tesoro. Fundamentalmente, la norma define una lengua fuente y una serie de lenguas objetivos. Los conceptos de la lengua fuente se traducen a la lengua objetivo mediante una serie de reglas que dependiendo de si existe una equivalencia exacta, inexacta (términos que significan conceptos similares), parcial (un término tiene un significado más amplio que el otro), compuesta (un término de una lengua puede traducirse como combinación de dos términos en el otro idioma), o no-equivalencia.

información dinámica. Aplicación del estándar ISO/ICE: 13250:1999, del Plan General del Conocimiento. TIC 2000-2003.

¹ Este documento es la continuación del artículo titulado "Utilización de estructuras verbales en la identificación de relaciones y descriptores en tesauros", que aparece en el último número de la revista Ciencias de la Información

² ISO 2788-1986 Directrices para el establecimiento y desarrollo de Tesauros Monolingües. Revista Española de Documentación Científica, 12(4) 1989: 463-482 y 13(1) 1990: 601-629.

³ La norma ISO 2788-1986 también denomina a los descriptores como términos preferentes, por oposición a los no preferentes o sinónimos.

A principios de los 90, Georges Van Slype⁴ ya sugería que se añadieran nuevos tipos de relaciones, aunque bien, es cierto que muchas de ellas tenían una relación más sintáctica que semántica. Sin embargo, esta corriente de matizar en mayor medida la tipología relacional era el germen de lo que en los siguientes años será una de las principales tendencias en investigación.

En el trabajo de D. Tudhope, H. Alani y C. Jones⁵ sobre la expansión de las relaciones en los tesauros automatizados actuales es donde se puede apreciar el incremento real, desde el punto de vista pragmático, que han sufrido las posibles relaciones de un tesoro, en especial las de asociación que son las que más subtipos tienen en la actualidad. Estos autores parten de un estudio de la *American Library Association (ALA)*⁶, y dividen, de acuerdo al mencionado estudio, en nueve grandes subtipos las relaciones asociativas de primer nivel:

1. Ideas combinadas.
2. Términos relacionados conceptualmente.
3. Contigüidad.
4. Relaciones asociativas por definición.
5. Relaciones asociativas con diferente jerarquía o facetadas.
6. Relaciones asociativas traslapadas por significado.
7. Relaciones asociativas con idéntica jerarquía.
8. Cuestiones de finalidad.
9. Relaciones asociativas sin especificar.

Cada uno de estos subtipos se divide, a su vez, varias veces, alcanzando en algunos casos los 6 niveles de profundidad. Lo que ha multiplicado exponencialmente el número posible de relaciones asociativas presentes en un tesoro.

En la primera parte de nuestro documento se analizaban también algunas de las ventajas que presenta la incorporación de formas verbales que desempeñen el papel de descriptores, en

⁴ Slype, Georges Van: *Los lenguajes de indización: concepción, construcción y utilización de los sistemas documentales*. Fundación Germán Sánchez Ruipérez Madrid, 1991

⁵ Tudhope, Douglas, Alani, Harith y Jones, Christopher. "Aumenting Thesurus Relationships: Possibilities for Retrieval". *Journal of Digital Information* (1), febrero 2001. Dirección <http://jodi.ecs.soton.ac.uk/Article/v01/i08/Tudhope> Consultado el 27/01/02.

⁶ Appendix B (part 2) Taxonomy of Subject Relationships compiled by Dee Michel with the assistance of Pat Kuhr June 1996 draft (hierarchical display), en <http://ala.org/alcts/organization/ccs/sac/appendxb.html> Consultado: 08/02/02.

especial cuando estamos en presencia de la catalogación de ciertos materiales, como los vídeos, en los que el movimiento es una característica intrínseca al material tratado⁷. En el presente escrito, vamos a continuar el prometedor campo que se abre cuando la incorporación de las formas verbales tiene como función crear relaciones entre los conceptos.

II. SIMILITUDES Y DIFERENCIAS CON OTRAS PROPUESTAS

Dos son las corrientes que nos han servido de base para el desarrollo concreto de nuestra propuesta de inclusión de formas verbales en la indización y construcción de los tesauros. A continuación analizaremos las similitudes y diferencias con nuestro planteamiento.

La primera de ellas es la **Utilización de clasificaciones verbales**. Este tipo de planteamientos se fundamenta en la relación semántica para mejorar la recuperación de información. La mayoría de los trabajos que se orienta en esta línea se basan en las aportaciones realizadas por Levin⁸ (1993). Un buen ejemplo lo constituye los trabajos de Green⁹, en los que se combina la red semántica de WordNet¹⁰ con la clasificación verbal propuesta por Levin, generando así una red semántica ampliada con respecto a la de partida, WordNet.

Nuestra propuesta se aleja de estos desarrollos, pues consiste en la creación de un tesoro de formas verbales como complemento al tesoro clásico de sustantivos. Esto supone que no se parte de una mera ampliación de la red semántica WordNet, y tampoco se limita a la clasificación verbal de Levin. Nuestro desarrollo pasa por un proceso que no es meramente automático, sino que incluye otro proceso de naturaleza más lingüística en el que se puedan encajar las categorías verbales a modo de relaciones facetables. Y es este punto el que nos separa de la mayoría de proyectos de este tipo, puesto que, en la mayor parte de los trabajos llevados a cabo en esta línea,

⁷ Todas estas cuestiones han sido desarrolladas en el primero de estos dos documentos.

⁸ Levin, B. English verb classes and Alternations: a preliminary investigation, University of Chicago Press Chicago, Ill., 1993.

⁹ Green, Rebecca, Pearl, Lisa, Dorr, Bonnie J., and Resnik, Philip. Mapping Lexical entries in Verbs Database to WordNet Senses. En Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001), Toulouse, France, July 9-11, 2001.

¹⁰ Fellbaum Christiane (1998) Wordnet : An Electronic Lexical Database (Language, Speech and Communication). Cambridge (Massachusetts): MIT Press.

se trata de localizar un conjunto de términos (a los que denominan *clusters*), que resultan relevantes por su frecuencia de coaparición en determinado contexto.

La identificación del rol de una asociación mediante un verbo, que nosotros proponemos, permite un abanico de relaciones mucho más adaptable a dominios concretos. Por ejemplo, ofrece la posibilidad de salvar ciertos modos y tiempos verbales como matices de la relación con la entidad y esto, a su vez, concreta la identificación de un documento en cuanto a su tipología.

Esta nueva fórmula de relaciones permite una indización automática más flexible, es decir, la dinamización del tesoro mejorará en gran medida la representación de múltiples dominios y aumentará, al mismo tiempo, la precisión y eficacia en los sistemas recuperación. La idea de representar algunos aspectos de un dominio mediante la identificación de los verbos no es totalmente nueva, existen estudios anteriores que sugieren las ventajas de la inclusión de los verbos para las labores de recuperación como el de Rumbaugh¹¹ para OMT (*Object Modeling Technique*), un precursor de UML (*Unified Modeling Language*)¹². Éste es también el enfoque clásico de modelado en el diseño de bases de datos relacionales¹³. La simplificación de representación obtenida por los lenguajes de modelado de software hace posible esquematizar y hacer comprensible información abstracta y característica de él mundo informático. Dentro de los lenguajes de modelado destaca por su aceptación y cobertura el mencionado Lenguaje Unificado de Modelado (UML). Las posibilidades y la semántica que puede representar este lenguaje son diferentes a la de los tesauros tradicionales. Por ejemplo, se pueden mostrar los agentes que interaccionarán con el sistema y de qué modo (casos de uso) lo realizan. Así como también se pueden mostrar las secuencias temporales de interacción con la información, es decir diagramas de secuencia y actividad. La semántica de relaciones de UML ofrece, tal como se veía en los ejemplos anteriores, más posibilidades en sus diagramas de clases que las de los tesauros clásicos.

¹¹ Rumbaugh, James: Modelado y diseño orientado a objetos: Metodología OMT. Prentice Hall. Madrid, 1998.

¹² Para una revisión consultar el libro de Stevens, Perdita & Pooley, Rob. Using UML: Software Engineering with Objects and Components. Essex: Pearson Education Limited, 2000.

¹³ Miguel Castaño, Adoración de; Piattini Velthuis, Mario: *Fundamentos y modelos de bases de datos*. Ra-Ma. Madrid, 1997.

Por otra parte, la tipología de relaciones de asociación del lenguaje UML destaca por el número y su definición no ambigua, lo que va en consonancia con la corriente actual de modificar el número de relaciones de asociación de los tesauros. Así, por ejemplo, en este lenguaje existen tipos como la relación de agregación, en la que la desaparición del todo no implica la desaparición de las partes, y la de composición, en la que la desaparición del todo implica la desaparición de sus partes. Además, se puede añadir información de multiplicidad (es decir, cuántos objetos pueden interactuar en una misma relación), dirección de la relación, y tipificación de relaciones. La superioridad de esta aproximación para la integración verbal viene avalada por diversos estudios pedagógicos¹⁴, donde esta forma de relacionar los conceptos mediante verbos se denominan mapas conceptuales (en inglés *concept maps*)¹⁵.

La segunda de las corrientes se basa exclusivamente en la **Utilización de WordNet**. Esta herramienta conlleva una red semántica multidisciplinar en inglés. Su calidad y disponibilidad lo han convertido en la herramienta idónea en lingüística durante los últimos años. Desde el principio, su uso ha estado muy vinculado a la desambiguación conceptual. Dentro de esta tendencia, es una de las líneas que está dando mejores frutos es el empleo para desambiguar términos mediante verbos. Un ejemplo representativo de este tipo de desambiguación se puede encontrar en Moldovan¹⁶ donde se esquematiza de la siguiente forma¹⁷:

- i. Se seleccionan de la frase todas las parejas de sustantivo-verbo

¹⁴ Novak, J.D. Clarify with concept maps: A tool for students and teachers alike. In: *The Science Teacher*, 1991. 58(7), pp. 45-49.

¹⁵ Hacemos esta precisión terminológica para que no se confunda el concepto que queremos resaltar aquí con el de los *Topic Maps*, cuya traducción es idéntica terminológicamente, pero no desde el punto de vista semántico.

¹⁶ Moldovan también ha propuesto otros algoritmos de desambiguación basados en WordNet, véase Sanda M. Harabagiu & Dan I. Moldovan. Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text. In: *Natural Language Processing & Knowledge Representation*. Ed. Lucja M. Iwńska and Stuart C. Shapiro. Menlo Park (CA): AAAI Press, MIT Press, 2000. p. 301-333.

¹⁷ El autor denomina al método desambiguación semántica. Using WordNet and lexical operators to improve Internet searches. Moldovan-DI; Mihalcea-R. *IEEE-Internet-Computing*. vol.4, no.1; Jan.-Feb. 2000; p.34-43. Con este sistema Moldovan afirma que obtiene una considerable mejora frente a Yarowsky (1995) que desambiguaba el 94% de los sustantivos (no otros términos) y a Stetina (1998) y este trabajo, desambiguar el 80% de todas las partículas.

- ii. Se escoge el significado más probable del término (subproceso que Moldovan denomina Desambiguación Terminológica):
 1. Las palabras de la frase se agrupan en parejas.
 2. Se buscan en WordNet los distintos significados de cada término.
 3. Se forman todos los diferentes pares de conceptos posibles.
 4. Se busca cada par en Internet. Luego se ordenan los resultados según los conceptos más frecuentes, esto es, según el número de veces que en Internet nos aparezcan juntos dos conceptos determinados.
- iii. Teniendo en cuenta los conceptos más frecuentes, se seleccionan todos los sustantivos de los “glosarios” de cada verbo y sus descendientes jerárquicos.
- iv. Se calcula mediante una fórmula los conceptos comunes entre los sustantivos del punto anterior.
- v. Se ordenan todas las parejas de conceptos de sustantivo-verbo según el resultado de la fórmula.

Este proceso, aunque fructífero de cara a la desambiguación, no permite, sin embargo, la identificación de las relaciones que establecen los distintos verbos con los que trabajan, quedándose así en un plano semántico muy concreto, es decir, el relativo a la identificación terminológica.

III. IDENTIFICACIÓN DE CONCEPTOS DINÁMICOS

Una vez analizadas las corrientes más importantes en la inclusión de verbos en la indización y recuperación documental, pasamos a la exposición de nuestra propuesta, a la que denominaremos **Identificación de conceptos dinámicos**¹⁸. Nuestro primer objetivo es conseguir que los conceptos dinámicos indiquen a priori las relaciones clásicas de los tesauros

(principalmente jerarquía, asociación y equivalencia, aunque se tienda a extenderlas a todas las variantes del modelo que incluyan los restantes subtipos, como es el caso de las relaciones de asociación presentadas más arriba).

Para alcanzar este primer objetivo, se está automatizando la construcción de tesauros basándose en las concurrencias de determinados sustantivos del tesoro que aparecen junto a determinados conceptos dinámicos. Esta primera aproximación se está llevando a cabo en *corpora* documentales concretos, sin embargo, aunque el sistema esté en menor o mayor medida adaptado a un determinado dominio, se pretende que, en un futuro no muy lejano, sea aplicable a cualquier dominio. Así, el proceso comienza indizando un *corpus* documental, localizando los descriptores (principalmente sustantivos) que interrelacionan las estructuras verbales en cada frase¹⁹. En concreto, los pasos que sigue el método²⁰ para la elaboración del tesoro son:

- i. Un análisis de documentos relevantes para la extracción de su vocabulario (glosarios, diccionarios, estándares sobre vocabulario, etc)
- ii. Depuración manual del vocabulario extraído, obteniendo los descriptores de los tesauros
- iii. Indización con estos descriptores de documentos textuales relacionados con Calidad (manuales, estándares, artículos, etc).
- iv. En esta etapa se almacenan principalmente aquellas frases del documento en las que aparece uno o varios descriptores del tesoro en el Sintagma Nominal Sujeto, y uno o varios descriptores en el Sintagma Verbal. Los elementos que aparecen entre ambos descriptores son identificados como los conceptos dinámicos del documento.

¹⁸ Tal como señalamos en nuestro anterior trabajo, llamaremos conceptos dinámicos a las formas verbales.

¹⁹ Díaz Rodríguez, S. I., Esquemas de representación de información basados en relaciones: aplicación a la generación automática de representaciones de dominios, Tesis doctoral, Director, Juan Lloréns Morillo. Leganés: Universidad Carlos III de Madrid, Departamento de Informática, 2001.

²⁰ La metodología expuesta esta ampliamente descrita en la tesis doctoral: Díaz Rodríguez, S. I.: Esquemas de representación de información basados en relaciones: aplicación a la generación automática de representaciones de dominios. Tesis doctoral, Director, Juan Lloréns Morillo. Leganés: Universidad Carlos III de Madrid, Departamento de Informática, 2001.

- v. Posteriormente, los conceptos dinámicos se agrupan, clasifican y se asimilan a las relaciones del tesoro que se deseen identificar²¹.
- vi. Manualmente, se revisan en el tesoro las relaciones obtenidas.
- vii. Se implementará el resultado para su consulta en Web.

El modelo teórico que sustenta el método descrito tiene como idea marco realizar una distribución de las relaciones de los tesauros de manera lógica. Para conseguir lo anterior se han mantenido los tres tipos de jerarquía, equivalencia y asociación, y se han seleccionado unas pocas relaciones que, por su importancia y elevada presencia, constituyen, a nuestro entender, la base perfecta para comenzar la elaboración del tesoro que proponemos. El siguiente cuadro contiene estas relaciones:

²¹ Un estado del arte de este particular se puede encontrar en: Morato, J (2001) Utilización de Estructuras Verbales en Tesauros dentro de un Entorno de Recuperación Documental Automatizada: Estado del Arte y Propuesta de un Método. Informes Técnicos del Departamento de Informática de la Universidad Carlos III (UC3M-TR-CS-2001-03): 181-203.

ASOCIACIÓN	<i>DISCIPLINA</i>		Un campo de estudio y los objetos que estudia
	<i>INSTRUMENTAL</i>	<i>INSTRUMENTO</i>	Operación o proceso y su agente instrumental necesario para que se produzca la acción
		<i>MATERIAL</i>	Material u objeto sobre el que actúa un instrumento
	<i>CAUSALIDAD</i>	<i>CAUSA/EFEECTO</i>	Una causa o acción y el efecto que produce
		<i>CAUSA/OBJETO</i>	Una acción y su sujeto pasivo
	<i>MEDICIÓN</i>		Conceptos y sus unidades de medida
	<i>PROCEDENCIA</i>		Un concepto relacionado con su origen físico
		<i>LOCAL</i>	Un concepto relacionado con su origen espacial
		<i>PROCESO O TRANSICIÓN</i>	Cambio natural de un estado
		<i>TEMPORAL</i>	Un concepto relacionado con su origen temporal
			Sucesión en el tiempo
	<i>OPOSICIÓN</i>	<i>AGENTE</i>	Un concepto y un agente contrario
		<i>ANTONIMIA</i>	Oposición de conceptos: incluye todos los tipos de antonimia.
	<i>PROFESIÓN</i>	<i>OCUPACIÓN</i>	Un ser vivo y su ocupación
<i>AUTORIA</i>		Conceptos relacionados con su autor o persona física	
<i>PROPIEDAD/ATRIBUTO</i>		Conceptos y sus propiedades. La propiedad que indica el material que forma algo se pone dentro de la relación parte todo	
<i>PARTE-TODO</i>	<i>AGREGACIÓN</i>	El término "parte" no debe de poder sustituir siempre al "todo", no hay herencia de propiedades (coche-rueda, si puedo decir "un coche necesita gasolina" pero no puedo decir "una rueda necesita gasolina"). Suelen ser subdivisiones de estructuras sociales	
	<i>COMPOSICIÓN</i>	Agregación en la que una parte no puede pertenecer a más de un todo. Al contrario de la agregación, aquí la dependencia es tan fuerte que si desaparece el "todo" del tesoro, también lo harían las partes, pues dejarían de tener sentido	
	<i>PROXIMIDAD CONCEPTUAL</i>	Yuxtaposición de dos conceptos, conceptos similares pero no equivalentes. Puede existir un subtipo denominado relación idiomática (traducción a otro idioma de un término: "En inglés, car significa coche"). Esta relación también está muy próxima a los sinónimos	
EQUIVALENCIA	<i>SINÓNIMOS</i>	<i>ABSOLUTOS</i>	Un concepto y un equivalente exacto. Es una relación menos frecuente de lo que se pueda pensar, ya que normalmente se trata de "proximidad conceptual" entre términos.
		<i>ABREVIATURAS</i>	El mismo concepto abreviado
		<i>TECNICISIMOS/TÉRMINOS POPULARES</i>	Término exacto y su equivalente popular
		<i>VARIANTES DIALECTALES</i>	Algunas de estas variantes seguramente son sinónimos absolutos
JERARQUÍA	<i>GENERO-ESPECIE</i>		La relación jerárquica debe cumplir que el término específico puede sustituir al genérico y seguir siendo una frase sintáctica y semánticamente correctas (entre animal y pato, si puedo decir "un animal come" puedo decir "un pato come"). El específico "hereda" todas las propiedades del genérico
	<i>CLASE-INSTANCIA</i>		Enumeración de nombres propios que pertenecen a una misma clase. Hay herencia de propiedades.