Evolución histórica y tendencias observables en los tesauros

MIGUEL-ÁNGEL LÓPEZ ALONSO
Dpto. de Biblioteconomía y Documentación
Universidad Carlos III de Madrid

RESUMEN

Artículo de revisión no exhaustiva sobre el origen y desarrollo de los tesauros. Desde los años cincuenta este proceso se ha diferenciado y entremezclado con el nacimiento de los principios teóricos de la Documentación, los estudios sobre otros lenguajes documentales y la indización de documentos. Su evolución se plantea a partir de los estudios lingüísticos y los tesauros literarios del pasado, sigue con los tesauros documentales desarrollados desde mediados de este siglo, y se proyecta en los tesauros conceptuales del futuro.

SUMMARY

Literacy review, not exhaustive, on the origin and development of thesauri. From the fifty, this process has been differentiated and mingled with the birth of Information Science theoretical principles, the studies on other documentary languages and the theory of indexation. Its evolution is outlined from older linguistic studies and literary thesauri, it continues with documentary thesauri developed by mid this century, and it is projected over future conceptual thesauri.

Boletín Millares Carlo, núm. 17. Centro Asociado UNED. Las Palmas de Gran Canaria, 1998

Introducción

En estos momentos se vuelve a pensar en los tesauros como herramienta de precisión para la recuperación del conocimiento en las bases documentales de la red Internet. Nos parece por ello interesante hacer una revisión de la evolución histórica de los vocabularios controlados, especialmente a partir de 1852 cuando Roget publicara la primera edición de su reconocido *tesauro literario*[1], y de 1878 momento en el que Poole discutiera las características de su índice referidas al anterior[2].

Otro momento clave se dio en 1948 al proponer Bernier su definición de *tesauro documental* como "herramienta conceptual de relaciones entre términos de tipo postcoordinado"[3], así como al deslindarse los primeros desarrollos experimentales de los primeros tesauros operacionales en los años sesenta, para alcanzar hasta las más recientes propuestas para el desarrollo de tesauros conceptuales. Estos últimos tesauros difieren de los tesauros precoordinados compilados en los años setenta en estar compilados a partir del sublenguaje científico tomado del Lenguaje Natural en contextos concretos, y en ser utilizados principalmente en la recuperación de documentos. Se integran en los Sistemas de Gestión de la Información para *mejorar la pertinencia de las búsquedas*, debido a sus numerosas relaciones asociativas contextuales.

Estos nuevos tesauros son una superserie de sublenguaje controlado en un dominio científico específico, que se usan durante:

— El proceso de indización, como ayuda en la identificación de los conceptos, y

— en el proceso de recuperación, como fuente de nuevos términos controlados que identifiquen nuevos conceptos y aumenten la precisión de las búsquedas booleanas.

---

[1] Roget, P. M.: *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expresion of Ideas and Assists in Literary Composition*. London: Longman, 1852.

[2] Poole, W. F.: "The plan of the new Poole's Index", *Library Journal*, 3 (3), 1878, pp. 109-110.

[3] Bernier, C. L. y Crane, E.: "Indexing abstracts", *Industrial Engineering Chemistry*, 40 (4), 1948, pp. 725-730.

Antecedentes en los glosarios científicos y en los estudios lingüísticos

Para encontrar los antecedentes más remotos de los tesauros actuales, como instrumentos de control que son de la terminología utilizada para la indización y/o recuperación de los documentos, es preciso investigar en los glosarios incorporados en algunas de las Recopilaciones Jurisprudenciales del Período Alejandrino, en los primeros siglos de nuestra era. Será mucho después, con el Renacimiento y más extensamente con la Ilustración, cuando se retome el estudio de las terminologías jurídicas dentro de los Derechos Romano o Canónico, todavía en lengua latina.

Su desarrollo en la época moderna vendrá precedido por los primeros estudios lingüísticos con esquemas clasificatorios sistemáticos. En España, el «*Libro de los Epítomes*» y el «*Libro de Proposiciones*» de Hernando de Colón de 1530[4], en Francia, el «*Dictionnaire historique et poétique de toutes les nations, nommes, lieux, fleuves...*» de Charles Estienne de 1553[5].

En la historia europea habrá que esperar a las Codificaciones Jurídicas Napoleónicas y el posterior resurgimiento de los nacionalismos de mediados del siglo XIX para que, con el abandono de los preponderantes derechos generales en latín y el arraigo definitivo de los Derechos Comunes y Forales, se precise conocer en profundidad los vocabularios jurídicos en las diferentes lenguas vernáculas.

A comienzos del siglo XVIII, en Francia, Girard trató de dar solución a las dudas que presenta el empleo de términos afines en su libro «*Justesse de la langue française*» de 1718[6], que orientará en Europa el sucesivo tratamiento de la sinonimia, como intento de editar, para cada lengua nacional, libros que fijen el valor exacto de las distintas palabras con el mismo significado. Por su parte, en Alemania (Leipzig) Gottsched publicó su libro «*Observaciones sobre el uso y abuso de varios términos de la Lengua Alemana*» en 1748[7].

Crabb editó, en Inglaterra, su «*Dictionary of English Synonymes Explained*» en 1824[8], libro que todavía se reedita con adiciones y puestas al día de diversos autores. El médico Roget publicará en 1852 su tesauro literario, el «*Thesaurus of English Words and Phrases, classified and arran-*

---

4 De Colón, Hernando: *Libro de Proposiciones*. Sevilla, 1530.

5 Estienne, Ch.: *Dictionnaire historique et poétique de toutes les nations, nommes, lieux, fleuves...* Francia, 1553.

6 Girard, G.: *Justesse de la langue française*. París, 1718. Reeditado en 1741 con el título «Synonimes françois».

7 Gottsched, J. C.: *Grundlegung einer deutschen Sprachkunst*. Leipzig, 1748.

8 Crabb, J.: *Dictionary of English Synonymes Explained* (3ª ed.). Inglaterra, 1824.

ged so as to facilitate the expresion of ideas and assists in literary composition», del cual se han hecho más de treinta ediciones[9], incluso una reciente para su consulta en línea con el nombre de *«Roget's Thesaurus»*. Fue concebido como un esquema clasificatorio dividido en seis grandes categorías: conceptos abstractos, espacio, materia, formación y comunicación de ideas, intereses socio individuales y aficiones, que guardan cierta afinidad con las posteriores categorías de Ranganathan de personalidad, energía, espacio, materia y tiempo[10].

En España, Manuel Dendo y Ávila con su breve *«Ensayo de los synónimos»* en 1757[11] y, en Viena, José López Huerta con el *«Examen de la posibilidad de fixar la significación de los sinónimos de la lengua castellana»* en 1789[12] (reeditado en España varias veces), estimularon la afición por los estudios sinonímicos. A éstos siguieron los destacados trabajos de José Joaquín de Mora en la *«Colección de Sinónimos de la Lengua Castellana»* de 1855[13] y el, ya contemporáneo, de Richard Ruppert con *«Spanische Synonimik»* de 1940[14].

Nacimiento y desarrollo de los tesauros documentales

Después de la II Guerra Mundial, prestigiosos investigadores fundaron serias esperanzas en el acceso a la información de manera directa y no secuencial[15] (el desarrollo del primer sistema de indización con un tesauro incorporado, los unitérminos de Taube)[16], la «automatización del análisis de los documentos» como vía para la profundización en la indización por conceptos (los descriptores de Mooers)[17] y el

---

9 Roget, P.: *Roget's International Thesaurus*. New York: Thomas Cronwell, 1977.

10 Ranganathan, S. R.: *Prolegomena to library classification*. 3ª ed., Bombay: Asia Pub. House, 1967 , 232 pp.

11 Dendo y Ávila, Manuel: *Ensayo de los sinónimos*. Madrid, 1757.

12 López Huerta, J.: *Examen de la posibilidad de fixar la significación de los sinónimos de la lengua castellana*. Viena, 1789.

13 De Mora, J. J.: *Colección de Sinónimos de la Lengua Castellana*. Madrid, 1855.

14 Ruppert, R.: *Spanische Synonimik*. Heilderberg, 1940.

15 Moreiro González, J. A.: "De la Documentación a la Ciencia de la Información: evolución de los conceptos y aplicaciones documentales", *Seminario de Humanidades Agustín Millares Carlo, Homenaje a Antonio de Bethencourt Massieu*, 1995, p. 10.

16 Formado por términos del Lenguaje Natural: unitérminos y sus relaciones paradigmáticas, a partir del fichero invertido de Mooers.

Taube, M.: *Studies in coordinate indexing*. Washington: Documentation Incorporated, 1953.

17 Mooers, C. N.: «The Theory of Digital Handling of Non-Numerical Information and Its Implications to Machine Economics», *Zator Technical Bulletin*, 48, 1950.

empleo de los índices permutados tipo KWIC («keyword in context») de Luhn[18].

Serán tiempos de consolidación de los principios teóricos atribuidos a renombrados investigadores: Taube, Howerton, Mooers, Brownson, Farradane, Jolley, etc., a través del análisis de los conceptos más recientemente acuñados: recuperación, tesauro, descriptor, resumen, relevancia, indización, etc.

Aunque las primeras referencias escritas se atribuyen a Peter Luhn[19], Bernier y Heumann[20], y Joyce y Needham[21]; fue Eugenio Wall[22] quien definió los principales contenidos lingüísticos de los tesauros documentales: sintaxis, semántica, género, sentido, etc. Y fue Helen Brownson quien utilizó el término «tesauro» por primera vez el 14 de mayo de 1957 en la Dorking Conference, citándolo como una herramienta para la recuperación de la información[23].

Desde finales de los cuarenta, con el desarrollo de los primeros ordenadores que facilitaban el procesamiento de la indización coordinada, comenzaron en Estados Unidos las investigaciones para el desarrollo de nuevos lenguajes documentales codificados que reemplazaron a los tradicionales esquemas clasificatorios.

Aunque no se documentara o discutiera ampliamente, el primer trabajo con un tesauro experimental lo realizó Whelan en el Royal Radar Establishment, Malvern (Inglaterra, 1955)[24]. Al año siguiente, la Unidad de Investigación Lingüística de Cambridge, CLRU (Inglaterra, 1956), avanzó sus hipótesis para mejorar la traducción automática de la información: la aplicación del concepto de tesauro en la combinación de los descriptores y el uso de la lógica booleana[25].

El primer sistema documental diseñado para localizar extensos textos legales se dio en el Centro Jurídico de la Universidad de Pittsburg. Utili-

---

[18] LUHN, H. P.: «A statistical approach to mechanized encoding and searching of literary information», *IBM Journal of Research and Development*, 1(4), 1957, pp. 309-317.

[19] *Ibidem*.

[20] BERNIER, C. L. y HEUMANN, K. F.: «Correlative indexes. III. Semantic relations among semantemes - the technical thesaurus», *American Documentation*, 1957, 8, pp. 211-220.

[21] JOYCE, T. y NEEDHAM, R. M.: «The thesaurus approach to information retrieval», *American Documentation*, 9, 1958, pp. 192-197.

[22] WALL, E.: «Information systems», *Chemical Engineering Progress*, 1959, 55, pp. 55-59.

[23] BROWNSON, H.: *Proceedings of the International Study Conference on Classification for Information Retrieval*. ASLIB, 1957, pp. 99-100.

[24] WHELAN, S.: "Library retrieval", *RRE Journal*, 42, october 1958, pp. 59-68.

[25] MASTERMAN, M. "Potentialities of a mechanical thesaurus", *MIT Conference on Mechanical Translation*, CLRU Typescript, 1956.

zaba como tesauro una mera compilación de términos con significados similares (unitérminos), que se parecía más al Roget's Thesaurus de principios de siglo que a los tesauros documentales de la actualidad.

El primero de estos tesauros a escala completa, totalmente operacional y con vistas a su utilización en la recuperación automatizada, se utilizó por la sociedad E.I. Du Pont Nemours and Co., Inc. en 1959 (Wilmington, USA)[26]. El primero de los publicados fue el ASTIA (de la Armed Services Technical Information Agency, USA, 1960) que luego se reconvirtió en el TEST (del Defense Documentation Center, 1967)[27], que comprendió 17.800 términos preferentes, 5.554 relaciones y numerosos términos asociados. Otro de los primeros ejemplos fue el del American Institute of Chemical Engineers en 1961[28].

Uno de los más visibles avances los dio Barhydt y su grupo de investigación de la Universidad Western Reserve cuando desarrollaron un tesauro de términos educacionales que utilizó el análisis de facetas puro. Con alrededor de 4.500 términos lo enviaron a ERIC (Office of Education, 1966) y fue rechazado, aunque, más tarde fue revisado y publicado por el Servicio de Publicaciones de la citada Universidad.

Modelos facetados mixtos

La primera solución a las dificultades que planteaba un tesauro de facetas puro la adelantarían Aitchison y sus colaboradores de la English Electric al presentar en 1969 el primer "Tesaurofacetas"[29], que adoptaría una *solución mixta* mucho más eficaz al aunar la clasificación por materias con una subdivisión por facetas que comprendía dos entradas:

— la relación alfabética de los descriptores del tesauro, con las relaciones TE, TG, TR, y el reenvío por un código de tres caracteres a la parte facetada, y

---

[26] HOLM, B. E. y RASMUSSEN, L. E.: «Development of a technical thesaurus», *American Documentation*, 1961, 12, pp. 184-190.

[27] ENGINEERING JOINT COUNCIL Y DEPARTMENT OF DEFENSE: "Thesaurus of Engineering and Science terms. A list of engineering and related scientific terms and their relationships for use as a vocabulary reference in indexing and retrieving technical information". Nueva York: EJC, 1967.

[28] AMERICAN INSTITUTE OF CHEMICAL ENGINEERS: "Chemical engineering thesaurus: a wordbook for use with the concept co-ordination system of information storage and retrieval". Nueva York: AIChE, 1961.

[29] AITCHISON, J. *et al.*: *Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects*, Whetstone, UK: The English Electric Co. Ltd, 1969.

— la clasificación por materias, en que los términos son reagrupados en facetas o funciones fundamentales.

Con este modelo se localiza el contexto en que se ha elaborado el tesauro y se integran en un solo sistema las ventajas de los tesauros alfabéticos y una nueva presentación sistemática de los términos, que tiene en cuenta también las relaciones entre las propias jerarquías o categorías. Analiza los términos de un campo temático en clases o conjuntos con una característica común, según los tipos básicos de funciones principales o facetas que representan, y abandona los campos de interés por disciplinas científicas usados tradicionalmente. Los objetos concretos pueden subdividirse en facetas, ordenadas en un orden lógico desde la más general a la más especializada y compleja[30], mientras que los campos temáticos se dividen por disciplinas: Agricultura, Medicina, Economía, etc.

Este modelo facetado mixto constituyó la aportación de la doctrina de indización americana al desarrollo de los tesauros, pero, exigía una disciplina mental muy rigurosa del compilador para que se crearan estructuras dignas de crédito[31]. Tiene como sus ejemplos más recientes el BSI ROOT Thesaurus[32], el Art and Architecture Thesaurus y el International Thesaurus of Refugee Terminology.

En la década de los sesenta y en la siguiente, se multiplicaron los *estudios sobre Lenguajes Documentales*:

— Spark Jones y Needham estudiaron los vocabularios controlados precoordinados, dentro de las investigaciones sobre procedimientos de clasificación automática[33],

— Coyaud elaboró una estructura teórica para el análisis de los lenguajes de indización, cuyos constituyentes están tomados de la terminología lingüística (fonemas, semas, etc.)[34],

— Dahlberg emprendió una búsqueda interdisciplinar de la «noción de clasificación», y rebasó el campo de la Biblioteconomía con la

---

[30] Situaciones ideales con existencia independiente que no son parte de un determinado texto: entidades, partes, propiedades, acciones (procesos y operaciones), agentes, aplicaciones, etc.

[31] AITCHISON, J.: «Thesaurofacet: a new concept in subject retrieval schemes», *Proceedings of an international symposium*, University of Maryland, 1971. Westport(Ct.): Greenwood Press, 1972, pp. 72-98.

[32] BRITISH STANDARDS INSTITUTION: *BSI ROOT thesaurus*, Milton Keynes: BSI, 1981.

[33] NEEDHAM, R. M., SPARCK JONES, K.: «Keywords and Clumps», *J. Documentation*, 1964, 20 (1), pp. 5-15.

[34] COYAUD, M.: *Introduction à l'étude des langages documentaires*. París: Klinchksieck, 1966, 148 pp.

Filosofía, la Epistemología, la Lingüística, las Teorías Científicas, etc.[35],

— Hutchins desarrolló una profunda introducción a las «Estructuras Lingüísticas Generales de los Lenguajes de Indización comparados con los Lenguajes Naturales» en sus diversos aspectos formales, semánticos, pragmáticos, etc.[36]

Estado actual de los tesauros

A partir de los años setenta, los especialistas tomaron conciencia de los inconvenientes del crecimiento desmesurado de los fondos documentales, de la parcelación del saber y de la expansión del léxico, y, por tanto, de la necesidad de poner al día los lenguajes documentales para una precisa indización y recuperación relevante de los primeros. La informática se difundió y aparecieron en el mercado ordenadores, cada vez más manejables y económicos, que potenciaban las tecnologías de la información y obligaban al desarrollo de los tesauros documentales para la automatización del procesamiento de la información.

Modelos lingüísticos y matemáticos

Desde Ferdinand de Sausurre[37], Bernard Pottier[38] o M. Coyaud[39], las razones que han empujado a los documentalistas a interesarse por las teorías lingüísticas han sido numerosas, dado que las propiedades de los Lenguajes Documentales se parecen mucho a las de los Lenguajes Naturales y derivan de éstos últimos más o menos profundamente.

Uno de los primeros *modelos lingüísticos* fue el del «Triángulo Semántico» de Ogden y Richards (1923)[40], modificado profundamente por Long

---

[35] Dahlberg, I.: *Grundlagen universaler Wissensordnung*. München: Verlag Dokumentation, 1974, 366 pp.

[36] Hutchins, W. J.: *Lenguages of indexing and classification: a linguistic study of structures and functions*. Librarian-ship information studies, 3). Stevenage (Herts.): Peregrinus, 1975, 148 p.

[37] Saussure, F. de: *Curso de Lingüística General* (Trad. del francés: Paris: Fayot, 1916). Barcelona: Planeta-Agostini, 1985.

[38] Pottier, B.: *Lingüística general, teoría y descripción*. Madrid: Gredos, 1976, 426 p.

[39] Coyaud, M.: *Introduction à l'étude des langages documantaires*. *Ibid.*, cit. n.° 18.

[40] Ogden, C. K., Richards, I. A.: *The meaning of meaning; a study of the influence of language on thought and of the science of symbolism*. Nueva York: Harcourt, Brace & Company, Inc., 1923.

(1980)[41], en el que los tesauros estaban constituidos por relaciones entre los conceptos (significados), los objetos (referentes), las expresiones o significantes gráficos y los significantes fonológicos.

La modelización matemática proporciona bases sólidas a los Lenguajes de Indización, al definir rigurosamente los términos empleados: clases, descriptores, relaciones, etc. Puede cubrir bien todo el Sistema de Información, un tipo de Lenguaje de Indización o un aspecto concreto de uno de éstos[42]. La función matemática empleada se suele derivar de la Teoría de Conjuntos, cuya terminología recuerda a la de los Lenguajes de Indización.

Entre estos últimos modelos, destaca el desarrollado y más tarde sintetizado por D. Soërgel (1985), que cubre todos los Lenguajes de Indización, en el que:

a) Se describe una base lógico-matemática general para la construcción de un sistema formal que distinga claramente entre el metalenguaje matemático (interpretación del modelo) y el lenguaje documental que representa (significado del modelo), y
b) se da un léxico y unas reglas de formación de expresiones de las que se deducen teoremas al nivel del metalenguaje matemático[43].

Hasta casi finales de los ochenta se ralentizó el desarrollo de nuevos tesauros, debido en parte al descenso en el ritmo de evolución de los ordenadores personales y a la creencia de que, aunque las tecnologías de la información estaban cambiado radicalmente, los tesauros podían permanecer inalterables. Se produjo su resurgimiento con el auge de las búsquedas en línea, pero como herramienta de precisión para la formulación de las ecuaciones de búsqueda más que como herramienta de indización.

Robert Fugmann (1974) exportó la estructura y las propiedades de la Teoría de los Grafos al Sistema de Indización Bidimensional TOSAR, y trató de definir tanto las relaciones semánticas como otras de orientación, distancia entre términos, etc.[44] En los tesauros tradicionales propuso los

---

[41] LONG, B.: «Linguistique et indexation», *Documentaliste-Sciences de l'Information*, 1980, 17 (3), pp. 99-106.

[42] VICKERY, B. C.: *Retrieval language models. Information systems*. London: Butterworths, 1973, pp. 203-222.

[43] SOËRGEL, D. (1985): «Index Language Structure. I: Conceptual», en *Organizing information: principles of data base and retrieval systems*. Orlando (Fl.): Academic Press, 1985, p. 269.

[44] FUGMAN, R. *et al.*: «Representation of the concept relations using the TOSAR system in the IDC», *Journal of the ASIS*, 1974, 25 (5), pp. 287-307. Cfr. 1.2.2.5.

esquemas de flechas para representar los descriptores y sus relaciones semánticas. Sintetizó su Teoría de la Indización en cinco axiomas que incluyen una dimensión ética, *el concepto de indización imperativa,* y obliga al indizador a elegir el término más apropiado, en contraposición con la indización habitual que dejaba la posibilidad de escoger varios términos, más o menos adecuados[45].

Soërgel avanzó su nueva definición de Lenguaje de Indización como:

> conjunto de descriptores, de relaciones y de reglas para la formación de expresiones condensadas del documento original, con la finalidad de reducir el volumen de datos de dicho texto[46],

y distinguió en el tesauro, términos no descriptores que conducen a los términos descriptores.

La indización se convirtió en multimodelo y se validó el mismo descriptor desde contextos y aproximaciones diferentes[47], a pesar de ser diferentes los usuarios de las distintas áreas del conocimiento. Para ello, los tesauros tradicionales fueron redefinidos para incorporar los avances más recientes de campos como la Lingüística, la Inteligencia Artificial, las Técnicas de Programación o el Diseño Informático.

Modelos semánticos o conceptuales

La modelización de los principios teóricos que presiden la estructura de los tesauros ha seguido preferentemente los modelos lingüísticos o matemáticos, sin embargo, las recientes teorías giran alrededor de *la noción del motivo o materia* de la que tratan los textos, es decir, del concepto semántico o conceptual.

Maniez propugnó (1976) un modelo de tesauro en el que las relaciones no sean lingüísticas: paradigmáticas (pertenecientes a la lengua, fuera de todo contexto), o sintagmáticas (pertenecientes al discurso, integradas en su contexto), sino extrasemánticas o asociativas; de forma que aúnen términos y conceptos reales por su similitud de sentido en el contexto específico del usuario[48]. En su libro de síntesis «*Los Lenguajes Documentales y de Clasificación...*» (1987), parte de la oposición entre tema

---

[45] FUGMANN, R.: «The five-axiom theory of indexing and information supply», *Journal of the ASIS*, 1985, 36 (2), pp. 116-129.

[46] *Ibidem*.

[47] Subdividiendo los conceptos por facetas, según características particulares comunes a un grupo de ellas.

[48] MANIEZ, J.: *Los lenguajes documentales y de clasificación*. Madrid: Pirámide, 1993, p. 214.

y comentario (es decir entre «de lo que se habla» y «lo que se dice en ese hablar»), propia de los lingüistas como Chomsky[49], para concluir que:

> La tematización por medio de los Lenguajes Documentales es una actividad informativa esencial, mientras que la enunciación tiene poco valor documental.[50]

Dewèze formalizó (1981) la representación de las relaciones semánticas, con la adopción de una *teoría de red semántica extralexical* que situó a un nivel superior al de los lenguajes naturales, en la perspectiva de construir tesauros multilingües. En esta teoría, un significado se define como «un conjunto de semas a los que se pueden atribuir relaciones lexicales en varios idiomas»[51].

Las relaciones de los semas se describe con la Teoría de los Grafos que representa las diferentes configuraciones sémicas. En un sistema documental semántico una materia se representa por un grafo, los conceptos son las cumbres y las relaciones son los arcos. Los "parámetros de demanda" se representan también mediante un grafo, cuyos arcos y cumbres son más o menos precisos. En las búsquedas documentales un programa compara el grafo de la demanda con los grafos de los documentos registrados en memoria, y retiene aquellos que contienen conceptos con estructura más parecida.

Tomando el concepto de red semántica de Dewèze, Schaüble (1989) propone una nueva estructura de la información, el Espacio Conceptual. Y construye una teoría de los Tesauros Conceptuales, como sistema formal a partir de la lógica matemática del dominio algebraico, que revela una estrecha relación entre los tesauros y el modelo espacial multidimensional, en la que las relaciones entre términos son definidas con más precisión que en los tesauros jerárquicos.

Producción de tesauros documentales en lengua española

En España, se aprecian dificultades para la compilación de tesauros, aunque, los contados ejemplares de finales de los setenta se convertieron en cerca de 50 en 1984 y alcanzaron la cifra de 187 a finales de 1988[52].

---

[49] CHOMSKY, N.: *Aspects of the Theory of Syntax*. Cambridge: MIT, 1965.

[50] *Ibidem*, pp. 205-208.

[51] DEWÈZE, A.: *Réseaux sémantiques: essai de modélisation; aplication à l'indexation et à la recherche documentaire*. Lyon: Universidad Claude Bernard. Tesis doctoral, 1981.

[52] ÁLVARO BERMEJO, C. *et al.*: «Evaluación de los Tesauros Disponibles en Lengua Española», *Revista Española de Documentación Científica*, 1989, 12 (3), pp. 283-297.

No existió correspondencia entre la carrera por la automatización en la mayoría de las Instituciones y la elaboración de los vocabularios controlados para una adecuada Recuperación Documental. Fueron pocas las Instituciones que acometieron esta tarea con el adecuado rigor científico y proliferaron microtesauros excesivamente específicos o glosarios que no alcanzaron a cubrir los requisitos de los Lenguajes Controlados.

La mayor continuidad y rigurosidad se percibió en el ICYT y en el ISOC, ahora fusionados en el CINDOC del CSIC, que, como antiguos institutos de información y documentación especializados en ciencias puras y humanas, respectivamente, tuvieron la oportunidad de atender las peticiones de compilación de tesauros venidas de muy variadas instituciones nacionales y extranjeras[53].

Conclusiones y prospectiva

1.ª) En la actualidad, *los sublenguajes científicos especializados* tienden a generarse automáticamente a partir del procesamiento del lenguaje natural de los documentos, y proporcionan alternativas a los términos de los usuarios durante las búsquedas.

Las bases de conocimientos terminológicas están formadas por tablas, con nombres y números de clasificación, que pueden visualizarse para la selección de los términos en los interfases de usuario, e incorporarse dentro de una estrategia de búsqueda. Estos lenguajes controlados pueden tener una mínima estructura de referencias cruzadas o de tablas correlacionadas, e incluso incluir tesauros multilingües para su uso en las Bases de Datos Internacionales.

Una de las herramientas de este tipo es la Base de Datos TERM, desarrollada por los Servicios de Recuperación Bibliográfica (BRS), compuesta de tablas de conceptos que incluyen términos controlados y texto libre[54]. Otro ejemplo es el Diccionario Experimental del Consejo de Europa, realizado por Universidades de Inglaterra, Alemania, Italia, Holanda, España y Servia desde 1988, que ha pasado a formar parte del Banco de Datos Terminológico EURODICAUTOM en la Dirección General XIII de la CEE.

Dado que en la compilación terminológica debe mantenerse un equilibrio de esfuerzos entre los procesos de diseño y utilización, previamente deberá establecerse una clara diferenciación entre los sistemas de re-

---

53 Gil Urdiciain, B.: «Orígenes y evolución de los tesauros en España», *Rev. General de Información y Documentación*, 1998, 8 (1), pp. 64-110.

54 Knapp, S. D.: «Creating BRS/TERM, a vocabulary database for searchers», *DATABASE*, 1984, 7(4), pp.70-75.

cuperación que utilizan tesauros, cuyos costes se producen al desarrollarlos, y los que procesan contextualmente el lenguaje natural de los documentos, cuyos costes se producen al realizar la búsqueda. Estos segundos se utilizan preferentemente en la recuperación de documentos, dado que convierten el lenguaje natural del usuario al sublenguaje científico de los textos, mediante los MAI (machine-aided indexing).

2.ª) Los Tesauros Documentales Conceptuales, diseñados específicamente para ayudar en la enunciación de las preguntas en la fase de Recuperación de la Información, propuestos por autores como Bates[55], Schmitz-Esser[56] o Milstead[57], palian en parte la indeterminación de las búsquedas en Lenguaje Natural, especialmente en aquellas Bases de Datos cuyos documentos con texto completo no han sido previamente indizados con ningún otro tesauro.

Se considera con Kristensen[58] y Larsson[59], que las recuperaciones que utilizan vocabularios postcontrolados, generados automáticamente a partir del Lenguaje Natural, obtienen muchas de las ventajas de los Lenguajes Controlados tradicionales y evitan algunos problemas lexicales de su uso directo: sinonimias, homografías, etc.

Diversos experimentos en que los usuarios son apoyados, en la enunciación de sus ecuaciones de búsqueda, con términos adicionales extraídos de un tesauro diseñado específicamente para la recuperación con Lenguaje Natural en grandes Bases de Datos; han aportado avances significativos en el conocimiento intrínseco de la relevancia de las recuperaciones[60], y se ha llegado incluso a doblar la precisión en el número de documentos recuperados si el usuario selecciona y usa los términos sugeridos por un tesauro como adicionales a sus propios términos[61].

---

[55] BATES, M. J.: «Subject Access in Online Catalogs: A Design Model», *Journal of ASIS*, 1986, 37 (6), p. 361.

[56] SCHMITZ-ESSER, W.: «New Approaches in Thesaurus Application», *International Classification*, 18 (3), 1991, pp. 143-147.

[57] MILSTEAD, J. L.: "Invisible Thesauri: the year 2000", *ONLINE & CDROM Review*, 1995, 19 (2), pp. 93-94.

[58] *Ibidem.*

[59] FREMER, E., LARSSON, B.: "SPIRS, WinSPIRS, and OVID: a question of free-text versus thesaurus retrieval?", [carta), *Bull. Med. Assoc.*, 1997, 85 (1), pp. 57-58.

[60] CROFT, W. B. y DAS, R.: «Experiments with query acquisition and use in document retrieval systems», en *Proceeedings of the 13th Conference on Research and Development in Information Retrieval*, Brussels, Belgium, 1990, sept.

KRISTENSEN, J.: «Expanded End-user's Query statements for free text searching with a search-aid thesaurus», *Information Processing & Management*, 1993, 29 (6), pp. 733-744.

[61] EKMEKCIOGLU, F. C., ROBERTSON, A. M. y WILLET, P.: «Effectiveness of query expansion in ranked-output document retrieval systems», *Journal of Information Science*, 1992, 18, pp.139-147.

A pesar de que los tesauros existentes se utilizan poco por la Lingüística Computacional o la Ingeniería del Conocimiento, en la búsqueda de soluciones para el procesamiento del lenguaje natural, se ha detectado un resurgimiento en sus principios que, como herramienta conceptual bien conocida y establecida, les obliga a incorporar aquellas relaciones que faciliten la adaptación a sus nuevos usuarios (los agentes expertos o MAI) y a sus nuevas técnicas, y sustituyan los criterios predominantes en los expertos humanos[62].

[62] LÓPEZ ALONSO, M.-A.: "Un Tesauro Conceptual para la recuperación de la información jurídica comercial", *Revista Española de Documentación Científica*, 1998, 21 (2), pp. 164-173.