



AVANCES EN

LEXICOGRAFÍA, TERMINOLOGÍA Y TRADUCCIÓN

Marisela Colín Rodea
Erika Ehnis Duhne
(coordinadoras)



La presente obra está bajo una licencia de:
<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es>



Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Advertencia](#).

Usted es libre de:

Compartir — copiar y redistribuir el material en cualquier medio o formato

Adaptar — remezclar, transformar y construir a partir del material

La licenciente no puede revocar estas libertades en tanto usted siga los términos de la licencia

Bajo los siguientes términos:



Atribución — Usted debe dar [crédito de manera adecuada](#), brindar un enlace a la licencia, e [indicar si se han realizado cambios](#). Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciente.



NoComercial — Usted no puede hacer uso del material con [propósitos comerciales](#).



CompartirIgual — Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la [misma licencia](#) del original.

Esto es un resumen fácilmente legible del:
texto legal de la licencia completa

En los casos que sea usada la presente obra, deben respetarse los términos especificados en esta licencia.



ENALLT ESCUELA NACIONAL
DE LENGUAS, LINGÜÍSTICA
Y TRADUCCIÓN UNAM

Los tesauros y la informática: puntos de encuentro

JESÚS VALDÉZ RAMOS

Introducción

El tema de la informática en la elaboración de un *tesauro* ha sido estudiado de manera abundante y de forma exhaustiva por parte de los ingenieros en informática y cómputo, así como por científicos de la información. Este tema se ha estudiado desde diferentes puntos de vista. Por ejemplo, desde el diseño interno de la herramienta —es decir las características y condiciones que se deben tomar en cuenta para que un tesauro sea una herramienta sólida y consistente con los principios teóricos del diseño de instrumentos informáticos para la recuperación de la información—, también desde el punto de vista del uso, aplicación y eficacia de la herramienta en la recuperación de la información. O bien desde el punto de vista de las ventajas que tiene el uso del tesauro en la indización, que es la contraparte del proceso del análisis documental en un sistema de almacenamiento y recuperación de la información. Asimismo, se ha investigado sobre la construcción automática de los tesauros y su impacto en los niveles de eficiencia y calidad en la recuperación de información en un sistema, o sobre los tesauros electrónicos en los que se incluyen interfaces que utilizan herramientas de apoyo multimedia, entre otros aspectos.

En este trabajo presentaré los resultados de la investigación bibliográfica realizada sobre las diferencias que se encontraron en la definición del término tesauro en diferentes obras de consulta especializadas en informática, en oposición a las definiciones del término en obras de consulta especializadas en lingüística. Asimismo, señalaré los principios básicos que deben considerarse para la construcción de tesauros y mencionaré cuatro momentos o etapas

importantes en el desarrollo de los tesauros que se realizaron con el apoyo de la Informática y en el contexto de la recuperación de la información.

Tesauros, una definición desde la informática

En primera instancia, iniciaré señalando las variantes que se encontraron en la definición del término tesoro en diferentes obras de consulta especializadas en informática, en oposición a las definiciones del término en obras de consulta especializadas en lingüística.

Desde la lingüística, Hartmann, R.R.K. (2006) señala que el término tesoro «...ha tenido diferentes extensiones en el significado a lo largo de los años. Ahora puede decirse que el término tiene al menos tres diferentes sentidos: 1) lista especial de palabras o *lexicon*, 2) diccionario semántico o *nomenclatur* y 3) base de datos terminológica o índice”.

La primera acepción, de acuerdo con Hartmann, ahora es obsoleta y se asocia con los proyectos lexicográficos relacionados con diccionarios literarios como el de Thomas Cooper, *Thesaurus Linguae Romance et Britannicae* (1665); o el *Thesaurus Linguae Latinae* (1894). La segunda acepción está relacionada con la idea de un diccionario organizado semánticamente; un producto típico de este enfoque es el famoso *Thesaurus of English Words and Phrases*, elaborado a mediados del siglo XIX por el famoso médico británico Peter Mark Roget.

La tercera acepción es la más reciente y se puede asociar con la idea de un índice para un vocabulario de un campo específico del conocimiento especializado que se puede utilizar para el estudio y la codificación de terminología técnica y la recuperación de la información. Un ejemplo de una obra que ha sido elaborada de acuerdo con esta acepción es *Thesaurus of Subject Headings for Television* (1969) de Sharon Black.

Para la informática el término tesoro ha sido definido desde diferentes enfoques, veamos. En el *Diccionario de Informática* (Illingworth, 1993) se define como «...característica de los sistemas de tratamiento de textos por la cual las expresiones similares

y sinónimos pueden visualizarse en pantalla e incorporarse en el texto. Los sistemas de recuperación de texto completo pueden poseer *búsqueda de tesauros* como opción por la cual pueden establecerse también términos similares en significado a los buscados». Asimismo y de acuerdo con Virga y Mesters (1994) un tesauro es «...un léxico jerarquizado de términos comprensibles por el programa para la búsqueda de una información. Un *tesauro* es, por tanto, una herramienta indispensable para consultar de forma eficaz un banco de datos. Para cada uso en particular, es necesario crear un nuevo tesauro (no se consulta de igual forma un banco de datos jurídico que uno médico)». En el *Dictionary of Computing* (Collin, 1995: 280) un *tesauro* es «...un archivo que contiene sinónimos que se presentan como alternativas para solucionar *Thesaurus of Subject Headings for Television* mediante un *spell-check*».

Después de analizar las definiciones que se han presentado se puede hacer una serie de consideraciones: se observan diferencias en la forma y en el enfoque al presentar la información. Por una parte, las obras de consulta especializadas en lingüística y lexicografía que fueron consultadas definen el término *tesauro* con mayor profundidad, presentan información sobre el origen y desarrollo histórico del término, señalan de manera precisa las diferentes acepciones y usos que tiene el término. Por otra, las definiciones presentan también datos sobre las características que tiene un tesauro, es decir, ser un léxico especializado organizado bajo criterios semánticos y de manera jerárquica y sobre los diferentes usos que tiene este léxico (es decir sirve para el estudio y codificación de terminología técnica y la recuperación de la información).

Las definiciones que fueron consultadas en diccionarios de Informática presentan información más escueta y, sobre todo enfocada hacia el uso y aplicación de los tesauros en los sistemas para la recuperación de la información, es decir el uso de la herramienta para propósitos específicos. De lo anterior podemos concluir que en la definición del término tesauro que presentamos desde la Lingüística y la Informática hay diferencias, en la forma y el contenido de la información, es decir en una se presenta información sobre el desarrollo histórico de la herramienta y sus características lingüísticas y en la otra se hace énfasis en la aplicación y uso que se hace de ella.

Para la lingüística un tesoro es un léxico especializado que se organiza de una forma diferente a un diccionario o a un glosario, es decir, su organización interna está regulada bajo criterios semánticos primordialmente y con una estructura jerárquica del tipo género-especie, en la que se establecen relaciones de asociación, equivalencia y jerarquía.

Para la informática un tesoro es una herramienta lógico conceptual a través de la cual se representa el contenido de los documentos. Se trata de un sistema complejo en el cual se establecen relaciones entre los términos para crear una red semántica que permite al sistema realizar tareas de indización y recuperación de información.

Pasos básicos para la elaboración de un tesoro

Una vertiente que he investigado sobre la relación entre la Informática y los tesoros es la que se refiere a los principios básicos para su construcción. Desde la perspectiva de la informática, la elaboración de un tesoro tiene como punto de partida el reconocimiento de grupos de conceptos que comparten características. Cada grupo está organizado como una jerarquía de conceptos y representado por términos en lenguaje natural adecuados al contexto en el que se utilizará el tesoro. Asimismo y para una comunicación eficiente entre el usuario y el sistema de búsqueda, se requiere un vocabulario regulado con relaciones conceptuales claramente definidas y notas de alcance que favorezcan el uso consistente de la herramienta.

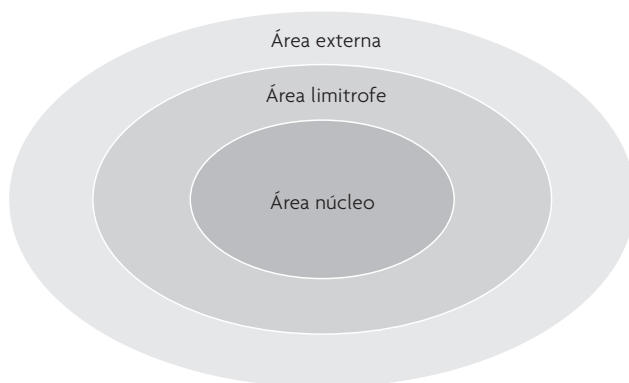
En los manuales consultados se sugiere que antes de iniciar la construcción de un tesoro es conveniente plantearse las siguientes preguntas:

- ¿Quiénes y cuántos serán los usuarios del tesoro?
- ¿Cuáles son las expectativas de los usuarios potenciales del tesoro?
- ¿Saben usar un tesoro para la búsqueda de la información?
- ¿Podrán usar un mapa conceptual para encontrar información?

Los pasos básicos que la Informática propone —así como la bibliotecología— para la elaboración de un tesoro son los siguientes:

Definir el rango y profundidad de un tesauro

La profundidad se define por el número de términos y su especificidad. Un campo de información puede ser dividido en tres áreas: 1) Área núcleo, que contiene temas y términos totalmente pertinentes al tema principal. 2) Área limítrofe, que contiene temas y términos que apoyan al tema principal, pero que no están directamente relacionados. 3) Área externa, que contiene temas y términos que pueden apoyar al tema principal, pero que no aportan conocimiento experto al mismo. De manera gráfica podría expresarse de la siguiente forma:



Identificar las fuentes para construir el vocabulario

Se recomienda utilizar las siguientes fuentes:

- Una muestra de la literatura clásica sobre el tema del tesauro.
- Una muestra del área, como memorias de congresos, publicaciones académicas reconocidas del que giran en torno al tema central.
- Publicaciones de organismos oficiales y reconocidos.
- Resúmenes de las principales publicaciones periódicas del tema.
- Una muestra de las preguntas de los usuarios.
- Tesauros e índices existentes en el tema.

Hasta aquí pareciera que esta es la forma tradicional de elaborar un tesauro, todavía no interviene la informática en la tarea, es decir es una tarea manual. A partir de ésta ya intervendrá; a continuación presento algunos datos acerca de la manera en que lo hará.

Extracción del vocabulario o términos candidatos

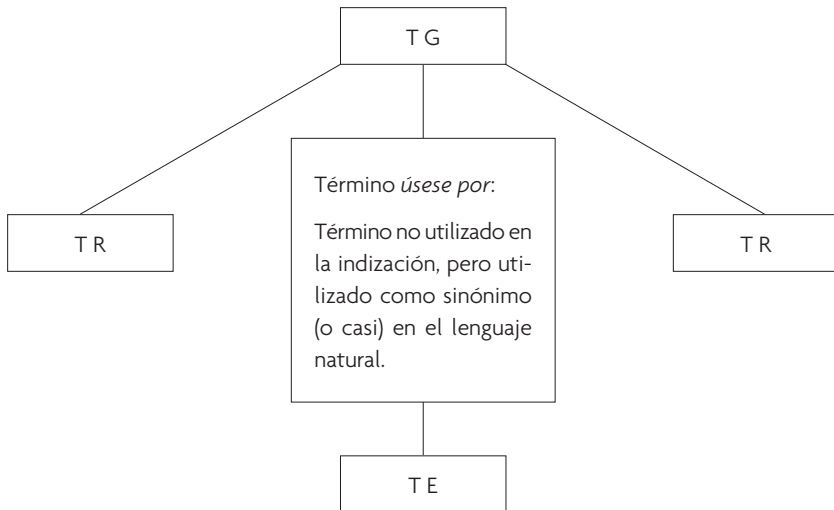
El criterio inicial para seleccionar los términos que formarán el tesauro es tomar los términos directamente de las fuentes documentales. Así, si aparecen los términos *salario*, *jornal*, *cuota* que significan *pago por un trabajo hecho*, se deben tomar debido a que aparecieron en los documentos fuente. Ya existen herramientas informáticas que pueden extraer términos de manera automática de algunas fuentes documentales.

Organizar el vocabulario en grupos de términos asociados

La organización de términos, de manera breve, consiste en el agrupamiento de éstos atendiendo a criterios semánticos de afinidad. Por ejemplo, *madera*, *plástico*, *metal*, se agruparían bajo el concepto de *materiales*. De forma general podemos decir que los términos se agrupan en tres categorías: asociación, equivalencia y jerarquía. Esta etapa del trabajo, que implica un trabajo manual, es la que consume la mayor parte del tiempo de construcción de un tesauro. Sin embargo, existen herramientas informáticas que si se pueden utilizar por las características, acelerarían esta etapa.

Acomodo de grupos de términos

El acomodo de grupos de términos se realiza a través del establecimiento de relaciones entre los términos, que son de tipo jerárquico y son del tipo padre-hijo; o relaciones de tipo hermanos, mismas que corresponden a las entradas de TG, TE y TR. Que de manera gráfica podría expresarse de la siguiente forma:



Establecer un sistema de notación

Se propone definir una notación a través de un código simbólico en el que se enuncie la jerarquía de los términos. Sayers, W.C. Berwick. (1992) señala que la notación debe ser fácil de pronunciar, escribir y recordar. Al definir un sistema de notación se debe tomar en cuenta las perspectivas de crecimiento del tesauro.

El tratamiento de los términos pre-coordinados o de más de una palabra

A medida que se elabora un tesauro de forma automática se presentan situaciones o momentos en los que se deben incluir algunos términos que están formados de más de una palabra. Este hecho tiene algunas implicaciones importantes para el tratamiento automático de los términos. Entre estas implicaciones se pueden distinguir al menos tres situaciones: 1) cuando el término de más de una palabra se utiliza en un campo de conocimiento determinado, 2) cuando la frase de más de una palabra aparece frecuentemente en los documentos que se han tomado como muestra para la indización, y 3) cuando es necesario utilizar la frase de más de una palabra para hacer más claro el significado de la misma.

Refinamiento del vocabulario

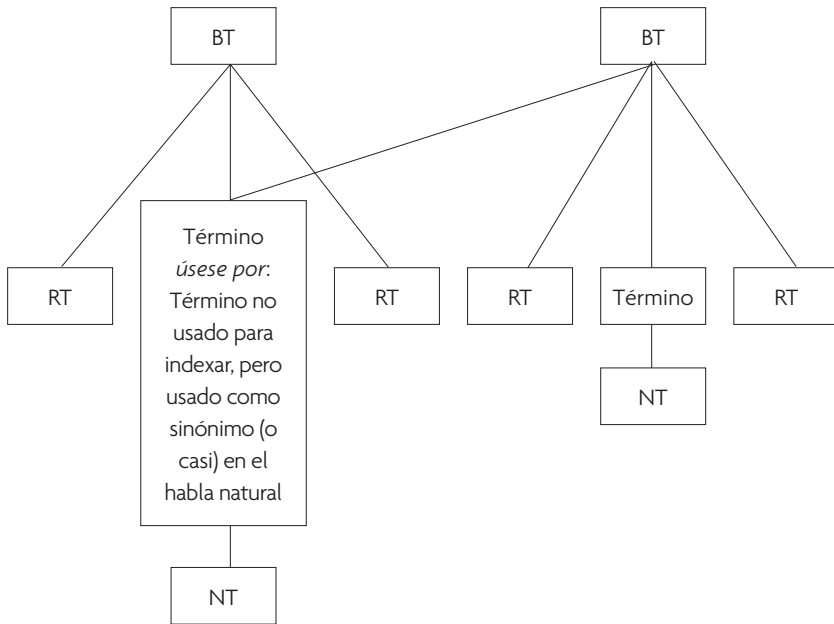
En esta fase de la elaboración de un tesauro se deben realizar al menos las siguientes acciones:

1. Decidir sobre el nivel de detalle necesario para el área del conocimiento que se quiere representar.
2. Transformar o normalizar el vocabulario extraído de los documentos a través de la indización en términos aceptados por la comunidad de usuarios meta del tesauro.
3. Distinguir de entre la misma palabra los diferentes significados, usos y aplicaciones que tiene.
4. Hacer más clara la forma de uso de un término añadiendo siempre una nota de alcance. Cabe mencionar que las notas de alcance no son definiciones más amplias de los términos, aunque en algunas ocasiones las definiciones pueden ser buenas notas de alcance. Ejemplo, la definición sicológica de *adolescencia* puede no ser útil para un indizador o para los usuarios, sin embargo si se utiliza una forma del tipo «Úsese para un(a) joven de entre 13 y 18 años», puede resultar útil a manera de nota de alcance.

Creación de referencias entre los términos

La creación de referencias en un tesauro se refiere al establecimiento de las relaciones entre los términos, las relaciones que establecen entre ellos son del tipo género-especie (TG-TE) o de jerarquía básica y de semejanza o similitud como: PADRE-HIJO o semejanza de familia. A partir del reconocimiento de estos tipos de relaciones es necesario hacer un análisis para determinar entre cuáles términos se pueden establecer más de una relación con un término general, esto sucede cuando varios tipos de jerarquías ocurren al mismo tiempo, como en el caso de COSA-TIPO de COSA y TODO-PARTE.

La creación de este tipo de referencias entre los términos no siempre se puede establecer a través de medios automáticos, entonces será necesario recurrir al trabajo humano de los documentalistas o lexicógrafos, según sea el caso del equipo de profesionales encargados de la elaboración de un tesauro.



Mantenimiento del tesauro

Para el mantenimiento de un tesauro existen en el mercado programas de cómputo que realizan esta tarea. En la investigación se mencionarán las características, ventajas y desventajas que tiene este tipo de programas.

Etapas importantes en el desarrollo de los tesauros que se realizaron con el apoyo de la informática y en contexto de la recuperación de la información

A continuación se exponen cuatro momentos o etapas importantes en el desarrollo de los tesauros que se realizaron con el apoyo de la Informática y en campo de la recuperación de la información.

El primer momento se identifica con los experimentos que realizó Gerard Salton y sus discípulos en los años cincuentas. Él creó la prueba llamada *Smart system* para llevar a cabo algunos experimentos sobre

la recuperación de la información en la que se incluía la construcción de un tesoro. Una característica del enfoque que utiliza este autor es la normalización, consistente del lenguaje para la recuperación de la información por medio del establecimiento de clases de conceptos. Salton experimentó con algunos métodos para la construcción automática de los tesauros, partiendo del análisis estadístico de los términos y el estudio de la construcción de jerarquías entre ellos.

El segundo momento es el logro de la construcción de los tesauros a partir de los conocimientos derivados de la semántica. Esto se representa por los trabajos realizados por Karen Spark-Jones en los años setenta. Ella está convencida de la utilidad del tesoro siempre y cuando éste sea organizado con base en un sistema automático de clasificación, entonces habrán encontrando así una alternativa al enfoque estadístico de concurrencia de términos desarrollado por Salton. Spark-Jones desarrolló su propia teoría sobre la clasificación semántica para la recuperación de la información. El elemento básico en la teoría de la clasificación semántica es un *run*, es decir, un grupo de formas de palabras que son intercambiables en algunos contextos. Sus experimentos sobre la recuperación de la información han sido calificados como efectivos.

El tercer momento, en los años ochenta, está representado por Edgard Fox y su enfoque de los inventarios de relaciones entre los conceptos. Él decidió que era importante considerar el uso de un amplio rango de relaciones en el proceso de la recuperación de la información. Su postura teórica se deriva de los conocimientos y experimentos realizados a partir del *modelo textual de significados* de Mel'cuk's en los años setenta. Éste contempla un gran inventario de relaciones que son de naturaleza diversa, morfológicas por ejemplo, que se dan en el eje paradigmático, esto es, relaciones entre las palabras que se expresan con una misma carga semántica o de significado. O bien otro grupo relaciones se aparece en el eje sintagmático, es decir aquellas que se expresan como un nexo entre las palabras que aparecen juntas en una oración. A partir de la implementación y uso del modelo, Fox encontró que, al combinar ambos tipos de relaciones en las preguntas que se le formulaban al sistema de recuperación de información, y utilizando el tesoro, se enriquecían y mejoraban significativamente los resultados de la búsqueda en el sistema.

El cuarto momento es el de la construcción de tesauros con la ayuda de la computadora. Éste ocurrió a partir de la primera mitad de los años setenta, en el contexto de las investigaciones realizadas para la construcción de algunos diccionarios electrónicos. Fue en los años ochentas cuando Fox y Evens llevaron a cabo algunos experimentos sobre la construcción de un tesoro para la recuperación de la información a partir de los datos almacenados en un diccionario electrónico, específicamente de *Webster's Seventh Collegiate Dictionary*. Dichos resultados no fueron satisfactorios en un inicio.

La ambigüedad en las definiciones contenidas en el diccionario electrónico fue un obstáculo importante al que se enfrentaron los investigadores que realizaron sus experimentos para la recuperación de la información a través de un tesoro electrónico. Un hallazgo que se puede derivar de dicho obstáculo para el tratamiento de la ambigüedad en las definiciones, fue el crear una herramienta de la lingüística computacional que permitiera superar el fenómeno de la ambigüedad. Sólo hasta entonces les fue claro que para la construcción automática de un tesoro se debe incluir una herramienta que permitiera hacer transparente la ambigüedad del sentido en una definición. Piénsese, por ejemplo, que permitiera distinguir e identificar la diferencia entre *pine cones* y *ice-cream cones*.

Otro avance en el tema fue el que realizó Nicoletta Calzonari en 1988, de la Universidad de Pisa, quien investigó centrándose en el problema de la ambigüedad léxica, señaló que «...el diccionario y el tesoro pueden y deben combinarse». Así, ella construyó una base de datos del léxico en italiano de más de un millón de palabras que formaron 120,000 radicales o raíces, la cual ha sido utilizada para diferentes propósitos, como la traducción automática, la generación de textos, el análisis sintáctico y la recuperación de la información. Asimismo, se han desarrollado proyectos para la recuperación de la información a partir del estudio de fenómenos semánticos como la sinonimia, y su contraparte la antinomia, así como la derivación morfológica y las relaciones morfológicas explícitas que aparecen en un diccionario.

Consideraciones finales

A la luz de los resultados de la investigación documental realizada podemos concluir que existen diferentes puntos de encuentro entre la informática y los tesauros. Estas convergencias se hacen evidentes en las definiciones que se analizaron en el estudio presentado aquí. Se establecieron las diferencias entre la manera de definir el término tesoro desde la Lingüística y la Informática.

El lugar de encuentro entre la informática y la bibliotecología son los principios básicos que comparten ambas disciplinas para la creación de un tesoro. Dichos principios no solamente norman el procedimiento, sino que regulan las relaciones conceptuales entre los términos para que de esta manera se pueda establecer una comunicación eficiente entre el usuario y el sistema de búsqueda y recuperación de la información.

Finalmente, la revisión que se presentó de cuatro etapas o momentos en los que se desarrollaron investigaciones sobre la construcción de modelos y enfoques específicos para la creación de tesauros, nos permite ver otro punto de encuentro y colaboración entre la informática y la bibliotecología orientadas a la aplicación en la recuperación de la información.

Referencias

- COLLIN, S. M. H. (1994). *Dictionary of Computing* [2ª edición] (p. 280). Cambridge: Peter Collin.
- HARTMANN, R.R.K. (2006). Thesauruses. En Brown, K. (ed.). *Encyclopedia of Language & Linguistics.*, [2a ed., vol. 12] (p 668). Amsterdam: Elsevier.
- ILLINGWORTH, V. (ed.) (1993). *Diccionario de Informática* (p. 638). Madrid: Oxford University Press/Díaz de Santos.
- VIRGA & J.P. Mesters (1994). *Diccionario de Microinformática: todas las nuevas tecnologías informáticas* (pp. 404-405). Madrid: Thompson / Paraninfo.
- MILLER, U. (1997). Thesaurus Construction: problems and their roots. *Information Processing & Management*, 33(4), pp. 481-493.
- SAYER, Berwick W.C. (1992). *A Manual Of Classification For Librarians And Bibliographers* [versión en línea]. Recuperado de <https://archive.org/details/manualofclassifi007579mbp>