Revista General de Información y Documentación
Vol. 8, n.° 2 - 1998: 193-199
ISSN: 1132-1873

EVALUACIÓN SEMÁNTICA Y ESTRUCTURAL DE TESAUROS

BLANCA GIL URDICIAIN
Escuela de Biblioteconomía y Documentación
Universidad Complutense de Madrid

Resumen: El artículo presenta un análisis de las características semánticas y estructurales de una muestra de tesauros españoles, cuyos resultados revelan su capacidad para recuperar información relevante.

Palabras clave: Evaluación de tesauros, Recuperación de información, Relevancia, Pertinencia, Utilidad.

Abstract: The article presents an analysis of the semantic and estructural characteristics of a sample of Spanish thesauri, the results of which point out their capacity to retrieve relevant information.

Key words: Evaluation of thesauri, Information retrieval, Relevance, Pertinence, Utility.

La recuperación de información en bases de datos es un proceso interactivo en el que intervienen numerosas variables. Se trata de una compleja operación cuyo éxito depende, entre otros factores, del sistema de representación documental, de la calidad de la indización y de la propia organización de la colección o base de datos. La dificultad para localizar la documentación que se encuentra en determinado Servicio de Recuperación de Información (SRI) puede, por lo tanto, ser debida a muchas causas: desde errores cometidos al indizar los documentos, en la formulación de estrategias de búsqueda, o en las características mismas del lenguaje de recuperación utilizado. A todo esto hay que añadir la capacidad del usuario para expresar adecuadamente su necesidad de información, quien, a veces, no tiene una idea exacta de la información que precisa, por lo que consul-

ta, en ocasiones, en más de una base de datos y en una o varias sesiones, hasta depurar la consulta, para la que no siempre es seguro que encuentre una respuesta adecuada.

Por las razones apuntadas, la valoración de un SRI no puede limitarse a la comparación de los resultados de una búsqueda determinada con la cuestión formulada, sino que implica el estudio de todos los detalles que conforman el proceso y que son, en definitiva, los que influyen en la cantidad y calidad de la información recuperada.

Los parámetros que se tienen en cuenta para evaluar dicha información son relevancia y pertinencia, que se confunden a menudo. La relevancia hace referencia a la valoración que realizan una o varias personas en relación con una determinada solicitud de información, encontrando coincidencia entre pregunta y respuesta. Pertinencia sería la valoración que hace el usuario de una respuesta dada por un SRI a una necesidad concreta de información, formulada por él mismo. Es decir, una respuesta es pertinente si es relevante y además apropiada para la persona que formuló la pregunta. Álvarez Ossorio[1] define ambos criterios de forma breve: *Existe relevancia cuando hay adecuación a la petición de información; existe pertinencia cuando la respuesta se adecúa a la necesidad de información.* Puede, como es natural, haber coincidencia en cuanto a las estimaciones de los dos índices, pero existe un sutil matiz diferenciador entre ambos: un documento considerado pertinente, no necesariamente puede ser valorado como relevante por una persona ajena a la formulación de la cuestión. Algunos especialistas (Cooper[2], Soergel[3] y Fugmann, entre otros) tienen en cuenta además el factor utilidad. Fugmann considera que *un documento es útil si es pertinente y aporta al usuario más información de la que conocía. La utilidad* —añade— *debe medirse en términos monetarios (¿Hasta qué punto le merece la pena al usuario conseguir el documento?)*[4]. Un documento, puede, en efecto, responder al ti-

---

[1] PÉREZ ÁLVAREZ OSSORIO, J. R.: *Introducción a la Información y Documentación Científica*. Madrid, Alhambra, 1988, p. 62.

[2] COOPER, W. S.: "On selecting a measure of retrieval effectiveness. Part. I. Philosophy", *Journal of the American Society for Information Science*, n.° 24 (1973), pp. 87-100. "Part II. Implementation of the philosophy", *Journal of the American Society for Information Science*, n.° 74 (1973), pp. 413-424.

[3] SOERGEL, D.: "Indexing and retrieval performance: The logical evidence", en *From classification to "knowledge organization". Dorking revisited or "past prelude"*. Edited by Alan Gilchrist. The Hague, International Federation for Information and Documentation (FID) 1997, p. 86.

[4] FUGMANN, R.: *Subject Analysis and Indexing: Theoretical Foundation and Practical Advice*. Frankfurt/Main, Indeks Verlag, 1993, p. 46.

po específico de pregunta planteada por el usuario y, no obstante, no ser útil por varias razones:

- *Por no ser de su interés*
- *Por ser similar a los que ya tiene en su poder*
- *Estar publicado en una revista de difícil o imposible localización*
- *Estar escrito en un idioma de difícil acceso*
- *Puede que la formulación de la pregunta se haya realizado de forma demasiado general por el usuario y, por tanto, haber causado una recuperación demasiado imprecisa*[5].

No siempre es posible controlar a priori —antes de la etapa de recuperación— todos los elementos que intervienen en el proceso documental y ofrecer al usuario información útil, pero sí algunos de ellos. Hay, por ejemplo, como indicaba anteriormente, muchas características de la indización que afectan a la recuperación. De los recursos estructurales y sintácticos que se pueden utilizar en dicha operación vamos a tomar en cuenta aquí sólo los primeros, y de ellos, el más importante, es decir, aquel derivado de la estructura del lenguaje de indización utilizado para la representación documental. Vamos a centrar la atención concretamente en las características semánticas y estructurales del tesauro, prototipo de lenguaje controlado empleado en bases de datos especializadas, ya que la evaluación de los tesauros utilizados para indizar y recuperar documentos antes de su puesta en funcionamiento o en el momento de su actualización periódica favorece, en buena medida, la obtención de resultados relevantes, pertinentes y de utilidad.

Los aspectos a tener en cuenta en la valoración varían según las fuentes, pero por lo común, coinciden con los señalados por Lancaster[6], a saber: composición, tamaño, relaciones de equivalencia, tasa de enriquecimiento, reciprocidad de las relaciones de equivalencia, jerárquicas y asociativas, cantidad de notas aclaratorias registradas por tesauro, número de niveles jerárquicos, morfología de las palabras, aspectos estéticos de composición y tipografía, y nivel de precoordinación.

Para valorar la composición se tendrá en cuenta si en la estructura del tesauro figuran, al menos, dos de las formas de presentación básicas: alfa-

---

[5] FUGMANN, R., *op. cit.*, p. 46.

[6] LANCASTER, F. W.: "Evaluación de los tesauros", en *El control del vocabulario en la recuperación de información*. Valencia, Universidad, 1995, pp. 173-174.

bética y sistemática o alfabética y gráfica. También se comprobará si incluye algún índice permutado auxiliar y una introducción explicativa de sus características y ámbito de aplicación. El tamaño hace referencia al número de descriptores y no descriptores que componen la terminología.

Para conocer el volumen de relaciones de equivalencia y hallar la tasa de enriquecimiento, que mide cuantitativamente la proporción entre el número de relaciones jerárquicas y/o asociativas y el número total de descriptores, se procede al recuento de todas las relaciones de este tipo. Los valores aconsejados en ambos casos se sitúan entre el 0,5 y 2 para los casos de sinonimia, y entre 2 y 5 relaciones jerárquicas y/o asociativas por descriptor. En cuanto a las notas aclaratorias, históricas o de aplicación, constituyen un valor cuantificable numéricamente, cuantas menos incluya un tesauro, mayor será el número de descriptores que puedan generar ambigüedad. Si en el lenguaje incluyen notas menos del 10 por ciento de los descriptores, se puede pensar que hay ambigüedad notable.

El número de niveles jerárquicos permite conocer el grado de especificidad de los términos que componen el lenguaje documental; se trata de un valor relativo en función de la especialización de la materia. La morfología de las palabras se refiere al uso que se hace en el tesauro de las formas de las palabras, del empleo del singular y el plural y de si los descriptores se expresan por medio de entradas directas, sin inversión de términos, respetando el orden natural de las expresiones. En cuanto a los aspectos estéticos de composición y tipografía, se evalúan cualidades formales tales como tipo de letra utilizado para representar los términos, sangrados, empleo de recursos para destacar los términos preferentes de los no preferentes, etc...

Por último, la tasa de precoordinación (t), o número medio de palabras significativas por descriptor, cuyos valores aconsejados oscilan entre 1'5 y 2, se obtienen a través del control de la frecuencia de los descriptores unitérminos (a), bitérminos (b), tritérminos (c), etc., y la aplicación de la siguiente fórmula:

$$t = \frac{a + 2b + 3c + 4d...}{a + b + c + d...}$$

lo que implica la minuciosa tarea de contabilizar uno a uno todos los descriptores componentes del tesauro. En cualquier caso es aconsejable calcular dichos índices sin recurrir al muestreo para conocer con exactitud los valores y, por extensión, su incidencia en el proceso de recuperación.

Los resultados de un reciente trabajo experimental[7] realizado con una muestra de seis tesauros españoles (tabla I), con objeto de comprobar su capacidad para recuperar información relevante, pusieron de manifiesto un predominio de bajos niveles de relaciones de equivalencia, así como un escaso índice de notas aclaratorias en casi todos ellos y, como consecuencia, un alto índice de ambigüedad. En cuanto a la tasa de enriquecimiento, es decir, la existencia de relaciones jerárquicas y asociativas entre los descriptores, la mitad de los tesauros analizados (Biología animal, Tesauro Electrotécnico y Tesauro Mujer) no alcanzan el mínimo aconsejado, entrando los otros tres dentro de los límites. La tasa de precoordinación es adecuada en todos excepto en el de Biología animal. La baja tasa que caracteriza a dicho lenguaje podría justificarse por el carácter unívoco de la terminología científica que lo compone.

Por último, se puede confirmar la existencia de consistencia interna en todos los tesauros estudiados, esto es, existe reciprocidad en las relaciones sintagmáticas y paradigmáticas, y hay uniformidad de criterio al emplear la forma y el número de las palabras.

Tabla I

	Relaciones de equivalencia	*Tasa de enriquecimiento*	*Notas aclaratorias*	*Tasa de precoordinación*
Biología animal	0,3	0,3	3	1,36
Electrotécnico	0,1	1,35	3,7	1,9
Medio ambiente	0,26	2,5	2,8	1,68
Mujer	0,56	1,32	17	1,5
Psicología	0,57	2,9	26	1,63
Asuntos sociales	0,25	2,09	1,4	1,54
Valores aconsejados	**0,5-2**	**2-5**	**10**	**1,5-2**

[7] GIL URDICIAIN, B.: *Evolución histórica de los tesauros españoles y análisis de su rendimiento en el proceso de recuperación de información.* Tesis doctoral dirigida por José López Yepes. Madrid, Universidad Complutense, 1997.

Las dos conclusiones más significativas que se pueden extraer del mencionado trabajo son, por una parte, que cuanto más alto es el nivel de precoordinación del tesauro, se consiguen resultados con mejores índices de precisión. Asimismo la precoordinación de los términos del lenguaje documental permite definir los temas de las preguntas y, por lo tanto, recuperar menos documentos irrelevantes. Por otra parte, el uso de un tesauro inadecuadamente específico, es decir, con una tendencia a excluir términos genéricos, produce resultados con bajo nivel de relevancia y afecta a la exhaustividad en aquellos casos en los que no hay términos para describir conceptos significativos.

Para terminar, quisiera apuntar que cuando se comparan los resultados de búsquedas realizadas en lenguaje libre y controlado se perciben las desventajas de uno y otro. Prescindir del controlado parece de todo punto imposible si se quiere aportar agilidad y precisión a la tarea, pero de dicha comparación también se deduce la existencia de casos en los que el lenguaje controlado no consigue documentos relevantes por la imposibilidad de su puesta al día permanente. De todo ello se puede concluir que uno y otro son complementarios. Dado que no todas las bases de datos ponen en práctica búsquedas combinadas con ambos procedimientos, sería aconsejable construir tesauros capaces de sumar las ventajas que ofrece el lenguaje natural en el momento de construir estrategias de búsqueda. En esa línea se está investigando actualmente en la Universidad de Sevilla, en donde se acaba de realizar un tesauro basado por completo en la postcoordinación, principio en el que se funda el lenguaje libre. Se trata del Tesauro de Patrimonio, editado por la Consejería de Cultura del Instituto Andaluz del Patrimonio Histórico y coordinado por Antonio García Gutiérrez. Este lenguaje documental, para cuya elaboración se ha aplicado como metodología básica la gramática de casos, está compuesto por 16.000 términos. Quizá éste sólo sea el comienzo de una nueva tendencia en el desarrollo de lenguajes documentales, que permita cubrir la necesidad de útiles cada vez más efectivos y adaptables a búsquedas interactivas. De hecho, el tesauro al que nos referimos no es más que la primera puesta en práctica de lo que García Gutiérrez denomina lenguaje epistemográfico, cuyas bases teóricas se recogen en su obra: *Principios del lenguaje epistemográfico: la representación del conocimiento sobre patrimonio histórico andaluz*. En ella se hace una crítica de la norma *UNE-50-106-90*, equivalente a *ISO 2788-1986, Directrices para el establecimiento y desarrollo de tesauros monolingües* y se propone un lenguaje asociativo, con nuevas estructuras —se considera el tesauro como un texto— acordes con la tecnología que existe actualmente.

Tal vez esto suponga un paso adelante en el acercamiento del lenguaje libre y el controlado, lo cual, sumado a la creciente incorporación de interfaces cada vez de más fácil manejo, beneficiaría al usuario, desconocedor de las técnicas documentales y, como dice el profesor Félix del Valle, *acostumbrado a consultar sistemas de análisis y recuperación de documentos sobre documentación diseñados por documentalistas para ser utilizados por documentalistas.*