Entropía, relevancia y pertinencia del descriptor en el mensaje comunicativo-documental. Nota tipológica

ANTONIO LUIS GARCIA GUTIERREZ
Profesor de Documentación
Facultad de Ciencias de la Información
Universidad Complutense de Madrid

INTRODUCCION

El proceso documental es un proceso comunicativo en el que intervienen un sujeto emisor, un canal, un mensaje y un receptor[1]. A diferencia de la lengua natural utilizada en la comunicación humana, la documentación utiliza un lenguaje controlado[2], creado por el hombre, capaz de generar diálogo (hombre, fichero u ordenador) a un subnivel comunicativo. De hecho, Susan Artandi, ratificada por Landry, considera el producto de la indización como mensaje y, en consecuencia, como objeto de estudio de las Ciencias de la Comunicación al señalar que la indización implica la comprensión del significado de los mensajes y la expresión de esta comprensión en la forma de nuevos mensajes (índices)[3]. En este proceso, el indizador, hombre o máquina, funciona como fuente y destino, como codificador y decodificador de mensajes.

Los lenguajes documentales provienen del lenguaje humano aun en el caso de artificialización de sus signos y grafemas. De hecho el lenguaje humano se hace permanente *en el* documento, mientras que

---

[1] Así consta en la obra del Dr. López Yepes, *Teoría de la Documentación*. Pamplona, EUNSA, 1978, pág. 132, y de Desantos Guanter, J. M., *La documentación científica como objeto de la información*, «Boletín del Fondo para la investigación económica y social», vol. II, fasc. 3, abril-junio, 1970, págs. 165-175.

[2] El lenguaje humano evoluciona por impulsos inconscientes psicosociales, mientras que el lenguaje controlado es codificado conscientemente por el hombre.

[3] Artandi, Susan: *Machine indexing: linguistic and Semiotic implications*, en «Journal of the ASIS», julio-agosto, 1976, pág. 236. Por su parte, Landry considera que la fuerza motriz generadora de diálogo documental necesita de un proceso intelectual que implica «reflexión sobre el orden de los conceptos que existen en el flujo informativo y debería servir para identificar la interrelación entre ellos». Véase el trabajo de Landry, B. C., y Rush, J. E.: *Toward a theory of indexing*, «Proceedings of the ASIS», vol. 5, 1968, pág. 62.

el lenguaje documental no es más que el medio de expresión *de los* documentos. Quiere esto decir que los documentos tienen sus códigos identificatorios, se hacen independientes y universales con el único límite de sus contenidos. Es labor humana manipular con las posibilidades lingüísticas de los colectivos documentales y crear, mediante codificación, términos controlados para la comunicación documental, es decir, descriptores. Estos elementos léxico-documentales permitirán el diálogo con los depósitos de documentos, manuales o automáticos, gracias a combinaciones potenciales de ellos, o frases documentales, las cuales constituyen el núcleo del lenguaje documental. Sólo estructurando una eficaz teoría de la indización (proceso intelectual de extracción de conceptos) podrá accederse a campos conceptuales y terminológicos óptimos que permitan construir lenguajes documentales útiles[4]. En este sentido el Thesaurus aparece hoy como una alternativa válida[5] para el control de documentos en áreas científicas y técnicas.

1. LA INDIZACION POR DESCRIPTORES: ELEMENTOS LINGUISTICOS Y COMUNICATIVOS PARA UN CONCEPTO

No es posible hablar de descriptores sin hacer referencia al thesaurus, pues aquéllos no tendrían operatividad fuera de contextos semánticos controlados. Sin embargo, el término es antiguo como afirma el brasileño Silva: descriptor es una vieja voz latina, en desuso en la lengua común, pero existente en los diccionarios españoles, portugueses, franceses e ingleses[6]. No obstante, fue Mooers[7], matizando el método creado por Taube en 1955[8], el primero en utilizar la denominación descriptor para referirse al producto de la indización basada en conceptos, aunque ciertos autores, como Perreault, rechazan esta denominación desde una concepción lógico-lingüística[9].

---

[4] Véanse los aspectos teóricos de la indización en el contexto de la lingüística documental en mi reciente trabajo: *Normalización de la Documentación informativa. Propuesta de Tesauro español de las Ciencias de la Información.* Madrid, Universidad Complutense, 1982, XIII, 1293 págs.

[5] La creación de lenguajes de estructura gramatical como el Syntol, Semantic Code, etcétera, a nivel internacional y en campos científicos de terminología inestable es una empresa inviable en la actualidad.

[6] Silva, Benedicto: *Origem e evoluçao dos descritores,* Río de Janeiro, Fundaçao Getúlio Vargas, 1972, pag. 27.

[7] Mooers, Calvin: *Zatocoding Applied to Mechanical Organization of Knowledge,* «American Documentation», vol. 2, núm. 1, 1951.

[8] Taube, Mortimer: *The Uniterm System of Indexing Operating Manual.* Washington, Documentation Inc., 1955, 47 pág.

[9] Perreault realiza una matización muy razonada sobre el término descriptor: «Si un vocabulario está controlado, éste adscribe conceptos al documento, por tanto, la indización por conceptos controlados produce adscriptores. Si el vocabulario es libre, se

Por otra parte, diversos investigadores como Chastinet[10] o Fondin[11] proponen definiciones y enfoques aislados del concepto de descriptor, el cual, a nuestro entender, no puede ser contemplado más que desde el campo linguo-documental, e inserto en mensajes documentales (provenientes de la indización en el análisis o en la recuperación), lo que nos permite definirlo como la *unidad significativa mínima del mensaje documental.* El thesaurus como marco léxico del descriptor es, por tanto, un *dispositivo que correlaciona los mensajes humanos y los documentales.*

Un solo descriptor puede constituir frase documental aunque la mayoría de los sistemas de indización atribuyen varios descriptores a cada documento. Esta afirmación está en directa relación con el concepto de profundidad de la indización. Según Bird, «la profundidad depende del número de conceptor indizables en el documento. En los diferentes sistemas de indización, los descriptores son generalizados o especificados en sus contenidos, pueden ser recuperables simplemente o en combinación mediante la lógica booleana. Todos estos factores más el sistema al que se aplican influirán en las normas de indización»[12]. Sobre la misma cuestión, Bradford, comentando a Alexander Pope, advierte que las palabras son como hojas y cuanto más abundantes menos se encuentra bajo ellas el punto del sentido[13].

Estas máximas hacen pensar en un excesivo subjetivismo en la indización que escapa al control del lenguaje documental. Sin embargo, es posible establecer unas conclusiones flexibles que permitan la coordinación multilateral en la tarea de los indizadores sobre todo cuando se vislumbran posibilidades de automatización, proceso que exije un alto nivel de normalización. Sin duda, la existencia de un

---

indiza describiendo el documento, y se producen los descriptores. La clasificación no describe sino que prescribe una posición en una organización conceptual e inscribe el documento en esa posición. La clasificación supone «scripción» pero no de los elementos del documento, sino de los documentos como totalidad. En resumen, tanto la clasificación como la indización «scriben» pero la primera prescribe (a la clase) e inscribe (de la clase); mientras que la segunda adscribe o describe; por tanto, desde una concepción lógico-lingüística, descriptor es el término proveniente de una indización en lenguaje libre», *Documentary relevance and Structural Hierarchy,* «American Documentation», julio, 1966, pág. 137.

[10] Chastinet Duarte, Yone: *Uso do kwic em Indexaçao bibliográfica. Curso de introduçao à tecnologia dos descritores.* Río de Janeiro, Fundaçao Getúlio Vargas, 1970, 8 folios.

[11] Fondin, Hubert: *La structure et le vocabulaire de l'analyse documentaire. Contribution pour une mise au point,* «Documentaliste», vol. 14, núm. 2, 1977, págs. 11 y ss.

[12] Bird, P. R.: *The Distribution of Indexing Depth in Documentation Systems,* «Journal of Documentation», vol. 30, núm. 4, 1974, pág. 381. El autor analiza en 57.000 documentos la distribución y estabilidad de los descriptores partiendo de la fórmula de Poisson y concluyendo que es preciso acudir a la descripción matemática de la indización para obtener el grado de dispersión de descriptores en un sistema.

[13] Bradford, S. C.: *Documentaçao,* 2.ª ed. Prólogo de Donker Duyris. Traducción de M. E. de Mello. Río de Janeiro, 1961, pág. 79.

Thesaurus es el primer paso para aproximarse a la univocidad de criterios en la indización.

2. PRECOORDINACION DE DESCRIPTORES: LA FRASE DOCUMENTAL

La precoordinación de descriptores es un problema relacionado con la profundidad del propio descriptor, con el número de ellos y con sus combinaciones a la hora de recuperar información. La precoordinación es beneficiosa cuando se trata de eliminar aisladamente la ambigüedad de un concepto y perjudicial cuando se realiza a priori con el fin de reducir las falsas combinaciones de la búsqueda. Sin embargo, su relativa utilidad puede variar según se trata de recuperación manual o automática.

En el caso de sistemas automáticos el proceso de poscoordinación es el más eficaz, ya que es la máquina la que dota de profundidad a la frase documental (por intersección booleana, etc.), a pesar de la posibilidad de falsas combinaciones. Pero en la recuperación manual, como, por ejemplo, índices simples, permutados y acumulativos que aparecen al final de la bibliografía, la precoordinación es el único medio de encontrar la referencia concreta. En definitiva, la poscoordinación de descriptores no es útil en la búsqueda manual de documentación.

Ejemplo:

— Descriptor compuesto para eliminar su propia ambigüedad.
/centro/+/información/=/centro de información/
— Descriptores simples no ambiguos a poscoordinar por la máquina.
/cine/+/España/+/cámara/+/dirección/+/productora/
— Descriptores precoordinados en el índice tradicional
/cine español/
/cámara cinematográfica/
/productora/

Sin embargo, al elaborar un lenguaje controlado es preciso prever la automatización, por lo cual es recomendable utilizar descriptores no precoordinados sólo en caso de irrelevancia o ambigüedad, que aparezcan descompuestos en las referencias y en las demandas y sólo sintagmatizados en los índices manuales. Aquellos descriptores que no sean lo suficientemente significativos como para situarse en campos facetados y concretos del Thesaurus deberían encontrarse en índices y tablas auxiliares de ese lenguaje documental.

3. TIPOLOGIA DE DESCRIPTORES REFERENCIALES Y EN THESAURUS

Además de los conocidos descriptores unitérminos y sintagmáticos o compuestos, las anteriores reflexiones nos llevan a la propuesta de la siguiente tipología:

— Descriptor primario: aquel concepto con significado propio incluso fuera de contexto o de frase documental.
Ej./comunicación de masas/, /persuasión/, etc.

— Descriptor secundario o auxiliar: aquel concepto ambiguo, genérico o no pertinente a la estructura del Thesaurus. Debe encontrarse en índices y tablas, auxiliares del mismo Thesaurus en orden alfabético permutado. Es preciso distinguir dos tipos de descriptores secundarios: los causados por ambigüedad.
Ej. /modelo/, /centro/ y por homonimia, ej. /índice/, /bibliografía/, en este segundo caso pueden indicar que el documento es una bibliografía o un estudio sobre el concepto de bibliografía, índice, etc. En consecuencia, los descriptores secundarios deben cumplir dos requisitos:
  1.º Aparecer en distinta grafía o posición determinada (por ejemplo en cursiva) en la bibliografías impresas.
  2.º Aparecer con características diferenciadoras de los descriptores principales en las bibliografías automatizadas mediante recursos previstos en el lenguaje utilizado.
Ambas posibilidades no son más que un intento de gramaticalización de la frase documental con objeto de reducir la distorsión en la recuperación.

— Descriptor informativo: son descriptores que pueden no estar en el Thesaurus, en los índices de las bibliografías, pero que son útiles en el contexto de descriptores que acompañan la referencia. Por ejemplo, /Ecología/ o /concepto/ en una bibliografía sobre cine. Aisladamente inducen a confusión, pero en el contexto de la referencia aclaran la frase documental. Todos los descriptores son informativos aunque pueden ser no pertinentes con respecto a determinados campos (dispersión).

— Descriptor referencial o direccional: son aquellos descriptores que permiten la localización de las referencias de forma colectiva mediante índices, demandas indizadas, etc. Todos ellos deberán formar parte del Thesaurus o aparecer en las listas de candidatos para su inclusión.

Desde el punto de vista de la elaboración del Thesaurus (thesaurificación) sea usando un método analítico (indización de muestra aleatoria de documentos) o global[14] (extracción de descriptores de diccionа-

[14] Sobre las diferentes metodologías véase Iso: *Príncipes directeurs pour l'établissement et le développement de thesaurus monolingues.* Genève, 1974, III, 14 págs. 502-788.

rios y terminologías especializadas) obtendremos una nueva tipología de descriptores.

1.º Descriptores confirmados: ratificados por la indización. Su uso posibilita la recuperación real de documentos que existen.

2.º Descriptores preventivos o potenciales: su función es prever cualquier análisis o búsqueda potencial, ya que no han sido extraídos de los documentos científicos existentes. Sin embargo, estos descriptores son muy eficaces para que el Thesaurus cumpla su función de ampliación o focalización en la indización. Según este criterio tendremos *descriptores preventivos de nivelación,* cuando su fin es cubrir lagunas o establecer puentes entre encabezamientos, y *descriptores de ampliación,* cuando su fin es ampliar el campo de descriptores confirmados.

4. RELEVANCIA, PERTINENCIA Y ENTROPIA DE LA INDIZACION

Hay que tener también presente el concepto de entropía documental referido a la cantidad de información aportada por un descriptor o por la frase documental con el menor número de signos. No parece recomendable utilizar, al indizar, exclusivamente el descriptor más ajustado al tema que trata el documento, ya que esto impide recuperar ese mismo documento usando criterios de búsqueda más amplios. En consecuencia, no caeríamos en excesiva redundancia, fenómeno opuesto a la entropía y economía lingüísticas, si empleamos en determinados casos el concepto inmediatamente más genérico según debe indicar el Thesaurus. Dentro del campo entrópico de un descriptor pueden estudiarse diversos niveles de comportamiento relacionados con la profundidad conceptual, la relevancia y la pertinencia.

El concepto de profundidad de la indización va ligado con lo que denomina Maron amplitud terminológica. El autor dota a esta última noción de dos interpretaciones: una *intensional,* basada en el significado de los términos en cuestión (exploración semántica) y otra *extensional,* relativa al número de documentos que el mismo término puede indizar. Maron indica que la amplitud semántica de un término (más generalidad) no guarda relación con el número de documentos que abarca (lo cual ya fue señalado por Jones)[15]. Incluso, «términos más específicos pueden indizar más documentos que los genéricos», lo cual, según Maron, no es paradójico[16]. De acuerdo con la vertiente extensional, mencionada por el autor, la profundidad se refiere al

[15] Jones, K. P.: *Towards a Theory of Indexing,* «Journal of Documentation», vol. 32, núm. 2, 1976, pág. 118.

[16] Maron, M. E.: *Depth of Indexing,* «Journal of the ASIS», julio, 1979, pág. 224.

número de términos asignados a un documento[17] y si se asignan muchos términos a un documento se dice que éste estará indizado más profundamente.

Por otra parte, y siguiendo la terminología del mismo investigador, desde el punto de vista intensional, «aquel documento indizado con términos más específicos estará intensionalmente indizado con más profundidad que el que contenga términos más genéricos»[18]. Basándose en los dos principios mencionados, Maron intenta definir el nivel óptimo de indización analizando el número medio de términos de indización por campos científicos y por documentos individuales, a partir del estudio de una colección de documentos y su potencial demanda por la comunidad de usuarios. Tras el estudio puede concluirse que un buen sistema de recuperación debe extraer sólo documentos relevantes para la demanda, usando una indización precisa, lo cual ha de ser determinado por una estricta selección, instrumentación rigurosa de normas de indización a hombres u ordenador, y encuestas frecuentes para averiguar las necesidades de la demanda.

Dentro de la importante división maroniana, se ha mencionado el concepto de *relevancia* documental. El mismo Landry considera el documento desde el punto de vista de la indización como una «colección de unos cuantos conceptos relevantes»[19]. Pensamos que la indización debe establecer el orden de la relevancia de esos conceptos, recortados por los diversos niveles de la profundidad de su aplicación. Perreault cita a Hillman como uno de los mayores teóricos de la relevancia documental[20]. No obstante, si Hillman intenta precisar objetivamente la relevancia de un documento a través de la indización y de la demanda, Perreault, acertadamente, pone en duda ese argumento planteando el carácter subjetivista de la relevancia de los conceptos, hasta tal punto, que llega a afirmar lapidariamente que «una característica propia de la relevancia es que cualquier juicio sobre ella ya es valorativo y, por tanto, no objetivo»[21].

En nuestra opinión, la relevancia de un documento está en función de su carga entrópica, es decir, de la aportación de información unitaria que comporta y, consiguientemente, es preciso determinarla en el contexto donde se inscribe cada unidad documental. Sin embargo, debe analizarse la relevancia de un núcleo para valorar los documentos relativos a ese mismo campo, obtenido con posterioridad. Es

---

[17] Denominado por Sparck Jones «exhaustividad de la indización» en *Does Indexing Exhaustivity Matter?*, «Journal of the ASIS», vol. 24, núm. 5, 1973, págs. 313, 316.

[18] Maron, M. E.: *Op. cit.*

[19] Landry, B. C., etc.: *Toward..., Op. cit.*, pág. 59.

[20] Hillman, Donald J.: *The Notion of Relevance*, «American Documentation», vol. 15, 1964, págs. 26-34.

[21] Perreault, J. M.: *Documentary...*, op. cit., pág. 136.

importante, también, apreciar la diferencia entre información relevante y relevancia de la información. Esta última es fácilmente comprobable a través de los fondos documentales mediante tests y sondeos. Los índices de *ruido* (documentos extraídos que no interesan) y de *silencio* (documentos existentes que no han podido ser extraídos por una deficiente indización) se oponen a la relevancia de la indización y, por tanto, la cantidad de información relevante obtenida por el usuario no coincide con la depurada por el sistema documental.

Sin embargo, el subjetivismo impreso en el concepto de relevancia de los documentos es limado por la inercia del colectivo documental y terminológico en cada campo científico. Si un indizador duda ante la elección o relevancia de un término, habrá de mantener cautela y observar los impulsos de los contenidos documentales sucesivos. El avance lógico del cuerpo investigador debe resolver tajante y mecánicamente lo que en un momento dado puede comportar dudas inevitables. Por tanto, una actitud a mantener a la hora de establecer la relevancia de un descriptor es la expectativa,una vez superada la fase predictiva que extrae el término en bruto del grueso del campo conceptual.

En ese mismo sentido se desgrana el término *pertinencia* de la indización. Un descriptor es pertinente, en principio, paralelamente al grado de adecuidad del documento que lo contiene. Sin embargo, aquí se plantea el problema de la *dispersión* de descriptores. En efecto,del mismo modo que existe una dispersión de los artículos científicos entre las publicaciones de temas relacionados o diferentes (dispersión controlable, por otra parte, según Bradford por su ley de los círculos concéntricos donde se disponen los artículos en orden decreciente de pertinencia o con el tema de las revistas mientras éstas crecen exponencialmente)[22], los descriptores pueden aparecer, con mayor facilidad, en contextos de poca relación con los conceptos que representan. Visto a la inversa, podemos encontrar en documentos especializados ciertas palabras claves de gran relevancia pero de dudosa pertinencia. ¿Cómo saber la vinculación de ese potencial descriptor con el tema que nos incumbe? La solución viene dada por el procedimiento de la observación de frecuencia de aparición. Inevitablemente si una palabra clave aparece en diversas ocasiones en documentos afines, debe ser considerada como probable descriptor de ese campo documental, a pesar de los probablemente justificados criterios de los documentalistas.

Sin embargo, hay que concluir que es condición indispensable para

[22] Bradford, S. L.: *Documentation*, 2.ª ed. London, Crosby Lockwood, 1952, págs. 49-55. Cit. por López Yepes, José: *Bibliometría*, «Cuadernos de trabajo del Departamento de Documentación», vol. II, núm. 3, 1980, págs. 36-40.

estudiar la pertinencia de una palabra clave, el haber demostrado previamente su índice de relevancia ad hoc, de acuerdo al nivel de profundidad establecido como norma en la indización de un determinado campo. Existe una directa relación entre las cotas de relevancia y pertinencia alcanzada por los descriptores y la eficacia de éstos en la recuperación. El índice de pertinencia debe ser definido en función de dos dimensiones cartesianas: la primera *horizontal* y desdoblada en un factor de relación que indica la consonancia del descriptor con respecto a un campo científico y en un factor de inclusión que lo sumerge en ese campo, y la segunda *vertical* relacionada con el índice de profundidad que define el grado de amplitud sémica del descriptor. Además hay que definir el índice de relevancia que no puede establecerse por comparación como el anterior, sino por medición objetiva de su carga semántica.

CONCLUSIONES

La documentación plantea los problemas clásicos de los procesos comunicativos. De hecho su principal obstáculo y objetivo es la *transmisión o circulación de información científica. Se hace necesaria* la utilización de criterios únicos de indización para facilitar el flujo de información científica, sobre todo en campos de terminología poco unívoca e inestable, como el de la comunicación de masas (con la excepción de los aspectos tecnológicos de la comunicación).

En base a las reflexiones precedentes puede afirmarse lo siguiente:

1.º Las técnicas de indización son diferentes según se destinen a procedimientos de recuperación manuales o automáticos. Sin embargo es posible llegar a una conciliación que haga compatible el uso combinado de ambos métodos.

2.º En caso de recuperación manual no es perjudicial el uso de descriptores precoordinados en la indización, y en búsqueda mediante ordenador es más efectiva la poscoordinación.

Tanto en el Thesaurus empleado como en las referencias bibliográficas los descriptores deben ser unitérminos en la medida de lo posible. En los índices, la sintagmatización permite recuperar sin ambigüedad. Sin embargo, y en general, los términos compuestos sólo deben admitirse para evitar la ambigüedad del concepto y no de la demanda.

3.º Deben establecerse medidas tendentes a eliminar la ambigüedad en las frases documentales y la posibilidad de falsas combinaciones. Es deseable, por tanto, llegar a una base mínima de gramaticalización de la indización (método posicional, diferente grafía, etc.).

4.º Los descriptores secundarios deben utilizarse para matizar el *sentido de los primarios. Sin embargo, no deben aparecer en la*

estructura facetada del Thesaurus sino en tablas alfabéticas permutadas y auxiliares. Tampoco son útiles si aparecen aislados en los índices manuales de las bibliografías.

5.º En el caso de no existir un descriptor en el Thesaurus común utilizado por un sistema internacional es conveniente, si así lo precisa la indización, emplear el descriptor nacional o regional pero acompañado de un genérico o asociado (BT o RT) presente en el Thesaurus con el fin de facilitar su recuperación a otros centros. Los descriptores candidatos deberán ser sometidos a un estudio de frecuencia y propuestos en las revisiones del lenguaje común.